Selección de algoritmos de preprocesamiento de datos del Hospital Delicia Concepción Masvernat (Concordia, provincia de Entre Ríos) que permita el desarrollo de un componente de software para predicción de enfermedades cardiológicas.

María Elizabeth Silva Layes[1], Marcelo Gabriel Benedetto[1], Duval Horacio Benítez[1], Elio Darío Costen[1], Joaquín Diez[1], Juan José Aguirre[1], Marcelo Alejandro Falappa[2], Jesús Fabián Frola[1]

[1] Facultad de Ciencias de la Administración – Universidad Nacional de Entre Ríos
Monseñor Tavella 1424 – Concordia, Entre Ríos (3200) - Tel.: +54(0345)4231433
{elizabeth.silva, marcelo.benedetto, horacio.benitez, elio.costen, juan.aguirre, fabian.frola}@uner.edu.ar, joako.10.diez@gmail.com

[2] Departamento de Ciencias e Ingeniería de la Computación - Universidad Nacional del Sur
Avenida Alem 1253 - Bahía Blanca (B8000CPB) - Tel.: +54(0291)4595135
mfalappa@cs.uns.edu.ar

Resumen. El sector sanitario, sin lugar a dudas es uno de los ámbitos en el que se administran grandes volúmenes de datos; principalmente en el área clínica. Esto conduce a identificar una importante necesidad de encontrar maneras de administrar, integrar, analizar e interpretar ese gran conjunto de datos; procurando identificar patrones de comportamiento que sean de utilidad en la toma de decisiones médicas. El proyecto de investigación[1] en el que se enmarca este artículo plantea como principal objetivo desarrollar un componente de software capaz de generar, con aprendizaje automatizado, un modelo con capacidades predictivas sobre enfermedades cardiológicas; que permita un mejor soporte a decisiones de diagnóstico clínico y un avance significativo en la medicina preventiva. Este artículo presenta una revisión exhaustiva de las herramientas de preprocesamiento de datos para analizar datos sanitarios masivos, en términos de la imputación de valores perdidos, detección de valores atípicos, reducción, escalado, transformación y partición de datos. Además, se proponen herramientas de ciencia de datos en el campo sanitario. Se ha presentado un análisis en profundidad para describir los pros y los contras de las herramientas existentes para abordar los desafíos prácticos. Los resultados obtenidos son útiles para el desarrollo de investigaciones basadas en predicción de enfermedades en el campo sanitario.

[1] Este trabajo fue realizado por la Universidad Nacional de Entre Ríos (UNER) y la Universidad Nacional del Sur (UNS), (PID 7060 - "Bioingeniería informática aplicada a la predicción de enfermedades cardiológicas y su implementación en el Hospital Delicia Concepción Masvernat de la ciudad de Concordia, provincia de Entre Ríos").

Palabras claves: Herramientas de Preprocesamiento, Inteligencia Artificial, *Machine Learning*, Sistemas de Soporte a Decisiones Clínicas.

1 Introducción y Motivaciones

La inteligencia artificial es una herramienta que nos ayuda en el análisis de datos, sirviendo de apoyo tanto en el área de cuidados de la salud como en la investigación médica a través de su aplicación, que abarca desde el diagnóstico médico hasta la capacidad para analizar datos desde una variedad de perspectivas tendientes a descubrir patrones ocultos.

Actualmente, existen diversas investigaciones que permiten encarar tratamientos de enfermedades específicas, aunque siguen siendo estudios puntuales que, si bien tienen muy buenos resultados, no son de aplicabilidad en la práctica cotidiana de consultas médicas. Teniendo en cuenta esta realidad, la investigación que se plantea persigue lograr el desarrollo de un componente de software que permita incorporar la predicción de probabilidad de riesgos de enfermedades cardíacas como sistema de soporte a las decisiones clínicas incorporada al acto asistencial; así como también en el proceso de gestión de medicina preventiva. Si bien existen diferentes proyectos aplicados que permiten conocer sobre la probabilidad de riesgos, dichos proyectos no terminan siendo integrados a la Historia Clínica Electrónica, de manera de tener una aplicación inmediata en el acto asistencial.

La primera etapa del proyecto se refiere a (1) la identificación de las variables a ser consideradas en el *dataset* sobre el que se trabajará, y (2) la selección, refinamiento y preprocesamiento de los datos.

Respecto al punto 1), hemos podido identificar las variables vinculadas al dominio del problema planteado; para lo cual trabajamos con expertos en el área de cardiología, con el objetivo de seleccionar las variables con las que comenzaremos a trabajar.

Referente al punto 2), el preprocesamiento de datos, es relevante destacar que su correctitud es fundamental en el proceso de descubrimiento de información o *KDD* (*Knowledge Discovery in Databases*).

La calidad de los datos utilizados en *Machine Learning* está incidida por diferentes factores, como el caso de la precisión, integridad, consistencia, puntualidad, credibilidad e interpretabilidad de estos factores [1]. La inexactitud e inconsistencia de los datos son bastante comunes en las grandes bases de datos que se manejan en el mundo real, ya sea por tener datos con valores erróneos, incompletos, duplicados, o por carecer de ellos. Estos problemas pueden deberse a diversos factores, entre ellos errores humanos o informáticos producidos al ingreso o transmisión de los datos, así como también, al momento de integrar datos de diferentes fuentes.

El preprocesamiento de datos involucra diferentes pasos a tener en cuenta: la limpieza, integración, reducción y transformación de datos. La limpieza de datos hace referencia a la completitud de los valores faltantes, la identificación o eliminación de valores atípicos, así como también, la resolución de las inconsistencias que se puedan presentar. Por otro lado, la integración de los datos se refiere a la integración de diferentes bases de datos, y archivos. La reducción de datos tiene como objetivo

obtener un menor volumen de datos, pero cuyos resultados analíticos sean los mismos o muy cercanos a los del volumen inicial, a la que también se aplican técnicas de compresión, selección de atributos, así como también, construcción de nuevos atributos; teniendo en cuenta también la reducción de numerosidad, es decir, el reemplazo de datos, utilizando modelos paramétricos para representaciones más pequeñas. Mientras que en la transformación de datos, se hace referencia a la normalización, discretización, y jerarquizaciones posibles de los datos.

Teniendo presente que la preparación de los datos consume entre un 60 y 80 % [2] de un proyecto analítico, es de vital importancia seleccionar las herramientas adecuadas para efectuar el preprocesamiento de los datos del proyecto. Por tal motivo, el paso inicial es efectuar la evaluación de diferentes herramientas de software que nos permitan llevar a cabo de la mejor manera posible cada una de las instancias que forman el preprocesamiento de datos.

2 Metodología para el análisis de herramientas de preprocesamiento de datos. Criterios y características involucradas

Cada una de las herramientas seleccionadas son analizadas teniendo como base estudios previos realizados por expertos sobre herramientas utilizadas para realizar minería de datos [3][4]. Con esta base se plantean cuatro categorías principales a evaluar: *performance*, funcionalidad, usabilidad, soporte de determinadas tareas, a la que se le sumó la evaluación de otras dos categorías: documentación, y actualización. Si bien nos basamos mayoritariamente en considerar el análisis de los criterios planteados en [3][4] para las diferentes categorías, algunos de ellos fueron descartados para esta etapa, y se consideraron otros referidos específicamente a la etapa de preprocesamiento de datos.

A continuación, se presentan los criterios previstos para cada una de las categorías:

Tabla 1: Criterios de *Performance*

Criterio	Descripción
Plataformas en las que corre	¿El software se ejecuta en una amplia variedad de plataformas informáticas?
Requerimientos del sistema para ejecutarse	Requerimientos mínimos necesarios para poder ejecutarse.
Arquitectura del *software*	¿El *software* utiliza una arquitectura cliente-servidor o una arquitectura independiente? ¿Puede el usuario elegir entre las diferentes arquitecturas? ¿Puede la herramienta trabajar *off line*?

Acceso a datos heterogéneos	¿Qué tan bien interactúa el *software* con una variedad de fuentes de datos? ¿Requiere algún *software* auxiliar para hacerlo?
Tamaño de los datos	¿Qué tan bien se adapta el *software* a grandes conjuntos de datos? ¿El rendimiento es lineal o exponencial? ¿Qué tan bien funciona el *software* con conjuntos de datos a gran escala?
Eficiencia	¿Los resultados producidos por el *software* son en tiempos razonables de acuerdo al tamaño del conjunto de datos?
Interoperabilidad	¿La herramienta interactúa fácilmente con otras herramientas de soporte de *KDD*? Si es así, ¿qué tan compleja es esa interacción?
Robustez	¿La herramienta funciona de manera consistente sin fallar?

Tabla 2. Criterios de Funcionalidad

Criterio	**Descripción**
Técnicas para preprocesamiento de datos	Variabilidad de técnicas que posee para la etapa de preprocesamiento de datos.
Flexibilidad en el tipo de datos	¿Permite manejar una amplia variedad de tipos de datos?
Exportación de datos	¿Permite exportar los datos preprocesados en diferentes formatos?
Problemas que permite resolver	¿Permite la herramienta resolver una variedad de problemas de limpieza de datos?
Modelado	Funcionalidades de modelado que soporta la herramienta (por ejemplo *pipeline*, *workflows*).

Tabla 3. Criterios de Usabilidad

Criterio	**Descripción**

Interface de usuario	¿Es la interfaz de usuario de fácil uso?
Curva de aprendizaje	¿La herramienta es fácil de aprender? ¿La herramienta fácil de usar de manera correcta?
Visualización de los datos	¿Qué tan bien presenta la herramienta los datos?
Tipo de usuarios	¿Puede el *software* ser utilizado por usuarios con diferentes niveles de experticia?
Reporte de errores	¿Qué tan significativo es el informe de errores? ¿Qué tan bien ayudan los mensajes de error al usuario a depurar los problemas?

Tabla 4. Criterios soporte de tareas determinadas

Criterio	**Descripción**
Modificación de valores falsos	¿Qué tan bien el software permite detectar/modificar valores falsos en los datos?
Sustitución de datos	¿Permite la sustitución global de un dato por otro de acuerdo a determinados criterios?
Filtrado de datos	¿Permite el software la selección de un conjunto de datos de acuerdo a un criterio determinado?
Eliminación de datos	¿Permite la herramienta la eliminación de todos los registros que pueden ser incorrectos?
Atributos derivados	¿Permite la creación de atributos derivados basados en los atributos existentes?

Tabla 5. Criterios de Documentación

Criterio	**Descripción**
Calidad, cantidad e idiomas	¿Qué profundidad tiene la documentación existente? ¿Existe una vasta documentación acerca de la herramienta?

	¿La documentación se encuentra disponible en diferentes idiomas?
Revisiones	¿Con qué periodicidad se publican revisiones y actualizaciones?
Comunidad	¿Existe una vasta comunidad de apoyo?

Tabla 6. Criterios de Actualización

Criterio	**Descripción**
Frecuencia	¿Con qué frecuencia se realizan actualizaciones de la herramienta?
Soporte técnico	¿El soporte técnico es adecuado?
Bug tracking	Posee mecanismos, procedimientos, análisis y/o respuestas para la detección de errores?

Cada uno de estos criterios son evaluados mayoritariamente de forma subjetiva, identificando cuáles y cómo son satisfechos cada una de ellos. Dichos criterios fueron valorados considerando dos escalas, dependiendo de las características a evaluar. Se utilizó para algunos de éstos la evaluación de acuerdo a una escala de 1 a 5, de acuerdo a la siguiente ponderación:

1. No cumple/No posee/No adecuado.
2. Cumple/posee/adecuado de manera suficiente.
3. Cumple/posee/adecuado de manera satisfactoria.
4. Cumple/posee/adecuado de manera muy satisfactoria.
5. Cumple/posee/adecuado de manera excelente.

Como segunda escala se consideró la evaluación 1- SI, y 2- NO, de manera de establecer la presencia o ausencia del criterio analizado.

En una segunda instancia, a cada criterio se le asignó un peso relativo dentro de la categoría, y finalmente un peso relativo a cada una de las categorías evaluadas, tomando como referencia los pesos relativos considerados por Ken Collier [3] y Luna [5].

Se presentan a continuación los pesos establecidos para los criterios utilizados para cada característica y criterios asociados.

Tabla 7. Pesos establecidos para los criterios

Característica	**Criterio**	**Peso**

Performance (0.30)	Plataformas en las que corre	0.05
	Requerimientos del sistema para ejecutarse	0.10
	Arquitectura del *software*	0.05
	Acceso a datos heterogéneos	0.10
	Tamaño de los datos	0.20
	Eficiencia	0.15
	Interoperabilidad	0.05
	Robustez	0.30
Funcionalidad (0.20)	Técnicas para preprocesamiento de datos	0.25
	Flexibilidad en el tipo de datos	0.15
	Exportación de datos	0.15
	Problemas que permite resolver	0.25
	Modelado	0.20
Usabilidad (0.30)	Interface de usuario	0.25
	Curva de aprendizaje	0.20
	Visualización de los datos	0.25
	Tipo de usuarios	0.15
	Reporte de errores	0.15
Soporte de determinadas tareas (0.10)	Modificación de valores falsos	0.25
	Sustitución de datos	0.20
	Filtrado de datos	0.25
	Eliminación de datos	0.10
	Atributos derivados	0.20

Documentación (0.05)	Calidad, cantidad e idiomas	0.35
	Revisiones	0.25
	Comunidad	0.40
Actualización (0.05)	Frecuencia	0.40
	Soporte técnico	0.35
	Bug tracking	0.25

3 Selección y descripción de los algoritmos

Primariamente se consideraron para su evaluación 6 herramientas de *software* que permiten preprocesamiento de datos:

- *KNIME*
- *RapidMiner*
- *WEKA*
- *Talend Open Studio*
- *Data Civilizer*
- *Data Cleaner*

La preselección de estas herramientas se realizó considerando principalmente que estos fueran "*open source*". Esta decisión se basó en el acceso necesario a las herramientas durante el período que dura el proyecto, considerando que muchas herramientas propietarias limitan su acceso a un período de tiempo, a conjuntos de datos de determinado tamaño, o restringen el uso de determinadas funcionalidades.

Luego de efectuar un análisis preliminar de cada una de las herramientas, se decidió en desestimar las herramientas *Data Civilizer* y *Data Cleaner*, por las razones que se esgrimen a continuación.

Para el caso de *Data Civilizer* y en base a la información recopilada desde distintas fuentes podemos determinar que este *software* no registra actualizaciones recientes siendo la última hace dos años (2019). Adicionalmente esta herramienta no cuenta con un soporte técnico razonable, escasa documentación y de baja calidad, incompatibilidad con otras herramientas *KDD*, bajo desempeño y un alto grado de dificultad al momento de comenzar a adquirir conocimientos en este *software* [13][14][15].

Respecto a *Data Cleaner*, es un proyecto del gobierno argentino por el cual su equipo de datos abiertos está creando herramientas que administren metadatos para los catálogos publicados por las diferentes oficinas públicas argentinas. Dicha herramienta es una librería en *Python* aplicada a la limpieza de datos, según

estándares del Equipo de Datos Argentina. Su primera versión "0.1.0" fue publicada el 18 de febrero del 2016, siendo sus últimas versiones "0.1.18" del 30 de abril del 2016, "0.1.20" del 28 de diciembre del 2018, "0.1.21" del 18 de marzo del 2019 y 0.2.0 del 8 de agosto del 2019.

Teniendo en cuenta la documentación relevada de la herramienta y su historial de versiones, se puede determinar que este paquete aún se encuentra en una etapa temprana de desarrollo [16][17][18].

A continuación, se realiza un resumen de cada una de las herramientas evaluadas.

KNIME (*Konstanz Information Miner*) es un software de manipulación de datos basado en la idea de nodos y conectores, haciéndolo totalmente visual en la construcción de flujos de preprocesamientos y análisis de datos. Cada nodo encapsula distintos tipos de algoritmos para trabajar sobre tablas de datos. A través de "*plugins*" *KNIME* se integra con varios proyectos o*pen source* tales como *Python*, R y *WEKA*, además de trabajar con *Deep Learning*, BI, H2O y *Tableau*. El carácter abierto de *KNIME* hace posible su extensión mediante la creación de nuevos nodos que implementan algoritmos a medida del usuario [6][7].

Es una herramienta *open source* que puede ser descargada y utilizada gratuitamente bajo los términos de licenciamiento GPL con una versión de pago para soluciones inmediatas. La compañía que lo desarrolla ofrece, adicionalmente, la posibilidad de contratar servicios de soporte en varios niveles, además de brindar servicios de consultoría y formación. Posee muchas herramientas de preprocesamiento de datos. Cada nodo que implementemos está debidamente documentado con sus casos de usos y ejemplos generales [18][19].

Rapidminer es un software de análisis de datos con licencia *open source* bajo licencia AGPL-3. Provee distintas experiencias de la mano de distintas distribuciones libres como *Rapidminer Studio*, así como distribuciones que requieren el pago de licencias.

Posee funcionalidades de *machine learning*, es multiplataforma, y se encuentra construido sobre una arquitectura cliente-servidor, que se ejecuta sobre Java. Actualmente, se encuentra en su versión número 9.8, con una frecuencia alta de actualizaciones, ya sea para la corrección de errores, como también para adicionar nuevas funcionalidades.

En cuanto al preprocesamiento de los datos, este *software* provee funcionalidades para la normalización, discretización, búsqueda y reemplazo de datos; así como también la inserción de datos en lagunas de datos que se puedan encontrar dentro del conjunto de datos.

Por otra parte, el software provee de integración con los lenguajes de programación *Python* y R, una interoperabilidad que permite ejecutar scripts de estos lenguajes como parte de los procesos de preprocesamiento.

WEKA (*Waikato Environment for Knowledge Analysis*) es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Es un software de código abierto emitido bajo Licencia Pública General (GNU), desarrollado en la Universidad de Waikato, Nueva Zelanda, y escrito en lenguaje Java. Posee

herramientas para la preparación de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización. Actualmente existen una versión estable (*Weka 3.8*), y una versión de desarrollo (*Weka 3.9*). Para la etapa de preprocesamiento, *WEKA* tiene integrado una gran diversidad de filtros que permiten realizar manipulaciones sobre los datos en dos niveles: atributos e instancias [8][9].

Talend Open Studio (*TOS*) es una suite que proporciona herramientas de gestión y desarrollo unificadas para integrar y procesar datos de cualquier tipo o volumen. Esta herramienta *open source* con licencia Apache, desarrollada en Java se encuentra actualmente en la versión 8.0.1. Posee una interfaz gráfica de usuario y múltiples componentes y conectores prediseñados, que facilitan su uso. En cuanto a la documentación, para los usuarios que utilizan el software de manera gratuita, se encuentra disponible la documentación provista por la "comunidad", así como también la ayuda que la empresa habilita, con ciertas restricciones para la modalidad *open source*.

Se pueden instalar paquetes adicionales, incluidas bibliotecas de terceros y controladores de bases de datos. En caso de no existir un componente para alguna fuente en particular, se pueda construir el componente e incorporarlo a la paleta de componentes. Se destaca la existencia de un amplio abanico de componentes que permiten una gran variedad de técnicas de preprocesamiento [10][11].

Para la realización de la evaluación de las diferentes herramientas se trabajó en base a un *dataset* proporcionado por el Hospital Delicia Concepción Masvernat. Dicho conjunto de datos que se describe a continuación, forma parte de una selección primaria de variables consideradas claves por el experto del área, así como surgen del análisis de características o provienen de investigaciones previas realizadas [12]:

Tabla 8. Características del *dataset*

Variable	Tipo	Descripción
Edad	Continuo	Edad en años
Sexo	Discreto	0- Femenino, 1- Masculino
Raza	Discreto	
Diabetes	Discreto	0- SI, 1- NO
Hipertensión arterial	Discreto	0- SI, 1- NO
Fumador	Discreto	0- SI, 1- NO
Apnea	Discreto	0- SI, 1- NO
Enfermedad renal crónica	Discreto	0- SI, 1- NO
Estrés	Discreto	0- SI, 1- NO

Antecedente de enfermedad cardíaca	Discreto	0- SI, 1- NO
Enfermedad vascular periférica	Discreto	0- SI, 1- NO
Enfermedad cerebrovascular	Discreto	0- SI, 1- NO
Actividad Física	Discreto	0- SI, 1- NO
Presión arterial diastólica	Continuo	
Presión arterial sistólica	Continuo	
Frecuencia cardíaca	Continuo	
Índice masa corporal	Continuo	
PCR (Prueba de Proteína C Reactiva)	Continuo	
Creatinina	Continuo	
Proteinuria	Continuo	
Triglicéridos	Continuo	mmol/L
Colesterol LDL	Continuo	mmol/L
Colesterol HDL	Continuo	mmol/L

Las pruebas fueron realizadas en tres instancias con cantidades diferentes de datos con el objetivo de analizar la *performance* de cada una de las herramientas evaluadas:

- Primera corrida: 1.001.790 registros
- Segunda corrida: 2.114.890 registros
- Tercera corrida: 3.227.990 registros

Es de destacar que la recolección de datos ha insumido mucho tiempo, por la dificultad en sí del proceso en el ámbito sanitario, que además, fue afectado notablemente por la distorsión ocasionada por la pandemia de COVID-19 y su impacto en el sistema de salud.

4 Estudio Comparativo

Cabe destacar que desde el proyecto de investigación PID-UNER 7060, se solicitó oportunamente la adquisición de equipamiento acorde para poder trabajar con este estudio comparativo de manera más conveniente; así como también para el desarrollo

de las demás actividades planificadas. Por razones presupuestarias, hasta el momento trabajaremos con la configuración que a continuación se explicita; dejando las pruebas con el equipamiento solicitado para desarrollarlas en trabajos futuros.

Para desarrollar las pruebas de preprocesamiento de los datos médicos y efectuar el análisis comparativo, las distintas herramientas se evaluaron sobre máquinas virtuales del software *VirtualBox*© (*Hipervisor* de tipo dos). Bajo esta herramienta se realizó la instalación del sistema operativo *Windows* y las aplicaciones de preprocesamiento, sin alterar nuestro sistema base.

Las especificaciones técnicas del ambiente de prueba se detallan a continuación:
Especificaciones técnicas de la PC donde se instaló *VirtualBox*©:

- Procesador: Intel® *Core*™ i3-10100 CPU 3.60 GHz.
- Memoria RAM: 8,00 GB.
- Sistema Operativo: *Windows* 10 Pro de 64 bits.

Especificaciones Técnicas de la máquina virtual en *VirtualBox*©:

- Procesador: Intel® Core™ i3-10100 CPU 3.60 GHz.
- Memoria RAM: 4,00 GB.
- Almacenamiento: 100 GB.
- Sistema Operativo: *Windows* 10 Pro de 64 bits.

Aplicaciones para el preprocesamiento de datos instaladas en la máquina virtual:

- *Weka 3.8.5* (weka-3-8-5-azul-zulu-windows.exe).
- *Rapidminer Studio 9.9.2* (rapidminer-studio-9.9.2-win64-install.exe).
- *Knime 4.4.0* (KNIME 4.4.0 Installer (64bit).exe).
- *Talend Open Studio 7.3.1*.
- (*TOS_DI-20200219_1130-V7.3.1-windows-installer.exe*).

A continuación, y a partir del análisis realizado, se presentan dos planillas que reflejan el estudio comparativo realizado sobre las herramientas y que reflejan de manera detallada los valores obtenidos para cada una de las categorías (una corresponde a la valoración 1 a 5 y la otra a SI/NO):

Tabla 9. Estudio de valoración realizado sobre las herramientas

	Software	***Rapidminer***	***Knime***	***Weka***	***Talend Open Studio***
	Características a evaluar				
Performance	Plataformas en las que corre	*5*	*5*	5	5
	Requerimientos del sistema para ejecutarse	3	5	5	5

	Cantidad de datos y diversidad de fuentes de datos que maneja	5	5	4	5
	Eficiencia de los algoritmos de acuerdo al tamaño del *dataset*	5	4	3	5
	Interoperabilidad con otras herramientas *KDD*	5	5	5	1
	Robustez de la herramienta	5	5	5	5
	Que tan bien funciona el software con *dataset* en gran escala	5	5	4	5
Funcionalidad	Variabilidad de técnicas que posee para la etapa de preprocesamiento	4	3	5	4
	Facilidad de uso (metodología de fácil entendimiento)	5	4	5	4
	Variedad de tipos de datos que permite manejar	5	4	4	5
	Variabilidad de las formas de exportación de los datos obtenidos	5	4	4	5
	Pipelines / Workflows.	5	5	5	5
	Variedad de problemas que la herramienta puede resolver	5	5	5	5
Usabilidad	Sencillez de uso de la interfaz	5	4	5	5
	Curva de aprendizaje necesaria para aprender a utilizar la herramienta	5	4	5	4
	Visualización de los datos (claridad)	5	4	4	3
	Adecuación del *software* a diferentes tipos de usuarios	4	4	5	4

	Claridad de lectura de los mensajes de error	5	4	3	5
Tareas adicionales soportadas	Modificación de valores falsos en los datos	5	5	5	5
Documentación	Cantidad, calidad e idiomas soportados	5	4	5	4
	Periodicidad de revisiones y actualizaciones	5	4	5	5
	Comunidad de apoyo a la herramienta	5	5	4	5
Actualización	Frecuencia	5	5	4	4
	Soporte Técnico	4	3	5	3
	Bug tracking (*feedback*, análisis y procedimiento).	4	3	4	4

Tabla 10. Estudio de habilidad realizado sobre las herramientas

	Software	***Rapidminer***	***Knime***	***Weka***	***Talend Open Studio***
	Características a evaluar				
Performance	Puede la herramienta trabajar *off line*	SI	SI	SI	SI
Tareas adicionales soportadas	Permite sustitución global de un dato por otro	SI	SI	SI	SI
	Eliminación completa de registros	SI	SI	SI	SI
	Permite la herramienta agrupar datos continuos	SI	NO	SI	NO

	para mejorar la eficiencia del modelado				
	Permite la creación de atributos derivados basados en los atributos existentes	SI	SI	SI	SI
Arquitectura (cliente-servidor, corrida *stand alone*)	Arquitectura C/S, *stand alone*, ambas Si posee ambas es 2 si posee solo una es 1	1	1	1	1
Soporte para *data cleaning*	Soporte Total es 2, Soporte parcial 1	2	2	2	2

Tabla 11. Valoración general final obtenida

Software/ Características	*Performance*	Funcionalidad	Usabilidad	Tareas adicionales soportadas	Documentación	Actualización	Puntuación Total
RapidMiner	0.2355	0.14	0.291	0.1	0.05	0.044	0.8605
Knime	0.2415	0.118	0.2328	0.1	0.044	0.0380	0.7743
Weka	0.2463	0.138	0.2072	0.1	0.046	0.0435	0.7810
TOC	0.2715	0.14	0.1720	0.1	0.0465	0.0365	0.7665

5 Conclusión

Las dos planillas de evaluaciones comparativas presentadas en el apartado anterior sometidas a los análisis correspondientes al set de datos primario receptados desde el Hospital Delicia Concepción Masvernat, referidas a características consideradas relevantes para nuestro estudio y ponderadas de acuerdo a los criterios expuestos previamente permiten arribar a las siguientes conclusiones:

La característica Documentación, que refiere a la calidad, cantidad, idiomas soportados, periodicidad de revisiones y actualizaciones y comunidad de apoyo a la herramienta, es contemplada de manera excelente por la herramienta *RAPIDMINER*, mientras que las herramientas *WEKA, KNIME* y *TALEND OPEN STUDIO* la contempla de manera muy satisfactoria.

La característica Actualización, que refiere a la frecuencia, soporte técnico, mecanismos para detección de errores, procedimiento para reportarlos, análisis y/o respuestas a éstos; es contemplada de manera muy satisfactoria por las herramientas

RAPIDMINER y *WEKA*, mientras que las herramientas *KNIME* y *TALEND OPEN STUDIO* la contempla de manera satisfactoria.

La característica *Performance*, que refiere a plataformas en las que corre, requerimientos del sistema para ejecutarse, acceso a datos heterogéneos, tamaño de los datos, eficiencia, interoperabilidad y robustez es contemplada de manera muy satisfactoria por las herramientas *KNIME* y *TALEND OPEN STUDIO*, mientras que las herramientas *RAPIDMINER* y *WEKA* la contempla de manera satisfactoria.

En lo que respecta a las operaciones de carga, aplicado de filtros y grabación del nuevo set de datos se trabajó en tres instancias registrando los tiempos de las pruebas de los 4 softwares con 1.001.790, 2.114.890 y con 3.227.990 registros. En la primera instancia de prueba el mejor tiempo de respuesta registrado fue *WEKA*, seguido por *TALEND OPEN STUDIO* y *RAPIDMINER* y posteriormente *KNIME* con una diferencia importante de tiempo. En la segunda y tercera instancia, *WEKA* no pudo ejecutarse por falta de memoria, siendo el mejor tiempo de respuesta registrado por *RAPIDMINER* y *TALEND OPEN STUDIO* y nuevamente *KNIME* con una diferencia importante de tiempo.

La característica Funcionalidad, que refiere a variantes de técnicas para preprocesamiento de datos, flexibilidad en el tipo de datos, facilidad de uso, exportación de datos, problemas que permite resolver y *Pipelines* / *Workflows*, es contemplada de manera excelente por las herramientas *RAPIDMINER* y *TALEND OPEN STUDIO*, mientras que la herramienta *WEKA* la contempla de manera muy satisfactoria y *KNIME* la contempla de manera satisfactoria.

La característica Usabilidad, que refiere a interface de usuario, curva de aprendizaje, visualización de los datos, tipo de usuarios y reporte de errores, es contemplada de manera excelente por la herramienta *RAPIDMINER*, de manera muy satisfactoria por WEKA; mientras que las herramientas *KNIME* y *TALEND OPEN STUDIO* la contempla de manera satisfactoria.

La característica de Soporte a tareas específicas, que refiere a modificación de valores falsos, es contemplada de manera excelente por todas las herramientas en estudio. Todas las herramientas analizadas permiten trabajar con sustitución, filtrado, eliminación de datos, agrupamiento, atributos derivados y soporte para data cleaning, con excepción de *KNIME* y *TALEND OPEN STUDIO* en lo que respecta específicamente al agrupamiento de datos continuos para mejorar la eficiencia del modelado.

De una evaluación general, contemplando todas las características, pruebas de *performance* de los diferentes set de datos y en base a las ponderaciones establecidas con los criterios precedentes, no podemos establecer que exista una herramienta que contemple todo y de la mejor manera. Lo que sí podemos determinar son las bondades de cada una de ellas y combinar su utilización de acuerdo a las necesidades que nos plantee el proyecto en cuestión en la etapa de preprocesamiento.

Se destaca que a pesar de no haber identificado una herramienta que para cada una de las variables analizadas supere al resto, sí identificamos que Rapidminer se destaca con una puntuación general total de 0.8605, siendo superior a todas las herramientas evaluadas. Concluimos que para el proyecto que desarrollamos (PID 7060) optamos por utilizar la herramienta *Rapidminer*.

6 Trabajos Futuros

Para la continuidad de las actividades del proyecto de investigación, se requiere la realización de un estudio para la elección de los parámetros de los algoritmos y aplicar mecanismos de optimización, a fin de mejorar los resultados en general para la clasificación. Se presentó una forma de elegir las herramientas que proporcionaron mejores resultados en escenarios de datos sanitarios, validado en conjunto de datos reales estudiados exhaustivamente desde nuestra propuesta de selección de variables. Las evaluaciones empíricas realizadas nos muestran resultados satisfactorios a partir del enfoque utilizado. Asimismo, se debe concluir con la identificación de variables que se tomarán en cuenta para realizar el preprocesamiento de datos definitivo.

Referencias

1. Han, J., Kambar, M., Pai, J. Data Mining. Concepts and Techniques. 3ra. edición. Morgan Kaufmann Publishers pp. 84-84 (2012).
2. Munson, M. A.: A Study on the Importance of and Time Spent on Different Modeling Steps. En: ACM SIGKDD Exploration Newsletter 13(2), 65–71 (2012).
3. Collier, K., Carey, Sautter, D., B.,Marjaniemi, C.: A Methodology for Evaluating and Selecting Data Mining Software. En Proceedings del 32 Annual Hawaii International Conference on System Sciences. pp.11 (1999) .
4. Giraud-Carrier, C., Povel, O.: Characterising Data Mining software. Intelligent Data Analysis. Intelligent Data Analysis, 181-192. (2003).
5. Altalhi, A.H., Luna, J.M., Vallejo, M.A. and Ventura, S.: Evaluation and comparison of open source software suites for data mining and knowledge discovery. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery (7) (2017).
6. RapidMiner, https://rapidminer.com/, último acceso 01/06/2022.
7. RapidMiner Community, https://community.rapidminer.com/, último acceso 31/04/2022.
8. Weka 3: Machine Learning Software in Java, https://www.cs.waikato.ac.nz/~ml/weka/index.html, último acceso 25/04/2022.
9. Weka Wiki, https://waikato.github.io/weka-wiki/, último acceso 31/04/2022.
10. Talend, https://www.talend.com/es/, último acceso 18/02/2022.
11. Talend Community, https://community.talend.com/s/?language=en_US, último acceso 24/02/2022.
12. King, W.: Is there evidence of social inequity in healthcare for coronary heart disease? an electronic-cohort analysis using record-linked, routine data (Doctoral dissertation, Cardiff University) (2015).
13. Data System and AI Lab, http://dsail.csail.mit.edu/index.php/data-civilizer/, último acceso 30/10/2021.
14. Kindi, E., Cao, L., Stonebraker, M., Simonini, G., Tao, W., Madden, S., Ouzzani, M., Tang, N., Elmagarmid, A.: Data Civilizer 2.0: A Holistic Framework for Data Preparation and Analytics. Proceedings of the VLDB Endowment. 1954-1957 (2019).

15. Deng, D., Castro Fernandez, R., Abedjan, Z., Wang, S., Stonebraker, M., Elmagarmid, A., Ilyas, I., Madden, S., Ouzzani, M., Tang, N.: The Data Civilizer System.8th Biennial Conference on Innovative Data Systems Research (CIDR '17) (2017).
16. Data-Cleaner github, https://github.com/datosgobar/data-cleaner, último acceso 25/11/2021.
17. Datos Argentina, https://www.datos.gob.ar/, último acceso 04/10/2021.
18. KNIME, https://www.knime.com/, último acceso 17/04/2022.
19. KNIME Documentation, https://docs.knime.com/, último acceso 18/04/2022.