Facultad de Ingeniería,
Universidad de la República

Enfoque integrado para la minería de procesos y datos organizacionales

Maestría en sistemas de información y tecnologías de gestión de datos

Centro de Posgrados y Actualización Profesional en Informática.
Instituto de Computación, Facultad de Ingeniería, Universidad de la República

Autor

Darío Martín Rubio Mata

Tutores

Andrea Delgado, Libertad Tansini

Montevideo, Uruguay, Setiembre 2022

Índice

Índice de figuras

Índice de tablas

Resumen

En las últimas décadas, los sistemas de información han cobrado un rol preponderante en las organizaciones. Mediante el uso de estos sistemas buscan el apoyo a la operativa y gestión que les permiten llevar a cabo sus objetivos. Cada uno de estos sistemas producen un gran volumen de datos en forma constante, estos datos son heterogéneos en cuanto a su formato y además provienen de diversas fuentes. A la hora de la toma de decisiones basada en evidencia es necesario transformar este enorme conjunto de datos en información valiosa, y para esto hace falta la utilización de herramientas que ayuden a esta tarea.

La Ciencia de Datos enmarca un gran número de disciplinas que permiten la gestión, análisis y descubrimiento de la información a partir de grandes volúmenes de datos. Dentro de estas disciplinas se encuentran la minería de datos y la minería de procesos.

Estas disciplinas suelen considerarse en forma independiente dependiendo de cuál es la fuente de los datos, si son datos generados por los sistemas tradicionales generalmente se utiliza la minería de datos, y para los datos asociados a ejecución de procesos de negocio se utiliza la minería de procesos. Pero la realidad de las organizaciones es única, sin importar de donde provengan los datos, esta división puede generar visiones parciales a la hora de obtener resultados a partir del análisis.

Por esto surge la necesidad de poder tener una visión integral de la información y para esto se necesita contar con las herramientas que permitan realizar el análisis de forma integrada. En este trabajo se realiza una investigación acerca de los avances existentes con respecto a esta concepción unificada de la minería de datos y la minería de procesos para luego realizar una definición general de la problemática. A su vez, se define una propuesta que permite crear las herramientas que brinden la posibilidad de tener esta visión integral. Mediante la implementación de un prototipo y la realización de validaciones mediante un caso de estudio se prueba la validez y factibilidad de una propuesta con dichas características.

1. Introducción

En las últimas décadas, se ha producido una democratización del uso de sistemas de información que dio lugar a la digitalización de las organizaciones y de la sociedad en general. Estos sistemas se encargan de cubrir las necesidades apoyando a la gestión de los datos que permiten a las organizaciones modelar su realidad y cumplir con sus propósitos. Las organizaciones se encuentran produciendo datos en forma constante mediante transacciones, compras, emails, imágenes y una innumerable diversidad de datos estructurados y no estructurados. Llevar a cabo la tarea de transformación de estos datos en información no es una tarea trivial, el volumen y la heterogeneidad de los mismos suelen ser de tal magnitud que resultaría imposible de asimilar e interpretar por una persona. Es por esto que se han desarrollado técnicas que permiten gestionar y transformar ese gran volumen de datos en información de manera sistematizada, permitiendo obtener indicadores, relaciones y patrones relevantes que ayuden a la toma de decisiones basada en evidencia.

La Ciencia de Datos [21; 3] es un campo interdisciplinario que ha emergido en los últimos años como respuesta a la problemática de la gestión, análisis y descubrimiento de información en grandes volúmenes de datos, generados a gran velocidad y con gran variedad (las tres V) [17] considerando también la veracidad [28]. Entre las disciplinas que componen a la Ciencia de Datos se encuentran la minería de datos [19] y la minería de procesos [3].

La minería de datos [19] es una disciplina madura y consolidada que tiene por objetivo analizar grandes volúmenes de datos buscando reglas generales y patrones de comportamiento, así como proveer predicciones con base en los datos de entrada. Las metodologías de modelado utilizadas pueden ser supervisadas (existe un conjunto de datos etiquetados del cual aprender las respuestas correctas) o no supervisadas (no existen dichas etiquetas). Las técnicas de minería de datos son clasificadas como descriptivas i.e. describen el conjunto de datos bajo análisis, o predictivas i.e. proveen predicciones sobre nuevos datos con base en datos existentes. El foco de la minería de datos ha sido tradicionalmente los datos organizacionales gestionados en

sus sistemas informáticos de soporte.

Por otro lado, la minería de procesos [3] es una disciplina relativamente reciente que utiliza técnicas de minería de datos y otras, desde una perspectiva de procesos. Un proceso de negocio es un conjunto de actividades o tareas que son llevadas a cabo por las organizaciones en un entorno organizacional y técnico, para cumplir con sus objetivos [35]. El objetivo de la minería de procesos también es extraer conocimiento a partir de los datos, pero utilizando como fuente de datos las trazas generadas por la ejecución de los procesos en su entorno operativo. La minería de procesos provee tres perspectivas de análisis: i) descubrimiento de modelos desde las trazas de ejecución (instancias) del proceso; ii) conformidad de modelos i.e. chequear relación entre modelos y trazas de ejecución; y iii) enriquecimiento de los modelos (enhance) i.e. agregando información descubierta de la ejecución. También es posible realizar análisis de performance de la ejecución de los procesos, como cuellos de botella, duración de las instancias, tiempos promedios, etc.

1.1. Marco del proyecto

Esta tesis fue realizada en el marco del proyecto "Minería de procesos y datos para la mejora de procesos en las organizaciones", financiado por la Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, Uruguay [13]. En este proyecto se presenta un framework que integra minería de procesos y datos, técnicas y algoritmos para el análisis de la ejecución de procesos y datos, integración de datos organizacionales y datos de procesos, calidad de datos y cumplimiento de procesos. Este marco multidisciplinario tiene como objetivo proveer soporte a la mejora basada en evidencia en las organizaciones. El framework propuesto se denomina PRICED (Process and Data sCience for oRganIzational improvEment) donde se definen tres grandes fases: de Ejecución (Enactment), de Datos (Data) y de Minería (Mining), donde los datos son registrados, extraídos, integrados, depurados, se aplica calidad de datos, realizando finalmente minería de procesos y datos.

En la fase de ejecución es donde operan los diferentes sistemas que pertenecen a las

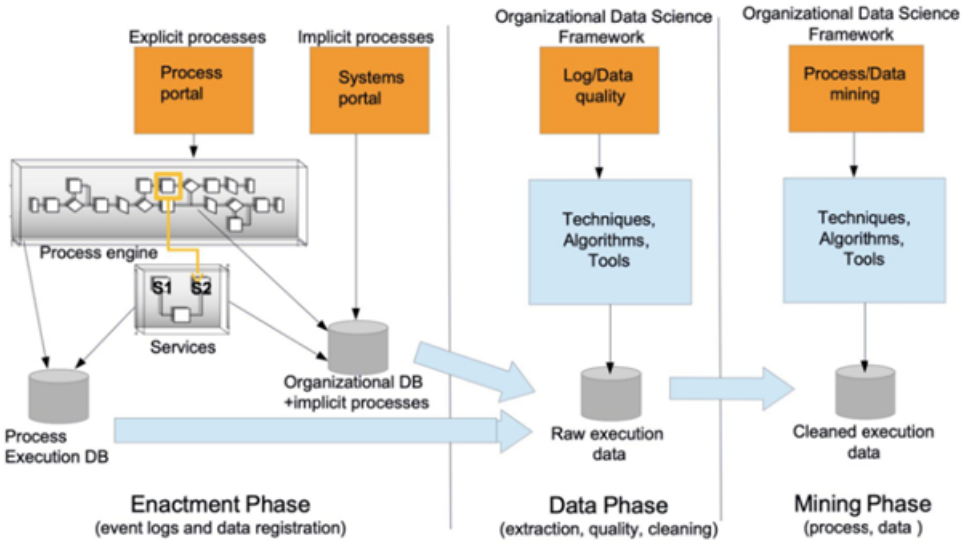

Figura 1: Fases del framework PRICED [13]

organizaciones, se pueden clasificar principalmente en dos clases: por un lado, los sistemas basados en procesos, donde se tienen explícitamente modelados los procesos de negocio y, por otro lado, están los sistemas de información tradicionales donde suele registrase la operativa transaccional. Más allá de la distinción entre los dos tipos de sistemas, existe una alta correlación entre los datos generados por ambos, ya que modelan una misma realidad para la organización. Comúnmente esta correlación se da de manera implícita, lo que dificulta el posible análisis o visión integrada de ambas fuentes de datos.

La fase de datos, enmarca la preparación de los datos con el objetivo de ser utilizados para la siguiente etapa. Esta preparación consiste en la extracción de los datos desde las fuentes. Una vez obtenidos estos datos con el formato correcto, comienza una etapa de evaluación de calidad de los mismos, de forma de eliminar o minimizar la presencia de datos que perjudiquen en el análisis que será realizado posteriormente.

Finalmente, en la última etapa del framework propuesto se encuentra la fase de minería, el objetivo de esta fase es brindar a las organizaciones una visión general de los datos obtenidos de su operación con una visión integral. Esta visión integral se basa en utilizar en forma complementaria las técnicas de minería de procesos y minería de datos. Para llevar a cabo esta visión integral resulta indispensable contar

con los datos de las fases anteriores, ya que proporcionan una visión integrada de los mismos en un formato que facilita su análisis y además cuentan con una calidad de datos que permite obtener conclusiones confiables.

Esta tesis se posiciona en la tercera fase del framework, es decir, en la fase de Minería de los datos.

1.2. Motivación

Las organizaciones cuentan con gran variedad de sistemas de información, cada uno de estos sistemas cumple un rol en la organización mediante la administración y generación de datos, apoyando a la operativa diaria. Si bien a priori estos sistemas pueden considerarse de forma independiente, modelan y reflejan muchas veces una realidad única pero con diferentes visiones. Por esto, en la toma de decisiones basada en evidencia es importante tomar en cuenta la visión más general posible y para esto es necesario poder integrar las diferentes perspectivas de los datos. La recopilación, limpieza e integración de los datos son etapas esenciales para poder analizar los datos, estas fases están contempladas en el framework descrito anteriormente, por lo tanto, para esta tesis se asumen dadas.

A la hora de analizar los datos y transformarlos en información es necesario contar con técnicas y herramientas adecuadas para realizar esta tarea. Tanto la minería de procesos como la minería de datos, suelen aplicarse en forma independiente y con objetivos finales disjuntos. Sin embargo, tanto las trazas de procesos como los datos organizacionales generados por los sistemas tienen una alta correlación, pertenecen a un mismo contexto operativo que refleja la realidad de las organizaciones, por este motivo resulta de gran importancia contar con herramientas que permitan analizar los datos desde ambas perspectivas en forma conjunta.

Actualmente, el número de artículos y propuestas existentes que contemplen este análisis de los datos en forma integral es escaso y, por lo tanto, lo son las herramientas que faciliten esta tarea. Este trabajo busca proponer un enfoque que permita

llevar a cabo el análisis desde las dos perspectivas de forma integral, brindando las herramientas necesarias para realizarlo.

1.3. Objetivos

El objetivo de esta tesis es definir una propuesta para el análisis de datos integrados de procesos y organizacionales como parte del framework para ciencia de datos organizacional, integrando técnicas de minería de procesos y datos para analizar la ejecución de procesos con algoritmos y herramientas adecuados para la mejora de las organizaciones basada en evidencia.

Como objetivos específicos se definieron:

- **(O1)** Generar un relevamiento del estado del arte en la temática de frameworks integrados para la minería de procesos y datos.
- **(O2)** Evaluar, analizar y seleccionar técnicas de las existentes y/o definir nuevas para el análisis y obtención de información de valor asociada a la ejecución de procesos y datos en las organizaciones.
- **(O3)** Evaluar, analizar y seleccionar herramientas existentes (con preferencia de tipo open source o shareware) que soporten las técnicas existentes identificadas como clave en el objetivo anterior.
- **(O4)** Definir estrategia general para el enfoque integrado del framework y las técnicas existentes o nuevas para la obtención de información de valor asociada a la ejecución de procesos y datos en las organizaciones, incluyendo metodologías y herramientas necesarias.
- **(O5)** Aplicar el framework de ciencia de datos organizacional definido en (O4) en casos de estudio de organizaciones reales para validar la propuesta e identificar oportunidades de mejora.

1.4. Publicaciones de la tesis

Una revisión sistemática que fue resultado de esta tesis fue publicado [32] y presentado en el track ESELAW (Experimental Software Engineering Track) en la Conferencia Iberoamericana de Ingeniería de Software (CIbSE) edición año 2021. Actualmente, se está trabajando para la presentación del plug-in desarrollado en conferencias asociadas al área.

1.5. Estructura del documento

El documento se compone por ocho capítulos, donde en el capítulo 2 se describen conceptos previos utilizados a lo largo del documento. En el capítulo 3 se lleva a cabo una revisión sistemática de forma de obtener el estado del arte en la problemática planteada. En capítulo 4 se realiza un planteo formal del problema para el análisis de datos organizacionales y de procesos y una propuesta de solución, la cual es llevada a la práctica mediante una implementación en el capítulo 5. Se realiza el análisis de un caso de estudio en el capítulo 6. En el capítulo 7 se plantea el análisis de los resultados obtenidos en base a los objetivos planteados y finalmente en el capítulo 8 se describen las conclusiones y trabajo a futuro respectivamente.

2. Marco teórico

En esta sección se describen los principales conceptos y definiciones que se utilizan a lo largo del informe. Se presentan conceptos acerca de la Ciencia de Datos donde existe un gran número de disciplinas que surgieron de ámbitos diferentes. Muchas comparten definiciones, metodologías y algoritmos. Dentro de esas disciplinas se encuentran la minería de datos y minería de procesos. En los apartados 2.1 y 2.2 se describen los principales aspectos de las mismas.

2.1. Minería de datos

La minería de datos [19] proviene de una comunidad fuertemente ligada a las bases de datos, que busca extraer información desconocida a partir de grandes volúmenes de datos. Si bien no existe un consenso en cuanto a la definición conceptual de la disciplina, comúnmente la minería de datos no es considerada de forma aislada. Sino que es necesario abarcar más allá de los límites de la extracción de la información. También se toman en cuenta aspectos que anteceden a dicha obtención de información desconocida, como lo son la gestión de datos o el preprocesamiento de los mismos (por ejemplo, mediante depuración, mediciones de calidad, etc.). A su vez, también se consideran aspectos posteriores, como por ejemplo el post-procesamiento de los resultados y visualización, por mencionar algunos.

Suele definirse a la minería de datos como un proceso de descubrimiento de patrones a partir de los datos, este proceso en algunas ocasiones puede ser automático o semiautomático, este último es el más habitual [36]. Estos patrones que se busca determinar tienen como objetivo convertirse en una herramienta que permita predecir comportamientos frente a nuevos datos o poder explicar características de los mismos, como por ejemplo posibles relaciones o similitudes.

La minería de datos se considera como un paso esencial dentro del proceso descubrimiento de conocimiento en bases de datos, conocido como KDD (Knowledge Discovery from Database) [16]. Este es un proceso iterativo que consta de 5 eta-

pas (Figura 2): selección, preprocesamiento, transformación, minería de datos y por último, interpretación y evaluación.

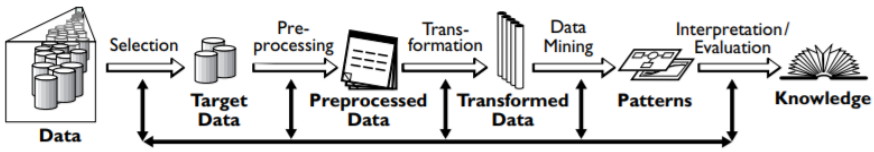

Figura 2: Etapas del proceso KDD[16]

- **Selección:** Consiste en determinar las fuentes y los conjuntos de datos que serán extraídos para ser utilizados en las siguientes etapas del proceso KDD.
- **Preprocesamiento:** Se depuran los datos seleccionados en la etapa anterior mediante reducción de ruidos, inconsistencias, etc.
- **Transformación:** En esta etapa se transforman los datos para satisfacer las necesidades de los algoritmos, facilitar la manipulación o hacerlos más comprensibles.
- **Minería de datos:** Una vez obtenidos los datos, depurados y transformados, están en condiciones de aplicar técnicas de minería de datos. Es en esta etapa donde se descubren los patrones y relaciones desconocidas para luego poder diseñar modelos genéricos que representen los datos.
- **Interpretación y evaluación:** Finalmente, una vez obtenido el modelo, se procede a una evaluación e interpretación de los resultados obtenidos. Como resultado de esta evaluación puede considerarse realizar una nueva iteración sobre algunas de las fases anteriores con el fin de mejorar los resultados.

Las primeras tres etapas es donde se preprocesan los datos, es decir que los mismos son preparados para convertirlos en los datos a minar. En la etapa de minería de datos es donde se obtienen los patrones que finalmente en la última etapa serán interpretados y sometidos a una evaluación. Generalmente, las técnicas de minería de datos parten de datos tabulares y tienen como resultado reglas de asociación, árboles de decisión, agrupaciones, entre otros.

Existen diversas formas de clasificar las diferentes técnicas de minería, a lo largo del tiempo se han creado diversas disciplinas y enfoques que utilizan muchas veces los mismos cimientos que definen a la minería de datos. Es por esto que resulta difícil establecer los límites que separan a dichas disciplinas. Por ejemplo, cuando hablamos de aprendizaje automático muchas veces se utilizan técnicas de minería de datos y en el contexto de aprendizaje automático existe una clasificación en donde se separan las técnicas en aprendizaje supervisado o no supervisado. En esta ocasión utilizaremos una de las formas más comunes de agrupar los diversos enfoques en donde se definen dos grandes conjuntos, por un lado, las técnicas descriptivas y por otro las técnicas predictivas. Entonces, basados en estos conceptos, podríamos decir que una técnica de minería, por ejemplo, es una técnica descriptiva de aprendizaje no supervisado.

En las siguientes subsecciones se describen estas clasificaciones y se muestran algunos ejemplos.

2.1.1. Técnicas descriptivas

Las técnicas descriptivas buscan relaciones que permitan caracterizar el conjunto de datos que se está analizando. Esta tarea se realiza mediante la identificación de patrones que explican o resumen los datos, extrayendo propiedades de los datos para un posterior análisis. A diferencia de las técnicas predictivas, las técnicas descriptivas no tienen como objetivo predecir nuevos datos, sino que describir los datos existentes. Una de las técnicas más utilizadas es el agrupamiento (clustering) donde se busca agrupar los datos de manera de obtener conjuntos que minimizan la distancia intra-conjuntos y maximizan la distancia inter-conjuntos, es decir, conjuntos de datos similares según algún criterio establecido (Figura 3). El clustering también se define como aprendizaje no supervisado, debido a que el conjunto de datos de entrada no necesita ser previamente etiquetado o categorizado de alguna manera. Existen diversos algoritmos de agrupamiento, siendo uno de los más populares el K-Means.

Por ejemplo, supongamos que una cadena de venta de productos pretende maxi-

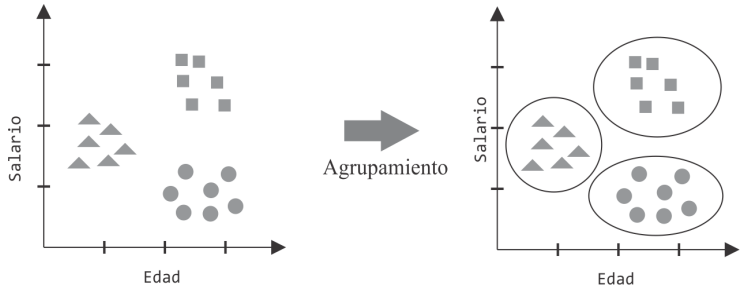

Figura 3: Agrupamiento

mizar sus ventas. Para esto se propone realizar campañas de publicidad que sean efectivas y atraigan a la mayor cantidad posible de clientes. Para llevar a cabo esta tarea, deciden realizar múltiples campañas de publicidad centrándose en diferentes segmentos de población y de esta forma se pueden definir diferentes canales de comunicación por los cuales dichos segmentos se ven atraídos de forma más frecuente. Los algoritmos de agrupamiento permiten obtener subconjuntos de elementos que tienen características similares. Estas características podrían ser la edad, los ingresos o cualquier otra característica relevante de los potenciales clientes, los conjuntos obtenidos se podrían interpretar como sectores de mercado. De esta forma, utilizando técnicas de agrupamiento se podrían obtener cuáles son los principales sectores, y conociendo las diferentes características de cada sector es posible planificar dichas campañas y maximizar los resultados.

Otro ejemplo de modelos descriptivos son los obtenidos con las reglas de asociación, donde se buscan dependencias y relaciones sobre atributos nominales. Las reglas de asociación es una de las técnicas más utilizadas y al igual que el agrupamiento se clasifica como una técnica de aprendizaje no supervisado. Una regla de asociación es una proposición probabilística [19] que define la probabilidad de ocurrencia de determinado patrón, por ejemplo, si se cumplen las condiciones X entonces, existe una probabilidad P de que ocurra Y, X =>Y. Existen dos medidas que permiten conocer la calidad de una regla, estas son la cobertura (support) y la confianza (confidence). La cobertura de una regla indica en cuantas de las instancias existentes en todo conjunto de datos se cumple la regla, y la confianza es el porcentaje de veces en que se cumple la premisa (X) de la regla, también se cumplen las conclusiones (Y).

Un uso frecuente de estas reglas de asociación se da en el sector de ventas, donde se buscan relaciones entre la venta de productos. A modo de ejemplo, supongamos nuevamente el escenario anterior. Donde una cadena de venta de productos en su tarea de maximizar sus ventas busca también poder sugerir productos que potencialmente el cliente puede estar interesado. Generalmente, se le conoce como listas de sugerencias. Mediante reglas de asociación es posible crear estas listas de sugerencias, donde se pueden obtener reglas que relacionen la compra de un producto A con una probabilidad de compra de un producto B. De esta manera, es posible establecer un orden en las góndolas de forma de tener los productos A y B juntos o también en el caso de un sistema de compras web, cuando un cliente añada el producto A a su carrito de compra, el sistema podrá sugerirle el producto B, sabiendo que generalmente los clientes compran A y B juntos.

2.1.2. Técnicas predictivas

Permiten crear modelos a partir de un conjunto de datos capaces de predecir resultados o comportamientos frente a nuevos casos. Los datos utilizados como evidencia cuentan con un atributo particular donde se materializa el resultado esperado, es decir, un valor particular que se busca predecir (clase, categoría o un valor). A las técnicas que utilizan el conjunto de datos y la salida esperada se les denomina técnicas supervisadas. Existen diversas técnicas predictivas, pero entre ellas se destacan la clasificación y la regresión.

En clasificación se busca crear modelos predictivos que permitan clasificar un conjunto de datos en clases discretas conocidas. Una vez obtenido este modelo se podrá predecir a cuál de las posibles clases discretas pertenece un nuevo dato. Dentro de las técnicas más populares de clasificación se encuentran los árboles de decisión, Random Forest o Nayve Bayes por mencionar algunos. Por ejemplo, mediante árboles de decisión es posible obtener un modelo que permita predecir si una solicitud de un préstamo será aprobada o no por un banco. Esta decisión puede depender de atributos como lo son si el solicitante ya es cliente del banco, si el cliente tiene

deuda previa, cuál es su nivel de ingresos y la antigüedad laboral. En la figura 4 se puede visualizar un árbol de decisión que describe un posible modelo para el caso mencionado. Cada nodo representa los atributos y las aristas posibles valores. Por

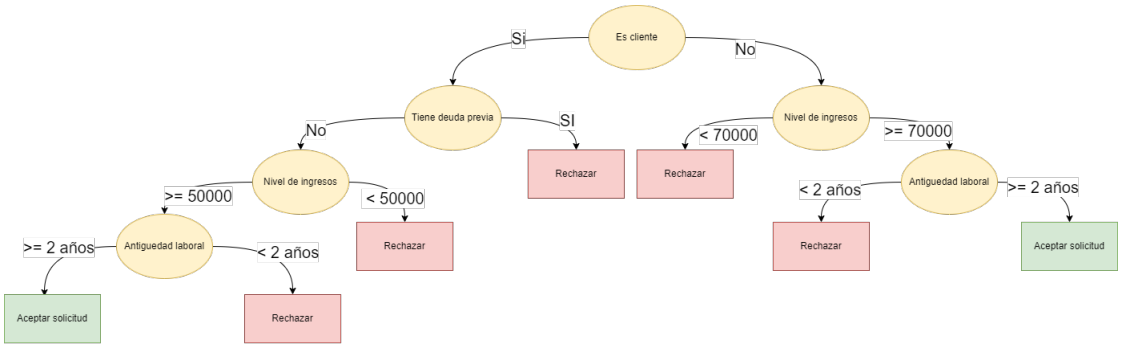

Figura 4: Árbol de decisión, solicitud de préstamo bancario.

último, las hojas representan la clase, en este caso si la solicitud será aceptada o rechazada.

Por otro lado, los modelos de regresión buscan predecir un valor no discreto (numérico) por ejemplo un número real. En clasificación se utilizan algoritmos de aprendizaje automático como árboles de decisión, redes neuronales, máquinas de soporte vectorial (support vector machine), K-vecinos más cercanos (K-nearest neighbours), entre otros. Las metodologías de regresión permiten obtener modelos que predicen valores continuos, por ejemplo el precio de un inmueble según metros cuadrados.

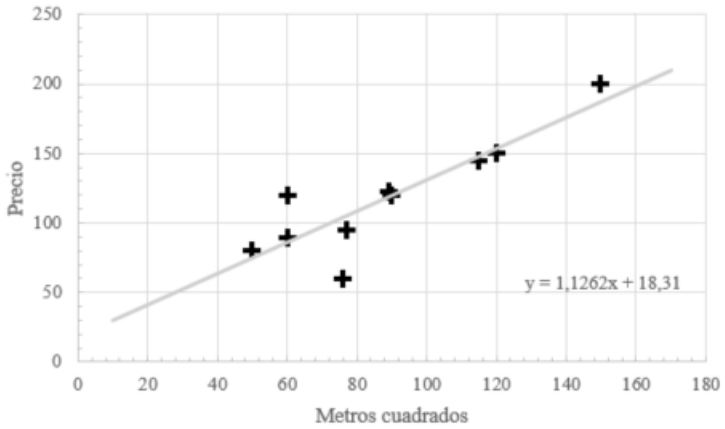

Figura 5: Ejemplo de regresión para el cálculo del precio de un inmueble.

Existen múltiples herramientas que ofrecen diversos algoritmos de minería de datos, una de las más populares es Weka [14] debido a que es software libre (GPL) y es

ampliamente utilizado en estudios académicos. Suele considerarse a Weka como la herramienta que cuenta con el estado del arte de los algoritmos de minería conocidos, ya que provee implementaciones de una gran cantidad de algoritmos. Además, ofrece las herramientas necesarias para visualizar y preparar los conjuntos de datos a los cuales se les desea aplicar minería. Una vez aplicado alguno de los algoritmos disponibles, también cuenta con facilidades de visualización que la convierten en una herramienta interactiva y fácil de usar. R [1] es otra de las herramientas más importantes que abarca aspectos de minería de datos, también tiene un amplio uso a nivel estadístico. Otro ejemplo es RapidMiner[2], que también tiene versiones de libre acceso y versiones de pago.

2.2. Minería de procesos

La minería de procesos [3] surge en comunidades ligadas a procesos, donde se usan algunas bases de minería de datos para analizar datos más específicos de la ejecución de los procesos. Para las organizaciones resulta esencial conocer sus procesos, poder medir la calidad de los mismos, encontrar posibles fallas o puntos de mejora y la minería de procesos es una herramienta que permite llevar a cabo esta tarea. Un proceso de negocio es un conjunto de actividades o tareas que son llevadas a cabo por las organizaciones en un entorno organizacional y técnico, para cumplir con sus objetivos [35]. Formalmente, es posible encontrar diferentes definiciones de procesos de negocio, por ejemplo [20] define un proceso de negocio como:

Un proceso de negocio es un conjunto de actividades, que impulsadas por eventos y ejecutándolas en una cierta secuencia, crean valor para un cliente (interno o externo).

Estos procesos pueden haber sido diseñados e implementados de manera explícita, y su ejecución se realiza utilizando herramientas BPMS (Business Process Management Systems) [11], o también existen los denominados procesos implícitos, donde no hay un diseño claro e intencional del mismo, pero en la práctica es posible distinguir un conjunto de actividades, que se ejecutan en forma cronológica y cumplen con las definiciones de un proceso de negocio.

En ambos casos, como resultado para las ejecuciones de los procesos, es posible obtener logs con datos que pueden ser una fuente de análisis. Estos datos tienen particularidades especiales, como por ejemplo un orden cronológico, tareas definidas o usuarios y roles asociados a las tareas, por mencionar algunas. La minería de procesos fue diseñada con el fin de contemplar estas particularidades y de esta forma proveer técnicas específicas que permiten extraer información a partir de esos datos de ejecuciones.

Existen múltiples herramientas que implementan diversos algoritmos de minería de procesos, una de las más populares es ProM[1], debido a que es software libre (GPL) y es ampliamente utilizado en estudios académicos ya que permite gran flexibilidad para incorporar nuevos plugins que implementen nuevos algoritmos o variaciones de algoritmos existentes. Otras herramientas son Disco[2] y Celonis[3] que cuentan con opción académica y de pago.

El principal insumo para la minería de procesos es el log de eventos (event log) resultantes de la ejecución de los distintos procesos. Un log de eventos corresponde a un solo proceso y se compone de casos (instancias o trazas) únicos y cada caso se compone a su vez de una secuencia de eventos. Los eventos dentro de un log deben estar ordenados cronológicamente y además cada evento cumple con las siguientes propiedades: i) cada evento refiere a una única actividad ii) cada evento pertenece a un único caso iii) cada evento tiene asociado un recurso que lo originó y iv) cada evento tiene una marca de tiempo (timestamp) asociada a su ejecución.

En la figura 6 se muestra una representación de la estructura de un log de eventos.

Partiendo de estos logs de eventos se utiliza la minería de procesos [3] para descubrir conocimiento, con tres enfoques: descubrimiento (discovery), conformidad (conformance) y enriquecimiento (enhancement).

- **Descubrimiento**: a partir de las trazas en el log de eventos se infiere de

[1]http://www.promtools.org/
[2]https://fluxicon.com/disco/
[3]https://www.celonis.com/

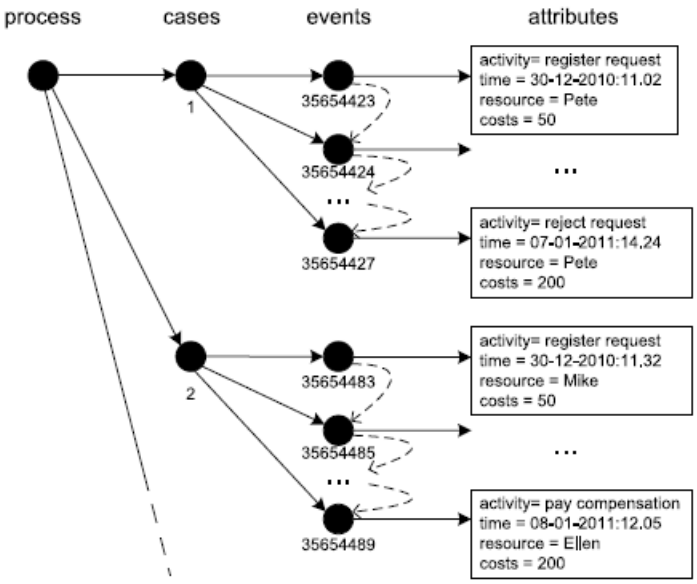

Figura 6: Estructura del log de eventos[3]

forma automática un modelo que represente dichas ejecuciones, por ejemplo un modelo especificado en una red de Petri[26] o BPMN[27]. Existe una variedad de algoritmos de descubrimiento con distintos enfoques (heurísticas, algoritmos evolutivos, etc.) y que han sido mejorados y extendidos en las últimas décadas [6].

- **Conformidad**: Estas técnicas permiten comparar un modelo de proceso (descubierto, existente) contra la ejecución real en el log de eventos. Mediante distintos enfoques, (por ejemplo: replay del log de eventos sobre el modelo, alineación de segmentos) permite conocer la concordancia entre el modelo y la ejecución, con indicadores como cobertura y precisión.

- **Enriquecimiento**: El objetivo es extender o mejorar un proceso utilizando la información de las trazas. Por ejemplo, utilizando las marcas de tiempo se puede conocer cuáles son las actividades que toman más tiempo y qué pueden significar un cuello de botella. También se pueden descubrir patrones de interacción entre las personas o recursos que realizan las actividades.

2.2.1. Estándar XES

Como se menciona anteriormente, los logs de procesos son los insumos principales para la minería de procesos. Estos conjuntos de datos son representados utilizando el estándar XES (IEEE 1849-2016) [18], donde se define una gramática basada en etiquetas que permite tener una visión unificada y extensible para este tipo de datos. La estructura de los archivos XES se describe mediante un esquema XML (XMLS). En la figura 7 se muestran los principales elementos que componen el estándar y sus relaciones.

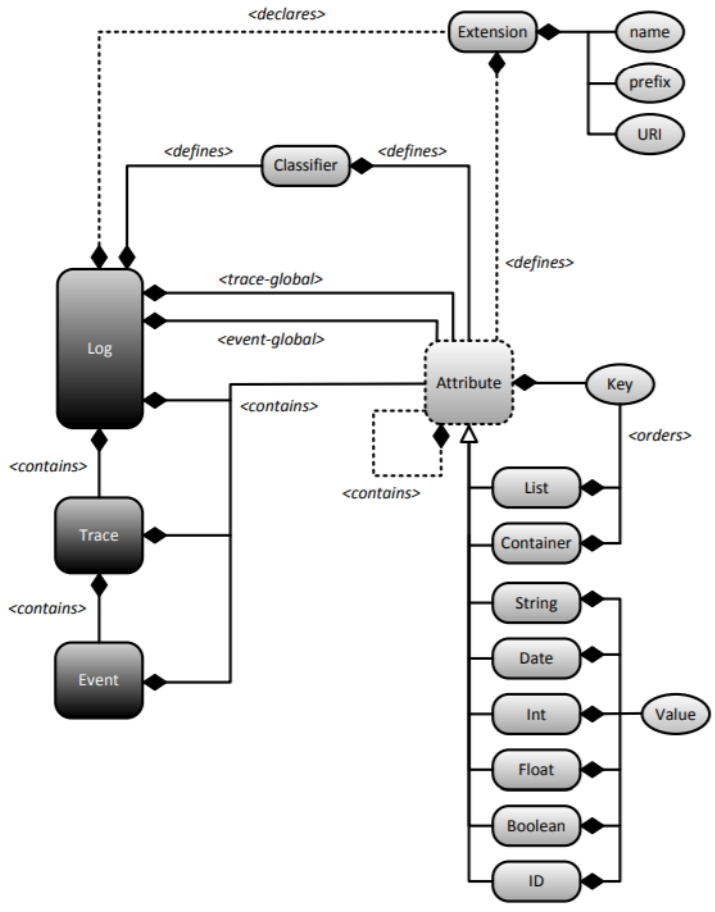

Figura 7: Estándar XES

El elemento principal dentro del un log con formato XES, es el elemento Log. Este elemento a su vez está compuesto por trazas y las trazas por eventos.

A su vez, cada uno de estos elementos puede tener atributos que faciliten y permitan

incorporar información acerca de los mismos. Estos atributos pueden ser de diferentes tipos de datos, como por ejemplo números enteros, cadena de texto (string), fechas, entre otros.

También se puede apreciar la posibilidad de extensión que brinda a nivel de atributos. Mediante estas extensiones se permite la posibilidad de modelar realidades específicas dado algún contexto particular.

2.3. Evaluación de la calidad de los modelos

Tanto en minería de datos como en minería de procesos es necesario evaluar la calidad de los resultados obtenidos. Dependiendo del tipo de modelo existen diferentes medidas de evaluación que permiten conocer características de interés.

En minería de datos, para tareas de clasificación medidas habituales son la ***precisión*** [19] y ***cobertura (recall)***. La precisión se calcula[4] como el número total de instancias del conjunto de pruebas clasificadas correctamente dividido el número total de instancias del conjunto de pruebas clasificadas como positivas (Ecuación 1).

$$Precision = \frac{VP}{VP + FP} \tag{1}$$

La cobertura (recall) se define como el número de instancias a las que una regla se aplica y predice correctamente (Ecuación 2).

$$Recall = \frac{VP}{VP + FN} \tag{2}$$

Para las técnicas de regresión, la evaluación más habitual es mediante el error cuadrático del valor predicho frente al valor esperado (ver Ecuación 3, donde x_i es el valor predicho y y_i es el valor esperado).

En tareas de agrupamiento, las medidas suelen considerar la separación inter-grupos

[4]VP=Verdadero positivo, FP= Falso positivo, FN= Falso negativo

y la cercanía intra-grupos.

$$\sum_{i=1}^{D} |x_i - y_i| \tag{3}$$

En minería de procesos son cuatro los criterios de evaluación que se utilizan habitualmente: ***adecuación (fitness o recall)***, ***simplicidad (simplicity)***, ***precisión (precision)*** y ***generalización (generalization)*** [3]. La adecuación de un modelo, es la capacidad de poder reproducir el comportamiento observado en el log de eventos. La simplicidad busca representar el nivel de simplicidad del modelo. La precisión es la capacidad de ajustarse solo a las instancias similares a las presentes en los logs. La generalización mide la capacidad de generalizar el comportamiento al soportar ejemplos nuevos no presentes en los logs utilizados para obtener los modelos.

Si solo se permite el comportamiento observado, está sobre-ajustado (overfitting), o sub-ajustado (underfitting) si permite mucho más comportamiento sin real soporte. Esto también se observa en minería de datos.

3. Estado del arte

Uno de los objetivos específicos (O1) definidos fue el relevamiento del estado del arte para conocer propuestas y estudios existentes que tomen en cuenta a la minería de procesos y la minería de datos en forma integrada. A la hora de recolectar esta información es importante evitar sesgos y obtener estudios de calidad, por este motivo el relevamiento se realizó en forma de revisión sistemática que brinda un método explícito, estructurado, repetible donde se minimizan posibles sesgos a la hora de la recolección y resumir la información existente. El procedimiento de revisión sistemática seguido se basa en los lineamientos propuestos por Kitchenham [22][23], donde se propone una metodología que consta de tres grandes etapas: planificación, ejecución y análisis.

A continuación, en las secciones 3.1 y 3.2 se presentan las etapas de planificación y ejecución respectivamente, mientras que el análisis y descripción de resultados se realiza en la sección 3.3.

3.1. Planificación

En esta fase de la metodología se definen los lineamientos generales de la revisión, como son la formulación de la pregunta de investigación y la definición de los criterios de inclusión y de exclusión para determinar los estudios primarios. También en esta etapa se definen las fuentes que se utilizarán para realizar las búsquedas y el periodo de tiempo para el cual se realizarán las búsquedas. La necesidad de la revisión sistemática se identifica en el contexto de la definición del framework en desarrollo, que fue delineado en [13]. El resultado de esta revisión permite identificar elementos clave en la integración de la minería de procesos y la minería de datos a ser considerados en el framework.

3.1.1. Pregunta de investigación

Una de las actividades más importantes en una revisión sistemática es la formulación de la pregunta de investigación [22]. Por este motivo, la pregunta de investigación busca representar el objetivo de investigación que involucra el presente estudio, y se define como:

¿Qué propuestas de frameworks y/o enfoques de aplicación integrada de minería de procesos y minería de datos existen para brindar soporte al análisis de datos e inteligencia del negocio basada en evidencia en las organizaciones?

Partiendo de la pregunta de investigación, se diseñó la cadena de búsqueda de forma de obtener la mayor cantidad de resultados, si bien esto puede generar falsos positivos, se reduce la posibilidad de no tener en cuenta estudios relevantes. Para esto se realizó un proceso iterativo, donde en cada etapa se buscó refinar y evaluar la cadena ejecutada en las distintas fuentes. Para la evaluación de estas etapas se utilizaron algunos estudios conocidos como grupo de control, y de esta forma medir la performance de la búsqueda en cada iteración. Se utilizaron los términos específicos de minería de procesos y minería de datos. No agregamos Ciencia de Datos o inteligencia de negocio, ya que estos últimos actúan de "paraguas", incluyendo otras disciplinas que no interesan en el contexto de esta revisión y aumentarían innecesariamente la cantidad de resultados no relacionados. La cadena de búsqueda fue definida como:

"process mining"
AND "data mining" AND ("business process" OR "business process management")
AND framework

Las palabras clave "business process" y "business process management" fueron in-

cluidas, ya que únicamente son de interés los enfoques de minería de datos que aplican o incluyen datos de procesos de negocio, no así los enfoques tradicionales que operan únicamente sobre datos organizacionales. El término framework se incluye en la cadena de búsqueda para enfocar los resultados en ese sentido, pero resultados que presenten propuestas menos abarcativas o no completamente definidos no son excluidos, por ejemplo con foco metodológico.

3.1.2. Criterios de inclusión y exclusión

Los criterios de inclusión definidos incluyen: i) estudios posteriores al año 2000; ii) estudios escritos en inglés; iii) estudios con aplicación integrada de minería de datos y minería de procesos (frameworks, metodologías). Los criterios de exclusión definidos incluyen: i) estudios que solo tratan sobre minería de procesos o solo tratan sobre minería de datos; ii) estudios no accesibles en formato digital.En la Tabla 1 se presentan los criterios de inclusión y exclusión definidos.

Tabla 1: Criterios de inclusión y exclusión

Criterios de Inclusión	Criterios de Exclusión
Estudios con aplicación integrada de minería de datos y minería de procesos	Estudios que solo tratan de minería de procesos o solo de minería de datos
Estudios escritos en Inglés	Estudios no accesibles en formato digital
Estudios posteriores al año 2000	

3.1.3. Fuentes y período de búsqueda

Para la revisión se utilizaron cinco fuentes principales de artículos científicos que permiten acceso web a los mismos, seleccionadas entre las fuentes generalmente utilizadas y su accesibilidad. Estas fuentes fueron seleccionadas en base a su cubrimiento de publicaciones internacionales tanto de conferencias como revistas de primer nivel en el área: Springer, IEEE Xplore, Scopus, y Science direct, que se muestran en la Tabla 2.

Las búsquedas se realizaron durante los meses de noviembre y diciembre de 2019,

Tabla 2: Fuentes de búsqueda

Fuente	**Sitio web**
Springer	http://www.springer.com/
IEEE Xplore	https://ieeexplore.ieee.org/
Scopus	http://www.scopus.com/
Science Direct	http://www.sciencedirect.com/

cubriendo el período 2000 a 2019, al inicio del trabajo de investigación para analizar el estado del arte en la temática. En marzo de 2021 se actualizaron las búsquedas con el fin de renovar los resultados y contrastar avances.

3.2. Ejecución

La ejecución de la revisión consistió en llevar a cabo las etapas mencionadas anteriormente, lo que se describe a continuación.

3.2.1. Proceso de selección de estudios

En la primera etapa de recuperación se ejecutaron las consultas en las diferentes fuentes, realizando las adaptaciones necesarias, obteniendo inicialmente 2868 artículos. En segundo lugar, se procedió a la lectura de títulos, resúmenes y palabras clave de cada uno para aplicar los criterios de inclusión/exclusión definidos, y además descartar duplicados, resultando en un total de 32 estudios seleccionados.

En una tercera etapa se llevó a cabo la lectura detallada de los estudios aplicando los criterios de inclusión/exclusión al texto de los artículos relevantes. En esta última etapa se descartaron 22 estudios, obteniendo como resultado de la revisión 10 estudios primarios que contestan la pregunta de investigación planteada. En la tabla 3 se muestran las cantidades de estudios obtenidas en cada fuente seleccionada, relevantes y primarios.

El criterio de exclusión por el cual se descartaron más estudios fue el enfoque, donde se descartó un total de 9 estudios relevantes. Luego se descartaron 6 estudios

Tabla 3: Resultados obtenidos en las distintas fuentes

Fuente	Encontrados	Relevantes	Primarios
Springer	859	15	7
IEEE Xplore	40	0	0
Scopus	1798	12	2
Science Direct	171	5	1
Total	2868	32	10

porque solo trataban a la minería de procesos, seguido de 5 estudios debido a que solo trataban a la minería de datos. Los restantes dos documentos descartados se debieron a que no estaban accesible en las fuentes utilizadas.

3.3. Análisis de resultados

En esta sección se presentan los resultados obtenidos, en primer lugar la descripción general y en segundo lugar las categorías analizadas. En la Tabla 4 se presentan los estudios primarios y sus referencias en el marco de su análisis.

3.3.1. Descripción de resultados

Las consultas realizadas en las fuentes retornaron un total de 2868 artículos, incluyendo artículos duplicados, de los cuales se identificaron los 10 estudios primarios que se presentan en esta sección. A pesar de cubrir el período 2000 a 2020 con las búsquedas realizadas, los estudios relevantes seleccionados inician a partir del año 2008. En la Figura 8 se presentan los estudios primarios por año de publicación.

La baja cantidad de trabajos primarios integrando las dos áreas confirma lo reciente de la integración de las disciplinas de minería de datos tradicional con la más nueva minería de procesos, siendo un área de investigación incipiente.

Los estudios primarios obtenidos fueron analizados desde cinco perspectivas que surgen a partir del contenido de las propuestas. Estas categorías son: los tipos de

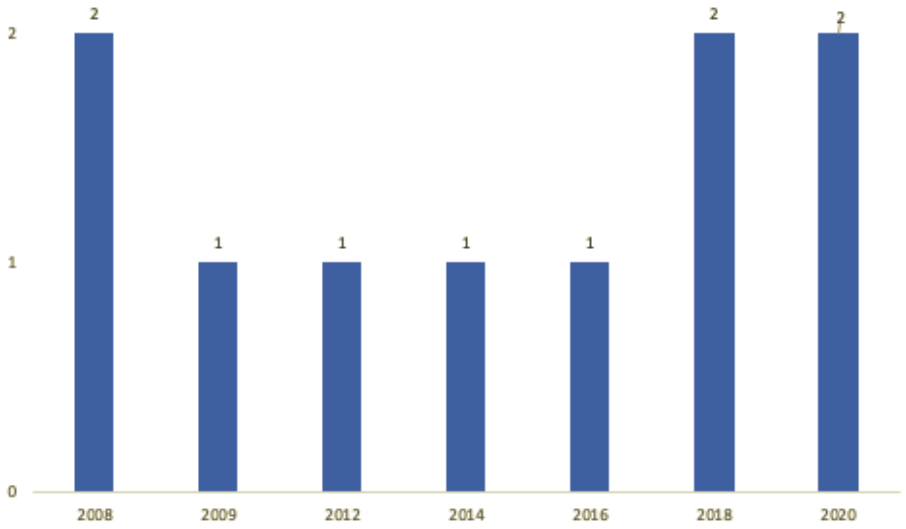

Figura 8: Años de publicación de los estudios primarios

propuestas (framework, metodología), la presencia y origen de los algoritmos utilizados (minería de procesos, minería de datos), la integración de los datos utilizados (log de eventos, datos organizacionales), tipo de minería (descriptiva, predictiva) y aspectos relacionados a la calidad de datos en los enfoques de minería utilizados. En la Tabla 4 se presentan los estudios primarios y su aporte en cada una de las categorías analizadas.

3.3.2. Tipo de propuesta

Las propuestas de tipo metodológico pueden ser más concretas y específicas para algún tipo de escenario particular, donde puede enfocarse solo en algunos aspectos del ciclo de vida de los procesos y datos. Tanto en [33] como en la primera parte de la metodología planteada en [15], el foco está centrado en logs de eventos de los procesos, con foco en agrupamiento de trazas. Con otra perspectiva, [10] utiliza algunos datos del contexto e información de procesos para crear un modelo de estimación de calidad. En esta propuesta se presenta una primera etapa donde crean modelos que permiten predecir qué camino tomará una instancia en ejecución. Por otra parte, [24] utiliza técnicas propias de data mining aplicadas directamente sobre logs de eventos de procesos, vinculado al ámbito de datos provenientes de una institución educativa.

Tabla 4: Análisis de estudios primarios

Estudio	**Propuesta**		**Algoritmos**		**Integración**		**Minería**		**Calidad**
	F	M	PM	DM	Logs	Datos	D	P	
van Zelst et. al. [38]	●		●	●	●	●	●		
Križanić, S. [24]		●		●	●	●	●		●
Roudjane et al. [31]	●			●	●			●	
Ezpeleta et al. [15]		●		●	●		●	●	
Teinemaa et al.[34]	●			●	●	●		●	
Bevacqua et al. [7]	●			●	●	●	●	●	
Bui et al. [9]	●			●	●				●
Song et al. [33]		●	●	●	●		●		
Radeschütz et al.[30]	●				●	●			
Cardoso J. [10]		●		●	●	●	●	●	

Las propuestas que presentan frameworks tienen un planteamiento más abarcativo y profundo, donde puede ser considerado el contexto de donde provienen los datos, se toma en cuenta todo el ciclo de vida de los procesos y datos o se proponen generalizaciones que permiten aplicarlo en diferentes entornos. En [34] se propone un framework para monitoreo de procesos en tiempo real, para esto se combinan técnicas de minería de textos y clasificación. También con el objetivo de analizar procesos en tiempo real, la propuesta de [31] busca detectar desviaciones en flujos de eventos. Por otro lado, en [7] se presenta un framework para poder predecir la performance de procesos en cuanto al tiempo de ejecución y la cantidad de actividades que se ejecutaran en el mismo. En [38] se presenta un framework con lineamientos para obtener agrupamiento de trazas basados en atributos de los procesos. En [30] se presentan lineamientos con el fin integrar datos operacionales y el log de eventos de procesos de forma manual o semi-automática. También se pueden encontrar propuestas genéricas donde no importa el escenario concreto, por ejemplo, [9] busca ampliar el espectro de algoritmos que se utilizan tradicionalmente partiendo de la

estructura XML en las que los logs se almacenan comúnmente.

3.3.3. Algoritmos

En cuanto a los algoritmos utilizados en las propuestas, se categorizaron de forma de establecer si los mismos provienen de la minería de datos o de la minería de procesos. En esta categorización se encontró que la mayoría de los trabajos que explicita el uso de algún algoritmo, utiliza algoritmos que provienen de la minería de datos [24], [31], [15], [34], [7], [9], [33], [10], [38], aplicados a logs de eventos de procesos. Principalmente, se destaca el uso de clustering y mayormente se utilizan variantes de K-Means con el objetivo de obtener conjunto de trazas para facilitar su posterior estudio.

En [24], luego de obtener clusters utilizando K-Means, aplican árboles de decisión para cada cluster con el objetivo de obtener una representación que permita un estudio más profundo de los datos. En [38] y [33] se hace mención además a algoritmos de minería de procesos, particularmente a algoritmos de descubrimiento . En ambos se descubren los modelos de procesos asociados a cada cluster de trazas para comparar su comportamiento a la luz del atributo seleccionado, en el segundo se aplica el algoritmo de minería heurística (heuristic miner).

3.3.4. Integración de datos

Los datos toman un rol primario tanto en minería de procesos como en minería de datos, ambos enfoques se basan en la utilización de datos de procesos y/u organizacionales, por lo que un aspecto importante en la aplicación de ambas minerías es buscar la forma en que se integran los datos utilizados.Como tercera forma de análisis de los estudios primarios se consideró la fuente de datos utilizada y el nivel de integración de los datos asociados a los procesos, los datos asociados al proceso y al negocio en sí i.e. organizacionales.

Los estudios [9], [15], [31], [33] sólo consideran como fuente de datos a los logs de

eventos de procesos, en cambio [7], [10], [24], [30], [34], [38] consideran en algún aspecto algunos datos asociados al contexto operacional (organizacionales). Cabe destacar que estos datos organizacionales suelen estar limitados a variables utilizadas en los procesos en sí, no considerándose en profundidad la información de negocio presente en otros sistemas. En contraposición, en [30] proponen diversas alternativas para la integración de datos organizacionales con los datos de procesos. En dicha propuesta se muestran alternativas manuales y semi-automáticas que permiten realizar esta tarea con el objetivo de integrar los datos para ser incorporados en un datawarehouse.

3.3.5. Tipo de minería

Los estudios se clasificaron según el principal enfoque de minería que utilizan: minería descriptiva y minería predictiva.

Las propuestas de minería descriptiva buscan el agrupamiento de trazas para facilitar y mejorar el análisis de los procesos. La mayoría de las aplicaciones de minería de procesos sobre logs de eventos se realizan para el descubrimiento de modelos. Las técnicas utilizadas muchas veces dan como resultado lo que se denomina procesos spaghetti, con gran cantidad de nodos y relaciones que dificulta su interpretación y extracción de conocimiento. Este problema no solo afecta a los resultados en el descubrimiento, sino que también en las aplicaciones enfocadas en la conformidad y mejora de procesos. Las propuestas en esta categoría toman logs de eventos de procesos y sus datos asociados, y buscan agruparlos de forma de obtener conjuntos más pequeños de trazas homogéneas, según algún criterio (i.e. atributo, como comportamiento, algún dato particular, etc.), que faciliten el entendimiento y obtención de resultados. En este concepto se basan seis de los estudios primarios [7], [10], [15], [24], [33] y [38]

En este caso, las propuestas utilizan la información existente para obtener modelos que permitan predecir el comportamiento frente a una nueva ejecución. Las predicciones se dividen en dos tipos, por un lado, se busca predecir cuál será el camino o

flujo que tomara un proceso y por otro la predicción de la performance que tendrá el mismo. Conocer qué camino dentro de un modelo tomará un proceso en momentos tempranos de su ejecución significa un gran valor para la toma de decisiones, permite evitar activamente cuellos de botella o anticiparse a situaciones antes de que estas sucedan. En este concepto se basan los estudios [15], [31] y [34]. De la misma manera que se puede predecir el camino que tomara una ejecución, se pueden crear modelos que permitan predecir el desempeño, por ejemplo, a nivel de costo o tiempo. En las propuestas [7], [34], [10], [31], [15] presentan diferentes técnicas que permiten llevar a cabo esta tarea.

3.3.6. Calidad de datos

Como se menciona en la sección anterior, los datos cumplen un rol fundamental en cualquier esfuerzo de minería para obtener información de interés para la organización. A la hora de trabajar con datos es clave que tengan características de calidad deseable, y para conocer la calidad de los mismos es necesario aplicar análisis de calidad de datos. Por tal motivo se buscó en los estudios relevantes la presencia de lineamientos en este sentido. Solo en [9] y en [24] se explicita como parte del framework propuesto una etapa de pre-procesamiento donde puede considerarse el estudio de la calidad. En los otros estudios se asume que los datos cuentan con una calidad aceptable como para realizar los diferentes procesamientos y análisis, y no se explicita cómo, cuándo y de qué tipo deberían ser los análisis de calidad de datos en la aplicación de las propuestas.

3.4. Amenazas a la validez

En esta sección presentamos las amenazas a la validez según [37]: construcción, interna, conclusión y externa.

Como amenaza a la validez de construcción, la selección de fuentes o la cadena de búsqueda pueden no recuperar todos los artículos relevantes. Para minimizarlas,

seleccionamos fuentes reconocidas en base a su cubrimiento de publicaciones internacionales tanto de conferencias como revistas de primer nivel en el área, y optamos por una cadena de búsqueda que maximizara la inclusión de artículos utilizando minería de procesos y minería de datos como términos clave incluyendo procesos de negocio como foco. Se realizaron búsquedas exploratorias en las fuentes y discusión entre los autores para ajustes previos.

La validez interna tiene como amenaza la reproducibilidad y el sesgo en la selección de estudios. Para minimizarlas, el protocolo fue claramente especificado, incluyendo las fuentes, la cadena de búsqueda y los criterios de inclusión / exclusión. Los resultados fueron discutidos entre los autores en cada etapa para prevenir posible sesgo en la selección.

Como amenaza a la validez externa, la cantidad y relevancia de los estudios seleccionados puede afectar la generalización de resultados. Para minimizar esta amenaza, los artículos fueron seleccionados mediante los criterios de inclusión/exclusión definidos en el protocolo y discusiones entre los autores. Si bien la cantidad de artículos seleccionados en comparación a los recuperados puede parecer baja, el foco de la revisión está en la aplicación conjunta de minería de procesos y minería de datos, que es un área incipiente de investigación, por lo que consideramos que los resultados presentados representan el estado actual en nuestro mayor conocimiento. De todas formas podría haber artículos no publicados o indexados al momento de las búsquedas que no estén incluidos.

La validez de las conclusiones se ve amenazada por la dependencia del punto de vista de los autores realizando la revisión. Para minimizar esta amenaza, la información recuperada de los artículos analizados es la indicada en cada propuesta y si en la extracción en las categorías identificadas había distintas visiones estas fueron discutidas entre los autores.

3.5. Conclusiones del relevamiento

Se realizó un relevamiento de los artículos existentes que tienen como foco a la minería de procesos y a la minería de datos. El mismo fue realizado en forma de revisión sistemática y tuvo como resultado 32 estudios relevantes, de los cuales 10 fueron seleccionados como estudios primarios. El estudio se centró en trabajos que proponían o mostraban ambos enfoques en forma complementaria e integrada. Según los resultados obtenidos, no hay un número importante de artículos publicados en la actualidad, registrándose entre uno o dos por año, lo cual confirma lo incipiente de la investigación el área.

En cuanto al contenido específico de las propuestas y el uso integrado de la minería de datos y de procesos, la integración de ambos enfoques se considera baja, ya que la mayoría de las propuestas se basan fuertemente en uno de los enfoques y utiliza el otro en alguna etapa puntual con un fin determinado y acotado. Principalmente, se detectó la aplicación de algoritmos de minería de datos sobre logs de eventos de procesos, en forma previa al descubrimiento de modelos para agrupación de trazas homogéneas, o para predicción de distintos elementos del proceso considerando las trazas en los logs de eventos.

También puede destacarse que los estudios primarios, aún en las propuestas más generales, no se tomaba en cuenta todo el ciclo de vida de los datos, sino que se asumían dados sin tomar en cuenta un posible proceso de obtención de los mismos. También se asumen con una calidad aceptable como para trabajar con ellos, sin indicaciones específicas de características de calidad deseables en la aplicación de los enfoques de minería. En este sentido, en el framework PRICED propuesto [13], se define claramente este ciclo de vida y las fases asociadas.

El trabajo de revisión sistemática presentado en esta sección, se vio reflejado en el artículo denominado “Process mining and data mining integration frameworks for evidence-based business intelligence: a systematic review” [32] el cual fue presentado y publicado en la Conferencia Iberoamericana de Ingeniería de Software (CibSE) edición 2021.

4. Propuesta

En el presente capítulo se plantea el problema de forma extensiva y la propuesta de solución definida en esta tesis que permita atacarlo.

4.1. Definición del problema

En la actualidad, las organizaciones se encuentran inmersas en ecosistemas heterogéneos que están formados por diferentes sistemas de información. Los datos asociados a cada uno de estos sistemas cumplen un rol fundamental en la toma de decisiones. Estas decisiones se deben basar en información lo más confiable posible y los datos generados y registrados por los sistemas son el principal insumo.

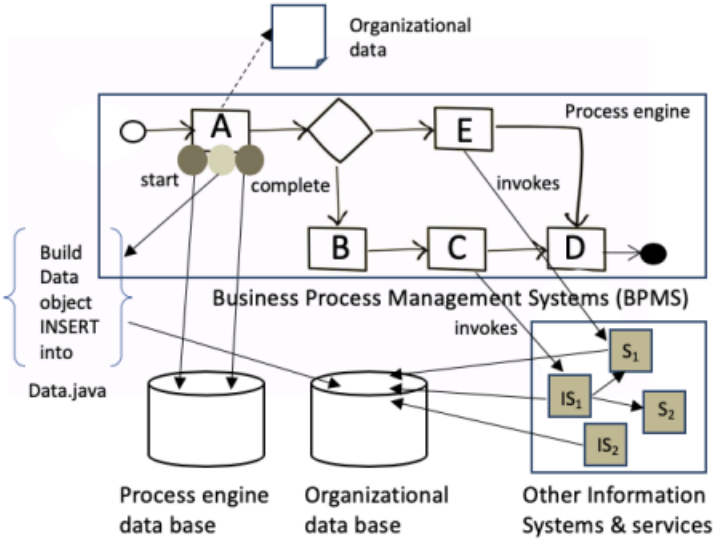

Figura 9: Relación entre datos de procesos y datos organizacionales [12]

Existen diversas fuentes de datos de donde se debe extraer la información. Un posible enfoque permite dividir estas fuentes de información en dos grandes tipos, los sistemas que modelan en forma explícita los procesos de negocio y los sistemas de información tradicionales se focalizan en el registro de datos operacionales. Si bien en la práctica muchas veces se puede observar claramente esta división, las organizaciones no deberían tomar en cuenta esto para la obtención de la información, su

realidad es única sin importar en qué tipo de sistemas se está modelando o registrando.

Esta división entre los tipos de sistemas se debe a diversos factores, generalmente el ecosistema de una organización cuenta con variados sistemas, implementados en diferentes tecnologías, sistemas legados que conviven con sistemas en plena evolución, muchas veces distribuidos geográficamente. En la figura 10 se muestra un esquema donde se puede visualizar esta división entre los enfoques de minería de datos y minería de procesos, dichos enfoques varían dependiendo de la fuente de datos asociada a cada sistema de las organizaciones. Además, cada sistema puede tener foco en perfiles de usuarios completamente diferentes, donde puede haber usuarios operativos que se encargan de registrar la operativa diaria de la organización, auditores que llevan controles en diferentes áreas, usuarios que analizan la información y toman decisiones, por mencionar algunos. Estos sistemas pueden comunicarse entre sí, intercambiando información o interactuando y de esta forma se generan puntos de dependencia y agregando más complejidad a dicho ecosistema.

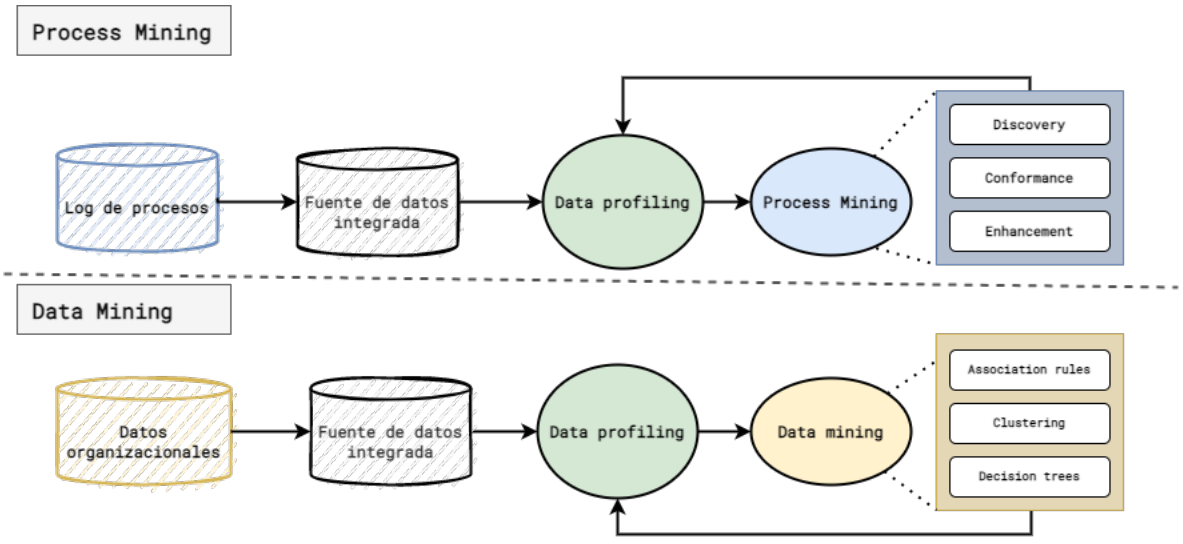

Figura 10: Minería desde las perspectivas de procesos y datos organizacionales

Además de la complejidad asociada a la heterogeneidad de los sistemas, también surgen los problemas asociados a los datos en sí, a la calidad de los mismos. Si contamos con datos de los cuales no conocemos su calidad o si la calidad de los mismos no es suficiente, afectará directamente a la calidad de la información que

se podría extraer. Partir de datos cuya correctitud sea baja, implicará conclusiones erróneas. Lo mismo pasa con la completitud, si en el conjunto de datos hay un bajo nivel de completitud se verá afectado en forma negativa los resultados que se pueden obtener. Otro aspecto importante es el nivel de frescura de los mismos, es decir, qué tan actualizados son, muchas veces la obtención de los mismos son resultado de procesos que insumen mucho tiempo, lo que dificulta contar con datos actualizados.

A la hora de que un tomador de decisiones quiera tener una visión global de su organización debe tomar en cuenta la información que está presente en todos los ámbitos de su organización, ello implica tomar en cuenta todas las complejidades que se mencionaron, lo que lo hace una tarea de extrema complejidad, y si tomamos en cuenta que el perfil de usuario que toma las decisiones generalmente no es un perfil técnico informático agudiza aún más el problema.

Es por esto que surge la necesidad de contar con mecanismos que faciliten la tarea de analizar la información asociada a los ecosistemas de las organizaciones, estos mecanismos deben permitir abstraerse de la complejidad descrita anteriormente y mostrar la información como una sola pieza. Tanto la minería de datos como la minería de procesos, cuentan con técnicas que permiten extraer información a partir de los datos, pero cada enfoque se basa en un tipo de información particular, lo que no permite a priori tener esta visión global que se mencionaba.

A modo de ejemplo, dado un evento de un proceso se cuenta con datos como el identificador de la instancia, nombre del evento, timestamp, recurso, entre otros. En la tabla 5 se muestra de forma simplificada algunos datos de un proceso donde para el evento “Register application”, se conoce su timestamp “2021-05-13T11:37:31” y fue ejecutada por “Pablo”.

Tabla 5: Ejemplo de datos de procesos

Instance	**Event**	**TimeStamp**	**Resource**
100	Register application	2021-05-13T11:37:31	Pablo
101	Define mobility program call	2021-05-13T11:45:23	Carlos
102	Requirements assesment	2021-05-13T12:07:01	Luis

Por otro lado, se tienen los datos organizacionales, datos que se registran en los diferentes sistemas organizacionales a medida que se ejecuta el proceso. Por ejemplo, para el caso del evento de la instancia con identificador 100 podemos suponer que tiene asociada en forma implícita una entidad "Programa" cuyos datos están almacenados en un sistema organizacional. Alguno de estos datos asociados a dicha entidad podrían ser los que se muestran en la tabla 6.

Tabla 6: Ejemplo de datos organizacionales

Id	**Name**	**Year**	**Date**
851	Erasmus	2022	04-06-2022
563	Itesm	2021	10-06-2021
...	...	...	...

Tener estos datos en forma disjunta obstaculiza la posibilidad de analizarlos en totalidad, por ejemplo, poder responder a las preguntas: ¿Cuál es el programa que más veces es instanciado en las distintas ejecuciones del proceso? ¿Existe un programa el cual no fue utilizado en ninguna instancia de proceso? ¿Existe una relación entre las variantes del proceso y los datos organizacionales?

Este es un ejemplo trivial, pero que muestra la problemática y el tipo de preguntas que se podrían realizar y responder contando las herramientas adecuadas y el conjunto de datos necesario. No contar con estas herramientas puede tener como consecuencia la visión parcial a la hora de analizar la información generada a partir de los datos.

En la siguiente sección se presenta una propuesta de solución cuyo objetivo es brindar un posible enfoque que permita crear un entorno donde se pueda analizar la información de las organizaciones en forma integral.

4.2. Propuesta de solución

En el capítulo anterior se pudo determinar cuáles eran las principales aplicaciones de la minería de datos y la minería de procesos, así como en qué aspectos eran consi-

deradas visiones complementarias sobre una misma realidad. Para esto se realizó un análisis de la información disponible en cuanto a propuestas y estudios que utilicen esta visión complementaria, pero se encontraron pocos resultados relevantes.

Además, en la sección anterior se realizó un planteo general del problema que supone la visión parcial que puede brindar la minería de datos y minería de procesos si se utilizan en forma separada sobre una realidad única.

En esta sección se busca plantear una propuesta que permita atacar el problema en base a utilizar los avances existentes en minería de datos y minería de procesos en forma complementaria. Dicha propuesta permite realizar el análisis de los datos integrados en forma evolutiva e incremental y además, en cada etapa dependiendo de la información que se esté buscando, permite utilizar indistintamente cualquiera de los enfoques. La propuesta se trata de una herramienta que facilite llevar a cabo la tarea de análisis desde un punto de vista integral. En la figura 11 se muestra un esquema que materializa una visión general, que se divide en tres grandes pilares. Lo primero es determinar la fuente de datos, luego debe permitir visualizar y realizar un perfilado de los datos (data profiling), y finalmente brindar la posibilidad de aplicar técnicas de minería de datos o minería de procesos dependiendo del estudio que se quiera realizar.

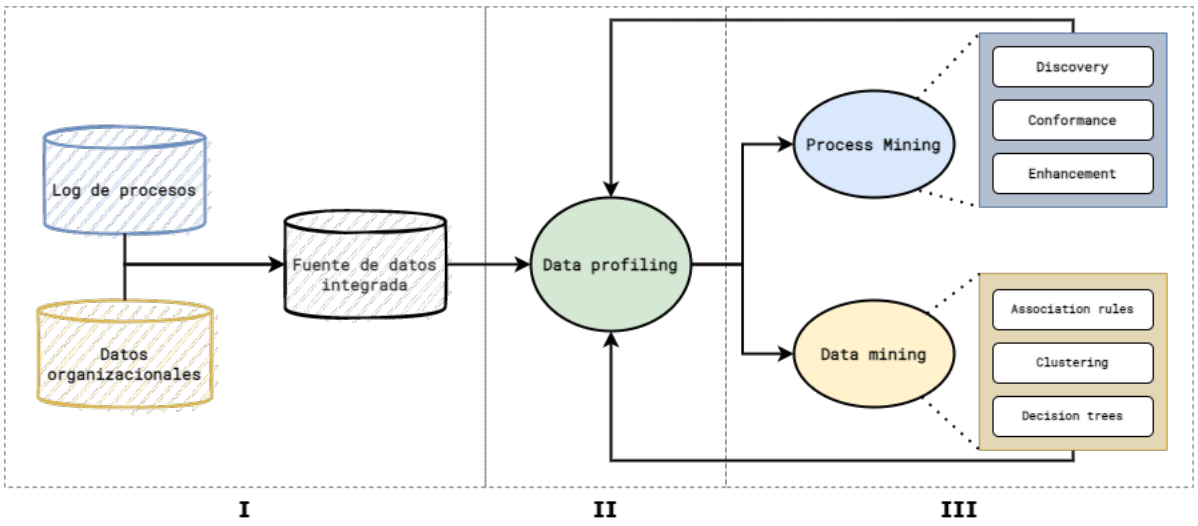

Figura 11: Diagrama de la propuesta para la integración de la minería de procesos y minería de datos

4.2.1. Fuente de datos

El origen de la fuente de datos es heterogénea, puede provenir desde sistemas de información tradicionales o de sistemas enfocados en los procesos. Como primera etapa de la solución debe contar con una fuente de datos integrada, que unifique y relacione las diferentes fuentes de datos. Esta integración de datos debe encontrar asociaciones entre las diferentes entidades existentes entre los diferentes sistemas, de forma de crear una visión única de los datos que modelan una misma realidad desde diferentes perspectivas. Considerando el contexto del presente trabajo y la posibilidad de contar con una extensión al estándar XES, la fuente de datos a utilizar debe ser el denominado log extendido.

En la definición del estándar XES no se establecen atributos específicos asociados a los elementos (log, traza, evento). Esto hace que la estructura sea genérica y se adapte a las realidades más heterogéneas, pero en escenarios específicos se puede requerir modelar alguna semántica concreta de la realidad en la que se está aplicando. Por este motivo, el estándar contempla la posibilidad de extender su definición con el fin de poder crear atributos asociados a un log, traza o evento de forma de modelar alguna semántica específica. En base a esta extensibildiad es que en la propuesta [12] se define un metamodelo que busca integrar los datos de los procesos con los datos organizacionales. La semántica reflejada en dicha propuesta busca generar una visión integral entre los datos organizacionales y los datos de procesos. Dentro de la propuesta se define un modelo en base a cuatro grandes componentes. En la Figura 12 se muestra un diagrama del modelo, donde se distinguen los cuatro componentes en forma de cuadrantes.

Por un lado, el cuadrante Process-Definition, describe las relaciones entre los elementos del proceso. Luego, el cuadrante Process-Instance, representa la ejecución de un proceso, es aquí donde se encuentra la información acerca de las instancias de ejecución (Case, VariableInstance, etc.).

El tercer cuadrante, Data-Definition, describe el modelo de datos donde se encuentran las entidades y los atributos. En la propuesta de integración, las entidades se

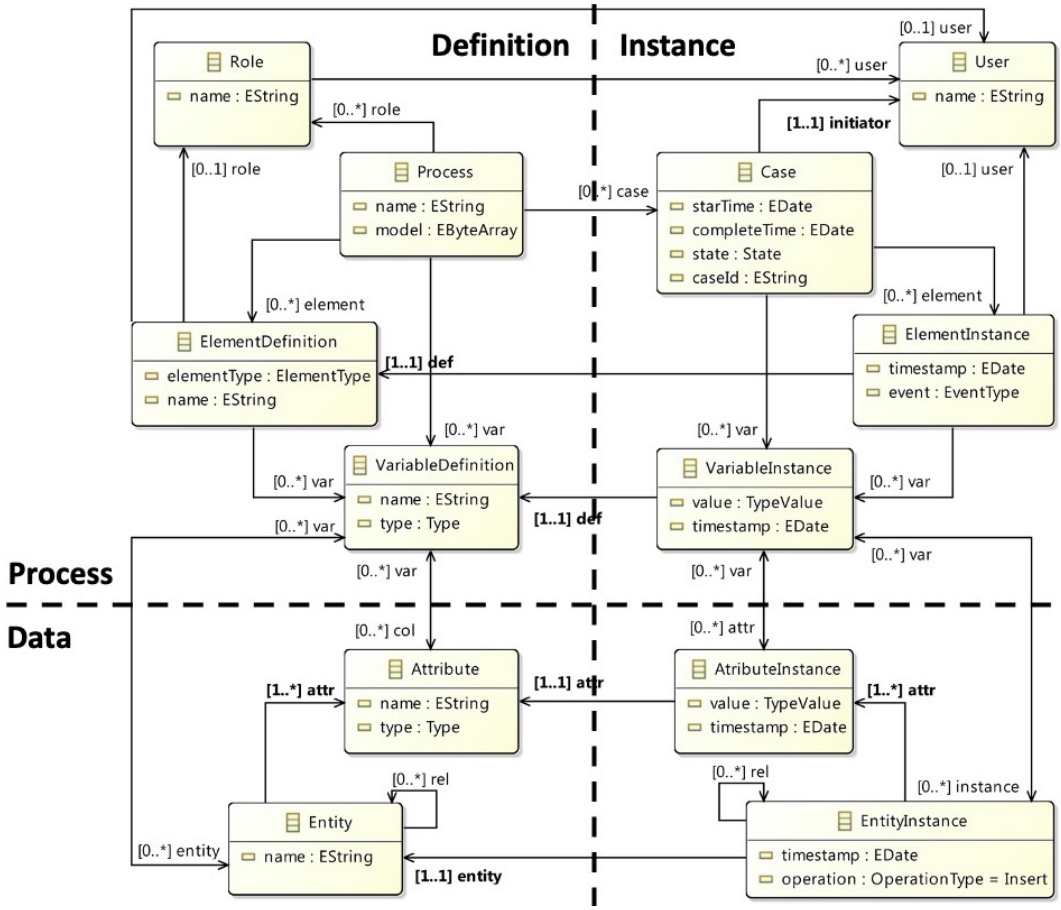

Figura 12: Modelo de datos integrados del log extendido [12]

.

corresponden con las tablas del modelo de datos de los sistemas de información y los atributos se obtienen a partir de las columnas de dichas tablas.

Por último se encuentra el cuadrante Data-Instance, en este se describen las instancias de estos elementos, es donde se encontrarán las instancias con valores de los atributos y entidades en que se encuentren presentes en los datos organizacionales. En el modelo, los cuatro cuadrantes contienen nexos que generan interconexiones entre sí, esto es lo que permite integrar los datos de los procesos con los datos orga-

Tabla 7: Ejemplo de datos de procesos y organizacionales integrados

Process data				**Program**			...
Event	TimeStamp	Resource	...	idProgram	Name	year	...
Define mobility ...		Carlos	...	851	ERASMUS	2022	...

nizacionales. Es decir, que se crean relaciones entre las instancias de ejecuciones de los procesos y las instancias de entidades y atributos de los datos organizacionales.

Figura 13: Nodo evento de log extendido

Utilizando el ejemplo mencionado en la sección anterior, la tabla 7 muestra una posible asociación entre los datos del proceso y organizacionales. Luego, en la figura 13 se puede visualizar parte de un log extendido que modela dichos datos de forma integrada. Esta porción de log muestra los datos del ejemplo asociados al evento del proceso. El nombre del evento se muestra bajo la clave “concept:name” y toma el valor “Define mobility program call”. La persona asociada a la tarea es Carlos, esto se puede ver bajo la etiqueta “org:resource”. Dentro de la lista “orgdata:varlist"se define la lista de variables asociadas, en el caso del ejemplo, se muestra una lista de profesores que contiene únicamente a un docente. Luego, en “orgdata:entlist” se presentan las entidades, en el caso de la figura, la entidad es Program. Además, dentro de la entidad se listan sus atributos. Para los atributos se define un nombre, tipo, valor, timestamp y asociación a una variable bajo los tags “concept:attname”, “orgdata:attValue”, “time:timestamp” y “orgdata:refVariable” respectivamente. En la figura 13, la entidad programa cuenta con dos variables, una es el nombre con valor Erasmus y la otra es el año con valor 2022.

4.2.2. Visualización y perfilado de los datos

En el segundo pilar se encuentra el data profiling o perfilado de los datos, que consiste en un proceso en el cual el usuario tiene un contacto directo con los datos. Donde tiene la posibilidad de poder conocerlos en profundidad para posteriormente poder llevar a cabo una tarea de análisis. Para esto debe saber en qué formato se encuentran, si son datos numéricos, caracteres, fechas, etc. También conocer la variabilidad de un atributo particular, saber si este toma o no una cantidad finita de valores. También a nivel de datos asociados a los procesos, resulta importante saber cuantas tareas se ejecutaron en una instancia o conocer el número total de instancias. Además, en este momento también se puede tener un acercamiento para conocer la calidad de los datos, conociendo la cantidad de posibles datos faltantes, la frescura y demás características de calidad.

Este paso es esencial para permitirle al usuario determinar qué información relevante podría extraer a partir de los datos disponibles.

4.2.3. Etapa de minería

Finalmente, el tercer pilar, aquí es donde permite al usuario aplicar indistintamente minería de procesos o minería de datos. En el capítulo 3 se obtuvieron cuáles son las técnicas que son utilizadas con mayor frecuencia y de las que permiten extraer la información más valiosa de los datos. Tanto desde el punto de vista de la minería de procesos, donde encontramos técnicas de descubrimiento, conformidad o mejora, como desde la visión de minería de datos utilizando reglas de asociación, agrupamiento o árboles de decisión. De esta manera poder contestar preguntas como por ejemplo:

- ¿Cuáles son las instancias que utilizaron un tipo de dato particular?
- ¿En qué instancias un dato tomó un valor específico?
- ¿Cuáles son los datos que pueden relacionarse a determinadas variantes de

procesos?

- ¿Las instancias que no finalizan correctamente cumplen con algún patrón en sus datos organizacionales?

Estos son algunos ejemplos de preguntas que podrían encontrar respuestas utilizando la minería de datos y minerías de procesos en forma complementaria.

Una vez que se aplique cualquiera de los enfoques, la solución debe permitir volver a la etapa de data profiling y de esta manera continuar realizando el análisis de forma iterativa. Muchas de las técnicas de ambos enfoques permiten obtener conjuntos de datos más pequeños y con características que los diferencian del resto, podríamos decir que son conjuntos de datos más homogéneos, donde el análisis incremental, permite realizar un análisis desde el punto de vista más general hacia puntos de vista más específicos sobre conjuntos de datos acotados ha determinado escenario.

Es importante que la solución sea extensible, que permita incorporar nuevos algoritmos o variaciones de los existentes, y de esta forma asegurar la evolución y futuras adaptaciones.

5. Desarrollo de un prototipo

En base a la definición del problema y la propuesta descrita en el capítulo 4, se diseñó un prototipo que busca cumplir con las características allí definidas. El prototipo busca implementar un marco de trabajo funcional donde se satisfaga el objetivo 4 (O4) y al mismo tiempo permite validar la factibilidad para este tipo de herramientas. La propuesta abarca las características mencionadas anteriormente y además sienta las bases para futuras investigaciones, extensiones y mejoras.

Dentro de los tres pilares que se mencionan en la propuesta de solución, el primero es la fuente de datos, donde se debe contar una visión integrada de los datos de la ejecución de los procesos y de los datos organizacionales. Esta visión integrada es lo que brinda el log extendido que se presentó en secciones anteriores. El prototipo que se presenta en esta sección se enmarca en el framework PRICED donde uno de los productos generados es dicho log extendido.

La herramienta propuesta permite integrar implementaciones existentes tanto de algoritmos de minería de procesos como de minería de datos basada en el log extendido. También es extensible de forma de permitir incluir nuevas implementaciones y brindar facilidades a nivel de interacción con el usuario que permitan llevar a cabo la tarea de análisis de la forma lo más clara posible.

Para llevar a cabo el prototipo se desarrolló en JAVA un plugin para PROM versión 6.9. PROM es la herramienta Open Source más utilizada sobre todo en marcos de investigación y académicos, ya que a través su entorno de desarrollo permite crear nuevos plugins e integrar implementaciones existentes en un ambiente diseñado especialmente para la minería de procesos. Esto hace accesible una gran cantidad de herramientas que facilitan la lectura e interpretación de los archivos de logs(.xes), como por ejemplo visualizadores y algoritmos de minería de procesos. Por otro lado, en cuanto a la minería de datos, se utiliza la biblioteca que provee Weka donde ofrece implementaciones de los principales algoritmos de minería de datos. Esta biblioteca se encuentra implementada en JAVA y cuenta con las interfases necesarias que facilitan la integración con la plataforma PROM.

5.1. Alcance

El alcance del prototipo busca cubrir los aspectos principales de análisis provistos por la minería de procesos y la minería de datos. Actualmente, PROM cubre las necesidades en torno a la minería de procesos. En la figura 14 se puede ver el alcance de la propuesta, donde a partir de un log extendido(.xes) se puede realizar data profiling para luego aplicar minería de procesos o de datos en forma indistinta.

Desde la perspectiva de la minería de procesos, permite realizar el análisis basado en el descubrimiento, conformidad y mejora. Como está implementado dentro de la plataforma de desarrollo de PROM, quedan a disposición todos los algoritmos, herramientas y visualizadores que provee PROM para el análisis de los logs de procesos.

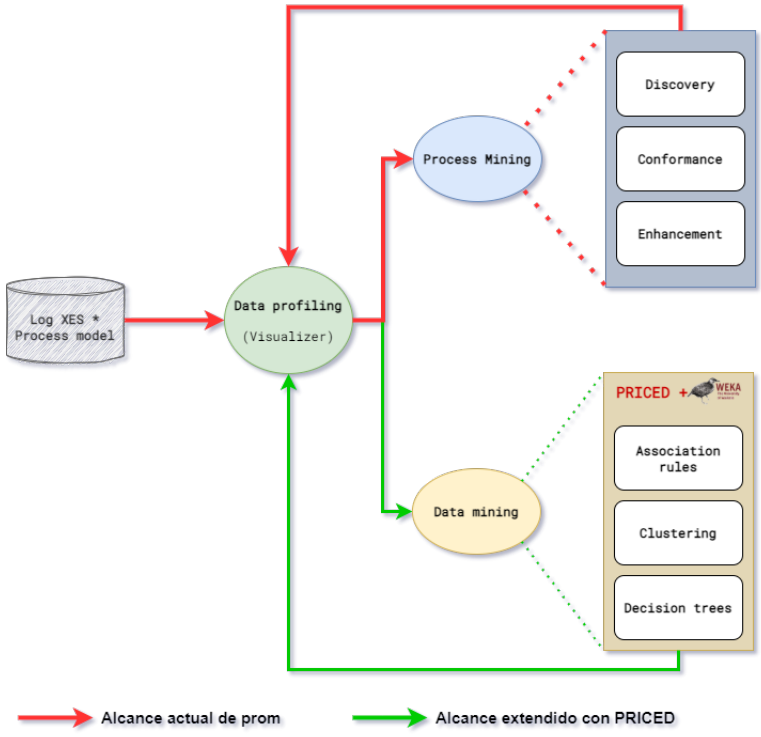

Figura 14: Alcance actual de prom y alcance extendido con PRICED

Desde el punto de vista del análisis de la minería de datos, el prototipo busca cubrir los enfoques más comunes como lo son las reglas de asociación, agrupamiento y árboles de decisión. Como se trata de un prototipo que busca validar la factibilidad,

se decidió integrar un algoritmo para cada una de estas metodologías. Dentro de las implementaciones de las técnicas provistas por WEKA, se eligieron los algoritmos más populares para cada una, esto es: para las reglas de asociación se integró el algoritmo Apriori [4], como técnica de agrupamiento se incorpora al algoritmo K-Means [5] y por último se seleccionó al algoritmo J48 [29] como implementación de algoritmo de árboles de decisión.

Luego es posible volver a realizar data profiling y continuar con el análisis de forma iterativa. El objetivo del prototipo es complementar el alcance actual de PROM añadiendo la visión que brinda la minería de datos. De esta forma se enriquece la etapa de data profiling, ya que además de los datos de la ejecución de los procesos, se cuenta con los datos de integrados de los diferentes sistemas de información que estén presente en el ecosistema de la organización a través del log extendido. Al incorporar las técnicas de minería de datos a la visión de minería de procesos, complementa el ciclo de análisis, brindando la posibilidad de, en cada etapa, utilizar un enfoque u otro de forma indistinta. También incluye una visualización de los resultados donde se integran las visiones de procesos y de los datos organizacionales de forma de facilitar la interpretación de los mismos.

5.2. Diseño

Para el diseño del prototipo se tomaron en cuenta los principales componentes a la hora de definir un plugin para PROM. Estos componentes encapsulan el diseño para la importación de archivos de log, ejecución de los distintos algoritmos y la visualización de información y resultados. Además de estos aspectos, también se consideró la extensibilidad del plugin que permitirá la evolución futura del mismo.

En las siguientes secciones se describen los principales aspectos de diseño de cada uno de estos componentes.

5.2.1. Importación

En el componente de importación se encuentra la lógica que permite la interpretación y posterior manipulación de los datos brindados por el log extendido. Como parte de esta interpretación del log, se diseñó un metamodelo con el objetivo de materializar en una estructura de datos los diferentes componentes que se encuentran en el log extendido.

Nativamente, OpenXes utiliza el modelo estándar XES (ver figura 15), donde existe un alto grado de similitud con el metamodelo definido más adelante. La necesidad de crear un nuevo metamodelo radica en que este último modela de forma nativa las pautas definidas en la propuesta PRICED, evitando complejidades asociadas a la generalidad del estándar XES.

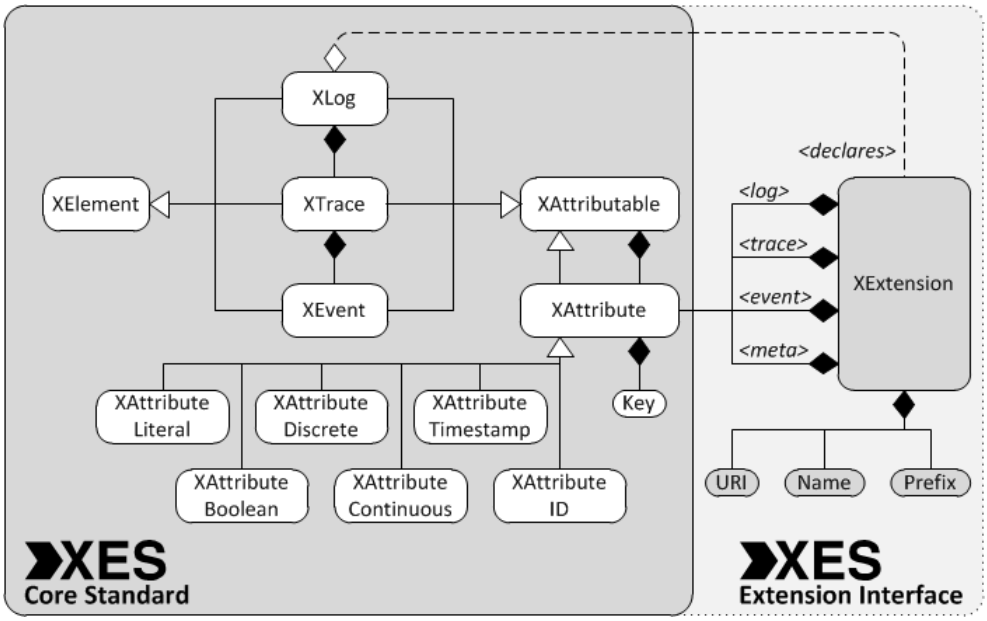

Figura 15: Modelo UML estándar XES con extensión

Dicho metamodelo se denominó PricedMetaModel, donde se materializa el proceso en su totalidad (Figura 16). Dicho proceso está compuesto por casos, a su vez, dentro de cada caso se encuentran los eventos. Para cada evento se conoce el tipo el usuario y rol asociado, así como un nombre que lo define. También los eventos tienen sus propias instancias de variables, donde de las variables se almacena, nombre, tipo, timestamp y valor.

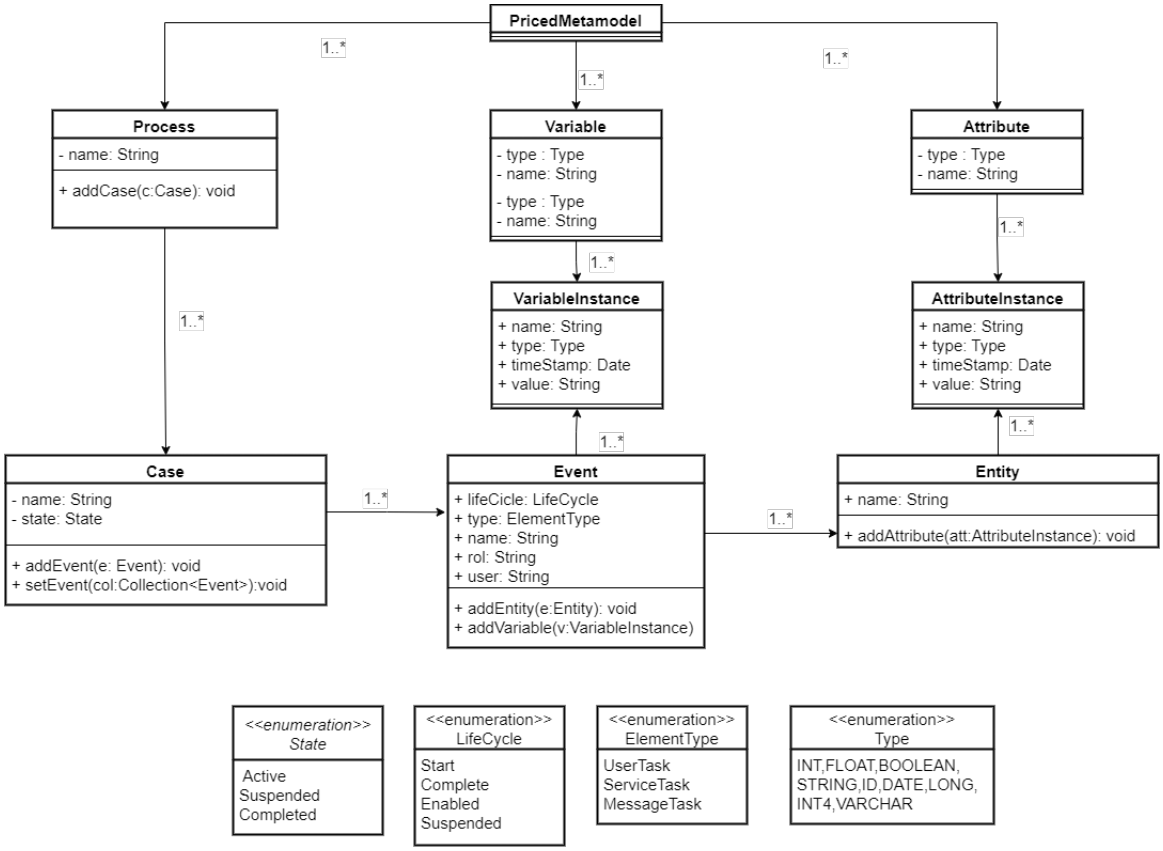

Figura 16: Diagrama de clases del metamodelo

Luego, dentro del evento se encuentran las entidades, de las que se almacena un nombre y una lista de instancias de atributos. Para cada atributo se conoce su nombre, tipo y timestamp y valor. Estos son los elementos principales que permiten materializar la información existente en el log extendido, integrando los datos de los procesos (Proceso, caso, evento) y con los datos organizacionales (Eventos y atributos).

Para obtener dicho metamodelo se debe procesar el archivo del log extendido, para esto se diseñó un parser de importación. Este parser recibe como entrada un archivo .xes, el cual contiene la extensión definida para el log extendido "dataorg", y retorna una instancia de PricedMetamodel que modela la información del log tal como se describió anteriormente.

5.2.2. Ejecución

La ejecución se presenta mediante un wizard que brinda una interfaz de usuario amigable. Mediante diferentes pantallas, guía al usuario en el proceso de análisis de los datos, brindando la posibilidad de elegir los diferentes algoritmos y seleccionar los datos sobre los cuales se quiere realizar el análisis. Se compone de cinco diálogos: diálogo de bienvenida, diálogo de selección de algoritmos, diálogo de selección de eventos, diálogo de selección de atributos y finalmente el diálogo de personalización de parámetros de los algoritmos (ver figura 17).

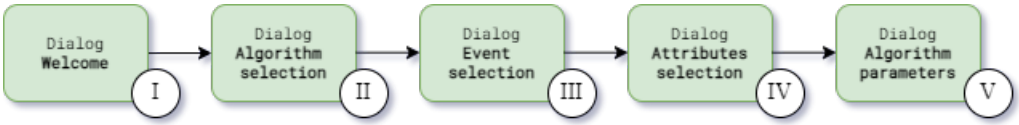

Figura 17: Interfases de usuario del wizard

El primer diálogo es un diálogo de bienvenida donde le debe permitir al usuario seleccionar el tipo de análisis a realizar con los tres enfoques de la minería de datos.

Luego, en la siguiente interfaz se selecciona el algoritmo a utilizar entre las diferentes implementaciones que estén disponibles.

El tercer paso muestra al usuario la posibilidad de aplicar el análisis sobre un evento particular del proceso.

En el cuarto diálogo se realiza la selección de atributos a utilizar y por último, en la última interfaz se pueden configurar los diferentes parámetros asociados al algoritmo seleccionado en el segundo diálogo.

Finalmente, se procederá a la ejecución del algoritmo seleccionado para el tipo de análisis utilizando los atributos que eligió el usuario. Como resultado de la ejecución se tendrá una salida PROM de tipo PricedResult (Figura 18), el cual cuenta con un visualizador dedicado que se presentará en secciones posteriores.

5.2.3. Visualización

Dentro del diseño de la visualización se pueden distinguir dos tipos de visualizadores, por un lado, el visualizador de log extendido, que permite realizar data profiling y visualizar los datos asociados al log extendido, como son atributos, variables, eventos. El visualizador se integra en PROM de forma de ser un visualizador más disponible para los elementos de tipo XLOG, el visualizador se denomina "PRICED Visualizer".

El objetivo principal es permitir profundizar en la visión de los datos organizacionales que no se acceden a través de los visualizadores estándar de PROM, donde se pueden ver vistas detalladas de los procesos, eventos, etc. Para los atributos se muestra una lista de los mismos y también se puede acceder a los distintos valores que toman. Por el lado de las entidades y atributos se listan sus nombres.

Por otro lado, se destaca el visualizador de resultados. En este caso, consiste en un visualizador que despliega al usuario los resultados obtenidos en una ejecución. Mediante diferentes tipos de interacciones permite al usuario interactuar con los resultados y además poder volver a ejecutar diferentes análisis sobre los sublogs obtenidos

Como resultado de la aplicación de los algoritmos se busca tener una agrupación de los casos que subdivida el log de procesos utilizado. Esta agrupación dependerá de cada tipo de análisis, por ejemplo, para el clustering se obtendrá una agrupación que modela los diferentes clusters obtenidos y para cada cluster se tendrán cuales son los casos, donde existen eventos y dentro de las entidades cuyos atributos cumplen las propiedades necesarias como para pertenecer al cluster. También a modo de ejemplo en el caso de las reglas de asociación, dicha agrupación representa las distintas reglas inferidas, y para cada regla se conocen los casos, cuyos eventos tienen alguna entidad tal que sus atributos cumplen con la regla en cuestión.

El contenido de dicho resultado variará en base al tipo de análisis, pero en todos los casos se cuenta con una agrupación de casos. Para esto se define la clase PricedDataMiningResultNode (Figura 18) cuyo objetivo es agrupar los casos para cada tipo

de análisis, en el caso del clustering cada nodo representará un cluster y dentro del cluster se encontrarán los casos asociados. Para las reglas de asociación cada nodo representa una regla y dentro de la misma se encuentran los casos asociados a dicha regla. Para el caso de los árboles de decisión, el mismo cuenta con una estructura particular (PricedTree) que permite agrupar los casos, donde se modela el árbol resultante de la aplicación del algoritmo. Dentro de cada nodo del árbol se cuenta con una lista de los casos relacionados.

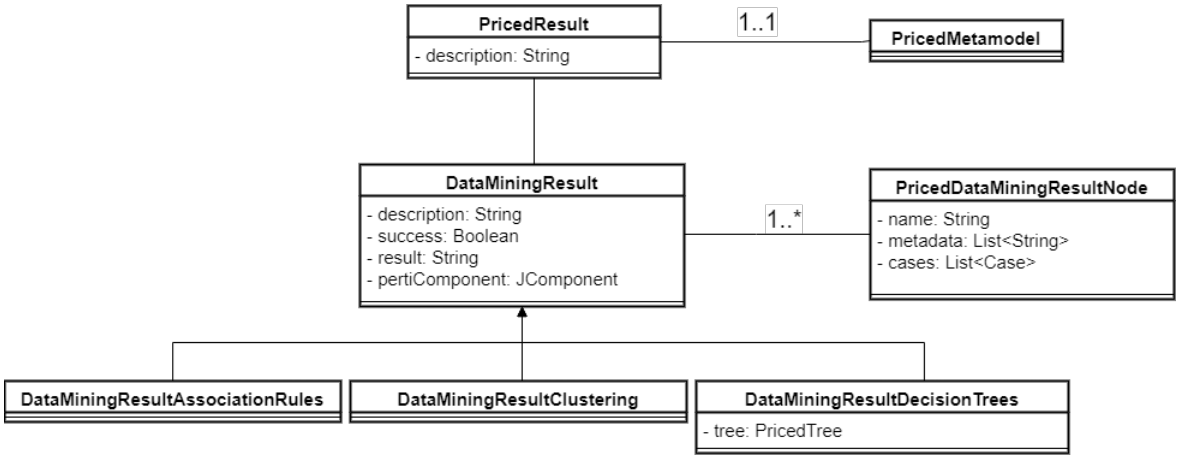

Figura 18: Diagrama de clases resultado

Luego, para cada tipo de análisis de data mining se define una clase abstracta que extiende DataMiningAlgorithm estas clases son: DMAlgorithmAssociationRules, DMAlgorithmCLustering, DMAlgorithmDecisionTree (Figura 18).

Para facilitar la interpretación de los resultados, particularmente en la visualización del proceso incluido en el log, se integra una red de petri obtenida a partir de un algoritmo de descubrimiento ya provisto por PROM. Este algoritmo de descubrimiento es Inductive Miner [25]. El objetivo de obtener una red de petri del proceso es poder brindar al usuario una visión gráfica del proceso a la hora de analizar los resultados en conjunto con la visón de minería de datos provista por el algoritmo que se esté utilizando.

5.2.4. Extensibilidad

Al tratarse de un prototipo que valida los mecanismos, no incluye un conjunto importante de implementaciones de algoritmos de minería de datos, resulta imprescindible la posibilidad de que la herramienta sea fácilmente extensible. Las futuras extensiones pueden incluir directamente implementaciones de los algoritmos, integración de nuevos algoritmos provistos por Weka o cualquier otra herramienta. Para permitir la posibilidad de extender se define las clases abstractas para cada tipo de algoritmos: DMAlgorithmAssociationRules, DMAlgorithmClustering y DMAlgorithmDecisionTree (figura 19).

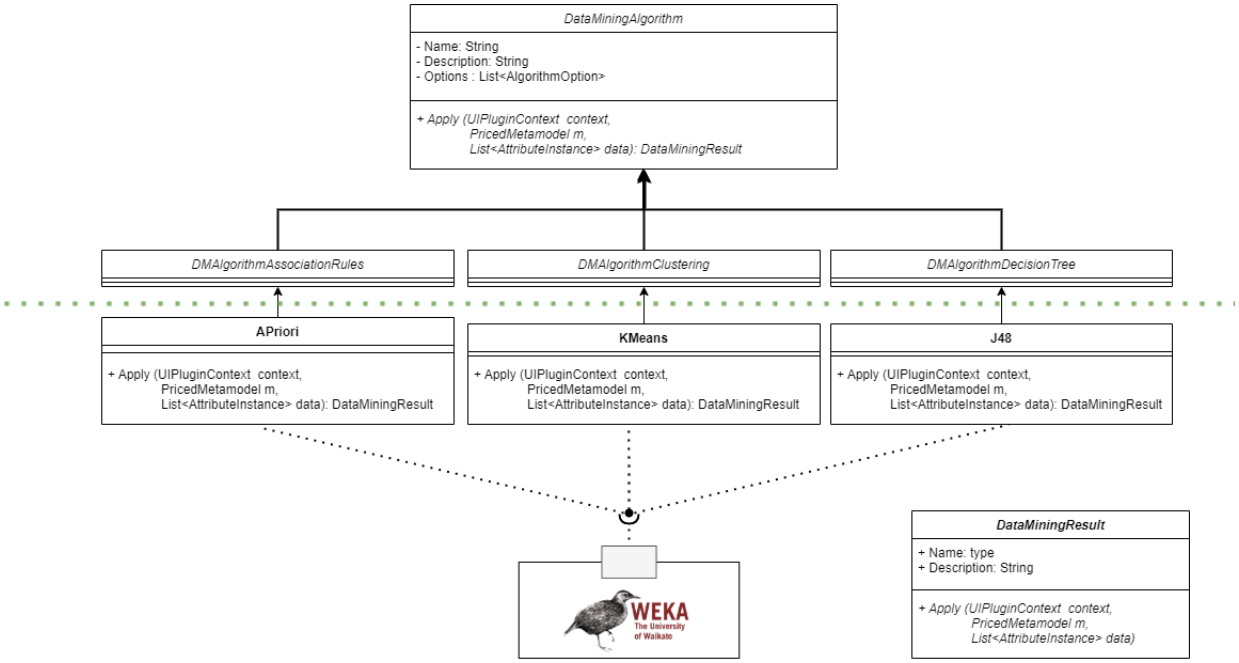

Figura 19: Diagrama integración con herramientas de minería de datos

Mediante la implementación de estas clases se permite la incorporación de nuevos algoritmos, ya sea mediante implementación directa o con la integración de otras herramientas.

5.3. Implementación

En sección anterior, se detallaron los principales aspectos de diseño del plugin, donde se destacaban las posibilidades que se brindan al usuario a la hora de ejecutar la herramienta. En esta sección se detallarán los principales aspectos asociados a la implementación del prototipo y de las interacciones descritas anteriormente. Dentro de esta descripción se encuentra el entorno de desarrollo necesario para integrar el plugin con PROM, la integración con Weka y los principales elementos para poder llevar a cabo los diferentes componentes descritos en la sección anterior.

Como se mencionó anteriormente, el plugin se desarrolla para la versión PROM 6.9. Esto determina aspectos a la hora de crear el entorno de desarrollo en cuanto a tecnologías y versiones de las mismas. La implementación se realizó utilizando la plataforma JAVA 1.8. Como IDE de desarrollo se seleccionó Eclipse en la versión 4.16 y como gestor de dependencias por defecto PROM utiliza Ivy, por lo que se seleccionó dicha tecnología. Para las interfases de usuario se utiliza la tecnología Swing.

Siguiendo la documentación disponible para la integración con PROM se utilizó como base el código fuente publicado en SVN[5] donde se encuentra el núcleo de PROM que permite la integración y desarrollo de personalizaciones y de nuevos plugins.

Para crear el nuevo plugin, se debe crear una clase que lo definirá (Listing 1). A partir de esta definición es que PROM puede detectar la nueva implementación. Para llevar a cabo esta tarea, se creó la clase "Priced"donde se encuentra la definición del mismo. Aquí es donde especifican los principales aspectos del plugin y además le permitirá a PROM reconocer la implementación como un nuevo plugin, registrándolo al inicio, quedando así a disposición del usuario.

Listing 1: Definición del plugin

```
1 @Plugin(
2       name = "PRICED, mining phase",
```

[5]https://svn.win.tue.nl/trac/prom/browser/Packages/GettingStarted

```
3       parameterLabels = { },
4       returnLabels = { "PRICED" },
5       returnTypes = { PricedResult.class },
6       userAccessible = true,
7       help = "An integrated view on process and data mining
8       is provided based on the extended log
9       connecting process and organizational data analysis."
10      )
```

En las siguientes secciones se muestran los detalles más relevantes de la implementación de cada uno de los componentes diseñados.

5.3.1. Importación

Luego, en la definición del método constructor de la clase Priced, es donde se comienza a definir el comportamiento del mismo. Dicho constructor toma como entrada un parámetro de tipo UIPluginContext, cuya función es brindar el contexto de ejecución. También como entrada tiene un parámetro de tipo XLog que representa el log que se desea analizar, en el contexto de este trabajo este parámetro tendrá un log extendido. Como resultado de dicho método se retorna el tipo de datos PricedResult (figura 18)

La importación consta de un parser que contiene la lógica capaz de interpretar el archivo de log extendido y conviertirlo en una instancia del metamodelo definido (figura 16).

La implementación del parser recibe un parámetro de tipo XLog con la referencia al log extendido importado por el usuario desde PROM, este XLOG tiene una estructura jerárquica propia del estándar XES de forma de modelar los logs de procesos con sus trazas, eventos, atributos etc. El parser se encarga de recorrer esta estructura y crear el metamodelo tomando en cuenta las extensiones asociadas al log extendido. A continuación se muestra un pseudocódigo que permite visualizar la recorrida jerárquica que se realiza sobre el XLog y la forma en que se construye

el metamodelo asociado.

Algorithm 1 PricedXLogParser pesudocode

```
procedure importFromXLog(UIPluginContext context,XLog l)
    /*Initialize PricedMetamodel Model and Process*/
    Process p = new Process(l.getName())
    PricedMetamodel m = new PricedMetamodel(p)

    for t in l.Traces() do
        Case c = new Case (t.getId() t.getName())
        p.addCase(c)

        for event in t.getEvents do
            Create a new priced event
            Get the event variables and add to de priced event
            Add priced event to the process

            for entity in e.getEntities do
                Create a new priced entity
                Get all entity attributes
                Add attributes to the priced entity
                Asociate the entity event to the priced event
            end for
        end for
    end for
    return m /* return metamodel */
end procedure
```

Como resultado de la ejecución del parser se obtiene una instancía del metamodelo que queda cargada en memoria y será utilizada en las siguientes etapas de ejecución del plugin.

5.3.2. Ejecución

Como parte de la ejecución se implementaron las diferentes interfases en forma de wizard que guían al usuario en el proceso de ejecución del plugin. La primera interfaz es un diálogo de bienvenida donde se permite al usuario seleccionar el tipo de análisis a realizar dentro de las opciones tradicionales de la minería de datos: Association Rules, Clustering y Desision Trees. Para cada uno de estos enfoques se muestra una

breve descripción acerca de las técnicas (figura 20).

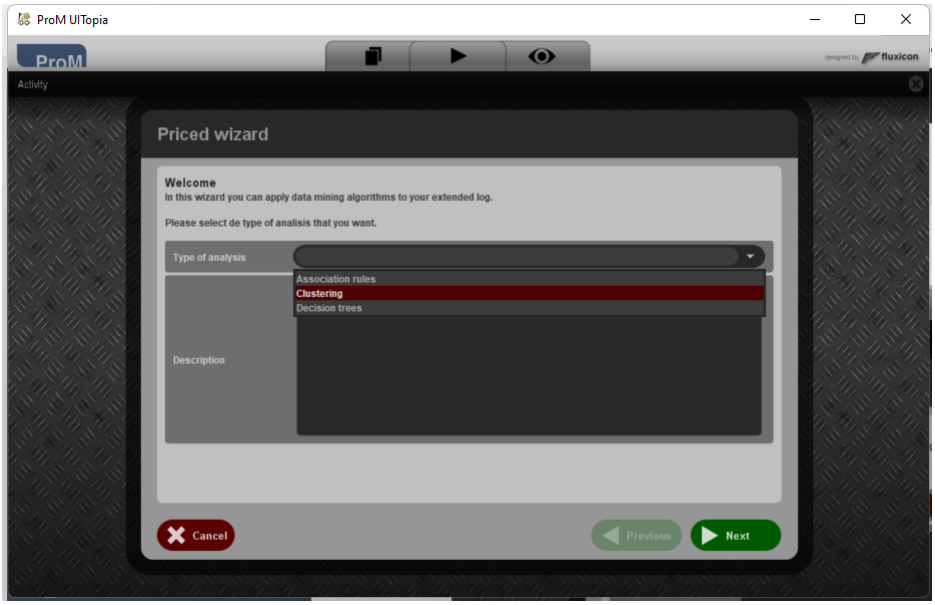

Figura 20: Diálogo de bienvenida

Una vez seleccionado el tipo de análisis a realizar, se muestran los algoritmos disponibles (figura 21). Para cada algoritmo se muestra una breve descripción del mismo que permite al usuario interpretar el objetivo del algoritmo.

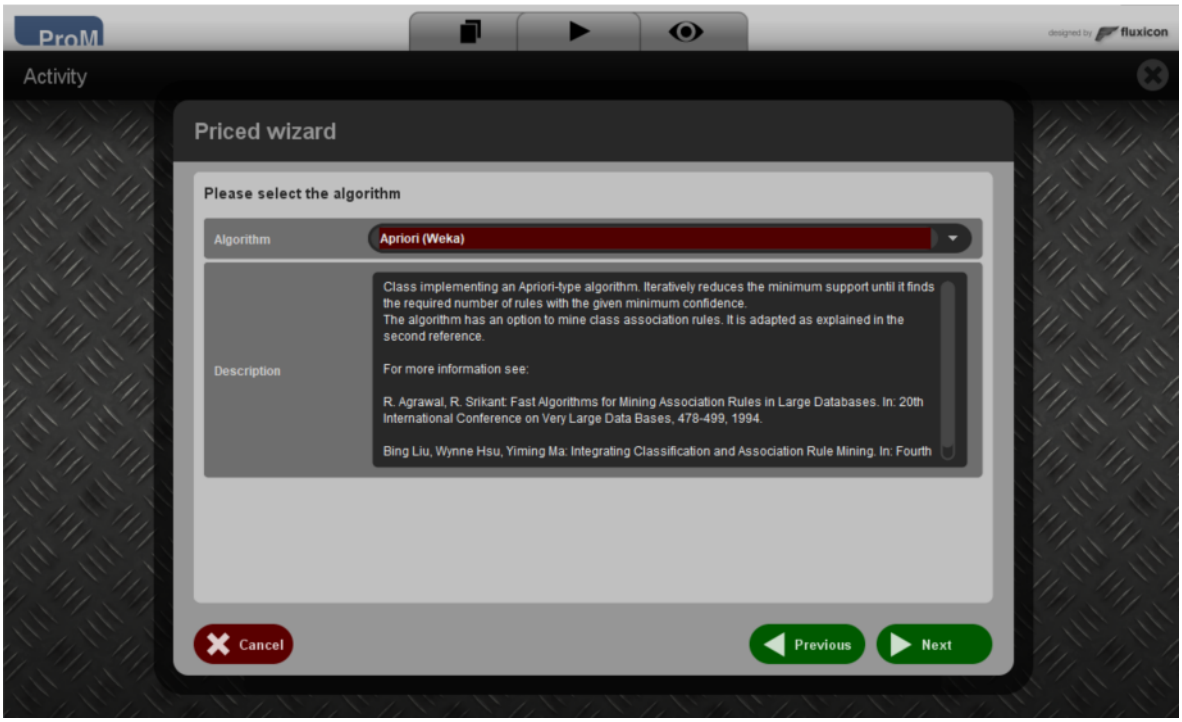

Figura 21: Diálogo de selección de atributos

Luego, a partir del algoritmo, en el tercer diálogo le permite determinar si se desea realizar el análisis sobre un evento en particular presente en el proceso, tomando en cuenta solo los atributos presentes en dicho evento o si, por lo contrario, el análisis se desea aplicar sobre todos los eventos, tomando en cuenta todos los atributos presentes en el log extendido (figura 22).

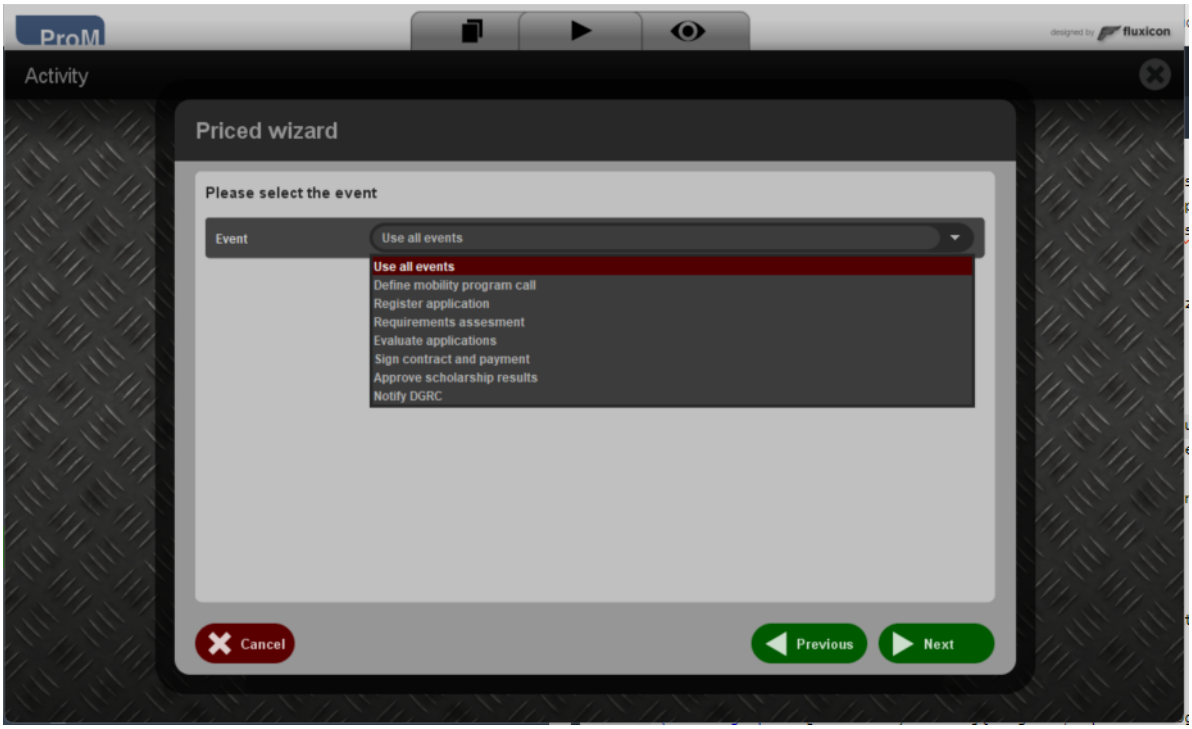

Figura 22: Diálogo de selección de evento

A continuación, se muestran los atributos y su tipo de datos presentes sobre los cuales se aplicará el algoritmo de análisis seleccionado (Figura 23).

Una vez seleccionado el o los atributos a tener en cuenta, se procede a la siguiente pantalla donde se muestra la configuración del algoritmo. Es aquí donde se puede personalizar la ejecución mediante la edición de las diferentes variables de ajuste provistas por los algoritmos.

A su vez, en la ejecución es donde se deben crear las herramientas necesarias para la integración de los algoritmos de minería de datos a PROM ya sea implementando directamente un algoritmo o integrando implementaciones provistas por otras herramientas. Para permitir esta extensiblilidad se definen clases abstractas donde se determinan los métodos y pautas a seguir a la hora de incorporar nuevas im-

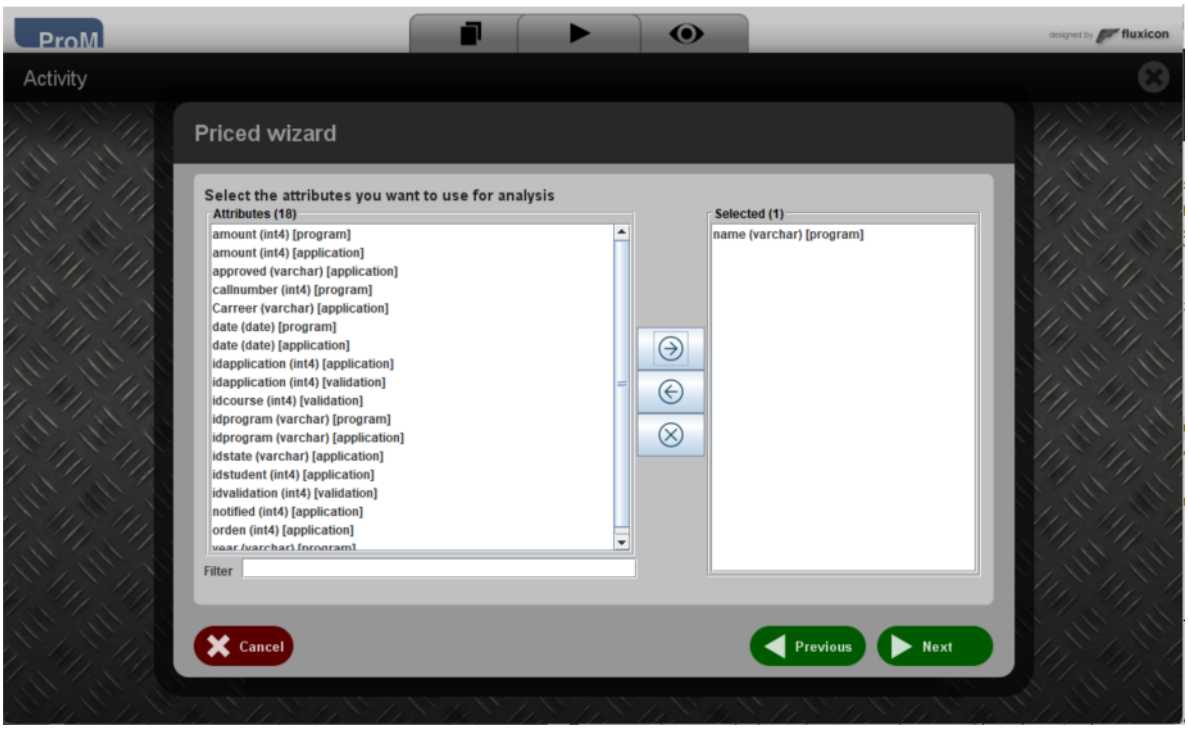

Figura 23: Diálogo de selección de atributos

plementaciones de algoritmos de minería de datos. La principal clase abstracta es la denominada DataMiningAlgorithm que contiene el método Apply(), en la implementación de este método se deberá invocar al algoritmo de minería de datos que corresponda. En el prototipo, estas implementaciones se encargan de integrar el plugin con la librería de Weka con el fin de utilizar las implementaciones de algoritmos provistas. Este método recibe tres parámetros: el contexto de ejecución de PROM, un metamodelo de tipo PricedMetamodel y la lista de atributos que serán utilizados para realizar el análisis. A su vez, retorna un resultado de tipo DataMiningResult (Figura 18) donde se materializa el resultado de aplicar el algoritmo.

A modo de validación y prueba de concepto se implementaron extensiones para las tres clases de modo de cubrir los tres enfoques. Los algoritmos que se integraron son: aPriori, KMeans y J48. En cada una de estas clases, en el método apply se realiza la manipulación y preparación de los datos necesaria para luego invocar a la biblioteca de WEKA con el algoritmo deseado. Una vez invocado, se manipula el resultado devuelto por la biblioteca para retornar una instancia de DataMiningResult. A continuación se presenta un pseudocódigo a modo de ejemplo, con la implementación

del método apply para el algoritmo SimpleKMeans.

Algorithm 2 Apply pesudocode

```
procedure APPLY(PricedMetamodel m, List<Attribute>attributes)
    /*Initialize Weka Model*/
    Model model = new SimpleKMeans()

    /*Initialize result variable*/
    DataMiningResultClustering result = new DataMiningResultClustering()

    /*Get data source as a matrix based on the priced metamodel and the selected
attributes. Columns represents the attributes, and the rows the instances.*/
    DataSource source = getDataAsMatrix(m,attributes)
    if data contains non nominal values then
        TransformAllDataToNominal(source)
    end if

    /*Set algorithm parameters*/
    model.setOptions()

    /*Apply Weka Algorithm*/
    model.build()

    /* Process the results and create a PricedResult instance. In clustering exam-
ple, weka return cluster assignaments to each instance */
    result.createClusters(model.getAssignments())
    return result
end procedure
```

Mediante la implementación de las clases abstractas DMAlgorithmAssociationRules, DMAlgorithmCLustering, DMAlgorithmDecisionTree (Figura 19) se permite integrar al plugin cualquier algoritmo disponible por otras herramientas o también existe la posibilidad de realizar una implementación propia. Como se mencionó anteriormente, en el caso del prototipo implementado en este proyecto, se implementó una integración para cada tipo de análisis a través de la librería brindada por WEKA. En la figura 19 se muestra las clases descritas anteriormente.

5.3.3. Visualización

Para implementar los visualizadores de log extendido, del metamodelo y de los resultados, se creó una clase con la anotación @Visualizer de forma que PROM la tome como un visualizador. A su vez, debe definirse añadiendo la anotación @Plugin donde además se deben especificar algunas características como nombre, etiquetas y el tipo de datos que retorna. En este caso el visualizador retornará un JComponent de Swing de forma de desplegar en pantalla la interfaz de usuario (Listing 2).

Listing 2: Definición del visualizador

```
1 @Plugin(
2     name          = "PRICED Visualizer",
3     returnLabels = { "Viewer" },
4     returnTypes   = { JComponent.class },
5     userAccessible = true)
6 @Visualizer
7 public class PricedVisualizer {}
```

En esta clase se definen tanto el visualizador del log extendido, un visualizador del metamodelo como el visualizador de los resultados. Para definirlo se implementan tres métodos visualize que se diferencian en sus parámetros. Todos reciben el contexto UIPluginContext, pero uno recibe XLog, otro PricedResult y el último PricedMetamodel (Figura 24).

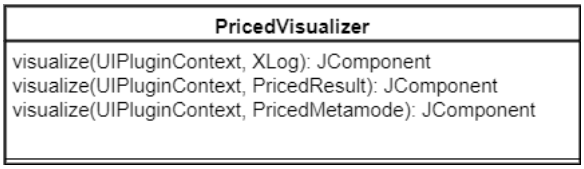

Figura 24: Clase del visualizador y sus métodos

En todos casos se crean a partir de componentes de tipo JPanel, donde se implementan los diferentes controles de UI. Estos controles son los provistos por Swing en forma nativa, a excepción del visualizador de redes de petri utilizado en el visualizador resultados. Este visualizador se implementó reutilizando otro visualizador provisto en PROM, PetriNetVisualization (@author bfvdonge). De esta forma se

evita la re-implementación de esta interfaz.

Visualizador del log extendido Este visualizador permite realizar data profiling y visualizar los datos asociados al log extendido, como son atributos, variables, eventos. El visualizador se integra en PROM de forma de ser un visualizador más disponible para los elementos de tipo XLOG, el visualizador se denomina "PRICED Visualizer".

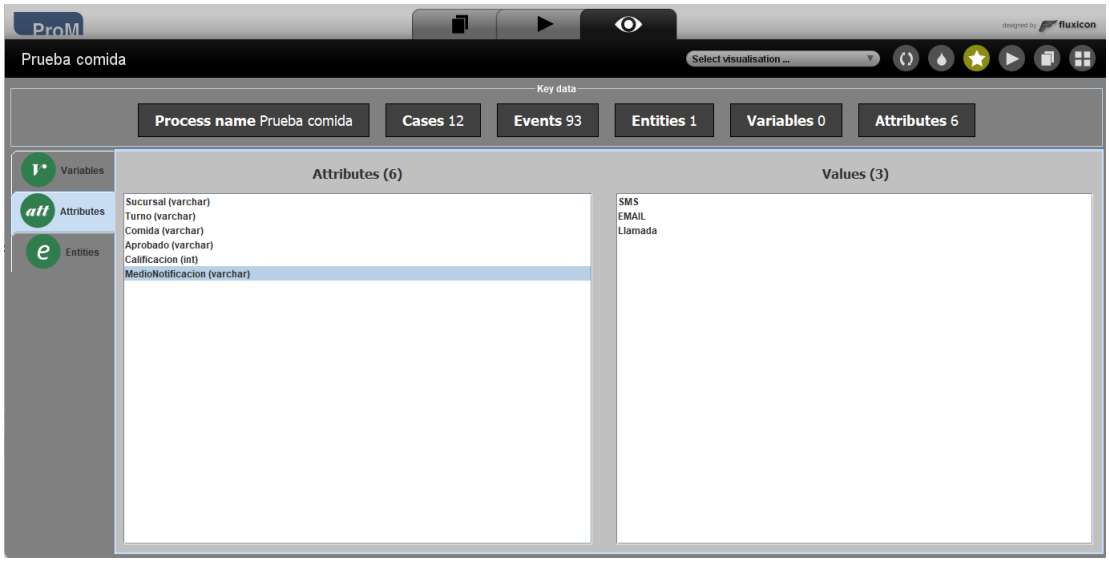

Figura 25: Interfaz del visualizador de log extendido

En la sección superior muestra el nombre del proceso, la cantidad de casos y eventos. Esta información se extrae a partir de los datos de los procesos. Por otro lado, también en dicha sección se muestran la cantidad de entidades, variables y atributos, en este caso, dicha información se extrae a partir de los datos integrados de los sistemas de información.

Visualizador de resultados Por último, cuenta con un visualizador de resultados, el cual se basa en un menú lateral con tres opciones: Result, Extended log y Summary. En la sección result se muestra el resultado obtenido por el algoritmo aplicado mediante una perspectiva visual e interactiva y también se brinda una salida en texto plano. En extended log, se cuenta un visualizador análogo al descrito en la sección 5.3.3 que permite tener una perspectiva general del log extendido utilizado

(Figura 26).

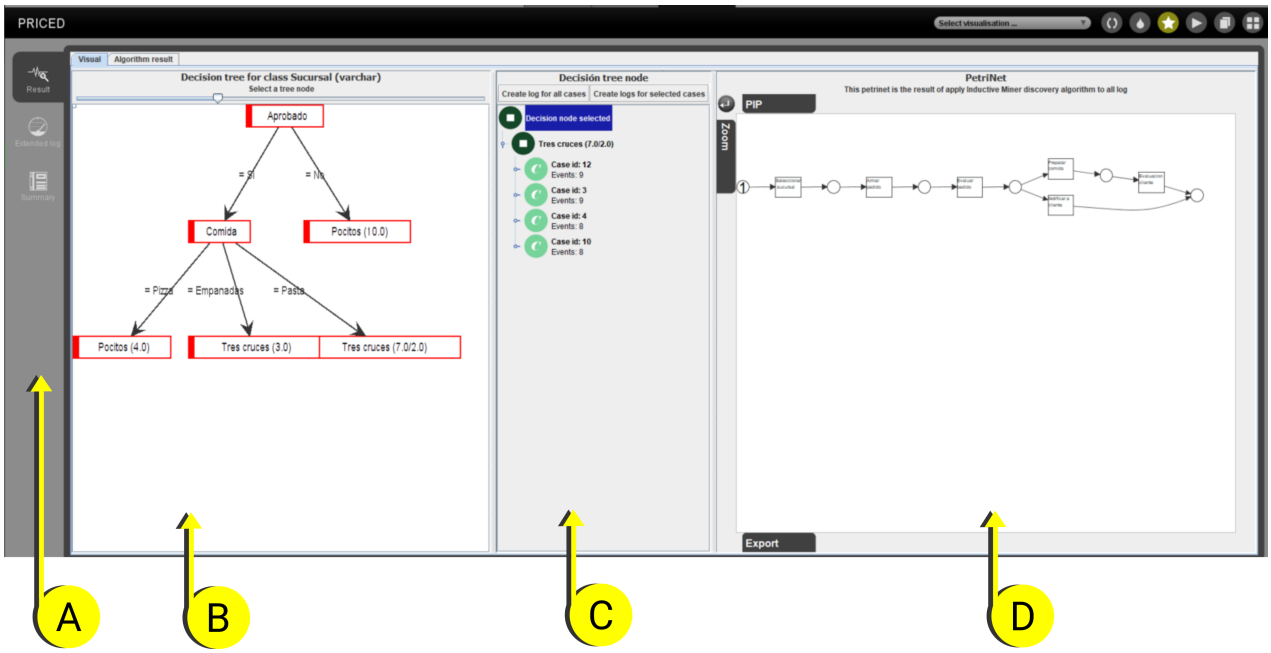

Figura 26: Secciones en visualizador. A. Menú lateral, B. Visualizador de árboles de decisión, C. Visualizador de agrupaciones de casos, D. Visualizador de la red de petri

La primera opción en el menú lateral es Result, esta sección se caracteriza por tener por una visión en texto plano de los resultados. Esta salida es provista directamente por Weka y es la salida que obtendría el usuario al utilizar Weka en forma interactiva. Por otro lado, se encuentra disponible una salida visual. Dicha interfaz se caracteriza por estar dividida en hasta tres secciones en forma vertical. En la figura 26 se pueden apreciar estas tres secciones, donde la sección de la izquierda se corresponde con un árbol que solo se muestra en el tipo de análisis de árboles de decisión. En dicha sección se muestra el árbol resultante obtenido a partir del algoritmo aplicado, donde cada nodo representa un atributo, y cada arista un posible valor. Las hojas son los valores asociados al atributo seleccionado como clase en el algoritmo. También cuenta con un control de slider que permite realizar zoom, para ampliar o reducir el tamaño del árbol que se está visualizando.

Luego, en la sección del medio, se tiene una visualización jerárquica de nodos. Aquí, dependiendo del tipo de análisis, el nodo raíz y los nodos de nivel 1 pueden tener diferente interpretación, estas son: un cluster para el caso de clustering, una regla

para el caso de reglas de asociación o un posible valor para el caso de árboles de decisión. En este último caso, para los árboles de decisión, el árbol que se muestra en la sección intermedia dependerá del nodo seleccionado en el árbol de decisión.

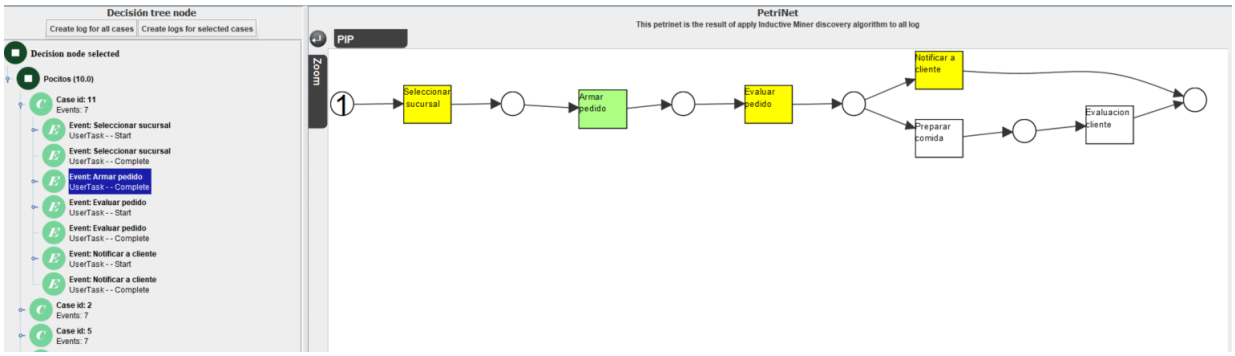

Figura 27: Interfaz de visualización de los resultados

A partir de los nodos de nivel 2, en todos los casos se muestra una jerarquía que representa los elementos de un proceso, donde para cada caso se tendrán sus eventos, para cada evento las entidades relacionadas. Cada entidad, a su vez, contienen sus atributos y finalmente cada atributo contiene información acerca de su valor, tipo de datos y timestamp (Figura 28).

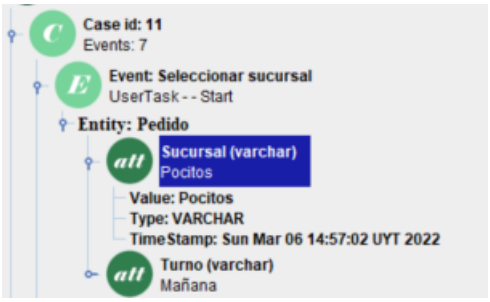

Figura 28: Interfaz de visualización de la jerarquía de casos

Esta sección cuenta con dos acciones: “Create log for all cases” y “Create logs for selected cases”. Estas acciones permiten crear nuevos logs extendidos que se obtienen a partir de subdividir el log extendido inicial. Esta partición se realiza a partir de los resultados obtenidos por el algoritmo, por ejemplo, para el caso de clustering se podrá generar un sublog que contenga solo las trazas (o casos) que pertenezcan a un cluster seleccionado. Para las reglas de asociación se podrán obtener sub logs que contengan solo las trazas que cumplan con determinada regla y para el caso de los árboles de decisión permite obtener el sublog que contiene las trazas que se

encuentran relacionadas con el nodo seleccionado en el árbol de decisión. Los logs extendidos generados a partir de estas acciones quedan disponibles en la interfaz de Prom como un archivo XLog como posible insumo para nuevamente ser analizado. Ya sea utilizando las herramientas de minerías de procesos que disponibiliza la herramienta en forma nativa, como también la posibilidad de aplicar nuevamente el plugin PRICED.

Por último, en la sección derecha de la interfaz, se encuentra el visualizador de una red de petri. La red de petri que se muestra, es el resultado de aplicar el algoritmo de descubrimiento Inductive Miner utilizando el log extendido. A su vez, cuando el usuario navega seleccionado uno de los nodos de la jerarquía, en la red de Petri se colorean en amarillo los eventos que se encuentran en los casos relacionados con nodo. Si la selección dentro de la jerarquía representa un evento o atributos de un evento, el mismo se coloreará de verde dentro de la red de petri. Los eventos que no estén presentes dentro de los casos relacionados con el nodo seleccionado se muestran en color blanco. En la figura 27 se muestra un ejemplo de selección donde se encuentran en amarillo todos los eventos relacionados con el caso y en verde el evento seleccionado.

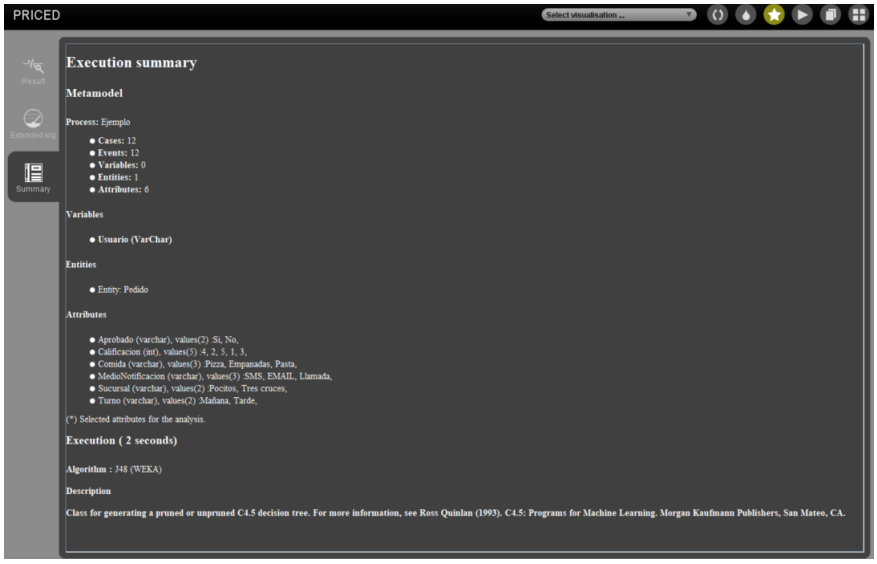

Figura 29: Resumen de ejecución

De esta forma, mediante una interfaz interactiva se le permite al usuario navegar a través de los resultados, interactuar con los mismos, visualizando los eventos que se

relacionan al nodo que esté analizando y a su vez obtener sublogs del proceso que le permitirán seguir iterando en el análisis.

Por último, en Summary, tal como se muestra en la figura 29, cuenta con un resumen de la ejecución en donde se brinda información acerca del metamodelo generado a partir del log extendido. La información que se despliega aquí consiste en un resumen con el total de casos, eventos, variables, entidades y atributos. Luego una sección donde se detallan cada uno de estos elementos. Por último se muestra el tiempo de ejecución insumido al aplicar el algoritmo de análisis y una descripción del algoritmo.

6. Caso de estudio

De forma de probar la validez del prototipo se llevaron a cabo pruebas prácticas en un caso de estudio. El objetivo de las pruebas es cubrir en un escenario empírico los principales aspectos asociados a la propuesta tanto a nivel teórico como a nivel práctico mediante la utilización del plugin implementado. Para el caso de estudio se utilizó el caso denominado Student Mobility, este fue presentado en la propuesta general de integración de datos de procesos de negocio y organizacionales [12]. Esta propuesta conceptual fue implementada en la práctica utilizado el caso mencionado anteriormente en la tesis de grado denominada "Modelos y algoritmos para minería de procesos y datos" [8]. En dicha tesis, se lleva a cabo una implementación de un algoritmo de matcheo que permite unificar los datos organizacionales y los datos de procesos.

A partir de esta implementación, se realizan ejecuciones con el fin de crear un log extendido, este log es el utilizado como input para llevar a cabo las pruebas del presente caso de estudio.

Cabe destacar que en la propuesta se definió el objetivo número 5, donde se planteaba la utilización de un caso de estudio con datos provenientes de la realidad, debido a que no fue posible obtener datos de casos reales, se utilizó el caso de estudio que se describe a continuación.

6.1. Descripción del proceso

El proceso de negocio describe el escenario de programas de movilidad estudiantil, donde cada año se crean llamados a programas para que los estudiantes pueden presentar sus intenciones de aplicar al mismo. Luego, mediante un proceso que consta de varias etapas de evaluación y controles, se procede a aceptar o rechazar dicha aplicación. Los estudiantes cuya aplicación sea aceptada procederán a la firma de un contrato y recibirán una determinada suma de dinero.

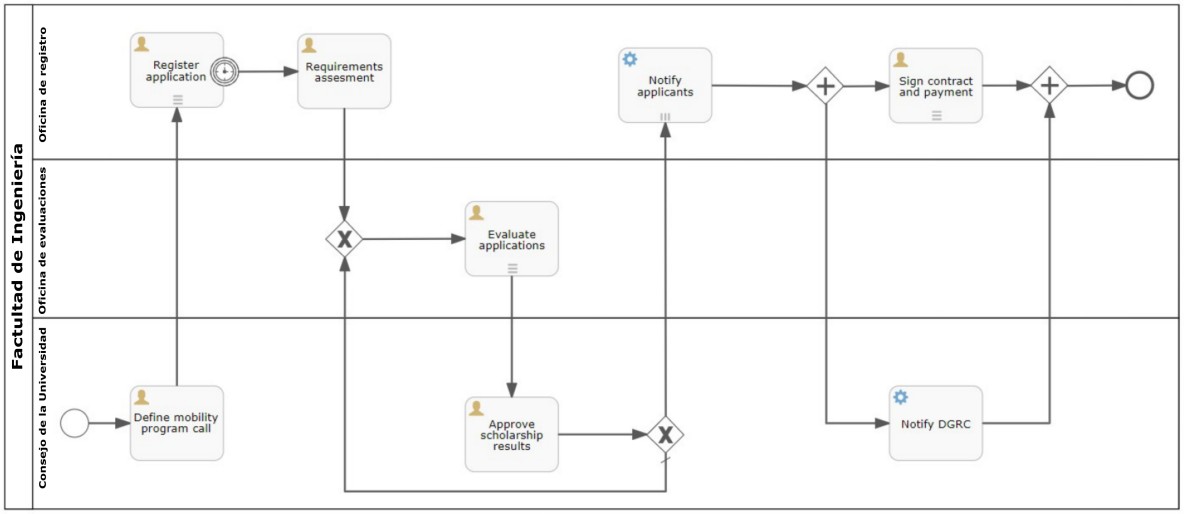

Figura 30: Proceso de negocio del caso Students Mobility. Adaptado de [8].

Este proceso comienza con la actividad “Define mobility program call”, en esta tarea es donde se define el llamado de inscripción para un programa de movilidad (Figura 30). Luego, en la tarea “Register Application”, se encarga de recibir las postulaciones de los estudiantes durante 15 días. Por cada una se registra una nueva Aplicación en la base de datos organizacional. El estado inicial para dicha aplicación será “initiated”, y la misma queda asociada al estudiante y programa correspondiente.

La siguiente tarea del proceso es “Requirements assessment”, como lo indica su nombre, en este paso se evalúan los requerimientos de las aplicaciones y se actualiza el estado de las mismas. En caso de ser aceptada, el estado será “confirmed” y “rejected” en caso contrario. Las aplicaciones rechazadas culminan su proceso en ese momento.

Luego, en la tarea “Evaluate applications”, se procede a evaluar cada aplicación y se establece un orden de prioridad, a su vez se determinan posibles titulares y sustitutos dependiendo del número de plazas disponibles para el llamado. Como resultado de esta tarea, el estado de la misma podrá ser “holder” o “substitute” dependiendo de la evaluación realizada.

A continuación, el consejo evalúa cada aplicación en la tarea “Approve scholarship results” y se confirma el estado de titular o sustituto para cada aplicación.

Finalmente, en la tarea “Notify applicants” se notifica el resultado del proceso de evaluación, cada estudiante relacionado. En caso de que la aplicación sea aprobada, se procede a la firma del contrato y recepción de pago por parte de los titulares seleccionados en la tarea “Sign contract and payment”. Para cada titular de aplicación se creará un nuevo registro “Mobility” en la base de datos organizacional.

6.2. Datos organizacionales

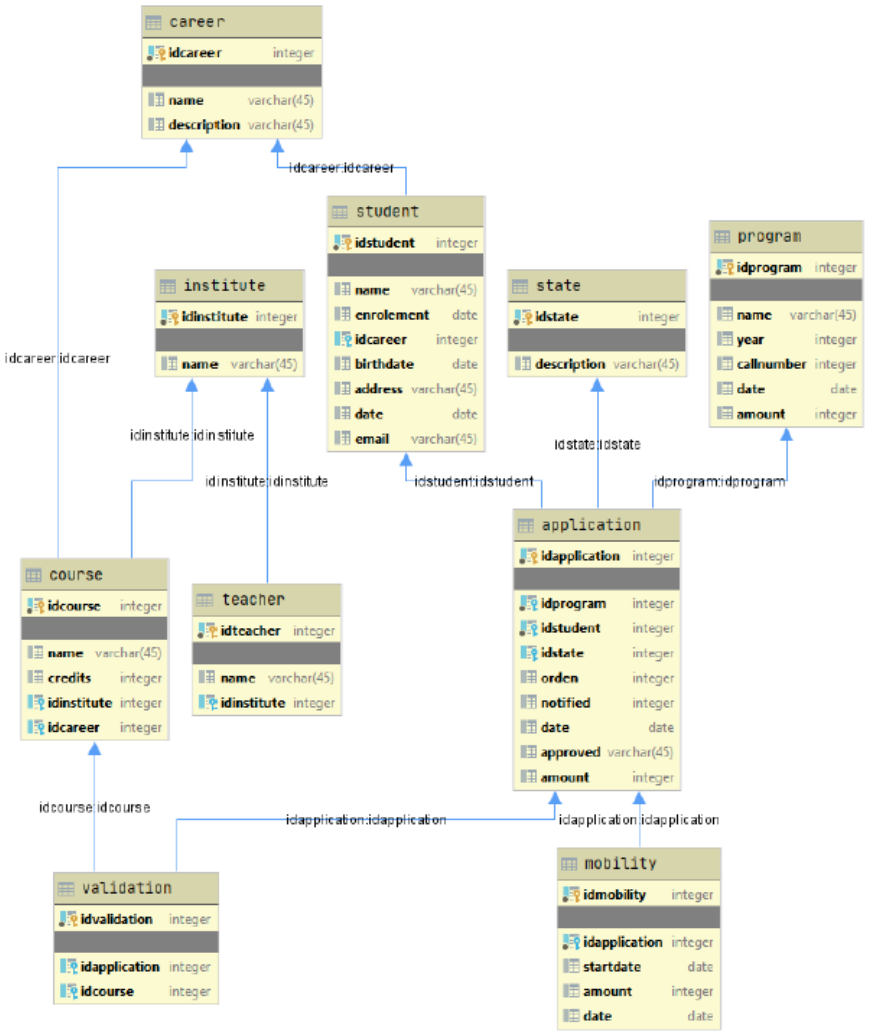

Figura 31: Modelo de datos del caso Students Mobility [8]

En el caso de estudio, además del proceso de negocio, se cuenta con un sistema organizacional donde se registran y materializan diferentes conceptos asociados a la

realidad. En la figura 31 se muestra el modelo de datos asociado. En él se pueden apreciar los diferentes conceptos, entidades y relaciones entre ellos.

6.3. Log extendido

En el proyecto de grado que se hace referencia, se implementó tanto el modelo presentado en la figura 30 en Activiti 6.0, como el modelo de datos presentado en la figura 31 utilizando el dbms PostgreSQL. Además, mediante la implementación de una herramienta de generación de instancias de procesos, se generaron datos de ejecución en forma automática. Un total de 100 instancias de procesos, que contienen ejecuciones con datos aleatorios, pero que respetan reglas determinísticas con el fin de imitar la realidad. De esta manera se obtuvo un log de procesos, este log de procesos es un log extendido, ya que contiene datos organizacionales de forma integrada. Esta integración se llevó a cabo mediante la puesta en práctica de la propuesta realizada, generando un modelo integrado de datos de ejecución de procesos y datos organizacionales.

Ese log extendido es el que se utilizó para realizar la evaluación práctica del presente trabajo. Cabe destacar que con el fin de simplificar la visualización y muestra del potencial del plugin implementado se decidió simplificar el log utilizado. La simplificación consiste en que en cada instancia de procesos, se encuentra presente una única postulación. En la implementación del modelo, por cada instancia de proceso se permitían crear múltiples postulaciones y esto hace que el trabajo de minería de procesos y de datos se vuelva más complejo. Esta simplificación permite mostrar casos prácticos de forma sencilla y fácilmente entendible sin perder generalidad.

En la Figura 32 se muestra parte de este log extendido donde por ejemplo se puede distiguir el evento Register application y los datos organizacionales. Además, está presente la lista con las variables studentId y selectCourses. También se encuentra en dicha figura la lista de entidades en donde se puede ver la entidad application y la lista de atributos asocidados a dicha entidad. Entre los atributos que se pueden visualizar están: date, idapplication e idprogram.

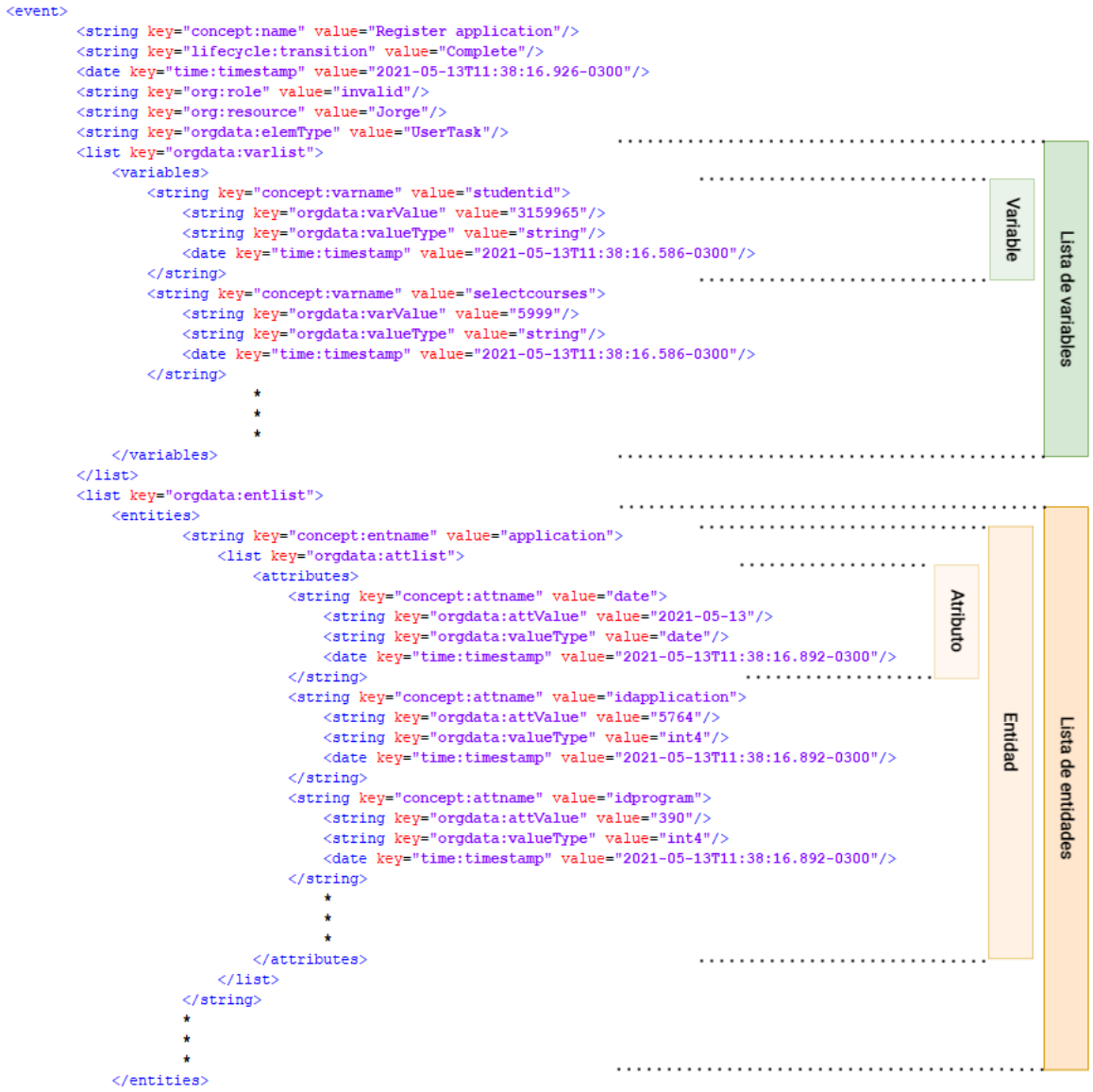

Figura 32: Parte del log extendido del caso de estudio

6.4. Evaluación práctica

El objetivo de esta evaluación es mostrar en la práctica la ejecución del plugin. Aquí se podrá constatar el potencial del análisis de forma integral del conjunto de datos, tanto desde el punto de vista de la minería de procesos como la minería de datos.

A modo de ejemplo se muestran tres diferentes aplicaciones donde se aplica una técnica de agrupamiento, reglas de asociación y árboles de decisión.

6.4.1. Data profiling

La primera etapa de todo análisis de datos, es el data profiling. Es decir, tener un primer contacto con los datos, poder visualizarlos, y navegar a través de ellos. Para llevar a cabo esta tarea, se procede a importar el log extendido en Prom.

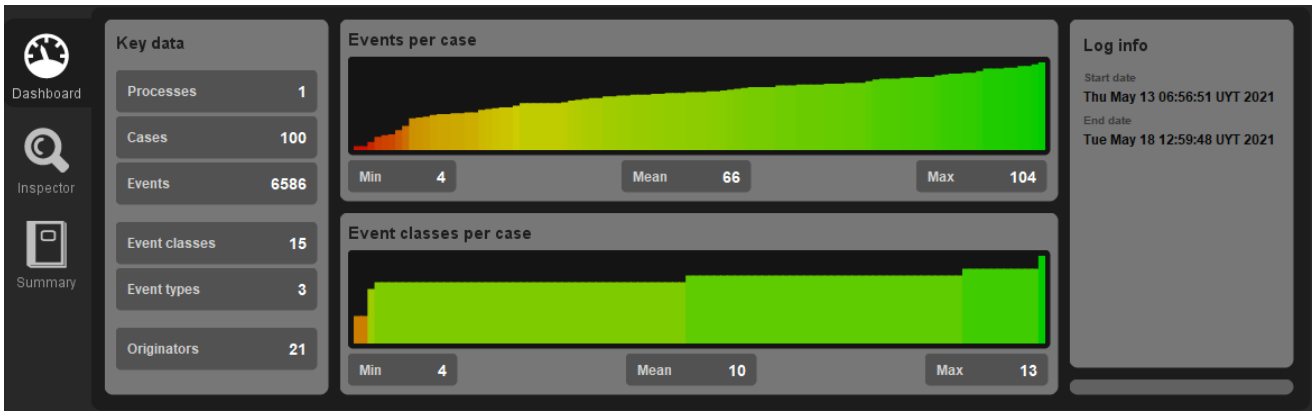

Figura 33: Visualizador estándar de log de procesos. [8]

Por defecto podemos utilizar el visualizador estándar, donde por ejemplo podemos ver la cantidad de procesos involucrados, la cantidad de casos y eventos. Este visualizador es el estándar que provee Prom, por lo cual no se entrará en mayores detalles en cuanto la funcionalidad que posee (Figura 33).

En este visualizador se puede ver fácilmente los datos asociados a los procesos, pero no los datos asociados a la extensión, es decir, los datos organizacionales. Cuando se selecciona el visualizador "Priced Visualizer", se aprecian fácilmente los datos organizacionales asociados. A nivel de cantidades generales, además de los datos que ya se podían visualizar en el estándar, como por ejemplo cantidad de casos y eventos, el visualizador priced permite ver la cantidad de entidades asociadas, variables y atributos. En el caso de ejemplo se cuenta con tres entidades: application, program, y validation. Existe un total de 16 variables y 19 atributos.

A modo de ejemplo, en la figura 34 se pueden ver los datos antes mencionados y además se listan los atributos presentes en el conjunto de datos. Dentro de los datos se encuentra el atributo name, asociado a la entidad program y los valores existentes son "Erasmus", "Itesm" y "Quercus".

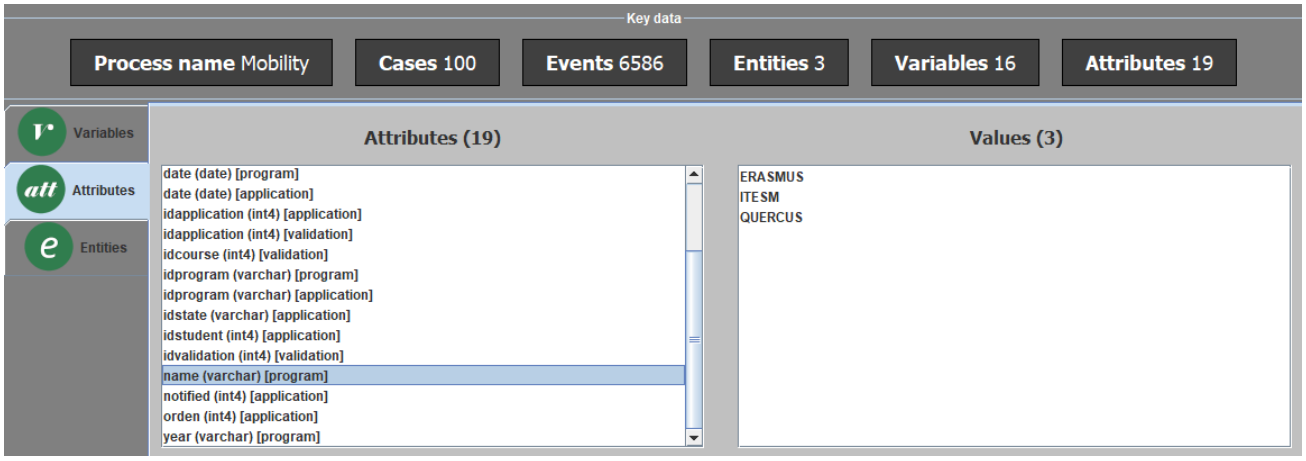

Figura 34: Visualizador priced

De la misma manera, se puede apreciar que los posibles valores para el atributo “approved” asociado a la entidad application son: “true” y “false”. En el caso de “Career” existen dos valores, “Ingeniería en computación” y “Licenciatura en informática”. Si analizamos el atributo “idstudent” vemos que existe un total de 81 estudiantes y por último inspeccionando el atributo “year” vemos que los años existentes son 10: desde 2010 al 2019.

6.4.2. Agrupamiento

Un posible primer enfoque, luego de haber realizado data profiling, es contraponer la cantidad de casos (100) contra la cantidad de estudiantes (81). Esta contraposición nos indica que hay estudiantes que han aplicado en más de una ocasión a algún llamado de movilidad. Entonces sería interesante poder determinar cuáles son los estudiantes que se encuentran en este escenario. Una posibilidad es aplicar clustering, en este caso mediante el algoritmo k-means utilizando el atributo “idstudent” y con un k=81 (la constante k determina la cantidad de clusters a obtener).

Como resultado se obtiene, por un lado, la red de petri descubierta a partir del log (figura 35). A su vez, se listan los 81 clusters obtenidos. En el caso de análisis interesa ver los clusters que contienen más de un caso, por ejemplo en el cluster número 8, se encuentran 10 casos. En la figura 36 se puede apreciar que para el estudiante con identificador 3159965 se corresponden los casos: 17, 43, 49, 52, 55,

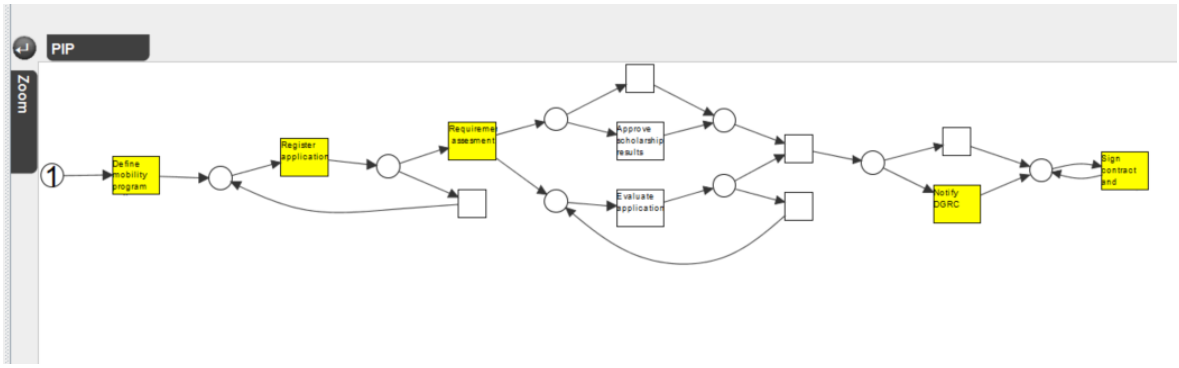

Figura 35: Red de petri descubierta

55, 56, 64, 66, 79.

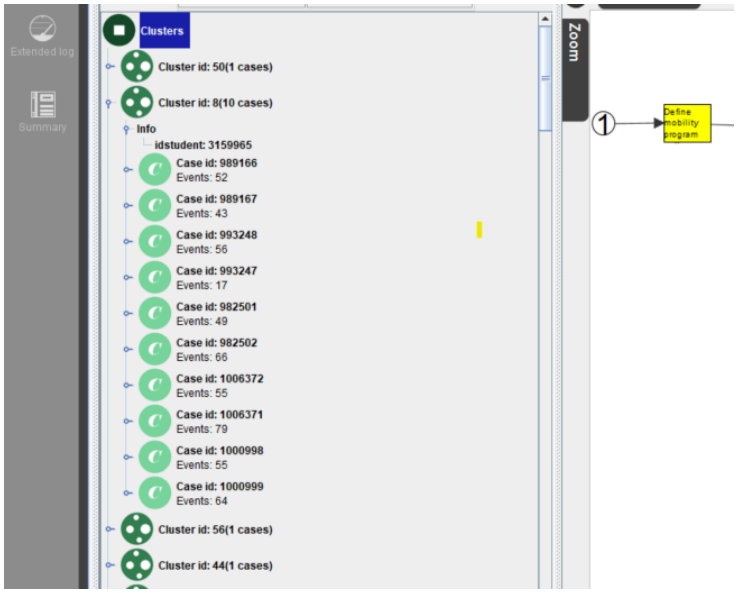

Figura 36: Clusters obtenidos a partir del atributo idstudents y k=81

Del total de 81 estudiantes existen 11 que se postularon más de una vez, dentro de esos se destaca el estudiante con identificador 3159965 y sus 10 postulaciones (Figura 36).

En la tabla 8 se muestra un resumen de los resultados obtenidos

Suponiendo que se quisiera ahondar en la particularidad del estudiante con 10 postulaciones, es posible obtener el sub log que contenga solamente los 10 casos asociados a dicho estudiante. Seleccionando el cluster número 8 y utilizando la funcionalidad "Create logs for selected cases" se puede llevar a cabo esta tarea. Utilizando este sublog en la siguiente sección se aplicará otra técnica de minería de datos de forma

Tabla 8: Resumen de resultados del agrupamiento

Estudiante	**Cantidad de casos**
3159965	10
67605695	2
59825845	2
69281187	2
11271158	2
38705940	2
56972348	2
18202902	2
48839231	2
54263655	2
39609457	2
50 restantes	1

de continuar realizando análisis del caso de estudio.

6.4.3. Reglas de asociación

Partiendo de la base del sublog obtenido anteriormente, es posible realizar nuevamente un dataprofiling. Por ejemplo, si visualizamos el atributo "idStudent" constatamos que solo toma el valor "3159965" lo cual es correcto, ya que estamos trabajando sobre un sublog que contiene solo los casos asociados al mismo. Por otro lado, si analizamos el atributo "approved", nuevamente vemos que toma como valores posibles "true" y "false". Esto indica que en las diferentes postulaciones hubo casos en que la solicitud del estudiante fue aprobada y casos en la que no. En cuanto al atributo "year" vemos que el estudiante se postuló en los años 2010,2012,2013,2015,2016,2017 y 2018. Asociado al atributo "name" que se corresponde con el nombre de los programas asociados, se encuentra que el estudiante solo se postuló a "ERASMUS" y a "QUERCUS", por lo que el estudiante nunca se postuló a un llamado de "ITESM".

Resultaría interesante ver si existe algún tipo de relación entre el programa al que se postuló, el año y si la escolaridad presentada fue aprobada o no. Para esto se puede utilizar las técnicas de reglas de asociación utilizando estos atributos y permitirá determinar si existe o no alguna relación entre los mismos.

Utilizando el algoritmo Apriori provisto y seleccionando los atributos “approved”, “name” y “year” se obtiene un total de 10 reglas. Dentro de estas 10 reglas hay dos que se destacan por la pertinencia y el nivel de confianza que se obtiene. Por un lado, se obtuvo la regla: *name=ERASMUS =>approved=true*. El valor confidence para esta regla es 1, quiere decir que en el 100 % de los casos en donde el programa es ERASMUS la escolaridad del estudiante fue aprobada.

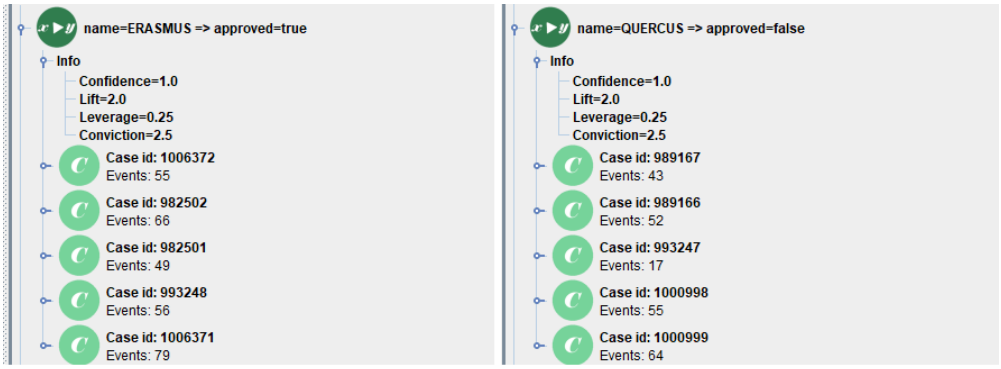

Figura 37: Reglas de asociación obtenidas

En contra parte se obtuvo la regla análoga: *name=QUERCUS =>approved=false*. Al igual que la regla anterior, el valor para confidence fue 1, lo que significa que en todos los casos en que el estudiante aplicó a un llamado QUERCUS su escolaridad fue rechazada (Figura 37).

Estas reglas podrían ser un ejemplo en donde a partir de los datos se pueden inferir algunas hipótesis, como que la escolaridad del estudiante es apta para los programas ERASMUS, pero no es suficiente para QUERCUS.

6.4.4. Árboles de decisión

A diferencia de las reglas de asociación, cuando aplicamos árboles de decisión debemos elegir un atributo, el cual será utilizado como atributo clase. Es decir, que es el atributo el cual queremos conocer las reglas existentes que permiten predecirlo. En el ejemplo del caso de estudio podríamos pensar en intentar predecir o conocer las reglas que permiten establecer a qué programa se postularán los estudiantes en base a por ejemplo el monto, carrera a la que pertenecen y el año.

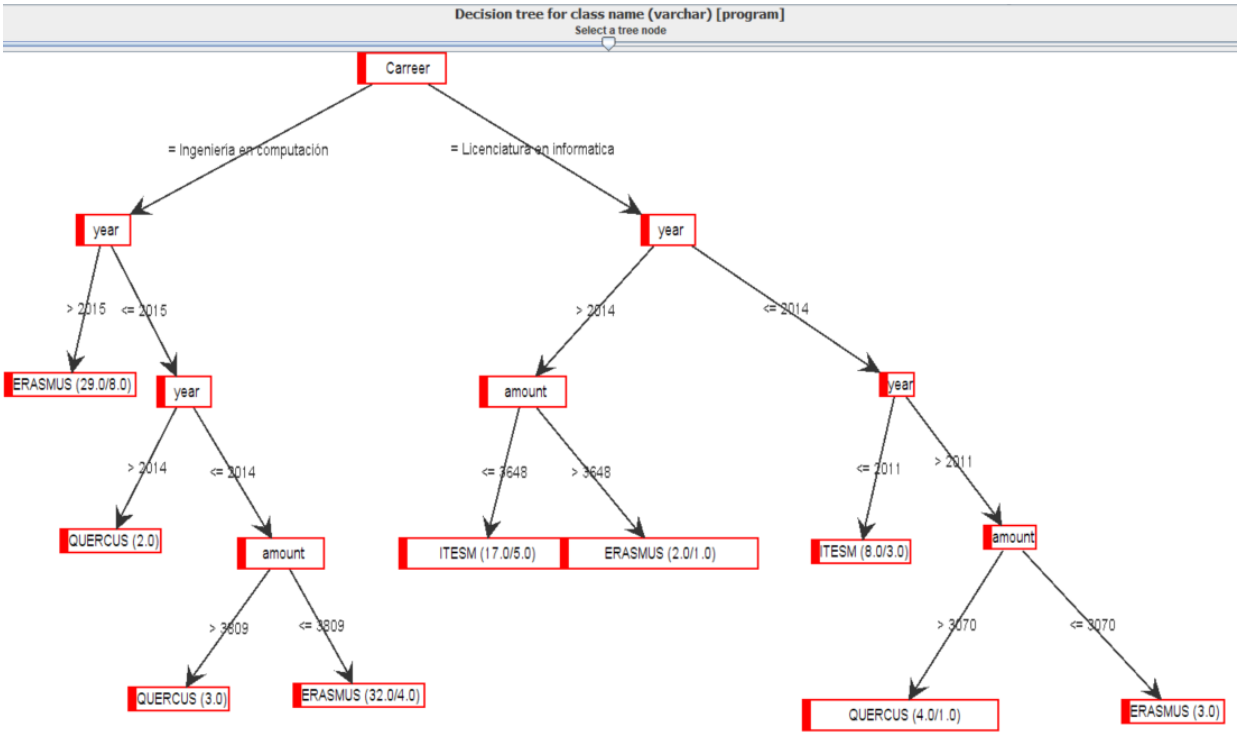

Figura 38: Árbol de decisión

Para llevar a cabo esta tarea se utilizó el conjunto de datos completo con el fin de tener resultados representativos. Una vez aplicado el algoritmo j48 y utilizando como clase el nombre de los programas, se obtuvo un árbol de profundidad 5 (figura 38). Los resultados en este caso no son tan determinísticos como en el caso de reglas de asociación, es decir que se obtuvieron relaciones que se cumplen en muchos casos pero no en el 100 % de los mismos. Esto no significa que haya un error, simplemente puede ser fruto de la aleatoriedad de los datos o que la en la realidad no exista una regla determinista. Para cada hoja del árbol se puede conocer el total de instancias que cumplen con las restricciones asociadas al camino y cuántas de estas no cumplen con la clase seleccionada. Esto se puede ver en la etiqueta de la hoja con el formato total-instancias/instancias-catalogadas-incorrectamente.

De todos modos nos permite conocer relaciones existentes entre los datos. Por ejemplo, cuando la carrera es Ingeniería en computación, se encontró que de los 29 casos de postulaciones de estudiantes que pertenecían a esta, si el año era mayor a 2015, 21 aplicaban al programa ERASMUS. Sin embargo, si el año era menor a 2014, el atributo amount cobra importancia en la determinación. En cambio, cuando la carrera

era licenciatura en informática, las reglas son más diversas, empiezan a depender del año y del monto asociado al llamado.

6.5. Conclusiones del caso de estudio

Dado el log extendido asociado al caso de estudio, se mostraron los tres posibles enfoques de minería de datos inmersos en un entorno de minería de procesos. Como puerta de entrada al análisis se muestra las posibilidades de dataprofiling a través del visualizador, esto permite conocer los datos y plantear posibles enfoques para analizarlos. Luego se muestra una aplicación de técnicas en forma iterativa, partiendo desde un conjunto de datos amplio y donde mediante la selección basada en resultados de minería de datos se obtenían subconjuntos de trazas homogéneas. Este resultado se utilizó como punto de partida de un nuevo análisis. De esta forma se encontraron patrones y relaciones entre los datos, con diferente nivel de certeza, pero que permiten diseñar diferentes hipótesis y conclusiones acerca de los datos.

En estas pruebas prácticas se puso foco en el alcance posible a través de la minería de datos, pero entre cada ciclo de análisis es posible utilizar cualquier herramienta provista nativamente por Prom con los enfoques tradicionales de la minería de procesos: Discovery, Conformance o Enhacement. Es por esto, que mediante la utilización del plugin se puede considerar que el análisis de los datos se realiza en forma integral, sin importar si son datos propios de los procesos o datos organizacionales, enriqueciendo y ampliando el universo de conclusiones que se pueden extraer a partir de los datos.

7. Resultados

En este capítulo se presentan los resultados obtenidos en el trabajo basándose en el objetivo general y a los cinco objetivos específicos planteados. El primer objetivo específico establecía la necesidad de realizar un relevamiento del estado del arte en el abordaje de la minería de procesos y minería de datos en forma integrada. Para cumplir con dicho objetivo se realizó una revisión sistemática, de donde se obtuvieron diez estudios primarios. De dichos estudios se pudo conocer cuáles eran los principales enfoques existentes acerca de la temática. Dicha revisión fue publicada y presentada en la Conferencia Iberoamericana de Ingeniería de Software (CibSE) edición 2021. Además, como producto de la revisión se pudo cumplir también con el objetivo específico número dos, que planteaba la necesidad de evaluar y analizar las técnicas existentes para el análisis y transformación de los datos en información a partir de la minería de datos y minería de procesos.

El tercer objetivo planteaba el análisis de las diferentes herramientas existentes, de las cuales PROM y WEKA son las que propiciaban el mayor beneficio gracias a que ambas son plataformas OpenSource y brindaban las herramientas con implementaciones de los algoritmos relevados en la revisión sistemática. Además, ambas herramientas compartían un mismo entorno utilizando las mismas tecnologías, lo que facilita la adaptación e integración entre sí.

El cuarto objetivo especifico planteaba definir una estrategia para tener un enfoque integral que permita obtener información de valor asociada a datos de los procesos y de los sistemas organizacionales. Se definió e implementó un prototipo que brinda las herramientas necesarias para poder llevar a cabo un análisis integral de la información utilizando minería de procesos y minería de datos en forma complementaria.

El último objetivo específico suponía aplicar la propuesta del objetivo anterior a un caso de estudio con datos reales con el fin de validar la propuesta. Esto no fue posible, ya que no se pudo contar con datos reales que cumplan con las necesidades del proyecto. Cabe destacar, que por motivos ajenos no se tuvo acceso a conjuntos de datos a los que al inicio del proyecto se tenía intenciones de acceder. Por este

motivo se realizó el análisis sobre un caso de estudio existente, producto de trabajos previos. A pesar de no contar con datos reales, el caso de estudio permitió probar y validar la herramienta y además dar una muestra del potencial de la misma.

Como resultado general, basado en cada uno de los objetivos específicos, se puede afirmar que se cumple con el objetivo general de definir una propuesta que permita el análisis de datos integrados de procesos y organizacionales como parte del framework de ciencia de datos que enmarca a este trabajo.

8. Conclusiones

En la actualidad, las organizaciones cuentan con un amplio abanico de oportunidades y posibilidades de crecimiento gracias a los avances tecnológicos, particularmente en los sistemas de información. Estos avances permitieron en las últimas décadas digitalizar y por ende automatizar gran cantidad de aspectos de las mismas. De esta forma, las organizaciones pueden poner foco en sus objetivos, permitiendo optimizar sus tareas. Además, estos avances provocan que exista mayor competitividad y, por lo tanto, mayores exigencias a nivel de gestión y optimización de sus procesos y recursos. Para llevar a cabo estas tareas se utilizan variados sistemas de información, como los son los ERP, CRMs y un amplio abanico de tipos de sistemas. Además, se suman los diferentes procesos de negocio, mediante los cuales se modelan las diferentes realidades. Todos estos sistemas generan datos, datos que son el fiel reflejo de la operativa diaria y poder interpretarlos es una tarea esencial a la hora de que los tomadores de decisiones puedan tener una visión general.

La Ciencia de Datos a lo largo del tiempo ha diseñado diversas herramientas que permiten llevar a cabo la tarea de análisis de datos. Cada fuente de datos tiene sus particularidades, lo que hace que la tarea de análisis sea todo un desafío. Si diferenciamos entre los sistemas de información tradicionales y los sistemas de gestión de procesos de negocio, las metodologías existentes para convertir los datos en información hoy en día disponibles son considerablemente diferentes. Por un lado, se encuentra la minería de datos, donde se encuentran metodologías de análisis y extracción de información de patrones más tradicionales, donde la fuente de datos generalmente se considera tabular, lo que permite realizar análisis sobre los datos proveniente de los sistemas de información más tradicionales. Por otro lado, se encuentra la minería de procesos, diseñada a partir de las particularidades de los datos generados a través de ejecución de procesos de negocio. Si bien cuenta con puntos en común con la minería de datos, el enfoque es totalmente diferente y los requerimientos asociados a los datos necesarios significan una diferencia considerable. Esto obliga a considerar en forma independiente el análisis de datos dependiendo del tipo

de fuente de información.

En este trabajo de tesis de maestría, se realiza una propuesta para lograr tener una visión única de todos los datos de una organización, ya que en la práctica, la realidad de una organización es una sola, sin importar si el modelado de la misma fue realizado en un sistema de gestión de procesos o en un sistema de información tradicional. Como primer objetivo, se planteó el relevamiento de propuestas que integren las dos visiones como una sola. Para esto se realizó una revisión sistemática de artículos existente que tengan como foco a la minería de datos y procesos en forma integral. Se consultaron las principales fuentes de artículos científicos que existen en la actualidad y se clasificaron los resultados en base a diferentes criterios. En total, como resultado de la búsqueda a través de la pregunta de investigación se obtuvieron, 2868 resultados, en donde 32 artículos fueron clasificados como relevantes y finalmente 10 fueron los estudios clasificados como primarios.

Estos estudios primarios fueron analizados en base a cinco perspectivas. Por un lado, se analizó el tipo de propuesta donde se buscó determinar si los artículos describían propuestas formales y generales o si planteaban una metodología no tan formal o general, donde quizás era aplicada a algún escenario especifico. Por otro lado, se buscó la presencia de algoritmos para obtener el origen del los mismos, es decir, conocer si el origen era la minería de datos o minería de procesos. También se buscaron aspectos relacionados a la integración de los datos, ya sean logs de eventos o datos organizacionales. Además, se clasificaron en base al tipo de minería, para conocer si los algoritmos utilizados pertenecían a la minería descriptiva o predicativa. Por último se analizaron aspectos relacionados a la calidad de datos en los enfoques de minería utilizados.

Como principales resultados de esta revisión sistemática se obtuvo que no existe un gran número de propuestas que tomen con una perspectiva integral a la minería de datos y minería de procesos. La mayoría de las aplicaciones se basaban fuertemente en uno de los dos y utilizaba al otro en alguna etapa puntual de la propuesta. La aplicación principal de los algoritmos fue de algoritmos de minería de datos sobre logs de eventos como etapa previa al descubrimiento de modelos. También algunos

casos de predicción de trazas en base a ejecuciones pasadas. Como resultado de esta revisión se presentó el artículo denominado “Process mining and data mining integration frameworks for evidence-based business intelligence: a systematic review” en la Conferencia Iberoamericana de Ingeniería de Software (CibSE) edición 2021 [32].

Basándose en la revisión sistemática, se constata la necesidad vigente de poseer herramientas que permitan la utilización de la minería de datos y de procesos en forma integral. Este trabajo se enmarca en un amplio proyecto donde se realiza una propuesta de Framework denominada PRICED (Process and Data sCience for oRganIzational improvEment)[12]. En dicha propuesta se presentan lineamientos y definiciones que abarcan las principales fases desde la generación de los datos, se toma en cuenta la calidad y procesos de depuración, integración y finalmente la explotación de los mismos. Uno de los productos generados es el denominado LogExtendido, que consiste en un log .XES que contiene en forma integrada los datos organizacionales y los datos asociados a la ejecución de procesos. Sobre la base de este log, se realizó una propuesta de herramienta donde se integran las técnicas disponibles de minería de procesos y minería de datos, donde se toma como única fuente de datos este log extendido. De forma de validar de forma la viabilidad de dicha propuesta, se implementó un plugin para Prom 6.9 que permite integrar al análisis tradicional de minería de procesos el análisis de datos utilizando los algoritmos disponibles en WEKA.

De esta manera, mediante la revisión sistemática se constató la falta de propuestas en el área y en los pocos estudios que se obtuvieron se analizaron los principales enfoques sobre los cuales se ha trabajado. También, se realizó una propuesta de herramienta general que utilice una visión integral de los datos sin importar si los mismos tienen origen en sistemas tradicionales o sistemas de gestión de procesos que fue implementada y materializada como un plugin para la herramienta Prom. Finalmente, mediante un caso de estudio se comprobó la validez del prototipo de plugin implantado y se muestra el potencial de la herramienta. Si bien esta implantación es un prototipo, fue diseñado de forma de poder extenderlo fácilmente para

incorporar nuevos algoritmos y funcionalidades.

8.1. Trabajo futuro

En este trabajo se presentaron las bases para posibles evoluciones, donde por un lado se realizó una revisión sistemática que eventualmente puede ser volver a realizarse de forma de comparar la evolución de los trabajos científicos asociados a la temática de esta tesis. En cuanto al prototipo implementado, el alcance actual es acotado. Se podrían considerar dos grandes enfoques, por un lado, ampliar la cobertura de algoritmos de minería de datos. Por ejemplo, incluyendo nuevas implementaciones de algoritmos provistos por Weka. También podrían considerarse integrar otras herramientas que provean algoritmos o incluso herramientas en la nube que en la actualidad poseen grandes ventajas frente a sistemas locales. Por otro lado, es posible potenciar las interacciones disponibles para el usuario a través del plugin, como posibles manipulaciones o posibilidad de aplicar filtros a los datos organizacionales.

Referencias

[1] R project, https://www.r-project.org

[2] Rapidminer, https://rapidminer.com/

[3] van der Aalst, W.: Process Mining: Data Science in Action, 2nd.ed. Springer (2016)

[4] Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of International Conference on Very Large Databases. pp. 487–499 (1994)

[5] Arthur, D., Vassilvitskii, S.: k-means++: The advantages of careful seeding (2006)

[6] Augusto, A., Conforti, R., Dumas, M., La Rosa, M., Maggi, F.M., Marrella, A., Mecella, M., Soo, A.: Automated discovery of process models from event logs: Review and benchmark. IEEE Trans. Knowledge &Data Eng. **31**(4), 686–705 (2018)

[7] Bevacqua, A., Carnuccio, M., Folino, F., Guarascio, M., Pontieri, L.: A data-driven prediction framework for analyzing and monitoring business process performances. In: Int. Conf. on Enterprise Inf. Systems (ICEIS). pp. 100–117. Springer (2012)

[8] Borges, A.; Artus, A.: Modelos y algoritmos para minería de procesos y datos (2021)

[9] Bui, D.B., Hadzic, F., Potdar, V.: A framework for application of tree-structured data mining to process log analysis. In: International Conference on Intelligent Data Engineering and Automated Learning. pp. 423–434. Springer (2012)

[10] Cardoso, J.: Applying data mining algorithms to calculate the quality of service of workflow processes. In: Intelligent Techniques and Tools for Novel System Architectures, pp. 3–18. Springer (2008)

[11] Chang, J.F.: Business process management systems: strategy and implementation. Auerbach Publications (2016)

[12] Delgado, A., Calegari, D.: Towards a unified vision of business process and organizational data. In: 2020 XLVI Latin American Computing Conference (CLEI). pp. 108–117 (2020). https://doi.org/10.1109/CLEI52000.2020.00020

[13] Delgado, A., Marotta, A., González, L., Tansini, L., Calegari, D.: Towards a data science framework integrating process and data mining for organizational improvement. In: 15th Int. Conference on Software Technologies (ICSOFT) (2020)

[14] Eibe Frank, M.A.H., Witten, I.H.: The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"(2016)

[15] Ezpeleta, J., Fabra, J., Álvarez, P.: On the use of log-based model checking, clustering and machine learning for process behavior prediction. In: 5th Int. Conf. Social Networks Analysis, Mgmt. and Security (SNAMS). pp. 209–214. IEEE (2018)

[16] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: The kdd process for extracting useful knowledge from volumes of data. Communications of the ACM **39**(11), 27–34 (1996)

[17] Furht, B., Villanustre, F.: Introduction to big data. In: Furht, B., Villanustre, F. (eds.) Big Data Technologies and Applications, pp. 3–11. Springer (2016)

[18] Group, X.W., et al.: Ieee standard for extensible event stream (xes) for achieving interoperability in event logs and event streams. IEEE Std **1849**, 1–50 (2016)

[19] Hernández Orallo, J., Ferri Ramirez, C., Ramirez Quintana, M.J.: Introducción a la Minería de Datos. Pearson Prentice Hall (2004)

[20] Hitpass, B.: BPM: Business Process Management: Fundamentos y Conceptos de Implementación 4a Edición actualizada y ampliada. Dr. Bernhard Hitpass (2017)

[21] IEEE: Task Force Data Science and Adv. Analytics. http://www.dsaa.co/ (2020)

[22] Kitchenham, B.: Procedures for performing systematic revs. (2004)

[23] Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering (2007)

[24] Križanić, S.: Educational data mining using cluster analysis and decision tree technique: A case study. International Journal of Engineering Business Management **12**, 1847979020908675 (2020)

[25] Leemans, S.J., Fahland, D., Van Der Aalst, W.M.: Process and deviation exploration with inductive visual miner. BPM (demos) **1295**(8) (2014)

[26] Murata, T.: Petri nets: Properties, analysis and applications. Proceedings of the IEEE **77**(4), 541–580 (1989)

[27] OMG: Business Process Model and Notation (BPMN) v2.0 (2011)

[28] Ong, K.L., De Silva, D., Boo, Y.L., Lim, E.H., Bodi, F., Alahakoon, D., Leao, S.: Big Data Applications in Engineering and Science, pp. 315–351. Springer (2016)

[29] Quinlan, R.: C4.5: Programs for machine learning morgan kaufmann publishers inc. San Francisco, USA (1993)

[30] Radeschütz, S., Mitschang, B., Leymann, F.: Matching of process data and operational data for a deep business analysis. In: Enterprise Interoperability III. pp. 171–182. Springer (2008)

[31] Roudjane, M., Rebaïne, D., Khoury, R., Hallé, S.: Real-time data mining for event streams. In: 22nd Int. Enterprise Distributed Object Computing Conference (EDOC). pp. 123–134. IEEE (2018)

[32] Rubio, M., Delgado, A., Tansini, L.: Process mining and data mining integration frameworks for evidence-based business intelligence: a systematic review. In:

24th Iberoamerican Conference on Software Engineering, CIbSE 2021, San Jose, Costa Rica, August 20 - September 3, 2021. pp. 28–41. Curran Associates (2021)

[33] Song, M., Günther, C.W., Van der Aalst, W.M.: Trace clustering in process mining. In: Int. Conf. Business Process Management (BPM). pp. 109–120. Springer (2009)

[34] Teinemaa, I., Dumas, M., Maggi, F.M., Di Francescomarino, C.: Predictive business process monitoring with structured and unstructured data. In: International Conference on Business Process Management (BPM). pp. 401–417. Springer (2016)

[35] Weske, M.: Business Process Management - Concepts, Languages, Architectures, Third Edition. Springer (2019)

[36] Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., DATA, M.: Practical machine learning tools and techniques. In: Data Mining. vol. 2 (2005)

[37] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B.: Experimentation in Software Engineering. Springer (2012)

[38] van Zelst, S.J., Cao, Y.: A generic framework for attribute-driven hierarchical trace clustering. In: International Conference on Business Process Management (BPM). pp. 308–320. Springer (2020)