Instituto de Computación - Facultad de Ingeniería - Universidad de la República

Montevideo, Uruguay, 2016

Proyecto de Grado

Relevamiento y obtención de datos sobre emergencias en Uruguay para análisis predictivo

Tutores Libertad Tansini

Sandro Moscatelli

Ignacio Chiazzo

Felipe Garcia

Guillermo Leopold

1 Resumen

Uruguay, a pesar de ser uno de los países más estables en materia de desastres naturales, sufre de manera recurrente de inundaciones dentro de su territorio, los cuales repercuten negativamente principalmente en su población, pero también en su economía debido a pérdidas materiales. Por lo tanto, resulta deseable poder predecirlas con la mayor exactitud posible y minimizar sus consecuencias a todo nivel.

Los datos públicos que el estado Uruguayo expone sobre meteorología y el estado de sus corrientes fluviales no son en todos los casos claros ni consistentes, y en algunos, casi inexistentes. Interesa entonces investigar sobre su disponibilidad y la posibilidad de consumirlos y almacenarlos en un data warehouse central con una calidad aceptable para su posterior uso en análisis predictivo.

Con esas dos premisas, se plantea el siguiente trabajo como proyecto de grado divido en tres objetivos principales a cumplir. En una primera etapa, el objetivo es el de realizar una investigación sobre los datos públicos disponibles y explorar y establecer medios para obtener aquellos no disponibles, tanto históricos como del presente de manera continua.

Como segunda etapa, se propone procesar dichos datos para darles una mayor calidad calidad y almacenarlos de manera adecuada y de que resulten útiles de alguna manera para asistir decisiones. Se logró una base de datos muy amplia, con más de un millón de registros concernientes a condiciones climatológicas que afectan a los eventos de inundación en el período 1983-2014.

En la última etapa del proyecto se investigan técnicas de predicción en base a estadística, llamadas de aprendizaje automático, mostrando el potencial de los datos obtenidos y procesados a la hora de monitorear este tipo de desastres. Luego de las pruebas se seleccionó el algoritmos de clasificación SVC con kernel RBF por tener mejor precisión y dar menor error en la mayoría de los casos de prueba.

Para desplegar geolocalizadamente éstas predicciones y gestionar los datos almacenados, se desarrolla además una pequeña aplicación web con un sistema de información geográfica integrado para mostrar de manera gráfica las posibles inundaciones en el territorio de alguno de los 19 departamentos por separado y con la posibilidad de una ejecución aplicada a todo el país con el fin de obtener un panorama general del territorio nacional.

Palabras Clave

Desastres Naturales, Inundación, Riesgo, Predicción, Aprendizaje Automático, SIG (sistema de información geográfica), Data Warehouse, Obtención de datos, Limpieza de datos, Calidad de datos

2 Contenido

3 Introducción

3.1 Antecedentes y Motivación

El presente proyecto se encuentra dividido en tres grandes etapas entre las cuales existe una relación de dependencia. La primera, sobre la obtención de información vinculada a desastres y en particular aquellos de inundaciones; la segunda, se basa en el procesamiento de esos datos y la última etapa se enfoca en el análisis de los mismos con el fin de mostrar el potencial de los datos para realizar predicciones sobre futuras inundaciones.

En cuanto a la obtención de los datos, la principal motivación del proyecto es obtener un panorama general sobre el actual estado de la información pública y privada sobre la temática, y finalmente, la implementación de una solución que permita, de manera continua, alimentar de información (datos) un sistema de datawarehouse centralizado a la última etapa: un análisis predictivo que arroje resultados sobre la posibilidad de ocurrencia de inundaciones. Para eso, los datos obtenidos tienen que pasar por una etapa de procesamiento y limpieza.

Con respecto a la tercer etapa, sobre predicción de futuras inundaciones, se sabe que los modelos hidrológicos existentes que atacan este problema, que involucran flujos hidráulicos, son complejos e involucran un conjunto muy grande de variables vinculadas a la situación geográfica, climática y demográfica de cada curso de agua. Así, estos modelos no resultan portables a otros problemas ya que quedan fuertemente vinculados, debido al nivel de especificación de sus parámetros, a un único curso de agua. Es así que surge la necesidad de investigar un enfoque de solución más accesible y con datos para el problema.

3.2 Objetivos

Dada la motivación planteada en la sección anterior, el objetivo de este trabajo es obtener información histórica lo más completa posible sobre los factores influyentes en desastres de inundación, y establecer mecanismos automáticos de acceso y almacenamiento de más información del mismo tipo, pero contemporánea, con el fin de mantener actualizada la base de información.

Además, investigar y presentar un camino alternativo a los modelos hidrológicos, mediante un enfoque de estadística y aprendizaje automático para realizar un análisis predictivo sobre estos datos que arroje resultados sobre los potenciales eventos de inundaciones, y finalmente presentarlos de manera amigable e intuitiva sobre un mapa para un usuario final mediante una aplicación web.

3.3 Organización del Documento

El presente documento se encuentra dividido en capítulos de forma de organizar su contenido.

El primer capítulo consiste de un breve **resumen** que intenta explicar la extensión del trabajo realizado. Seguido de un **índice** de contenidos que nos guía dentro del informe.

El Capítulo 3 es una **introducción** más detallada sobre el contexto del trabajo, su alcance y objetivos.

En el Capítulo 4 se presenta un resumen del **estado del arte**, el cual plasma el conocimiento adquirido por el grupo sobre la problemática enfrentada, la situación a nivel nacional y un breve contexto internacional, áreas de la computación estudiadas para su posible aplicación en el proyecto, trabajos similares y finalmente, dada la relevancia que tiene para el proyecto, un estudio de la situación de los datos públicos centrando en nuestro país.

En el Capítulo 5, luego de haber estudiado la problemática, su contexto y las herramientas a nuestro alcance para trabajar en una solución, se explica la propuesta para resolver los objetivos planteados. Desde la primera hasta la última etapa, describe el proceso, requerimiento, análisis y diseño, problemas encontrados, decisiones tomadas y una conclusión de implementación para cada etapa definida del proyecto. Además en este capítulo se detallan y registran las pruebas realizadas sobre el trabajo descrito, con el fin de validarlo.

En el Capítulo 6 se expresan las **conclusiones** del equipo sobre el trabajo realizado y se añaden posibles trabajos a futuro que pueden extender o complementar el trabajo de este proyecto.

El Capítulo 7 contiene un **glosario** de términos empleados a lo largo del informe que se consideraron dignos de ser aclarados, y en el capítulo 8 se listan las **referencias** utilizadas en este documento.

Finalmente, se anexan a este informe dos documentos adicionales. El Anexo 1 es una versión extendida y completa del estado del arte, cuyo resumen se presenta en este informe en forma de Capítulo 4. El Anexo 2 es un **manual de usuario** para la aplicación web implementada bajo los lineamientos descritos en el Capítulo 5.

4 Resumen del Estado del Arte

En éste capítulo se presenta un resumen del estado del arte del presente proyecto, adjuntando como primer anexo su versión completa.

Se comienza estudiando el contexto general acerca de los desastres naturales, definiendo qué son y trasladando la temática a la región y principalmente a Uruguay. Luego, una vez que resulta claro que el desastre que más daños provoca en el país son las inundaciones, se realiza un estudio de las mismas, pasando por una definición de qué se considera una inundación, los diversos tipos, causas y consecuencias y un breve repaso cronológico de eventos de este tipo ocurridos en Uruguay. Finalmente en lo que se podría considerar como el final de esta primera parte del estado del arte, se adentra en lo que es la gestión de riesgos de desastres, brindando una definición de lo que es la misma, describiendo sus etapas, cómo se aplica, así como desarrollando en detalle su contexto y funcionamiento en nuestro país.

En lo que se puede considerar la segunda parte de este capítulo, ya que a partir de acá el estudio se centra en las tecnologías, se comienza con una explicación de lo que es un sistema de información geográfica y su funcionamiento y utilidad. Luego, se estudian algunos conceptos previos a lo que es el aprendizaje automático, investigando acerca de la minería de datos (o data mining), análisis predictivo y modelo predictivo; para luego ahondar en el aprendizaje automático terminando con las opciones consideradas a la hora de implementar un modelo predictivo.

Por último, en una tercer y última parte de este capítulo se explican y mencionan algunos proyectos/trabajos previos que tienen objetivos de características similares al de este proyecto, es decir trabajar sobre las inundaciones, en estos casos buscando anticipar las mismas y no sus consecuencias. Estos brindaron una primer idea de como funcionan este tipo de proyectos pese a tener ambos caracteristicas completamente diferentes considerando que uno es un proyecto de la NASA y el otro es un proyecto de grado de la Facultad de Ingeniería.

4.1 Desastres Naturales

Se entiende por desastres naturales aquellos desastres ocasionados como consecuencia de un fenómeno natural [1].

Por otra parte la ONU [2] define que no toda ocurrencia de un fenómeno natural desemboca en un desastre natural, sino que es el resultado de una falta de planificación y prevención así como de la presencia de la acción del hombre.

A grandes rasgos, un desastre es el resultado de una relación directa entre una amenaza (fenómeno natural en este caso) y la vulnerabilidad (exposición y falta de capacidad de tolerancia y fortaleza ante el daño) ante la misma ($desastre = amenaza * vulnerabilidad$).

Al centrarnos en las inundaciones, la frecuencia de las mismas, además de tener altos niveles de ocurrencia, parece incrementarse tanto en incidencia como en intensidad en los últimos años. Según estimaciones de la CEPAL (Comisión Económica para América Latina y el Caribe) [3], en la región más de 150 millones de habitantes se han visto afectados por los desastres y más de 310.000 han fallecido como consecuencia de los mismos, además de haberse generado 30 millones de damnificados directos; con un monto total de daños acumulados estimado en unos 213.000 millones de dólares (datos válido para el año 2000), tomando como punto de partida el terremoto de Managua, capital de Nicaragua, en 1972 [1].

Finalmente, con el fin de reducir los efectos a largo plazo de los desastres, las acciones deben enfocarse en dos frentes paralelos según ONU [4]:

- En la previsión de un evento desastroso, teniendo en cuenta la asignación de recursos para la prevención y mitigación del impacto.
- Por otra parte, en asegurar que lo invertido a la reconstrucción, sea con miras a una reducción de la vulnerabilidad, garantizando un desarrollo sustentable.

EM-DAT International disaster database, centro de Investigación Epistemológico de Desastres

El Centro de Investigación Epistemológico de Desastres es un departamento de de la Universidad Católica de Louvain (UCL) que se ubica en Bruselas, Bélgica [5].

Dicho centro brinda públicamente su base de datos de desastres llamada EM-DAT, la cual es una base de datos global que almacena desastres naturales.

EM-DAT tiene como objetivos principales ayudar a la acción humanitaria tanto a nivel nacional como internacional, a racionalizar acerca de la toma de decisiones para la preparación de posibles desastres y proporcionar una base de datos objetiva para la evaluación de la vulnerabilidad y establecimiento de prioridades [6].

DesInventar (Inventario de desastres)

Surge con el fin de tener una base de datos normalizada sobre desastres naturales en América Latina, el Caribe, Asia y África, la Red de Estudios Sociales en Prevención de Desastres de América Latina desarrolló el Sistema de Inventario de los Efectos de los Desastres (DesInventar) en 1994 [7].

Actualmente, a pedido del SINAE, Uruguay se encuentra incluido entre los 30 países [8] que cuentan con una base de datos nacional de DesInventar.

4.1.1 Desastres Naturales en Uruguay

En nuestro país, desde finales de la década del 50 cuando sucedió el primer gran desastre natural, las inundaciones de 1959, el registro se ha llevado a cabo con mayor seriedad, registrándose números elevados tanto en pérdidas humanas como en población afectada.

Considerando los registros comprendidos entre 1967 y 2014, los desastres que más personas han afectado son las inundaciones, con 224263 afectados y las tormentas, con 200 personas; mientras que los desastres que más pérdidas humanas ocasionaron son las inundaciones con 23 pérdidas y las tormentas y temperaturas extremas con 11 cada una. En cuanto al apartado económico, las pérdidas por sequías han sido un total de 250 millones de dólares, 89 millones por inundaciones y 25 millones por tormentas [9].

En balance general, las inundaciones han sido el desastre que más ha afectado a la región y a nuestro país. Para reflejar esto de manera objetiva se recomienda estudiar las tablas de la sección 2.1 del Anexo 1; en las mismas se ven los daños de manera numérica considerando personas afectadas y daños económicos de los mayores desastres naturales registrados en Uruguay.

4.2 Inundaciones

Según el glosario internacional de hidrología (OMM/UNESCO, 1974), una inundación es el “aumento del nivel del agua por encima del nivel normal del cauce”. En este contexto, se entiende por nivel normal a la elevación de la superficie del agua tal que no provoca daños en sus alrededores [10].

Por otra parte, SINAE define una inundación de la siguiente manera: "*Una inundación es el avance de las aguas sobre zonas que habitualmente están secas. Puede producirse por el desborde de ríos, lagos y embalses a causa de lluvias torrenciales o por la rotura de diques o presas*" [11].

4.2.1 Clasificación de las inundaciones

Debido a que las inundaciones poseen diferentes características, como son la duración, la intensidad, el origen o causa, el impacto, etc., es común ver distintas clasificaciones de las mismas.

Las clasificaciones más habituales son según origen y según el impacto generado. Dentro de la clasificación por origen, se pueden distinguir inundaciones pluviales, fluviales, costeras y causadas por catástrofes; por otra parte, al clasificarse por impacto, se distinguen tres niveles: inundación ordinaria, extraordinaria y catastrófica [12][13].

4.2.2 Principales causas de las inundaciones

Cuando se habla de inundaciones, es común asociar a las mismas con un resultado producto de grandes lluvias. Sin embargo, si bien es correcto en el sentido que las lluvias son la principal causante, esta no es el único causante, sino que son variados los motivos y pueden ser tanto naturales como producto de la acción del hombre.

Las principales causas de inundaciones son [14]:

Lluvias

La lluvia es la principal causa de las inundaciones. El exceso de lluvias provoca que el agua fluya sobre la tierra, contribuyendo así a las inundaciones. Generalmente son las que ocurren con una intensidad fuerte y tienen una duración prolongada las que causan inundaciones.

Desborde de Rios

Tanto ríos como arroyos pueden desbordar sus orillas. Esto ocurre cuando el río tiene más agua corriente arriba de lo que tiene normalmente y fluye hacia zonas bajas (zonas de llanura aluvial generalmente).

Inundaciones Costeras

Este tipo de inundaciones suceden como consecuencia de largas tormentas o fenómenos naturales como huracanes, ciclones, tsunamis que provocan que el “cuerpo” del agua se precipite hacia la tierra.

Rotura de Represas

Las inundaciones en estos casos suceden cuando la represa se encuentra debilitada en sus muros y terminan rompiéndose por exceso de cantidad de agua, superando la capacidad de la represa.

Derretimiento de los Glaciares

En regiones frías, durante el verano las grandes cantidades de hielo y nieve que se producen y acumulan durante el invierno comienzan a derretirse, produciendo así grandes movimientos de agua en tierra.

Urbanización

La urbanización de las ciudades, en general implica que el suelo se cubra con una capa de concreto o asfalto. Si esto se combina con malos drenajes o con drenajes obstruidos por basura, es muy factible que el agua de lluvia se acumule provocando inundaciones.

Deforestación

La tala de árboles provoca que la cobertura vegetal del suelo se pierda y que el agua de lluvias arrastren la tierra. Con el paso del tiempo, esta tierra puede obstruir arroyos y/o ríos cambiando el flujo del agua o provocando que la misma se acumule en puntos que no están preparados para soportarlo.

4.2.3 Efectos Negativos de las inundaciones

La ocurrencia de una inundación en una zona urbanizada puede tener un gran impacto en la sociedad. A continuación se clasifican los efectos de una inundación según su perdurabilidad en el tiempo.

En el corto plazo

Si bien la pérdida de vidas por inundación no es el escenario más común en nuestro país, pues por lo general ocurren de forma paulatina producto de las lluvias prolongadas, en aquellos lugares con riesgos de tsunamis o tormentas muy fuertes, este escenario no es tan irreal. La pérdida de vidas y las personas heridas son, sin lugar a dudas, la peor consecuencia posible de una inundación.

A nivel material, muchas veces, el nivel del agua aumenta con tal rapidez que las personas se ven obligadas a evacuar sus hogares, pudiéndose llevar consigo un porcentaje muy pequeño de sus pertenencias.

En el largo plazo

Por lo general, los daños en infraestructura posterior a una inundación implican corte de calles, días sin luz eléctrica en las partes afectadas por la inundación, etc. Esto supone un gran impacto en la economía del país para poder superar la situación y volver a funcionar con normalidad.

Las inundaciones pueden también causar daños en zonas agrícolas lo cual repercute directamente sobre el suministro de alimentos y en el caso particular del Uruguay, sobre una de las actividades económicas más desarrolladas y esenciales. Además, muchas veces el agua desplaza a roedores y serpientes hacia lugares en los que no es frecuente encontrarlos; esto supone peligro tanto para la salud los humanos como para los animales.

Finalmente no solo está en riesgo la salud por las lesiones y las mordeduras o picaduras de animales peligrosos, sino que es muy común que las fuentes de agua potable se contaminen con materiales tóxicos provocando enfermedades [15].

4.2.4 Inundaciones a nivel nacional

Las características de penillanura (suaves pendientes) del suelo de Uruguay hace que los cursos de agua sean propensos a no generar crecidas violentas además de resultar bastante predecibles considerando parámetros como el volumen de las lluvias y su intensidad en conjunto con el factor tiempo. Gracias a esto y a la experiencia (de inundaciones pasadas) adquirida en las zonas más propensas a ser afectadas por inundaciones, es posible en muchos casos realizar evacuaciones preventivas, logrando poner a resguardo a la potencial población afectada y sus bienes, así como ganar tiempo para tomar otras medidas preventivas [16].

El registro histórico con mayor número de personas evacuadas es del año 1959 cuando ocurrieron las inundaciones más importantes en la historia de nuestro país, en aquel entonces la cantidad de evacuados ascendió hasta casi 45.000 personas. Estas inundaciones duraron un mes entero entre fines de marzo y fines de abril, generando así un desastre a nivel nacional [17].

La misma perjudicó al país en su totalidad, teniendo consecuencias catastróficas como la caída de redes telefónicas enteras, alteraciones en el sistema de transporte y ocasionando serios problemas con el abastecimiento de energía eléctrica. Esto último fue debido a la situación particularmente grave que enfrentó la Represa del Rincón de Bonete, la cual es clave en la generación de energía eléctrica en Uruguay, siendo sobrepasada por las aguas y quedando fuera de servicio.

Hasta la actualidad, no se ha repetido una inundación de estas magnitudes, sin embargo si han habido otras grandes inundaciones, con importantes números de evacuados en el país, destacándose los nueve meses comprendidos entre 1997 y 1998 donde casi todo el litoral del Río Uruguay permaneció bajo las aguas como consecuencia de las precipitaciones ocasionadas por el fenómeno ENOS (Niño Oscilación Sur) [17].

Por otra parte, desde el comienzo del siglo XXI, se ha estado trabajando en medidas preventivas y de mitigación en zonas que suelen ser afectadas por las inundaciones como lo son los departamentos de Salto, Paysandú y Soriano, mientras que en departamentos como Artigas, Cerro Largo y Durazno, se trabajó además en la elaboración de mapas de riesgo para la relocalización de población en situación constante de ser “evacuados potenciales” [16].

Finalmente, como dato respecto al impacto reciente de las inundaciones en nuestro país, es remarcable el hecho de se registraron más de 67.000 personas evacuadas en la década pasada debido a este tipo de eventos, siendo por lo tanto el más frecuente y con mayor impacto en nuestro país [11].

4.3 Gestión de Riesgos

Antes de estudiar la gestión de riesgos a nivel nacional, resulta necesario comprender y manejar algunos conceptos previos de la misma y tener en consideración que la gestión de riesgos no acepta una única definición sino que hay múltiples definiciones, cada una aceptada por diferentes gobiernos y organizaciones.

Conceptos previos

Se define el riesgo en el contexto de la gestión de riesgos como la probabilidad de que ocurra un evento de consecuencias socioeconómicas o ambientales en un espacio de lugar y tiempo específico con cierta duración. Dicho riesgo está vinculado a desastres y se da donde haya una sociedad afectada de forma importante por el impacto de algún evento físico de diverso origen, como pueden ser las inundaciones. Este riesgo trae consigo además la irrupción en las vidas cotidianas de la población afectada.

Finalmente, en este contexto, se considera al riesgo como el resultado de una interacción entre la **amenaza** y la **vulnerabilidad** [18] y se puede expresar de la siguiente forma

$$Riesgo = Amenaza * Vulnerabilidad.$$

Se entiende por amenaza como la probabilidad de que un fenómeno de origen natural, socionatural o antrópico se presente con cierta intensidad en un sitio específico y dentro de un período de tiempo, con potencial de producir efectos adversos sobre las personas, los bienes y el medio ambiente [18].

La vulnerabilidad, por su parte, expresa las características y circunstancias de una comunidad, sistema o bien, que los vuelven susceptibles o predispuestos a padecer los efectos dañinos de una amenaza, tanto de origen natural como del hombre [18].

Definición de Gestión de Riesgos

Un primer acercamiento a lo que es la gestión de riesgos, aplicada a la temática de este proyecto, es decir que la misma utiliza estrategias para disminuir la vulnerabilidad y promover acciones de conservación, desarrollo, mitigación y prevención frente a desastres tanto naturales como antrópicos [18].

La gestión de riesgos implica la implementación de un conjunto de medidas que permitan conocer y medir todos los elementos vinculados a los riesgos con el único objetivo de poder enfrentarlos para minimizarlos, o en el mejor de los casos, anularlos. Los métodos a utilizar varían según el contexto y marco en el cual se realice su gestión [18].

Conceptualizando, a continuación se presentan algunas de las definiciones de lo que es gestión de riesgos (por más que algunas puedan tener mucho en común), ofreciendo así diferentes puntos de vista sobre la misma.

La ONU define a la gestión de riesgos como la organización, planificación y aplicación de medidas para la preparación, respuesta y recuperación ante desastres. Resaltando que aún así, la gestión de riesgos no garantiza la eliminación o la evasión completa de una amenaza, centrándose más bien en la creación e implementación de planes de preparación como de disminución del impacto de los desastres, buscando una recuperación más ágil [19].

Drabek define la gestión de riesgos como la disciplina y la profesión de aplicar la ciencia, la tecnología, la planificación y la gestión para hacer frente a eventos extremos que pueden dañar o matar a un gran número de personas, dañando la vida de la comunidad [20].

En el contexto nacional, el SINAE (Sistema Nacional de Emergencias, ente público encargado de prevenir y actuar en situaciones de desastre, del cual se profundiza más adelante), define a la gestión de riesgos como un proceso coordinado entre varias instituciones para reducir, prevenir, responder y apoyar a la rehabilitación y recuperación frente a eventuales emergencias y desastres, en el marco de un desarrollo sostenible [21].

Concepción actual de la Gestión de Riesgos

La gestión integral de riesgos para desastres naturales se compone de cuatro fases elementales: mitigación, preparación, respuesta y recuperación. Una buena gestión de riesgos emplea todas estas fases y se definen de la siguiente forma, ver figura 1 [18]:

- **Mitigación**: consiste en la reducción y/o eliminación de la probabilidad de un fenómeno y sus posibles consecuencias. El objetivo es que el impacto de un fenómeno sea lo menos dañino posible.
- **Preparación**: consiste en equipar a los potenciales afectados por el desastre y a posibles entidades de ayuda con equipamiento, herramientas y recursos necesarios para incrementar las probabilidades de supervivencia y eliminar/minimizar pérdidas.
- **Respuesta**: consiste en la ejecución de ciertas acciones que tienen como objetivo la reducción y eliminación de los daños ocasionados por la ocurrencia del desastre (o que está ocurriendo aún), con la finalidad de reducir pérdidas tanto humanas como financieras (a causa de interrupción de procesos o daños en infraestructura).
- **Recuperación**: consiste en lograr que las víctimas vuelvan a sus vidas cotidianas normales después de las consecuencias del desastre natural. Esta fase suele empezar de forma inmediata una vez concluida la fase de respuesta y su duración es absolutamente variable, pudiendo incluso llevar años.

Figura 1. Ciclo de la gestión de riesgos [22].

4.3.1 Gestión de Riesgos en Uruguay

A modo de enmarcar la realidad de Uruguay, es importante hacer mención al hecho de que el área y estudio de gestión de riesgos es relativamente nueva en América Latina (menos de dos décadas). Esto implica que en los últimos años, lo que antiguamente era concebir a

los desastres como un hecho a causa de la voluntad de Dios o la naturaleza, actualmente se lo concibe de forma más integral, donde la gestión del riesgo se entiende como la probabilidad de que existan pérdidas y daños vinculados al suceso de un evento físico y la vulnerabilidad de la sociedad ante estos.

En Uruguay, tanto los riesgos de emergencias como la presencia de desastres (tanto de origen natural o antrópico) han sido históricamente desconsiderados y visto como algo ajeno al país. De hecho los desastres ocurridos durante el siglo pasado, como fueron las epidemias, las sequías y las inundaciones, permanecían fuera de la percepción social. Esto comenzó a cambiar a mediados de la primera década del siglo actual, siendo la creación del SINAE una muestra de esto.

SINAE y marco institucional en Uruguay

Creado por la ley 18.621 “de Creación del Sistema Nacional de Emergencias”
“La instancia específica y permanente de coordinación de las instituciones públicas para la gestión integral del riesgo de desastres en Uruguay es el Sistema Nacional de Emergencias (SINAE). Su objetivo es proteger a las personas, los bienes de significación y el medio ambiente de fenómenos adversos que deriven, o puedan derivar, en situaciones de emergencia o desastre, generando las condiciones para un desarrollo sostenible” [23].

El SINAE es la unidad encargada de la coordinación de las instituciones públicas y de la gestión integral de riesgos en Uruguay. Este tiene como meta asegurar la protección de la población, de las infraestructuras y del ambiente contra eventos adversos que puedan tener como consecuencia un desastre o situación de emergencia. El SINAE no está representado por un cuerpo específico sino que su accionar se ejecuta mediante todas las acciones del Estado vinculadas a la gestión de riesgos en sus diferentes fases, siendo esta una tarea interinstitucional como resultado de la legislación de un espacio de articulación en las instituciones existentes [24].

La instancia superior de coordinación y decisión del SINAE se encuentra en el Poder Ejecutivo mientras que la Dirección Nacional se encuentra en la Presidencia de la República [24].

En cuanto a lo que es nivel departamental, se encuentran los Comités Departamentales de Emergencia (CDE), estos trabajan en conformidad con las políticas del SINAE y se encargan de formular y aplicar de este modo las políticas y estrategias a nivel local. Sus instancias operativas departamentales son los Centros Coordinadores de Emergencias Departamentales (CECOED) [24].

Finalmente, con la ley de “Descentralización política y participación ciudadana” surge un nuevo nivel de gobierno que son los municipios y en conjunto con estos surge a la par un nuevo nivel de gestión de riesgos. Según esta ley, uno de los cometidos de los municipios es: *“Adoptar las medidas urgentes necesarias en el marco de sus facultades, coordinando y colaborando con las autoridades nacionales respectivas, en caso de accidentes, incendios, inundaciones y demás catástrofes naturales comunicándolas de inmediato al Intendente, estando a lo que éste disponga”* [24].

Manejo de la respuesta: ver figuras 2 y 3

NIVEL DE RESPUESTA	COORDINACIÓN DE LA RESPUESTA
Atención Primaria	Autoridad Idónea en el evento
Respuesta Departamental	Comité Departamental de Emergencias
Respuesta Nacional	Comando de Respuesta Nacional
Situación de Desastre	Poder Ejecutivo

Figura 2. Cuadro de nivel de coordinación de Respuestas frente a emergencias y desastres [24].

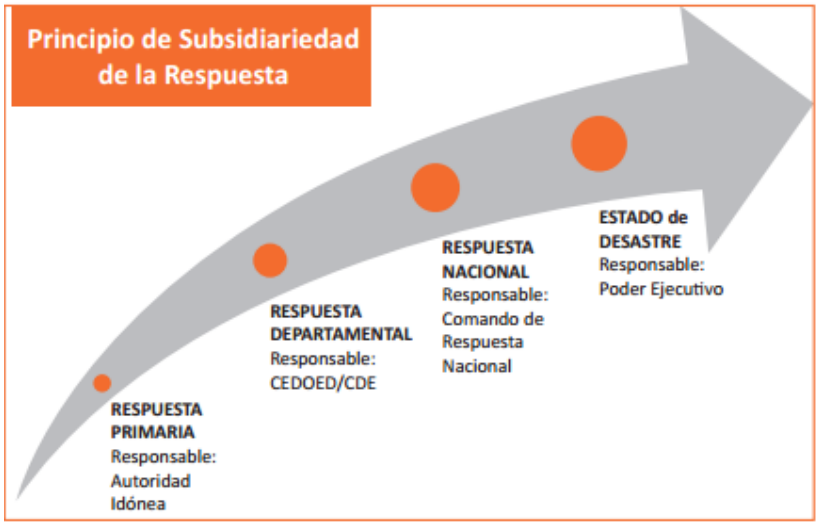


Figura 3. Avance de la respuesta y autoridad responsable [24].

INUMET

Creado por la ley 19158 de “Creación del Instituto Uruguayo de Meteorología”, el cual tiene dentro de sus principales fines [25]:

“A) Prestar los servicios públicos meteorológicos y climatológicos, consistentes en observar, registrar y predecir el tiempo y el clima en el territorio nacional y zonas oceánicas adyacentes y otros espacios de interés, de acuerdo a los convenios aplicables, con el objeto de contribuir a la seguridad de las personas y bienes y al desarrollo sostenible de la sociedad.

B) Coordinar las actividades meteorológicas de cualquier naturaleza en el país.

C) Representar a la República Oriental del Uruguay ante los organismos internacionales en materia de meteorología, así como cumplir con las obligaciones asumidas por el país ante los mismos.”

El INUMET tiene como misión institucional contribuir a la seguridad de la población y sus bienes, ayudando además así, al desarrollo sostenible de la sociedad. A su vez, dentro de su visión, se destaca la contribución a la gestión de riesgos meteorológicos y climáticos [26].

4.4 Data Warehouse

Para cumplir con los objetivos del proyecto se propone trasladar y almacenar la información en una base de datos, más precisamente en un data warehouse.

Data Warehouse (DW), o almacén de datos, es una colección de datos organizados enfocados a una organización o tarea específica, que proporciona información con el fin de ser analizada y de esta forma, asistir la toma de decisiones. Suelen ser un banco centralizado de información proveniente de distintas fuentes, que engloba la totalidad de la organización o proyecto.

Bill Inmon, uno de los principales autores y prácticamente un creador del término Data Warehouse junto con Ralph Kimball, presentó una definición de almacén de datos haciendo énfasis en propiedades de sus datos.

Inmon [27] propone como necesarias las siguientes propiedades para los datos alojados en un almacén, distinguiendo los siguientes atributos fundamentales:

- **No volátil:** los datos no se pueden modificar ni perder
- **Orientado a temas:** que exista vinculación entre los datos de manera de que éstos tengan sentido para describir una temática en cuestión
- **Variante:** se pueden agregar datos continuamente y las alteraciones realizadas sobre los mismos quedan registradas de manera de que éstos describan su evolución en el tiempo
- **Integrado:** los datos del almacén deben representar la totalidad de la información manejada por la organización o proyecto en cuestión

En particular, para este proyecto se consideraron como propiedades elementales del DW: orientado a temas, variante e integrado.

Por otro lado, Kimball [28] define a los almacenes de datos en función de su principal rol, definiéndose como el conjunto de toda la información referente a una organización, presentada de manera tal de facilitar su consulta para posteriores análisis.

4.5 Sistemas de Información Geográfica

Ante la necesidad de desplegar información geolocalizada (mediante coordenadas) sobre un mapa de manera de dar una vista panorámica de los resultados obtenidos sobre el territorio nacional, se debió investigar sobre las posibilidades para lograrlo. Dado el conocimiento del equipo sobre aplicaciones web y la popularidad de las mismas para este tipo de tareas, resultó inmediato investigar qué había a disposición por esa vertiente y en esta sección se exponen las posibles soluciones.

Un sistema de información geográfica es una herramienta que permite almacenar datos de tipo convencional que además posean una referencia geográfica que los ubique, además de desplegarlos sobre cartas de mapas de la tierra que éstos renderizan por defecto [29].

Los datos geolocalizados pueden ser de distintas características. Desde un punto único, una recta, conjunto de rectas o hasta polígonos delimitados por rectas, y más. Para todo ello cuentan con bases de datos relacionales especializadas en almacenar, consultar y acceder este tipo de datos, además de una organización de los mismos en formas de capas superpuestas que permiten la segregación de problemas. Estas capas representan una colección de datos y se puede trabajar sobre cada una de ellas independientemente del resto, además de activar o desactivar su visualización sencillamente, cosa que facilita el trabajo en la mayoría de los SIG [30].

Entre las varias aplicaciones de los SIG se tienen aplicaciones móviles que facilitan el traslado de personas, aplicaciones de cálculos de rutas, gestión de terrenos, cartografía de alta precisión, de asistencia hospitalaria, policial, y más.

Concentrándose en la utilización de un SIG para el presente proyecto, según Goodchild (1993) [31] un SIG destinado al análisis de datos y modelización ambiental debe incorporar un conjunto de herramientas para:

- Preprocesar grandes volúmenes de datos y prepararlos para su análisis
- Analizar los datos con el objeto de descubrir regularidades y desarrollar modelos
- Implementar estos modelos

También se considera como una funcionalidad posible en un SIG el permitir reorganizar los resultados en modo de tablas, gráficos o mapas de forma que sean útiles para el usuario.

4.6 Análisis Predictivo y Aprendizaje Automático

4.6.1 Minería de Datos

Se entiende por minería de datos o *data mining* como el proceso y las técnicas que permiten el estudio de grandes bases de datos tanto de manera automática como semi-automática con el fin de encontrar propiedades o patrones en los conjuntos de datos bajo cierto contexto [32].

En el marco de esta disciplina, se considera que la materia prima son los **datos**, los cuales se transforman en **información** una vez que reciben cierto valor (se les da un contexto y una utilidad); posteriormente estos datos (ahora información) pasan a integrar un modelo (en el cual la información es comparada o procesada), donde el resultado de su vinculación con el mismo, brindará un valor agregado que será referido como el **conocimiento** [32].

Aclarando que dentro del data mining no todos los casos son iguales, se puede decir que el procesamiento común se desglosa en los siguientes cuatro pasos [32]:

- Determinación de los objetivos
- Preprocesamiento de datos
- Determinación del modelo
- Análisis de los resultados

En la figura 4 se puede apreciar un gráfico comparativo que muestra la carga de trabajo en cada fase.

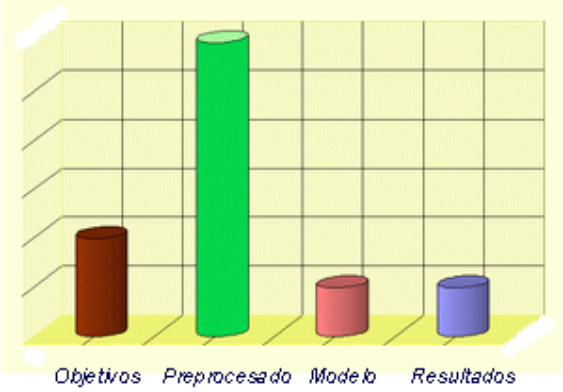


Figura 4. Carga de trabajo por fase [32].

4.6.2 Análisis Predictivo

El análisis predictivo es el campo en data mining [33] que pretende obtener información a partir de los datos con el objetivo de calcular las probabilidades y tendencias a futuro. Mediante el mismo, se pueden obtener conclusiones sobre la ocurrencia y las características de eventos desconocidos sin importar su ubicación temporal. No tienen por qué ser necesariamente en el futuro (aunque sí funciona bajo el concepto de utilizar datos anteriores al evento que se busca predecir, en ese sentido si predice el futuro).

Esto se logra, mediante el uso de métodos estadísticos, matemáticos y reconocimiento de patrones.

Este campo no es nuevo y además tiene aplicaciones de todo tipo. Por ejemplo, ya en la década de los 40, en plena guerra, el predictor de Kerrison podía apuntar un misil hacia un vehículo aéreo en movimiento que se le fuera indicado, tomando sólamente algunos

parámetros como la dirección de su trayectoria, velocidad, y ángulo observado desde el punto de lanzamiento, prediciendo la posición del mismo al momento de iniciar el lanzamiento del proyectil [34].

Entre tantas otras aplicaciones científicas, éstas técnicas se utilizan para, por ejemplo, detectar patrones de fraude en pagos electrónicos, optimización de campañas de márketing, reducción de riesgos comerciales y mejora de operaciones a todo nivel. Incluso ha prestado grandes aportes a la predicción de enfermedades crónicas y epidemias en el área de la salud. Se concluye que las aplicaciones del área son vastas mientras se tengan datos para alimentar el análisis [35].

Más detalles sobre la evolución del campo y otras aplicaciones del mismo pueden ser encontradas en la sección 7.2 de la versión completa del estado del arte presentada como Anexo 1 de este proyecto.

4.6.3 Modelado Predictivo y Aprendizaje Automático (Machine Learning)

En el análisis predictivo, el elemento central es el predictor, una variable que es medida para predecir el comportamiento en el futuro. No tiene porque haber un único predictor; de hecho, el uso de predictores múltiples conforman un modelo predictivo, los cuales bajo análisis, son utilizados para predecir con un nivel aceptable de fiabilidad [35].

Mediante el modelado predictivo se hace uso de estadísticas con el fin de obtener una predicción como resultado. Los mismos pueden utilizarse para cualquier tipo de evento desconocido sin importancia de su ubicación temporal [35].

En estos modelos, se recopilan los datos para los indicadores de relevancia, se formula un modelo estadístico, luego se hacen las predicciones y se valida el modelo con los datos adicionales que se encuentren disponibles. Dicho de otra manera, los modelos describen en función de variables medibles aquellos fenómenos que se quieren analizar y ésta metodología es muy utilizada actualmente en la tecnología de la información [35].

Una de las técnicas más populares para realizar un modelado predictivo es el aprendizaje automático o *Machine Learning* que es un subcampo de la computación, mediante el cual se busca resolver problemas donde hay un conjunto de datos muestra, y se intenta predecir propiedades de un conjunto de datos desconocido. La idea es que el programa “aprenda” (y de allí el nombre) de las muestras de datos que tiene como entrada, de manera tal de inducir conocimiento (modelar el comportamiento) sobre sus propiedades y que dicha base de conocimiento permita hacer las predicciones mencionadas sobre otros conjuntos de datos [36].

A los conjuntos de datos de entrada, se le llaman *training sets*, o conjuntos de entrenamiento ya que son éstos los que asisten al aprendizaje del programa. Y usualmente se tiene otro conjunto de datos “nuevo” sin procesar sobre el cual se aplican el modelo del primer conjunto, a los cuales se les suele llamar *testing sets*, o conjuntos de prueba [36].

Sobre los datos de entrada, pueden ser tanto valores numéricos simples, como vectores de varias columnas de datos de tipos mixtos, a los que se les llama características o atributos.

Dentro del aprendizaje automático, nos interesa para utilizar en este proyecto, un tipo de metodología en particular, llamados de **aprendizaje supervisado** [37] donde existen disponibles ciertos conjuntos de datos de entrada que poseen atributos que luego se querrá predecir para futuros conjuntos de datos. Es posible aplicar estas técnicas debido a que se tienen datos ya etiquetados.

Una variable continua es aquella que, dentro de un valor mínimo y máximo, puede tomar infinitos valores en el intervalo, esto podría ser, por ejemplo, números decimales dentro de un rango definido. Por el contrario, una variable discreta puede tomar solamente ciertos valores dentro del máximo y del mínimo, con un codominio de tamaño finito [37].

A su vez, dentro del aprendizaje supervisado, se encuentran dos tipos de problemas a resolver. Los problemas de **regresión** [36] son aquellos donde la salida es el valor de una variable continua. Los análisis de regresión son procesos que utilizando diversas técnicas matemáticas, estiman relaciones entre las variables o atributos de los datos de entrada. La estimación de estas relaciones y el aprendizaje sobre éstas en base a los datos de entrada permiten a las técnicas empleadas poder ganar conocimiento entre cómo ciertas variables independientes (o predictoras) afectan el cambio de una variable dependiente que se desea predecir, de manera tal de poder arrojar dicha predicción como resultado final.

Por otro lado, también dentro del aprendizaje supervisado se encuentra a los problemas de **clasificación** [36], donde se busca clasificar utilizando etiquetas a cierto conjuntos de datos (observación). Para esto se toma un conjunto de datos previamente etiquetados de igual estructura a la observación. Notar que en este caso de clasificación, la salida, o sea una etiqueta sobre un conjunto de datos dado, resulta una variable dependiente de tipo discreta, esto es, habrá un número finito y definido de posibles etiquetas para clasificar el nuevo conjunto de datos.

Algoritmos de predicción con Machine Learning

Generalmente, cuando se plantea un problema a resolverse mediante aprendizaje automático una de las partes más difíciles suele ser determinar el algoritmo correcto a utilizar. Cada estimador se adapta mejor a diferentes tipos de problemas con diferentes conjuntos de datos [38].

En la versión completa del estado del arte presentada como anexo 1 de este proyecto se puede encontrar una lista detallada de aquellos algoritmos investigados junto con una base teórica de cada una mientras que en éste resumen se presenta a continuación una lista enumerando estos algoritmos y clasificándolos según sean de regresión o clasificación.

Clasificación:

- Support Vector Classification
- NuSVC (Nu Support Vector Classification)
- Naive Bayes
- Stochastic Gradient Descent

- Nearest Neighbors

Regresión:

- Linear Regression
- Ridge
- Lasso
- Elastic Net
- SVR (Support Vector Regression)

4.6.4 Validación de Resultados

Con el fin de medir la calidad de un modelo predictivo, se recurre a un simple concepto, el método de retención (ver figura 5). Se toman todos los datos referentes al pasado disponibles, de los cuales ya conocemos su resultado (si existió o no un evento de inundación) y lo particionamos en dos conjuntos: un conjunto de datos de entrenamiento del modelo, y otro conjunto de prueba. La idea de las pruebas es simple y radica en entrenar un modelo que utilice un método dado con el conjunto de datos de entrenamiento que se obtuvo, luego predecir resultados utilizando el modelo previamente entrenado y alimentándose con los datos que se separaron para prueba, lógicamente “enmascarando” el hecho de que ya se tienen los resultados reales asociados a los mismos. Para concluir la prueba, se compara el resultado de dicha predicción con aquellos resultados reales y comprobados que se tienen para cada elemento del conjunto de datos de prueba. Como resultado de esta comparación se obtienen métricas de error de cada uno de los métodos que serán detalladas más adelante dentro de esta sección. Dado que se quiere aproximar el resultado lo máximo posible a la realidad, se busca reducir dicho error al mínimo y por lo tanto se seleccionarán aquellos métodos que lo minimicen.

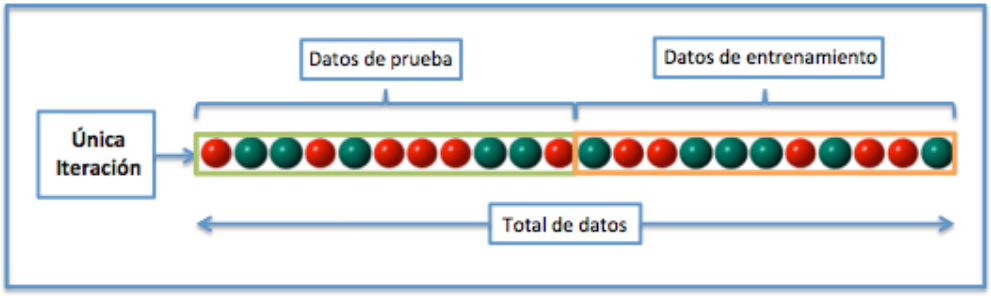


Figura 5. Método de retención [39].

Sin embargo, esta técnica de validación introduce una problemática y ésta es una dependencia entre los datos. Dado que los algoritmos de aprendizaje automático, como se describe con más detalle en la sección 7.3 del Anexo 1, funcionan buscando relaciones entre los conjuntos de variables de entrada y sus resultados asociados, si uno siempre utiliza un mismo conjunto de datos, puede caer en un modelo que se entrenó de manera muy dependiente a ésta relación y probablemente no funcione adecuadamente con un

nuevo conjunto de datos, que es justamente su fin.

Aquí se introduce el término de validación cruzada [39] que resultó vital para la selección final del método adecuado. La validación cruzada busca atacar la problemática expuesta en el párrafo anterior y es una técnica conocida para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba [40].

Existen tres tipos de validación cruzada que se desean destacar.

En K iteraciones *(K-Fold)*

Se realizan K iteraciones, donde dentro de cada una, se separa al conjunto de datos total en K subconjuntos contiguos. Uno de estos subconjuntos hará el papel de datos de prueba, mientras el resto (K-1) conjuntos, hacen el lugar de datos de entrenamiento. Dentro de cada iteración se computa un resultado, y finalmente se computa un promedio con el fin de dar un resultado único de precisión, y asociado a él, la **desviación estándar** con respecto a dicho promedio. Posee la desventaja de ser lento computacionalmente, pero dado que separa el mismo conjunto total de datos es diferentes combinaciones de subconjuntos de entrenamiento/prueba, contrarresta la potencial dependencia a desarrollarse entre éstos subconjuntos a la hora de evaluar un método [39]. Ver figura 6.

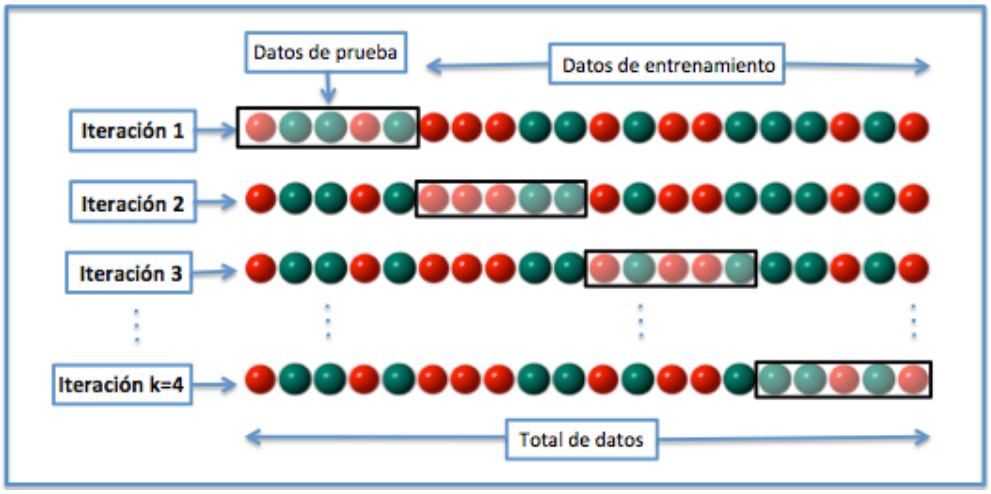


Figura 6. Validación K-Fold [39].

Aleatoria

En este caso los conjuntos de entrenamiento y prueba para cada iteración se conforman seleccionando elementos individuales de manera aleatoria.
A diferencia del tipo anterior, se tiene como ventaja que la división de conjuntos no depende del número de iteraciones, pero como contrapartida, puede suceder que algunos

elementos no sean probados nunca, o que algunos lo sean más de una vez [39]. Ver figura 7.

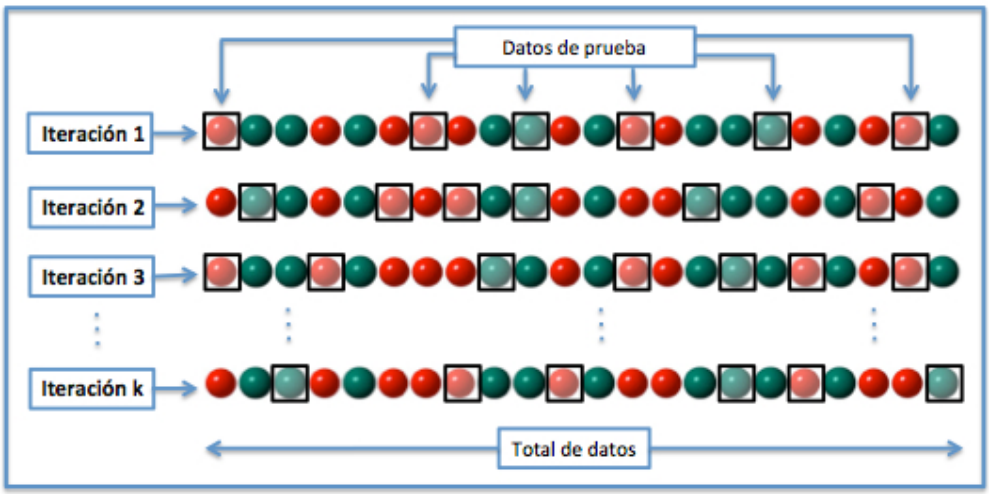


Figura 7. Método aleatorio [39].

Dejando uno fuera *(Leave One Out o simplemente, LOO)* [39]

Este tipo es un caso particular de la validación K-Fold con K = cantidad de muestras. Es así que se generan N iteraciones donde N es la cantidad de muestras, y dentro de cada iteración, uno de los datos hace el papel del de prueba, y todo el resto pasan a ser datos de entrenamiento. Como desventaja, presenta un alto costo computacional resultado de un mayor número de combinaciones a computar y el hecho de que el conjunto de datos de entrenamiento sea de mayor tamaño, que también hace que la tarea de entrenar un modelo tome más tiempo. Ver figura 8.

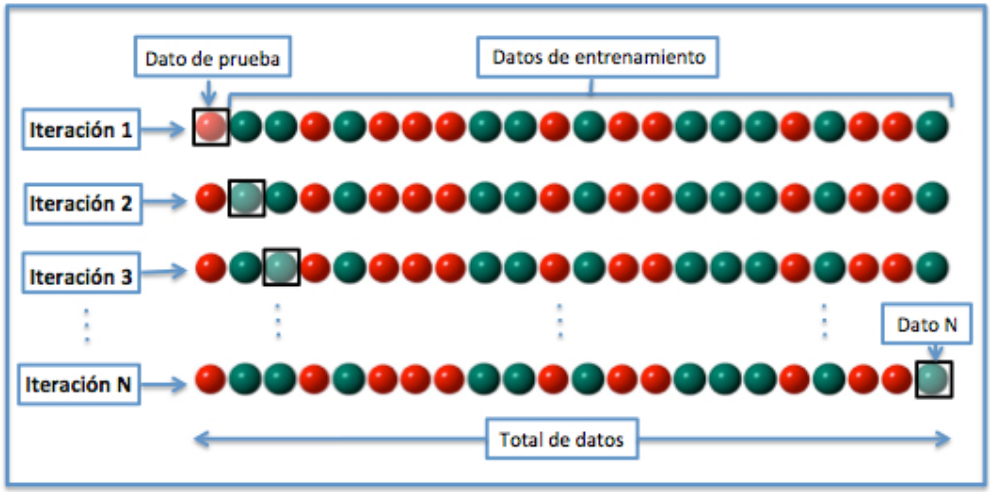


Figura 8. Método *Leave One Out* [39].

Es preciso ahora introducir aquellas métricas que miden la calidad de un modelo. Utilizamos los términos en inglés ya que son los utilizados por artículos científicos. Además,

aquí sólo se listan las métricas validadas con sus nombres, pero su descripción completa se encuentra en la sección 7.4 del estado del arte completo adjuntado como Anexo 1.

Métricas de validación para Clasificación

- **Accuracy**: Porcentaje de aciertos en las predicciones efectuadas.

Específicamente para la clasificación binaria, tenemos:

- **Verdaderos Positivos (TP):** numero de muestras positivas etiquetadas positivas (correcto).
- **Falsos Positivos (FP):** número de muestras negativas etiquetadas como positivas (error).
- **Verdaderos Negativos (TN):** número de muestras negativas etiquetadas como negativas (correcto).
- **Falsos Negativos:** número de muestras positivas etiquetadas como negativas (error)

Con estas nuevas definiciones, podemos redefinir el mencionado término de *accuracy* como:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Es importante notar que la exactitud (*accuracy*) es inversamente proporcional a la suma de falsos positivos y falsos negativos.

Para combatir estas tendencias, introducimos dos nuevas métricas:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

Métricas de validación para Regresión

Las principales métricas utilizadas para contabilizar la calidad de los métodos de regresión son R cuadrado, error medio absoluto y error cuadrático medio.

R cuadrado (R^2), o también llamado coeficiente de determinación, es una métrica estadística que representa la proporción de la variación del resultado que puede explicar el modelo entrenado, a partir de las variables de entrada. Su valor está en el rango del 0 al 1, donde un 0 indica que el modelo no puede explicar nada de la variación del resultado, y un 1 significa que describe la variación perfectamente.

Error Medio Absoluto (*mean absolute error*) es una medida de distancia entre dos variables continuas. Si tenemos las variables de observación x_i y aquellas predichas y_i, entonces definimos al Error Medio Absoluto entre la observación y lo predicho como:

$$EMA = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

Finalmente, el **Error Cuadrático Medio** (*mean squared error*) es muy similar al anterior, el Error Medio Absoluto, pero con la particularidad de que se suman los cuadrados de las diferencias entre lo observado y lo predicho. Por lo tanto, lo definimos como:

$$EMC = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

Donde su diferencia con el error medio absoluto previamente presentado reside en el cuadrado de la diferencia entre cada observación y realidad. Este enfoque claramente pone mayor “peso” o gravedad en los errores y de esta manera, “castiga” de peor manera a los mismos.

4.7 Herramientas para el Análisis de Datos

En esta sección se presentan las herramientas más importantes que presentan una alternativa al desarrollo utilizando Machine Learning. Dichas herramientas en general ya cuentan con implementaciones que buscan aportar soluciones a problemas comerciales que requieren de un modelado predictivo además de resultar más accesibles para usuarios sin capacitaciones técnicas en el ámbito de la programación.

4.7.1 Weka

Weka es una herramienta conformada por un conjunto de algoritmos de machine learning para realizar tareas de data mining. Estos algoritmos son aplicables directamente a un conjunto de datos así como también utilizables desde un proyecto Java. Además contiene herramientas para preprocesado de datos, clasificación, regresión, agrupamiento (o análisis de grupos), reglas de asociación y visualización [44]. A su vez cuenta con una interfaz gráfica que permite unificar todas las herramientas disponibles para un uso más simple.

Como principales características, se puede describir a Weka como una herramienta muy adaptable a las necesidades mediante la aplicación de tareas básicas de data mining. Los datos que utiliza pueden tener como origen ficheros planos con un número fijo de atributos por registro así como accediendo a instancias de base de datos mediante SQL gracias a JDBC [45].

4.7.2 R

R es un lenguaje de programación publicado al igual que Weka bajo licencia GNU y por lo tanto libre. El mismo conforma un entorno de análisis estadístico para la manipulación de datos, sus cálculos y la posterior visualización de estos y los resultados obtenidos

mediantes gráficos. Este es considerado como una implementación más moderna y libre de S (lenguaje de programación estadístico desarrollado en 1976) [46].

R provee una amplia variedad de estadísticas (modelado lineal y no lineal, tests estadísticos, análisis de series temporales, clasificación, agrupación, etc), además es muy extensible contando con más de 10.000 paquetes disponibles [46][47].

En la actualidad, R se ubica como el software con mayor cantidad de recursos, grado de desarrollo y aceptación, siendo usado por empresas como Google y Facebook entre otras y dentro de la comunidad científica para la realización de investigaciones [47].

Algunas de las principales características con las que cuenta son [48]:

- Almacenamiento y manipulación de datos
- Herramientas de análisis de datos
- Gráficos para análisis de datos
- Operadores de cálculo para arreglos y matrices

4.7.3 Pentaho

Pentaho es una plataforma de código abierto enfocada a la realización de análisis de datos e informes. A través de su implementación mediante Java, el mismo brinda la posibilidad de moldearse a necesidades específicas de cada organización [49].

Al ser un software orientado al negocio (Business Intelligence o simplemente BI) brinda entre sus funciones la realización de reportes intuitivos, análisis OLAP (procesamiento analítico en línea), cuadros de mando, integración de datos, minería de datos y plataforma BI [49].

Sin profundizar en detalles, algunas de las principales características de Pentaho son [50]:

- El concepto de “Governed Data Delivery”, según Pentaho refiere a la capacidad de poder acoplar datos confiables y oportunos para brindar análisis a escala para todos los usuarios en todos los entornos.
- Posee una fácil integración a cualquier flujo de trabajo gracias a tener una arquitectura multi-tenant.
- Capacidad para acceder y combinar información para otorgar información analítica lista para el consumo de los usuarios finales en conjunto con una interfaz “drag and drop” para eliminar complejidad.
- Integración nativa mediante una capa que se integra a cualquier fuente de datos, incluyendo Hadoop y NoSQL entre otros.

4.8 Antecedentes

Como parte de este proyecto también se estudió sobre lo que ya se ha investigado e implementado a nivel mundial y local en cuanto a predicción de inundaciones y trabajos relacionados.

4.9 Global Flood Monitoring System

Según la agencia estadounidense NASA, predecir inundaciones es algo de lo más complicado e involucra una combinación de parámetros de tipo ambiental, de estado del tiempo reciente y actuales [51]. En 2012, la misma financió, y trabajó en conjunto con el Dr. Huan Wu de la Universidad de Maryland en el Global Flood Monitoring System (GFMS) [52], un sistema experimental de monitoreo de inundaciones. El mismo basa su funcionamiento inicialmente en el Tropical Rainfall Measuring Mission satélite (TRMM), y posteriormente migró a utilizar el Global Precipitation Measurement satellite (GPM), ambos satélites de observación sobre la Tierra en tiempo real de precipitaciones, siendo el segundo de mejor precisión y tecnología.

El funcionamiento del sistema combina información brindada por el satélite en cuestión, un modelado de la composición de la superficie de la Tierra que incluyen densidad de vegetación, sedimentación de los ríos y el terreno de casi todo (rango de latitudes 50°N - 50°S) el planeta llamado Variable Infiltration Capacity, así como también un modelo de enrutamiento de excedentes de agua en cursos, el Dominant River Tracing based runoff-Routing (DRTR), para calcúlar cuánta de la lluvia que caerá será absorbida por la tierra y penetrará las capas terrestres, y cuánta terminará desplazándose por los cursos de agua, eventualmente saturando su caudal y resultando en una inundación.

El producto resultante es una aplicación web con funcionamiento continuo sin interrupciones, accesible mediante http://flood.umd.edu/ [53], donde los usuarios pueden visualizar, plasmadas en un mapa, las siguientes estadísticas:

- Inundación (profundidad del agua en mm)
- Flujos de agua en una grilla de resolución de 12km
- Flujos de agua en una grilla de resolución de 1km
- Flujos de agua sobre los máximos establecidos como normales
- Precipitaciones instantáneas en mm/h
- Acumulación de precipitaciones en mm de 3, 5 y 7 días previos

4.10 Sistema de Alerta Temprana Prohimet-Yi

En Uruguay, se ha trabajado previamente en esta temática a nivel de sistemas de información, motivados por desastres que han generado daño socioeconómico a la población.

Prohimet-Yi es un ejemplo de estos sistemas, siendo este un sistema de alerta temprana cuyo objetivo es tener una mejora en la gestión de las inundaciones que se dan en la ciudad de Durazno [54].

El mismo empezó a llevar a cabo sus etapas iniciales para ponerse en funcionamiento entre junio de 2009 y diciembre de 2011, siendo financiado por la OMM (Organización Meteorológica Mundial) y realizado por el Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA) de la Facultad de Ingeniería de la Universidad de la República. Además contó y actualmente cuenta con el apoyo de las instituciones nacionales involucradas en el proyecto y de los miembros de la red PROHIMET [54].

El objetivo principal de este proyecto de grado, fue el de investigar e implementar un sistema de información geográfica (SIG) que, además de alertar en caso de inundación, permita realizar consultas espaciales, (cantidad de personas a ser evacuadas, de qué regiones de la ciudad, etc.) en base a la información geográfica de la ciudad de Durazno. Los objetivos secundarios consisten en incorporar mejoras en la parametrización de los scripts del modelo actual, brindar una gran flexibilidad para generar consultas espaciales de forma dinámica, mejoras en la notificación de las alertas utilizando mensajes de texto, entre otros [55].

4.11 Datos Abiertos

El presente proyecto planteó la necesidad de obtener una gran cantidad de datos históricos (estados del tiempo, alturas de ríos, registros de desastres), como consecuencia de esto fue necesario investigar, buscar fuentes de datos públicos y ponerse en contactos con diferentes entes que resultaran oportunos para proporcionarnos la información requerida.

En este contexto resulta interesante comparar cómo se dan estos datos en Uruguay, la calidad de los mismos y su disponibilidad en comparación con países de primer mundo o generalizando a nivel internacional.

Como concepto previo, hay que mencionar a los Datos Abiertos como una tendencia moderna con gran potencial de desarrollo. Esto es, que cualquier ente (individual u organizacional) recaba información (datos) para realizar sus tareas. En particular, el Gobierno resulta vital en este contexto ya que recoge una amplia cantidad de datos de empresas, de la población, del medio ambiente, económicos, de salud, geográficos, etcétera [56].

Definiendo a los Datos Abiertos, estos son los datos que existe la posibilidad de que cualquier persona sea libre de utilizarlos, reutilizarlos y redistribuirlos. Siendo sus principales características las siguientes [56]:

- Disponibilidad y Acceso: completamente disponibles y con un costo de reproducción razonable, siendo la descarga gratuita a través de internet el caso óptimo.

- Reutilización y Redistribución: posibilidad de realizar un producto como derivado de estos datos, así como usarlos en conjunto con otras fuentes de información y distribuirlo de forma gratuita.
- Ausencia de Restricción Tecnológica: No debe haber obstáculos de carácter tecnológico para su uso y redistribución (Formato abierto).
- Participación Universal: Cualquier persona puede utilizar, reutilizar y redistribuir, sin restricciones respecto a las acciones posibles sobre estos o condicionantes.

Por otra parte, los datos de gobierno se consideran abiertos si cumplen los siguiente ocho puntos establecidos internacionalmente por Open Government Data [57]:

- Completos: Todos los datos públicos están disponibles. Se considera datos públicos a aquellos que no tienen restricciones de privacidad, seguridad o privilegio.
- Primarios: Los datos son obtenidos en la fuente, con el mayor nivel posible de granularidad, sin ser modificados ni agrupados.
- Periódicos Los datos quedan disponibles tan pronto como sea necesario para preservar su valor.
- Accesibles: Los datos quedan disponibles para la mayor cantidad posibles de usuarios y propósitos
- Procesables: Los datos están estructurados (tanto como es posible), permitiendo así un procesamiento automatizado.
- No Discriminatorios: Disponibles a todo publico sin necesidad de registros
- Sin Licencia: Los datos no se encuentran sujetos a ningún tipo de regulación de derechos. Pudiéndose permitir únicamente en casos razonables restricciones de privacidad, seguridad o privilegio.

Además, el cumplimiento de estos puntos, debe ser comprobable.

4.11.1 Datos Abiertos en Uruguay

En los últimos años, se ha promovido la liberación de datos gubernamentales a nivel público, esto forma parte de los avances realizados en el marco de lo que se conoce como e-gobierno (gobierno electrónico). Estos avances se han realizado principalmente mediante la AGESIC (Agencia para el Desarrollo del Gobierno de Gestión Electrónica y la Sociedad de la Información y del Conocimiento), que es el organismo encargado de planificar e implementar estrategias de Gobierno Electrónico, avanzando en la concreción de políticas de acceso a la información y de datos abiertos en el Estado [58].

Con el fin de regularizar la liberación de datos es que en 2008 se crea la Ley de Acceso a la Información (Ley 18.381), la cual explica en su primer artículo tiene como objetivo [59]

"*promover la transparencia de la función administrativa de todo organismo público, sea o no estatal, y garantizar el derecho fundamental de las personas al acceso a la información pública*".

Más allá de que nuestro país no cuenta con una política definida en cuanto a la publicación de información pública en el catalogo de datos abiertos (en catalogodatos.gub.uy [60]), si existe una licencia definida que presenta las siguientes condiciones [60]:

"El Usuario reutilizador de los datos de este sitio, que no disponga de licencia específica indicada por la entidad pública que los aporta (Usuario catalogador), deberá cumplir, al menos, las siguientes condiciones básicas:
1) Mantener el sentido original de la información,
2) Citar siempre la fuente
3) Explicitar la fecha de la última actualización.
Se permite cualquier explotacion de los datos abiertos, incluyendo una finalidad comercial, así como la creacion de obras derivadas, estando permitida su distribucion sin ninguna restriccion. La utilización, reproducción, modificación o distribución de los conjuntos de datos supone siempre la obligación de reconocer, citar y/o enlazar a la entidad de que se trate como la fuente de los conjuntos de datos. Deben conservarse, y por tanto no alterarse ni suprimirse los metadatos sobre la fuente, fecha de actualización y las condiciones de reutilización aplicables incluidos, en su caso, en el documento puesto a disposición para su utilización o reutilización. AGESIC puede solicitar a un usuario que cese en el uso de esta licencia y cualquier forma de distribución de datos que se realice bajo esta, en el caso que considere que ha existido una violación a los términos y condiciones descriptos. Ello, sin perjuicio de las medidas de carácter legal que AGESIC pueda llegar a adoptar con el Usuario, que haya incumplido los Términos y Condiciones generales definidos."

Vale aclarar que dicho reglamento únicamente aplica a los datos dentro del portal de catálogo de datos abiertos, no a todos los datos en poder del Estado.

Información más detallada sobre los datos públicos del Uruguay se puede encontrar en la sección 10.1 del Anexo 1 de este informe, la versión completa del estado del arte.

5 Análisis, diseño e implementación de la solución

En este capítulo se detalla la solución propuesta en este trabajo y detalles de su análisis, diseño e implementación.

En términos generales la arquitectura de la solución (ver figura 9) consiste en un data warehouse (DW) central que aloja la totalidad de la información obtenida, tanto de manera dinámica como los históricos estáticos, ya procesada y con formato definitivo para entrar al mismo y un servidor web que consume datos de este DW, procesa estos datos mediante un motor predictivo implementado en el servidor y finalmente, entrega los resultados de manera interpretada para fácil lectura del usuario final a través de un mapa embebido en el sitio. Además, este mismo servidor web tendrá implementado los workers o scripts de obtención automática descritos en la sección 5.2 y éstos podrán ser ejecutados desde el mismo.

La alimentación del DW se da tres maneras. A través de bases de datos obtenidas de manera artesanal, descrito en la sección 5.1, de manera dinámica a través del consumo de APIs, o de manera dinámica también pero mediante scraping de fuentes web, ambos descritos en la sección 5.2.

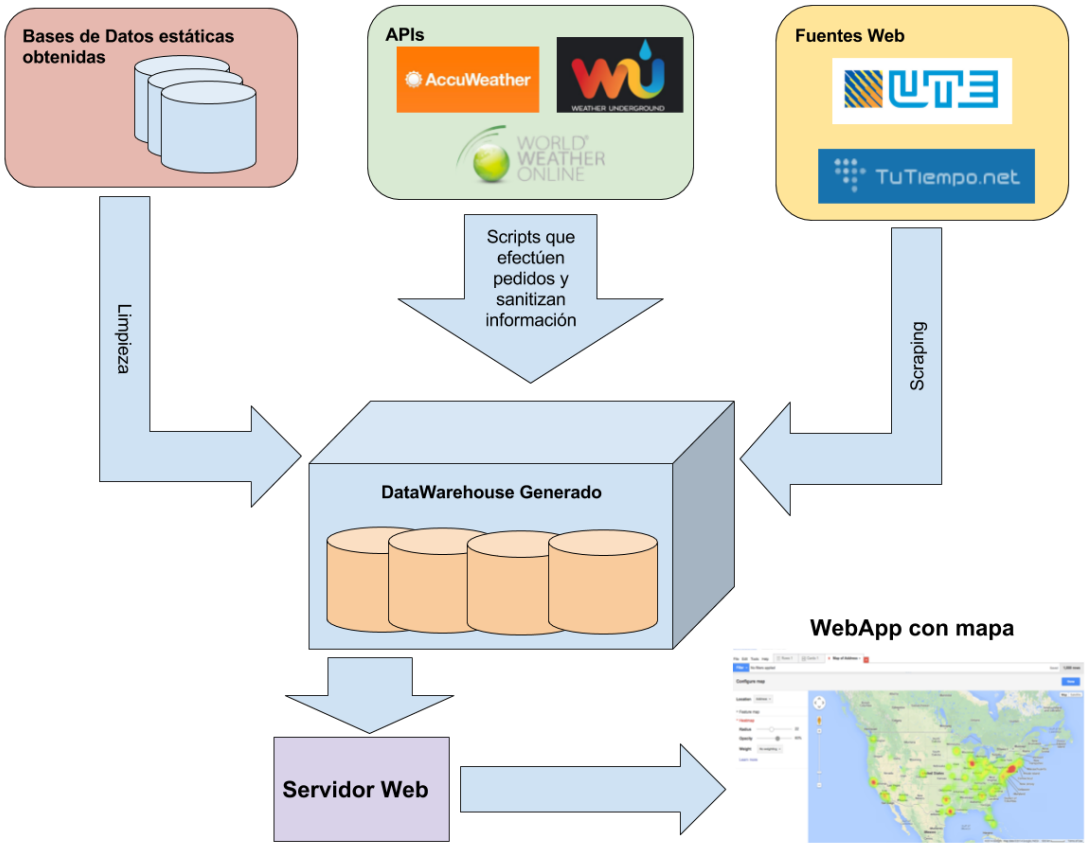


Figura 9. Arquitectura básica de la aplicación final.

5.1 Obtención de Datos

Dado que para el cumplimiento de los objetivos de este trabajo se decidió usar técnicas de aprendizaje automático, fue necesario generar conjuntos de datos de entrada que cumplan el papel de los mencionados conjuntos de entrenamiento o *training sets* de dichas técnicas. Es razonable entonces que, dado el objetivo del proyecto y su planteamiento de predecir desastres en base a datos históricos, éstos conjuntos de datos sean de sucesos reales pasados, de la misma naturaleza de aquellos que se intenta anticipar aquí.

Es así que se decidió trabajar en la tarea de obtención de éstos datos contactando fuentes oficiales relacionadas con la temática de desastres que afectan el territorio nacional.

Lamentablemente, los datos que estos organismos nos pudieron proveer de manera automática son nulos. El estado se encuentra, en estos tiempos, en un importante e incipiente proceso de digitalización e integración de todo tipo de información estatal dado que la situación a la fecha de elaborado este trabajo es inaceptable. Los distintos organismos estatales manejan información incoherente entre sí, incompleta y que no cumple con mínimos estándares de calidad de datos que se mencionan en la sección 5.3 del presente informe. Además, en el caso particular de aquellos organismos que poseen los datos que se desean para este trabajo, no presentan infraestructura alguna para servir estos datos de manera automática mediante internet, como por ejemplo servicios web.

A continuación se describen los datos que sí se pudieron obtener y los medios para ello.

5.1.1 Datos Estáticos

Como consideración previa, se menciona como datos estáticos a aquellos obtenidos a partir de documentos excel brindados por los diferentes entes con los que se mantuvo contacto (SINAE, UTE, INUMET).

De esta manera, la tarea de obtención de datos se realizó de forma manual en el sentido de tener que entablar una comunicación mediante correo electrónico con aquellos organismos de los cuales era deseable extraer información de interés, e incluso en más de una ocasión, concretar una reunión presencial con algún jerarca de la misma para realizar la solicitud.

Los organismos considerados para la recabación, y los datos obtenidos de cada uno de ellos fueron los siguientes.

SINAE

SINAE es uno de los principales actores en la temática de desastres en el territorio Uruguayo y así fue que lógicamente se presentó como una de las principales opciones a manejar para solicitud de los datos requeridos. Mediante una reunión, se acordó el intercambio de un archivo de formato Excel de su propiedad, en el cual ellos tienen un registro de desastres sucedidos en territorio nacional, comprendidos entre 1983 y 2014, junto con sus consecuencias a niveles humanos y materiales en términos numéricos, la

mayoría de los cuales ellos a su vez recabaron del histórico de medios nacionales de comunicación, principalmente periódicos. También poseía comentarios textuales sobre lo sucedido, y una localización geográfica para cada evento, en formato de coordenadas de latitud y longitud, lo cual fue una gran ventaja para el cumplimiento del objetivo de geolocalización de la información.

INUMET

Se acudió a dicho organismo en busca de información relacionada a pronósticos y registros de estado del tiempo históricos en el país, siendo estos un factor determinante en los desastres sucedidos en el pasado, y representando así un valioso conjunto de datos para el entrenamiento de los algoritmos de aprendizaje utilizados en este trabajo.

Mediante intercambio de correos electrónicos y el haber presentado una carta de solicitud formal de los datos, firmada por los docentes tutores del presente trabajo, el organismo accedió a enviar de manera gratuita información en formato Excel con registros de precipitaciones con nivel de granularidad periódica, en unidad de milímetros cúbicos, correspondientes a las mediciones realizadas en diversas estaciones meteorológicas controladas por el organismo a través del territorio nacional comprendidas en el período desde 1960 hasta 2016.

Esta información complementa con mayor exactitud la obtenida mediante el SINAE; además de agregar exactitud en un factor determinante en la probabilidad de inundación de una localidad dada, como lo es la cantidad de precipitación que se dió en el lugar o cercanías. La vinculación de cada pieza de información con una estación de medición, de la cual se conoce públicamente su posición geográfica, hace que se pueda considerar también geolocalizada.

UTE

La Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE), creada por Ley N° 4.273, de 21 de octubre de 1912:

“La Administración Nacional de Usinas y Trasmisiones Eléctricas (UTE), es una empresa propiedad del Estado uruguayo que se dedica a las actividades de generación, trasmisión, distribución y comercialización de energía eléctrica, prestación de servicios anexos y consultoría.” [61].

UTE nos resultó particularmente útil ya que están a cargo de las estaciones de mediciones de alturas de ríos, otra variable de vital importancia en un suceso de inundación, ya que justamente, éstos se dan, en su gran mayoría, cuando dichos niveles exceden las cotas normales de un curso de agua dado. Éstas estaciones, algunas operadas de manera manuales por operarios empleados del organismo, y las restantes de ellas automáticas y electrónicas, registran las mediciones con granularidad a nivel horaria en las estaciones de medición a lo largo de los cursos principales de agua en el país, principalmente el Río Negro y Río Uruguay. Cabe destacar que es en los bordes de éstos ríos donde se dan la mayoría de los eventos de inundación que afligen al país.

A través de un intercambio de correos electrónicos y finalmente una reunión presencial en su sede oficial, se llegó a un acuerdo en el que se comprometieron al envío de información en formato Excel conteniendo dichas mediciones, comprendidas en el período desde 1996 hasta 2016. Los datos de altura de ríos obtenidos se encontraban en unidad de metros. En conjunto con estos datos, también brindaron mediciones de precipitaciones en una gran cantidad de pluviómetros administrados por UTE, ubicados su mayoría en el Río Negro y cercanías, estos datos vienen dados en milímetros al igual que los datos proporcionados por INUMET y son mediciones diarias.

Gracias a estos datos y sumados a los proporcionados por INUMET, se destaca que se logró recabar más de un millón de registros de precipitaciones comprendidos entre 1960 y 2016.

EM-DAT

El Centro de Investigación Epistemológico de Desastres proporciona, de manera pública y consultable por cualquier persona que desee hacerlo mediante una interfaz web, una base de datos la cual posee información registrada sobre desastres naturales que ocurrieron en el Uruguay, entre otros países. Cabe mencionar que los eventos registrados fueron sólo aquellos eventos que tuvieron un alto nivel de repercusiones, se encontraron alrededor de 60 eventos naturales de distinto índole que ocurrieron en el Uruguay desde el año 1900 hasta el presente [62].

5.1.2 Problemas encontrados para la obtención de datos

Como era de esperar, el estado de la información recibida, en términos de cumplimiento de los estándares de calidad de datos que se conocen y estudian en la actualidad, no fue bueno y esto derivó en un importante trabajo de calidad de datos (detallado en secciones 5.3.1 y 5.3.3) y su conversión a un formato deseado, para posteriormente ingresar en el DW destinado a ser la base central de información de nuestro proyecto.

Además, también se realizaron solicitudes al **SOHMA** (Servicio de Oceanografía, Hidrografía y Meteorología de la Armada) y **DINAGUA** (Dirección Nacional de Agua), pero sin éxito ya que éstos no poseían datos preparados para ser entregados, ni los recursos humanos para asignar a la preparación de los mismos, y en algunos casos, tampoco tenían los datos solicitados.

5.1.3 Datos obtenidos

Sobre los datos finalmente obtenidos, cabe destacar que, de cada estación de medición se conoce su posición en términos de coordenadas de latitud y longitud (tanto para UTE como INUMET), por lo tanto estos datos también se consideran geolocalizados.

También, para los eventos de SINAE y EM-DAT se realizó un minucioso y extenso trabajo para completar la información y conseguir la localidad de cada registro así como de limpieza de datos que no eran de interés, ver sección 5.3.3.

A continuación se hace un detalle de la estructura de la información recibida de cada ente.

SINAE

Se nos proporcionó un archivo en formato Excel de 3675 filas y 42 columnas las cuales se especifican en la Tabla 1.

Columna	**Descripción**
Serial:	Código compuesto por dos números: año del desastre y un número identificador para cada entrada
Disasterid	Identificador del desastre en la base de datos del SINAE (un desastre puede tener varias entradas, por ubicación, que se referencian teniendo el mismo disasterid)
Fecha_crea	Fecha en que se registró el desastre en la base de datos
Fuentes	Fuente de donde se recabó la noticia
Nombre_geo	Localización geográfica del desastre (puede ser departamento o departamento/ciudad)Localización geográfica representada por un nombre de departamento y ciudad del desastre
Cod_geogr	Código identificador del lugar geográfico del desastre
Sitio	Ubicación con mayor granularidad del desastre (barrio, pueblo, zona, parque)
Longitud	Coordenada de longitud de la localización exacta
Latitud	Coordenada de latitud de la localización exacta
Pre_coor	Característica de la precisión de la ubicación de las coordenadas
Tipo_de_ev	Tipo del evento ocurrido (inundación, incendio, tormenta, etc)
Sub_even	Dato de mayor precisión respecto al evento ocurrido
Observacio	Texto con mayor descripcion respecto al tipo de desastre ocurrido

Fecha_inic	Fecha en que comenzó el desastre
Duracion	Duración en días del desastre
Nivel_re	Tipo de respuesta al evento (nacional, departamental, primaria)
Observac_3	Texto descriptivo acerca de las medidas tomadas contra el evento ocurrido
Tipo_de_ca	Tipo de catástrofe (natural, causado por el hombre, etc)
Observac_2	Texto descriptivo de consecuencias generadas por el desastre
muertos	Cantidad de fallecidos en el desastre
desapareci	Cantidad de desaparecidos en el desastre
heridos	Cantidad de heridos en el desastre
damnificad	Cantidad de afectados por el desastre
evacuados	Cantidad de evacuados por el Sistema de Emergencias, Policía, Bomberos)
autoevac	Cantidad de autoevacuados (gente que abandonó la zona por sus medios)
observac_1	Detalles acerca de la población afectada por el evento
viviendas	Cantidad de viviendas afectadas
viviendas_1	No aplica, columna casi sin datos
cultivos	Medición en hectáreas de cultivos afectados
ganado	Cantidad de ganado afectados
vias_afect	Metros de vías afectadas

valor_perd	Valor estimado de las pérdidas a causa del desastre
valor_pe_1	No aplica, columna casi sin datos
otras_perd	Descripción textual y más detallada de las pérdidas ocasionadas por el evento
transporte	No aplica, columna casi sin datos
edificios	No aplica, columna casi sin datos
energia	Numero de viviendas afectadas por corte de energia
comunicaci	No aplica, columna casi sin datos
acueducto	No aplica, columna casi sin datos
otros	Descripción de consecuencias a nivel de infraestructura del desastre
autor	Encargado de registrar el desastre

Tabla 1. Datos SINAE.

UTE

UTE proporcionó datos en archivos con formato excel. Dichos archivos se describen a continuación.

- **Niveles y caudales:** Dos archivos vinculados al registro de los niveles y caudales de Ríos del Uruguay. Cada archivo contiene los campos detallados en la tabla 2.

Columna	**Descripción**
Paso	Nombre en formato de sigla código del paso donde se realizó la medición
Fecha	Fecha de la medición
Hora	Hora de la medición

Cod.Disp	No corresponde, campo mayormente vacio
Altura	Medición en metros de la altura del río en el paso
Caudal	Registro del caudal. Dato desestimado para el trabajo.
Observación	No corresponde, campo mayormente vacio

Tabla 2. Datos niveles y caudales UTE.

- **Precipitaciones:** Cuatro archivos que describen las precipitaciones en distintos puntos del Uruguay. Cada archivo contiene los campos descritos en la Tabla 3.

Columna	**Descripción**
Fecha	Fecha en que se registró la precipitación
Nro Oficial	Código utilizado como identificador de la estación
Codigo	Código en sigla de la estación
Localidad	Nombre de la localidad donde se realizó la medición
Medida	Medición de las precipitaciones en milímetros
Cod Disp	No corresponde, campo mayormente vacio
Observación	No corresponde, campo mayormente vacio

- **Estaciones :** Archivo excel que contiene la ubicación de las estaciones en el Uruguay que UTE posee. Las columnas se describen en la Tabla 3.

Columna	**Descripción**

No. Oficial	Identificador de la estación
Nombre de estación	Nombre de la estación, como indica la columna
Operador	Ente cargado de manejar la estación
Tipo	Tipo de estación (automática o convencional)
Latitud	Coordenada de latitud de ubicación de la estación
Longitud	Coordenada de longitud de ubicación de la estación

Tabla 3. Datos estaciones UTE.

INUMET

Se basa en un excel que cuenta con 10 hojas, donde cada una de éstas representa los registros de una estación meteorológica particular de INUMET. A su vez, cada hoja cuenta con tres columnas que son las descritas en la Tabla 4.

Estación	Estación en la cual se midieron las precipitaciones. Dicho valor es el mismo para todas las filas dentro de cada hoja de cálculo.
Fecha	Corresponde a la fecha en que se registró la precipitación. Dicha fecha identifica el año, mes y el día, siendo la hora y el minuto datos omitidos ya que cada estación de INUMET mide la precipitaciones a las 7:00 AM cada día.
Mm:	Cantidad de milímetros de agua registrados.

Tabla 4. Datos INUMET.

Cada fila corresponde a una precipitación registrada en una fecha y estación particular.

EM-DAT

Se obtuvo acceso a la base de datos [62], la cual registra distintos desastres naturales ocurridos en el Uruguay. Se registra un total de 15 inundaciones donde cada una de ellas presenta la información descrita en la Tabla 5.

Atributo	**Descripción**
Start Date	Fecha de comienzo del desastre
End Date	Fecha de finalización del desastre
Country	País donde ocurrió el evento
Location	(En Uruguay) Departamento/s donde ocurrió el evento
Dis Type	Tipo de desastre (Inundación, terremoto, tornado, etc)
Total Damage (USD)	Costo de los daños en dólares
Dis SubType :	Subtipo del desastre (tipo de inundación por ejemplo)
Total Death :	Cantidad de fallecidos
Total Affected :	Cantidad de afectados
Source :	Origen del dato

Tabla 5. Datos EM-DAT.

5.2 Obtención Dinámica de Datos

Fue necesario la realización de un trabajo en paralelo de obtención dinámica y automática de datos que complementen la información estática conseguida a través de los entes mencionados anteriormente (INUMET, SINAE, UTE, EM-Dat). La motivación detrás de la realización de esta tarea fue la siguiente:

- Complementar los datos históricos conseguidos, buscando conseguir registros de temperaturas de años anteriores.

- Buscar una forma de lograr mantener actualizada la base de datos con datos actuales, a modo que con el paso del tiempo no haya un salto o ausencia de información entre la fecha actual y la fecha del dato más reciente.

Para llevar a cabo este trabajo se abordaron dos posibilidades con el fin de conseguir distinta información: solicitudes a APIs de servicios meteorológicos y web-scraping en sitios de interés. Buscando además, la posibilidad de que para cumplir con el segundo objetivo, se logrará automatizar el proceso y así realizar una ejecución diaria del mismo.

5.2.1 Requerimientos de nivel de calidad de los datos

Como se mencionó en capítulos anteriores, se requiere cierto nivel de calidad para los datos que lleguen al almacén de datos central del trabajo. Es así que este requerimiento se traslada directamente sobre los datos obtenidos de manera dinámica.

Estos datos ya deben llegar a la base de datos cumpliendo con los requerimientos de calidad de datos mínimos establecidos (ver secciones 5.3.1 y 5.3.3), de forma que éstos coincidan exactamente con el formato definido para la misma.

5.2.2 Fuentes

Para el cumplimiento de los requerimientos establecidos, se buscaron fuentes públicas, gratuitas y de confianza dada por su popularidad que nos pudieran suministrar lo deseado. Es así que se dio con la siguiente lista de fuentes.

Weather Underground

Organización comercial que provee servicios meteorológicos en todo el mundo. Su oficina principal se ubica en San Francisco, California. La Organización brinda información meteorológica desde el año 1993 [63] y fue adquirida por IBM en el año 2015 [64].

Toda la información generada por dicha organización es propia de su sistema de pronosticación llamado BestForecast [65].

Weather Underground cuenta con personal en distintas estaciones de Estados Unidos y en diferentes partes del mundo, donde más de 180.000 personas se ubican en estaciones dentro de Estados Unidos y alrededor de 29.000 personas se encuentran en estaciones de distintos países [65].

Weather Underground provee una interfaz pública mediante la cual los usuarios pueden consumir y obtener distintos datos meteorológicos y datos de geolocalización. Para la utilización de la misma, basta con crear una cuenta en su sitio web oficial. Cabe mencionar que existen 3 tipos de cuentas, que ofrecen distintos niveles de información, siendo el nivel más bajo totalmente gratuito. Dicho tipo de cuenta ofrece servicios de clima, geolocación, y hasta un límite de 3 días de pronóstico. El mismo posee un límite de 10 pedidos HTTP por minuto [66].

AccuWeather

Empresa estadounidense que brinda servicios de pronosticación en todo el mundo. La oficina central se encuentra en State College, Pensilvania. Además, la empresa cuenta con otras grandes oficinas en Nueva York y en Fort Washington, Pennsylvania [67].

AccuWeather permite la integración de terceras aplicaciones mediante una interfaz pública [68]. Dicha interfaz es limitada, teniendo un mínimo de pedidos dependiendo del plan que el usuario posee [69]. A grandes rasgos el plan gratis que brinda la aplicación permite realizar un total de 500 pedidos por usuario.

Dicha interfaz provee información de localización, pronosticación del tiempo, estado del tiempo actual, así como también alertas meteorológicas en distintas partes del mundo. La aplicación no brinda registros de datos históricos.

World Weather Online

Dedicada a proveer información climática y de pronosticación a distintas compañías y páginas web. Entre la información que provee se destaca [70]:

- Condiciones marítimas
- Registros históricos del clima
- Pronósticos del tiempo
- Alertas climatológicas

La información obtenida proviene de distintas partes del mundo, particularmente, cubre aproximadamente 3 millones de ciudades [70].

La información de pronósticos del tiempo es utilizada por una gran variedad de clientes incluyendo Intelligent Plante, Sonitus Systems, Weathercare entre otras [70].

World Weather cuenta con dos centros de data ubicados en Dinamarca y Alemania, mientras que sus oficinas se encuentran en Reino Unido, Estados Unidos e India [70].

Dicha compañía brinda una interfaz que permite a los desarrolladores acceder al contenido que esta provee. Al igual que los servicios descritos anteriormente, este cuenta con ciertos planes que se describen en su web principal [71].

UTE-i

Sección de UTE dedicada a proveer información acerca de gestión, consumo, facturación y estado de los servicios, entre otros [72].

El nombre con la “i” corresponde a una serie de conceptos por los cuales se identifica UTE-i.

Dentro de las secciones de información que brinda el servicio, resultan de utilidad para el proyecto los datos referidos a los registros de altura y de precipitaciones. Estos datos son registrados varias veces diariamente en una amplia cantidad de estaciones, la mayoría

concentradas en el Río Negro. Estos datos son recolectados de forma manual o automática, según la estación y los mismos se presentan en tablas.

Debido a que UTE, como se menciona anteriormente, es una empresa del estado, cuya información es oficial y que el mismo es el encargado de administrar la mayor parte de las estaciones pluviométricas y de registro de altura de ríos, se consideró como prioritario encontrar una forma de obtener los datos actuales del servicio para mantener alimentada la base de datos con datos diarios del mismo que siguieran la misma línea y estructura de los registros históricos proporcionados por el ente.

Tutiempo.net

Web donde es posible consultar pronósticos del tiempo para los próximos 15 días, así como también cuenta con un registro histórico que varía en su antigüedad según la ciudad consultada o el aeropuerto (accesibles mediante su código ICAO).

Como contrapartida, la web no brinda la posibilidad de consumir sus datos mediante una API y sus administradores aclaran que no están dispuestos a compartir los archivos XLS que hacen de base de datos [73] ni el acceso al tiempo mediante XML como realizaban antiguamente, al menos al momento de consultar [74].

Debido a esto último, con la finalidad de fortalecer el registro histórico de la base de datos en el apartado de registro de temperaturas, mediante web-scraping se obtuvo datos diarios que van hasta 1980 para diversos puntos del país.

A su vez, también se obtuvo el registro climático desde 2013 para nueve aeropuertos de Uruguay.

Es importante aclarar que los datos brindados por el portal son reales y obtenidos mediante un programa especial de la Organización Meteorológica Mundial (WMO). La base de datos del clima cuenta con más de 118.000.000 de registros climáticos a nivel mundial, con un registro que en algunos casos va hasta el año 1929, siendo la fuente de estos las estaciones de la Organización Meteorológica Mundial [75].

5.2.3 Análisis, Diseño e Implementación de la obtención dinámica de datos

Para entrar en detalles acerca de la implementación de la obtención de datos, es importante distinguir con claridad que la misma se realizó tanto mediante el consumo de APIs así como utilizando Scrapy para la extracción de datos. A continuación se describen también las herramientas utilizadas para llevar a cabo la tarea.

Python

Python es un lenguaje interpretado, de alto nivel y orientado a objetos. Debido a su facilidad de uso, a su manejo dinámico, al hecho de que no es un lenguaje desarrollado para un ambiente específico (se utiliza para desarrollo de sitios webs, aplicaciones, scripts, consumo de APIs, etc), en los últimos años ha ganado mucho mercado, resultando

especialmente atractivo para el desarrollo de aplicaciones de mediano y pequeño porte [76].

Como consecuencia de este atractivo para los desarrolladores en la actualidad, es que se ha ido formando una amplia comunidad en la cual la gente intercambia, comenta, colabora con el desarrollo de nuevas librerías y se ayuda, en la cual se puede encontrar respuesta a casi cualquier duda que pueda surgir a la hora de desarrollar una aplicación [76].

Para el proyecto, se decidió utilizar Python considerando el proyecto en toda su extensión, es decir que se tomó dicha decisión con la finalidad de buscar utilizar un único lenguaje (del cual se tenía un conocimiento medio) que resultase oportuno para la implementación de scripts, para la obtención y extracción de datos en la web, para el modelo predictivo mediante aprendizaje automático y la implementación de una aplicación web final. Esto se debe a que cuenta con librerías que se detallarán en las siguientes secciones, que están muy bien documentadas y recomendadas, como es el caso de Scrapy que será descrita en el siguiente párrafo, scikit-learn, además de la presencia de un framework orientado al desarrollo web como es Django. Asimismo, estas librerías gracias a que tienen gran uso de parte de la comunidad, siendo todas las más usadas en su cometido (a excepción de Django), tienen una amplia documentación y posibilidad de consulta en base a experiencias, preguntas e ideas de usuarios anteriores, el cual sirve de base y ayuda a la hora de comenzar un proyecto que involucre su uso.

Scrapy

Scrapy es una librería de código abierto, escrita en Python; la misma proporciona un framework para la descarga (crawling) y extracción (scraping) de datos de sitios web.

Un proyecto implementado bajo este framework consta de una estructura de proyecto propia donde resaltan tres componentes principales: Spiders, Ítems y Pipelines [77].

Los Spiders son clases, piezas de código y módulos escritos con el fin de descargar el código fuente de la URL objetivo, y recorrer la información del sitio web descargado, para luego depositarlo en los Ítems. Por cada sitio web que se planea scrapear, se implementa una spider propia para ese sitio, ya que en estas vienen dadas las reglas de qué componentes y elementos se deben scrapear (por ejemplo mediante tags HTML o CSS) [78].

Los Ítems son los contenedores de la información resultante de la extracción mediante Spiders. Es un DSL mediante el cual se define la estructura de los datos a extraer. Generalmente su estructura coincide con la utilizada por una eventual salida de los datos en un archivo o en consola, o con una tabla a nivel de base de datos [78].

Los Pipelines se encargan de procesar al Ítem una vez que fue extraído por una Spider. Por cada Ítem se implementa un Pipeline que consta de un método encargado de procesar, evaluar ciertas condiciones o descartar el Ítem. Generalmente se usan para limpiar datos HTML, validar datos, búsqueda de resultados duplicados o para almacenar en una base de datos [78].

Se optó por utilizar esta librería como herramienta para realizar la extracción de datos en las webs UTE-i y tutiempo.net ya que se contaba con experiencia en su manejo, además la misma resultaba adecuada por tener una estructura de proyecto fácil de comprender y manejar, contando con buena documentación de base y por ser una librería de Python, su integración con el proyecto principal Django resulta más sencilla, permitiendo importar modelos del mismo en el proyecto Scrapy.

Obtención de datos vía APIs

Se dividirá también en descripciones sobre la implementación según cada una de las fuentes para ser consistentes con el punto anterior.

Weather Underground

Por un lado se creó un worker (script que corre de manera continua o cuando es invocado) que es ejecutado una vez por día el cual obtiene de la API distintos factores del clima como lo son, máxima temperatura, mínima temperatura y precipitaciones registrados el día anterior. Estos datos son almacenados en nuestra base de datos, para procesarlos posteriormente.

AccuWeather

Se realizaron consultas a la interfaz para la obtención de los distintos puntos del Uruguay en el cual el servicio brinda datos de pronosticación. Una vez que se obtuvieron las localidades, se almacenaron en la base de datos las estaciones correspondientes a esas ubicaciones obtenidas.

Se creó un script que obtiene información del pronóstico del tiempo para los siguientes cinco días en las localidades obtenidas anteriormente. El mismo es utilizado para importar información relevante del estado del tiempo del día en el que se desea predecir un evento, así como también la predicción meteorológica para esos días, siempre y cuando no se tengan ya.

World Weather Online

Se crearon dos scripts, el primero recolecta información histórica del clima en distintos puntos del Uruguay, particularmente se encontraron 19 puntos, uno por cada departamento. El segundo, recolecta información de pronósticos del tiempo de los próximos 5 días al día consultado. El primer script fue ejecutado una primera vez para la obtención de datos climáticos del Uruguay desde el año 2008 hasta el día en que fue ejecutado, una vez obtenidos los datos se guardaron en nuestra base de datos, a su vez el script es ejecutado diariamente para la obtención del registro del clima del día anterior manteniendo el registro del clima al dia en nuestra base de datos.

El segundo script es ejecutado manualmente con el mismo fin que el mencionado anteriormente para AccuWeather, esto es, obtener información del estado del tiempo para una fecha en concreto y las predicciones para los días siguientes.

Implementación de la obtención de datos mediante APIs

Con el fin de mantener una estructura y formato unificados para los datos que entren en la base de datos y la facilitación de la incorporación de una nueva entidad meteorológica, se desarrollaron scripts en lenguaje Python en los cuales se utilizó un patrón similar al patrón Adapter [79] (ver figura 10) para implementar el acceso a las distintas APIs, que individualmente todas poseen distintas formas con respecto a cómo obtener datos de las mismas; no todas tienen las mismas caracteristicas y a su vez retornan un formato propio para los datos, y mediante el patrón se expuso una única interfaz de llamada de los métodos, y un normalizador de los datos que se reciben de cada una que transforma los mismos al formato unificado definido.

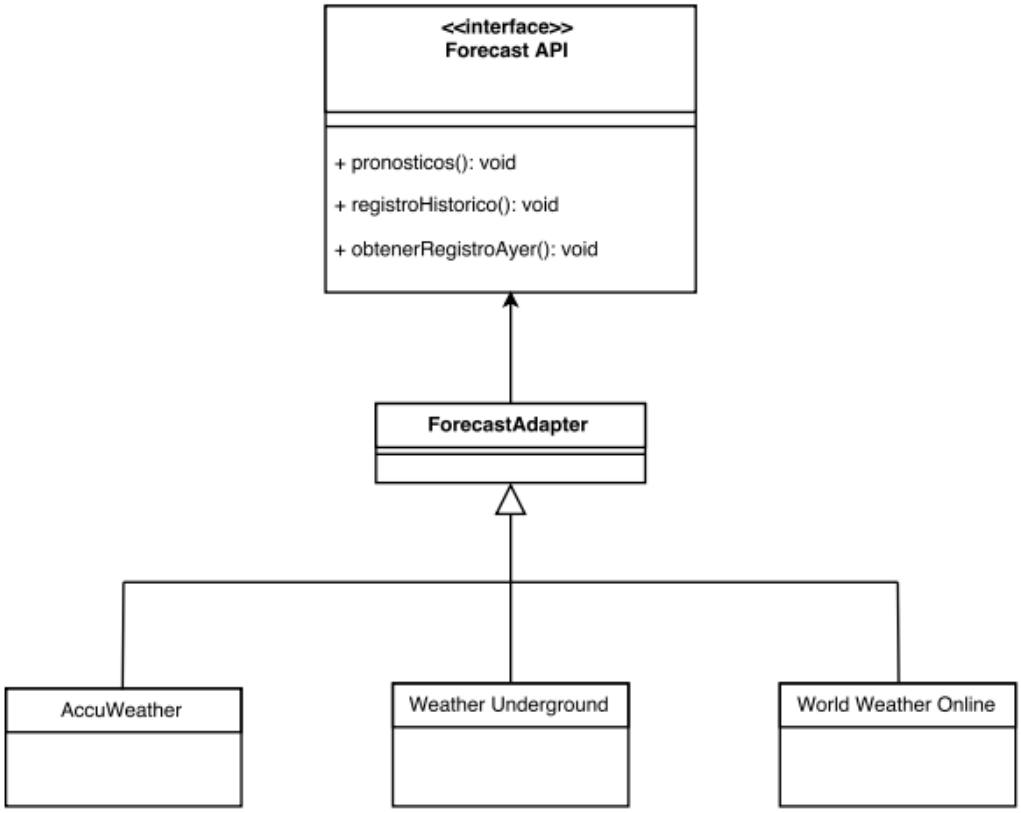


Figura 10. Patrón de diseño aplicado a la obtención de datos de las distintas APIs.

Obtención de datos vía scraping

Para realizar la extracción de datos de los sitios webs de UTE-i y TuTiempo.net se creó un proyecto scrapy. Mediante el mismo los datos que se deseaban obtener eran: precipitaciones mayormente en el Río Negro así como la medición altura del agua en distintos puntos del mismo; y por otra parte el registro histórico de estados del tiempo que hay disponible en TuTiempo.net para más de quince puntos de Uruguay. El proyecto consta de tres spiders, cada una con la función de obtener uno de los datos anteriormente mencionados. Además por cada spider, hay definido un ítem y un pipeline, que se encarga de mapear el objeto almacenado en el ítem en la base de datos de la manera esperada.

Finalmente este proyecto scrapy fue embebido dentro del proyecto principal django con la el objetivo de publicar un único proyecto en la nube y que la comunicación entre ambos fuese lo más sencilla posible.

Con el fin de explicar de forma un poco más detallada la extracción de los datos, se separa la implementación en dos secciones, la correspondiente a UTE-i ya que ambas spiders son muy similares y una para TuTiempo.net

UTE-i

Tanto para las precipitaciones como para las alturas, la web de la UTE presenta la información en tablas similares, por ende el trabajo realizado en cada spider fue muy similar. Una vez estudiado el código fuente del sitio, se procedió a iterar sobre la tabla principal donde se encuentran los datos, accediendo a la misma mediante selectores HTML [80] principalmente, estos funcionan descargando el código fuente del sitio web, inspeccionando el DOM del sitio. En el caso particular de las tablas, se puede acceder directo a las mismas gracias a que tienen ids (ver figura 11), que tienen que ser únicos y así facilitan el acceso a sí mismas. La iteración se realizó sobre la primer fila con el objetivo de obtener el nombre de las diferentes estaciones y sobre la siguiente fila para obtener los milímetros de precipitaciones en las últimas veinticuatro horas.

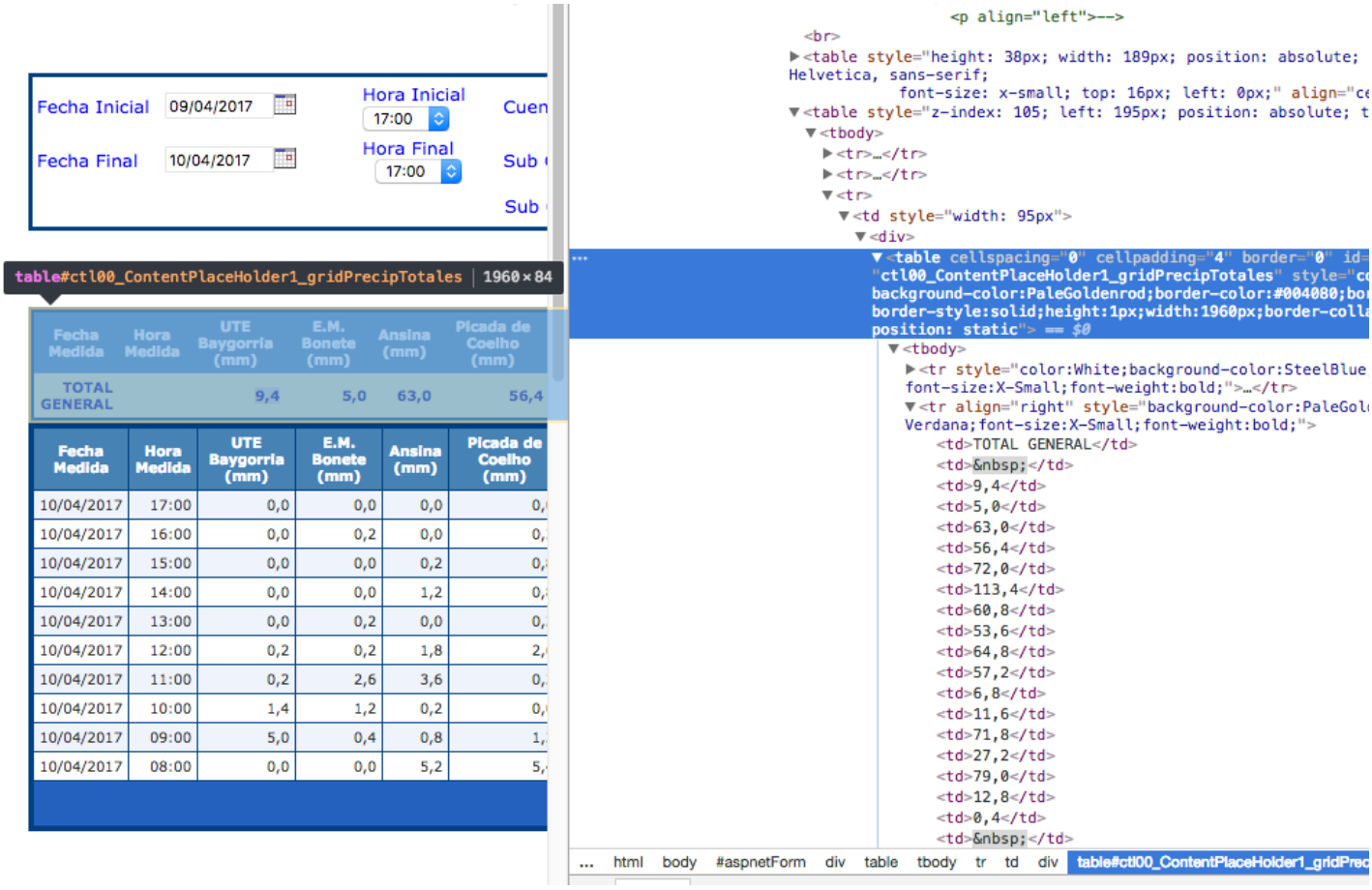


Figura 11. Tabla de precipitaciones de UTE y su código fuente

Una vez obtenidos los datos, los mismos son almacenados en memoria en objetos ítems que tienen únicamente campos scrapy sin tipo definido; finalmente estos objetos son almacenados en la base de datos mediante un pipeline específico para cada ítem. Tanto para el pipeline de precipitaciones como para el de alturas, se definió un diccionario de estaciones, para mapear el nombre de la estación como viene dado en la web con su id en la base de datos y el objeto en sí es mapeado con objetos importados de Django que son los que definen el esquema de la base de datos.

Cada spider es ejecutada de manera automática desde el servidor Heroku. Con el objetivo de tener los datos lo más completo posible, dichas ejecuciones se realizan cada una hora, ya que no todas las estaciones publican datos sobre la altura o las precipitaciones en todas las horas. Se almacena únicamente un registro por cada fecha y estación, descartandose los datos obtenidos para una estación en la que ya se obtuvo un dato en otra corrida anterior para ese día.

TuTiempo.net

El concepto para este caso fue el mismo, la spider accede a una tabla que contiene varios datos acerca del estado del tiempo, las tablas corresponden al mes de un lugar y un año dado. En este caso para lograr iterar sobre todos los meses, se formó un conjunto de URLs de comienzo para scrapear, correspondientes a todas las combinaciones de meses, años y lugares disponibles en el sitio. El número total de URLs a los que debe acceder la spider es de aproximadamente 12 mil.

El acceso a la tabla principal se logra de forma directa gracias a que la misma tiene una clase única. En la misma se itera sobre todas las filas excepto la primera que tiene los títulos y la última que tiene la media del mes, obteniendo de cada una las temperaturas mínima, máxima y la precipitación junto al día del mes.

La información de la tabla se complementa con la fecha (años-mes) que viene dada en la URL y el lugar que se encuentra en un elemento *html h2*. Ver figura 12.

Para cada día del mes, se genera un item que es almacenado en la base de datos mediante un procedimiento análogo al explicado en el punto anterior.

Esta spider es ejecutada una única vez con la finalidad de recabar un amplio conjunto de datos históricos.

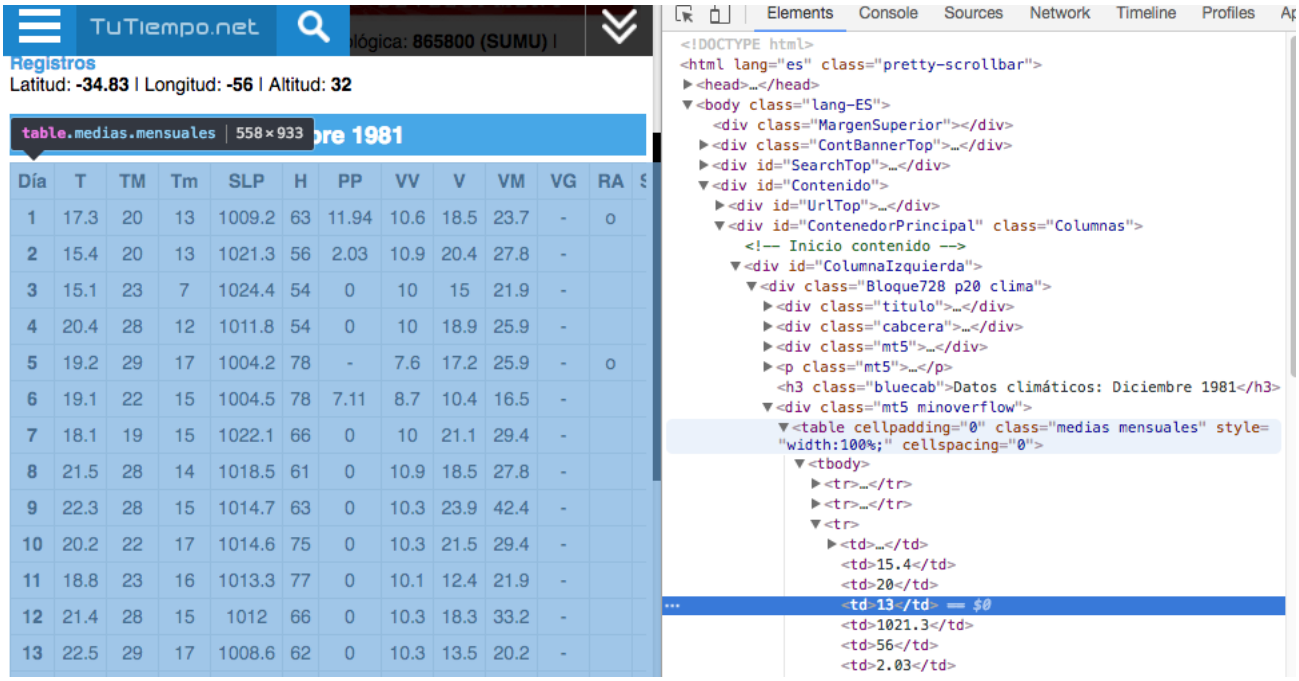


Figura 12. Tabla de precipitaciones de UTE y su código fuente

5.2.4 Problemas encontrados en la obtención de datos

Cada una de las APIS maneja sus propias métricas y dimensiones por lo que se tuvo que procesar los datos de manera de hacerlos uniformes. A su vez, no todas las APIS proveen los mismos servicios, a modo de ejemplo AccuWeather carece de un servicio de obtención de datos meteorológicos históricos.

Para prevenir sobrecarga de pedidos, las APIs proveen un máximo de pedidos en una determinada cantidad de tiempo definido. Es por ello que en caso de que se necesitan realizar más pedidos de lo permitido, se efectúan cada cierto tiempo respetando el máximo de requests por tiempo.

En particular, un problema que presenta la obtención de datos mediante web-scraping, es la fragilidad que presenta ante los cambios. Esto significa que si en el futuro la página de la UTE realiza cambios en cómo se presentan los datos, esto implicaría una reescritura del código scrapy ya que es muy probable que el mismo deje de funcionar si se cambia por ejemplo, el id de la tabla o el nombre de las estaciones.

Este problema no es menor si se considera la posibilidad de que la aplicación tenga un uso a largo plazo.

5.3 Data Warehouse y Calidad de Datos

En el caso del presente proyecto, se optó por implementar un almacén de datos con forma de una única base de datos central manejada por Postgres y a la cual vierten sus datos los diversos métodos de obtención de datos empleados explicados en el capítulo anterior.

Esta base de datos aloja el total de la información recabada en la etapa de obtención de datos descrita en el capítulo anterior, y representa la totalidad de la información a disposición del trabajo de análisis predictivo que le sigue.

5.3.1 Requerimientos para el DW

La misma definición vista en el punto anterior de un DW, presenta requerimientos que fueron exigidos a la base de datos que se busca generar. Esto es, debe ser no volátil, íntegra, variante y orientada a un tema.

Además, es oportuno introducir en este momento el concepto de **calidad de datos** [81]. La calidad de los datos es un requerimiento para el almacén a construir, ya que hay propiedades deseables que se pretenden establecer y/o mantener.

En este contexto, el concepto de calidad de datos, hace referencia a procesos, técnicas y operaciones destinadas a mejorar la calidad de los datos almacenados, con el fin de que éstos cumplan el propósito para el cual están siendo almacenados.

La calidad de los datos se suele evaluar teniendo en cuenta distintas dimensiones o propiedades de éstos. Estas son algunas dimensiones sobre las cuales este proyecto puso énfasis con el objetivo de que efectivamente los datos almacenados cumplan su función:

Exactitud: La exactitud tiene que ver con la cercanía que tiene un valor real y su representación en los datos almacenados. Cuanto mayor sea esta cercanía, más exactos sean los datos, y por consiguiente más útiles serán para el cometido del sistema.

Unicidad: La unicidad se relaciona con la representación duplicada de entidades del mundo real en su representación de datos, con distintas claves. Además, puede existir contradicción si además de estar un dato duplicado, éstos duplicados tienen distintas características, creando ambigüedad. Lo ideal sería que cada entidad del mundo que se quiera representar tenga una única instancia en la base de datos.

Completitud: La medida en que los datos son de suficiente alcance y profundidad. Tiene dos aspectos, la cobertura, que es la porción de los datos que se encuentran contenidos en el sistema, y la densidad de la misma, que es la cantidad de información contenida, y la faltante. Por ejemplo, los valores nulos constituyen una causa frecuente para la baja completitud de un conjunto de datos.

Consistencia: Cumplimiento de las reglas semánticas que son definidas sobre los datos. Que éstos representen realmente a la entidad del mundo real que se quiera.

Actualidad: Actualización de los datos y su vigencia.

A continuación se destacan los principales factores y causas identificados por los cuales los datos se consideran “sucios” o de baja calidad :

- El principal factor identificado es que las fuentes de datos relevados no son sistemas de bases de datos, sino que son archivos escritos manualmente en hojas de cálculo.
- Los datos relevados provienen de distintas fuentes externas, que si bien manejan los mismos conceptos, los datos son almacenados de distintas maneras, es decir son heterogéneos.
- Falta de metadatos, ausencia de restricciones.
- Errores de digitación.
- Datos repetidos.
- Datos omitidos.

5.3.2 Análisis y Diseño del DW

Es así que para cumplir lo requerido, se presenta el diagrama de diseño de la Base de Datos (ver figura 13).

Modelo de Dominio

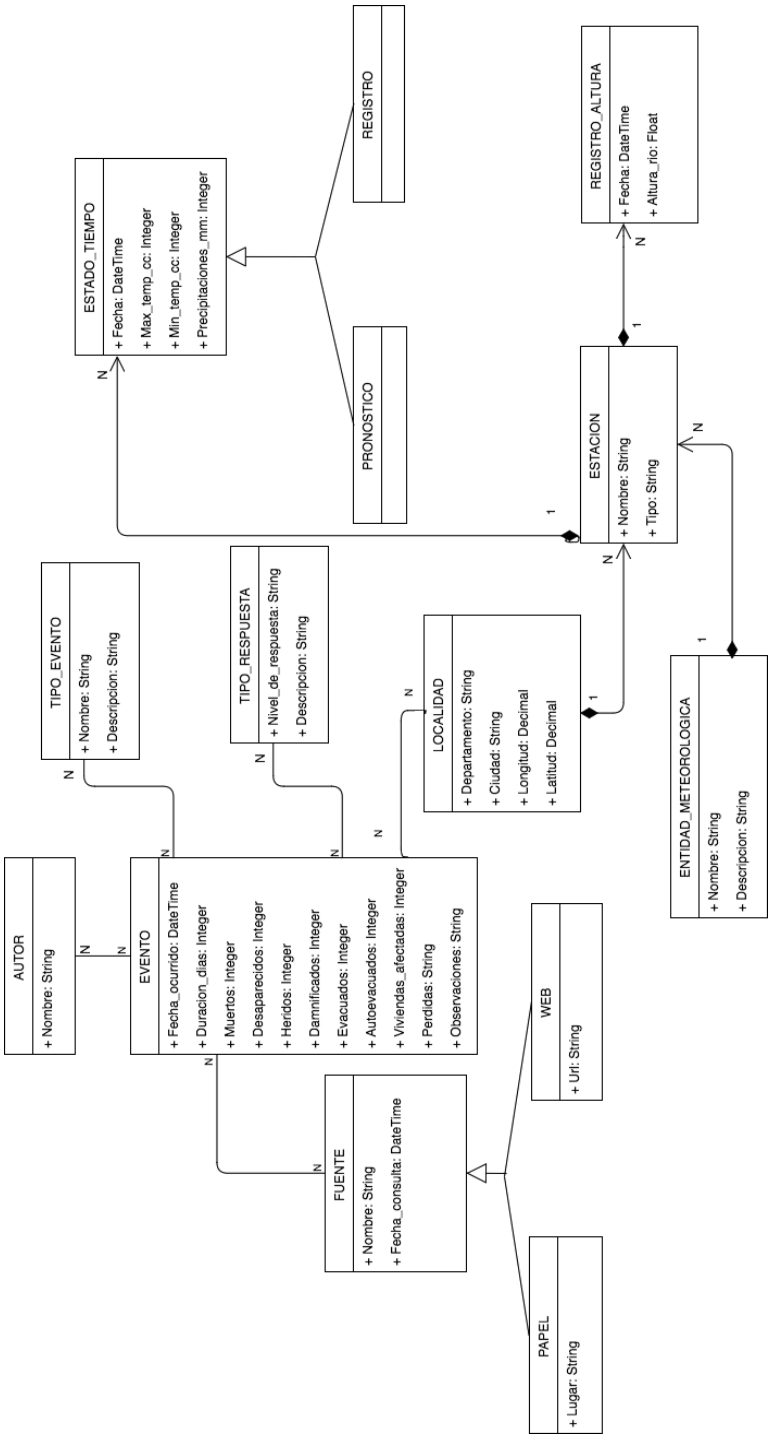


Figura 26. Modelo de Dominio de la aplicación

Se describe a continuación en la tabla 6 cada una de las clases, no así los atributos ya que se consideró que es más clara la explicación de estos en el esquema de la base de datos.

Clase	**Descripción**
Autor	Representa los autores que fueron encargados de registrar el evento. Un evento puede ser creado y/o editado por varios autores.
Fuente	Fuente de donde es extraída la información correspondiente al evento meteorológico. Se tienen dos tipos de fuentes, Web y Papel. En el caso de que la fuente sea Web, se registra la URL y en el caso que sea papel, se registra el lugar del mismo. Una fuente puede registrar varios eventos.
Evento	Provee la información relacionada con el evento, como la fecha que ocurrió, duración días, observaciones y sus consecuencias (muertos, desaparecidos, heridos, damnificados, evacuados, autoevacuados, viviendas afectadas y perdidas). A su vez, un evento es creado editado por varios autores, tiene una o más fuentes, uno o más tipos de eventos, uno o más tipos de respuestas y por último una o más localidades.
Tipo_Evento	Tipo de evento ocurrido. Se registran dos tipos: Tormenta e inundaciones. Cada tipo de evento puede tener más de un evento asociado.
Tipo_Respuesta	Tipo de respuesta efectuada ante el evento meteorológico ocurrido. Cada tipo de respuesta puede tener más de un evento asociado.
Localidad	Localidad en donde ocurrió el evento. En una localidad pueden registrarse varios eventos, a su vez en una localidad puede ubicarse más de una estación.
Estación	Estaciones meteorológicas en el Uruguay. Cada estación meteorológica pertenece a una entidad meteorológica y a una localidad y se encarga de brindar datos históricos y de pronosticación del clima registrando estados del tiempo y altura de los ríos.
Estado Tiempo	Estado del tiempo registrado. Puede ser tanto un pronóstico como un registro histórico del clima. Un estado tiempo es registrado por una estación.
Entidad_Meteorologica	Entidad meteorológica detectada. Se encontraron un total de cinco entidades meteorológicas que son: UTE, INUMET, Wunderground, Accu Weather y World Weather. Una entidad meteorológica tiene varias estaciones ubicadas en el Uruguay.

Registro_Altura	Registro del nivel de caudal de los ríos. Es registrado por una estación.

Tabla 6. Clases del DW.

5.3.3 Implementación

A continuación se describe las herramientas y la metodología utilizadas para el procesamiento de los datos obtenidos y la creación de una base de datos centralizada en donde se almacenan los datos obtenidos ya procesados.

OpenRefine

OpenRefine es una herramienta de código abierto para el manejo de datos. Fue desarrollada por la empresa Metaweb Technologies y fue apoyada por Google bajo el nombre de Google Refine. En el año 2012 Google decidió no darle más soporte al proyecto, por lo que desde ese entonces la misma fue mantenida por un conjunto de desarrolladores en un repositorio de código abierto bajo el nombre OpenRefine [82] [83].

Las funcionalidades de dicha herramienta van desde limpieza, organización y transformación en diferentes formatos, hasta la posibilidad de extender los datos a través de web services y relacionar con bases de datos como Freebases [84].

OpenRefine también permite la manipulación de datos mediante expresiones que pueden ser escritas en Python o en GREL (General Refine Expression Language) lenguaje creado por los desarrolladores de la herramienta [85].

RefinePro

Herramienta que provee instancias de máquinas virtuales en la nube, en la cual en cada una de ellas se está ejecutando el servicio OpenRefine. RefinePro permite la manipulación de OpenRefine en la nube, brindando la posibilidad de que varios usuarios puedan acceder a la herramienta de manera colaborativa en tiempo real [86].

Calidad de Datos

Lamentablemente, los datos históricos recibidos de parte de las distintas entidades no poseían una calidad de datos aceptable en los términos descritos en la sección 5.2.1. En particular aquellos relacionados a los eventos sucedidos en el pasado obtenidos en formato Excel desde el SINAE fueron los que presentaron mayores deficiencias. De esta manera resulta conveniente usamos estos datos como ejemplo para explicar la metodología aplicada.

Con el fin de imponer las condiciones mínimas de calidad de datos anteriormente mencionadas, se usaron OpenRefine y RefinePro (ver figura 14) [86] en primera instancia sobre la fuente de datos sobre eventos en formato Excel, por conveniencia y comodidad en cierto tipo de operaciones sobre las columnas, y se termina de procesar los datos mediante SQL para otro tipos de operaciones, principalmente aquellas con carga semántica sobre los

datos a manejar y que por lo tanto requerían consultas complejas para establecer relaciones entre los mismos.

En OpenRefine, se cargaron los datos en formato Excel de manera directa ya que la herramienta posee una funcionalidad para ello.

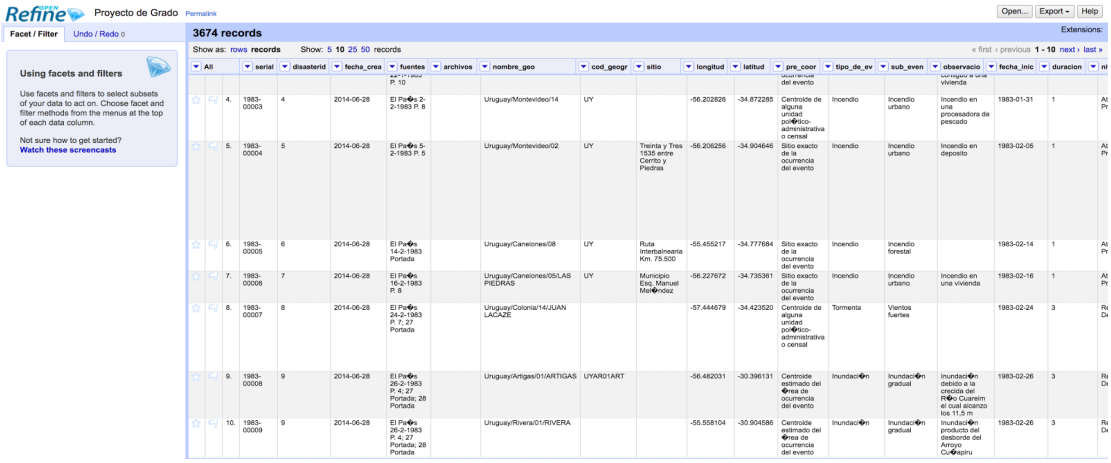


Figura 14. Open Refine: Datos de SINAE cargados.

Una vez que los datos fueron cargados a la herramienta y era posible visualizarlos y manipularlos desde la plataforma, se comenzó con el tratamiento de los mismos. Dado que se procedió de una forma similar para las distintas columnas, se detalla solamente el procedimiento realizado a la columna localidades.

Una falencia destacable de los datos era su poca precisión en cuanto a la nomenclatura de las localidades geográficas donde estaban geolocalizados, así como otras columnas. Existían diversas versiones para varios de los nombres de las localidades dependiendo de los errores de escritura y el formato en que estos son escritos según la persona a cargo de ingresarlos en el archivo Excel. A modo de ejemplo, algunos de los formatos encontrados fueron: /pais/departamento/ciudad, departamento/ciudad ó solo texto sin seguir ningún patrón como por ejemplo “Las piedras Canelones”, “Canelones Las Piedras”, “Can, Las Piedras”.

Algunos de ellos con mayúsculas, otros sin ninguna o algunas de ellas. Algunos con tildes (los tildes representan un punto común de errores de precisión cuando se trata con textos del idioma español), y otros sin tilde, o con tildes en algunas de las letras donde correspondía. Claramente se enfrentó a un problema de pobre unicidad y consistencia de los datos.

Para atacar estos problemas, se usó la funcionalidad de clusterización provista por la herramienta. La clusterización permite, mediante la identificación de patrones, agrupar filas de información bajo determinado criterio sobre un campo particular de las filas. En esta problemática en particular, se procedió a generar clusters utilizando distintos algoritmos de clusterización soportados por la herramienta que se describe a continuación:

- **Colisión de claves**: Este método se basa en la idea de crear una clave para cada uno de los valores, donde en este caso particular los valores son cadenas de caracteres. Luego se agrupan aquellas sentencias que comparten una misma clave [87]. Para la creación de las claves se utilizaron distintas funciones descritas a continuación:

    - **Huella dactilar**: Es un simple y rápido método el cual se dice que es la función que produce menor cantidad de falsos positivos en promedio [87]. Dicho método sigue la siguiente serie de pasos :

        1. Remueve los espacios en blanco que están al inicio y al final de la sentencia.
        2. Cambia todos los caracteres a minúscula.
        3. Remueve todos los símbolos de puntuación y caracteres de control.
        4. Dada la sentencia, se crea un arreglo donde cada celda corresponde a una palabra ubicada en la sentencia, siguiendo el mismo orden.
        5. Se ordena el arreglo de palabras alfabéticamente.
        6. El arreglo se convierte en una cadena de caracteres separando cada celda por un espacio en blanco.
        7. Por último se normalizan los caracteres occidentales extendidos como por ejemplo "gödel" → "godel".

    La clave es generada una vez que la sentencia pasa por los 7 pasos anteriores, todos ellos en el orden presentado anteriormente.

    - **N-Gram Huella Dactilar**: Este algoritmo es muy similar al de huella dactilar solo que éste recibe un entero como parámetro (n-gram) y en el paso 4 el arreglo es creado por cadenas de caracteres del largo del parámetro ingresado, es decir, si se tiene la sentencia “Hola, cómo estás” y n-gram es 2 el arreglo generado va a ser el siguiente: [Ho, la, có, mo, es, tá, s]. Los restantes pasos iguales [87].

    A su vez, la herramienta posee otras funciones de obtención de claves que utilizan la similitud entre la pronunciación de las sentencias, es decir si dos sentencias se pronuncian muy parecidamente, siguiendo un tolerancia definida por la herramienta, tendrán la misma clave. Sin embargo, dichas funciones no se pudieron utilizar ya que no soportaba el idioma español.

- **Vecino más cercano**: Dado que el algoritmo de colisión de claves es muy estricto ya que para que dos sentencias se agrupen tiene que tener la misma clave, se procedió por utilizar uno más flexible como lo es el algoritmo del Vecino más cercano. Dicho algoritmo tolera cierta diferencia en las sentencias fijando un límite ingresado por el usuario (conocido como radio o k). Cualquier par de sentencias que difieran en menos que el radio serán agrupadas. Una gran deficiencia encontrada en este algoritmo con respecto al anterior es que es muy lento ya que compara dos a dos, es decir dadas n sentencias se van a ejecutar $\frac{n(n-1)}{2}$ comparaciones.

Una manera de evitar tanta carga de ejecución de operaciones es ejecutando primero el algoritmo de colisión de claves y luego para cada uno de los grupos (siempre y cuando este tenga más de una sentencia) aplicarle la función del vecino más cercano [87]. Cabe mencionar que la ejecución de estos dos algoritmos fue utilizado ocasionalmente ya que la cantidad de datos no es tan grande como para tomarse más de 5 minutos de ejecución, tiempo de espera que consideramos razonable.

Para determinar las distancias entre sentencias se utilizaron las siguientes funciones:

- **Distancia de Levenshtein**: La distancia entre dos palabras se calcula como el mínimo número de operaciones que transforma a una de las palabras en la otra. Las tres operaciones son eliminación, inserción y eliminación de un carácter. Variamos el parámetro k (radio) utilizando diferentes valores para obtener diferentes resultados dependiendo del largo de las palabras, a modo de ejemplo si la mayoría de las palabras son de entre 3-4 caracteres no tiene sentido utilizar una distancia 3 ya que de ese modo gran parte de las palabras van a ser agrupadas [88].

- **PPM:** Algoritmo implementado por OpenRefine que se basa en una publicación realizada por la Universidad de Waterloo para estimar la similaridad entre distintas secuencias de caracteres utilizando la complejidad de Kolmogorov [89].

  El algoritmo estudia el contenido de cada sentencia y fundamenta que si dos cadenas de caracteres A y B son idénticas, si se comprime A + B (concatenación de las dos sentencias) debe poseer una muy pequeña diferencia con respecto a A y B, siendo la función de compresión determinada por la complejidad de Kolmogorov [90]. Siguiendo esta idea, la herramienta calcula la distancia como:

  $$d(A,B) = comp(A + B) + comp(B + A) \,/\, comp(A + A) + comp(B + B)$$

  Donde $comp$ es la función de compresión determinada por la complejidad de Kolmogorov.

Una vez visualizados estos datos en la plataforma de manera agrupada, comprobando manualmente que efectivamente les correspondía la misma localidad, se les asignó a cada grupo un único nombre para la localidad, escrito de manera correcta.

Además, algunas piezas de datos fueron estandarizadas y normalizadas desde el punto de vista relacional [91] también para proporcionar una mayor consistencia a los datos. Por ejemplo, el tipo de un evento (de una lista predefinida de tipos de inundaciones) era un campo más de cada fila en el formato inicial donde los tipos se encontraban en formato texto, separados por coma uno de otros. Así, se procedió a crear una tabla extra (en OpenRefine una “tabla” se representa como una nueva hoja del documento Excel) donde se definieron los distintos tipos de evento y se les asignó un identificador numérico único e

incremental. Luego, se creó una tabla intermedia que establece vinculaciones entre los eventos y sus tipos, dado que éstos pueden ser de más de un tipo a la vez.

De manera similar a la descrita para el dato de tipo de evento, se efectuó de manera análoga un trabajo de normalización de Localidades, Autores y Fuentes generando así una nueva tabla para cada una de las entidades asegurando unicidad de las mismas y referencias uniformes.

También, con respecto a la dimensión de la calidad de datos que trata sobre la exactitud de los mismos, se encontraron localidades cuya información espacial en coordenadas no coincidía con el nombre que se les había asignado o simplemente no poseían nombre asignado pero sí coordenadas geográficas. Fue así que se desarrolló un script que recorría todas estas coordenadas y comprobaba que éstos coincidieran, mediante consumo de la API de geocodificación reversa (resolución de coordenadas a una dirección) de Google [92], y en caso de no ser así, se asignó a cada par de coordenadas el nombre de la localidad que arrojaba dicha API.

A continuación se resume, en una tabla descriptiva (tabla 7), alguno de los valores de los datos del SINAE antes y después del procesamiento de los mismos para ilustrar sus resultados.

	Antes del procesamiento	**Después del procesamiento**
Cantidad de Tablas	1	11
Total de Celdas	3675	No aplica.
Tota de Columnas	42	No aplica.
Total Celdas Vacías	154350	0 (no se toma en cuenta campos en donde el valor puede ser nulo como es el caso del campo “url” en la tabla fuentes que puede ser nulo ya que la fuente puede ser en formato papel.
Total de localidades distintas	552 localidades detectadas, siendo 114 localidades identificadas como “Uruguay”. Dado que Uruguay no es una localidad específica tuvimos que identificar dichas localidades mediante las coordenadas geográficas y/o por la descripción del evento como describimos anteriormente.	366 localidades distintas, con su correspondiente, departamento, ciudad y coordenadas.

Columna Fuentes	Todas las celdas distintas, es decir, 3675 fuentes diferentes.	6 fuentes distintas.
Cantidad de localidades por evento	1 o ninguna.	Más de una.
Cantidad de eventos	3674	1604

Tabla 7. Comparativa procesamiento calidad de datos.

PostgresSQL

Entre los DBMS, se seleccionó a PostgreSQL para conformar la base de datos centralizada de nuestro sistema.

PostgreSQL es un sistema de código abierto de gestión para base de datos relacionales. El mismo utiliza un sistema de cliente/servidor, siendo mediante el servidor que se realiza la comunicación con la base de datos, este funciona bajo un sistema de multiprocesos [93].

Su funcionamiento es similar al de cualquier sistema de este tipo.

El mismo utiliza el lenguaje de consultas estructuradas SQL, que permite un sencillo manejo (alta, baja y modificación) de la información, mediante consultas , que permite agrupar y seleccionar de manera precisa las piezas de información que se desean manejar, y aplicarle los cambios necesarios. Además, mediante SQL se realizan consultas al DBMS que permiten recuperar y consultar la información deseada en forma de tablas, algo que se alinea de manera excepcional con los requerimientos del presente proyecto [93].

Finalmente, destacar que se analizó la compatibilidad entre PostgreSQL y el framework Django a ser utilizado como cliente final de la aplicación, y resultó que la compatibilidad es completa ya que Django soporta a PostgreSQL como uno de sus manejadores de información por defecto, algo que también resultó clave en su elección.

Entidades

Como parte central del DW propuesto en esta sección, se detallarán cada una de las tablas presentadas en el diagrama de la figura 15 que formarán parte de la base de datos final, omitiendo los identificadores de las clases, así como también las relaciones entre ellas.

Diagrama de la base de datos

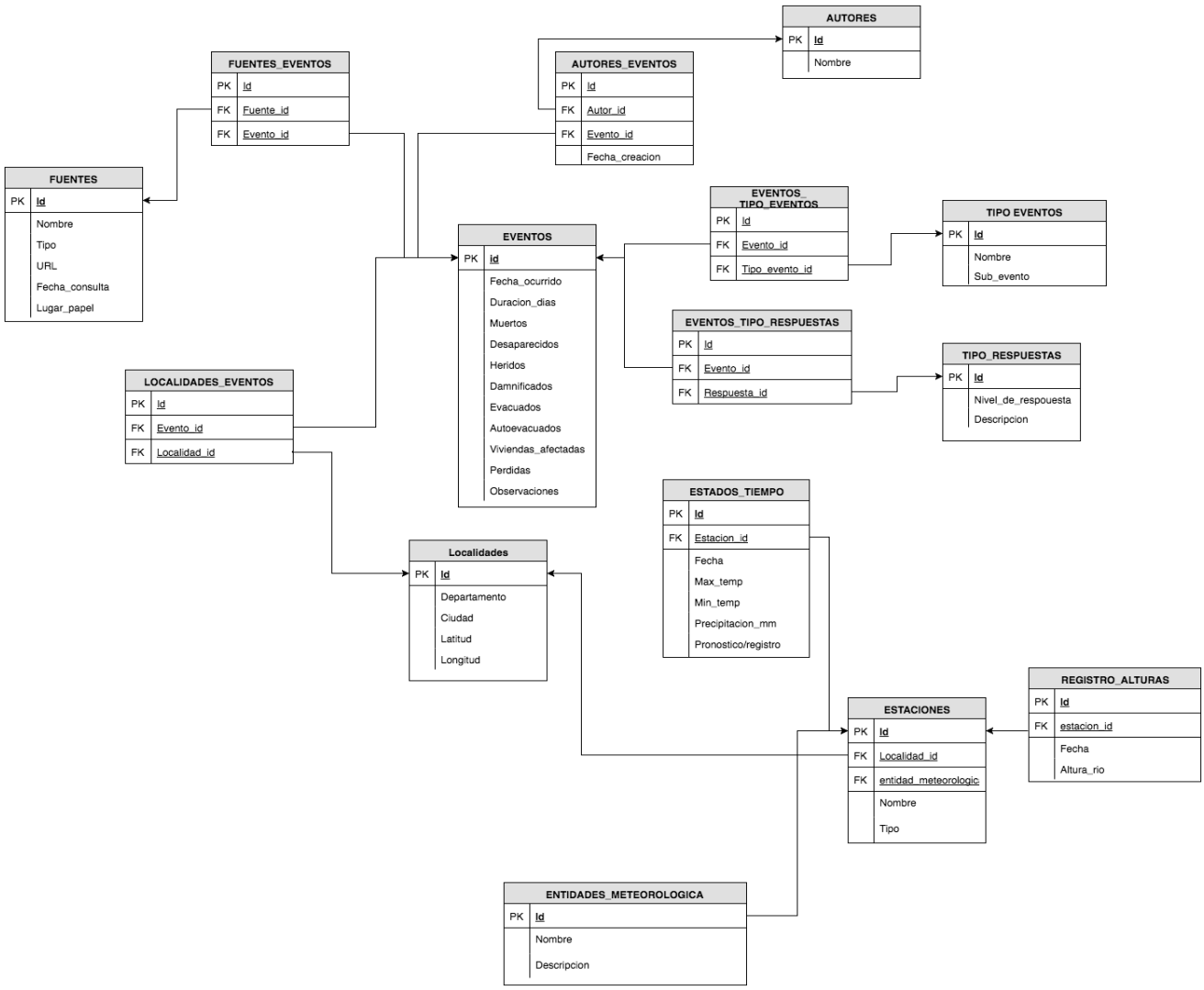


Figura 15. Diagrama de la base de datos

Eventos

Provee la información relacionada con el evento.

Campos:

- Fecha_ocurrido: Fecha en la que se inició el fenómeno
- Duracion_dias: Duración del evento en días.
- Muertos: Número de personas fallecidas.
- Desaparecidos: Número de personas desaparecidas.
- Heridos: Número de personas heridas.
- Damnificados: Número de personas damnificadas.

- Evacuados: Número de personas evacuadas.
- Autoevacuados: Número de personas auto evacuadas.
- Viviendas_afectadas: Número de viviendas afectadas.
- Perdidas: Descripción de pérdidas materiales.
- Observaciones: Observaciones y detalles del evento.

Autores

Representa los autores que fueron encargados de registrar el evento.

Campos:

- Nombre: Nombre del autor.

Autores_Eventos

Tabla intermedia que relaciona los eventos con los autores. Un evento puede ser editado/creado por varios autores, y a su vez un autor puede crear/editar varios eventos.

Fuentes

Fuente de información del evento.

Campos:

- Nombre: Nombre de la fuente de Información.
- Tipo de fuente: Se presentan dos tipos de fuentes de información, fuente del tipo web, o fuente del tipo papel.
- URL: En el caso de que la fuente sea del tipo WEB, ésta debe con una URL asociada.
- Fecha de consulta: En el caso de que la fuente sea del tipo WEB, se cuenta con la fecha de consulta de la URL.
- Lugar_papel: En el caso de que la fuente sea del tipo Papel Representa la ubicación del evento dentro de una fuente del tipo papel.

Fuentes_Eventos

Tabla intermedia que vincula fuentes con eventos. Un evento puede tener varias fuentes y a su vez una fuente puede estar vinculado con varios eventos

Tipo_Eventos

Tipo del evento detectado.

Campos:

- Nombre: Nombre del evento.
- SubEvento: Subevento asociado.

Eventos_Tipo_Eventos

Relación N a N entre los eventos y los tipos de eventos.

Tipo_respuestas

Tipo de respuesta ante el evento.

Campos:

- Nivel de respuesta: Nivel de respuesta asociado al evento. Se detectaron 4 tipos de respuesta: Atención Primaria, Estado de Desastre, Respuesta Departamental y Respuesta Nacional.
- Descripción : Descripción del tipo de respuesta.

Eventos_Tipo_Respuestas

Relación N a N entre los eventos y los tipos de respuestas.

Localidades

Representa localidades en distintos puntos del Uruguay.

Campos:

- Departamento: Departamento de la localidad.
- Ciudad: Nombre de la localidad.
- Latitud: Latitud geográfica de la localidad.
- Longitud: Longitud geográfica de la localidad.

Localidades_Eventos

Relación N a N entre los eventos y las localidades. Un evento es registrado en distintas localidades, así como también, en una localidad se pueden registrar más de un evento.

Estaciones

Estaciones meteorológicas en el Uruguay. Cada estación meteorológica pertenece a una entidad meteorológica y a una localidad.

Campos:

- Nombre: Nombre de la estación.
- Tipo: Tipo de la estación (convencional o automática).

Estados_Tiempo (estado del tiempo para una fecha dada en términos de temperatura y precipitaciones)

Estado del tiempo registrado.

Campos:

- Fecha: Fecha del estado del tiempo registrado.
- Max_temp: Máxima temperatura en grados Celsius (**°C**).
- Min_temp: Minima temperatura en grados Celsius (**°C**).
- Precipitacion_mm: Precipitación registrada en milímetros.
- Pronóstico/Registro: Indicador de tipo de estado del tiempo, de tipo registro (estado del tiempo real) o pronóstico.

Entidad_Meteorológica

Entidad meteorológica detectada.

Campos:

- Nombre: Nombre de la entidad.
- Descripción: Descripción de la entidad.

Registro_Alturas

Registro del nivel de caudal de los ríos.

Campos:

- Fecha: Fecha en la cual se registró la altura.
- Altura Río: Crecida del río en metros con respecto al nivel del mar.

Una vez creada la base de datos en el manejador Postgres, estando ésta vacía, se necesitó migrar datos desde las distintas fuentes de información explicadas en los dos puntos anteriores del presente capítulo, no sin antes aplicarles los procedimientos necesarios para cumplir con los requerimientos mencionados en cuanto a la calidad de los datos.

Fue así que en una primera instancia se procedió por la limpieza de cada una de las bases de datos (en formato Excel, mencionadas en los puntos anteriores del presente capítulo) descritas anteriormente. Una vez que se finalizó con la limpieza individual se continúo por la creación de una base de datos central *proyecto_de_grado* y la limpieza de ésta, de manera de generar una base de datos centralizada con distintos datos provenientes de las distintas fuentes consultadas.

Dicha limpieza consistió entre otras de las siguientes funciones :

- Eliminación de datos duplicados.
- Normalización de los datos.
- Identificar y eliminar inconsistencias, discrepancias y errores en datos, para mejorar la calidad.

Una vez finalizado la etapa de limpieza de datos, se continuó con la centralización de los datos migrando los mismos hacia la base de datos centralizada en Postgres.

Pasaje de formato Excel al DBMS

El formato de hojas de cálculo Excel es aquel en el que se recibió la totalidad de los datos estáticos solicitados. Esto se debe a que los sistemas ofimáticos que manejan este formato gozaron de amplia popularidad en los últimos años y se utilizaron como base de información en entes estatales y de toda índole.

Sin embargo, este formato no es el deseado en el presente proyecto. Se decidió utilizar sistemas de manejo de base de datos (DBMS) para esto por diversas razones.

Una de las principales razones para realizar una migración de formato es que al día de la fecha son éstos los sistemas más usados para el manejo de información en cualquier tipo de sistemas informáticos permitiendo el crecimiento y la sostenibilidad del sistema.

Para migrar la información en formato Excel hacia un DBMS PostgreSQL, se llegó a la conclusión de que la manera más efectiva era utilizando otro formato (uno mucho más sencillo y básico) como paso intermedio entre uno y otro. Esta necesidad se volvió evidente al ver que no hay formas directas de migrar de Excel a PostgreSQL, a pesar de que se tenía representada la información en Excel en manera similar en la que se tendría en el DBMS, esto es, en forma de tablas separadas.

Se eligió el formato CSV. CSV, o Comma Separated Values, es como dice su nombre, un formato mediante el cual a cada fila de una tabla de información, se la representa como simplemente los valores de cada una de sus columnas, de manera explícita, separados simplemente mediante una coma. En este formato, cada fila se representa con una línea distinta del archivo. Los gestores de formato Excel suelen permitir una exportación directa de su contenido a formato CSV y fue de esta manera que se logró llenar archivos CSV con cada una de las tablas que se tenían en Excel.

En el otro extremo de la transformación, PostgreSQL facilitó mediante psql, su cliente de líneas de comando, la migración desde los archivos CSV intermedios a sus tablas relacionales mediante el comando \copy [94], que permite el copiado de información desde un archivo a una tabla relacional del DBMS. Entre los formatos de entrada que soporta se encuentra CSV y no Excel.

Es así que, partiendo de una estructura de base de datos (tablas) definida por la aplicación Python en la creación de sus modelos, y separadamente, los archivos en formato CSV generados a partir de los excel, se ejecutó un script de bash linux compuesto de una secuencia de comandos copy que copian, uno a uno, los archivos a las tablas correspondientes.

De esta manera, como resultado del proceso de conversión, se finaliza con una base de datos SQL en el manejador PostgreSQL, precargada con los datos estáticos que se obtuvieron en etapas más tempranas del proyecto, como se menciona en la sección 5.1.1.

5.4 Motor Predictivo

El objetivo de esta etapa del proyecto, alineándose con los descritos en la sección 3.2, es el de dar un valor agregado a los datos obtenidos en la parte previa. La investigación, obtención, integración y limpieza de los datos constituyó la mayor parte del tiempo y esfuerzo del presente trabajo. En esta etapa final se busca introducir y aplicar técnicas predictivas sencillas, demostrando que los datos obtenidos poseen utilidad a la hora de gestionar el riesgo de un desastre de inundación. Dicho de otra manera, se buscará predecir la presencia o no de un evento de inundación utilizando como base dos elementos principales: la información referente a lo sucedido en el pasado, y la información ambiental obtenida para el día de hoy.

5.4.1 Análisis y Diseño del Motor Predictivo

Una vez que se plantearon ciertos objetivos y ya con el conocimiento de qué opciones de tecnologías habían para lograr implementar el motor predictivo deseado, se asistió a una conferencia sobre Machine Learning brindada por la empresa Tryolabs [107] en la Facultad de Ingeniería [95]. En esta conferencia se brindaron los conceptos básicos de *machine learning,* dándose algunos ejemplos de ayuda para entender cómo funciona la tecnología y por otra parte mostrando el avance de la misma, sus usos y su posicionamiento tanto en investigación como comercial, en nuestro país.

Posteriormente y con el fin de tener un entendimiento más concreto del funcionamiento y uso de las herramientas en el contexto del presente problema, y para poder estudiar su viabilidad y posibilidades en conjunto con gente que tuviese experiencia, se pudo concretar una reunión con personal de la empresa encargada de brindar la conferencia. En dicha reunión se describió la idea en mente, lo ya trabajado y el objetivo que se tenía planteado para así obtener comentarios, opiniones y/o sugerencias que sirvieran de ayuda para la implementación del motor. Los resultados de esta reunión fueron muy positivos ya que ayudaron a plantear un objetivo posible (replanteando el que se tenía hasta el momento) y a comprender mejor cómo vincular los datos disponibles.

Como ya se mencionó en el resumen del estado del arte del presente informe, en su sección 4.6, se intenta convertir los datos obtenidos en conocimiento. De esta manera, se decidió que el conocimiento que se quiere alcanzar como mínimo es el de saber si habrá o no un evento de inundación dados:

1. Un contexto geográfico y temporal: lugar y una fecha
2. Condiciones ambientales dentro del contexto mencionado

5.4.2 Implementación del Motor Predictivo

Es así que después de la investigación llevada a cabo, con sus resultados plasmados en forma de resumen en el capítulo 4 de este informe y en forma completa adjunto como Anexo 1, nos decidimos en favor de técnicas de aprendizaje automático, y en particular a utilizar el lenguaje de programación Python como medio, su librería Sci-kit learn dedicada a

técnicas de aprendizaje automático y también su framework asociado Django (ver sección 5.3.3) el cual permite una sencilla escritura del código en programación orientada a objetos y una rápida presentación de una aplicación web.

En el marco de la programación orientada a objetos y las entidades que nuestra realidad presenta, se implementaron las clases descritas en el diagrama de la sección 5.3.2 y su subsecuente conexión a la base de datos central alojada en Postgres, descrita en la sección 5.3.

El mapeo objeto-relacional que Django provee conecta directamente las clases previamente mencionadas con las tablas de la base de datos central.

Una vez completado lo anterior, se obtuvo, mediante el framework y su mapeo objeto-relación, acceso sencillo a la información en forma de instancias de clases, así como también la habilidad de consultar datos de éstos modelos según se requiera.

A su vez, mediante la instalación e importación de la librería Sci-kit learn, se logró fácil acceso también a distintos algoritmos de aprendizaje automático a utilizarse con la información proveniente de los modelos.

Por lo tanto se desarrolló un script con el fin de seleccionar aquellas técnicas, dentro de las que presenta la librería y aquéllas investigadas (sección 4.6.3) que mejor sirvieran al propósito descrito.

5.4.3 Pruebas Realizadas para Validar Motor Predictivo

Para la selección de aquellos algoritmos de aprendizaje automático que mejor se ajusten a la realidad propuesta y a los datos que se tienen, se enfrentó a un desafío de validación de los métodos probados. Esto es, probar de alguna manera si los métodos arrojan un resultado al menos dentro de un error esperado, y dentro de los resultados observados, seleccionar los mejores para su uso definitivo. Además, cabe recordar aquí que la mayoría de éstos algoritmos de aprendizaje automática poseen parámetros ajustables y cuyo valor óptimo dependerá de la naturaleza de los datos que se tengan para cada problema, por lo tanto, dentro de cada método, si corresponde, también se busca encontrar aquellos valores de parámetros que den los resultados más ajustados a la realidad posible. Por supuesto, estos resultados se podrían mejorar con un estudio más profundo de los parámetros y su sensibilidad, pero dentro del alcance del proyecto, se busca la mejor aproximación posible.

Iteraciones de validación

Luego de estudiadas las posibilidades para la validación de los modelos (sección 4.6.4), se decidió utilizar K-Fold a los efectos de validar y encontrar el método que mejores resultados arroje. Ésta decisión fue basada en la popularidad del método, el hecho de que sea el método por defecto a la hora de validar este tipo de problemas, y el alcance del proyecto.

Cabe mencionar también que fue encontrada en la documentación de la librería Sci-kit learn una variante de estos métodos ya implementada llamada StratifiedKFold [96]. Ésta

manera de realizar una separación de la muestra en validación cruzada, selecciona los distintos subconjuntos de prueba y de entrenamiento de manera tal de que éstos mantengan una proporción lo más pareja posible de muestras de cada clase de resultados (positivos y negativos).

A partir de la decisión de utilizar K-Fold como método de validación cruzada, resulta necesario encontrar valores óptimos para la cantidad de subconjuntos en los que se separarían las muestras de datos, esto es, el valor K. Es así que en la investigación sobre cómo encontrarlo, se dió con el trabajo [97] de Ron Kohavi de la Universidad de Stanford, quien elaboró una publicación científica dedicada justamente a definir esto para diferentes problemas. En su trabajo, prueba varias métricas para medir la precisión de un método, y en particular aquellas mencionadas en este trabajo. Consideró también para sus pruebas la técnica de muestreo Bootstrapping [98], que no fue considerada para el presente trabajo dada su complejidad y el tiempo disponible para su desarrollo. De todas maneras, dicha técnica finalmente no arroja los mejores resultados. Luego de tener en cuenta no sólo la *accuracy* y cantidad de veces que cada método acertaba, sino también la relación *precision/recall* que tomaba cada método para uno u otro resultado y la desviación que posee cada uno, termina concluyendo lo siguiente:

"Nuestros resultados indican que la estratificación es en general un mejor acercamiento tanto en términos de precisión/recall como de desviación comparado a un cross-validation normal. Bootstrap posee baja desviación pero puede llegar a valores de precisión/recall demasiado malos en algunos problemas. Para seleccionar un modelo, recomendamos usar validación 10-fold estratificado" [97]

Dada la gran aceptación general del enunciado salvo casos particulares, se decidió alinear el procedimiento de este trabajo con su tesis. Esto es, usar 10-fold cross-validation estratificado para nuestras pruebas de validación de los diferentes métodos.

La implementación éstas pruebas de 10-Fold se vio simplificada por la librería Sci-kit learn. Ésta provee un método llamado `cross_val_score` [99] que toma como parámetros un algoritmo de aprendizaje ya instanciado con sus respectivos hiperparámetros, el conjunto de datos y un número K de iteraciones a realizarse. Luego, una vez se llama a dicho método, éste corre las K iteraciones, encargándose de la división en subconjuntos, la computación del error para cada una, y finalmente arrojar un valor de *accuracy* promedio referente a todo el conjunto y todas las iteraciones como ya se mencionó. La idea es correr estas iteraciones instanciando los distintos algoritmos de aprendizaje y en cada de uno de ellos, probar distintos hiperparámetros de manera de seleccionar aquellos que arrojaron los mejores resultados equilibrando *accuracy* y *recall*.

Para las pruebas se utilizó la división política del país en departamentos y se generaron conjuntos de datos totales, con granularidad diaria, que contengan datos en cantidades balanceadas (estratificados) tanto de días donde sucedieron eventos, como de días en los que no sucedió nada, tomados estos de manera aleatoria. Cabe destacar que si para un departamento en cuestión la cantidad de elementos de datos con resultado positivo (se lo clasifica como 1), esto es, días en los que haya sucedido un evento, es muy reducida

entonces el conjunto total de datos resultará muy reducido y puede suceder que: el departamento sea descartado o las divisiones del conjunto (el valor K, en este caso 10) sean pocas para no terminar con subconjuntos de datos muy pequeños y los modelos se entrenen pobremente. Es también esta la razón principal por la cual se utilizó una división por departamentos y no más específica, ya que la cantidad de datos de eventos existente para una localidad en particular es generalmente muy baja (menor a las 10 muestras) salvo un par de casos particulares.

Como métricas de puntaje, luego de estudiadas las posibilidades (ver sección 4.6.4), se utilizó el promedio de los aciertos para la clasificación binaria ya que resulta trivial e inmediata dada la naturaleza de la clasificación binaria. El resultado es correcto (1) o incorrecto (0), por lo tanto en una corrida se harán 10 iteraciones y se promediaran sobre todos los resultados, resultando en un número entre 0 y 1 donde lo más cercano a 1 posible será lo mejor.

Para el caso de regresión, se decidió utilizar la métrica de error medio absoluto, como ya se mencionó la misma es la distancia entre el valor real y el valor predecido y por ende el resultado es mejor cuanto más cercano a 0 sea el mismo. Al igual que en clasificación, se realizan 10 corridas y se promedian los resultados.

Otra parte fundamental del trabajo con aprendizaje automático es lo que en inglés se llama ***selección de variables (feature engineering)*** [100] y en español no parece tener acuñado ningún término aún. Este término hace referencia al trabajo que se debe realizar, utilizando conocimiento del dominio donde será aplicado aprendizaje automático, para extraer de los datos que se tienen, variables de entrada para que, luego de entrenado con éstas, la predicción mediante el algoritmo arroje los resultados más cercanos a la realidad posible. Es deseable entonces que estas variables de alguna manera tengan vinculación con lo que se desea predecir, o dicho de otra manera, que los valores que tomen realmente afecten o asistan de manera adecuada la predicción, mejorandola y llevándola hacia mayor exactitud. Éste es un proceso semi-artesanal donde intervienen el conocimiento del negocio, sentido común y la creatividad de quienes lo lleven a cabo, y por lo tanto no es una ciencia exacta.

Es por todo esto que es considerado por algunos autores como una temática “informal”, e incluso un arte, pero sin lugar a dudas todos los autores coinciden en que es uno de los factores más determinantes para el buen funcionamiento de técnicas de aprendizaje automático a la hora de resolver un problema dado. Se trata de un problema de representación, de transformar los datos que se tienen en una estructura que mejor represente la situación de la realidad sobre la que se trabajó y que permita que el algoritmo de aprendizaje pueda entender y al que le pueda sacar mayor provecho en pos de resultados que se ajusten a la realidad. Descomponer los datos y componer estructuras que describan de la mejor manera posible el problema ante estos algoritmos. Es así que una variable de entrada debe ser un atributo de utilidad para describir la realidad [100].

Así es que en base a lo investigado, conocimiento obtenido sobre los eventos desastrosos de inundación en el territorio nacional (sección 4.2) y los datos que se obtuvieron

(secciones 5.1 y 5.2), definimos las siguientes variables de entrada de las cuales seleccionaremos aquellas que sean más tenidas en cuenta por los algoritmos.

Las variables de entrada definidas para ser probadas son:

PH: Precipitaciones del día corriente
PS: Precipitaciones de la semana previa incluyendo al día corriente, promediados
P-1: Precipitaciones del día anterior
P-2: Precipitaciones de del día antes al anterior
Estación: Estación del año discretizada en 4 variables booleanas
Mes: Mes del año discretizado en 12 variables booleanas
Alturas de Río: Alturas de todos los ríos, una variable de entrada por cada una de las estaciones de medición

La metodología tomada en este trabajo para llevar a cabo la tarea de selección de variables de entrada, dentro de aquellas que fueron definidas, fue la de almacenar aquellas combinaciones de variables de entrada que se hayan destacado en su rendimiento, para luego contabilizar sus apariciones y determinar cuales fueron utilizadas una mayor cantidad de veces, y por ende, más relevantes para la ejecución del algoritmo.

Por lo tanto, se escribieron en el script encargado de correr las validaciones cruzadas, iteraciones adicionales anidadas para cada una de las anteriores iteraciones recién mencionadas, variando la combinación de las variables de entrada a considerar, de manera tal de determinar cuales son las que mayor aporte generan a la hora de predecir. Las variables de entradas a considerar se toman desde el data warehouse (sección 5.3) del proyecto y es responsabilidad del script que corre las iteraciones el traerlas desde allí, procesarlas y alimentar al modelo con las mismas.

5.4.4 Resultados de la Selección de Variables

A continuación se presentan los resultados obtenidos para cada uno de los métodos probados en formato de tabla y a partir de ellas y su visión comparativa para cada departamento, al final del capítulo se dan conclusiones con respecto a él o los métodos a utilizarse finalmente.

Para cada departamento, junto al nombre del departamento también se especifica la cantidad de muestras que el mismo posee, esto es, el doble de la cantidad de eventos registrados para ese departamento.

La tabla para cada departamento está constituida de tres partes: hiperparámetros, variables de entrada y puntajes. A su vez, cada fila de la tabla corresponde a un método y contiene los datos asociados a la ejecución de ese método que tuvo más éxito de todas las realizadas en validación cruzada, como ya se dijo, buscando un equilibrio entre la métrica de *accuracy* y *recall*. Se agregó a la tabla la métrica de *precisión* de modo simplemente ilustrativo.

Los hiperparámetros son aquellos valores encontrados como óptimos, esto es, que dieron los mejores resultados en las pruebas de validación cruzada realizadas, como se explicó en la sección anterior. En la sección de variables de entrada de cada tabla, se especifica de manera booleana (Sí o No) cuáles de éstas fueron utilizadas en la corrida de mayor éxito que se está registrando. Finalmente en la sección de Puntajes se tienen las tres métricas ya mencionadas en el punto anterior, *accuracy*, *precision* y *recall*.

Hiperparámetros:

C, Alfa, Dual, L1 ratio (ver Anexo 1, sección 7.3 para más detalles sobre éstos): valores óptimos hiperparámetros correspondientes a cada uno de los métodos como visto en la sección 4.6.3. Si no aplica, se indica con un guión (“-”).

Variables de entrada:

PH: Precipitaciones del día corriente
PS: Precipitaciones de la semana previa incluyendo al día corriente, promediados
P-1: Precipitaciones del día anterior
P-2: Precipitaciones de del día antes al anterior
Estación: Estación del año discretizada en 4 variables booleanas
Mes: Mes del año discretizado en 12 variables booleanas
Alturas de Río: Alturas de todos los ríos

Tipos de algoritmos:

Métodos de Regresión
Mejor método de Regresión
Métodos de Clasificación
Mejor método de Clasificación

Artigas (tamaño de muestra: 111)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	Si	No	Si	Si	0.41	-	-
Lasso	-	0	-	-	Si	Si	No	Si	No	Si	Si	0.42	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**No**	**0.32**	**-**	**-**
Ridge	-	0.1	-	-	Si	Si	No	Si	Si	No	No	0.41	**-**	**-**
ElasticNet	-	0.04	-	0.1	Si	No	No	Si	Si	No	No	0.41	-	-
Linear SVC	1	-	No	-	Si	No	Si	Si	Si	Si	No	0.77	0.77	0.74
RBF SVC	**1**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**0.79**	**0.75**	**0.85**
Sigmoid SVC	1	-	-	-	Si	No	No	No	Si	No	No	0.72	0.7	0.76
K Neighbors	-	-	-	-	Si	No	No	No	Si	No	No	0.66	0.63	0.83

Tabla 8. Mejores resultados algoritmos en Artigas.

Canelones (tamaño de muestra: 283)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	Si	No	Si	Si	0.34	-	-
Lasso	-	0	-	-	Si	Si	Si	Si	Si	Si	Si	0.34	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**Si**	**No**	**Si**	**0.26**	**-**	**-**
Ridge	-	10	-	-	Si	Si	No	Si	Si	Si	Si	0.34	-	-
ElasticNet	-	0.05	-	0.1	Si	Si	No	Si	No	Si	Si	0.34	-	-
Linear SVC	1	-	No	-	Si	Si	No	No	Si	No	Si	0.80	0.85	0.78
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**No**	**No**	**Si**	**0.82**	**0.76**	**0.92**
Sigmoid SVC	10	-	-	-	Si	No	Si	No	No	No	No	0.79	0.73	0.86
K Neighbors	-	-	-	-	Si	Si	Si	No	Si	Si	No	0.81	0.80	0.80

Tabla 9. Mejores resultados algoritmos en Canelones.

Cerro Largo (tamaño de muestra: 167)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	Si	No	No	Si	0.42	-	-
Lasso	-	0	-	-	Si	Si	Si	Si	Si	No	Si	0.44	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**Si**	**No**	**No**	**No**	**Si**	**0.34**	**-**	**-**
Ridge	-	0.1	-	-	Si	No	No	Si	No	Si	Si	0.42	-	-
ElasticNet	-	0.28	-	0.1	Si	No	Si	Si	No	Si	Si	0.44	-	-
Linear SVC	10	-	Si	-	Si	Si	Si	No	No	Si	No	0.67	0.65	0.65
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**Si**	**No**	**Si**	**Si**	**No**	**No**	**0.70**	**0.65**	**0.77**
Sigmoid SVC	10	-	-	-	Si	No	Si	Si	No	No	Si	0.59	0.59	0.56
K Neighbors	-	-	-	-	Si	No	Si	No	No	Si	No	0.71	0.64	0.68

Tabla 10. Mejores resultados algoritmos en Cerro Largo.

Colonia (tamaño de muestra: 163)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	Si	Si	Si	Si	0.39	-	-
Lasso	-		-	-	Si	Si	Si	Si	Si	Si	Si	0.39	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**Si**	**Si**	**No**	**0.31**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	Si	Si	Si	Si	0.37	-	-
ElasticNet	-	0.01	-	1.0	Si	Si	Si	Si	Si	Si	Si	0.38	-	-
Linear SVC	1	-	No	-	Si	Si	No	No	No	Si	Si	0.75	0.77	0.74
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**No**	**Si**	**No**	**Si**	**Si**	**Si**	**0.80**	**0.76**	**0.81**
Sigmoid SVC	10	-	-	-	Si	No	Si	No	No	No	No	0.70	0.70	0.69
K Neighbors	-	-	-	-	Si	No	No	Si	Si	No	Si	0.78	0.79	0.80

Tabla 11. Mejores resultados algoritmos en Colonia.

Durazno (tamaño de muestra: 139)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	No	No	Si	Si	0.41	-	-
Lasso	-	0	-	-	Si	No	No	Si	Si	Si	No	0.42	-	-
SVR	**-**	**-**	**-**	**-**	**No**	**Si**	**No**	**No**	**No**	**Si**	**No**	**0.33**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	No	No	No	Si	0.40	-	-
ElasticNet	-	0.004	-	1	Si	Si	Si	No	No	No	Si	0.42	-	-
Linear SVC	1	-	No	-	Si	No	Si	Si	No	Si	No	0.73	0.73	0.73
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**Si**	**Si**	**No**	**0.80**	**0.78**	**0.82**
Sigmoid SVC	1	-	-	-	Si	No	No	No	Si	No	No	0.69	0.72	0.68
K Neighbors	-	-	-	-	Si	No	No	Si	No	Si	No	0.79	0.79	0.72

Tabla 12. Mejores resultados algoritmos en Durazno.

Flores (tamaño de muestra: 37)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	No	Si	No	Si	Si	No	0.30	-	-
Lasso	-	0	-	-	Si	No	No	Si	No	No	Si	0.32	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**Si**	**Si**	**No**	**0.25**	**-**	**-**
Ridge	-	1	-	-	Si	Si	Si	Si	No	Si	No	0.33	-	-
ElasticNet	-	0.01	-	1.0	Si	Si	Si	Si	No	Si	No	0.32	-	-
Linear SVC	1	-	No	-	Si	Si	Si	Si	No	Si	Si	0.79	0.85	0.94
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**No**	**0.92**	**0.85**	**0.94**
Sigmoid SVC	1	-	-	-	Si	No	No	Si	No	Si	No	0.84	0.76	0.88
K Neighbors	-	-	-	-	Si	Si	Si	Si	Si	no	Si	0.90	0.88	0.83

Tabla 13. Mejores resultados algoritmos en Flores.

Florida (tamaño de muestra: 159)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	No	Si	No	Si	No	Si	0.38	-	-
Lasso	-	0	-	-	Si	No	Si	No	Si	No	Si	0.37	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**Si**	**Si**	**No**	**0.34**	**-**	**-**
Ridge	-	1	-	-	Si	Si	Si	No	No	No	Si	0.38	-	-
ElasticNet	-	0.01	-	1.0	Si	Si	No	No	Si	Si	Si	0.38	-	-
Linear SVC	1	-	No	-	Si	No	Si	No	No	Si	Si	0.71	0.71	0.74
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**No**	**Si**	**Si**	**No**	**Si**	**Si**	**0.79**	**0.72**	**0.86**
Sigmoid SVC	1	-	-	-	Si	No	Si	No	No	No	No	0.71	0.67	0.83
K Neighbors	-	-	-	-	Si	No	No	No	No	Si	Si	0.75	0.72	0.84

Tabla 14. Mejores resultados algoritmos en Florida.

Lavalleja (tamaño de muestra: 59)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	No	No	No	No	No	No	0.45	-	-
Lasso	-	0	-	-	Si	No	Si	No	No	No	No	0.45	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**No**	**0.38**	**-**	**-**
Ridge	-	0.1	-	-	Si	Si	Si	No	Si	Si	Si	0.43	-	-
ElasticNet	-	0.11	-	0.1	Si	Si	Si	Si	Si	No	No	0.45	-	-
Linear SVC	50	-	Si	-	Si	Si	Si	No	Si	No	No	0.52	0.52	0.62
RBF SVC	10	-	-	-	Si	Si	No	No	No	No	No	0.69	0.62	0.75
Sigmoid SVC	**1**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.60**	**0.56**	**0.93**
K Neighbors	-	-	-	-	Si	Si	Si	Si	No	Si	No	0.66	0.63	0.65

Tabla 15. Mejores resultados algoritmos en Lavalleja.

Maldonado (tamaño de muestra: 185)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	No	No	Si	No	0.35	-	-
Lasso	-	0.01	-	-	Si	Si	Si	Si	No	Si	No	0.35	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**No**	**No**	**No**	**0.25**	**-**	**-**
Ridge	-	0.1	-	-	Si	No	Si	Si	Si	Si	Si	0.35	-	-
ElasticNet	-	1.69	-	0.1	Si	Si	Si	Si	Si	Si	Si	0.35	-	-
Linear SVC	50	-	Si	-	Si	Si	Si	No	No	No	Si	0.68	0.67	0.80
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**No**	**Si**	**No**	**No**	**No**	**Si**	**0.84**	**0.82**	**0.84**
Sigmoid SVC	1	-	-	-	Si	No	No	Si	No	No	No	0.72	0.66	0.85
K Neighbors	-	-	-	-	Si	No	No	No	Si	No	Si	0.78	0.83	0.72

Tabla 16. Mejores resultados algoritmos en Maldonado.

Montevideo (tamaño de muestra: 235)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	Si	Si	No	Si	0.37	-	-
Lasso	-	0	-	-	Si	No	Si	No	Si	Si	Si	0.38	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**Si**	**No**	**No**	**0.26**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	Si	Si	Si	Si	0.37	-	-
ElasticNet	-	0.01	-	1.0	Si	Si	Si	Si	Si	No	Si	0.37	-	-
Linear SVC	10	-	Si	-	Si	Si	No	No	Si	Si	Si	0.79	0.79	0.82
RBF SVC	**1**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.79**	**0.73**	**0.90**
Sigmoid SVC	1	-	-	-	Si	No	Si	No	No	No	No	0.74	0.68	0.91
K Neighbors	-	-	-	-	Si	No	No	No	Si	No	Si	0.83	0.82	0.87

Tabla 17. Mejores resultados algoritmos en Montevideo.

Paysandú (tamaño de muestra: 139)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	No	No	No	Si	0.43	-	-
Lasso	-	0.01	-	-	Si	No	No	No	Si	No	No	0.42	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**No**	**0.32**	**-**	**-**
Ridge	-	1	-	-	Si	No	Si	Si	No	Si	Si	0.42	-	-
ElasticNet	-	0.01	-	0.90	Si	Si	Si	No	Si	Si	No	0.42	-	-
Linear SVC	10	-	No	-	Si	Si	No	Si	No	Si	No	0.63	0.67	0.73
RBF SVC	1	-	-	-	Si	Si	Si	Si	No	Np	Si	0.68	0.60	0.73
Sigmoid SVC	1	-	-	-	Si	No	No	No	Si	No	No	0.64	0.62	0.81
K Neighbors	-	-	-	-	Si	No	No	No	No	No	Si	0.71	0.73	0.71

Tabla 18. Mejores resultados algoritmos en Paysandú..

Soriano (tamaño de muestra: 191)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	Si	Si	No	No	0.36	-	-
Lasso	-	0	-	-	Si	Si	Si	Si	Si	Si	Si	0.36	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**No**	**Si**	**Si**	**0.33**	**-**	**-**
Ridge	-	0.1	-	-	Si	No	Si	Si	No	Si	Si	0.36	-	-
ElasticNet	-	0.01	-	0.5	Si	Si	Si	Si	Si	Si	Si	035	-	-
Linear SVC	50	-	Si	-	Si	Si	Si	Si	Si	Si	Si	0.61	0.56	0.66
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.70**	**0.67**	**0.8**
Sigmoid SVC	10	-	-	-	Si	No	No	Si	No	No	Si	0.63	0.62	0.69
K Neighbors	-	-	-	-	Si	No	Si	Si	Si	No	Si	0.72	0.68	0.72

Tabla 19. Mejores resultados algoritmos en Soriano.

Tacuarembó (tamaño de muestra: 115)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	No	No	No	No	0.42	-	-
Lasso	-	0	-	-	Si	Si	Si	Si	No	No	No	0.40	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**Si**	**No**	**0.33**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	Si	Si	Si	Si	0.42	-	-
ElasticNet	-		-		Si	No	Si	No	Si	Si	Si	0.42	-	-
Linear SVC	1	-	-	-	Si	No	Si	Si	No	Si	Si	0.72	0.72	0.68
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**No**	**Si**	**Si**	**No**	**No**	**No**	**0.80**	**0.74**	**0.89**
Sigmoid SVC	10	-	-	-	Si	No	No	No	Si	Si	Si	0.74	0.76	0.79
K Neighbors	-	-	-	-	Si	Si	No	No	Si	Si	Si	0.74	0.76	0.72

Tabla 20. Mejores resultados algoritmos en Tacuarembó.

Salto (tamaño de muestra: 205)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	No	Si	Si	Si	No	No	0.37	-	-
Lasso	-	0.07	-	-	Si	Si	Si	Si	Si	Si	No	0.39	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**0.30**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	No	Si	Si	Si	0.37	-	-
ElasticNet	-	0.01	-	0.5	Si	Si	Si	No	Si	No	Si	0.38	-	-
Linear SVC	-	-	-	-	Si	No	No	Si	No	Si	Si	0.69	0.67	0.75
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**Si**	**Si**	**Si**	**0.78**	**0.73**	**0.83**
Sigmoid SVC	-	-	-	-	Si	No	Si	Si	No	No	Si	0.65	0.69	0.74
K Neighbors	-	-	-	-	Si	Si	No	Si	Si	No	No	0.75	0.72	0.81

Tabla 21. Mejores resultados algoritmos en Salto.

Rivera (tamaño de muestra: 161)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	No	No	Si	Si	0.38	-	-
Lasso	-	0.002	-	-	Si	Si	Si	Si	No	Si	No	0.38	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**0.33**	**-**	**-**
Ridge	-	0.1	-	-	Si	Si	Si	Si	Si	Si	Si	0.37	-	-
ElasticNet	-	0.01	-	1	Si	Si	Si	No	Si	Si	No	0.38	-	-
Linear SVC	1	-	-	-	Si	No	Si	No	Si	Si	No	0.71	0.67	0.79
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**No**	**Si**	**No**	**Si**	**0.74**	**0.71**	**0.88**
Sigmoid SVC	1	-	-	-	Si	No	No	No	No	No	No	0.73	0.73	0.73
K Neighbors	-	-	-	-	Si	No	No	No	Si	No	Si	0.75	0.69	0.725

Tabla 22. Mejores resultados algoritmos en Rivera.

Rocha (tamaño de muestra: 169)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	Si	No	No	No	0.35	-	-
Lasso	-	0.01	-	-	Si	Si	Si	No	No	Si	Si	0.35	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**No**	**No**	**No**	**0.30**	**-**	**-**
Ridge	-	10	-	-	Si	Si	Si	Si	Si	Si	Si	0.35	-	-
ElasticNet	-	0.09	-	0.1	Si	Si	Si	Si	Si	Si	Si	0.35	-	-
Linear SVC	10	-	-	-	Si	No	Si	Si	Si	Si	No	0.68	0.72	0.75
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.79**	**0.72**	**0.92**
Sigmoid SVC	10	-	-	-	Si	No	Si	No	No	No	Si	0.68	0.63	0.88
K Neighbors	-	-	-	-	Si	No	Si	No	Si	No	Si	0.77	0.75	0.75

Tabla 23. Mejores resultados algoritmos en Rocha.

Río Negro (tamaño de muestra: 107)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	No	Si	No	No	0.36	-	-
Lasso	-	0.01	-	-	Si	No	No	No	Si	Si	Si	0.39	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**Si**	**0.35**	**-**	**-**
Ridge	-	10	-	-	Si	Si	No	Si	Si	Si	Si	0.38	-	-
ElasticNet	-	0.07	-	0.1	Si	Si	No	No	No	Si	No	0.39	-	-
Linear SVC	1	-	-	-	Si	No	No	Si	Si	Si	No	0.75	0.72	0.79
RBF SVC	**1**	**-**	**-**	**-**	**Si**	**Si**	**No**	**Si**	**No**	**No**	**Si**	**0.72**	**0.66**	**0.89**
Sigmoid SVC	1	-	-	-	Si	No	No	No	No	No	No	0.67	0.66	0.68
K Neighbors	-	-	-	-	Si	No	No	No	No	No	Si	0.74	0.75	0.74

Tabla 24. Mejores resultados algoritmos en Río Negro.

Treinta y Tres (tamaño de muestra: 113)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	No	Si	No	Si	Si	0.44	-	-
Lasso	-	0.01	-	-	Si	No	No	Si	No	Si	Si	0.44	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**No**	**Si**	**No**	**No**	**Si**	**0.38**	**-**	**-**
Ridge	-	1.0	-	-	Si	Si	Si	Si	Si	Si	Si	0.42	-	-
ElasticNet	-	0.06	-	0.1	Si	No	No	Si	No	No	No	0.43	-	-
Linear SVC	50	-	-	-	Si	Si	No	Si	No	Si	No	0.66	0.61	0.66
RBF SVC	**50**	**-**	**-**	**-**	**Si**	**Si**	**No**	**No**	**Si**	**No**	**Si**	**0.70**	**0.67**	**0.73**
Sigmoid SVC	10	-	-	-	Si	No	No	No	No	No	No	0.68	0.69	0.64
K Neighbors	-	-	-	-	Si	No	No	No	Si	No	No	0.47	0.47	0.79

Tabla 25. Mejores resultados algoritmos en Treinta y Tres.

San José (tamaño de muestra: 123)

Método	Hiperparámetros				Variables de entrada							Puntajes		
	C	Alfa	Dual?	L1 ratio	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Mean Error / Accuracy*	*Precision*	*Recall*
Linear	-	-	-	-	Si	Si	Si	No	Si	No	Si	0.34	-	-
Lasso	-	0.01	-	-	Si	Si	Si	Si	Si	No	Si	0.33	-	-
SVR	**-**	**-**	**-**	**-**	**Si**	**No**	**Si**	**No**	**Si**	**Si**	**No**	**0.27**	**-**	**-**
Ridge	-	0.1	-	-	Si	Si	Si	Si	Si	Si	Si	0.34	-	-
ElasticNet	-	0.02	-	0.5	Si	Si	Si	Si	Si	Si	Si	0.33	-	-
Linear SVC	1	-	-	-	Si	No	Si	Si	Si	Si	No	0.73	0.73	0.77
RBF SVC	**10**	**-**	**-**	**-**	**Si**	**No**	**Si**	**No**	**Si**	**No**	**Si**	**0.85**	**0.79**	**0.92**
Sigmoid SVC	1	-	-	-	Si	No	Si	No	No	No	No	0.72	0.65	0.90
K Neighbors	-	-	-	-	Si	No	No	No	No	Si	No	0.79	0.76	0.82

Tabla 26. Mejores resultados algoritmos en San José.

Como se puede apreciar en las tablas, en materia de algoritmos de clasificación, el SVC con kernel RBF fue aquel que superó al resto en la gran mayoría de los casos, 17 de 19, y en algunos casos con buen margen.

Uno de los dos casos donde no resultó el mejor fue en el departamento de Lavalleja, donde la cantidad de muestras es pequeña (59) comparado con los demás departamentos y por este motivo no es considerado un departamento muy determinante para tomar de referencia con respecto a eventos de inundaciones en el país.

En cuanto a los algoritmos de regresión, el SVR fue el que dió mejor resultados en los 19 departamentos. En general los resultados no son tan positivos incluso con SVR, obteniéndose en promedio un error de entre 0.30 y 0.40 en la mayoría de los casos. Es interesante además, el hecho de que para el resto de los algoritmos de regresión los resultados son casi idénticos entre si para cada departamento. El hecho de que en un dominio tan acotado como éste la variación sea de 0.30, hace que se considere estos no adecuados para el problema en cuestión.

Por otra parte, al realizarse una investigación más profunda del tema que brindará una explicación a estos malos resultados, se encontró que no es recomendado utilizar regresión para problemas que tienen una salida binaria. Los principales motivos para esto son los siguientes [101]:

- Varias suposiciones que son base en regresión lineal son violadas en un problema de salida binaria: Continuidad y distribución normal de Y (la salida), la varianza de la media constante.
- La predicción puede estar fuera de rango. Esto es que un algoritmo entrenado de regresión, puede dar resultados fuera del rango (0,1) de probabilidades ya que su salida es, por naturaleza una variable continua.

Es por estas razones que se decidió descartar los métodos de regresión para las corridas definitivas de predicción a través de la aplicación web final. Sin embargo, el grupo considera que el haber probado estas técnicas sirvió como aprendizaje y para un mayor entendimiento de la tecnología y el área en general.

Dado que el algoritmo de SVC con kernel RBF fue el método elegido, se realizó un estudio particular de los resultados arrojados por el mismo, con el fin de determinar el valor óptimo del hiperparámetro C del método, así como también aquellas variables de entrada que tuvieron mayor influencia en el resultado y por ende, serán consideradas en el método a usarse definitivamente cuando sea necesario realizar predicciones.

En la tabla 27 se ilustran los mejores resultados obtenidos con el método elegido para cada departamento con el fin de poder compararlos más sencillamente.

En cuanto al hiperparámetro C, en vista de resultados oscilantes, se observa que éste no representó grandes cambios en los resultados ni hubo una inclinación muy grande por ninguno de los valores probados (aquellos sugeridos por la librería sci-kit learn). Es así que de manera intuitiva simplemente se decidió usar aquel que fue encontrado como óptimo mayor cantidad de veces. Por lo tanto se selecciona el valor 10, con una aparición de 9 veces.

Respecto a las variables de entrada, se observaron comportamientos dentro de lo esperado. Cuando se realizó el diseño de las corridas se determinaron las variables de entrada a probarse mediante intuición y los datos que se poseían. Estas fueron detalladas en el final del punto anterior sobre pruebas realizadas.

Las variables de precipitación del día corriente (PH), precipitación del día anterior (P-1), precipitación del día anterior al anterior (P-2) y alturas de río (alturas) fueron efectivamente utilizados en un rango de veces mayor al 68% de los casos, en el peor de los casos, que fue el de la precipitación del día anterior. Por una parte, se esperaba que éstas fueran importantes para la predicción, y de alguna manera estos resultados lo confirman, por lo tanto serán utilizadas en la predicción final.

Estudio particular método SVC con kernel RBF

Departamento (nº. muestras)	C	PH	PS	P-1	P-2	Estación	Mes	Alturas de Río	*Accuracy*	*Precision*	*Recall*
Artigas (111)	**1**	**Si**	**Si**	**Si**	**No**	**No**	**No**	**Si**	**0.79**	**0.75**	**0.85**
Canelones (283)	**50**	**Si**	**Si**	**Si**	**Si**	**No**	**No**	**Si**	**0.82**	**0.76**	**0.92**
Cerro Largo (167)	**10**	**Si**	**Si**	**No**	**Si**	**Si**	**No**	**No**	**0.70**	**0.65**	**0.77**
Colonia (163)	**10**	**Si**	**No**	**Si**	**No**	**Si**	**Si**	**Si**	**0.80**	**0.76**	**0.81**
Durazno (139)	**50**	**Si**	**No**	**No**	**Si**	**Si**	**Si**	**Si**	**0.80**	**0.78**	**0.82**
Flores (37)	**10**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**No**	**0.92**	**0.85**	**0.94**
Florida (159)	**50**	**Si**	**No**	**Si**	**Si**	**No**	**Si**	**Si**	**0.79**	**0.72**	**0.86**
Lavalleja (59)	10	Si	Si	No	No	No	No	No	0.69	0.62	0.75
Maldonado (185)	**10**	**Si**	**No**	**Si**	**No**	**No**	**No**	**Si**	**0.84**	**0.82**	**0.84**
Montevideo (235)	**1**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.79**	**0.73**	**0.90**
Paysandú (139)	1	Si	Si	Si	Si	No	No	Si	0.68	0.60	0.73
Soriano (191)	**10**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.70**	**0.67**	**0.8**
Tacuarembó (115)	**10**	**Si**	**No**	**Si**	**Si**	**No**	**No**	**No**	**0.80**	**0.74**	**0.89**
Salto (205)	**50**	**Si**	**Si**	**Si**	**No**	**Si**	**Si**	**Si**	**0.78**	**0.73**	**0.83**
Rivera (161)	**10**	**Si**	**Si**	**Si**	**No**	**Si**	**No**	**Si**	**0.74**	**0.71**	**0.88**
Rocha (169)	**50**	**Si**	**Si**	**Si**	**Si**	**Si**	**No**	**Si**	**0.79**	**0.72**	**0.92**
Río Negro (107)	**1**	**Si**	**Si**	**No**	**Si**	**No**	**No**	**Si**	**0.72**	**0.66**	**0.89**
Treinta y Tres (113)	**50**	**Si**	**Si**	**No**	**No**	**Si**	**No**	**Si**	**0.70**	**0.67**	**0.73**
San José (123)	**10**	**Si**	**No**	**Si**	**No**	**Si**	**No**	**Si**	**0.85**	**0.79**	**0.92**
	C	**PH**	**PS**	**P-1**	**P-2**	**Estación**	**Mes**	**Alturas de Río**	***Accuracy***	***Precision***	***Recall***
Totales	1 - 4 veces **10 - 9 veces** 50 - 6 veces	**19/19 100%**	**13/19 68%**	**14/19 73%**	**11/19 57%**	**10/19 52%**	**5/19 26%**	**15/19 78%**	**Promedio 0.77**	**Promedio 0.72**	**Promedio 0.84**

Tabla 27. Mejores resultados SVC con kernel RBF en cada departamento.

Además, se quiso agregar información referente a la época del año en que se encontraba cada dato, ya que de manera intuitiva se considera que la situación climática estacional hace variar los eventos de inundación, y en cada uno de éstos se presentan con mayor o

menor frecuencia. Es así que se decidió agregar dos variables de entrada y determinar si estaban involucradas realmente o no en los resultados. De hecho ninguna de estas dos variables es una variable en sí, sino un conjunto de variables booleanas discretizadas. El primero a utilizarse fue la estación del año. La estación puede ser verano, otoño, invierno o primavera en base a la fecha del día. Por otro lado, se discretizó también el mes al que pertenecía el día para probar su efectividad.

Los resultados arrojados no son claros ni determinantes, como suele suceder con las técnicas de aprendizaje automático, aunque hay uno que si se considera más fuerte que el resto. La discretización de los meses fue utilizada en tan sólo 5 corridas de las mejores 19, y por lo tanto con un porcentaje de 26% de los casos, se concluye que no realizó un mayor aporte a la predicción. Sin embargo, las estaciones se utilizaron un 57% de las veces, lo cual deja cierta incertidumbre sobre su utilización o no. Como grupo, se decidió mantener esta variable discretizada ya que consideramos puede llegar a ser de mayor utilidad con tamaños de muestra mayores y tampoco parece ser de influencia negativa.

Para concluir, se define como algoritmo a ser utilizado para las predicciones finales el SVC con kernel RBF, instanciándolo con hiperparámetro C en valor 10 y las variables de entrada a utilizarse para entrenar el modelo y para luego predecir nuevas ocurrencias de eventos serán: precipitación del día corriente, precipitación del día anterior, precipitación del día antes al día anterior, estación del año (discretizada en booleanos) y altura de los ríos.

5.5 Aplicación Web

Como parte final del proyecto y para completar la arquitectura general del proyecto descrita en la sección 5, se acordó en incluir el desarrollo de una aplicación web que permita visualizar el resultado del análisis predictivo aplicado sobre los datos que se obtuvieron en etapas anteriores.

5.5.1 Requerimientos para la Aplicación Web

Dada la naturaleza geolocalizada de la información manejada en el proyecto, se requirió como parte del mismo que los resultados obtenidos por el análisis predictivo realizado (sección 5.4) se muestren plasmados en un mapa del territorio Uruguayo, haciendo uso del aspecto geográfico mencionado.

Adicionalmente, el grupo se propuso desarrollar una herramienta de administrador mediante la cual los administradores del sistema pueden acceder a todas las entidades de la base de datos, en este caso ya en forma de modelos del framework Django (ver sección 5.5.3 a continuación), para poder realizar consultas, alta, baja y modificación de los datos de cada una.

Esto último resulta particularmente importante ya que consideramos tiene las prestaciones y la robustez necesaria para que eventualmente se use de manera oficial. Haciendo referencia a la sección 4.11 que habla de los datos abiertos en Uruguay y en particular para el caso de estudio de este proyecto, recordamos que la situación de manejo de datos y su

calidad en el país es pobre. Es así que diseñamos esta interfaz que posee potencial para que entidades oficiales del estado la utilicen con el fin de alimentar continuamente la base de datos de eventos ocurridos, a medida que vayan sucediendo en el correr del tiempo.

También se decidió permitir al usuario desplegar todas las estaciones existentes en la base de datos mediante marcadores ubicados en su posición. Además estos marcadores deben poder responder al evento (sección 4.5) de click y ante él, desplegar información básica de la estación como su nombre y ubicación.

Para la ejecución de la predicción de eventos, se presentarán al usuario dos opciones. La primera consiste en correr el algoritmo predictivo seleccionado para cada departamento del país, para el día corriente, de forma tal que retorne una predicción para el día siguiente. La otra alternativa será la de ingresar pronósticos futuros manualmente mediante un formulario que posea entradas para las diferentes variables de entradas determinadas. Se recurre a ésto ya que es lógicamente imposible obtener pronósticos de alturas de ríos y precipitaciones cuando no existen estaciones que las midan ni fuentes que las provean. Ésto permitirá a los usuarios obtener una aproximación de una predicción futura, sin poseer los datos referentes a la fecha sobre la cual se quiere predecir.

5.5.2 Análisis y Diseño de la Aplicación Web

Dados los requerimientos establecidos en la parte anterior, se decidió mostrar la ocurrencia o no de un evento en una localidad dada mediante el coloreado de su departamento mediante verde (negativo ó 0) cuando no sucedería ningún evento, y rojo (positivo ó 1) en el caso contrario. Para representar los departamentos se dibujará una capa (ver sección 4.5) que contenga todos los polígonos asociados a cada uno de los 19 departamentos de manera tal de que éstos puedan ser asignados un color de fondo.

Además, como los departamentos representan porciones territoriales bastante grandes y sin embargo se comprueba que la gran mayoría de los eventos se han dado en ciertas zonas reiteradas de éstos, resultó interesante destacar dichas localidades con marcadores clusterizados (ver sección 4.5) del mismo color que se coloree el departamento. De esta manera, se hace un énfasis y se muestran, de alguna manera, las zonas donde es más probable tener casos de inundación. Para ésto, se utilizará la funcionalidad de agrupación de marcadores (ver sección 4.5) provista por Google Maps para representar grupos de marcadores donde cada uno de ellos haga referencia a un evento sucedido en el pasado. Así se verán coloreados en tonos de gravedad desde colores más claros (marcadores agrupados con menor cantidad de eventos) hasta colores más oscures aquellos que contengan cantidades mayores.

5.5.3 Implementación

Django

Django es un framework de código abierto para el desarrollo en Python de aplicaciones web. El mismo tiene como primicia el desarrollo ágil y limpio, encargándose de manejar gran parte de las molestias que puede traer el desarrollo web (ruteo, conexión entre vistas

y controladores, etc), con el fin de que el desarrollo sea centrado en implementar la parte funcional de la aplicación.

Sus principales características son:

- La posibilidad de desarrollar de una manera rápida, minimizando la cantidad de configuraciones que deben realizarse en el proyecto.
- Centrado en la seguridad, evita que se cometan errores de seguridad comunes.
- Escalable y flexible [102].

Por otra parte, el hecho de ser un framework de Python es lo que nos permite la utilización e integración de librerías Python mencionadas anteriormente como son scikit-learn (machine learning) y scrapy. Pudiendo aprovechar de la ya mencionada amplia comunidad que tiene detrás el lenguaje, lo cual provee una abastecida fuente de material, tutoriales, guías y ejemplos.

En base a lo mencionado anteriormente, se decidió utilizar Django como framework para la implementación de la aplicación web, ya que considerando sus características, el mismo nos permitía reducir el tiempo invertido en configuraciones así como gracias a su estructura de proyecto, la posibilidad de integrar dentro del mismo en diferentes “apps” (término con el que se denomina a cada sección de un proyecto django): el motor predictivo, la obtención de datos mediante APIs y la interfaz web.

django-admin y django-admin-tools [103] [104]

Con el fin de gestionar las entidades y proveer una interfaz que facilite la consulta, alta, baja y modificación de la información, se decidió agregar una aplicación de administración de las entidades que se manejan en el sistema.

Para ello, el framework Django tiene como librería más popular a django-admin, y su librería separada, opcional, django-admin-tools que agrega una capa extra de personalización para la aplicación de administración.

Además, para gestionar accesos a dicha aplicación, se utilizó **django authentication system** [105], el cual facilita la gestión de la autenticación y los usuarios administradores del sistema mediante el mismo admin. Todas las vistas requieren login para ser desplegadas al usuario.

Google Maps (JavaScript)

Como referente de la tecnología global que es Google, y en particular su popular software SIG web llamado Maps [106], éste fue el elegido para, mediante una importación de su SDK en JavaScript, desplegar los resultados de las predicciones de manera geolocalizada.

Dado que en la sección 5.4.4 se resolvió utilizar una granularidad a nivel departamental, para desplegar los resultados de las predicciones se optó por crear una capa (ver sección 4.5) de tipo vectorial donde se dibujarán en forma de polígonos los diferentes departamentos, coloreando a los mismos con tonalidades diferentes en base a la potencial existencia de eventos de inundación en su territorio.

Además, como la ocurrencia de éstos eventos históricamente se reitera en ciertas ubicaciones y en otras nunca sucedió, se decidió indicar aquellas localidades donde sí existieron inundaciones utilizando agrupación de marcadores (ver sección 4.5). Se utiliza la característica de agrupación con el fin de presentar al usuario con marcadores más grandes en aquellos lugares donde hubieron más eventos y más chicos donde menos, llamando la atención de acuerdo a la probabilidad de una nueva ocurrencia.

Manual de Usuario

Finalmente y con el objetivo de ilustrar y complementar lo aquí descrito se incluye un documento como Anexo 2, al presente informe, un manual de usuario con un recorrido con capturas de pantalla e instrucciones para llevar a cabo todas las acciones posibles.

6 Conclusiones

En este capítulo se comenta el desarrollo del proyecto, así como las conclusiones y posibles trabajos futuros. Se hace una separación de cada sección de acuerdo a los temas que han sido abordados en el proyecto: la obtención de datos estáticos y dinámicos, el procesamiento de los mismos, la creación de un motor predictivo, y por último una aplicación Web donde se puedan visualizar los mismos.

6.1 Desarrollo del Proyecto

Este proyecto constó de la elaboración de una investigación resultante en el presente informe, además del desarrollo del código descrito también a través del mismo, con el fin de generar una base de datos que cumpla ciertos mínimos de calidad de datos explicados en la sección 5.3 que permitan un análisis predictivo sobre los mismos. Dicho análisis busca predecir la ocurrencia de eventos de inundación, con el objetivo de sentar un precedente en este enfoque probabilístico de resolución del problema, a fin de que éstos puedan ser extendidos en el futuro.

Se considera que el proyecto cumplió con su alcance y obtuvo muy buenos resultados. A nivel de obtención de datos se logró una base de datos amplísima, con más de un millón de registros concernientes a condiciones climatológicas que afectan a los eventos de inundación. También se logró una predicción dentro de parámetros aceptables que intenta predecir la ocurrencia o no de un evento de ese tipo, y finalmente, se desarrolló una aplicación web, incluyendo un Sistema de Información Geográfica mediante el cual se presentan de manera amigable los resultados predichos recientemente mencionados.

El grupo de proyecto ya contaba con experiencia en desarrollo web y manejo de base de datos. Sin embargo, no poseía experiencia en la limpieza y calidad de los datos, así como tampoco en el lenguaje de programación utilizado.

En cuanto a las técnicas de aprendizaje automático, fue un área totalmente nueva para todo el equipo, siendo éste un primer contacto en el transcurso del proyecto y del cual el grupo considera que se lleva un valioso conocimiento sobre un área incipiente dentro de las tecnologías de la información y que seguramente sea utilizada incluso en los software más cotidianos.

Adicionalmente, se estuvo en contacto con entes y organizaciones públicas con los cuales ninguno de los miembros del equipo había entablado conversaciones y sobre los cuales también se aprendió mucho, en particular sobre el estado de sus centros de cómputo y la información que éstos manejan.

6.2 Conclusiones

En primer lugar, resulta importante dar una conclusión grupal sobre el estado y accesibilidad de los datos públicos en Uruguay, y en particular sobre aquellos que tratan sobre meteorología y eventos de desastres naturales.

Dentro del marco teórico dado en las secciones 4.11 y 5.3.1 sobre datos públicos y calidad de datos respectivamente, se notó que la situación de los datos públicos del Uruguay es ampliamente mejorable. Si bien el estado Uruguayo se encuentra medianamente bien posicionado en los ranking internacionales elaborados sobre el tema, y libera una buena parte de los datos, una de sus principales problemáticas reside justamente en su usabilidad. Además, en el caso particular de aquellos datos requeridos para el proyecto, su disponibilidad tampoco fue buena.

Se trabajó más de lo esperado para obtener datos históricos tanto de métricas meteorológicas como de ocurrencia y registro de eventos, que si bien existen, no están accesibles de manera automática a quién los quiera consultar, y además poseen en su gran mayoría un formato ilegible de manera automatizada lo cual compromete su usabilidad. Además, desde el 2014 en adelante ya ningún organismo público se encarga de registrar y documentar eventos desastrosos y sus consecuencias, lo cual genera incompletitud y desactualización de la información, y, a los efectos de la aplicación que se quiere dar en el presente trabajo, dejan a los últimos 3 años inútiles en materia de datos ya que no se sabe qué sucedió en cada uno de los días.

En cuanto a la calidad de los datos, también se notaron carencias en la normalización, unicidad y presentación de los mismos, muchas veces teniendo éstos faltas graves en los mencionados aspectos, como fue desarrollado en la sección 5.3.1. El grupo considera que dados los importantes y recientes avances y la sencillez con las cuales se pueden crear servicios web accesibles pública y automáticamente, dichos organismos, no sólo para ésta pero para todas las áreas, deberían trabajar en la normalización, mejora de calidad y exposición de los datos de manera que la población pueda consultarlos, aumentando su transparencia, y generar contenido en base a los mismos, por ejemplo con la posibilidad de aumentar la calidad de vida de los ciudadanos.

En este contexto, este proyecto aporta una completa base de datos ya normalizada, de fácil acceso mediante el sitio de administración de la aplicación web, la cual puede servir como base para proyectos futuros y como un buen ejemplo de integración de datos, contando la misma con datos de alturas, precipitaciones, eventos registrados y la ubicación geográfica de todas las estaciones entre otros, además de actualización diaria en lo que respecta a los datos de precipitaciones y alturas.

En lo que respecta a la segunda parte del proyecto, se quiere destacar la aplicabilidad y utilidad de las técnicas de aprendizaje automático a la predicción de este tipo de eventos de la realidad, no sin antes mencionar que es un área en desarrollo y crecimiento. Teniendo ésto último en cuenta, el grupo de investigación se vió gratamente sorprendido no sólo por los resultados alcanzados, sino también con la disponibilidad y accesibilidad con la que se puede hacer uso de estas técnicas. Además, luego de tener contacto con aprendizaje

automático resultó inmediato imaginar diversas aplicaciones para este relativamente nuevo enfoque de resolución de problemas que de otra manera resultaban más complejos o costosos computacionalmente. Particularmente sobre scikit-learn, se encontró una completa y detallada documentación, una librería extensa, probada y sencilla de usar, que no requiere a quién le quiera dar uso un profundo conocimiento de las técnicas sobre las cuales basa su funcionamiento y expone de manera clara los métodos necesarios tanto para el preprocesamiento de los datos como la utilización de los mismos aplicando los algoritmos descritos con más en la sección 4.6.3 del Anexo 1. Además, ante todo inconveniente durante la realización de ésta etapa, nos encontramos al consultar con una importante comunidad científica que trabaja con esto, lo cual facilita el desarrollo.

En cuanto a los resultados obtenidos para las predicciones, se consideran aceptables para lo acordado en el alcance del proyecto, pero sin lugar a dudas tienen espacio para la mejora, tanto en la etapa de armado de datos de entrada, como en el entrenamiento de los modelo usados, ya que la complejidad que presentan es amplia y las posibilidades de ajuste de los mismos son infinitas. Con el algoritmo de predicción seleccionado se obtuvieron valores de precisión entre un 70 y 80% en las validaciones dependiendo del departamento en cuestión con una muestra de datos por departamento desde 40 entradas hasta casi 300 en otros, número que también puede ser mejorado con la obtención de más datos de desastres en el país. Finalmente, se quiere aclarar que aún no se consideran a estos resultados como algo realmente aplicable a la realidad en forma de alerta pública de ningún tipo, si no que el enfoque fue probar la utilidad de los datos obtenidos en forma de predicción de un fenómeno que tanto ha afectado el territorio nacional, punto que se demostró.

6.3 Trabajos futuros

No se descarta que mediante un trabajo futuro que valide y alcance menores desviaciones de error en las predicciones, estas técnicas pueden usarse como método alternativo y de apoyo a las técnicas meteorológicas e hidrológicas clásicas aplicadas hasta el momento. Para esto, debería trabajarse con mayor detalle sobre cada uno de los modelos entrenados buscando obtener los mejores resultados, con menor error posible, para cada situación. Por ejemplo, con respecto al armado de los datos de entrada para cada modelo, se puede trabajar sobre el tema tanto como los datos y la creatividad, experiencia y conocimiento del desarrollador lo permitan. Los métodos de aprendizaje automático no tienen una única manera de recibir datos, como se vio en la sección 5.4.3, y esto entonces presenta muchas más posibilidades de las que pudieron ser probadas y validadas en el proyecto por motivos de alcance y duración.

También se podría trabajar en modificar la granularidad geográfica con la cual se agrupan los datos previo a alimentar los modelos. Como ya se mencionó, en este proyecto se optó por utilizar la división política del país en departamentos y utilizamos dicha clasificación geográfica para agruparlos. Consideramos que fue un buen punto de partido sin embargo

no dudamos que modelos ajustados a regiones geográficas más específicas o agrupadas en base a otros criterios puedan dar mejores resultados.

Como trabajo futuro también se pueden expandir las fuentes para el obtención de datos de forma dinámica, con el fin de tener fuentes y consecuentemente datos de respaldo en caso de que alguna falle o cambie la estructura de su sitio web, algo que a lo largo del informe se explica como riesgo. En este sentido es importante recordar que la obtención de datos de precipitaciones y alturas de las webs de TuTiempo y de la UTE se realiza mediante web scraping, lo cual brinda cierta fragilidad ya que ante cualquier cambio en la web de donde se obtiene la información, es muy probable que ya no se puedan obtener los datos sin realizar cambios en la implementación del sistema.

7 Glosario

Desastre: Evento caracteristicas destructivas, repentino o previsible, que trastorna el funcionamiento habitual de la sociedad, teniendo como consecuencia pérdidas humanas, materiales, económicas o ambientales.

Prevención: es la acción anticipada para impedir que ocurra un fenómeno peligroso, o para evitar su incidencia negativa sobre la población, los bienes y el ambiente.

Mitigación: son las medidas para atenuar el impacto de los fenómenos adversos asumiendo que no siempre es posible evitarlos.

Preparación: son las actividades orientadas a asegurar la disponibilidad de los recursos y la efectividad de los procedimientos para enfrentar una situación de emergencia.

Atención de emergencias: es el conjunto de acciones de respuesta para proteger a la población, los bienes y el ambiente ante la ocurrencia de un evento adverso.

Rehabilitación: es la puesta en funcionamiento, en el menor tiempo posible, de los servicios básicos afectados por un evento adverso.

Recuperación: luego de un evento adverso, es el esfuerzo por promover condiciones de vida adecuadas y sostenibles, incluyendo la reactivación del desarrollo económico y social de la comunidad en condiciones más seguras

Web Scraping: El web scraping es una técnica que sirve para extraer información de páginas web de forma automatizada. Esta técnica se aplica recorriendo el DOM de una web de la que se desea extraer información, utilizando su estructura y composición para acceder a la información deseada.

DOM: Document Object Model, es un modelo y estándar que define cómo se deben estructurar y conectar los elementos que componen un sitio web. Además presenta una interfaz para que otros lenguajes (por ejemplo JavaScript) puedan manipular dichos elementos.

Multi-tenant: es un modelo de arquitectura que consiste en una única instancia corre en un servidor y varios clientes se conectan a está.

JDBC: Java Database Connectivity, interfaz para conectar una aplicación Java con una base de datos.

Heroku: es un servicio que permite desarrollar y alojar aplicaciones en la nube, soportando diversos lenguajes de programación, entre ellos Python e incluyendo el uso de una base de datos postgres por proyecto. En su versión gratuita, permite el uso de múltiples complementos que son herramientas y servicios para desarrollar, extender y operar su aplicación. Un ejemplo de complemento es el que se utiliza para la realización de tareas automáticas cada cierto periodo de tiempo.

E-gobierno: el Gobierno electrónico (E-gobierno) puede definirse, según los conceptos más recibidos, como una nueva forma de interacción o relación entre los Gobiernos de los distintos países y sus respectivos ciudadanos o personas que eventualmente tengan contacto con ellos. Esta nueva forma consiste en la implementación, desarrollo y aplicación de las herramientas informáticas tales como las tecnologías de la información y las comunicaciones.

Datos: Información dispuesta de manera adecuada para su tratamiento por un ordenador.

8 Referencias

[1] Manual para la Evaluación de Desastres, ONU/CEPAL, 2014.
Consultado en: repositorio.cepal.org/bitstream/handle/11362/35894/S2013806_es.pdf
Accedido en: Marzo/2017

[2] Acerca de la ONU
Consultado en: http://www.un.org/en/sections/about-un/overview/index.html
Accedido en: Mayo/2017

[3] Acerca de la CEPAL
Consultado en: http://www.cepal.org/en/about
Accedido en: Mayo/2016

[4] Manual para la evaluación del impacto socioeconómico y ambiental de los Desastres, ONU/CEPAL, 2003.
Consultado en: repositorio.cepal.org/bitstream/handle/11362/2781/5/S2003652_es.pdf
Accedido en: Marzo/2017

[5] Université Catholique de Louvain
Consultado en: https://www.uclouvain.be/index.html
Accedido en: Septiembre/2016

[6] EM-DAT: Preguntas frecuentes
Consultado en: http://www.emdat.be/frequently-asked-questions
Accedido en: Septiembre/2016

[7] ¿Que es Desinventar?
Consultado en: http://www.desinventar.net/whatisdesinventar.html
Accedido en: Abril/2017

[8] Bases de datos de Desinventar
Consultado en: http://www.desinventar.org/es/database
Accedido en: Abril/2017

[9] EM-DAT: The OFDA/CRED International Disaster Database. V11.8
Université catholique de Louvain, Brussels, Bélgica.
Consultado en: www.emdat.be
Accedido en: Septiembre/2016

[10] Glosario Hidrológico Internacional.
Consultado en: unesdoc.unesco.org/images/0022/002218/221862M.pdf
Accedido en: Diciembre/2016

[11] Inundaciones en Uruguay, SINAE.
Consultado en: sinae.gub.uy/sistema-de-informacion/amenazas/inundaciones/
Accedido en: Diciembre/2016

[12] Clasificación de inundaciones.
Consultado en: www.floodup.ub.edu/clasificacion/
Accedido en: Febrero/2017

[13] Tipos de Inundaciones
Consultado en: https://www.gob.mx/presidencia/articulos/tipos-de-lluvias-e-inundaciones
Accedido en: Febrero/2017

[14] Causas de Inundaciones
Consultado en:
www.eartheclipse.com/natural-disaster/what-is-flood-and-what-causes-flooding.html
Accedido en: Febrero/2017

[15] Artículo de Vital Perú, 2017
Consultado en:
vital.rpp.pe/salud/la-leptospirosis-enfermedad-infecciosa-de-las-inundaciones-noticia-1038222
Accedido en Marzo: 2017

[16] Síntesis histórica de emergencias ocurridas en Uruguay
Consultado en: http://archivo.presidencia.gub.uy/sne/htm/sne_historico02.htm
Accedido en: Abril/2017

[17] Inundación de 1959 en Uruguay
Consultado en: http://www.miuruguay.com/2009/11/inundaciones-en-uruguay.html
Accedido en: Abril/2016

[18] Gestión de Riesgos en Desastres. :
Lizardo Narváez, Allan Lavell, Gustavo Pérez Ortega, Secretaria General de la Comunidad Andina, Perú, ISBN: 978-9972-787-88-1.
Consultado en: http://www.comunidadandina.org/predecan/doc/libros/procesos_ok.pdf

[19] Definición de Gestión de Riesgos de la ONU
Consultado en: https://www.unisdr.org/we/inform/terminology
Accedido en: Mayo/2017

[20] The evolution of emergency management.
Artículo de T. E.Drabek, 1991. Washington, DC: International City Management Association.

[21] Gestión Integral del Riesgo
Consultado en: sinae.gub.uy/conceptos-basicos/gestion-integral-del-riesgo/
Accedido en: Noviembre/2016

[22] Ciclo de la Gestión de Riesgos
Consultado en:
https://anotherverse.wordpress.com/2014/12/22/icts-in-disaster-risk-management/
Accedido en: Noviembre/2016

[23] Definición del SINAE
Consultado en: http://sinae.gub.uy/institucional/definicion-del-sinae/
Accedido en: Febrero/2017

[24] Sistema Nacional de Emergencias y Gestión del Riesgo de Desastres en Uruguay. Presidencia y SINAE, Uruguay.
Consultado en:
http://www.uy.undp.org/content/dam/uruguay/docs/MAyE/Gu%C3%ADa_1_GIR.pdf

[25] Ley 19158 de “Creación del Instituto Uruguayo de Meteorología”
Consultado en: http://meteorologia.gub.uy/reportes/institucional/LEY_19158.pdf
Accedido en: Febrero/2017

[26] Misión de INUMET
Consultado en: http://www.meteorologia.com.uy/institucional/mision
Accedido en: Abril/2017

[27] DW2.0 – Architecture for the Next Generation of Data Warehousing W.H. Inmon, Derek Strauss, Genia Neushloss. Morgan-Kaufman, 2008. ISBN: 978-0-12-374319-0

[28] The Data Warehouse Toolkit
R. Kimball. John Wiley & Sons, 2002. ISBN: 0-471-20024-7

[29] Definición SIG, Curso Introducción a los Sistemas de Información Geográfica de la FIng.
Consultado en: https://www.fing.edu.uy/inco/cursos/sig/clases/Generalidades160812.pdf
Accedido en: Abril/2017

[30] Modelo de un SIG, ESRI
Consultado en: http://downloads2.esri.com/ESRIpress/images/147/think3_ch1_SP.pdf
Accedido en: Abril/2017

[31] Goodchild, M. F. (1993). The state of GIS for environmental problem-solving. Environmental modeling with GIS, 8-15. ISO 690

[32] Información de Data-Mining
Consultado en: http://www.sinnexus.com/business_intelligence/datamining.aspx
Accedido en: Abril/2017

[33] ¿Que es Data-Mining?
Consultado en:
anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
Accedido en: Mayo/2016

[34] Kerrison Predictor
Consultado en:
https://www.revolvy.com/main/index.php?s=Kerrison%20Predictor&item_type=topic
Accedido en: Mayo/2017

[35] Predictive Analytics
Consultado en: https://www.sas.com/en_us/insights/analytics/predictive-analytics.html#
Accedido en: Mayo/2017

[36] Introducción a Machine Learning
Consultado en: http://scikit-learn.org/stable/tutorial/basic/tutorial.html
Accedido en: Junio/2016

[37] Machine Learning: Aprendizaje Supervisado
Consultado en:
scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html
Accedido en: Junio/2016

[38] Machine Learning: Mapa de algoritmos
Consultado en: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html
Accedido en: Junio/2017

[39] Cross Validation: evaluating estimator performance
Consultado en: http://scikit-learn.org/stable/modules/cross_validation.html
Accedido en: Junio/2017

[40] Cross Validation
Consultado en: https://www.cs.cmu.edu/~schneide/tut5/node42.html
Accedido en: Junio/2017

[41] SVR
Consultado en: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
Accedido en: Junio/2017

[42] Interpretación de R2
Consultado en:
http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
Accedido en: Julio/2017

[43] Coeficiente de determinación decisión
Consultado en:
http://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination
Accedido en: Julio/2017

[44] Weka Software
Consultado en: http://www.cs.waikato.ac.nz/ml/weka/index.html
Accedido en: Abril/2017

[45] Caracteristicas WEKA
Consultado en: http://cor-mineriadedatos.blogspot.com.uy/2011/06/weka.html
Accedido en: Abril/2017

[46] R-Project
Consultado en: https://www.r-project.org/about.html
Accedido en: Abril/2017

[47] Librerías distribuidas para R
Consultado en: https://cran.r-project.org/web/packages/
Accedido en: Abril/2017

[48] Introducción a la programación estadística con R para profesores.
Jose Contreras, Elena Molina, Pedro Arteaga, ISBN: 978-84-693-4859-8.
Consultado en: http://www.ugr.es/~batanero/pages/ARTICULOS/libroR.pdf

[49] Portada Pentaho Business Intelligence
Consultado en: http://pentaho.almacen-datos.com/
Accedido en: Abril/2017

[50] Pentaho información
Consultado en: http://www.pentaho.com/about
Accedido en: Abril/2017

[51] Predicting Floods, NASA
Consultado en: http://science.nasa.gov/science-news/science-at-nasa/2015/22jul_floods/
Accedido en: Febrero/2017

[52] Global Flood and Landslide Monitoring
Consultado en: http://pmm.nasa.gov/trmm/flood-and-landslide-monitoring
Accedido en: Febrero/2017

[53] Global Flood Monitoring System
Consultado en: http://flood.umd.edu/
Accedido en: Febrero/2017

[54] Informe sobre la respuesta del SINAE frente a los eventos hidrometeorológicos de Febrero de 2014
Consultado en:
http://reliefweb.int/sites/reliefweb.int/files/resources/Informe%20sobre%20la%20respuesta%20del%20SINAE%20frente%20a%20los%20eventos%20hidrometeorologicos%20de%20febrero%20de%202014.pdf
Accedido en: Febrero/2017

[55] Prohi: Extensión del sistema Prohimet-Yí de alerta temprana de inundaciones en la ciudad de Durazno.
Patricia Martínez Suiffet, Martín Mochetti Grolero, Agustín Nin Castro, tutor Bruno Rienzi. Proyecto de Grado, Facultad de Ingeniería de la UDELAR, Junio 2013.

[56] Conceptos Básicos de Datos Abiertos
Consultado en:
https://www.agesic.gub.uy/innovaportal/v/1544/1/agesic/conceptos-basicos-de-datos-abiertos.html
Accedido en: Junio/2017

[57] Open Data Government: Los 8 principios de los datos Abiertos.
Consultado en: https://opengovdata.org/
Accedido en: Junio/2017

[58] Información Pública y datos abiertos en Uruguay. ¿Qué licencia usamos?
Patricia Diaz, Matias Jackson, 2015. Facultad de Ingeniería de Uruguay
Consultado en:
https://eva.fing.edu.uy/pluginfile.php/99011/mod_resource/content/1/ETHICOMP%202015%20Arti%CC%81culo%20Jackson%20Di%CC%81az.pdf

[59] Ley 18.381
Consultado en: http://archivo.presidencia.gub.uy/_web/leyes/2008/10/EC1028-00001.pdf
Accedido en: Junio/2017

[60] Catálogo de Datos Uruguay
Consultado en: https://catalogodatos.gub.uy/
Accedido en: Junio/2017

[61] ¿Que es la UTE?
Consultado en: http://portal.ute.com.uy/institucional/qui%C3%A9nes-somos
Accedido en: Febrero/2017

[62] EM-DAT: Perfiles de países
Consultado en: http://www.emdat.be/country_profile/index.html
Accedido en: Septiembre/2016

[63] Información compañía Wunderground
Consultado en: https://www.wunderground.com/about/our-company
Accedido en: Enero/2017

[64] Información Wunderground
Consultado en: www.wunderground.com/blog/JeffMasters/comment.html?entrynum=3170
Accedido en: Enero/2017

[65] Información datos Wunderground
Consultado en: https://www.wunderground.com/about/data
Accedido en: Enero/2017

[66] API Wunderground
Consultado en: https://www.wunderground.com/weather/api/d/docs?d=data/index
Accedido en: Enero/2017

[67] Informacion Accuweather
Consultado en: http://www.accuweather.com/en/about
Accedido en: Enero/2017

[68] API Accuweather
Consultado en: http://developer.accuweather.com/accuweather-forecast-api/apis
Accedido en: Enero/2017

[69] Pricing y paquetes Accuweather
Consultado en: http://developer.accuweather.com/pricing-and-data-packages
Accedido en: Enero/2017

[70] Información Worldweather
Consultado en: https://www.worldweatheronline.com/aboutus.aspx
Accedido en: Enero/2017

[71] API Worldweather
Consultado en: https://developer.worldweatheronline.com/premium-api-explorer.aspx
Accedido en: Enero/2017

[72] Información Portal UTE-i
Consultado en: http://portal.ute.com.uy/institucional/ute-i
Accedido en: Diciembre/2016

[73] Información datos climáticos Tutiempo
Consultado en: http://blog.tutiempo.net/informacion-de-datos-climaticos/
Accedido en: Enero/2017

[74] Información acceso datos climáticos Tutiempo
Consultado en: http://blog.tutiempo.net/acceso-listados-xml-del-tiempo/
Accedido en: Enero/2017

[75] Validez de los datos de Tutiempo
Consultado en: https://blog.tutiempo.net/clima-tutiempo/
Accedido en: Abril/2017

[76] ¿Que es Python?
Consultado en: https://www.python.org/doc/essays/blurb/
Accedido en: Septiembre/2016

[77] Repositorio de Scrapy
Consultado en: https://github.com/scrapy/scrapy
Accedido en: Septiembre/2016

[78] Documentación Scrapy
Consultado en: https://doc.scrapy.org/en/latest/index.html
Accedido en: Septiembre/2016

[79] Diseño de patrones, Patrón Adapter.
Consultado en: http://arantxa.ii.uam.es/~eguerra/docencia/0809/03%20Adapter.pdf
Accedido en: Marzo/2016

[80] Documentacion selectores Scrapy
Consultado en: https://doc.scrapy.org/en/latest/topics/selectors.html
Accedido en: Septiembre/2016

[81] Un caso de estudio en Calidad de Datos para Ingeniería de Software Empírica
Bruno Bianchi Gallo, María Carolina Valverde Corrado
Proyecto de Grado, Facultad de Ingeniería de la UDELAR, Diciembre 2009.

[82] Repositorio de OpenRefine
Consultado en: https://github.com/OpenRefine/OpenRefine
Accedido en: Julio/2016

[83] Historia de OpenRefine
Consultado en: http://openrefine.org/2013/10/12/openrefine-history.html
Accedido en: Julio/2016

[84] Artículo de La Nación: ¿Como usar OpenRefine?
Consultado en:
http://blogs.lanacion.com.ar/data/datos-abiertos/como-usar-google-refine-para-trabajar-una-base-de-datos/
Accedido en: Julio/2016

[85] Wiki de General Refine Expression Language
Consultado en:
https://github.com/OpenRefine/OpenRefine/wiki/General-Refine-Expression-Language
Accedido en: Julio/2016

[86] Sitio web RefinePro
Consultado en: http://refinepro.com/
Accedido en: Julio/2016

[87] Clusterización en profundidad
Consultado en: https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth
Accedido en: Mayo/2017

[88] Algoritmos de Similaridad y Distancia
Consultado en:
http://www.kramirez.net/wp-content/uploads/2012/02/Algoritmos-de-Similaridad-y-Distancia.pdf
Accedido en: Mayo/2017

[89] The similarity metric. Ming Li , Xin Chen, Xin Li, Bin Ma, Paul Vitanyi.
Consultado en: https://arxiv.org/abs/cs/0111054
Accedido en: Mayo/2017

[90] Complejidad de Kolmogorov
Consultado en: https://people.cs.uchicago.edu/~fortnow/papers/kaikoura.pdf
Accedido en: Mayo/2017

[91] Fundamentos de la normalización de bases de datos
Consultado en:
https://support.microsoft.com/es-uy/help/283878/description-of-the-database-normalization-basics
Accedido en: Mayo/2017

[92] Geocodificación Reversa
Consultado en:
https://developers.google.com/maps/documentation/javascript/examples/geocoding-reverse
Accedido en: Enero/2017

[93] Acerca de Postgresql
Consultado en: https://www.postgresql.org/about/
Accedido en: Octubre/2016

[94] Comandos PSQL
Consultado en: https://www.postgresql.org/docs/9.2/static/sql-copy.html
Accedido en: Agosto/2016

[95] Charla sobre Machine Learning en Facultad de Ingeniería
Consultado en:
http://www.ricaldoni.org.uy/noticia/charla-machine-learning-101-introduccion-tecnica-nueva-electricidad

[96] StratifiedKFold
Consultado en:
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
Accedido en: Junio/2017

[97] Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, pp. 1137-1145). ISBN:1-55860-363-8
Consultado en: http://frostiebek.free.fr/docs/Machine%20Learning/validation-1.pdf
Accedido en: Junio/2017

[98] Efron, B. (1979). Bootstrap methods: another look at the jackknife. The annals of Statistics, 1-26.
Consultado en: http://www.jstor.org/stable/2958830
Accedido en: Junio/2017

[99] cross_val_score
Consultado en:
http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html#sklearn.model_selection.cross_val_score
Accedido en: Junio/2017

[100] Discover Feature Engineering
Consultado en:
http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/
Accedido en: Junio/2017

[101] Why shouldn't I use linear regression if my outcome is binary?
Consultado en:
http://thestatsgeek.com/2015/01/17/why-shouldnt-i-use-linear-regression-if-my-outcome-is-binary/
Accedido en: Julio/2017

[102] Documentación Django
Consultado en: https://www.djangoproject.com/
Accedido en: Septiembre/2016

[103] Django Admin
Consultado en: https://docs.djangoproject.com/en/1.11/ref/contrib/admin/
Accedido en: Junio/2017

[104] Django Admin Tools
Consultado en: https://github.com/django-admin-tools/django-admin-tools
Accedido en: Junio/2017

[105] Django Authentication System
Consultado en: https://docs.djangoproject.com/en/1.11/topics/auth/
Accedido en: Junio/2017

[106] Google Maps
Consultado en:
https://developers.google.com/maps/documentation/javascript/?hl=es-419
Accedido en: Junio/2017

[107] TryoLabs
Consultado en: https://tryolabs.com/
Accedido en: Junio/2017

[108] La ONU y la Gestión de Riesgos de desastres
Consultado en:
http://www.un-spider.org/es/riesgos-y-desastres/ONU-y-gesti%C3%B3n-del-riesgo-de-desastres
Accedido en: Mayo/2017

[109] Eventos extremos de tiempo y clima en Uruguay.
INUMET, Octubre de 2014
Consultado en: http://www.meteorologia.com.uy/reportes/escuela/Eventos_extremos.pdf

[110] Artículo de El País, 2014
Consultado en: elpais.com.uy/informacion/resisten-riada-temor-robos-inundaciones.html
Accedido en: Marzo/2017

[111] Artículo de Diario Clarin, 2015
Consultado en:
clarin.com/sociedad/drama-inundacion-bichos-brote-enfermedades_0_Nysflp58x.html
Accedido en: Marzo/2017

[112] Sistema Nacional de Emergencias y Gestión del Riesgo de Desastres en Uruguay.
Presidencia y SINAE, Uruguay.
Consultado en:
http://www.uy.undp.org/content/dam/uruguay/docs/MAyE/Gu%C3%ADa_1_GIR.pdf
Accedido en: Abril:2016

[113] Acerca de la Cruz Roja
Consultado en: http://www.ifrc.org/es/nuestra-vision-nuestra-mision/
Accedido en: Mayo/2017

[114] Definición de la Cruz Roja de Gestión de Riesgos
Consultado en:
http://www.ifrc.org/es/introduccion/disaster-management/gestion-de-desastres/
Accedido en: Mayo/2017

[115] Acerca de la Organización Mundial de la Salud
Consultado en: http://www.who.int/about/es/
Accedido en: Mayo/2017

[116] Definición de la OMS de Gestión de Riesgos
Consultado en: http://apps.who.int/disasters/repo/7679.pdf
Accedido en: Mayo/2017

[117] Introduction to Emergency Management.
G. Haddow, J. Bullock, 2004. Amsterdam: Butterworth-Heinemann ISBN 0-7506-7689-2.

[118] Información acerca de SINAE e INUMET
Informe sobre la Respuesta del SINAE frente a los Eventos Hidrometeorológicos. Febrero de 2014.
SINAE, Presidencia, 2014.

[119] SIG aplicados al análisis y cartografía de riesgos climáticos
Consultado en: http://www.um.es/geograf/sigmur/cursos/SIG_clima.pdf
Accedido en: Mayo/2017

[120] Marker Clustering
Consultado en:
https://developers.google.com/maps/documentation/javascript/marker-clustering
Accedido en: Mayo/2017

[121] Electronic Numeric Integrator And Calculator
Consultado en: http://www.thocp.net/hardware/eniac.htm
Accedido en: Mayo/2017

[122] The first Numerical Weather Prediction on ENIAC
Consultado en:
http://www.easterbrook.ca/steve/2011/01/the-first-numerical-weather-prediction-on-eniac/
Accedido en: Mayo/2017

[123] A Look at the History and Future of Predictive Analytics and Big Data
Consultado en:
https://visual.ly/look-history-and-future-predictive-analytics-and-big-data?utm_source=visually_embed
Accedido en: Mayo/2017

[124] Algoritmos Support Vector Machine
Consultado en: http://scikit-learn.org/stable/modules/svm.html
Accedido en: Junio 2017

[125] Algoritmos Support Vector Classification
Consultado en: http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
Accedido en: Junio/2017

[126] Algoritmo NU SVC
Consultado en: http://scikit-learn.org/stable/modules/generated/sklearn.svm.NuSVC.html
Accedido en: Junio/2017

[127] Teorema de Bayes
Consultado en:
https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/
Accedido en: Junio/2017

[128] Algoritmo Naive Bayes
Consultado en: http://scikit-learn.org/stable/modules/naive_bayes.html
Accedido en: Junio/2017

[129] Algoritmo Stochastic Gradient Descent
Consultado en: http://scikit-learn.org/stable/modules/sgd.html
Accedido en: Junio/2017

[130] Algoritmo Nearest Neighbors
Consultado en: http://scikit-learn.org/stable/modules/neighbors.html
Accedido en: Junio/2017

[131] Linear Regression
Consultado en:

http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
Accedido en: Junio/2017

[132] General Linear Models
Consultado en: http://scikit-learn.org/stable/modules/linear_model.html
Accedido en: Junio/2017

[133] Ridge
Consultado en:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html
Accedido en: Junio/2017

[134] Lasso
Consultado en:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html
Accedido en: Junio/2017

[135] Elastic Net
Consultado en:
http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html
Accedido en: Junio/2017

[136] Support Vector Regression: Propiedades y Aplicaciones. Juan José Martín Guareño.
Consultado en:
https://idus.us.es/xmlui/bitstream/handle/11441/43808/Mart%C3%ADn%20Guare%C3%B1o%2C%20Juan%20Jos%C3%A9%20TFG.pdf?sequence=1&isAllowed=y
Accedido en: Junio/2017

[137] Real-time Global Flood Estimation using Satellite-based Precipitation and a Coupled Land Surface and Routing Model. Huan Wu, Robert F. Adler, Yudong Tian, George J. Huffman, Hongyi Li and JianJian Wang. Earth System Science Interdisciplinary Center, University of Maryland.

[138] Open Data Barometer
Consultado en: http://www.opendatabarometer.org/
Accedido en: Junio/2017

[139] Global Open Data Index
Consultado en: http://index.okfn.org/place/uruguay/
Accedido en: Junio/2017

Instituto de Computación - Facultad de Ingeniería - Universidad de la República

Montevideo, Uruguay, 2016

Proyecto de Grado - Anexo 1

Estado del Arte

Tutor Libertad Tansini

Co-tutor Sandro Moscatelli

Ignacio Chiazzo

Felipe Garcia

Guillermo Leopold

1 Introducción

En este anexo se presenta el estado del arte, haciéndose un estudio del contexto que rodea la problemática así como de las opciones y de las tecnologías actuales que fueron objetivo de estudio y que son requeridas o fueron consideradas a la hora de desarrollar la solución.

Se comienza presentando un contexto general acerca de los desastres naturales, definiendo qué son y trasladando la temática a la región y principalmente a Uruguay. Luego, una vez que resulta claro que el desastre que más daños provoca en el país son las inundaciones, se realiza un estudio de las mismas, pasando por una definición de qué se considera una inundación, los diversos tipos, causas y consecuencias y un breve repaso cronológico de eventos de este tipo ocurridos en Uruguay. Finalmente en lo que se podría considerar como el final de esta primera parte del estado del arte, se adentra en lo que es la gestión de riesgos de desastres, brindando una definición de lo que es la misma, describiendo sus etapas, cómo se aplica, así como desarrollando en detalle su contexto y funcionamiento en nuestro país.

En lo que se puede considerar la segunda parte de este capítulo, ya que a partir de acá el estudio se centra en las tecnologías, se comienza con una explicación de lo que es un sistema de información geográfica y su funcionamiento y utilidad. Luego, se estudian algunos conceptos previos a lo que es el aprendizaje automático, investigando acerca de la minería de datos (o data mining), análisis predictivo y modelo predictivo; para luego ahondar en el aprendizaje automático terminando con las opciones consideradas a la hora de implementar un modelo predictivo.

Por último, en una tercer y última parte de este capítulo se explican y mencionan algunos proyectos/trabajos previos que tienen objetivos de características similares al de este proyecto, es decir trabajar sobre las inundaciones, en estos casos buscando anticipar las mismas y no sus consecuencias. Estos brindaron una primer idea de como funcionan este tipo de proyectos pese a tener ambos caracteristicas completamente diferentes considerando que uno es un proyecto de la NASA y el otro es un proyecto de grado de la Facultad de Ingeniería.

2 Desastres Naturales

Se entiende por desastres naturales aquellos desastres ocasionados como consecuencia de un fenómeno natural [1].

Por otra parte, es importante comprender el concepto de la Oficina de las Naciones Unidas (ONU) [2] para la Reducción del Riesgo de Desastres [108]. Ésta explica que no toda ocurrencia de un fenómeno natural desemboca en un desastre natural, sino que un fenómeno para ser catalogado como desastre, implica la falta de planificación y prevención así como la presencia de la acción del hombre; a modo de ejemplo: una erupción volcánica es considerada un desastre natural si hay una población afectada, de lo contrario, no.

Un fenómeno natural con potencial destructivo sobre un área poblada es considerado una amenaza natural. Mientras que la vulnerabilidad (producto de la falta de prevención, planificación o de carencias económicas para afrontar situaciones de riesgo) es una condición previa, que queda expuesta al ocurrir un desastre, siendo esta un indicador de la exposición del capital, de la capacidad de tolerancia y de la fortaleza al daño por parte de personas, infraestructuras, comunidades e incluso estados [1]. Es decir, cuanto más alto es el nivel de vulnerabilidad, mayor es la exposición del capital y menor es la capacidad de tolerancia y de fortaleza al daño.

A grandes rasgos, un desastre es el resultado de una relación directa entre una amenaza (fenómeno natural en este caso) y la vulnerabilidad (exposición y falta de capacidad de tolerancia y fortaleza ante el daño) ante la misma ($desastre = amenaza * vulnerabilidad$). Se detallan y definen conceptos en la sección 4 sobre Gestión de Riesgos.

Con el fin de reflejar con claridad la relación anterior, existen casos que dejan en evidencia que no hay un vínculo directo entre que suceda un fenómeno natural y la ocurrencia de un desastre como consecuencia de este. Un ejemplo conocido es el del Río Nilo donde las inundaciones anuales son necesarias para el desarrollo de los asentamientos humanos que se encuentran ubicados en las riberas del mismo [1]. A su vez, una inundación de cierta intensidad no afecta a todas las sociedades de igual manera, un ejemplo sencillo de esto es nuestro país, donde una posible inundación en Montevideo no genera los mismos daños que ocasiona en Paysandú u otros departamentos más susceptibles ubicados sobre el Río Negro y Río Uruguay.

Sobre la frecuencia de las inundaciones, además de tener altos niveles de ocurrencia, parece incrementarse tanto en incidencia como en intensidad en los últimos años (ver figura 1). Según estimaciones de la CEPAL (Comisión Económica para América Latina y el Caribe) [3], en la región más de 150 millones de habitantes se han visto afectados por los desastres y más de 310.000 han fallecido como consecuencia de los mismos, además de haberse generado 30 millones de damnificados directos; con un monto total de daños acumulados estimado en unos 213.000 millones de dólares (datos válido para el año 2000), tomando como punto de partida el terremoto de Managua, capital de Nicaragua, en 1972 [1].

Considerando la realidad de Uruguay, un país en vías de desarrollo, es importante hacer mención que las estadísticas mundiales reflejan que los desastres causan daños desde el punto vista social más significativos e incluso irreversibles en los países en desarrollo.

Dentro de América Látina y el Caribe se han logrado avances en el campo de la gestión de riesgos pero aún así grandes segmentos de la población viven en condiciones de alta vulnerabilidad (sección 4 refiere a la Gestión de Riesgos). Hay que tener en cuenta en este contexto que los países de la región se encuentran dispuestos en zonas propensas a la incidencia de fenómenos naturales, tanto hidrometeorológicos como geológicos; trayendo consigo los daños mencionados anteriormente y el deterioro ambiental en la región [109]. Ver figura 2.

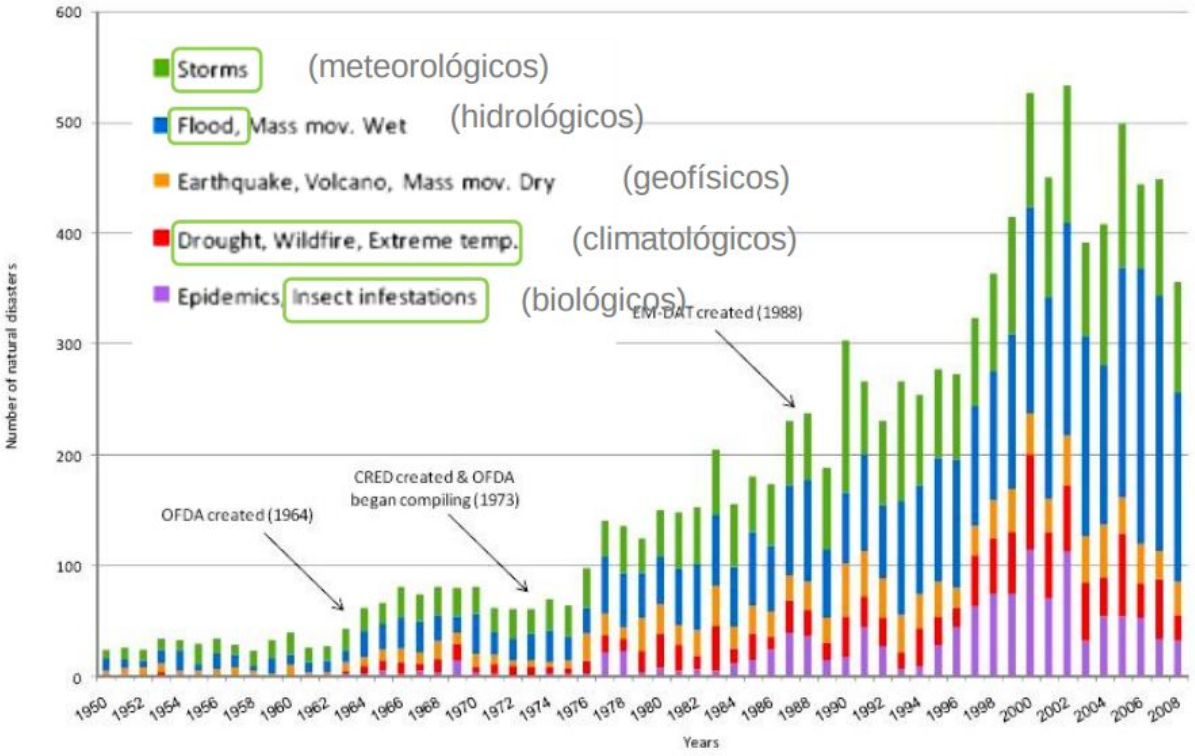


Figura 1. Gráfico de número de desastres por año [109].

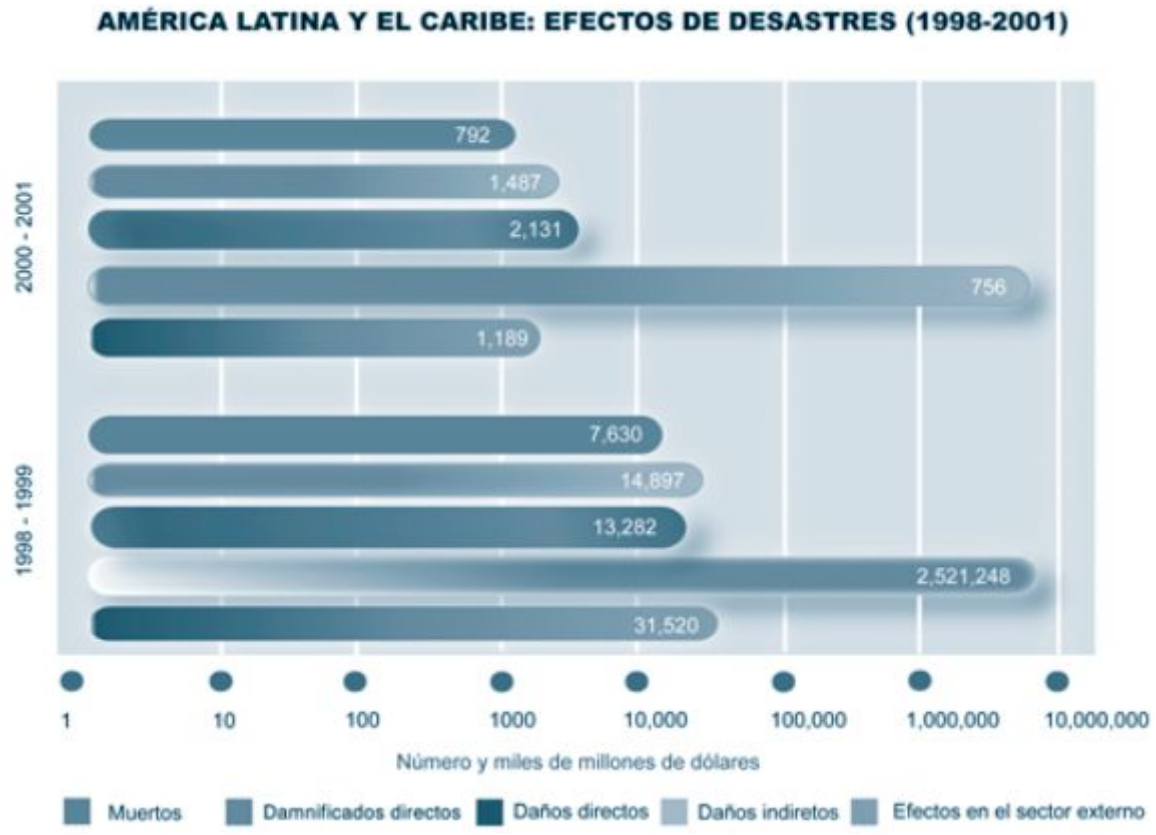


Figura 2. Gráfico efecto de desastres en América Latina [109].

Finalmente, con el fin de reducir los efectos a largo plazo de los desastres, las acciones deben enfocarse en dos frentes paralelos según ONU [4]:

- En la previsión de un evento desastroso, teniendo en cuenta la asignación de recursos para la prevención y mitigación del impacto.
- Por otra parte, en asegurar que lo invertido a la reconstrucción, sea con miras a una reducción de la vulnerabilidad, garantizando un desarrollo sustentable.

EM-DAT International disaster database, centro de Investigación Epistemológico de Desastres

El Centro de Investigación Epistemológico de Desastres es un departamento de de la Universidad Católica de Louvain (UCL) que se ubica en Bruselas, Bélgica [5]. El centro ha estado activo por más de 30 años en el campo de los desastres internacionales y temas relacionados con problemas en la salud [6].

Dicho centro brinda públicamente su base de datos de desastres llamado EM-DAT, la cual es una base de datos global que almacena desastres naturales. Se han registrado un total de 21,000 desastres en todo el mundo desde el año 1900 hasta el presente [6].

EM-DAT tiene como objetivos principales ayudar a la acción humanitaria tanto a nivel nacional como internacional, a racionalizar acerca de la toma de decisiones para la preparación de posibles desastres y proporcionar una base de datos objetiva para la evaluación de la vulnerabilidad y establecimiento de prioridades [6].

DesInventar (Inventario de desastres)

Con el fin de lograr juntar en un único lugar de manera normalizada los datos sobre desastres naturales en América Latina, el Caribe, Asia y África, la Red de Estudios Sociales en Prevención de Desastres de América Latina desarrolló el Sistema de Inventario de los Efectos de los Desastres (DesInventar) en 1994. Está contiene todos los desastres grandes y medianos ocurridos en la mayoría de los países de América Latina en los últimos 40 años [7].

En la actualidad cuenta con bases de datos nacionales de desastres de 30 países [8], incluyendo Uruguay que en los últimos años y por pedido del SINAE (Sistema Nacional de Emergencia, se describe en la sección 4.1) fue incluido en la misma.

2.1 Desastres Naturales en Uruguay

En nuestro país, desde finales de la década del 50 cuando sucedió el primer gran desastre natural, las inundaciones de 1959, el registro se ha llevado a cabo con mayor seriedad, registrándose números elevados tanto en pérdidas humanas como en población afectada.

Considerando los registros comprendidos entre 1967 y 2014, los desastres que más personas han afectado son las inundaciones, con 224263 afectados y las tormentas, con

200 personas; mientras que los desastres que más pérdidas humanas ocasionaron son las inundaciones con 23 pérdidas y las tormentas y temperaturas extremas con 11 cada una. En cuanto al apartado económico, las pérdidas por sequías han sido un total de 250 millones de dólares, 89 millones por inundaciones y 25 millones por tormentas [9].

En balance general, las inundaciones han sido el desastre que más ha afectado a la región y a nuestro país.

A modo de comprender mejor cómo han afectados los desastres a nivel nacional, se muestran los siguientes datos de distintos desastres que ocasionaron las mayores pérdidas en termino de personas afectadas y también pérdidas económicas, comprendidos entre 1967 y 2014 [9]:

Los 10 desastres con mayor número de afectados:

Tipo de desastre	**Fecha**	**Nro. de afectados**
Inundación	1967	38063
Inundación	4/5/1967	119200
Inundación	08/1986	18500
Inundación	12/06/1992	4700
Tormenta	08/09/1993	2000
Inundación	11/04/1998	9300
Inundación	16/05/2000	5000
Inundación	06/2001	5000
Inundación	01/04/2002	2500
Inundación	21/11/2009	22000

Los 10 desastres con mayor número de fallecidos:

Tipo de desastre	**Fecha**	**Nro. de fallecidos**
Inundación	1967	8
Tormenta	21/12/1997	1
Inundación	11/04/1998	1
Tormenta	29/06/1999	1
Temperatura extrema	07/2000	7

Tormenta	15/03/2002	2
Tormenta	23/08/2005	7
Inundación	04/05/2007	2
Inundación	21//11/2009	12
Temperatura extrema	07/2010	4

Los 5 desastres con mayor impacto económico

Tipo de desastre	**Fecha**	**Daño en dólares**
Inundación	1967	39.000.000
Inundación	11/04/1998	5.000.000
Sequía	06/1999	250.000.000
Tormenta	15/03/2002	25.000.000
Inundación	04/05/2007	45.000.000

3 Inundaciones

Según el glosario internacional de hidrología (OMM/UNESCO, 1974), una inundación es el "aumento del nivel del agua por encima del nivel normal del cauce". En este contexto, se entiende por nivel normal a la elevación de la superficie del agua tal que no provoca daños en sus alrededores [10].

Por otra parte, SINAE define una inundación de la siguiente manera: "*Una inundación es el avance de las aguas sobre zonas que habitualmente están secas. Puede producirse por el desborde de ríos, lagos y embalses a causa de lluvias torrenciales o por la rotura de diques o presas*" [11].

3.1 Clasificación de las inundaciones

Debido a que las inundaciones poseen diferentes características, como son la duración, la intensidad, el origen o causa, el impacto, etc., es común ver distintas clasificaciones de las mismas.

A continuación se mencionan las clasificaciones más habituales [12][13].

Según origen

Inundaciones pluviales: Son ocasionadas por lluvias muy intensas que sobresaturan el terreno y provocan que el agua se acumule sobre el mismo, pudiendo permanecer desde

horas hasta semanas antes de evaporarse por completo y permitir que el terreno recupere su capacidad de inflación.

Inundaciones fluviales: Este tipo de inundación se produce cuando los arroyos, ríos, etc. aumentan su volumen a tal punto que el agua se desborda y afecta áreas adyacentes que normalmente no tienen agua.

Inundaciones costeras: Los huracanes, ciclones, maremotos, etc. en las zonas costeras provocan que el agua aumente su oleaje a tal magnitud que el agua golpea con fuerza la costa cubriendo zonas muy extensas y provocando grandes daños.

Inundaciones por catástrofes: Este último tipo de inundaciones es el producto de siniestros en infraestructuras que manejan grandes volúmenes de agua, como por ejemplo presas, represas, etc.

Según impacto

Es el criterio de clasificación más utilizado a la hora de realizar relevos de informes y estudios de inundaciones históricas. Los tres tipos que se consideran son los siguientes:

Inundación o avenida ordinaria: son inundaciones que producen pocos daños y en las que el agua no supera los límites del cauce del río.

Inundación extraordinaria: en esta categoría el agua sí se desborda respecto al cauce habitual y provoca algunos daños de importancia variable, como la inundación de bajos, parkings, algunos cortes de luz, etc.

Inundación catastrófica: en estas inundaciones los daños son muy importantes, algún edificio o puente puede resultar parcial o completamente destruido, los cortes de servicios básicos son más extensos.

3.2 Principales causas de las inundaciones

Cuando se habla de inundaciones, generalmente se asocia a las mismas con un resultado producto de grandes lluvias. Esto en parte es así, ya que estas representan la gran mayoría de los grandes casos de inundaciones ocurridos.

Más allá de esta realidad, los causantes de las inundaciones son variados y no son producto únicamente de eventos naturales, sino que también han sido provocadas por la mano del hombre sobre el entorno.

Aca se mencionan las principales causas de inundaciones[14]:

Lluvias

La lluvia es la principal causa de las inundaciones. El exceso de lluvias provoca que el agua fluya sobre la tierra, contribuyendo así a las inundaciones. En particular, las lluvias causantes de estos eventos son lluvias con mucha intensidad y con una duración prolongada.

En algunos casos las lluvias de corta duración también pueden ocasionar inundación, esto depende de algunas variables del entorno como pueden ser la humedad del suelo, la distribución y la cantidad de lluvia.

Desborde de rios

Tanto ríos como arroyos pueden desbordar sus orillas. Esto ocurre cuando el río tiene más agua corriente arriba de lo que tiene normalmente y fluye hacia zonas bajas (zonas de llanura aluvial generalmente). Esto tiene como efecto la ocurrencia de descargas repentinas de agua en las tierras aledañas, provocando inundaciones.

Esto también puede ocurrir en presas que se ven desbordadas en su capacidad provocando un desborde de agua y generando inundaciones en las planicies vecinas.

Las inundaciones causadas por desbordamientos de ríos tienen la peligrosidad de que pueden provocar barridas enteras en su trayectoria río abajo.

Inundaciones costeras

Este tipo de inundaciones suceden como consecuencia de largas tormentas o fenómenos naturales como huracanes, ciclones, tsunamis que provocan que el “cuerpo” del agua se precipite hacia la tierra. Este tipo de desbordamientos son altamente destructivos, pudiendo destruir estructuras enteras como son casas, puentes, carreteras, etc.

Rotura de represas

Este tipo de estructuras son realizadas con la finalidad de contener al agua que fluye desde terrenos más elevados. Con el fin de aprovechar la fuerza con la que el agua se dirige hacia la presa, las mismas son utilizadas para producir energía eléctrica. Las inundaciones en estos casos suceden cuando la represa se encuentra debilitada en sus muros y terminan rompiéndose por exceso de cantidad de agua, superando la capacidad de la represa.

Es común que algunas veces con la intención de evitar el rompimiento de la represa, se libere agua deliberadamente para alivianar la presión.

Derretimiento de los glaciares

En regiones frías, durante el verano las grandes cantidades de hielo y nieve que se producen y acumulan durante el invierno comienzan a derretirse, produciendo así grandes movimientos de agua en tierra. Esto también ocurre en regiones con presencia de montañas con nieve en la cima cuando en el verano ocurre un proceso similar de derretimiento por aumento de temperatura atmosférica.

Urbanización

La urbanización de las ciudades, en general implica que el suelo se cubra con una capa de concreto o asfalto. Si esto se combina con malos drenajes o con drenajes obstruidos por basura, es muy factible que el agua de lluvia se acumule provocando inundaciones.

Es también normal que las ciudades crezcan y se construyan viviendas en las zonas costeras. Muchas veces, estas nuevas construcciones no están preparadas para soportar inundaciones producto del aumento del nivel del agua.

Deforestación

La tala de árboles provoca que la cobertura vegetal del suelo se pierda y que el agua de lluvias arrastren la tierra. Con el paso del tiempo, esta tierra puede obstruir arroyos y/o ríos cambiando el flujo del agua o provocando que la misma se acumule en puntos que no están preparados para soportarlo.

3.3 Efectos negativos de las inundaciones

Como se mencionó en la sección anterior, una inundación en una zona urbanizada puede tener un gran impacto en la sociedad. Esta sección busca clasificar los efectos de una inundación según su perdurabilidad en el tiempo.

En el corto plazo

Si bien la pérdida de vidas por inundación no es el escenario más común en nuestro país, pues por lo general ocurren de forma paulatina producto de las lluvias prolongadas, en aquellos lugares con riesgos de tsunamis o tormentas muy fuertes, este escenario no es tan irreal. La pérdida de vidas y las personas heridas son, sin lugar a dudas, la peor consecuencia posible de una inundación.

Las personas que tienen que caminar en zonas inundadas pueden sufrir lesiones debido a elementos tales como árboles y líneas de electricidad caídas o cualquier elemento cortante que se encuentre en el agua. En consecuencia, una lesión que puede parecer leve, como lo es un corte en una pierna, puede más adelante resultar letal debido a infecciones y el acceso limitado a hospitales que se puede dar en desastres de estas características.

A nivel material, muchas veces, el nivel del agua aumenta con tal rapidez que las personas se ven obligadas a evacuar sus hogares, pudiéndose llevar consigo un porcentaje muy pequeño de sus pertenencias. En la ciudad de Paysandú, en Julio de 2014, a pesar de que el nivel del Río Uruguay creció abruptamente, mucha gente convivió con más de un metro de agua en sus casas por miedo a que sus pertenencias sean robadas afectando así la calidad de su salud y en términos generales, calidad de vida de los damnificados [110].

En el largo plazo

Por lo general, los daños en infraestructura posterior a una inundación implican corte de calles, días sin luz eléctrica en las partes afectadas por la inundación, etc. Esto supone un gran impacto en la economía del país para poder superar la situación y volver a funcionar con normalidad.

Las inundaciones pueden también causar daños en zonas agrícolas lo cual repercute directamente sobre el suministro de alimentos y en el caso particular del Uruguay, sobre una de las actividades económicas más desarrolladas y esenciales. Además, muchas veces

el agua desplaza a roedores y serpientes hacia lugares en los que no es frecuente encontrarlos; esto supone peligro tanto para la salud los humanos como para los animales. En Diciembre del 2015, tras las inundaciones en Entre Ríos, Argentina, se registraron muchas personas con mordeduras de serpientes, picaduras de alacranes, etc[111].

No solo está en riesgo la salud por las lesiones y las mordeduras o picaduras de animales peligrosos, sino que es muy común que las fuentes de agua potable se contaminen con materiales tóxicos provocando enfermedades. Además, los agentes que se transmiten por el agua son amenazas para aquellos que tienen que caminar constantemente por las aguas profundas causadas por las inundaciones. Por ejemplo, es común que aparezcan casos de leptospirosis post inundación [15].

3.4 Inundaciones a nivel nacional

Las características de penillanura (suaves pendientes) del suelo de Uruguay hace que los cursos de agua sean propensos a no generar crecidas violentas además de resultar bastante predecibles considerando parámetros como el volumen de las lluvias y su intensidad en conjunto con el factor tiempo. Gracias a esto y a la experiencia (de inundaciones pasadas) adquirida en las zonas más propensas a ser afectadas por inundaciones, es posible en muchos casos realizar evacuaciones preventivas, logrando poner a resguardo a la potencial población afectada y sus bienes, así como ganar tiempo para tomar otras medidas preventivas [16].

El registro histórico con mayor número de personas evacuadas es del año 1959 cuando ocurrieron las inundaciones más importantes en la historia de nuestro país, en aquel entonces la cantidad de evacuados ascendió hasta casi 45.000 personas. Como consecuencia de la probabilidad y el temor de que la Represa de Rincón del Bonete colapsara (ubicada en el centro del país en Río Negro), se decidió evacuar a poblaciones enteras cercanas. Estas inundaciones duraron un mes entero entre fines de marzo y fines de abril, generando así un desastre a nivel nacional [17].

La misma perjudicó al país en su totalidad, teniendo consecuencias catastróficas como la caída de redes telefónicas enteras, alteraciones en el sistema de transporte y ocasionando serios problemas con el abastecimiento de energía eléctrica. Esto último fue debido a la situación particularmente grave que enfrentó la Represa del Rincón de Bonete (ver figura 3), la cual es clave en la generación de energía eléctrica en Uruguay, siendo sobrepasada por las aguas y quedando fuera de servicio. Como agravante, la UTE, que por aquel entonces se encontraba en obras construyendo la Represa de Baygorria, con el fin de no perjudicar dichas obras (y debido a que los pronósticos en aquel entonces no eran como los de ahora) decidió restringir la evacuación del lago artificial de Rincón del Bonete, el cual enfrentaba una situación alarmante y la apertura de sus compuertas hubiese disminuido la presión y el riesgo. Debido a esto, ante una situación que se tornó insostenible fue necesario dinamitar un terraplén de contención de agua, buscando evitar así que la Central Hidroeléctrica se desbordara y cubriera a las turbinas generadoras de electricidad, además de evitar así la ruptura del dique central. Como consecuencia de esta acción se logró evitar la ruptura del dique pero no lo primero, pero se tuvo que realizar una evacuación inmediata de la

población de Paso de los Toros y zonas cercanas. Por otra parte, al fracasar en el primer objetivo, las turbinas se detuvieron al ser cubiertas por agua y esto tuvo como consecuencia que una buena parte del país quedará privada de energía eléctrica [17].

Recién el 27 de abril de ese año, la represa comenzó a resurgir entre las aguas. No obstante, es importante destacar que la zona de la represa no fue la única perjudicada, si no que todo el país se vio afectado por las inundaciones, especialmente en la zona litoral de Salto, Paysandú y Fray Bentos, así como también Rivera que quedó únicamente comunicado mediante ferrocarril [17].

Hasta la actualidad, no se ha repetido una inundación de estas magnitudes, sin embargo si han habido otras grandes inundaciones, con importantes números de evacuados en el país, destacándose los nueve meses comprendidos entre 1997 y 1998 donde casi todo el litoral del Río Uruguay permaneció bajo las aguas como consecuencia de las precipitaciones ocasionadas por el fenómeno ENOS (Niño Oscilación Sur) [17].

Figura 3. La sala de máquinas anegada de la Represa de Rincón del Bonete [17].

Por otra parte, desde el comienzo del siglo XXI, se ha estado trabajando en medidas preventivas y de mitigación en zonas que suelen ser afectadas por las inundaciones como lo son los departamentos de Salto, Paysandú y Soriano, en los cuales gracias a préstamos internacionales se logró construir viviendas en zonas no inundables, mitigando en parte así grandes problemas para la población local. Así mismo, en departamentos como Artigas, Rivera, Cerro Largo y Durazno, se trabajó de igual forma, elaborando mapas de riesgo mediante los cuales se otorgase un enfoque serio a la relocalización de muchos habitantes que estaban en situación constante de ser “evacuados potenciales” [16].

Finalmente, como dato respecto al impacto reciente de las inundaciones en nuestro país, es remarcable el hecho de se registraron más de 67.000 personas evacuadas en la década pasada debido a este tipo de eventos, siendo por lo tanto el más frecuente y con mayor impacto en nuestro país [11].

4 Gestión de Riesgos

Antes de estudiar la gestión de riesgos a nivel nacional, resulta necesario comprender y manejar algunos conceptos previos de la misma y tener en consideración que la gestión de riesgos no acepta una única definición sino que hay múltiples definiciones, cada una aceptada por diferentes gobiernos y organizaciones.

Conceptos previos

Resulta pertinente remarcar previamente la definición de riesgo en el contexto de la gestión de riesgos como la probabilidad de que ocurra un evento de consecuencias socioeconómicas o ambientales en un espacio de lugar y tiempo específico con cierta duración. Dicho riesgo está vinculado a desastres y se da donde haya una sociedad afectada de forma importante por el impacto de algún evento físico de diverso origen, como pueden ser las inundaciones. Este riesgo trae consigo además la irrupción en las vidas cotidianas de la población afectada.
Finalmente, en este contexto, se considera al riesgo como el resultado de una interacción entre la **amenaza** y la **vulnerabilidad** [18] y se puede expresar de la siguiente forma:

$$Riesgo = Amenaza * Vulnerabilidad.$$

Se entiende por amenaza como la probabilidad de que un fenómeno de origen natural, socionatural o antrópico se presente con cierta intensidad en un sitio específico y dentro de un período de tiempo, con potencial de producir efectos adversos sobre las personas, los bienes y el medio ambiente [18].

La vulnerabilidad, por su parte, expresa las características y circunstancias de una comunidad, sistema o bien, que los vuelven susceptibles o predispuestos a padecer los efectos dañinos de una amenaza, tanto de origen natural como del hombre [18].

Es así que entonces, se puede concluir que el riesgo se presenta como una condición que existe en la medida que no se ve mitigada o reducida mediante intervención humana o por cambios ambientales en el entorno físico. Dicho riesgo existe si se da la existencia de

población humana, infraestructura y producción que se encuentra expuesta al impacto de los posibles eventos físicos [18].

De esto se desprende que los eventos físicos y la vulnerabilidad son lo que se conocen como **factores de riesgo** ya que la presencia de estos elementos son las que generan la existencia de un riesgo. Finalmente, remarcar que no todo riesgo puede conllevar un nivel de daños y pérdidas que puedan considerarse de **desastre**; este último concepto implica la presencia de daños y pérdidas que perjudican de manera sustancial el funcionamiento normal de la sociedad [112].

Definición de Gestión de Riesgos

Un primer acercamiento a lo que es la gestión de riesgos, aplicada a la temática de este proyecto, es decir que la misma utiliza estrategias para disminuir la vulnerabilidad y promover acciones de conservación, desarrollo, mitigación y prevención frente a desastres tanto naturales como antrópicos [18].

La gestión de riesgos implica la implementación de un conjunto de medidas que permitan conocer y medir todos los elementos vinculados a los riesgos con el único objetivo de poder enfrentarlos para minimizarlos, o en el mejor de los casos, anularlos. Los métodos a utilizar varían según el contexto y marco en el cual se realice su gestión [18].

Conceptualizando, a continuación se presentan algunas de las definiciones de lo que es gestión de riesgos (por más que algunas puedan tener mucho en común), ofreciendo así diferentes puntos de vista sobre la misma.

La ONU define a la gestión de riesgos como la organización, planificación y aplicación de medidas para la preparación, respuesta y recuperación ante desastres. Resaltando que aún así, la gestión de riesgos no garantiza la eliminación o la evasión completa de una amenaza, centrándose más bien en la creación e implementación de planes de preparación como de disminución del impacto de los desastres, buscando una recuperación más ágil. [19].

La Cruz Roja[113] explica a la gestión de riesgos como *“la organización y la gestión de recursos y responsabilidades para abordar todos los aspectos humanitarios de las emergencias, en particular la preparación, la respuesta y la recuperación a los desastres, a fin de reducir sus efectos”* [114]. Ver figura 4.

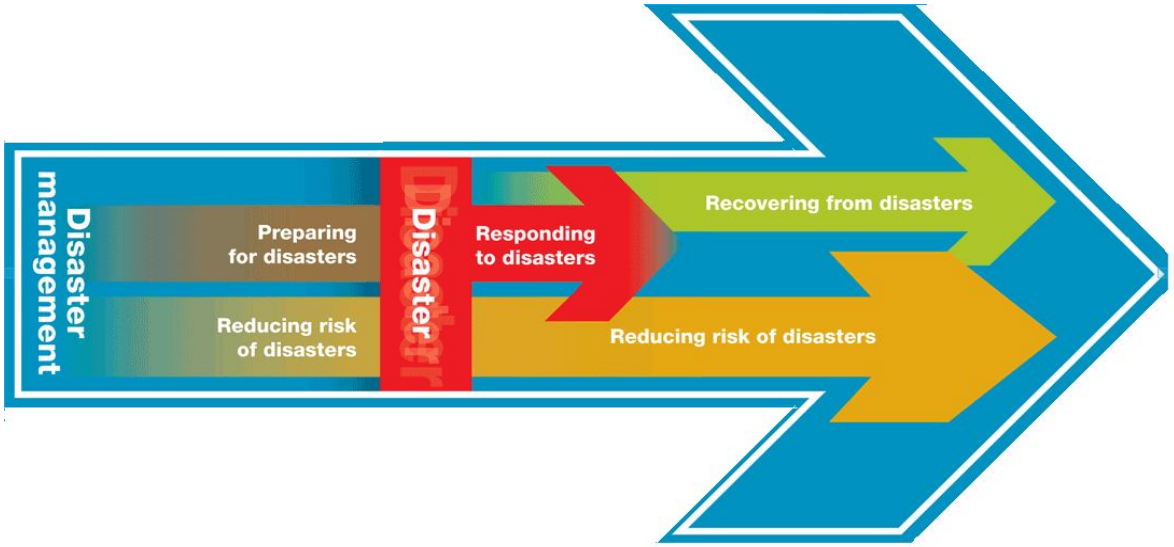


Figura 4. Flujo de la gestión de riesgos según la Cruz Roja [114].

La OMS (Organización Mundial de la Salud)[115] sugiere separar en dos términos a la gestión de riesgos: “prevención de desastres” y “gestión de emergencias”. Esta separación tiene su sustento en el hecho de que por definición los desastres no pueden ser gestionados si no que se pueden prevenir y posteriormente gestionar la emergencia. La gestión de emergencias se encarga de todas las actividades, desde la preparación hasta una vez ocurrido el evento y el proceso de rehabilitación incluido. A su vez la OMS hace mención a los términos de “recuperación” como el proceso que va desde el impacto del evento hasta la reconstrucción; “reducción de riesgos” como la etapa entre la reconstrucción y la preparación, y finalmente define tanto “alivio” que es todo aquello que se encuentra del lado derecho del ciclo mientras, como “desarrollo” que es todo lo que se encuentra en el lado izquierdo [116]. Ver figura 5.

Drabek define la gestión de riesgos como la disciplina y la profesión de aplicar la ciencia, la tecnología, la planificación y la gestión para hacer frente a eventos extremos que pueden dañar o matar a un gran número de personas, dañando la vida de la comunidad [20].

En el contexto nacional, el SINAE (Sistema Nacional de Emergencias, ente público encargado de prevenir y actuar en situaciones de desastre, del cual se profundiza más adelante), define a la gestión de riesgos como un proceso coordinado entre varias instituciones para reducir, prevenir, responder y apoyar a la rehabilitación y recuperación frente a eventuales emergencias y desastres, en el marco de un desarrollo sostenible [21].

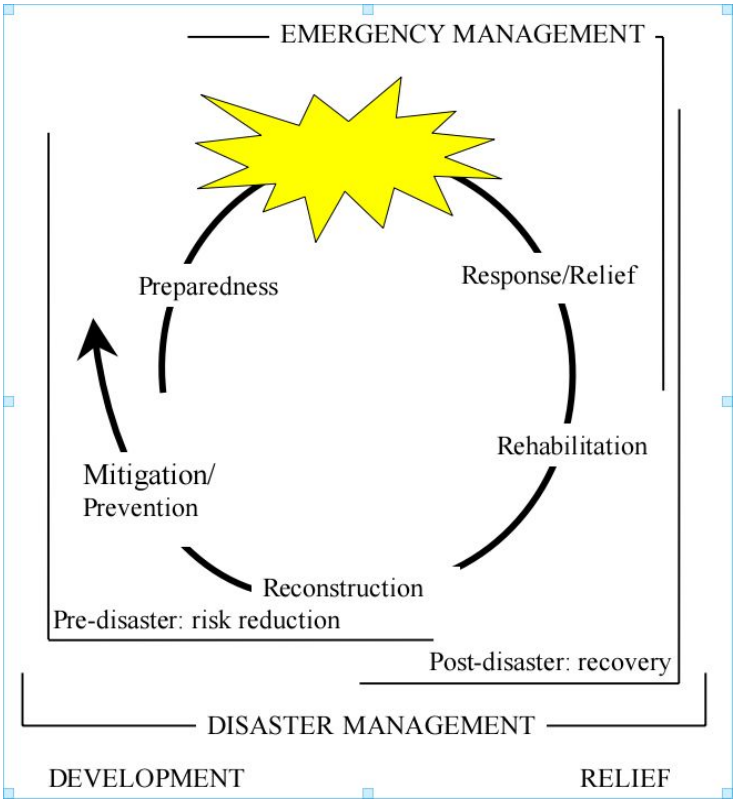


Figura 5. Modelo de gestión de riesgos de la OMS [116].

Concepción actual de la Gestión de Riesgos

La gestión integral de riesgos para desastres naturales se compone de cuatro fases elementales: mitigación, preparación, respuesta y recuperación. Una buena gestión de riesgos emplea todas estas fases y se definen de la siguiente forma [18], ver figura 6:

- **Mitigación**: consiste en la reducción y/o eliminación de la probabilidad de un fenómeno y sus posibles consecuencias. El objetivo es que el impacto de un fenómeno sea lo menos dañino posible.

- **Preparación**: consiste en equipar a los potenciales afectados por el desastre y a posibles entidades de ayuda con equipamiento, herramientas y recursos necesarios para incrementar las probabilidades de supervivencia y eliminar/minimizar pérdidas.

- **Respuesta**: consiste en la ejecución de ciertas acciones que tienen como objetivo la reducción y eliminación de los daños ocasionados por la ocurrencia del desastre (o que está ocurriendo aún), con la finalidad de reducir pérdidas tanto humanas como financieras (a causa de interrupción de procesos o daños en infraestructura).

- **Recuperación**: consiste en lograr que las víctimas vuelvan a sus vidas cotidianas normales después de las consecuencias del desastre natural. Esta fase suele empezar de forma inmediata una vez concluida la fase de respuesta y su duración es absolutamente variable, pudiendo incluso llevar años.

Figura 6. Ciclo de la gestión de riesgos [22].

4.1 Gestión de Riesgos en Uruguay

A modo de enmarcar la realidad de Uruguay, es importante hacer mención al hecho de que el área y estudio de gestión de riesgos es relativamente nueva en América Latina (menos de dos décadas). Esto implica que en los últimos años, lo que antiguamente era concebir a los desastres como un hecho a causa de la voluntad de Dios o la naturaleza, actualmente se lo concibe de forma más integral, donde la gestión del riesgo se entiende como la probabilidad de que existan pérdidas y daños vinculados al suceso de un evento físico y la vulnerabilidad de la sociedad ante estos.

En Uruguay, tanto los riesgos de emergencias como la presencia de desastres (tanto de origen natural o antrópico) han sido históricamente desconsiderados y visto como algo ajeno al país. De hecho los desastres ocurridos durante el siglo pasado, como fueron las epidemias, las sequías y las inundaciones, permanecían fuera de la percepción social. Esto comenzó a cambiar a mediados de la primera década del siglo actual, siendo la creación del SINAE una muestra de esto.

SINAE y marco institucional en Uruguay

Creado por la ley 18.621 “de Creación del Sistema Nacional de Emergencias”
“La instancia específica y permanente de coordinación de las instituciones públicas para la gestión integral del riesgo de desastres en Uruguay es el Sistema Nacional de Emergencias (SINAE). Su objetivo es proteger a las personas, los bienes de significación y el medio ambiente de fenómenos adversos que deriven, o puedan derivar, en situaciones de emergencia o desastre, generando las condiciones para un desarrollo sostenible.” [23]

El SINAE es la unidad encargada de la coordinación de las instituciones públicas y de la gestión integral de riesgos en Uruguay. Este tiene como meta asegurar la protección de la población, de las infraestructuras y del ambiente contra eventos adversos que puedan tener como consecuencia un desastre o situación de emergencia. El SINAE no está representado por un cuerpo específico sino que su accionar se ejecuta mediante todas las acciones del Estado vinculadas a la gestión de riesgos en sus diferentes fases, siendo esta una tarea interinstitucional como resultado de la legislación de un espacio de articulación en las instituciones existentes [112].

La instancia superior de coordinación y decisión del SINAE se encuentra en el Poder Ejecutivo mientras que la Dirección Nacional se encuentra en la Presidencia de la República [112].

En cuanto a lo que es nivel departamental, se encuentran los Comités Departamentales de Emergencia (CDE), estos trabajan en conformidad con las políticas del SINAE y se encargan de formular y aplicar de este modo las políticas y estrategias a nivel local. Sus instancias operativas departamentales son los Centros Coordinadores de Emergencias Departamentales (CECOED) [112].

Finalmente, con la ley de “Descentralización política y participación ciudadana” surge un nuevo nivel de gobierno que son los municipios y en conjunto con estos surge a la par un nuevo nivel de gestión de riesgos. Según esta ley, uno de los cometidos de los municipios es: *“Adoptar las medidas urgentes necesarias en el marco de sus facultades, coordinando y colaborando con las autoridades nacionales respectivas, en caso de accidentes, incendios, inundaciones y demás catástrofes naturales comunicándolas de inmediato al Intendente, estando a lo que éste disponga”* [112].

Manejo de la respuesta

En primera instancia, la atención de las emergencias (“respuesta”) es ejecutada de forma descentralizada por parte de las instancias de coordinación con el aporte de sus capacidades y recursos. Según el grado de riesgo y los requerimientos específicos de la respuesta así como por la intensidad y cobertura del impacto, es que se determinan el nivel de descentralización y la coordinación del SINAE para la atención durante la ocurrencia del fenómeno de riesgo [112]. Ver figuras 7 y 8.

NIVEL DE RESPUESTA	COORDINACIÓN DE LA RESPUESTA
Atención Primaria	Autoridad Idónea en el evento
Respuesta Departamental	Comité Departamental de Emergencias
Respuesta Nacional	Comando de Respuesta Nacional
Situación de Desastre	Poder Ejecutivo

Figura 7. Cuadro de nivel de coordinación de Respuestas frente a emergencias y desastres [112].

Figura 8. Avance de la respuesta y autoridad responsable [112].

Atención Primaria

A este nivel, la respuesta está a cargo de la autoridad idónea según el evento y está coordinada por la misma. Un ejemplo sencillo del mismo es el caso de un incendio, éste es responsabilidad de la Dirección Nacional de Bomberos, pudiendo ser apoyado por los municipios en su primera fase. Es necesario pasar a una respuesta departamental cuándo [112]:

- Esta primera respuesta se ve desbordada o inminentemente se verá superada.
- Es necesario el trabajo articulado de más de una institución durante la respuesta o en instancias.
- El impacto es significativo sobre personas, infraestructura y medio ambiente.
- El evento puede aumentar en su magnitud.

Respuesta Departamental

Esta respuesta es manejada por el CDE, donde cada uno cuenta con reglas de funcionamiento y tomas de decisiones preestablecidas. El mismo se debe encargar de definir qué acciones tomar según los planes de contingencia y encargarse de tomar todas las decisiones requeridas durante la respuesta. Además debe manejar el flujo de la información así como la comunicación con los medios y las recomendaciones a la población [112]

En cada CECOED existen Planes de Emergencia actualizados, estos tienen en cuenta las características de cada departamento y las principales amenazas en cada uno, ponderarlos con la relación entre su probabilidad de ocurrencia y su impacto. Contando además las amenazas principales con Planes de Contingencia [112].

Respuesta Nacional

Esta manejado por el Comando de Respuesta Nacional (CRN). Este toma la dirección nacional de la respuesta y entra en Sesión Permanente una vez que [112]:

- Las capacidades departamentales se ven superadas
- Es necesario una coordinación interinstitucional más compleja
- El impacto cubre más de un departamento o no esta geolocalizado
- Hay un alto riesgo de que aumento de las magnitud del evento.
- El impacto dañino del desastre es de gravedad en población, infraestructura y ambiente.

Por otra parte el CNR tiene la potestad de la toma central de decisiones para mejorar el desempeño del SINAE en su conjunto así como coordinar el flujo de información, comunicación con los medios y recomendaciones a la población.

Situación de Desastre

El Presidente de la República por medio del Poder Ejecutivo en mutuo acuerdo con los ministerios pertinentes es el encargado de declarar esta situación. La Situación de Desastre es establecida por la magnitud y el impacto del evento. Ante una situación que requiera este tipo de respuesta, es requerido la existencia de un nivel de organización, coordinación y asignación de recursos a gran escala con accionar inmediato, requiriendo colaboración de instituciones nacionales e incluso por parte de la comunidad internacional. En esta situación, la dirección de respuesta es asumida por parte del Poder Ejecutivo [112].

INUMET

Creado por la ley 19158 de “Creación del Instituto Uruguayo de Meteorología”, el cual tiene dentro de sus principales fines [25]:

“A) Prestar los servicios públicos meteorológicos y climatológicos, consistentes en observar, registrar y predecir el tiempo y el clima en el territorio nacional y zonas oceánicas adyacentes y otros espacios de interés, de acuerdo a los convenios aplicables, con el objeto de contribuir a la seguridad de las personas y bienes y al desarrollo sostenible de la sociedad.

B) Coordinar las actividades meteorológicas de cualquier naturaleza en el país.

C) Representar a la República Oriental del Uruguay ante los organismos internacionales en materia de meteorología, así como cumplir con las obligaciones asumidas por el país ante los mismos.”

El INUMET tiene como misión institucional contribuir a la seguridad de la población y sus bienes, ayudando además así, al desarrollo sostenible de la sociedad. A su vez, dentro de su visión, se destaca la contribución a la gestión de riesgos meteorológicos y climáticos [26].

Advertencias Meteorologicas de INUMET

Diariamente, el INUMET publica dos pronósticos del tiempo por día así como reportes cuando se cree conveniente. Como parte de estos boletines y con el objetivo de mejorar la

discriminación del grado de peligrosidad de estos fenómenos, se establecieron cuatro niveles definidos por colores [118]. Ver figura 9

- Verde: No existe ningún riesgo meteorológico.
- Amarillo: No existe riesgo meteorológico para la población en general aunque sí para alguna actividad concreta. Es una llamada de atención para los usuarios sobre la predicción meteorológica en vigor. En ciertos casos donde la peligrosidad rebasa ciertos umbrales, ya en esta alerta se pueden emitir “Advertencias sobre Eventos Meteorológicos Adversos”. Es en estos casos que el SINAE activa los protocolos de preparación y respuesta establecidos.
- Naranja: Existe un riesgo meteorológico importante (fenómenos meteorológicos no habituales y con cierto grado de peligro para las actividades usuales).
- Rojo: El riesgo meteorológico es extremo (fenómenos meteorológicos no habituales, de intensidad excepcional y con un nivel de riesgo para la población muy alto).

Cuando se da una advertencia, la misma contiene cuatro datos básicos [118]:

- QUÉ: Indica el tipo de fenómeno y su intensidad (descargas eléctricas, viento fuerte, lluvias intensas, ola de frío, etc.).
- POR QUÉ: indica las causas (frente, depresión, inestabilidad del aire húmedo, etc.).
- CUÁNDO: indica el momento de ocurrencia (desde y hasta qué momento tiene vigencia la advertencia).
- DÓNDE: indica los departamentos afectados.
- Además en caso de que sea pertinente y se trate de un evento aún no ocurrido, se indica la probabilidad de ocurrencia.

Con el objetivo de una mejor comunicación ante potenciales desastres y situaciones de riesgo, equipos técnicos de INUMET y del SINAE establecieron el “Protocolo de Comunicación Durante las Fases de Advertencia y Respuesta a Eventos Meteorológicos Adversos” (ver figura 10), definiendo en este la coordinación y comunicación durante la existencia de alertas o durante la ocurrencia de un desastre meteorológico. Asegurando así una comunicación adecuada con la finalidad de garantizar que se brinde información fiable y útil a la hora de tomar decisiones, tanto a nivel de entes encargados de tomar decisiones como para la población [118].

nivel de riesgo → ↓ fenómeno	nivel verde	nivel amarillo	nivel naranja	nivel rojo
Tormenta	No es necesaria especial atención	**"tormentas intensas"** Tormentas con lluvias abundantes y/o granizo Inferior a 2 cm de diámetro, y/o viento con rachas entre 60 y 75km/h	**"tormentas fuertes"** Tormentas con lluvias copiosas y/o granizo entre 2 y 3 cm de diámetro, y/o viento con rachas entre 75 y 120km/h	**"tormentas severas"** Tormentas con lluvias copiosas y/o granizo superior a 3 cm de diámetro, y/o viento con rachas mayores a 120km/h
Lluvia		**"precipitaciones abundantes"** de 20 a 50 mm en 6 h, o de 50 a 100 mm en 24 h	**"precipitaciones copiosas"** de 50 a 100 mm en 6 h, o de 100 a 200 mm en 24 h	**"precipitaciones torrenciales"** mayores a 100 mm en 6 h, o mayores a 200 mm en 24 h

Figura 9. Cuadro de rango de advertencias meteorológicas [118].

Figura 10. Imagen ilustrativa del flujo de comunicación entre el SINAE, INUMET y direcciones relevantes [118].

5 Data Warehouse

Para cumplir con los objetivos del proyecto se propone trasladar y almacenar la información en una base de datos, más precisamente en un data warehouse.

Data Warehouse (DW), o almacén de datos, es una colección de datos organizados enfocados a una organización o tarea específica, que proporciona información con el fin de ser analizada y de esta forma, asistir la toma de decisiones. Suelen ser un banco centralizado de información proveniente de distintas fuentes, que engloba la totalidad de la organización o proyecto.

Bill Inmon, uno de los principales autores y prácticamente un creador del término Data Warehouse junto con Ralph Kimball, presentó una definición de almacén de datos haciendo énfasis en propiedades de sus datos.

Inmon [27] propone como necesarias las siguientes propiedades para los datos alojados en un almacén, distinguiendo los siguientes atributos fundamentales:

- **No volátil:** los datos no se pueden modificar ni perder
- **Orientado a temas:** que exista vinculación entre los datos de manera de que éstos tengan sentido para describir una temática en cuestión
- **Variante:** se pueden agregar datos continuamente y las alteraciones realizadas sobre los mismos quedan registradas de manera de que éstos describan su evolución en el tiempo
- **Integrado:** los datos del almacén deben representar la totalidad de la información manejada por la organización o proyecto en cuestión

En particular, para este proyecto se consideraron como propiedades elementales del DW: orientado a temas, variante e integrado.

Por otro lado, Kimball [28] define a los almacenes de datos en función de su principal rol, definiéndose como el conjunto de toda la información referente a una organización, presentada de manera tal de facilitar su consulta para posteriores análisis.

Siendo estas las principales visiones sobre los almacenes de datos, se notó que éstos no están vinculados a ninguna forma de implementación específica. Si bien cada autor presenta diferentes enfoques para realizarlo, quedan a decisión de los implementadores los detalles de su diseño, mientras respeten su objetivo.

6 Sistemas de Información Geográfica

“Un Sistema de Información Geográfica (SIG) es la combinación de cinco componentes: personas especializadas, datos descriptivos y espaciales, métodos analíticos, hardware y software; organizados para analizar, manipular, procesar, almacenar, generar y visualizar todo tipo de información referenciada geográficamente” [29]

La figura 11 brinda de manera más clara y visible lo que es un SIG:

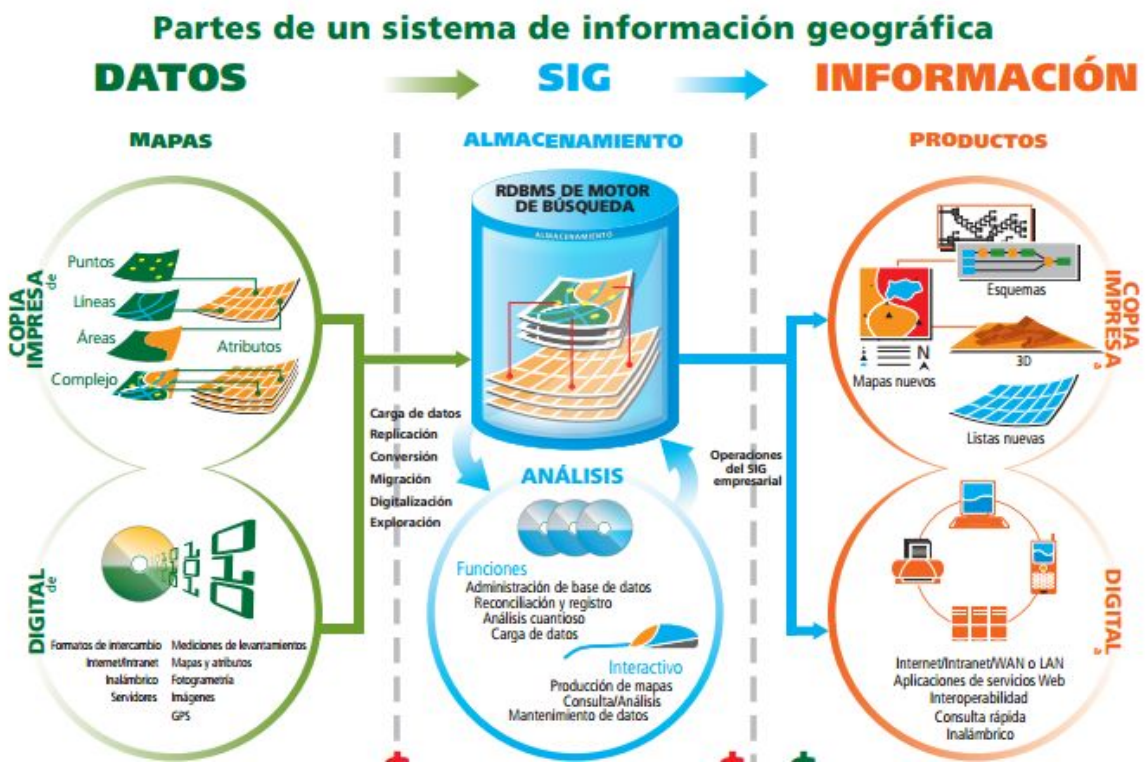


Figura 11. Modelo de un SIG [30].

Para poder entender con mayor claridad el diagrama anterior, hay que comprender el término “datos espaciales” como un conjunto de datos sin procesar que contienen un vínculo geográfico; es decir que un dato espacial está compuesto por una referencia (como puede ser una coordenada) a un lugar específico de la tierra y un conjunto de atributos (por ejemplo nombre, antigüedad, etc). Estos elementos son los que se almacenan en la base de datos de un SIG, comúnmente en **capas** y representan **puntos**, **líneas** o **polígonos**; mientras que los atributos son guardados como un atributo normal en cualquier base de datos relacional, en formato de tabla.

Las capas, como se ve en la figura, son colecciones de información geográfica, idealmente utilizadas para separar diferentes entidades o información de la realidad de manera de estructurar de mejor manera un SIG. En una capa se pueden tener elementos como los descritos anteriormentes y a su vez las capas se pueden activar/desactivar de una vista de

un SIG, así como también apilar de manera de poder visualizar más de una a la vez y superponer su información.

En el modelo, se puede observar en verde el origen de los datos espaciales anteriormente explicados, este suele ser mapas impresos y otra documentación no virtualizada que una vez digitalizada es posible relacionar con información ya anteriormente digitalizada, para así guardarla en base de datos geográficas y poder utilizarla en los SIG.

En azul, se observa el SIG en esencia, compuesto por la base de datos geográfica y el sistema que realiza las funciones, cálculos, análisis y trabaja comunicándose con la base de datos para procesar información y brindar los resultados que tenga como objetivo.

Finalmente en naranja se puede observar lo que comúnmente se puede denominar como clientes finales, esto es, las aplicaciones que consumen el SIG, utilizando y/o desplegando las respuestas que este da. Ejemplos de este puede ser un mapa en la web, un GPS, nuevos resultados impresos, etcétera.

El campo de los SIG es altamente pluridisciplinar, integrando ya diversas ciencias. Es ya tan amplio que podemos distinguir tres tendencias en la utilización de los SIG [119]:

Cartografía de alta precisión: utilizada principalmente para arquitectura e ingeniería, sistemas CAD.

Servidores de Mapas: a través de internet, aplicaciones que permiten gestionar facilidades ubicadas geográficamente, ordenación del territorio, por ejemplo, el popular Google Maps.

SIG para modelización ambiental: enlazado con herramientas de análisis de datos y modelización, con aplicaciones diversas en las ciencias de la tierra.

Dentro la anterior categorización, el último es el tipo de SIG con el que el presente proyecto se identifica. Aunque se utilice como base un servidor de mapas para dibujar sobre él, la combinación del data warehouse y el análisis predictivo (ver siguiente sección) que se le enlaza a fin de transmitirle la información, forman conjuntamente un SIG para modelización ambiental. [119]

Según Goodchild (1993) [31] un SIG destinado al análisis de datos y modelización ambiental debe incorporar un conjunto de herramientas para:

- Preprocesar grandes volúmenes de datos y prepararlos para su análisis
- Analizar los datos con el objeto de descubrir regularidades y desarrollar modelos
- Implementar estos modelos

También se considera como una funcionalidad posible en un SIG el permitir reorganizar los resultados en modo de tablas, gráficos o mapas de forma que sean útiles para el usuario

Como se explica con mayor amplitud en la sección 3.2 del informe principal, son justamente esas herramientas y la integración entre las mismas aquella que se busca lograr para

modelar geoespacialmente de la mejor manera posible una predicción del fenómeno de inundación.

Para lograrlo, se identifican tres tipos de modelados a realizarse. El modelado conceptual, acerca de cómo entendemos la realidad. Esto es la representación de la realidad que se quiere modelar, en términos de variables de tipo cualitativo (nombre de localidades, tipo de suelo, etc.) como de tipo cuantitativo (altura de ríos, precipitaciones, etc.)

En segundo lugar, un modelado lógico que determina cómo se muestrean y organizan las variables mencionadas para lograr una representación lo más adecuada posible. Existen básicamente dos grandes tipos de capas de información geográfica:

- **Raster:** se divide el espacio en un conjunto regular de celdas y a estas celdas se les asigna valores numéricos que representen algo que se quiera mostrar, o bien un identificador único si son de tipo cualitativo. Es normal ver que se utiliza también una escala de colores para colorear cada celda en base a su valor y dar una representación visual más amigable al usuario. (ver figura 12)

- **Vectorial:** cualquier entidad presente en el espacio geográfico se representará como un objeto de geometría euclidiana. Pueden ser clasificados por su dimensión en tres tipos: puntos, líneas y polígonos. Y combinar los mismos para generar otros tipos.

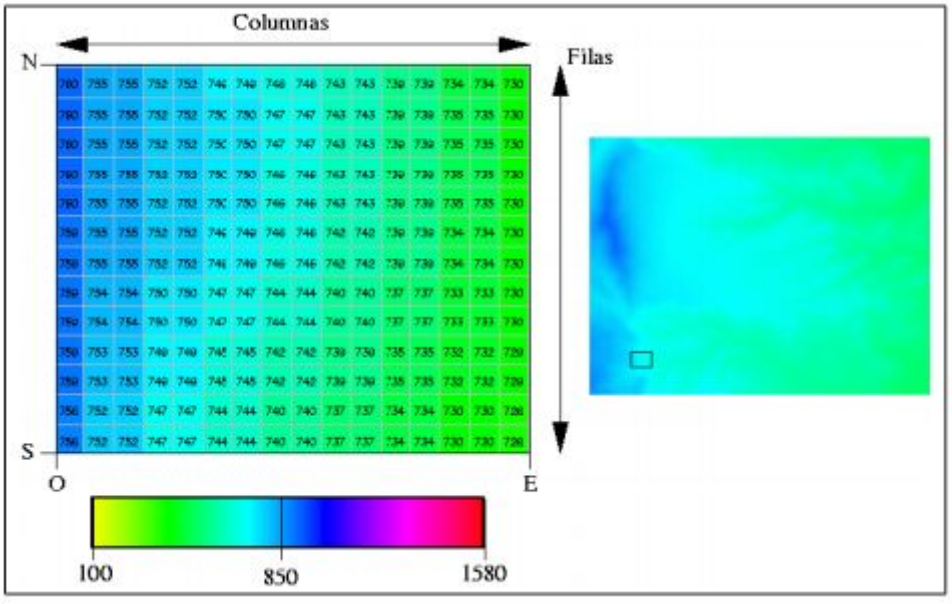


Figura 12. Representación raster de un SIG [30].

Un de las maneras más utilizadas para presentar información sobre un punto en una capa de un SIG son los *markers* o **marcadores**. Estos son ubicados en un punto específico del mapa y se visualizan de diferentes maneras dependiendo del sistema que soporte el SIG, y además muchos de ellos permiten que el programador defina un ícono gráfico personalizado para representarlo.

Estos markers, así como también otros elementos que pueden pertenecer a una capa, como los mencionados polígonos, rectas y puntos, y en general la mayoría de los elementos del mapa, tienen la posibilidad de responder a **eventos** generados por la interacción del usuario. Uno de los más utilizados e importantes es la reacción al evento de click de un usuario. Esto es, ante el evento de un click sobre la representación visual en el mapa de uno de éstos elementos, el desarrollador puede programar una reacción.

Finalmente, queremos mencionar ya que serán utilizados en el trabajo a la característica de ***marker clustering***, o agrupación de marcadores. [120] Ésta característica es una especie particular de marker y se utiliza para representar a la distancia un conjunto de markers mediante uno sólo que posee en su representación gráfica un número que representa cuántos markers contiene. A medida que el usuario se acerca con el zoom al marker, éste se puede dividir en otros sub-clusters que poseen menos dependiendo de cómo estén distribuidos geográficamente en la zona visible del mapa.

7 Análisis Predictivo y Aprendizaje Automático

7.1 Minería de datos

Se entiende por minería de datos o *data mining* como el proceso y las técnicas que permiten el estudio de grandes bases de datos tanto de manera automática como semi-automática con el fin de encontrar propiedades o patrones en los conjuntos de datos bajo cierto contexto [32].

Su surgimiento tiene el fin de lograr una mejora y una ayuda en la comprensión del contenido en un repositorio de datos. Para realizar esta tarea existen una amplia cantidad de procesos, algoritmos, prácticas estadísticas e incluso uso de inteligencia artificial, redes neuronales y técnicas como el aprendizaje automático [32].

En este contexto, la materia prima son los **datos**, los cuales se transforman en **información** una vez que reciben cierto valor (se les da un contexto y una utilidad); posteriormente estos datos (ahora información) pasan a integrar un modelo (en el cual la información es comparada o procesada), donde el resultado de su vinculación con el mismo, brindará un valor agregado que será referido como el **conocimiento** [32]. (Ver figura 13)

Figura 13. Pirámide de evolución de los datos [32].

A modo de ejemplo, en el marco del presente trabajo, los datos obtenidos por parte de los diferentes entes como la UTE, no son más que datos hasta que se les otorga una unidad de medición, pasaron una cierta limpieza y corrección de errores así y se les otorgó una finalidad que es la de ser utilizados para un modelo de predicción de consecuencias. Posteriormente esta información será derivada en conocimiento cuando sea procesada en el modelo y se tengan valores como resultado de la relación entre el modelo a emplear y esta información.

Aclarando que dentro del data mining no todos los casos son iguales, se puede decir que el procesamiento común se desglosa en los siguientes cuatro pasos [32]:

- Determinación de los objetivos
- Preprocesamiento de datos
- Determinación del modelo
- Análisis de los resultados

En la figura 14 se puede apreciar un gráfico comparativo que muestra la carga de trabajo en cada fase.

Figura 14. Carga de trabajo por fase [32].

7.2 Análisis Predictivo

El análisis predictivo es el campo en data mining [33] que pretende obtener información a partir de los datos con el objetivo de calcular las probabilidades y tendencias a futuro. Mediante el mismo, se pueden obtener conclusiones confiables sobre la ocurrencia y las características de eventos desconocidos sin importar su ubicación temporal. No tienen por qué ser necesariamente en el futuro (aunque sí funciona bajo el concepto de utilizar datos anteriores al evento que se busca predecir, en ese sentido si predice el futuro).

Esto se logra, mediante el uso de métodos estadísticos, matemáticos y reconocimiento de patrones.

Este campo no es nuevo y además tiene aplicaciones de todo tipo. Por ejemplo, ya en la década de los 40, en plena guerra, el predictor de Kerrison podía apuntar un misil hacia un vehículo aéreo en movimiento que se le fuera indicado, tomando sólamente algunos parámetros como la dirección de su trayectoria, velocidad, y ángulo observado desde el punto de lanzamiento, prediciendo la posición del mismo al momento de iniciar el lanzamiento del proyectil [34].

A modo de ilustrar la diversidad de aplicaciones, y también dar con un ejemplo más cercano a la temática del presente proyecto, nos interesa mencionar también la primera predicción numérica del estado del tiempo arrojada por la máquina ENIAC en la década de 1950. ENIAC fue de los primeros computadores de propósito general, desarrollado por la armada estadounidense con el fin de calcular tablas de balística. [121] Esto fue logrado por un grupo de meteorólogos liderados por John von Neumann. Von Neumann ya venía trabajando en el área de simulación atmosférica para estudiar la explosión de potenciales armamentos nucleares, y naturalmente pensó que ENIAC podía arrojar resultados sobre el estado del tiempo. Detrás de esto también hubieron intenciones bélicas ya que se veía en la predicción y control del estado del tiempo una poderosa arma. El cómputo inicial tomó 24 horas y arrojó una predicción del estado para las siguientes 24 horas. Este descubrimiento resultó importantísimo ya que significaba que con mayor capacidad de cómputo, la predicción del clima era posible [122].

En 1956, se resuelve por primera vez el problema del camino más corto, lo cual presentó grandes optimizaciones por ejemplo para el viaje aereo. Ya a partir de los 80 comienzan a aparecer las primeras soluciones comerciales para soporte de toma de decisiones empresariales utilizando técnicas de análisis predictivo. En los 90 con la aparición y masividad de las tiendas en línea, comenzó también una carrera para mejorar la experiencia del usuario y las sugerencias comerciales que se le presentaban, con el claro objetivo de aumentar también las ganancias de parte de las tiendas, basándose en acciones e información que el usuario dejaba al navegar en el sitio web [123].

Con el paso del tiempo, y a partir del nuevo siglo 21, esa información que se recababa ya sea de clientes, mediciones, y métricas en general dependiendo del área de aplicación, se pasó a denominar Big Data. Esto es, cada día, con la constante intervención de los usuarios, se generaban billones de bytes de datos día a día [123].

Así es que si bien el campo existe desde el surgimiento de las computadoras, en la actualidad están gozando de un auge por las siguientes razones [35]:

- Justamente el recién mencionado crecimiento de los volúmenes de información y un mayor interés general por almacenarlos.
- Evolución exponencial de la capacidad de cómputo y su accesibilidad general que permitió que tareas de cómputo de horas se realicen en de manera casi instantánea.

- Situación económica competitiva en todas las áreas, donde todos buscan sacar una diferencia y los resultados interpretados de los análisis sobre los datos de sus clientes presentan conocimiento agregado que sustenta la toma de decisiones.

Es así que hoy por hoy, entre tantas otras aplicaciones científicas como las anteriormente mencionadas, estas técnicas se utilizan también para, por ejemplo, detectar patrones de fraude en pagos electrónicos, optimización de campañas de márketing, reducción de riesgos comerciales y mejora de operaciones a todo nivel. Incluso ha prestado grandes aportes a la predicción de enfermedades crónicas y epidemias en el área de la salud. Se concluye que las aplicaciones del área son vastas mientras se tengan datos para alimentar el análisis. [35].

7.3 Modelado Predictivo y Aprendizaje Automático (Machine Learning)

En el análisis predictivo, el elemento central es el predictor, una variable que es medida para predecir el comportamiento en el futuro. No tiene porque haber un único predictor; de hecho, el uso de predictores múltiples conforman un modelo predictivo, los cuales bajo análisis, son utilizados para predecir con un nivel aceptable de fiabilidad [35].

Mediante el modelado predictivo se hace uso de estadísticas con el fin de obtener una predicción como resultado. Los mismos pueden utilizarse para cualquier tipo de evento desconocido sin importancia de su ubicación temporal [35].

Un modelo predictivo está compuesto por varios predictores, los cuales son factores variables que influyen (o pueden hacerlo) en el comportamiento de los resultados a futuro [35].

En estos modelos, se recopilan los datos para los indicadores de relevancia, se formula un modelo estadístico, luego se hacen las predicciones y se valida el modelo con los datos adicionales que se encuentren disponibles. Dicho de otra manera, los modelos describen en función de variables medibles aquellos fenómenos que se quieren analizar y ésta metodología es muy utilizada actualmente en la tecnología de la información [35].

Estos modelos pueden servir para una variedad de objetivos, pero no deben dirigir el análisis general que se quiere hacer, lo primero que se debe plantear y dejar en claro es qué problemática se apunta a solucionar mediante el análisis predictivo que se realiza [35].

Una de las técnicas más populares para realizar un modelado predictivo es el aprendizaje automático o *Machine Learning* que es un subcampo de la computación, mediante el cual se busca resolver problemas donde hay un conjunto de datos muestra, y se intenta predecir propiedades de un conjunto de datos desconocido. La idea es que el programa “aprenda” (y de allí el nombre) de las muestras de datos que tiene como entrada, de manera tal de inducir conocimiento (modelar el comportamiento) sobre sus propiedades y que dicha base de conocimiento permita hacer las predicciones mencionadas sobre otros conjuntos de datos [36].

A los conjuntos de datos de entrada, se le llaman *training sets*, o conjuntos de entrenamiento ya que son éstos los que asisten al aprendizaje del programa. Y usualmente se tiene otro conjunto de datos “nuevo” sin procesar sobre el cual se aplican el modelo del primer conjunto, a los cuales se les suele llamar *testing sets*, o conjuntos de prueba [36].

Sobre los datos de entrada, pueden ser tanto valores numéricos simples, como vectores de varias columnas de datos de tipos mixtos, a los que se les llama características o atributos.

Dentro del aprendizaje automático, nos interesa para utilizar en este proyecto, un tipo de metodología en particular, llamados de **aprendizaje supervisado** [37] donde existen disponibles ciertos conjuntos de datos de entrada que poseen atributos que luego se querrá predecir para futuros conjuntos de datos. Es posible aplicar estas técnicas debido a que se tienen datos ya etiquetados.

Una variable continua es una que, dentro de un valor mínimo y máximo, puede tomar infinitos valores en el intervalo, esto podría ser, por ejemplo, números decimales dentro de un rango definido. Por el contrario, una variable discreta puede tomar solamente ciertos valores dentro del máximo y del mínimo, con un codominio de tamaño finito [37].

A su vez, dentro del aprendizaje supervisado, se encuentran dos tipos de problemas a resolver. Los problemas de **regresión** [36] son aquellos donde la salida es el valor de una variable continua. Los análisis de regresión son procesos que utilizando diversas técnicas matemáticas, estiman relaciones entre las variables o atributos de los datos de entrada. La estimación de estas relaciones y el aprendizaje sobre éstas en base a los datos de entrada permiten a las técnicas empleadas poder ganar conocimiento entre cómo ciertas variables independientes (o predictoras) afectan el cambio de una variable dependiente que se desea predecir, de manera tal de poder arrojar dicha predicción como resultado final.

Por otro lado, también dentro del aprendizaje supervisado se encuentra a los problemas de **clasificación** [36], donde se busca clasificar utilizando etiquetas a cierto conjuntos de datos (observación). Para esto se toma un conjunto de datos previamente etiquetados de igual estructura a la observación. Notar que en este caso de clasificación, la salida, o sea una etiqueta sobre un conjunto de datos dado, resulta una variable dependiente de tipo discreta, esto es, habrá un número finito y definido de posibles etiquetas para clasificar el nuevo conjunto de datos.

Para aplicar estos algoritmos de aprendizaje sobre la información que se tiene, se debe usar esta información como semilla y así “alimentar” al algoritmo con un conjunto de datos de entrenamiento, que en nuestro caso es el data warehouse. Además, esta información debe estar previamente procesada de manera de cumplir los requerimientos mínimos de calidad de datos establecidos en la sección 5.3.1 del informe principal, dado que la mayoría de los algoritmos de aprendizaje tienen problemas cuando tratan con información faltante, o información con formato incorrecto o inesperado.

Algoritmos de predicción con Machine Learning

Generalmente, cuando se plantea un problema a resolverse mediante aprendizaje automático una de las partes más difíciles suele ser determinar el algoritmo correcto a

utilizar. Cada estimador se adapta mejor a diferentes tipos de problemas con diferentes conjuntos de datos [38].

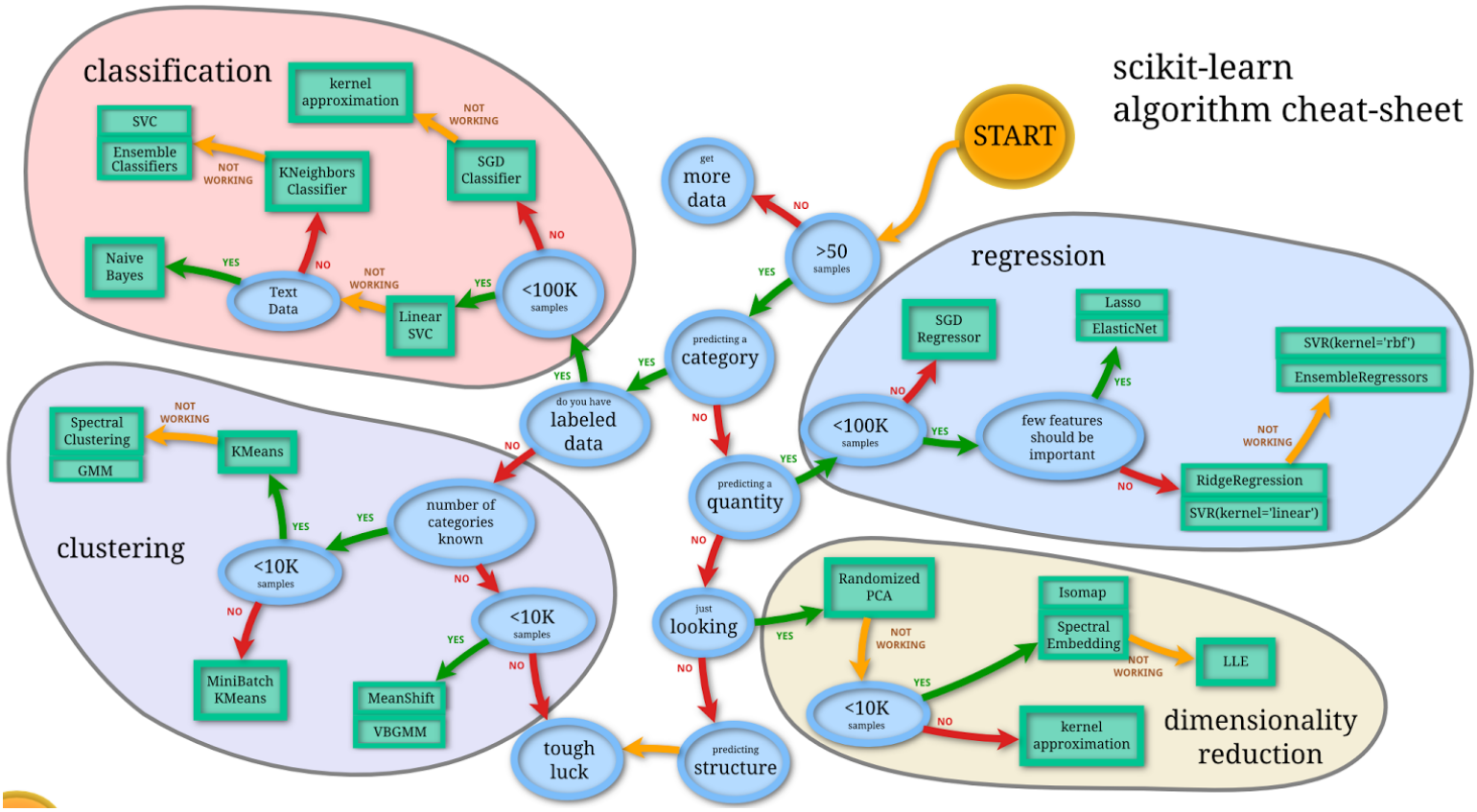


Figura 15. Mapa de selección de algoritmos [38].

La figura 15, busca servir como una guía a la hora de elegir qué algoritmos emplear según qué y cuántos disponibles hay disponibles, así como que se busca predecir. Como ya se mencionó, los algoritmos de regresión y clasificación pertenecen a la familia de algoritmos supervisados, mientra que los algoritmos de clustering (o agrupamiento) y dimensionality reduction (reducción de dimensionalidad) pertenecen a los algoritmos no supervisados. Los algoritmos no supervisados fueron dejados de lado, ya que la definición del mismo no se enmarca en el contexto de nuestro problema de estudio.

Clasificación: Support Vector Classification

El subgrupo de algoritmos SVC pertenecen a la familia de de algoritmos SVM (Support Vector Machines o Máquina de Vectores de Soporte en español). Este tipo de algoritmo presentan como ventaja que son efectivos en espacios con muchas dimensiones e incluso cuando el número de dimensiones es mayor al número de muestras. Su funcionamiento implica el uso de subconjuntos de puntos de entrenamiento en la función de decisión (estos son llamados vectores de soporte), convirtiéndolo de este forma un algoritmo eficiente en cuanto a memoria. En contrapartida, como aspectos negativos, este tipo de algoritmos presentan una rendimiento pobre en casos donde el número de variable de entradas es

mucho mayor al número de muestras. Finalmente, su salida no proporciona estimación de probabilidad directa [124].

Los algoritmos SVM presentan la versatilidad de poder aplicar distintos kernels (en la figura 16 se muestra la aplicación de diferentes kernel sobre un mismo conjunto de datos). Un kernel es la función de similaridad que se le proporciona al algoritmo, el mismo toma dos entradas y determina qué tan similares son las mismas, el mismo trabaja como un parámetro más de entrada en el algoritmo, estos son aplicados en algoritmos que trabajan con productos de puntos [124]

Los posibles kernels son los siguientes [124]:

- Linear: $\langle x, x' \rangle$.
- Polynomial: $(\gamma \langle x, x' \rangle + r)^d$
- Rbf: $\exp(-\gamma |x - x'|^2)$
- Sigmoid: $\tanh(\gamma \langle x, x' \rangle + r)$
- Ademas es posible definir nuevos kernels.

Los algoritmos SVC trabajan resolviendo el siguiente problema primal [125]:

Dados el vector de entrenamiento $x_i \in R^p,\ i = 1, .., n$ y un vector $y \in \{1, -1\}^n$,

$$min_{\omega.b,\zeta} \frac{1}{2}\omega^T\omega + C \sum_{i=1}^{n} \zeta i \quad \text{(C es mayor a 0 es la cota superior)}$$

$$sujeto\ a\ y_i(\omega^T \phi(x_i) + b) \geq 1 - \zeta i$$

$$\zeta i \geq 0,\ i = 1, ..., n$$

Y su dual es:

$$min_{\alpha} \frac{1}{2}\alpha^T Q\alpha - e^T\alpha$$

$$sujeto\ a\ y^T\alpha = 0$$

$$0 \leq \alpha_i \leq C,\ \ i = 1, ..., n$$

con *e* un vector de unos y *Q* es una matriz de $n * n$ definida por la función $Q_{ij} = y_i y_j K(x_i, x_j)$ donde $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ es el kernel. Y los vectores de entrenamiento (training) están implícitamente mapeados en un espacio dimensional mayor por la función ϕ.

Finalmente, la función de decisión es: $sgn(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho)$.

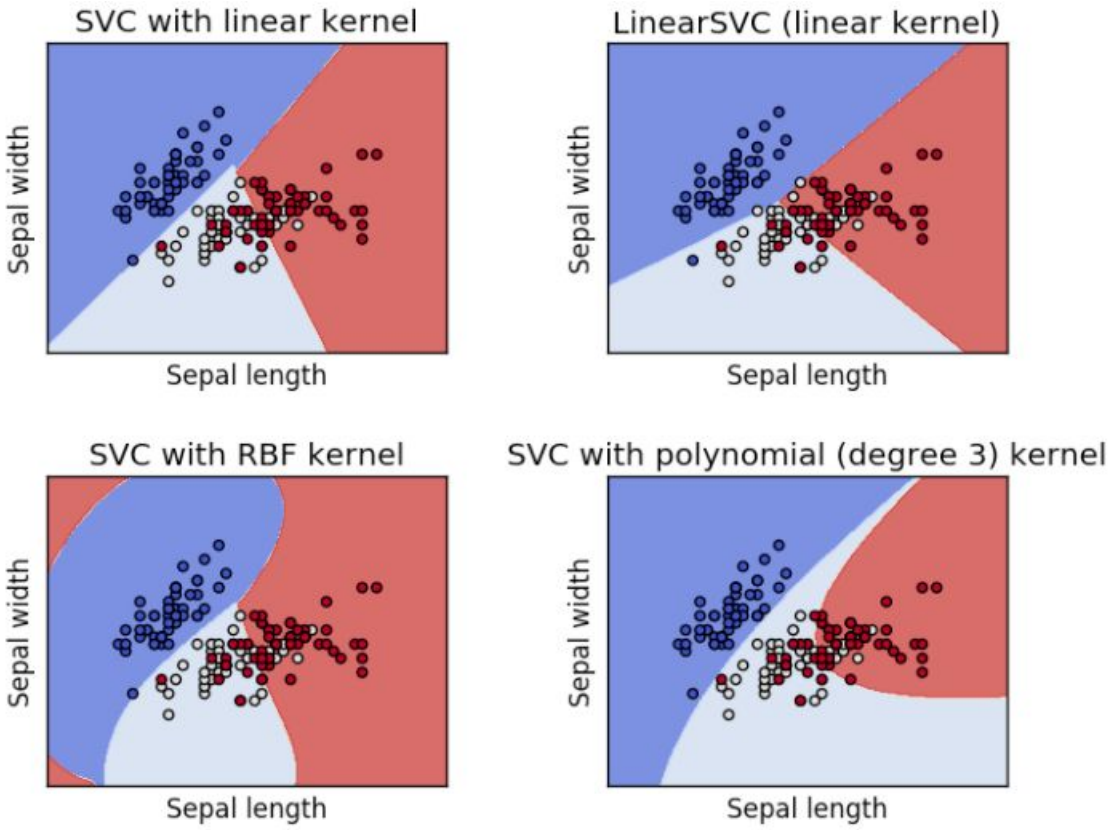


Figura 16. Visualización algoritmo SVC [125].

Cuando se trabaja con algoritmos SVC, es importante estudiar el conjunto de datos con el objetivo de determinar el valor de ciertos hiperparametros que son usados por el algoritmo, por ejemplo el valor C no debe ser muy grande en casos donde haya muchos datos con ruido. También se debe considerar si el conjunto de entrenamiento es balanceado o no y comunicarle al algoritmo en ese caso.

Clasificación: NuSVC

Este algoritmo es muy similar al anterior, con la variación de introducir un nuevo parámetro *v* que se encarga de controlar la cantidad de vectores de soporte y errores de entrenamiento. El mismo está comprendido en el rango $(0, 1]$, siendo este una cota superior para la fracción de la cantidad de errores de entrenamiento y una cota inferior en la fracción de vectores de soporte [126]

Clasificación: Naive Bayes

Los métodos Naive (ingenuo) Bayes son un conjunto de algoritmos que se basan en la aplicación del Teorema de Bayes[127], utilizando la asumpción de independencia entre todo par de variables de entrada (esa es la asumpcion 'ingenua') [128].

La aplicación del teorema se da del siguiente modo, dado una variable de clase y y una variable de entrada dependiente $x_1 ... x_n$, el teorema de Bayes afirma la siguiente relación [128]:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y)P(x_1, \ldots x_n \mid y)}{P(x_1, \ldots, x_n)}$$

Esta relación cuando se asume la independencia ingenua, se tiene para todo i

$$P(x_i|y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = P(x_i|y),$$

lo cual simplifica la ecuación a la forma:

$$P(y \mid x_1, \ldots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

Ademas, dado que $P(x_1, ..., x_n)$ es una constante dada como entrada, se puede aplicar la siguiente regla de clasificación:

$$P(y \mid x_1, \ldots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Finalmente se puede aplicar algún estimador de máximo para estimar $P(y)$ y $P(x_i|y)$, siendo el último la frecuencia relativa de la clase ***y*** en el conjunto de entrenamiento [128]

Naive Bayes funciona de manera efectiva como un algoritmo de clasificación pero su capacidad de predicción es mala. Un usos cotidiano de estos algoritmos es en el mailing para filtrar correos spam.

El algoritmo presenta algunas variaciones como lo son Gaussian Naive Bayes, Multinomial Naive Bayes y Bernoulli Naive Bayes [128] pero no fueron caso de estudio ya que, como se mencionó, este tipo de modelos no resultan útiles para la predicción.

Clasificación: Stochastic Gradient Descent

Más conocido como SGD, es un algoritmo simple y eficiente para realizar aprendizaje discriminatorio de clasificadores lineales, utilizando funciones de pérdidas convexas como SVM. Este algoritmo ha tenido éxito al aplicarse a gran escala y en problemas de aprendizaje automático esparsos, ejemplos de estos son los problemas de clasificación de texto y procesamiento de lenguaje natural [129]

Las ventajas que presenta este algoritmo es su eficiencia y su fácil implementación mientras que sus desventajas son que requiere un número grande de hiperparametros y es sensible a cuando la cantidad de variables de entrada escala [129].

Su implementación como algoritmo de clasificación, se describe visualmente en la figura 17.

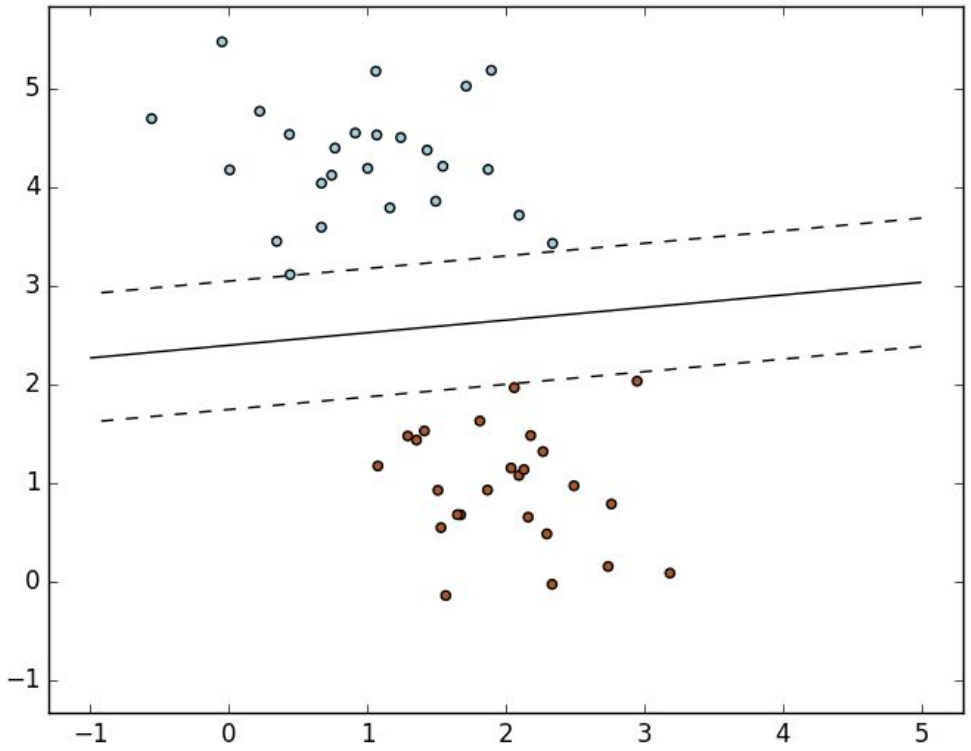


Figura 17. Gráfico modelo SGD [129].

En el algoritmo es posible elegir algunos parámetros importantes para su ejecución como los son las diferentes funciones de pérdida y las penalizaciones por clasificación [129].

Clasificación: Nearest Neighbors

La familia de los algoritmos de vecinos cercanos (nearest neighbors) es amplia y hay tanto algoritmos supervisados como no supervisados, así como tanto de clasificación como de regresión dentro de la categoría de supervisados. Se estudió el algoritmo de vecinos cercanos en la categoría de algoritmos de supervisados de clasificación ya que regresión en este tipo de algoritmos se aplica sobre salidas continuas.

La base de este tipo de métodos está en encontrar un número predefinido de muestras de entrenamiento cercanas al nuevo punto a determinar, prediciendo así su etiqueta. El número de vecino es posible definirlo (k-nearest neighbors) o determinarlo de manera variada según la densidad local (radius-based neighbors) [130].

Este tipo de algoritmos son conocidos como métodos no generalizadores de aprendizaje automático ya que funcionan recordando todos los datos de entrenamiento en estructuras de indexado rápido como pueden ser árboles) [130].

En el caso de clasificación, el resultado se computa a partir de una mayoría en los vecinos cercanos de cada punto, un punto de consulta se asigna a la clase de salida que tiene más representantes dentro de los vecinos más cercanos del punto [130].

Ambas implementaciones mencionadas anteriormente son aplicables en los algoritmos de clasificación (k-nearest y radius-based). El más usado generalmente es el enfoque de k-nearest, siendo k muy dependiente de los datos; como regla general para este algoritmo,

un k más alto tiende a eliminar los errores como resultado del ruido pero suele implicar que los límites entre cada clasificación son menos distintos [130].

Por otra parte, en casos donde los datos no están uniformemente muestreados, el enfoque basado en distancia puede resultar una mejor opción. Esto permite que los puntos en vecindarios esparsos (con pocos puntos o "vecinos") sean evaluados por menos vecinos pero manteniendo la similitud con esos pocos y no siendo afectado por un vecindario cercano con mayor cantidad de muestras de otra clase de la salida [130].

El algoritmo tiene dos variaciones (tanto para k-nearest como para radius-based) que están determinadas por un hiperparametro "weight" (peso), este puede tomar dos valores que son "uniform" y "distance". En el caso weight = uniform, cada vecino tiene la misma influencia sobre otro punto, sin importar la distancia; mientras que cuando weight = distance, se pondera la distancia y el peso es inversamente proporcional a la distancia entre el punto de consulta y el vecino. Ver figuras 18a y 18b.

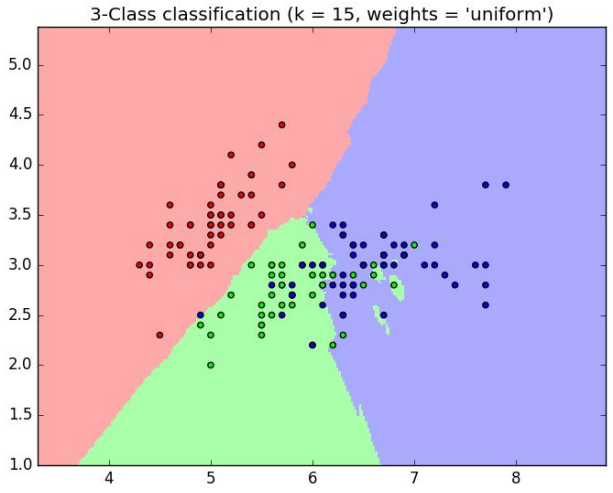


Figura 18a. Nearest neighbors con peso uniforme [130].

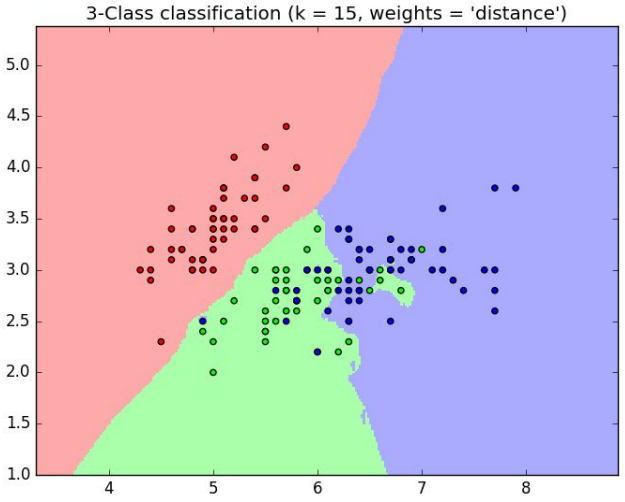


Figura 18b. Nearest neighbors con peso distancia[130].

Regresión: Linear Regression

Linear Regression [131] pertenece a la familia de modelos generales lineales, los cuales buscan la solución como una combinación lineal de las variables de entrada recibidas. Matemáticamente, si $y(w, x) = w_0 + w_1 x_1 + \ldots + w_p x_p$ es el resultado predecido, se nombran como coeficientes al vector w e interceptor al valor w_0. En otras palabras, es un problema ordinario de mínimos cuadrados lineales.

En el caso particular de Linear Regression, busca encontrar el vector de coeficientes w que minimice la suma de los residuos de los cuadrados entre los resultados reales observados (los datos etiquetados) y las respuestas predichas por la aproximación linear.

Matemáticamente e ilustrando lo que hace con una imagen, soluciona el problema como se ve en la figura 19.

$$\min_w ||Xw - y||_2^2$$

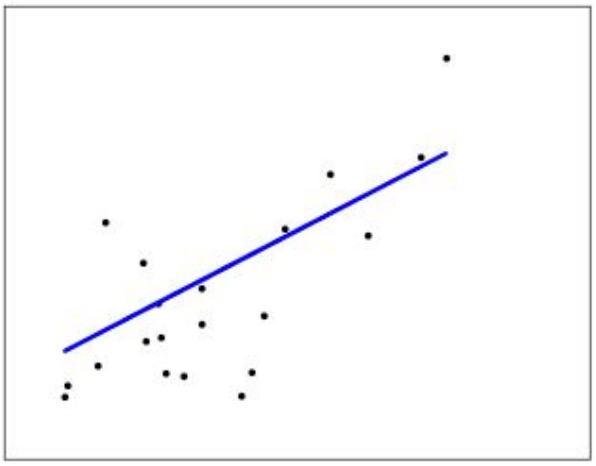


Figura 19. Linear Regression [132].

Regresión: Ridge

Ridge [133] también pertenece a la familia de modelos lineales generales, e intenta solucionar algunos de los problemas de los problemas de mínimos cuadrados ordinarios imponiendo una penalización al tamaño de los coeficientes. De esta manera, los coeficientes de Ridge buscan minimizar una versión penalizada de la suma cuadrada de residuos:

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2$$

Donde alfa ($\alpha \geq 0$) es un parámetro de complejidad que controla la cantidad de encogimiento: cuanto más grande es alfa, mayor encogimiento se impone y así, los coeficientes se vuelven más robustos a la colinealidad. Es así que alfa es un hiperparámetro del algoritmo, esto es, un parámetro que no forma parte de las variables de entrada y debe ser determinado con pruebas de validación buscando los mejores resultados.

Regresión: Lasso

Lasso [134] es un modelo lineal que estima coeficientes esparsos. A diferencia de Ridge, este algoritmo soporta coeficientes nulos, a modo de ejemplo, si se detecta que una de las variables de entrada utilizadas no mejora resultados se le puede asignar un coeficiente nulo asociado a esa variable, de manera que no afecte el resultado.
Resulta útil en algunos contextos dada su tendencia a preferir soluciones con menos parámetros de entrada, reduciendo efectivamente el número de variables sobre las cuales la solución es dependiente.

Expresado matemáticamente,

$$\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha||w||_1$$

Donde n es la cantidad de muestras de entrada y alfa, al igual que en Ridge, un hiperparámetro del algoritmo. Es entonces que Lasso soluciona un problema de minimización de cuadrados penalizados, pero sumado a $\alpha||w||$ donde $||w||$ es la norma 1 del vector de coeficientes w.

Regresión: Elastic Net

Este es un modelo de regresión lineal que es entrenado usando las normas 1 (utilizada en el hiperparámetro del algoritmo de Ridge) y norma 2 (utilizada en el hiperparámetro del algoritmo de Lasso) como regularizadores. [135] Esta combinación permite aprender un modelo esparzo donde sólo algunas de las variables de peso no valen cero como en Lasso, pero manteniendo las propiedades de la regularización de Ridge.

Para controlar la combinación convexa entre las normas 1 y 2, se utiliza un hiperparámetro llamado usualmente $l1\_ratio$.

Matemáticamente, Elastic-Net busca minimizar

$$\min_w \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

Donde n es la cantidad de muestras de entrenamiento, alfa α y ro ρ son hiperparámetros del algoritmo, a determinar.

Regresión: SVR (Support Vector Regression)

Dados los buenos resultados que brindaron las máquinas de vectores soporte en algoritmos de clasificación, se decidió probarlas en los casos de regresión, dando lugar a SVR. En este caso, la idea es seleccionar el hiperplano regresor que mejor se ajuste a nuestro conjunto de datos de entrenamiento, pero esta vez sin clases para separar como en los algoritmos de clasificación. Por lo tanto la idea se trata de considerar una distancia épsilon de modo que esperamos que todos los ejemplos se encuentren dentro de una banda o tubo (dependiendo de la dimensión de los datos de entrada) entorno al hiperplano.

No se ahondará en los fundamentos matemáticos detrás de SVR ya que son similares a los mencionados para SVC. Estos algoritmos también presentan variantes con diferentes funciones kernel, tal como los SVC [136].

7.4 Validación de Resultados

Con el fin de medir la calidad de un modelo predictivo, se recurre a un simple concepto, el método de retención. Se toman todos los datos referentes al pasado disponibles, de los cuales ya conocemos su resultado (si existió o no un evento de inundación) y lo particionamos en dos conjuntos: un conjunto de datos de entrenamiento del modelo, y otro conjunto de prueba. La idea de las pruebas es simple y radica en entrenar un modelo que utilice un método dado con el conjunto de datos de entrenamiento que se obtuvo, luego predecir resultados utilizando el modelo previamente entrenado y alimentándose con los datos que se separaron para prueba, lógicamente “enmascarando” el hecho de que ya se tienen los resultados reales asociados a los mismos. Para concluir la prueba, se compara el resultado de dicha predicción con aquellos resultados reales y comprobados que se tienen para cada elemento del conjunto de datos de prueba. Como resultado de esta comparación se obtienen métricas de error de cada uno de los métodos que serán detalladas más adelante dentro de esta sección. Dado que se quiere aproximar el resultado lo máximo posible a la realidad, se busca reducir dicho error al mínimo y por lo tanto se seleccionarán aquellos métodos que lo minimicen. Ver figura 20.

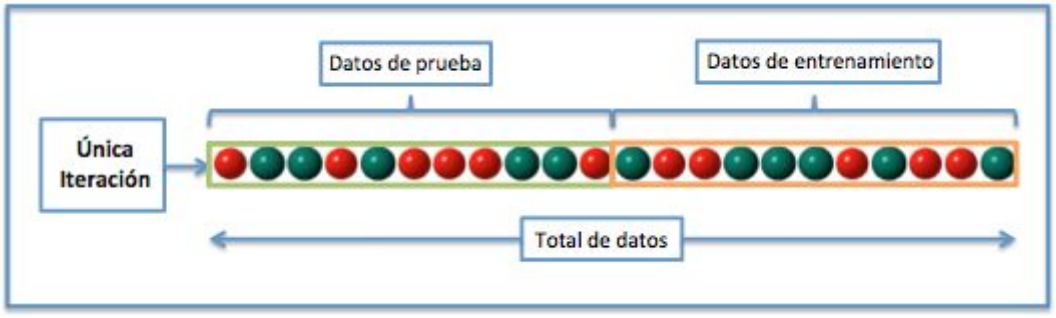


Figura 20. Método de retención [39].

Sin embargo, esta técnica de validación introduce una problemática y ésta es una dependencia entre los datos. Dado que los algoritmos de aprendizaje automático, como se describió en la sección 7.3, funcionan buscando relaciones entre los conjuntos de variables de entrada y sus resultados asociados, si uno siempre utiliza un mismo conjunto de datos, puede caer en un modelo que se entrenó de manera muy dependiente a ésta relación y probablemente no funcione adecuadamente con un nuevo conjunto de datos, que es justamente su fin.

Aquí se introduce el término de validación cruzada [39] que resultó vital para la selección final del método adecuado. La validación cruzada busca atacar la problemática expuesta en el párrafo anterior y es una técnica conocida para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba [40].

Existen tres tipos de validación cruzada que se desean destacar.

Validación cruzada en K iteraciones *(K-Fold)*

Se realizan K iteraciones, donde dentro de cada una, se separa al conjunto de datos total en K subconjuntos contiguos. Uno de estos subconjuntos hará el papel de datos de prueba, mientras el resto (K-1) conjuntos, hacen el lugar de datos de entrenamiento. Dentro de cada iteración se computa un resultado, y finalmente se computa un promedio con el fin de dar un resultado único de precisión, y asociado a él, la **desviación estándar** con respecto a dicho promedio. Posee la desventaja de ser lento computacionalmente, pero dado que separa el mismo conjunto total de datos es diferentes combinaciones de subconjuntos de entrenamiento/prueba, contrarresta la potencial dependencia a desarrollarse entre éstos subconjuntos a la hora de evaluar un método [39]. Ver figura 21.

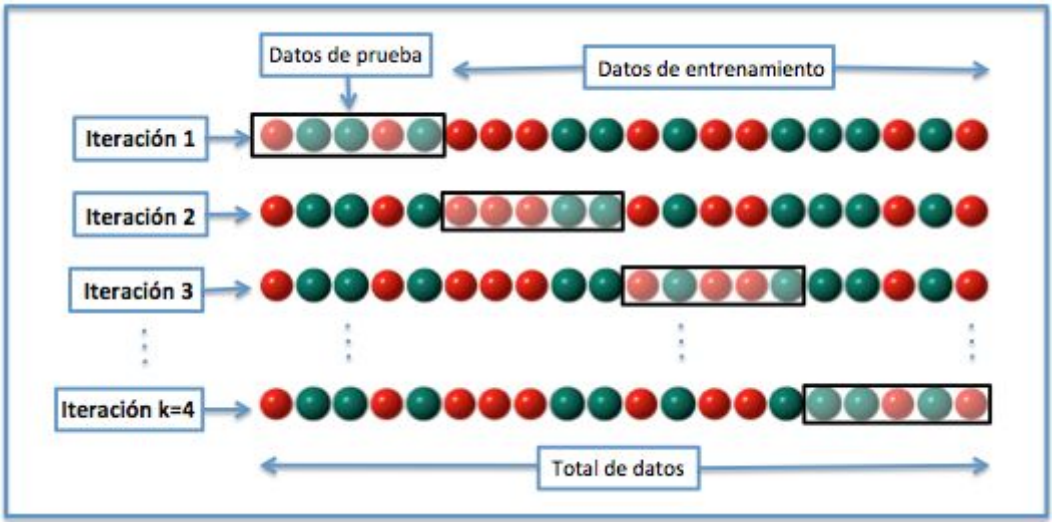


Figura 21. Validación K-Fold [39].

Aleatoria

En este caso los conjuntos de entrenamiento y prueba para cada iteración se conforman seleccionando elementos individuales de manera aleatoria.
A diferencia del tipo anterior, se tiene como ventaja que la división de conjuntos no depende del número de iteraciones, pero como contrapartida, puede suceder que algunos elementos no sean probados nunca, o que algunos lo sean más de una vez [39]. Ver figura 22.

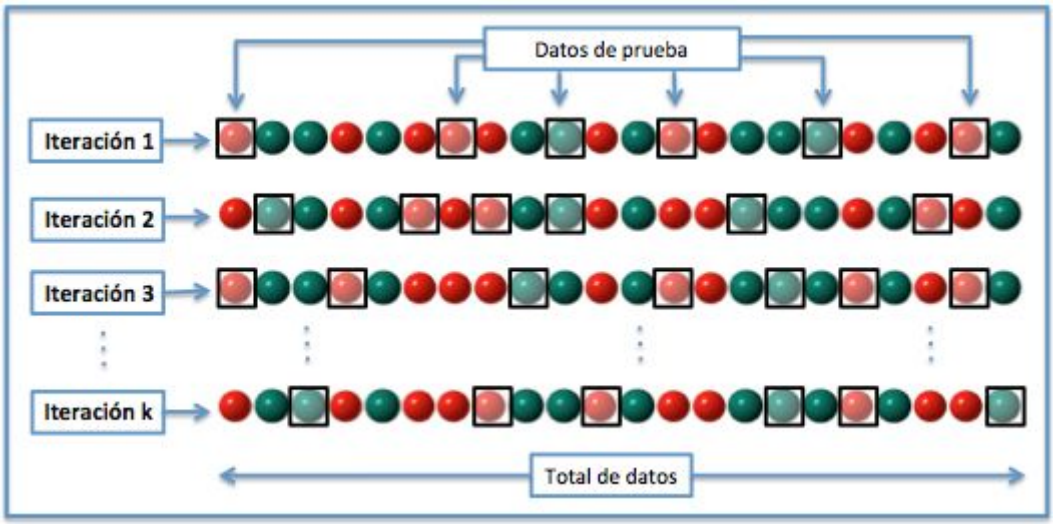


Figura 22. Método aleatorio [39].

Dejando uno fuera *(Leave One Out o simplemente, LOO)*

Este tipo es un caso particular de la validación K-Fold con K = cantidad de muestras. Es así que se generan N iteraciones donde N es la cantidad de muestras, y dentro de cada iteración, uno de los datos hace el papel del de prueba, y todo el resto pasan a ser datos de entrenamiento. Como desventaja, presenta un alto costo computacional resultado de un mayor número de combinaciones a computar y el hecho de que el conjunto de datos de entrenamiento sea de mayor tamaño, que también hace que la tarea de entrenar un modelo tome más tiempo [39]. Ver figura 23.

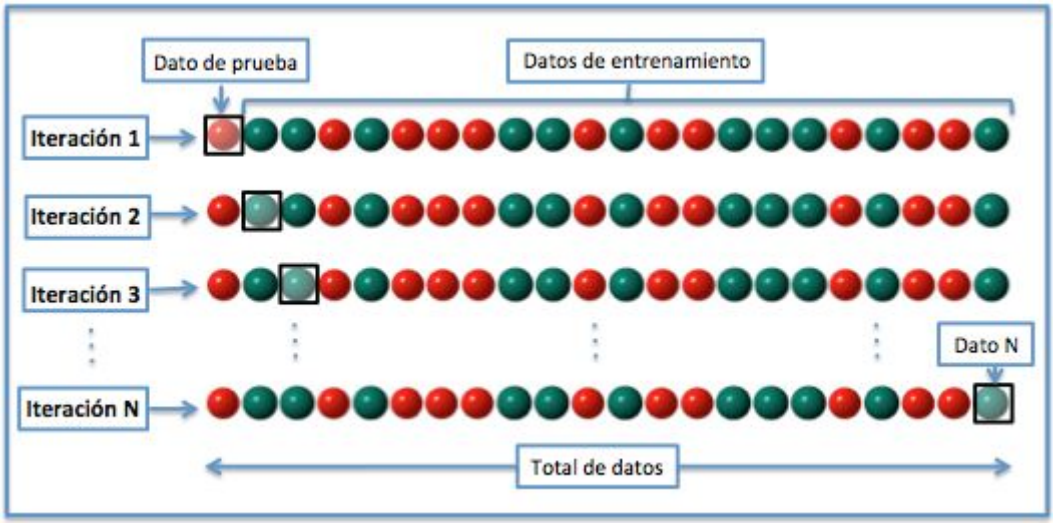


Figura 23. Método *Leave One Out* [39].

Es preciso ahora introducir aquellas métricas que miden la calidad de un modelo. Utilizaremos los nombres utilizados en la lengua inglesa ya que son los más utilizados en artículos científicos.

Métricas de validación para Clasificación

La primera, utilizada para las corridas de clasificación, y tal vez más intuitiva de ellas es la ***accuracy*** **(exactitud)** del método. Su definición es la proporción de resultados correctos que alcanzó el clasificador. Por ejemplo si el clasificador logró clasificar la mitad de las muestras de manera correctamente, tendrá una precisión de 0.5, o expresado porcentualmente, el 50%. Pero ésta métrica no necesariamente representa de manera única la calidad de un método.

En la clasificación binaria, como la que será utilizada en este proyecto, tenemos dos posibilidades: o bien el método etiquetó una muestra positiva como positiva, o cometió un error y lo marcó como negativo, y viceversa. Es así que definimos las siguientes métricas:

Verdaderos Positivos (TP): numero de muestras positivas etiquetadas positivas (correcto).
Falsos Positivos (FP): número de muestras negativas etiquetadas como positivas (error).
Verdaderos Negativos (TN): número de muestras negativas etiquetadas como negativas (correcto).
Falsos Negativos: número de muestras positivas etiquetadas como negativas (error)

Con estas nuevas definiciones, podemos redefinir el mencionado término de *accuracy* como:

$$Accuracy = (TP + TN) / (TP + TN + FP + FN)$$

Es importante notar que la exactitud (*accuracy*) es inversamente proporcional a la suma de falsos positivos y falsos negativos.

Para combatir estas tendencias, introducimos dos nuevas métricas:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

Entonces, *precision* responde lo siguiente: de todas las muestras que el clasificador etiquetó como positivas, ¿en qué proporción acertó? y *recall* responde: de todas las muestras que eran originalmente positivas ¿en qué proporción acertó?

Si un clasificador no comete errores, entonces precision = recall = 1.0. Sin embargo, esto es prácticamente imposible. Mirándolo de otra manera, si realizamos un clasificador que siempre etiquete las muestras como positivas, obtendremos una *precisión* perfecta pero probablemente una *accuracy* muy pobre, y análogamente sucederá con un clasificador que siempre etiquete negativamente y la métrica *recall*.

La conclusión que se saca de esto es que la medición de calidad de un método dado y la calibración de sus hiperparámetros es una cuestión de balancear lo que sea más importante en cada problema: *precision* o *recall*, dependiendo de la lógica de negocio sobre la que se

esté trabajando. Por ejemplo, en detección de fraudes en transacciones bancarias, importa tener un *recall* alto de manera de que la mayoría de las transacciones fraudulentas sean detectadas, incluso ante el costo de perder *accuracy* ya que es de mayor relevancia detectar y lanzar una alarma con respecto los fraudes que efectivamente se llevaron a cabo, que generar falsas alarmas en casos donde no lo son, algo solucionable con un análisis manual más detallado.

En el presente trabajo, se representará como positivo (1) al caso de existencia de evento de inundación, y como negativo (0) al caso contrario. Aplicando el razonamiento en el presente proyecto, consideramos que resulta importante, por lo menos, alertar en la mayoría de los casos positivos posibles ya que éstos pueden desencadenar grandes pérdidas como se detalla en la sección 2.1, incluso ante el costo de generar falsas alarmas. Es así que se volcó por la idea de buscar un equilibrio entre *accuracy* y *recall* lo más alto posible.

Métricas de validación para Regresión

Las principales métricas utilizadas para contabilizar la calidad de los métodos de regresión son R cuadrado, error medio absoluto y error cuadrático medio.

R cuadrado (R^2), o también llamado coeficiente de determinación, es una métrica estadística que representa la proporción de la variación del resultado que puede explicar el modelo entrenado, a partir de las variables de entrada. Su valor está en el rango del 0 al 1, donde un 0 indica que el modelo no puede explicar nada de la variación del resultado, y un 1 significa que describe la variación perfectamente. Sin embargo, éste valor por si sólo no representa un dato completo para la medición de calidad de una predicción ya que distintas áreas de estudio poseen distintas dificultades y aspiran a resultados diferentes a la hora de predecir. Por ejemplo, en el área de psicología humana, un R cuadrado de 0.2 puede llegar a ser un valor altamente aceptable ya que los comportamientos de la mente humana son muy complicados de establecer, mientras que en otras ciencias más exactas, un 0.9 puede aún ser un valor inaceptable. R cuadrado provee una medida de la fuerza de la relación entre el modelo y la variable de respuesta, pero por sí sólo no representa una única medida global [42][43]

Error Medio Absoluto (*mean absolute error*) es una medida de distancia entre dos variables continuas. Si tenemos las variables de observación x_i y aquellas predichas y_i, entonces definimos al Error Medio Absoluto entre la observación y lo predicho como:

$$EMA = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

Y por lo tanto representa un promedio del error absoluto que se cometió para cada observación particular, lo cual resulta, intuitivamente, una buena medida de calidad para un método de predicción simple.

Finalmente, el **Error Cuadrático Medio** (*mean squared error*) es muy similar al anterior, el Error Medio Absoluto, pero con la particularidad de que se suman los cuadrados de las diferencias entre lo observado y lo predicho. Por lo tanto, lo definimos como:

$$EMC = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

Donde su diferencia con el error medio absoluto previamente presentado reside en el cuadrado de la diferencia entre cada observación y realidad. Este enfoque claramente pone mayor “peso” o gravedad en los errores y de esta manera, “castiga” de peor manera a los mismos. Si la distancia entre la observación es cercana a cero, un valor cercano a cero al cuadrado se hace incluso menor, y si es mayor a uno, entonces su cuadrado será aún mayor.

8 Herramientas para el Análisis de Datos

En esta sección se presentan herramientas más importantes que presentan una alternativa al desarrollo utilizando Machine Learning. Dichas herramientas en general ya cuentan con implementaciones que buscan aportar soluciones a problemas comerciales que requieren de un modelado predictivo además de resultar más accesibles para usuarios sin capacitaciones técnicas en el ámbito de la programación.

8.1 Weka

Weka es una herramienta conformada por un conjunto de algoritmos de machine learning para realizar tareas de data mining. Estos algoritmos son aplicables directamente a un conjunto de datos así como también utilizables desde un proyecto Java. Además contiene herramientas para preprocesado de datos, clasificación, regresión, agrupamiento (o análisis de grupos), reglas de asociación y visualización [44] A su vez cuenta con una interfaz gráfica que permite unificar todas las herramientas disponibles para un uso más simple.

Como principales características, se puede describir a Weka como una herramienta muy adaptable a las necesidades mediante la aplicación de tareas básicas de data mining. Los datos que utiliza pueden tener como origen ficheros planos con un número fijo de atributos por registro así como accediendo a instancias de base de datos mediante SQL gracias a JDBC [45].

Su interfaz está conformada por una consola mediante la cual se puede acceder a las cuatro interfaces principales de la herramienta, a grandes rasgos estas se pueden describir de la siguiente forma [45]:

- **Simple Command Line Interface**: En esta consola es posible invocar cualquier función directa de Weka mediante Java.
- **Experimenter**: Interfaz que resulta favorable para la aplicación de pruebas a gran escala, que pueden requerir del uso de más de una función y la obtención de resultados numéricos.
- **Explorer**: Interfaz que brinda el acceso a todas las funciones de manera sencilla, agrupandolas en paneles.
- **KnowledgeFlow**: Aplicable para generar proyectos de data mining mediante flujos de información.

Entre las ventajas de esta herramienta, se destaca el libre acceso por parte de los usuarios debido a estar disponible bajo licencia pública GNU. Además posee una importante cantidad de técnicas tanto para modelado como para procesamiento de datos, las cuales son utilizables mediante una interfaz de usuario amigable y simple. Por otra parte, el aspecto más negativo a resaltar es que los resultados tienden a resultar difíciles de comprender cuando se utilizan métodos de combinación de modelos [45].

8.2 R

R es un lenguaje de programación publicado al igual que Weka bajo licencia GNU y por lo tanto libre. El mismo conforma un entorno de análisis estadístico para la manipulación de datos, sus cálculos y la posterior visualización de estos y los resultados obtenidos mediantes gráficos. Este es considerado como una implementación más moderna y libre de S (lenguaje de programación estadístico desarrollado en 1976) [46].

R provee una amplia variedad de estadísticas (modelado lineal y no lineal, tests estadísticos, análisis de series temporales, clasificación, agrupación, etc), además es muy extensible contando con más de 10.000 paquetes disponibles [46][47].

En la actualidad, R se ubica como el software con mayor cantidad de recursos, grado de desarrollo y aceptación, siendo usado por empresas como Google y Facebook entre otras y dentro de la comunidad científica para la realización de investigaciones [47].

Algunas de las principales características con las que cuenta son [48]:

- Almacenamiento y manipulación de datos
- Herramientas de análisis de datos
- Gráficos para análisis de datos
- Operadores de cálculo para arreglos y matrices

Además, a diferencia de otros lenguajes estadísticos, R maneja el uso de objetos como una entidad. Esto significa que toda expresión evaluada en R es realizada en una serie de pasos y sus resultados intermedios son guardados en objetos, utilizables para posteriores análisis. Finalmente, como consecuencia de ser una versión más moderna de S, también existe la posibilidad de utilizar una gran cantidad de programas realizados con S que pueden utilizarse en R [48]

8.3 Pentaho

Pentaho es una plataforma de código abierto enfocada a la realización de análisis de datos e informes. A través de su implementación mediante Java, el mismo brinda la posibilidad de moldearse a necesidades específicas de cada organización [49].

Al ser un software orientado al negocio (Business Intelligence o simplemente BI) brinda entre sus funciones la realización de reportes intuitivos, análisis OLAP (procesamiento

analítico en línea), cuadros de mando, integración de datos, minería de datos y plataforma BI [49].

Sin profundizar en detalles, algunas de las principales características de Pentaho son [50]:

- El concepto de “Governed Data Delivery”, según Pentaho refiere a la capacidad de poder acoplar datos confiables y oportunos para brindar análisis a escala para todos los usuarios en todos los entornos.
- Posee una fácil integración a cualquier flujo de trabajo gracias a tener una arquitectura multi-tenant.
- Capacidad para acceder y combinar información para otorgar información analítica lista para el consumo de los usuarios finales en conjunto con una interfaz “drag and drop” para eliminar complejidad.
- Integración nativa mediante una capa que se integra a cualquier fuente de datos, incluyendo Hadoop y NoSQL entre otros.

9 Antecedentes

Como parte de este proyecto también se estudió sobre lo que ya se ha investigado e implementado a nivel mundial y local en cuanto a predicción de inundaciones y trabajos relacionados.

9.1 Global Flood Monitoring System

Según la agencia estadounidense NASA, predecir inundaciones es algo de lo más complicado e involucra una combinación de parámetros de tipo ambiental, de estado del tiempo reciente y actuales [51] En 2012, la misma financió, y trabajó en conjunto con el Dr. Huan Wu de la Universidad de Maryland en el Global Flood Monitoring System (GFMS) [52], un sistema experimental de monitoreo de inundaciones. El mismo basa su funcionamiento inicialmente en el Tropical Rainfall Measuring Mission satélite (TRMM), y posteriormente migró a utilizar el Global Precipitation Measurement satellite (GPM), ambos satélites de observación sobre la Tierra en tiempo real de precipitaciones, siendo el segundo de mejor precisión y tecnología.

El funcionamiento del sistema combina información brindada por el satélite en cuestión, un modelado de la composición de la superficie de la Tierra que incluyen densidad de vegetación, sedimentación de los ríos y el terreno de casi todo (rango de latitudes 50°N - 50°S) el planeta llamado Variable Infiltration Capacity, así como también un modelo de enrutamiento de excedentes de agua en cursos, el Dominant River Tracing based runoff-Routing (DRTR), para calcúlar cuánta de la lluvia que caerá será absorbida por la tierra y penetrará las capas terrestres, y cuánta terminará desplazándose por los cursos de agua, eventualmente saturando su caudal y resultando en una inundación.

El producto resultante es una aplicación web con funcionamiento continuo sin

interrupciones, accesible mediante http://flood.umd.edu/ [53], donde los usuarios pueden visualizar, plasmadas en un mapa, las siguientes estadísticas:

- Inundación (profundidad del agua en mm)
- Flujos de agua en una grilla de resolución de 12km
- Flujos de agua en una grilla de resolución de 1km
- Flujos de agua sobre los máximos establecidos como normales
- Precipitaciones instantáneas en mm/h
- Acumulación de precipitaciones en mm de 3, 5 y 7 días previos

Presenta una interfaz gráfica que permite al usuario hacer zoom y mover la posición del enfoque sobre el mapa de manera tal de que pueda consultar la región de interés y ver sobre el mismo un coloreado del tamaño de la grilla seleccionada con una leyenda que vincula distintos colores a distintos valores, en relación al dato que se haya seleccionado para plottear [137].

El sistema es capaz de mostrar al usuario una gráfica de la tendencia de cualquiera de estas mencionadas estadísticas en un período de tiempo pasado que el usuario puede ingresar mediante la interfaz ofrecida [137].

Cuenta como inundada una celda donde el nivel de agua actual o proyectado (predecido) supere la cota máxima definida para esa celda, siendo ésta derivada de estadísticas basadas en 13 años de retrospectiva (estadísticas) de almacenamiento de agua [137].

Esta información brindada por el GFMS es habitualmente utilizada por entidades como las Naciones Unidas o la Cruz Roja, antes, durante y luego de los desastres naturales, y se suele combinar con datos demográficos a fin de calcular y dirigir de manera eficiente esfuerzos de recuperación y auxilio de las poblaciones adyacentes a los cursos afectados [53].

9.2 Sistema de Alerta Temprana Prohimet-Yi

En Uruguay, se ha trabajado previamente en esta temática a nivel de sistemas de información, motivados por desastres que han generado daño socioeconómico a la población.

Prohimet-Yi es un ejemplo de estos sistemas, siendo este un sistema de alerta temprana cuyo objetivo es tener una mejora en la gestión de las inundaciones que se dan en la ciudad de Durazno [54].

El mismo empezó a llevar a cabo sus etapas iniciales para ponerse en funcionamiento entre junio de 2009 y diciembre de 2011, siendo financiado por la OMM (Organización Meteorológica Mundial) y realizado por el Instituto de Mecánica de los Fluidos e Ingeniería Ambiental (IMFIA) de la Facultad de Ingeniería de la Universidad de la República. Además contó y actualmente cuenta con el apoyo de las instituciones nacionales involucradas en el proyecto y de los miembros de la red PROHIMET [54].

El proyecto original estuvo armado en 3 etapas [54]:

- Etapa 1 (Julio 2009 - Junio 2010): en esta primera etapa se trabajó en la recopilación de antecedentes e información hidrometeorológica. Además se realizaron avances en el relevamiento de las secciones transversales del Río Yi, de la modelación hidrológica del mismo y en la utilización del hidroestimador sobre la cuenca. Por último se realizó entrenamiento sobre cómo realizar aforos.

- Etapa 2 (Julio 2010 - Marzo 2011): se continuó trabajando en los avances realizados en la etapa 1 así como en la elaboración de las curvas IDF (establecen una relación matemática entre la intensidad, la duración y la frecuencia de las precipitaciones;) de la ciudad de Durazno y de un sistema de información geográfico de la cuenca del Río. Finalmente se realizó una recopilación de información, antecedentes y análisis del sistema urbano.

- Etapa 3 (Abril 2011 - Diciembre 2011): en esta etapa final se puso en operación el Sistema de Predicción de Niveles en la ciudad de Durazno y se comenzó con la capacitación para la utilización de la herramienta y de la interpretación de las salidas.

Además del objetivo de mejorar la gestión en las inundaciones de la ciudad de Durazno, el proyecto enmarca en sus metas lo siguiente [54]:

- Obtener un sistema de alerta temprana (SAT) operativo y de tiempo real con buena precisión y confiabilidad, como se requieren.

- Desarrollar un sistema espacial que apoye al Comité de Emergencia Departamental.

- Fomentar el fortalecimiento institucional de los organismos intervinientes en el sistema de alerta con una mejora de capacitación y entrenamiento y buscando sustentabilidad en el tiempo.

- Lograr un acople en cuanto a la información meteorológica, obtenida de distintas fuentes.

- Expandir el diálogo interinstitucional a nivel nacional en lo que respecta a definición de responsabilidades en las distintas etapas de alerta temprana.

Más adelante, en el año 2012, comienza el trabajo de tres estudiantes de ingeniería en un proyecto de grado con el fin de extender el sistema existente.

Éste consistió en una investigación y reelaboración de la herramienta existente para la alerta temprana de inundaciones en el Río Yí, buscando la mejora y el agregado de nuevas funcionalidades [55].

El objetivo principal de este proyecto de grado, fue el de investigar e implementar un sistema de información geográfica (SIG) que, además de alertar en caso de inundación, permita realizar consultas espaciales, (cantidad de personas a ser evacuadas, de qué regiones de la ciudad, etc.) en base a la información geográfica de la ciudad de Durazno.

Los objetivos secundarios consisten en incorporar mejoras en la parametrización de los scripts del modelo actual, brindar una gran flexibilidad para generar consultas espaciales de forma dinámica, mejoras en la notificación de las alertas utilizando mensajes de texto, entre otros [55].

10 Datos Abiertos

El presente proyecto planteó la necesidad de obtener una gran cantidad de datos históricos (estados del tiempo, alturas de ríos, registros de desastres), como consecuencia de esto fue necesario investigar, buscar fuentes de datos públicos y ponerse en contactos con diferentes entes que resultaran oportunos para proporcionarnos la información requerida.

En este contexto resulta interesante comparar cómo se dan estos datos en Uruguay, la calidad de los mismos y su disponibilidad en comparación con países de primer mundo o generalizando a nivel internacional.

Como concepto previo, hay que mencionar a los Datos Abiertos como una tendencia moderna con gran potencial de desarrollo. Esto es, que cualquier ente (individual u organizacional) recaba información (datos) para realizar sus tareas. En particular, el Gobierno resulta vital en este contexto ya que recoge una amplia cantidad de datos de empresas, de la población, del medio ambiente, económicos, de salud, geográficos, etc [56]

Definiendo a los Datos Abiertos, estos son los datos que existe la posibilidad de que cualquier persona sea libre de utilizarlos, reutilizarlos y redistribuirlos. Siendo sus principales características las siguientes [56]:

- Disponibilidad y Acceso: completamente disponibles y con un costo de reproducción razonable, siendo la descarga gratuita a través de internet el caso óptimo.
- Reutilización y Redistribución: posibilidad de realizar un producto como derivado de estos datos, así como usarlos en conjunto con otras fuentes de información y distribuirlo de forma gratuita.
- Ausencia de Restricción Tecnológica: No debe haber obstáculos de carácter tecnológico para su uso y redistribución (Formato abierto).
- Participación Universal: Cualquier persona puede utilizar, reutilizar y redistribuir, sin restricciones respecto a las acciones posibles sobre estos o condicionantes.

Por otra parte, los datos de gobierno se consideran abiertos si cumplen los siguiente ocho puntos establecidos internacionalmente por Open Government Data [57]:

- Completos: Todos los datos públicos están disponibles. Se considera datos públicos a aquellos que no tienen restricciones de privacidad, seguridad o privilegio.
- Primarios: Los datos son obtenidos en la fuente, con el mayor nivel posible de granularidad, sin ser modificados ni agrupados.

- Periódicos Los datos quedan disponibles tan pronto como sea necesario para preservar su valor.
- Accesibles: Los datos quedan disponibles para la mayor cantidad posibles de usuarios y propósitos
- Procesables: Los datos están estructurados (tanto como es posible), permitiendo así un procesamiento automatizado.
- No Discriminatorios: Disponibles a todo publico sin necesidad de registros
- Sin Licencia: Los datos no se encuentran sujetos a ningún tipo de regulación de derechos. Pudiéndose permitir únicamente en casos razonables restricciones de privacidad, seguridad o privilegio.

Además, el cumplimiento de estos puntos, debe ser comprobable.

10.1 Datos Abiertos en Uruguay

En los últimos años, se ha promovido la liberación de datos gubernamentales a nivel público, esto forma parte de los avances realizados en el marco de lo que se conoce como e-gobierno (gobierno electrónico). Estos avances se han realizado principalmente mediante la AGESIC (Agencia para el Desarrollo del Gobierno de Gestión Electrónica y la Sociedad de la Información y del Conocimiento), que es el organismo encargado de planificar e implementar estrategias de Gobierno Electrónico, avanzando en la concreción de políticas de acceso a la información y de datos abiertos en el Estado [58].

Con el fin de regularizar la liberación de datos es que en 2008 se crea la Ley de Acceso a la Información (Ley 18.381), la cual explica en su primer artículo tiene como objetivo [59] "*promover la transparencia de la función administrativa de todo organismo público, sea o no estatal, y garantizar el derecho fundamental de las personas al acceso a la información pública*".

Más allá de que nuestro país no cuenta con una política definida en cuanto a la publicación de información pública en el catalogo de datos abiertos (en catalogodatos.gub.uy [60]), si existe una licencia definida que presenta las siguientes condiciones [60]:

"El Usuario reutilizador de los datos de este sitio, que no disponga de licencia específica indicada por la entidad pública que los aporta (Usuario catalogador), deberá cumplir, al menos, las siguientes condiciones básicas:
1) Mantener el sentido original de la información,
2) Citar siempre la fuente
3) Explicitar la fecha de la última actualización.
Se permite cualquier explotacion de los datos abiertos, incluyendo una finalidad comercial, así como la creacion de obras derivadas, estando permitida su distribucion sin ninguna restriccion. La utilización, reproducción, modificación o distribución de los conjuntos de datos supone siempre la obligación de reconocer, citar y/o enlazar a la entidad de que se trate como la fuente de los conjuntos de datos. Deben conservarse, y por tanto no alterarse

ni suprimirse los metadatos sobre la fuente, fecha de actualización y las condiciones de reutilización aplicables incluidos, en su caso, en el documento puesto a disposición para su utilización o reutilización. AGESIC puede solicitar a un usuario que cese en el uso de esta licencia y cualquier forma de distribución de datos que se realice bajo esta, en el caso que considere que ha existido una violación a los términos y condiciones descriptos. Ello, sin perjuicio de las medidas de carácter legal que AGESIC pueda llegar a adoptar con el Usuario, que haya incumplido los Términos y Condiciones generales definidos."

Vale aclarar que dicho reglamento únicamente aplica a los datos dentro del portal de catálogo de datos abiertos, no a todos los datos en poder del Estado.

Comparando a nivel internacional, Uruguay se encuentra bien posicionado en los rankings de medición, siendo los principales el Open Data Barometer [138] y el Global Open Data Index [139], estando en el puesto 17 y 19 respectivamente en cada ranking.

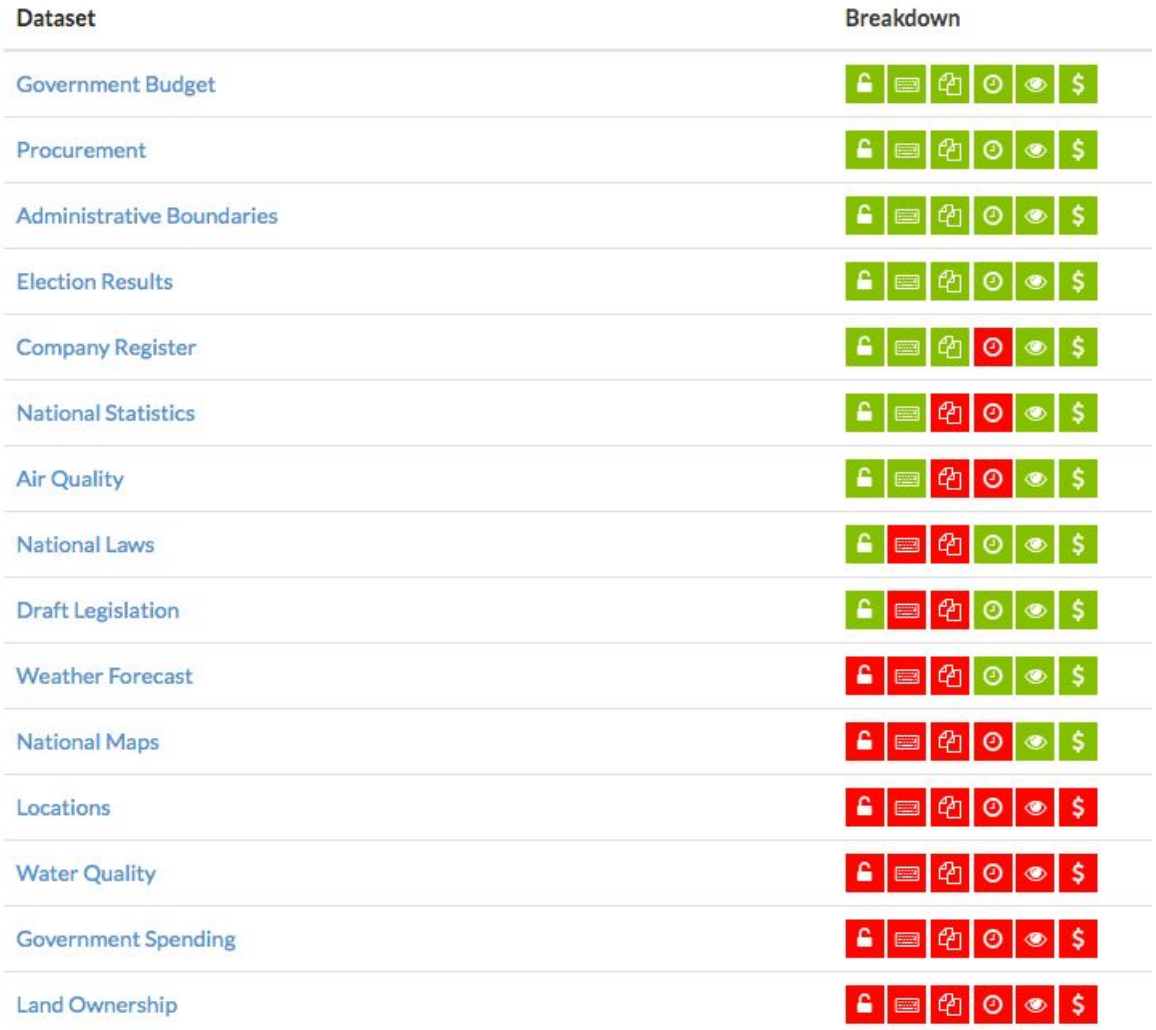


Figura 24. Desglose de los datos públicos en Uruguay por Global Open Data Index. [139]

En la figura 24, usando como ejemplo el caso de los pronósticos del tiempo, que son los datos que nos conciernen en este proyecto, la simbología en orden de izquierda a derecha representa lo siguiente:

- Licencia abierta: No.
- Legible por computadora: No. Esto significa que no hay una forma sencilla de consumirlos como puede ser una API o CSV.
- Descargables en un paso: No.
- Actualizados: Si.
- Disponibles publicamente: Si.
- Acceso gratuito: Si.

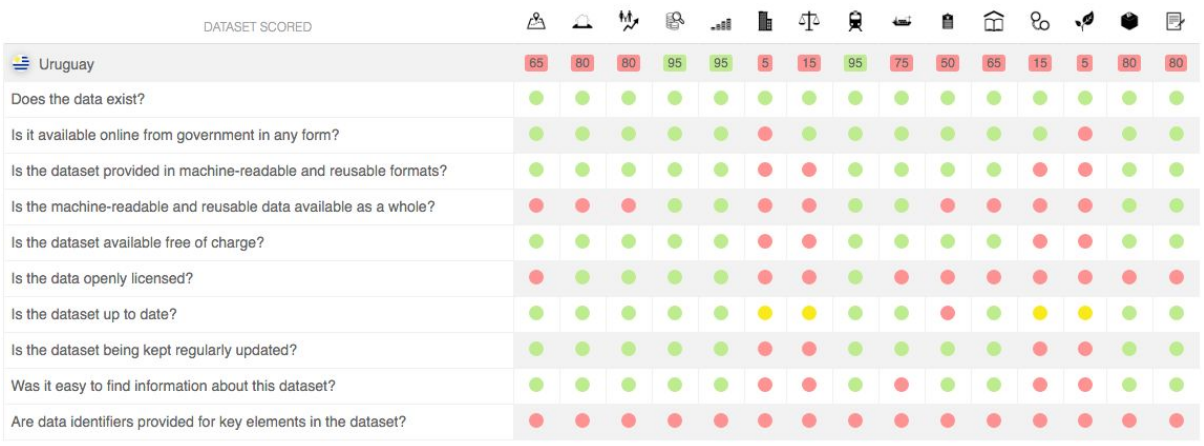


Figura 25. Desglose de información acerca de los datos públicos por Open Data Barameter. [138]

Como se puede observar en ambas tablas de información (ver figuras 24 y 25), Uruguay posee gran cantidad de información, la cual se encuentra mayormente accesible y eso es lo que favorece su posicionamiento en dichos rankings, pero se encuentra atrasado en cuanto a su fácil acceso o consumo, siendo en la mayoría de los casos imposible de obtener en una única descarga o de consumir mediante APIs (u otras tecnologías).

En nuestro país, el área que se encuentra más avanzada es la geográfica, habiendo mapas de todo el país, calles e incluso en la capital se puede obtener todos los cruces de las calles, semáforos y los horarios de los ómnibus con la posibilidad de ver en tiempo real su ubicación.

Instituto de Computación - Facultad de Ingeniería - Universidad de la República

Montevideo, Uruguay, 2016

Proyecto de Grado - Anexo 2

Manual de Usuario

Tutor Libertad Tansini

Co-tutor Sandro Moscatelli

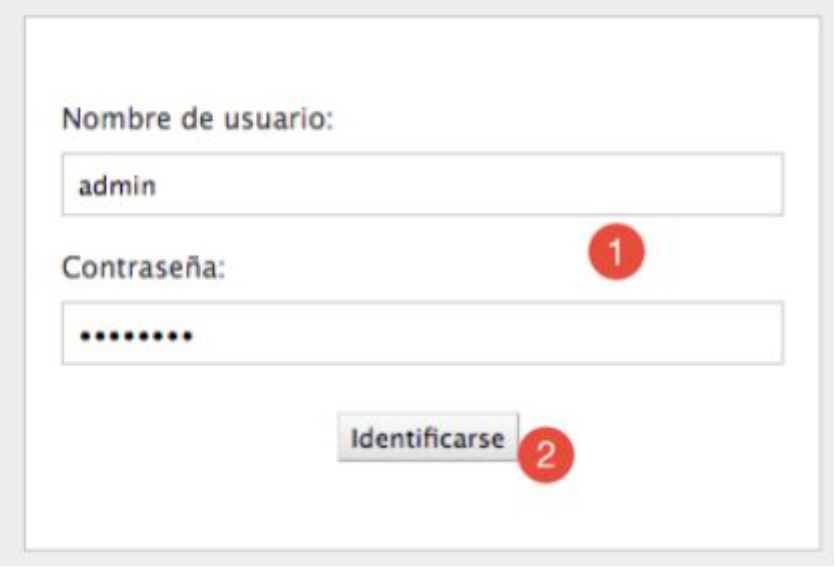
Ignacio Chiazzo

Felipe Garcia

Guillermo Leopold

1. Inicio de Sesión

La pantalla de inicio de sesión será la primera que un usuario encuentre al navegar a la raíz de la aplicación web. Para acceder al sistema el usuario debe iniciar sesión con una cuenta de administrador que se le será suministrada, mediante el uso de un nombre de usuario y contraseña.

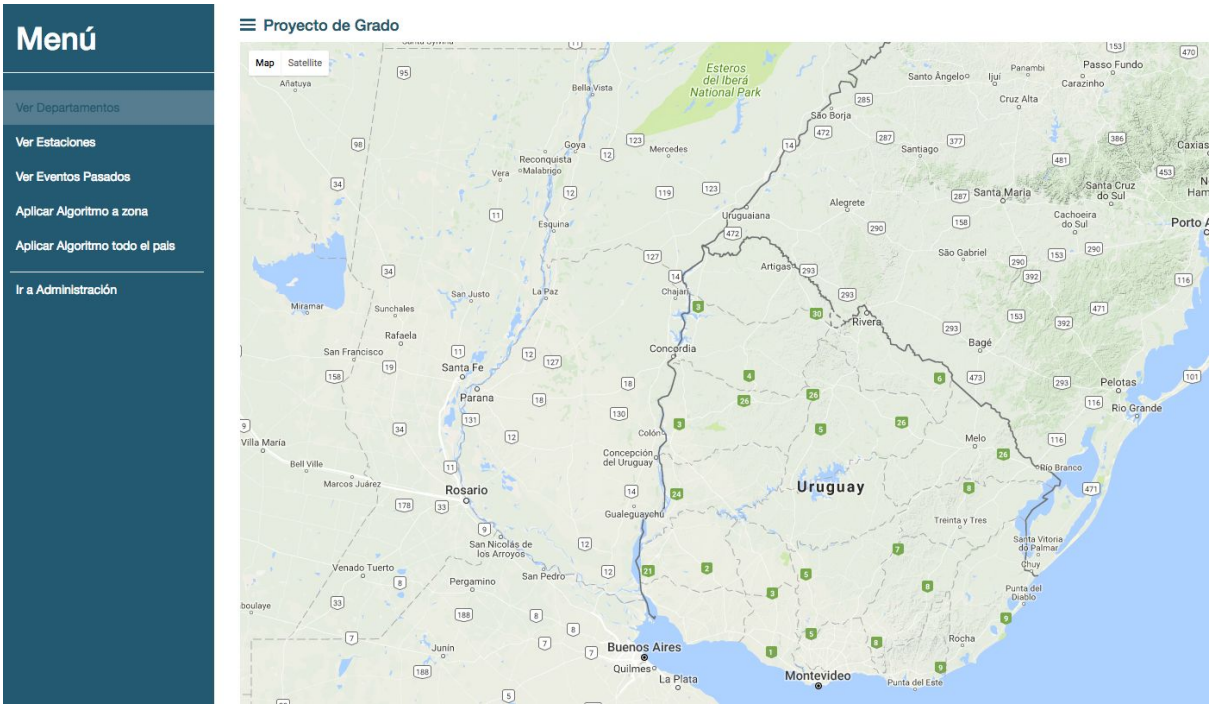


Para iniciar sesión debe:

1. Llenar el formulario con su nombre de usuario y contraseña correspondiente
2. Clickear en “Identificarse"

2. Pantalla Principal

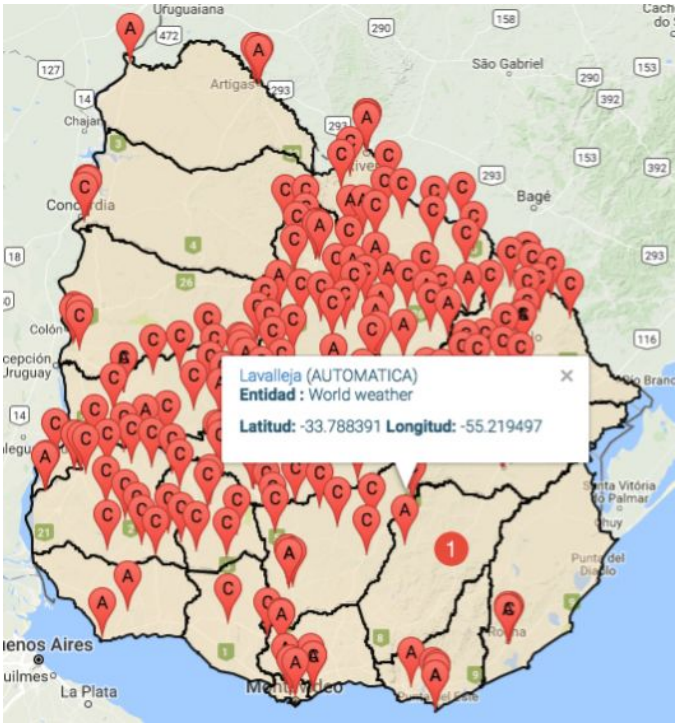
Una vez autenticado en el sistema como usuario, verá la siguiente pantalla como principal, con una vista del mapa embebido y enfocado por defecto en el territorio Uruguayo. En el costado izquierdo se encuentra un menú que permite utilizar funcionalidades y navegar a otras partes de la aplicación.



2.1. Ver Estaciones

En el mapa de la pantalla principal, podrá desplegar con marcadores la posición de todas las estaciones de medición que se tienen registradas. Para ello, debe:

1. Clickear en “Ver Estaciones” y verá todas las estaciones señaladas con un marcador. Si desea remover los marcadores del mapa, solo basta con volver a clickear el mismo botón. Aquellas estaciones de tipo automáticas (operadas sin intervención humana) tendrán una “A” en el marcador, mientras que las convencionales tendrán una “C”.

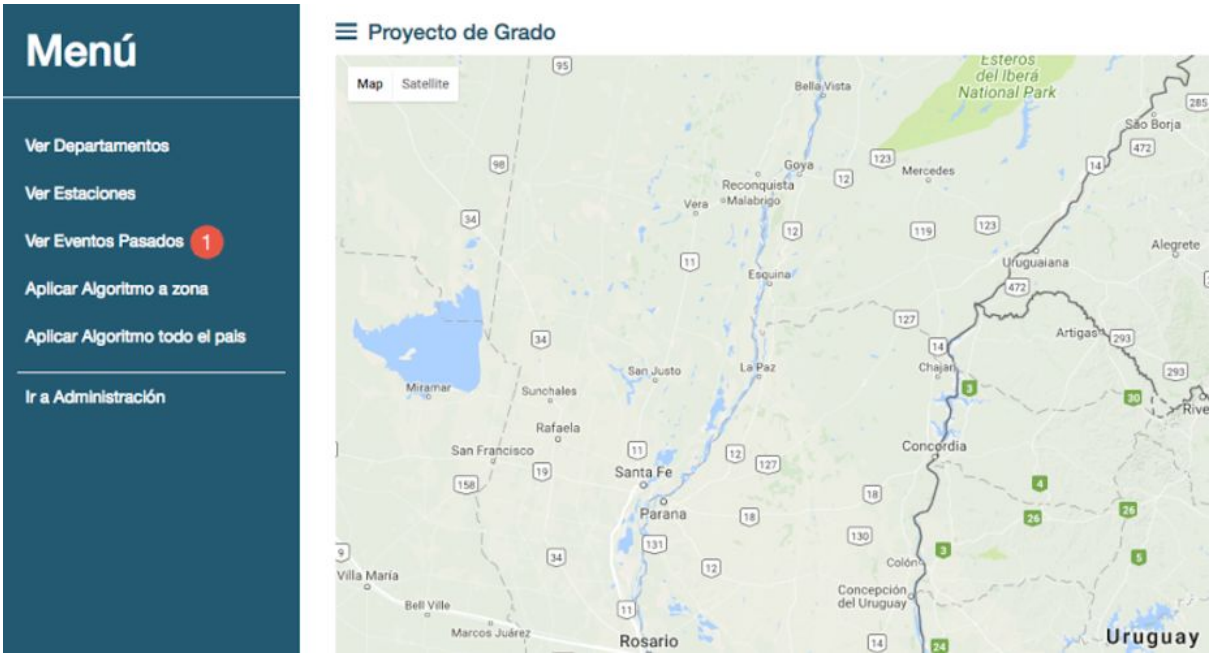


Para ver información sobre una estación en particular, debe:

1. Clickear sobre el marcador correspondiente y un InfoWindow se abrirá con la información correspondiente.

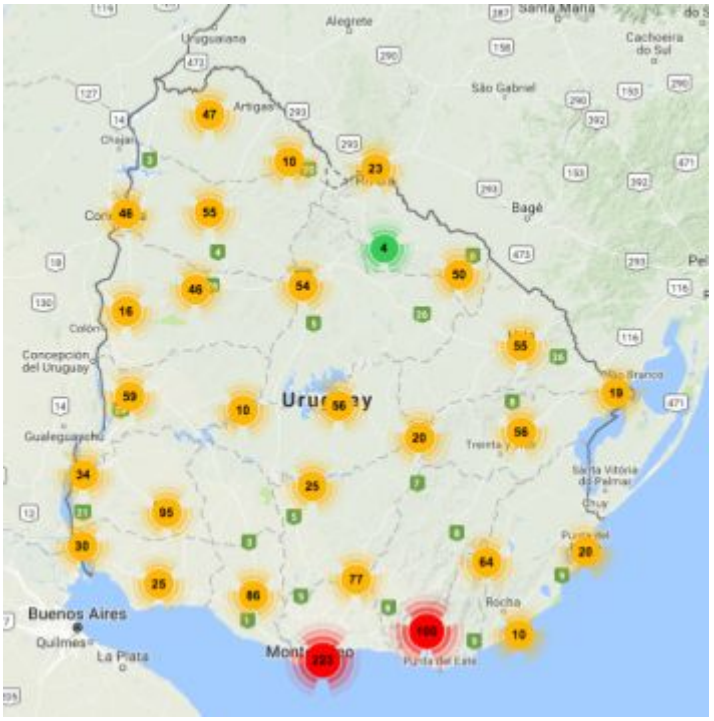
2.2. Ver Eventos Pasados

En la pantalla principal con la vista de mapa, si desea ver los eventos pasados geolocalizados mediante *marcadores agrupados*, debe:



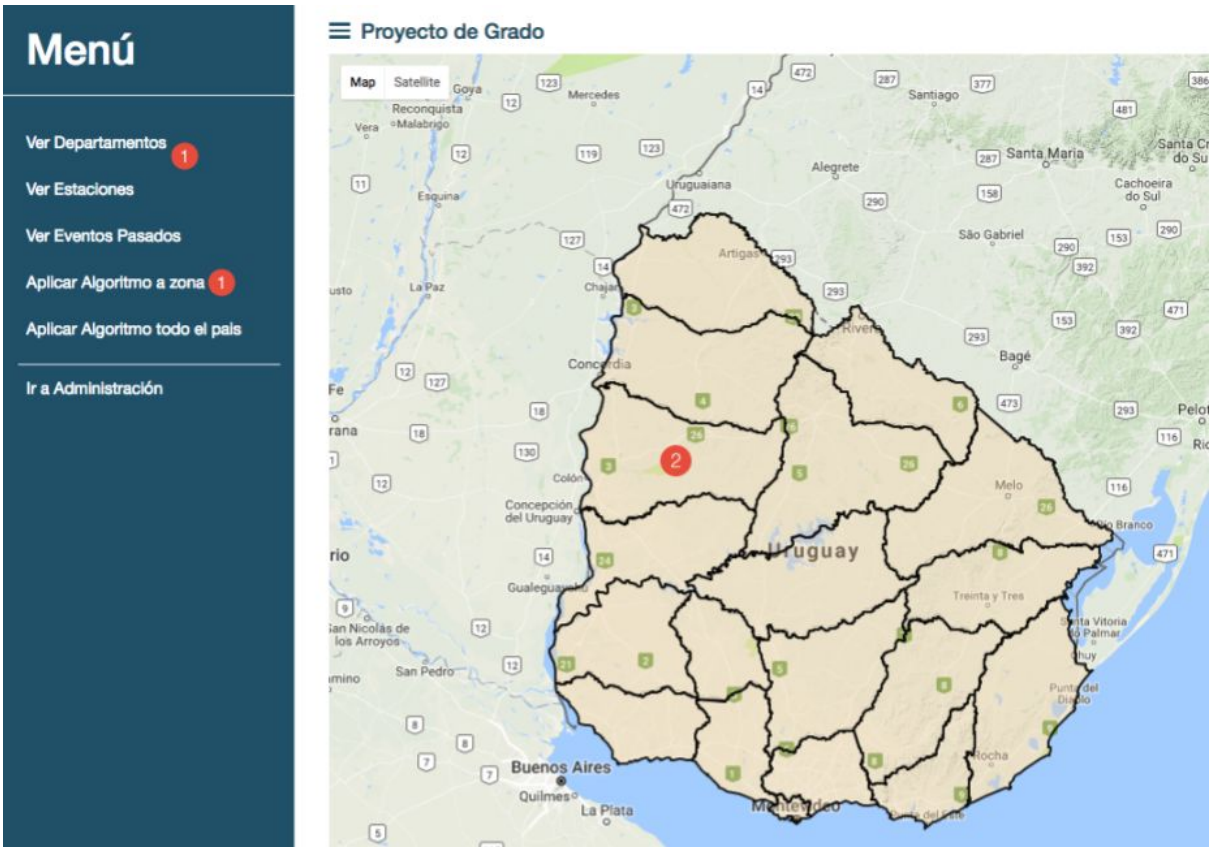
1) Clickear en “Ver Eventos Pasados” y la localidad donde fueron registrados esos eventos serán marcados con *marcadores agrupados* (ver sección 4.5 del informe principal del proyecto)

Y el resultado será desplegado en el mapa de la siguiente manera:



2.3. Aplicar Algoritmo a Zona

Una vez en la pantalla principal, el usuario tendrá la posibilidad aplicar al algoritmo predictivo a una zona, en este caso un departamento.



Para ello, debe:

1) Clickear ya sea en “Ver Departamentos” o “Aplicar Algoritmo a Zona”
2) Clickear en el departamento para el cual se quiere correr la predicción

Luego de seleccionada la zona, un dialogo modal se abrirá para seleccionar una fecha para la cual se quiere correr, y finalmente confirmar la ejecución del algoritmo. Para ello se tendrán dos opciones:

Una predicción real: ésta será posible para la fecha corriente o fechas anteriores a la misma (el pasado). Ésta predicción arrojará el resultado para el día corriente, y tomará como datos de entrada para entrenar al algoritmo, datos reales del datawarehouse generado como primera etapa principal del proyecto.

Una simulación: ésta será una alternativa experimental para fechas futuras. Dado que no se tienen datos climáticos futuros, por temas obvios (no han pasado), se le solicita al usuario que ingrese los datos de precipitaciones y alturas de ríos que tenga como pronóstico, o simplemente lo que desee. Consideramos esta característica como un área de

experimentación para observar el comportamiento de los resultados en base a los datos recibidos.

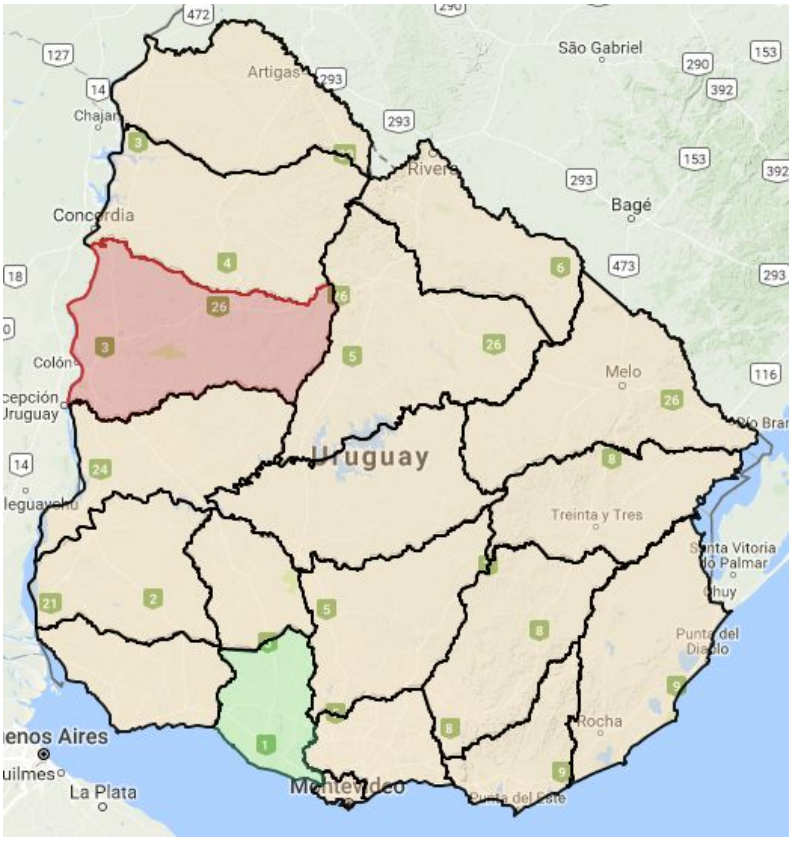
2.3.1. Correr Predicción Real

Una vez se despliega el cuadro de diálogo para aplicar el algoritmo, si se ingresa una fecha anterior a la corriente, se verá lo siguiente:

Para efectivamente ejecutar la corrida, se deberá:

1) Seleccionar el departamento objetivo. Por defecto será seleccionado aquel sobre el cual se clickeó en el paso (2) del ítem anterior del manual.
2) Seleccionar la fecha (pasada)
3) Clickear “Aplicar Algoritmo”

Como resultado se tendrá un mensaje de diálogo de confirmación de la corrida, y además se coloreará el polígono del departamento con ROJO, si hay altas probabilidades de un evento de inundación para la fecha ingresada, o VERDE en caso contrario. Recordar que la predicción arroja un resultado binario. Ejemplo a continuación:



2.3.2. Correr una Simulación

Para correr una simulación para una fecha futura, se debe seleccionar una fecha futura a la corriente, y automáticamente se desplegará en el cuadro de diálogo un formulario para que el usuario ingrese los valores climatológicos.

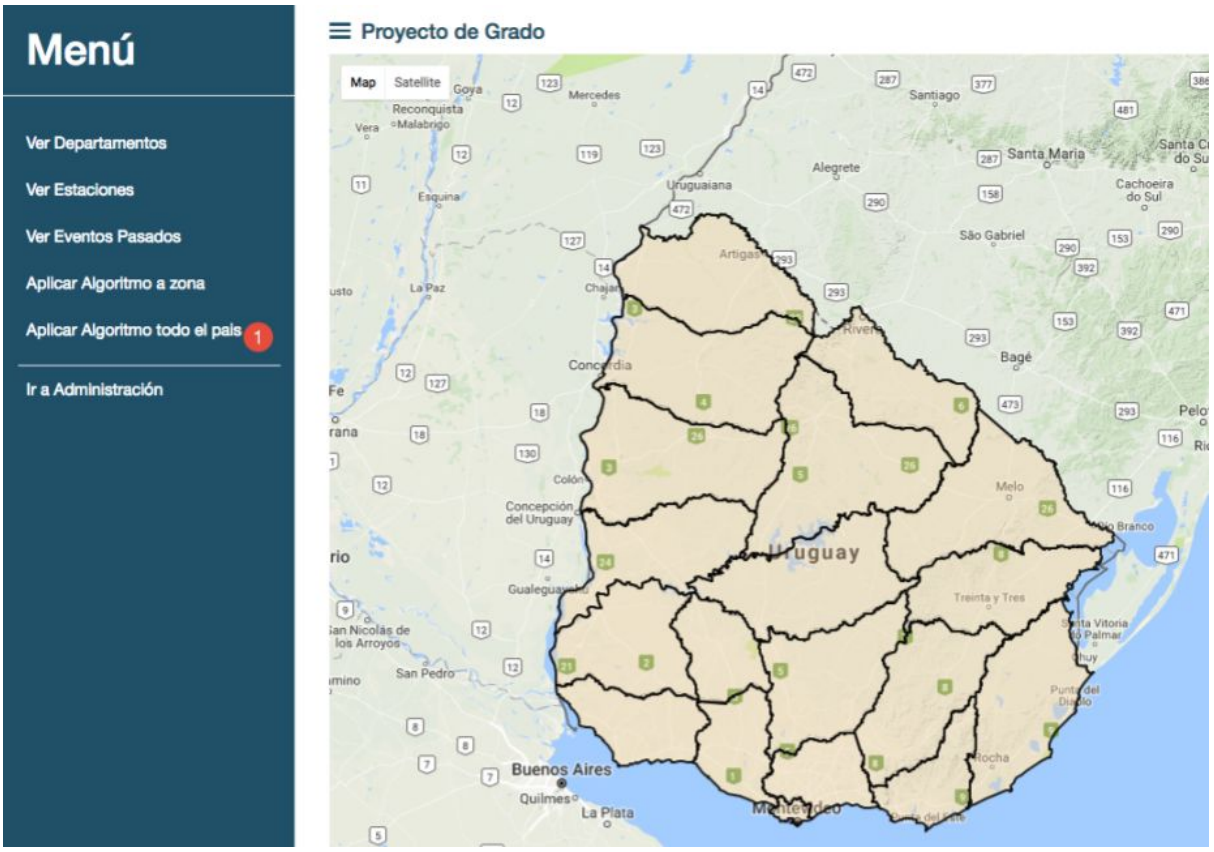
Para efectuar finalmente la corrida, deberá:

1) Seleccionar el departamento objetivo
2) Seleccionar una fecha (futura)
3) Clickear en el símbolo de menos (-) para desplegar el formulario
4) Llenar los valores numéricos de precipitaciones y alturas de río mediante sliders que facilitan su completado.
5) Clickear “Aplicar Algoritmo”

El resultado será análogo al punto anterior del presente manual.

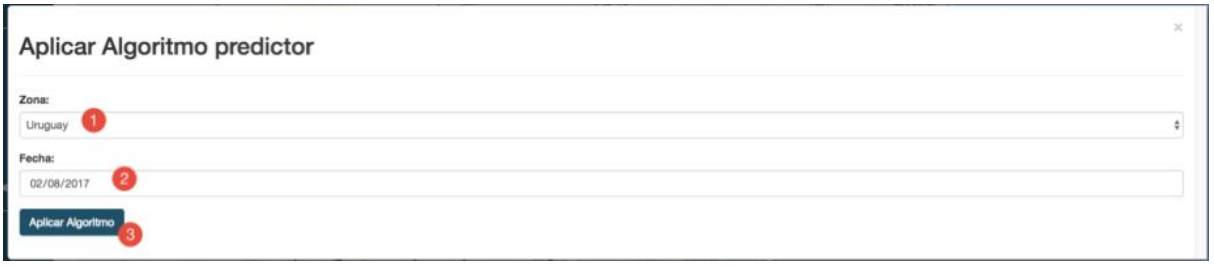
Aplicar Algoritmo a todo el País

Para aplicar el algoritmo a todo el país, deberá:



1) Clickear en el botón “Aplicar Algoritmo todo el país”

Luego de lo cual se verá ante un cuadro de diálogo similar al del punto anterior.

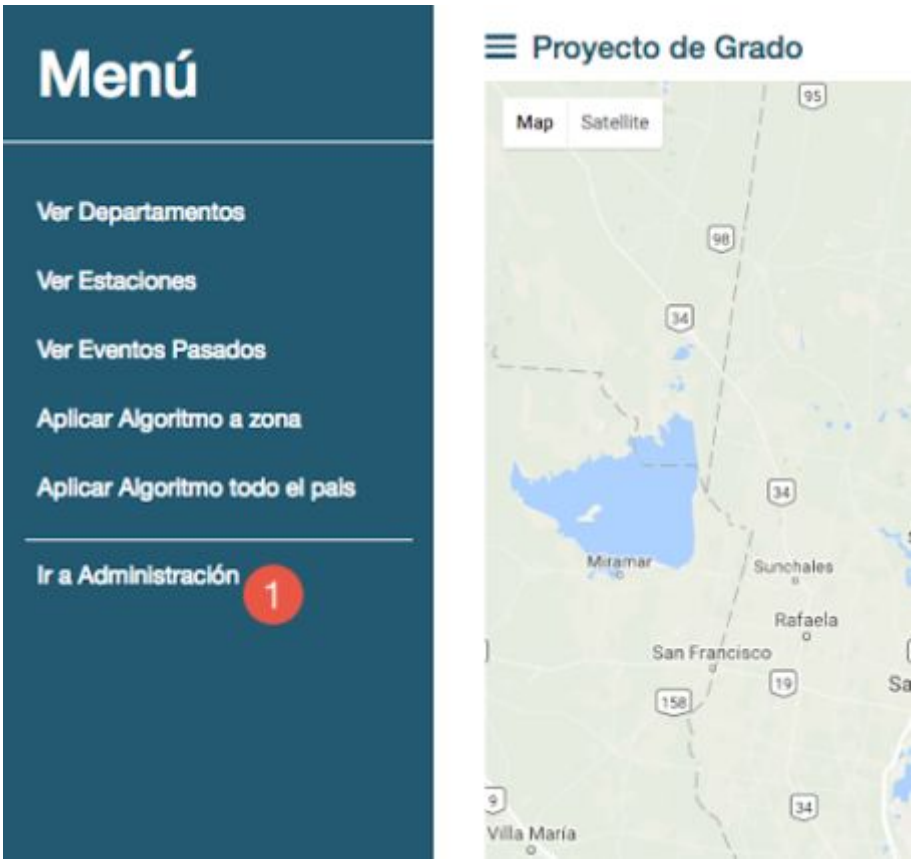


Donde deberá:

1) Seleccionar “Uruguay” como zona (por defecto si siguió el paso anterior)
2) Seleccionar una fecha pasada ya que no funcionarán las simulaciones, debido a que el volumen de datos requerido para cada departamento es muy voluminoso para ingresarlo manualmente.
3) Clickear el botón “Aplicar Algoritmo”

2.4. Ir a Administrador

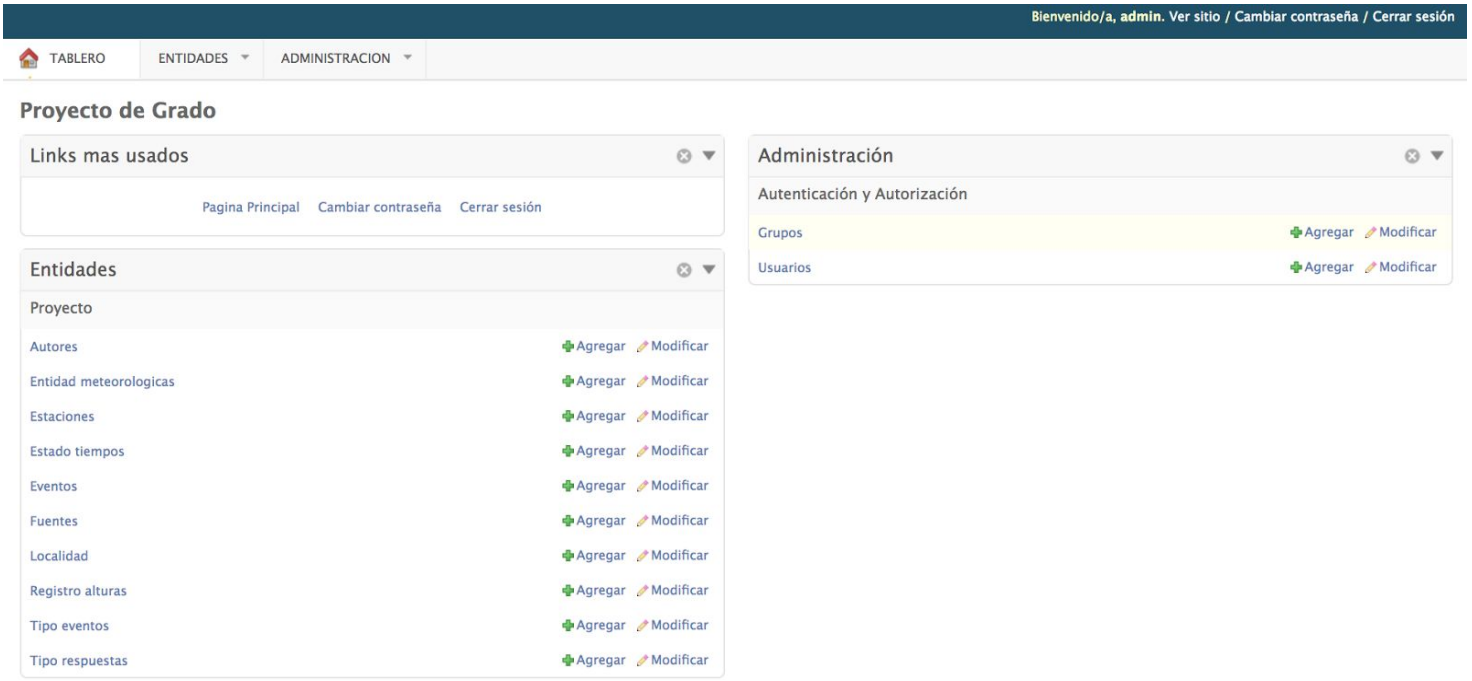
En la pantalla principal, si se desea acceder al panel de administración, se debe:



1) Clickear en “Ir a Administración”.

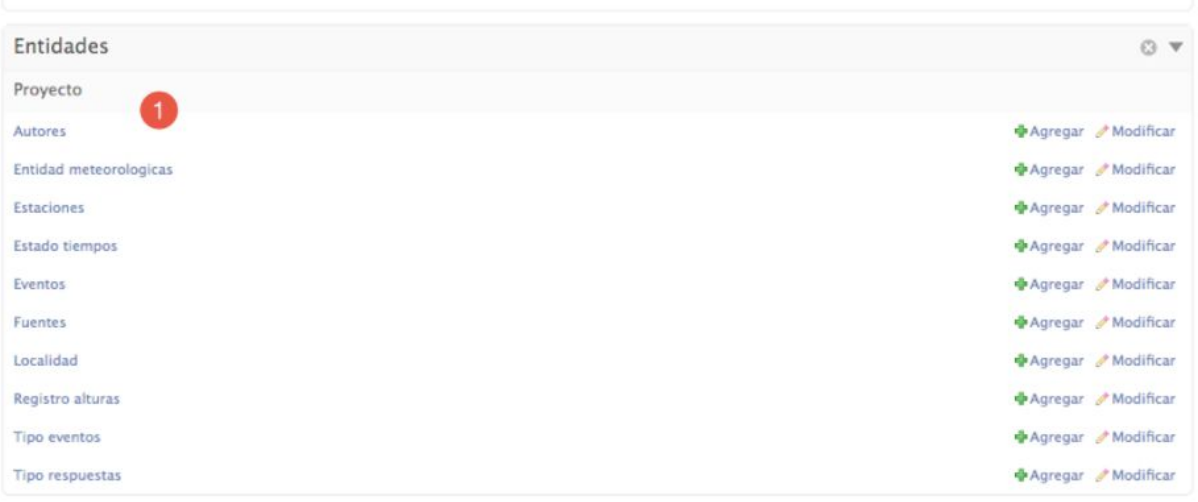
3. Pantalla de Administración

La pantalla de administración es la primera en mostrarse una vez accedido al administrador desde el botón de ir a administración mencionado anteriormente. Desde la misma se podrá acceder a todas las entidades registradas para figurar como disponibles, así como a la configuración de contraseña del usuario que se encuentra logueado.



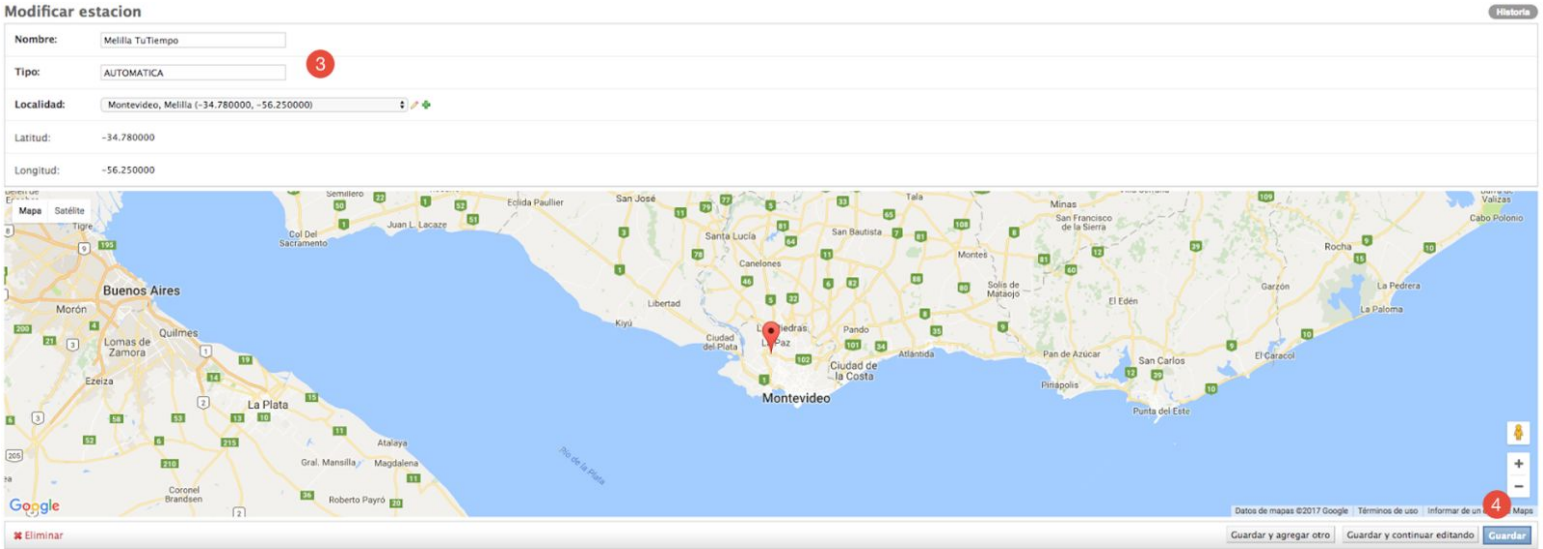
3.1. Ver/Editar objeto

Desde la pantalla de administración, es posible acceder a los datos almacenados como lo son las estaciones, localidades, eventos, estados del tiempo, registros de altura, etc.
Para esto se debe:



1) Clickear en la entidad (tabla) a la que se desea acceder a sus datos.

2) A continuación, se desplegará un índice con todos los elementos almacenados para esa entidad, el mismo presenta varias columnas de información. Para acceder a un objeto en particular, se debe clickear en la columna identificadora del mismo (en el ejemplo de la imagen, el nombre)



3) En esta pantalla, se puede ver la información de la estación, pudiéndose modificar los campos que se encuentran editables.

4) Finalmente, si se han realizado modificaciones, las mismas se guardan clickeando en el botón de Guardar en la esquina inferior derecho.

El procedimiento si se desea agregar algún elemento en alguna entidad, es similar. Clickeandose en el botón de editar disponible tanto en los índices de las entidades como a la derecha de las entidades en el panel principal de la pantalla de Administración.

3.2. Cerrar Sesión

Para cerrar la sesión en el sistema, deberá:

1) Clickear en el botón “Cerrar Sesión” ubicado en la esquina superior derecha de la pantalla de administración, y será redirigido a la pantalla de Inicio de Sesión