PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

CARRERA DE: SISTEMAS DE INFORMACIÓN

Trabajo de Titulación

Tema: Análisis de herramientas para minería de datos.

AUTOR:

Miguel Angel Flores Amores

QUITO DM, 2024

DEDICATORIA

Dedicado a mi familia y amigos, cuya paciencia, apoyo y aliento inquebrantable hicieron posible este logro. A mis profesores y mentores, por su sabiduría y orientación en cada paso del camino. Gracias por creer en mí y ayudarme a alcanzar mis metas.

También quiero expresar mi gratitud a todos mis compañeros de estudio, quienes compartieron este viaje académico y con quienes aprendí y crecí cada día. Este trabajo es el reflejo del esfuerzo colectivo y del espíritu de colaboración que nos ha unido durante estos años.

AGRADECIMIENTO

Quiero expresar mi más profundo agradecimiento a todos aquellos que me han acompañado en este viaje académico. En primer lugar, a mi querido padre, cuya presencia y apoyo incondicional han sido pilares fundamentales en mi vida y en la realización de esta tesis. Papá, tu sabiduría, tu guía constante y tu ejemplo de perseverancia me han inspirado a alcanzar mis metas. Esta obra es también un testimonio de tu amor y tu dedicación.

A mi madre, cuya partida dejó un vacío profundo en mi vida, pero cuyo legado de amor, sacrificio y fortaleza continúa siendo mi fuente de inspiración diaria. Mamá, aunque no estés físicamente presente, tu espíritu vive en mí y en cada logro que alcanzo. Esta tesis está dedicada a ti, en memoria de tu amor incondicional y tu eterno apoyo.

A mi tío, quien ha sido una figura paternal y un apoyo incondicional a lo largo de este camino académico. Gracias por tus palabras de aliento, por creer en mí y por compartir tu experiencia y conocimientos. Tu presencia ha sido invaluable y tu apoyo ha significado mucho para mí.

A mi querido perro Max, compañero fiel y fuente inagotable de alegría y consuelo. Gracias por estar siempre a mi lado, por tus ladridos de ánimo y tus miradas de complicidad durante las largas horas de estudio. Tu presencia ha sido una luz en los momentos de tensión y una fuente constante de amor incondicional.

A todos mis amigos, colegas y profesores, gracias por su apoyo, orientación y por compartir su sabiduría durante este proceso. A mi familia, por su paciencia, comprensión y amor incondicional en cada paso del camino.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

En esta tesis se realiza un análisis de tres herramientas prominentes para minería de datos: Orange, RapidMiner, y KNIME. Cada una de estas herramientas ofrece distintas características y funcionalidades que las hacen adecuadas para diferentes contextos y necesidades en el ámbito de la ciencia de datos. Se evalúan criterios clave como la flexibilidad, escalabilidad, facilidad de uso, capacidades de preprocesamiento, y soporte para algoritmos avanzados.

Orange destaca por su interfaz intuitiva y amigable, basada en widgets que permiten la creación rápida de flujos de trabajo mediante un sistema de arrastrar y soltar. Esta simplicidad hace que Orange sea especialmente útil para la enseñanza y el aprendizaje de conceptos básicos de minería de datos, así como para la exploración y prototipado rápido. Sin embargo, sus capacidades de preprocesamiento y soporte para algoritmos avanzados son más limitadas en comparación con las otras herramientas evaluadas.

RapidMiner, por su parte, proporciona una robusta plataforma con amplias capacidades de preprocesamiento y un extenso soporte para algoritmos de aprendizaje automático y minería de datos. Su interfaz también se basa en el concepto de flujo de trabajo, aunque con una curva de aprendizaje más pronunciada debido a la abundancia de opciones y configuraciones. Esta herramienta es adecuada para proyectos que requieren un análisis avanzado y detallado, así como para usuarios que necesitan manejar datos complejos y realizar tareas de integración.

KNIME sobresale por su flexibilidad y potencia, permitiendo la creación de flujos de trabajo complejos y repetitivos. Su capacidad para integrar múltiples fuentes de datos y soportar una amplia variedad de extensiones y plugins la hace ideal para entornos empresariales y proyectos de análisis de datos a gran escala. No obstante, su configuración puede ser compleja, lo que representa una barrera para usuarios sin experiencia previa, y puede requerir más recursos computacionales.

Las conclusiones de esta tesis indican que la elección de la herramienta adecuada depende en gran medida del contexto de uso y de las necesidades específicas del proyecto. Orange se recomienda para contextos educativos y proyectos de prototipado rápido, mientras que RapidMiner y KNIME son más adecuados para análisis avanzados y aplicaciones empresariales complejas. La tesis concluye con recomendaciones sobre cómo seleccionar la herramienta más apropiada basándose en las capacidades requeridas y el nivel de experiencia del usuario.

1. CAPÍTULO I: INTRODUCCIÓN

1.1. TEMA

Análisis de herramientas para minería de datos.

1.2. JUSTIFICACIÓN

La minería de datos tiene una gran importancia estratégica para una empresa, ya que permite entender de una manera más analítica, contextualizada y precisa el comportamiento de los clientes y los movimientos del mercado.

La minería de datos es una estrategia basada en la exploración y análisis de datos que permite analizar grandes volúmenes de información para descubrir patrones o reglas de importancia, incluso relaciones de comportamiento orientado en escenarios específicos.

Para elegir la herramienta adecuada para minería de datos se debe hacer un análisis previo basado en varias características de importancia como puede ser los tipos de datos o la compatibilidad de los sistemas, entre otras.

Los errores al usar herramientas de minería de datos son más frecuentes de lo que se esperaría. Estos problemas pueden comenzar desde la etapa de selección de la herramienta adecuada, afectando negativamente el desarrollo del análisis. Seleccionar una herramienta que no se alinea bien con las características de los datos o con los objetivos del análisis puede llevar a resultados subóptimos y dificultades en la interpretación de los datos (Han, Kamber, & Pei, 2011). Por lo tanto, es crucial elegir herramientas que sean compatibles con los datos disponibles y con los propósitos del análisis. Revisar y analizar las herramientas más populares puede proporcionar información valiosa sobre su aplicación efectiva en la minería de datos (Gartner, 2022).

1.3. PLANTEAMIENTO DEL PROBLEMA

Aunque la minería de datos es una poderosa herramienta para comprender el comportamiento y las decisiones de los clientes, la gran variedad de herramientas disponibles hoy en día puede complicar la elección de la más adecuada. Cada herramienta tiene sus propias fortalezas y debilidades, lo que

significa que la elección incorrecta puede llevar a errores en el análisis o a un uso ineficiente del tiempo (Tan, Steinbach, & Kumar, 2018). La falta de experiencia o conocimiento especializado en estas herramientas puede aumentar el riesgo de selección inapropiada, por lo que el asesoramiento experto es crucial para evitar estos problemas. De hecho, es posible que se necesite más de una herramienta para abordar todas las necesidades de un proyecto, dado que algunas herramientas son mejores para tareas específicas, como el análisis predictivo o el procesamiento de grandes volúmenes de datos (Gartner, 2022).

1.4. OBJETIVOS

1.4.1 General:

Analizar herramientas de minería de datos para generar una propuesta de aplicabilidad, enfocada en su idoneidad para entornos educativos y su capacidad para facilitar el aprendizaje de la lógica y técnicas de minería de datos.

1.4.2 Específicos:

1. Analizar algunas de las características y la funcionalidad de herramientas de minería de datos
2. Seleccionar tres herramientas de minería de datos
3. Realizar un análisis comparativo de las herramientas de minería de datos
4. Elaborar las conclusiones y recomendaciones

1.5. ALCANCE

El tema de titulación culminará con la entrega de un documento que analizará algunas de las características y funcionalidades de tres herramientas de minería de datos. Se realizará un análisis comparativo entre las herramientas seleccionadas, describiendo sus ventajas y desventajas. Las herramientas seleccionadas para este análisis son Orange, RapidMiner, y KNIME, elegidas por ser gratuitas y por ofrecer una gama completa de funcionalidades para el análisis de datos (Hofmann & Klinkenberg, 2013) (Van Rijin, Vanschgren, Torgo, & Brazdil, 2018). Estas herramientas son ampliamente reconocidas en la industria por su capacidad de soportar diversas técnicas de minería de datos, lo que las hace adecuadas para una variedad de aplicaciones (Demsar, Zupan, Leban, & Curk, 2013) (Knime, 2023).

Basándose en el análisis comparativo, se propondrá una herramienta recomendada para su uso en contextos educativos. La recomendación se fundamentará en la herramienta que mejor facilite el aprendizaje práctico de minería de datos, que sea intuitiva para los estudiantes, y que ofrezca el mejor equilibrio entre funcionalidad y facilidad de uso.

2. CAPÍTULO II: FUNDAMENTOS CONCEPTUALES

En este capítulo se dará una introducción con respecto a los temas a tratar durante el desarrollo y comparación del presente trabajo de titulación. Adicionalmente se tratarán a detalle las características esenciales de las herramientas a comparar.

2.1. MINERÍA DE DATOS

La minería de datos ha experimentado varios cambios significativos desde sus primeras menciones. El término "minería de datos" se popularizó a finales de la década de 1980 y principios de la década de 1990, con contribuciones clave de diversos investigadores. Uno de los primeros trabajos fundamentales fue presentado por Agrawal, Imielinski, y Swami en 1993, quienes desarrollaron el algoritmo de asociación Apriori, una técnica fundamental para descubrir patrones en grandes conjuntos de (Agrawal, Imielinski, & Swami, 1993). Originalmente, la minería de datos se propuso como una herramienta para ayudar en el tratamiento y análisis de grandes volúmenes de datos, con el objetivo de extraer conclusiones útiles que pudieran mejorar la toma de decisiones empresariales (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Actualmente la minería de datos se puede definir como el procedimiento con el cual inicia el análisis de una gran cantidad de datos, con el objetivo de encontrar patrones y tendencias orientados a mejorar el desempeño de los procesos y actividades de las empresas.

A lo largo del tiempo, los datos han adquirido una importancia creciente, y la manera en que se gestionan ha evolucionado considerablemente. El tratamiento de datos varía según los procesos aplicados y las herramientas utilizadas. En los últimos años, el aumento de la relevancia del análisis de datos se ha debido a varios factores, incluyendo la reducción de costos, la disminución en los tiempos de procesamiento, y el incremento en la velocidad de procesamiento de datos (Gantz & Reinsel, 2012) (Davenport & Patil, 2012). Estos avances han llevado al desarrollo de una amplia gama de herramientas, tanto comerciales como gratuitas, cada una diseñada para abordar diferentes aspectos del análisis de datos (Chaudhuri, Dayal, & Narsayya, 2011).

Se han desarrollado diversas técnicas avanzadas que mejoran el rendimiento y la capacidad de extraer información no accesible mediante los métodos tradicionales de minería de datos. La inteligencia artificial, junto con el análisis estadístico y el uso de representaciones gráficas, ha acelerado significativamente esta evolución, facilitando una comprensión más profunda del análisis de datos (Han, Kamber, & Pei, 2011) (Hastie, Tibshirani, & Friedman, 2009).

Es crucial distinguir entre la minería de datos y la estadística. Aunque ambas disciplinas pueden tener el objetivo común de construir modelos comprensibles, sus enfoques y metodologías son diferentes. La minería de datos se centra en descubrir patrones ocultos en grandes volúmenes de datos, mientras que la estadística se enfoca en la inferencia y la prueba de hipótesis (Han, Kamber, & Pei, 2011) (Aggarwal, 2015).

Como punto principal de diferenciación entre la minería de datos y la estadística sería el uso de la inteligencia artificial por parte de la minería de datos, sin mencionar que por lo habitual orienta sus métodos confirmatorios es decir que prueban teorías o hipótesis específicas que ya han sido establecidas. Por el contrario, la minería de datos usa métodos exploratorios para buscar patrones y tendencias sin la necesidad de probar una hipótesis previamente establecida.

Para facilitar la comprensión de que es minería de datos en comparación de la estadística se podría usar un ejemplo. Si una empresa desea saber qué factores afectan la venta de su mercancía tienen dos opciones que serían métodos estadísticos confirmatorios y por el contrario lo que sería el método de minería de datos por exploración.

En un análisis de datos, el enfoque confirmatorio puede involucrar la formulación de una hipótesis específica sobre cómo un producto afecta las ganancias, seguida de la aplicación de una regresión lineal sobre el historial de compra y venta de ese producto para validar la hipótesis (Montgomery, Peck, & Vining, 2012). En contraste, un enfoque exploratorio mediante la minería de datos podría utilizar algoritmos de clustering, como el “K-Means”, para identificar patrones agrupando productos de diferentes categorías, pero con precios similares, lo que permite descubrir relaciones que puedan informar estrategias para aumentar las ventas (Tan, Steinbach, & Kumar, 2018).

Esta diferencia refleja la distinción sutil pero importante entre la estadística tradicional y la minería de datos; mientras que la estadística se centra en pruebas confirmatorias y modelado basado en hipótesis, la minería de datos explora grandes volúmenes de datos para identificar patrones y relaciones (Hand, Mannila, & Smyth, 2001). La creciente relevancia de la minería de datos está impulsada por el manejo de grandes volúmenes de datos, la velocidad de procesamiento, y la variedad de fuentes de datos, conocidas como las "V" del big data (Chen, Mao, & Liu, 2014).

Una vez definido lo que es minería de datos y esclarecida las dudas con su respectivo contendiente en ámbitos de objetivos, podemos avanzar a conocer el proceso adecuado, sugerido y destacado para realizar una minería de datos.

2.2. PROCESO DE MINERÍA DE DATOS

Antes de iniciar el proceso analítico, es fundamental establecer un plan claro y detallado. Esto incluye definir la problemática y los objetivos que se deben cumplir para resolverla. Una planificación meticulosa asegura que cada etapa del proceso esté alineada con el objetivo final y evita malentendidos o desvíos durante la ejecución (Beltrán González, 2011). A continuación, se describen las etapas del proceso de minería de datos, basado en las directrices del CRISP-DM, el cual es ampliamente reconocido en este campo (Beltrán González, 2011).

2.2.1 Selección:

Primeramente, para iniciar con el proceso de la minería de datos se debe identificar los datos con los que trabajaremos, una vez identificados deben ser tratados y almacenados de manera correcta en un Data Warehouse para que sean aptos para el siguiente paso. Muchas veces al ser tantos datos hay algunos que su uso no contribuye por lo tanto hay que depurar estos datos antes de realizar las siguientes fases.

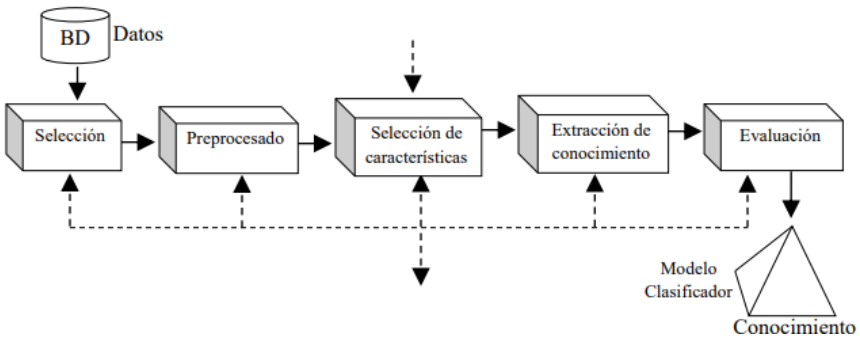

Figura 1: Proceso de Minería de Datos tomado de Beltrán (2024)

2.2.2 Preprocesado:

Inmediatamente después de elegir los datos debemos adecuar su formato ya que la gran mayoría del tiempo no son aptos para el proceso necesario que va a seguir el algoritmo, por lo tanto, en el

preprocesado descartamos los datos que no son compatibles o bien no tienen validez. También podemos agruparlos mediante clústeres cuando hay demasiados datos similares.

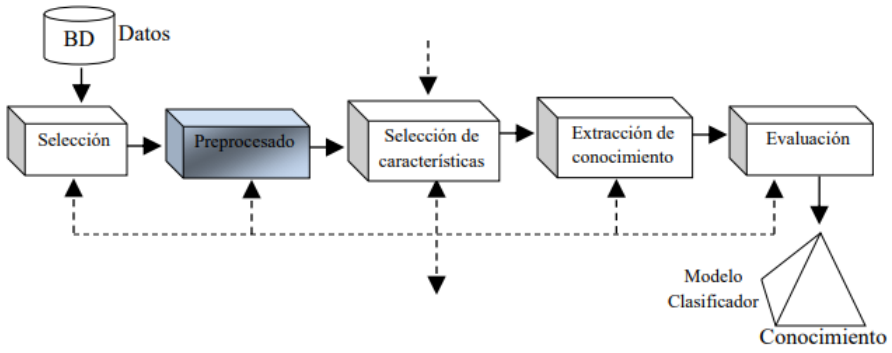

Figura 2: Preprocesado tomado de Beltrán (2024)

2.2.3 Selección de características:

Una vez preprocesados los datos destacan algunos más que otros, por lo tanto, elegiremos según nuestra conveniencia basados en cuales nos podrían ayudar a resolver el problema propuesto en un inicio.

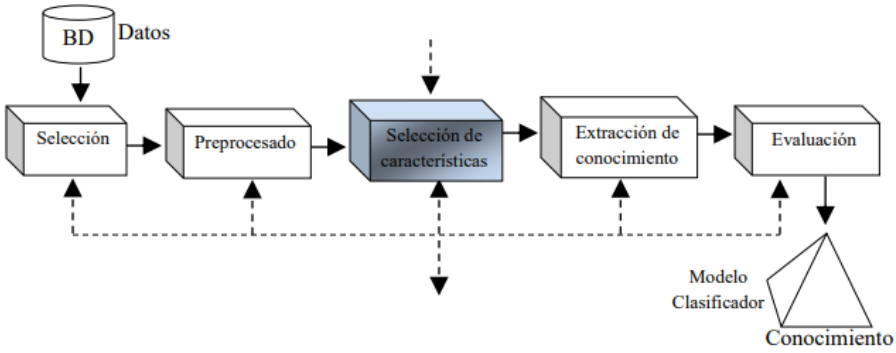

Figura 3: Selección de características tomado de Beltrán (2024)

2.2.4 Extracción de conocimiento:

El siguiente paso consiste en implementar técnicas de minería de datos para obtener modelos de conocimiento. En caso de necesitar más modelos, se deben emplear diferentes técnicas; sin embargo, es crucial ser cauteloso, ya que algunas técnicas requieren procesos de preprocesamiento distintos a

otros (Han, Kamber, & Pei, 2011). Una vez obtenidos los modelos, es posible identificar patrones o tendencias significativas, así como relaciones entre múltiples variables (Witten , Frank, Hall, & Pal, 2016).

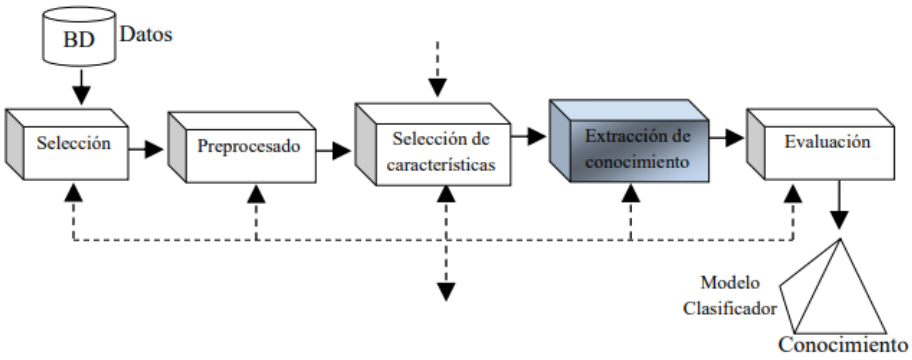

Figura 4: Extracción de conocimiento tomado de Beltrán (2024)

2.2.5 Evaluación:

Tras obtener los modelos deseados procederemos a evaluarlos en base a los criterios establecidos y confirmando que sus datos arrojen conclusiones satisfactorias. La ventaja de tener varios modelos en comparación de tener uno solo es clara ya que al tener varios resultados podemos confirmar si hay algún reincidente, en caso de que no sea ningún resultado satisfactorio debemos hacer una regresión de los pasos realizados y hacer cambios para obtener nuevos modelos.

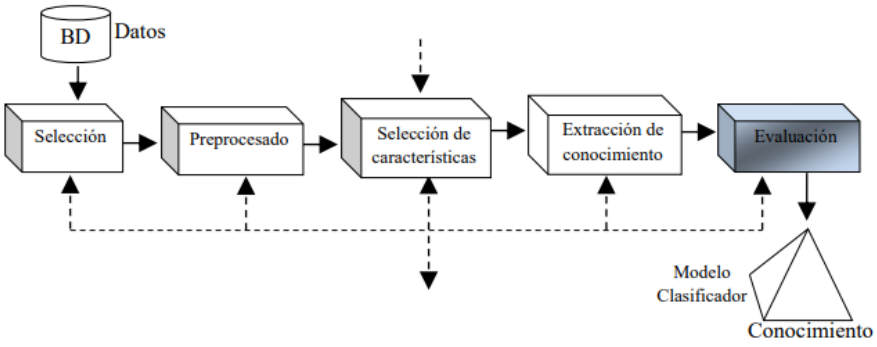

Figura 5: Evaluación tomado de Beltrán (2024)

2.3. APRENDIZAJE SUPERVISADO Y NO SUPERVISADO

En el gran mundo que es la minería de datos existen dos metodologías de aprendizaje que son las más representativas y cada una con sus respectivas técnicas. El aprendizaje supervisado y el no supervisado, una manera simple de describirlas seria que el método supervisado se espera que el modelo reproduzca un resultado conocido y por el contrario en el no supervisado, al no tener una salida conocida, se espera detectar patrones de similitud y tendencias.

Vale recalcar que muchas veces se confunde el aprendizaje supervisado con los métodos confirmatorios mencionados en el apartado 2.1 del segundo capítulo del documento. Según se expone en el estudio, el aprendizaje supervisado, representado por métodos como la regresión lineal (RL), comparte similitudes con los métodos confirmatorios en términos de uso de datos etiquetados e hipótesis previamente establecidas (Garcia, 2015). Discutir estos conceptos en la parte correspondiente del documento es esencial para fundamentar la metodología empleada y justificar la elección de técnicas en contextos específicos, asegurando la validez y aplicabilidad de los resultados obtenidos.

Las técnicas supervisadas se utilizan en aprendizaje automático para construir modelos predictivos a partir de datos etiquetados, como regresión lineal o árboles de decisión, mientras que los métodos confirmatorios en estadística se centran en probar hipótesis específicas utilizando datos recopilados con un diseño de estudio planificado, como pruebas de hipótesis o análisis de varianza, para validar teorías existentes.

2.3.1 Aprendizaje Supervisado:

El aprendizaje supervisado ha sido de las principales herramientas para desarrollar a profundidad el aprendizaje automático y la inteligencia artificial, destaca por su gran facilidad y capacidad para crear modelos de predicción los cuales son muy precisos a partir de datos etiquetados los cuales serían las hipótesis o salidas deseadas. Gracias a estas salidas deseadas el modelo obtiene la facilidad de relacionar los datos ingresados con la respuesta que esperaría, a esto se lo conocería como generalización y nos permite que el modelo pueda a futuro tener mayor precisión con entradas que no han sido ingresadas.

Las técnicas que emplearíamos en este modelo de aprendizaje se pueden clasificar en dos categorías principales: clasificación y predicción. Esta distinción nos ayudará a abordar adecuadamente las necesidades del problema.

2.3.1.1 Técnicas de aprendizaje supervisado por clasificación

Entre las técnicas más comunes de aprendizaje supervisado por clasificación se encuentran la regresión logística, los árboles de decisión, y el algoritmo de K-Nearest Neighbor (KNN), entre otras mencionadas, ampliamente reconocidas en la literatura (Hastie, Tibshirani, & Friedman, 2009) (Han, Kamber, & Pei, 2011):

1. Regresión logística: este método generalmente usado en problemas estadísticos ayuda a lograr identificar sin una instancia pertenece a una de dos clases, debido a su lógica simple y resultados efectivos es uno de los más usados (Hosmer & Lemeshow, 2000).
2. Arboles de decisión: Los árboles de decisión son modelos de aprendizaje automático que se utilizan tanto para clasificación como para regresión. Utilizan una estructura en forma de árbol donde cada nodo representa una característica del dato y cada rama representa una posible decisión basada en esa característica. Esta división jerárquica permite tomar decisiones de manera eficiente al seguir un camino directo en el árbol según las características de los datos. Esta flexibilidad los hace útiles para identificar patrones complejos y resolver problemas de manera efectiva (Breiman, Friedman, Olshen, & Stone, 1984).
3. K-Nearest Neighbor (KNN): algoritmo sencillo que clasifica en pares eligiendo al mayor del vecino más cercano de una instancia (Cover & Hart, 1967).

2.3.1.2 Técnicas de aprendizaje supervisado por predicción

Según se describe en la literatura especializada, algunas de las técnicas más comunes de aprendizaje por predicción incluyen las siguientes (Montgomery, Peck, & Vining, 2012) (Russell & Norvig, 2021) (Breiman, Random forests. , 2001):

1. Regresión Lineal: La regresión lineal es ampliamente utilizada en predicción y modelado. Este método nos permite visualizar y cuantificar las relaciones entre variables, asumiendo que estas relaciones son constantes y lineales. Es útil para explorar cómo una variable depende de otras variables independientes, incluso cuando no hay dependencia entre las variables predictoras (Montgomery, Peck, & Vining, 2012).

2. Redes Neuronales Artificiales (ANN): como su nombre lo sugiere son inspirados en las funciones neurológicas del cerebro, destacan por su amplia capacidad de aprendizaje y adaptación para buscar relaciones con variables de entrada y una predicción esperada. (Russell & Norvig, 2021).
3. Random Forest: Random Forest es un algoritmo de aprendizaje automático que combina varios árboles de decisión independientes. Cada árbol se entrena con una muestra aleatoria del conjunto de datos y utiliza diferentes conjuntos de características para las divisiones en cada nodo. A diferencia de un solo árbol de decisión, que puede ser propenso al sobreajuste, Random Forest reduce este riesgo mediante la combinación de múltiples modelos. Cada árbol en el bosque contribuye con sus predicciones individuales, que se combinan mediante votación o promedio para mejorar la precisión predictiva del modelo global. Esta técnica se destaca por su capacidad para manejar conjuntos de datos grandes y complejos, manteniendo una buena generalización y robustez en las predicciones (Breiman, Random forests. , 2001).

2.4. HERRAMIENTAS

2.4.1. Orange:

- Una de las herramientas con licencia GPL o de uso gratuito que se analizara es “ORANGE”, diseñada para hacer minería de datos en conjunto de análisis de datos.
- Desarrollada en el lenguaje de programación C++, Python, Cython y C en la “Universidad de Liubliana” ubicada en Eslovenia.
- Su última versión hasta le fecha (junio 2024) es la 3.36.2
- Está disponible para Windows, Mac y para su uso en repositorios virtuales mediante líneas de código.
- Para adquirir la aplicación, los usuarios deben dirigirse a la página web de Orange, la cual cuenta con varios apartados y una comunidad para resolver dudas. La descarga es fácil y no requiere registro previo. La aplicación puede descargarse desde el siguiente enlace: https://orange.biolab.si/download/.

Figura 6: Orange tomado de orangedatamining.com (2024)

Una vez dentro de la aplicación podemos encontrar las distintas herramientas o algoritmos, en el caso de Orange están catalogados de manera vertical organizado por grupos contando con una barra de búsqueda y una sección que al posar el ratón sobre cada ítem nos sale una descripción. Se podría observar de la siguiente manera:

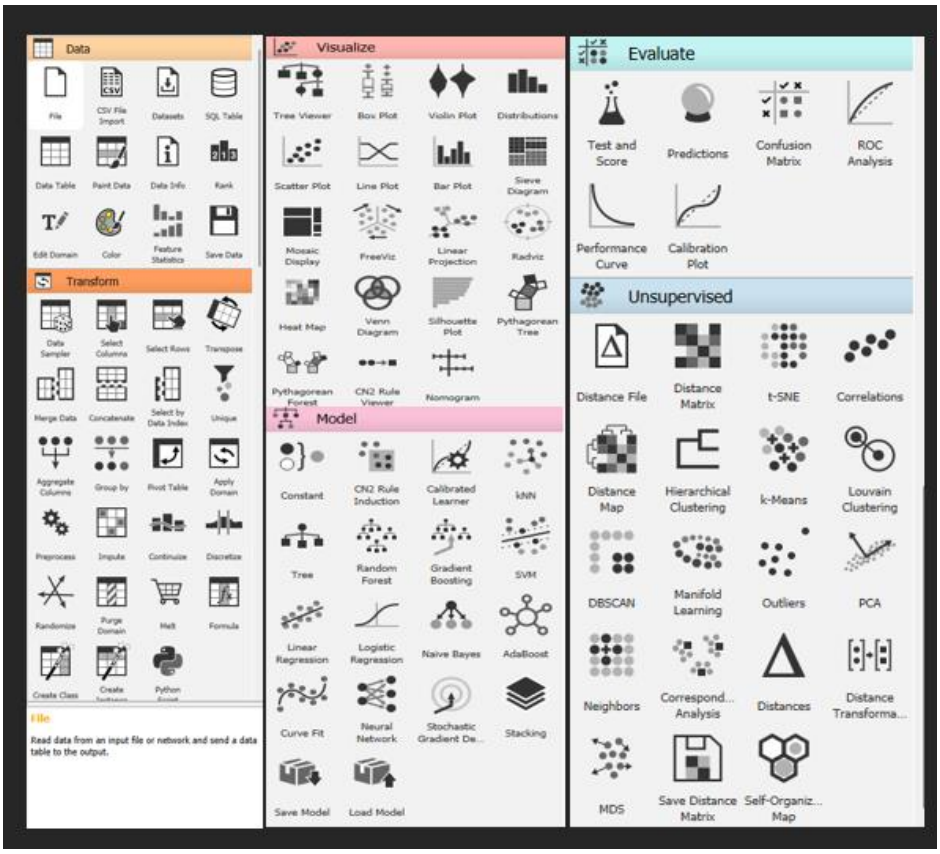

Figura 7: Herramientas disponibles en Orange

La herramienta consume los siguientes recursos: CPU, memoria, disco y red respectivamente.

Figura 8: Recursos consumidos Orange

2.4.2. Knime:

- Siglas de Konstanz Information Miner
- Plataforma para minería de datos desarrollado en eclipse y posee entorno visual.
- Creado en la "Universidad de Constanza" en Alemania.
- Herramienta de uso libre
- Su última versión hasta la fecha (junio 2024) es la 5.2.5
- Para adquirir esta aplicación deberemos dirigirnos a su página web: https://www.knime.com/downloads/download-knime la cual consta con varios apartados y una comunidad para resolver dudas, su descarga es fácil y no hay necesidad de registrarse.

Figura 9: Knime tomado de knime.com (2024)

Una vez dentro de la aplicación podemos encontrar las distintas herramientas o algoritmos, en el caso de Knime están catalogados de la siguiente manera:

Figura 10: Herramientas encontradas en Knime

Cuenta con un apartado vertical junto a una barra de búsqueda y un filtro, al ser tan extensas cada sección posee la opción de mostrar más. Se podría encontrar el resto de las herramientas disponibles con su respectiva descripción en su página oficial a través del siguiente enlace: https://hub.knime.com/search?type=Node&pk_vid=831d0ab9d9fbb9411719729474c90e21

La herramienta consume los siguientes recursos: CPU, memoria, disco y red respectivamente.

Figura 11: Recursos consumidos knime

2.4.3. Altair Rapid Miner:

- Programa de uso pago con acceso a prueba para minería de datos
- Creado en la "Universidad de Dortmund" en Alemania.
- Programado en Java.
- Necesidad de registrarse para poder descargarlo
- Su última versión hasta la fecha (junio 2024) es la 10.4.1.0
- Para descargarla debemos ir al siguiente enlace: https://rapidminer.com/get-started/ .

Figura 12: RapidMiner de rapidminer.com (2024)

Una vez dentro de la aplicación podemos encontrar las distintas herramientas o algoritmos, en el caso de Rapid Miner están catalogados de la siguiente manera:

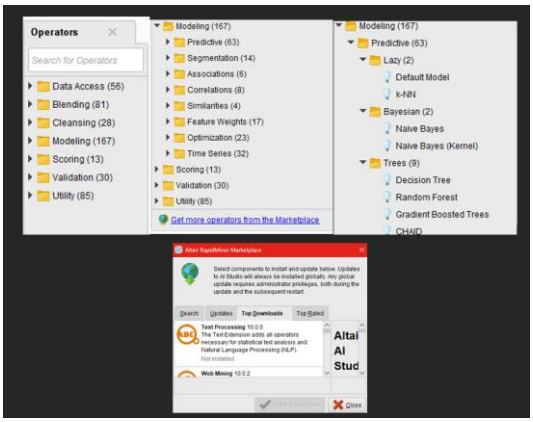

Figura 13: Herramientas disponibles en Rapid Miner

Se encuentran de manera vertical catalogadas en carpetas de acuerdo con el propósito al momento de trabajar, cuenta con una barra de búsqueda y es posible adquirir mas desde la misma aplicación. Se podría encontrar las herramientas disponibles con su respectiva descripción en su página oficial a través del siguiente enlace: https://docs.rapidminer.com/latest/studio/operators/

La herramienta consume los siguientes recursos: CPU, memoria, disco y red respectivamente.

Figura 14: Recursos consumidos rapidminer

3. CAPÍTULO III: ANÁLISIS COMPARATIVO

En este capítulo se describirá los procesos y puntos a evaluar para comparar las herramientas propuestas anteriormente, de manera que podamos concluir pros y contras de cada una. Especialmente para ver cual nos ayudaría a entender la lógica de la minería de datos de una manera más amigable para las personas que están en proceso de aprendizaje.

3.1. METODO DE ANALISIS

Para ver de manera más optima cada herramienta, emplearemos el método comparativo para destacar varios puntos de una manera equitativa. Incluyendo pruebas de rendimiento, adaptabilidad del usuario, facilidad de uso, etc.

3.1.1. Método Comparativo:

Para comparar estas herramientas debemos iniciar desde un punto equitativo y destacando los puntos que se van a evaluar, las pruebas iniciaran con cada aplicación desde su ambiente al iniciarlos.

Se concluirá con una tabla comparativa en la cual se destacará cada una de las herramientas siendo comparada con base en los siguientes parámetros:

- Características principales
- Caso de uso
- Facilidad de uso
- Escalabilidad
- Integración
- Soporte
- Costo
- Opinión

3.1.1.1. Enfoque General:

Al comparar herramientas podemos ver ventajas y desventajas de cada una, facilitándonos el uso correcto dependiendo nuestras necesidades.

Se eligieron Orange, KNIME y RapidMiner, tres herramientas destacadas en la minería de datos según un análisis reciente. Estas plataformas han sido reconocidas por su versatilidad, capacidad de

análisis avanzado y facilidad de uso, como se menciona en un artículo sobre software de data mining (IONOS, 2023).

Según el artículo, Orange, KNIME y RapidMiner son ampliamente utilizadas en diversas industrias debido a su capacidad para gestionar grandes volúmenes de datos. Orange se destaca por su facilidad de uso y visualización intuitiva; KNIME es reconocido por su flexibilidad y potencia en la integración de diversos tipos de datos; y RapidMiner es valorado por su robustez y capacidad para realizar análisis avanzados de datos. Estas herramientas facilitan la extracción de conocimientos significativos para la toma de decisiones estratégicas.

3.1.1.2. Proceso:

1. Se iniciará preparando el ambiente de cada una de las herramientas con sus versiones más recientes.
2. Por razones de tiempo se elegirán los datos que no necesiten tanta depuración, en caso de ser necesario pasarán por la etapa de preprocesamiento.
3. Se empleará un algoritmo adecuado para los datos y se empleará el mismo en todas las herramientas, también el mismo dataset.

3.1.1.3. Evaluación:

Dado que las conclusiones serán elegidas para un ámbito estudiantil, cada punto a evaluar será calificado con base en su uso de manera simple y con herramientas de acceso público, enfocándose en dos puntos generales, los cuales son: su facilidad de uso y la eficiencia de recursos.

3.1.2. Criterios de Evaluación:

3.1.2.1. Facilidad de Uso:

Se evaluará la interfaz ofrecida al usuario en conjunto con su curva de aprendizaje, es decir que tan rápido el usuario se acostumbra a su uso en un inicio. También en caso de no terminar de aprender la herramienta se buscará la disponibilidad de recursos adicionales como manuales o tutoriales.

3.1.2.2. Eficiencia de recursos:

Adicionalmente se evaluará el uso de memoria y del CPU durante el proceso.

3.2. BASE DE DATOS

Se procederá a usar las tres herramientas propuestas en casos de estudio supervisado y no supervisado. El dataset a usar es uno disponible al público de manera gratuita en un repositorio virtual y se puede acceder al archivo directamente desde el siguiente enlace: https://raw.githubusercontent.com/selva86/datasets/master/BreastCancer.csv.

La base elegida para los ejemplos tanto supervisados como no supervisados es una base de datos que viene por defecto, conocida como 'Breast Cancer Wisconsin (Diagnostic)'. Aunque su uso es generalmente aplicado para métodos supervisados debido a sus etiquetas, algunos de sus datos permiten agrupar en clústeres sin usar las etiquetas. Esto hace que sea apta tanto para técnicas supervisadas como no supervisadas, proporcionando flexibilidad en su aplicación.

3.2.1. Descripción de los datos:

Según el UCI Machine Learning Repository, el dataset "Breast Cancer Wisconsin (Diagnostic)" es ampliamente utilizado para la clasificación de tumores en benignos o malignos a partir de características extraídas de imágenes digitales de núcleos celulares de biopsias de seno. Es una herramienta fundamental en la investigación médica y en la enseñanza de técnicas avanzadas de análisis de datos, proporcionando un caso de estudio crucial para la aplicación de algoritmos de aprendizaje automático en la detección y diagnóstico preciso del cáncer de mama.

Se puede encontrar más detalles en el siguiente enlace: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

- Estructura del Dataset

    - Total, de registros: 569
    - Número de variables: 31 (30 características y 1 etiqueta)
    - Variables del Dataset

- ID del Paciente (id)

    - Descripción: Identificador único para cada paciente
    - Tipo de dato: Numérico (entero)

- Diagnóstico (diagnosis)

    - Descripción: Resultado del diagnóstico (M = maligno, B = benigno)
    - Tipo de dato: Categórica

- Características de los núcleos celulares

Estas características se calculan a partir de imágenes digitalizadas de células FNA:

- radio (mean_radius, se, worst_radius)
- textura (mean_texture, se_texture, worst_texture)
- perímetro (mean_perimeter, se_perimeter, worst_perimeter)
- área (mean_area, se_area, worst_area)
- suavidad (mean_smoothness, se_smoothness, worst_smoothness)
- compacidad (mean_compactness, se_compactness, worst_compactness)
- concavidad (mean_concavity, se_concavity, worst_concavity)
- puntos cóncavos (mean_concave_points, se_concave_points, worst_concave_points)
- simetría (mean_symmetry, se_symmetry, worst_symmetry)

- dimensión fractal (mean_fractal_dimension, se_fractal_dimension, worst_fractal_dimension)

- Descripción de las características:

  - Tipo de Dato: Continuo
  - Rango y descripción específica de cada una de estas características varía según si son promedios ('mean'), errores estándar ('se'), o peor caso ('worst').

3.3. USO DE HERRAMIENTAS CON METODO SUPERVISADO

Se usará uno de los métodos más comunes para mayor facilidad y entendimiento de la aplicación, un árbol de decisión permite evaluar de manera funcional las herramientas sin mencionar que es uno de los algoritmos más completos.

Para evaluar la efectividad de un modelo de clasificación, se utilizan varias métricas clave: precisión, que mide la proporción de predicciones correctas en la clase positiva; exactitud, que evalúa la proporción total de predicciones correctas; sensibilidad, que indica la capacidad del modelo para detectar casos positivos; especificidad, que mide la identificación correcta de los casos negativos; y valor F, que balancea precisión y sensibilidad (García, 2018). Estas métricas son cruciales para interpretar los resultados de un modelo de clasificación (García, 2018):

Donde se puede interpretar como:

- TP (True Positives): Son los verdaderos positivos. Casos en los que el modelo predice correctamente la clase positiva.
- TN (True Negatives): Son los verdaderos negativos. Casos en los que el modelo predice correctamente la clase negativa.
- FP (False Positives): Son los falsos positivos. Casos en los que el modelo predice la clase positiva incorrectamente (falsas alarmas).
- FN (False Negatives): Son los falsos negativos. Casos en los que el modelo predice la clase negativa incorrectamente (fallos en la detección).
- F1 (Medida F1): Es la media armónica de la precisión y la sensibilidad. Proporciona un balance entre ambas métricas y es especialmente útil en situaciones de clases desbalanceadas.

1. Precisión: Proporción de predicciones correctas en la clase positiva. Indica la calidad de las predicciones positivas.

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Figura 15: Fórmula de precisión

2. Exactitud: Proporción total de predicciones correctas, considerando tanto positivas como negativas. Mide la eficiencia general del modelo la cual nos ayuda a medir la capacidad de un modelo para realizar predicciones correctas de manera consistente y precisa.

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figura 16: Fórmula de exactitud

3. Sensibilidad: Proporción de verdaderos positivos identificados correctamente. Evalúa la capacidad del modelo para detectar casos positivos.

$$\text{Sensibilidad} = \frac{TP}{TP + FN}$$

Figura 17: Fórmula de sensibilidad

4. Especificidad: Proporción de verdaderos negativos identificados correctamente. Mide la capacidad del modelo para evitar falsos positivos.

$$\text{Especificidad} = \frac{TN}{TN + FP}$$

Figura 18: Fórmula de especificidad

5. Valor F: Media armónica de precisión y sensibilidad. Balancea ambas métricas, útil en situaciones de clases desbalanceadas.

$$F1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

Figura 19: Fórmula de valor f

3.3.1. Árbol de decisión en Orange:

Se construye el modelo:

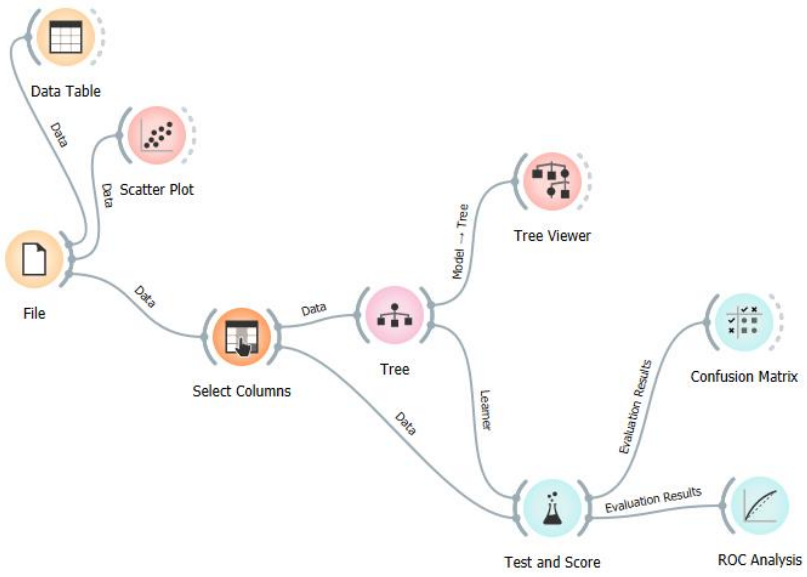

Figura 20: Árbol en Orange

Al conectar los datos desde el nodo “File” hacia el nodo select columns seleccionamos la columna a que queremos usar, en este caso elegimos diagnosis. Luego conectamos a tree y se entrenará automáticamente una vez se haya configurado.

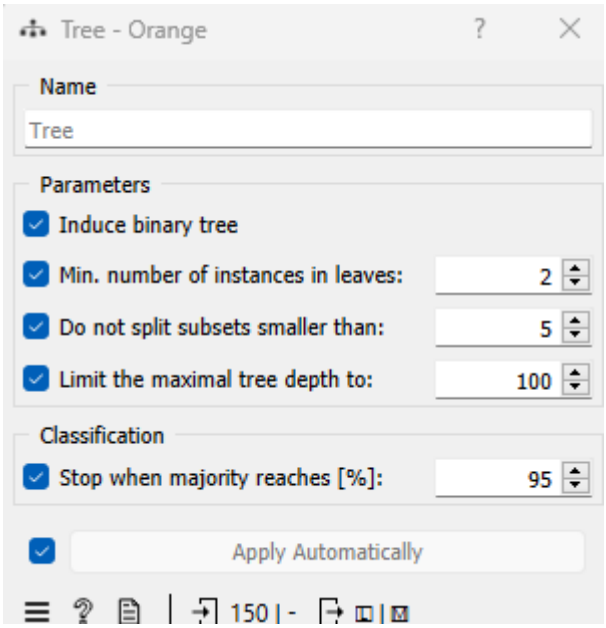

Figura 21: Configuración de Árbol en Orange

Usaremos el nodo “Test and Score” para evaluar el modelo el cual se realiza con una validación cruzada que viene en 5 particiones de manera predeterminada para la clase que vamos a predecir ya balanceada en cada una de las particiones.

Figura 22: Resultados del Árbol en Orange

Gracias a los nodos de matriz de confusión y Roc podremos evaluar la calidad de los resultados. La matriz de confusión la podemos ver en porcentajes y en valores absolutos dependiendo si queremos ver la proporción de lo predicho o el número de instancias respectivamente.

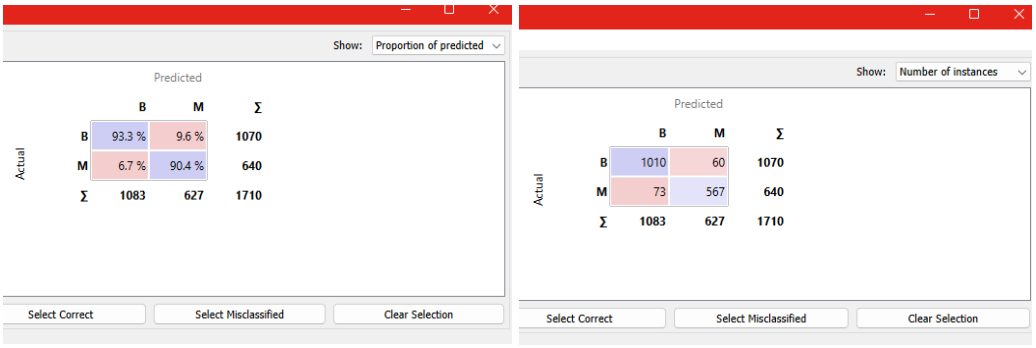

Figura 23: Matriz de confusión árbol en Orange

Cálculos con los resultados obtenidos en la Fig.23:

1. Precisión (Precision)

   Interpretación: La precisión del 93.61% indica que el 93.61% de las instancias clasificadas como positivas son realmente positivas. Esto sugiere que el modelo tiene una baja tasa de falsos positivos.

$$Precisión - \frac{TP}{TP + FP} - \frac{1070}{1070 + 73} \approx 93.61\%$$

2. Exactitud (Accuracy)

   Interpretación: La exactitud del 93.76% refleja que el modelo clasifica correctamente el 93.76% del total de instancias, lo cual es un buen indicador de rendimiento general.

$$Exactitud - \frac{TP + TN}{TP + TN + FP + FN} - \frac{1070 + 627}{1070 + 627 + 73 + 33} \approx 93.76\%$$

3. Sensibilidad (Recall)

   Interpretación: La sensibilidad del 97.00% muestra que el 97.00% de las instancias positivas reales fueron correctamente identificadas como positivas. Esto es importante en contextos donde es crucial minimizar los falsos negativos.

$$Sensibilidad - \frac{TP}{TP + FN} - \frac{1070}{1070 + 33} \approx 97.00\%$$

$$Sensibilidad - \frac{TP}{TP + FN} - \frac{1070}{1070 + 33} \approx 97.00\%$$

4. Especificidad (Specificity)

   Interpretación: La especificidad del 89.57% indica que el 89.57% de las instancias negativas reales fueron correctamente identificadas como negativas. Esto es importante para minimizar los falsos positivos.

$$Especificidad - \frac{TN}{TN + FP} - \frac{627}{627 + 73} \approx 89.57\%$$

5. Valor F (F1 Score)

   Interpretación: El valor F del 95.72% proporciona un equilibrio entre precisión y sensibilidad, lo cual es útil cuando se necesita una medida combinada de la calidad del modelo.

$$Valor\ F - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9361 * 0.9700)}{0.9361 + 0.9700} \approx 95.72\%$$

$$Valor\ F - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9361 * 0.9700)}{0.9361 + 0.9700} \approx 95.72\%$$

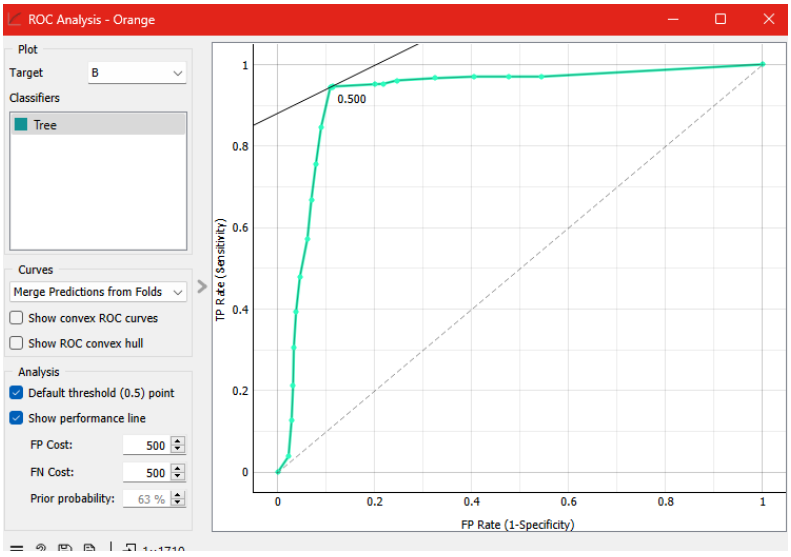

Figura 24: Análisis ROC de árbol en Orange

Tras obtener una alta precisión en la matriz de confusión, al comparar la proporción de predicciones correctas con el número total de instancias, y con el análisis ROC alcanzando una curva satisfactoria, se puede concluir que el modelo ha mostrado un buen rendimiento. Esto se refleja en el 90% de precisión obtenido en los cálculos.

3.3.2. Árbol de decisión en Rapid Miner:

Diseñamos el árbol en Rapid Miner.

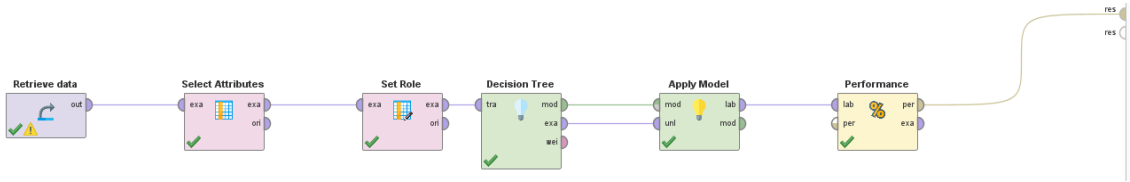

Figura 25: Arbol de decisión en rapid miner

En Rapid Miner hay que cargar los datos directamente y al hacerlo se crea una base en la herramienta con la cual podremos configurar las variables y luego cargar la data al diseño.

Figura 26: carga de datos en rapid miner

Así la etiqueta principal seria diagnosis entonces la destacaríamos como label.

Figura 27: etiquetar dato en rapid miner

Así vendría la configuración del árbol de manera predeterminada.

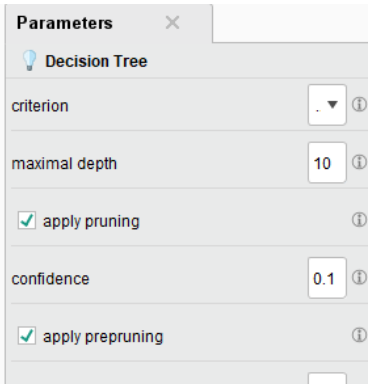

Figura 28: Configuración de árbol en rapid miner

Al ejecutar la herramienta nos mostraría la siguiente precisión siendo aceptable.

accuracy: 95.96%

	true M	true B	class precision
pred. M	198	9	95.65%
pred. B	14	348	96.13%
class recall	93.40%	97.48%	

Tabla 1: Precisión del árbol en rapid miner

Cálculos con los resultados de la Tabla 1:

1. Precisión (Precision)

   Interpretación: La precisión indica que el 95.65% de las instancias clasificadas como positivas (M) realmente son positivas. Esto refleja que el modelo es bastante confiable al predecir la clase positiva.

$$Precisión - \frac{TP}{TP + FP} - \frac{198}{198 + 9} \approx 95.65\%$$

2. Exactitud (Accuracy)

   Interpretación: La exactitud general del modelo es del 95.96%, lo que indica que el modelo clasifica correctamente la gran mayoría de las instancias. Sin embargo, no distingue entre la capacidad para predecir correctamente la clase positiva y negativa.

$$Exactitud - \frac{TP + TN}{TP + TN + FP + FN} - \frac{198 + 348}{198 + 348 + 9 + 14} \approx 95.96\%$$

3. Sensibilidad (Recall)

   Sensibilidad (M)

   Interpretación: La sensibilidad para la clase M es del 93.40%, lo que significa que el modelo identifica correctamente el 93.40% de las instancias verdaderamente positivas.

$$Sensibilidad(M) - \frac{TP}{TP + FN} - \frac{198}{198 + 14} \approx 93.40\%$$

   Sensibilidad (B)

   Interpretación: Para la clase B, la sensibilidad es del 97.48%, indicando una alta capacidad del modelo para detectar instancias verdaderamente negativas.

$$Sensibilidad(B) - \frac{TP}{TP + FN} - \frac{348}{348 + 9} \approx 97.48\%$$

4. Especificidad (Specificity)

   Interpretación: La especificidad muestra que el modelo identifica correctamente el 97.48% de las instancias verdaderamente negativas, lo cual es crucial en aplicaciones donde es importante minimizar los falsos positivos.

$$Especificidad - \frac{TN}{TN + FP} - \frac{348}{348 + 9} \approx 97.48\%$$

5. Valor F (F1 Score)

   Interpretación: El valor F para la clase M es del 94.51%, lo que combina precisión y sensibilidad en una sola métrica para medir la calidad del modelo al predecir la clase positiva. Para la clase B, el valor F es del 96.80%, indicando un excelente equilibrio entre precisión y sensibilidad.

$$Valor\ F(M) - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9565 * 0.9340)}{0.9565 + 0.9340} \approx 94.51\%$$

$$Valor\ F(B) - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9613 * 0.9748)}{0.9613 + 0.9748} \approx 96.80\%$$

3.3.3. Árbol de decisión en Knime:

Creamos el modelo en Knime.

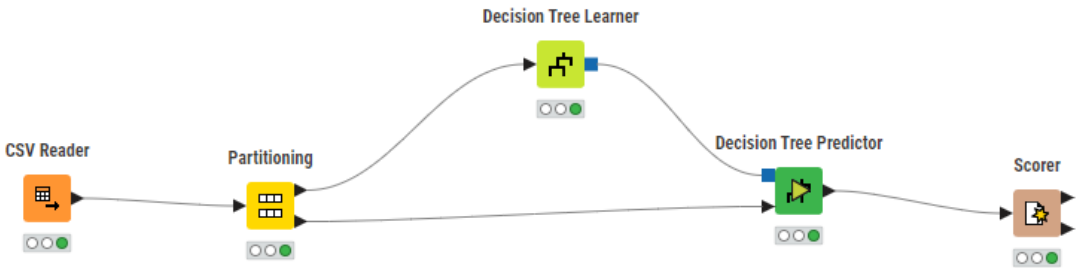

Figura 29: Diseño de árbol en K-nime

Con el nodo Partitioning configuramos el entrenamiento en 70/30 para entrenamiento y testeo respectivamente.

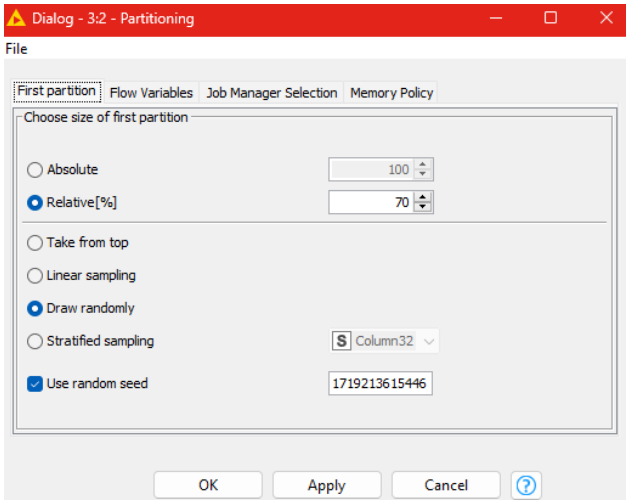

Figura 30: Distribución de datos en K-nime

Con el nodo Decision tree Predictor vemos como se armó el árbol.

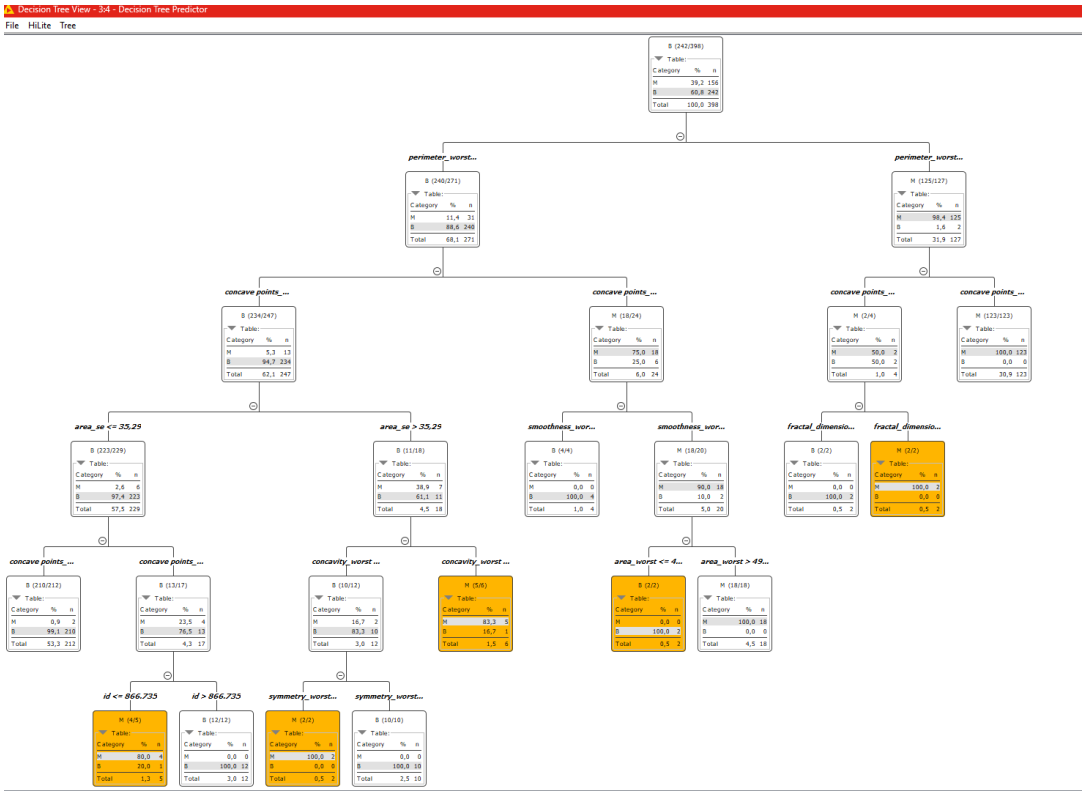

Figura 31: Arbol en K-Nime

En base a los resultados obtenidos del nodo score en el cual nos muestra la matriz de confusion pdoemos ver que es un modelo satisfactorio debido a su alta precision.

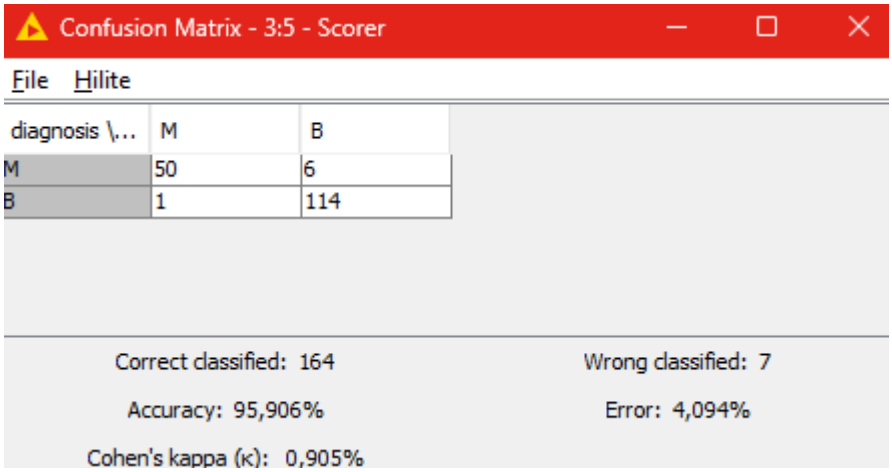

Figura 32: Matriz de confusión en K-Nime

Cálculos con los resultados obtenidos en la Fig.35:

1. Precisión (Precision)

   Interpretación: La precisión indica que el 98.04% de las instancias clasificadas como positivas (M) realmente son positivas. Esto muestra que el modelo es altamente fiable al predecir la clase positiva.

$$Precisión - \frac{TP}{TP + FP} - \frac{50}{50 + 1} \approx 98.04\%$$

2. Exactitud (Accuracy)

   Interpretación: La exactitud general del modelo es del 95.91%, lo que indica que clasifica correctamente la mayoría de las instancias.

$$Exactitud - \frac{TP + TN}{TP + TN + FP + FN} - \frac{50 + 114}{50 + 114 + 1 + 6} \approx 95.91\%$$

3. Sensibilidad (Recall)

   Sensibilidad (M)

   Interpretación: La sensibilidad para la clase M es del 89.29%, indicando que el modelo identifica correctamente el 89.29% de las instancias verdaderamente positivas.

$$Sensibilidad(M) - \frac{TP}{TP + FN} - \frac{50}{50 + 6} \approx 89.29\%$$

Sensibilidad (B)

Interpretación: Para la clase B, la sensibilidad es del 99.13%, reflejando una excelente capacidad del modelo para detectar instancias verdaderamente negativas.

$$Sensibilidad(B) - \frac{TP}{TP + FN} - \frac{114}{114 + 1} \approx 99.13\%$$

4. Especificidad (Specificity)

   Interpretación: La especificidad muestra que el modelo identifica correctamente el 99.13% de las instancias verdaderamente negativas, minimizando los falsos positivos.

$$Especificidad - \frac{TN}{TN + FP} - \frac{114}{114 + 1} \approx 99.13\%$$

5. Valor F (F1 Score)

   Interpretación: El valor F para la clase M es del 93.44%, lo que combina precisión y sensibilidad para medir la calidad del modelo en predecir la clase positiva. Para la clase B, el valor F es del 98.58%, indicando un excelente balance entre precisión y sensibilidad.

$$Valor\ F(M) - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9804 * 0.8929)}{0.9804 + 0.8929} \approx 93.44\%$$

$$Valor\ F(B) - \frac{2 * (Precisión * Sensibilidad)}{Precisión + Sensibilidad} - \frac{2 * (0.9913 * 0.9804)}{0.9913 + 0.9804} \approx 98.58\%$$

3.4. USO DE HERRAMIENTAS CON METODO NO SUPERVISADO

Se usará el método K Means ya que, al buscar la media de los puntos del clúster, nos ayuda a encontrar la estructura natural de los datos sin usar las etiquetas.

Para evaluar la efectividad de la agrupación analizaremos sus métricas individualmente ya que, a diferencia de los anteriores ejemplos en los resultados obtenidos en los métodos no supervisados, son diferentes y no muestran los mismos datos.

3.4.1. K-Means en Orange:

Se construye el modelo:

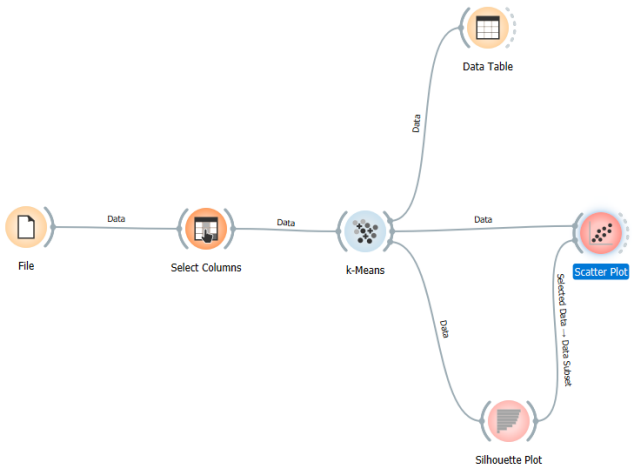

Figura 33: K-Mine en Orange

Dentro del nodo k-Means nos permite seleccionar la cantidad de clústeres que usaremos después de haber elegido diagnosis para los datos.

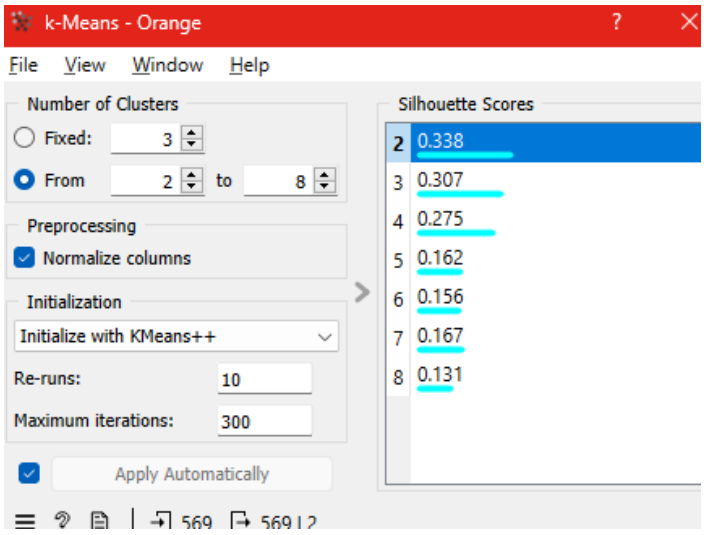

Figura 34: Configuración K-means en Orange

Un valor de silueta de 0.338 sugiere que los puntos están moderadamente bien agrupados. Es un indicativo de que la cohesión y separación de los clústeres es aceptable, pero no óptima.

1. Coeficiente de Silueta (Silhouette Coefficient)

El coeficiente de silueta evalúa la similitud de un punto con los puntos de su propio clúster en comparación con los puntos de otros clústeres. En la Fig. 38 se muestran los valores de silueta para cada instancia y para los centroides de los clústeres.

Interpretación:

Valores para instancias: Los valores individuales de silueta indican qué tan bien está cada punto agrupado dentro de su clúster.

2. Centroides de los Clústeres

La tabla 2 muestra las características de los centroides (C1 y C2) para cada clúster, lo cual es útil para entender las características promedio de los datos agrupados en cada clúster.

Data: **data**: 569 instances, 35 variables
Features: 32 numeric (3.1% missing values)
Target: categorical
Metas: 2 (1 categorical, 1 numeric)

	diagnosis	Cluster	Silhouette	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mea
1	M	C1	0.591922	842302	17.990	10.38	122.80	1001.0	0.11840
2	M	C1	0.517079	842517	20.570	17.77	132.90	1326.0	0.08474
3	M	C1	0.605839	84300903	19.690	21.25	130.00	1203.0	0.10960
4	M	C1	0.543174	84348301	11.420	20.38	77.58	386.1	0.14250
5	M	C1	0.554087	84358402	20.290	14.34	135.10	1297.0	0.10030
6	M	C1	0.509209	843786	12.450	15.70	82.57	477.1	0.12780

Centroids: **data centroids**: 2 instances, 33 variables
Features: 31 numeric (no missing values)
Metas: 2 (1 categorical, 1 numeric)

	Cluster	Silhouette	id	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	co
1	C1	0.545348	0.0771372	0.986498	0.49202	1.01867	0.974797	0.58712	1.014073	1.14
2	C2	0.629012	-0.0380624	-0.486776	-0.24278	-0.50265	-0.481002	-0.28971	-0.500382	-0.5

Tabla 2: Datos de K-Means en Orange

1. Distribución de los Clústeres (Scatter Plot):

- Cohesión: Los puntos están bien agrupados por color, sugiriendo que los clústeres son coherentes internamente.
- Separación: Hay una clara separación entre los grupos de colores, lo que indica una buena segmentación de los datos.

2. Coeficiente de Silueta (Silhouette Plot):

- Valores Positivos: La mayoría de los puntos tienen valores de silueta positivos, lo que indica que están bien agrupados.
- Valores Negativos: Los valores negativos indican puntos que pueden estar mal clasificados.
- Promedio de Silueta: Un valor promedio positivo cercano a 0.3 sugiere que, en general, los clústeres están razonablemente bien definidos, pero hay margen de mejora.

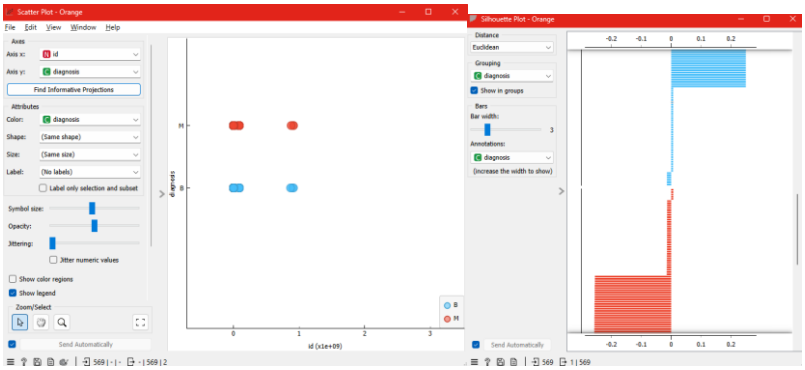

Figura 35: Esquemas scatter y silhoutte de K-Means en Orange

3.4.2. K-Means en Rapid Miner:

Se diseña el modelo en Rapid Miner:

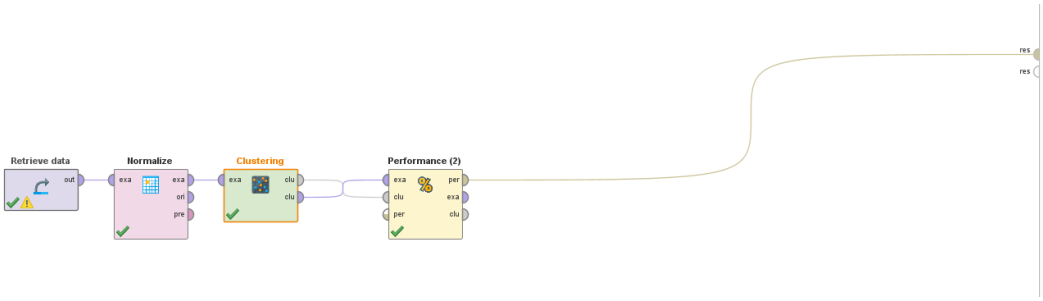

Figura 36: Diseño de Kmeans en Rapid miner

Podemos configurar los grupos que deseamos.

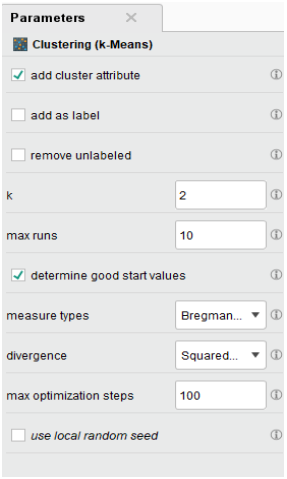

Figura 37: Configuración de Kmeans en Rapid Miner

Una vez echo el modelo podemos elegir entre varias graficos como el siguiente de scatter plot:

1. Cohesión: Los puntos de cada grupo están bien agrupados, lo que sugiere que radius_mean es una característica importante para la clasificación.
2. Separación: La separación clara entre los grupos indica que radius_mean diferencia efectivamente entre los diagnósticos M y B.
3. Patrones de Datos: La diferencia en los valores de radius_mean entre M y B sugiere que las instancias de M tienen características distintas en comparación con las de B.

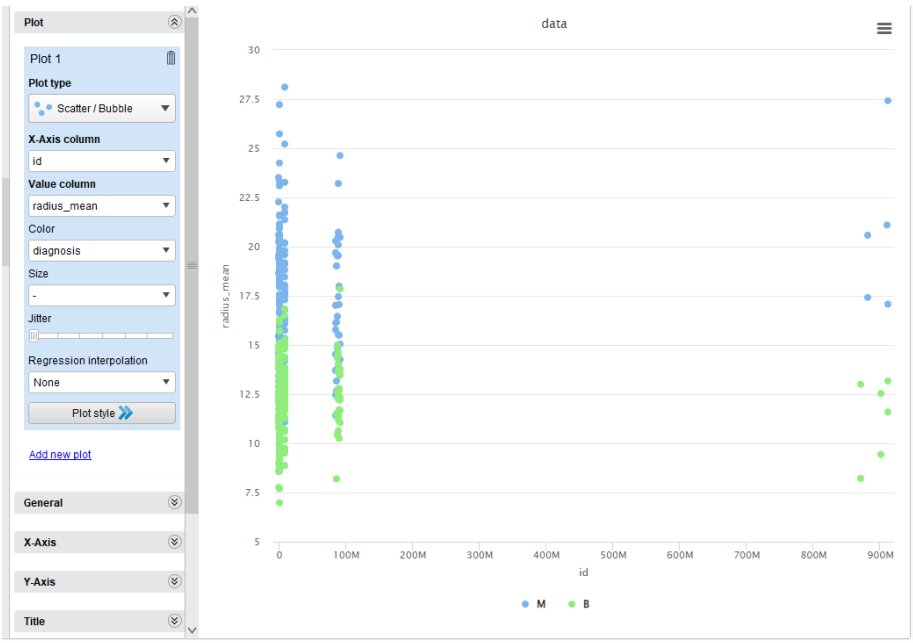

Figura 38: Figura de Scatter en Rapid miner

Al no poseer nodos de desempeño la manera de ver su precisión es viendo la distancia del centroide en la información que la misma herramienta nos ofrece.

1. Distancia Promedio dentro del Centroide:

- La distancia promedio dentro de todos los centroides (-20.343) y las distancias específicas para cada clúster (clúster 0: -33.115, clúster 1: -14.041) indican la cohesión de los datos dentro de cada clúster.
- Valores menores (en términos absolutos) sugieren que los datos están bien agrupados alrededor de los centroides.

2. Índice Davies-Bouldin:

- Un índice Davies-Bouldin de -1.309 indica la separación y compactación de los clústeres. Normalmente, un valor más bajo (positivo) indica una mejor agrupación. El valor negativo aquí debe ser revisado, pero generalmente se busca un valor cercano a 0.

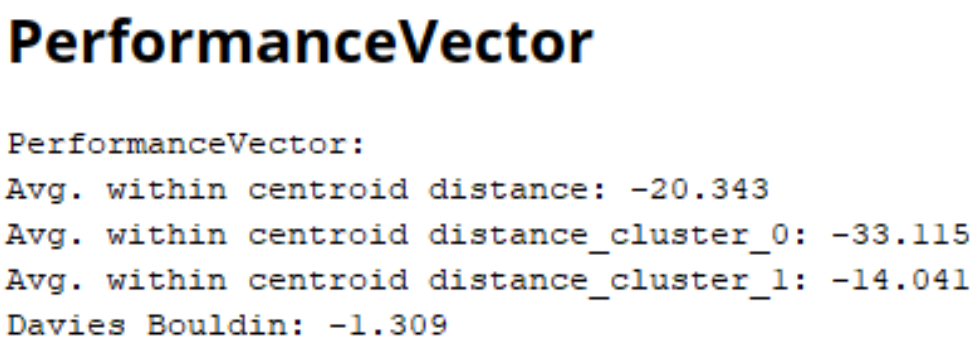

Figura 39: Vector de desempeño de algoritmo Kmeans en Rapid Miner

3.4.3. K-Means en Knime:

Diseñamos el modelo en Knime:

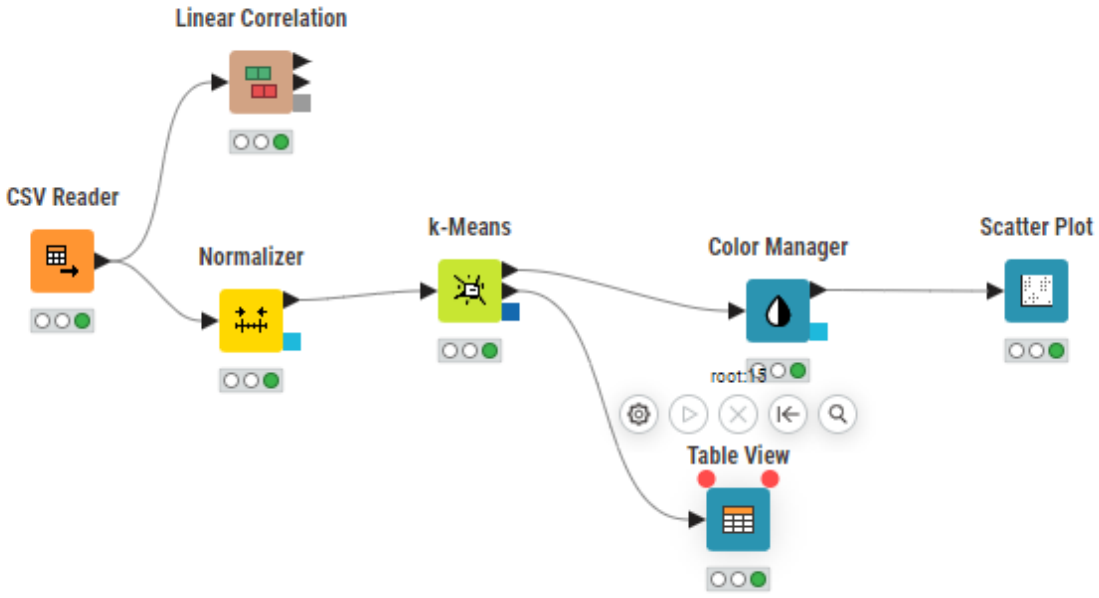

Figura 40: Diseño de Kmeans en K-Nime

Con el nodo Table View podemos ver los datos clasificados.

1. Valores Promedio de las Características:

Los valores promedio para cada característica indican cómo se agrupan las instancias dentro de cada clúster.

Interpretación:

- cluster_0 tiene valores promedio más altos en todas las características en comparación con cluster_1.
- Por ejemplo, radius_mean es significativamente mayor en cluster_0 (0.505) comparado con cluster_1 (0.255), lo que sugiere que las instancias en cluster_0 tienden a tener radios más grandes.

- Similarmente, perimeter_mean, area_mean, y otras características son mayores en cluster_0, indicando que este clúster puede agrupar instancias con características más grandes o prominentes.

2. Comparación de Clústeres:

- cluster_0 tiene características más pronunciadas, lo que puede indicar que agrupa instancias con valores más altos en las características analizadas.
- cluster_1 agrupa instancias con características menos pronunciadas en comparación.

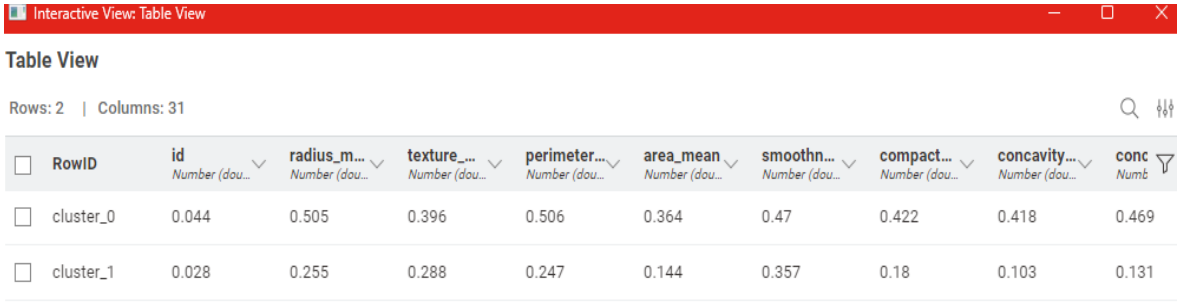

Figura 41: Tabla de clústeres en Knime

Y con el nodo Scatter nos permite ver el grafico de la clasificación.

1. Cohesión de los Grupos:

Los puntos correspondientes a cada diagnóstico (M y B) están bien agrupados en sus respectivas líneas horizontales.

- Interpretación: Esto sugiere una buena separación entre los dos grupos de diagnóstico en el eje Y, pero no proporciona información sobre la separación en otras dimensiones (como características del dataset).

2. Distribución de ID:

Los puntos se distribuyen a lo largo del eje X sin un patrón claro.

- Interpretación: El id no parece tener una relación directa con la variable diagnosis, lo que es esperado ya que id es una variable nominal que solo identifica instancias y no tiene un valor analítico directo en términos de diagnóstico.

3. Diagnóstico:

Los puntos azules corresponden a los diagnósticos M y B.

- Interpretación: Los datos muestran claramente las dos categorías de diagnóstico, lo que es útil para entender la distribución general de los datos según la variable diagnosis.

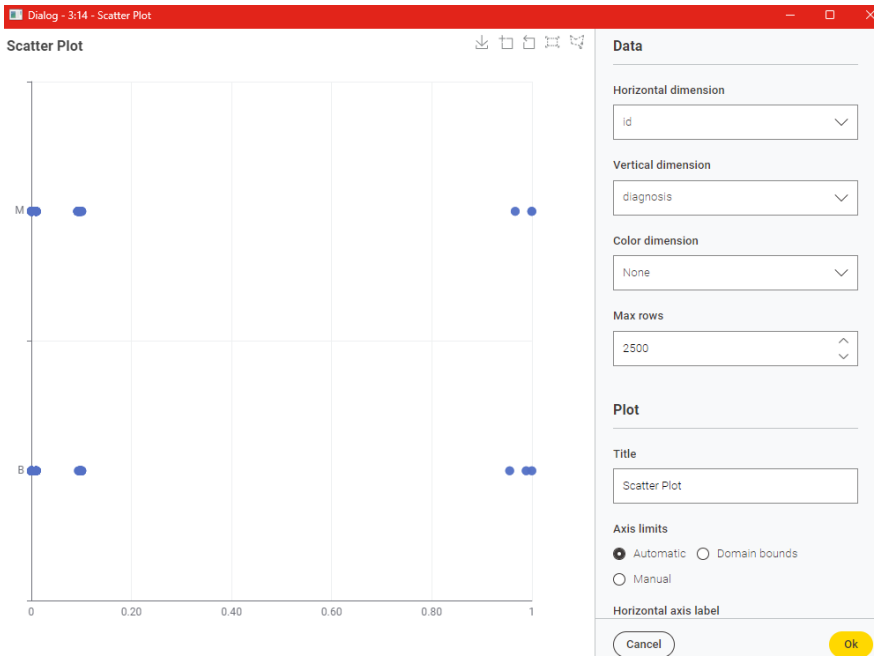

Figura 42: Scatter en K-Nime

4. CAPÍTULO IV: ANÁLISIS DE RESULTADOS

4.1. TABLA COMPARATIVA

A continuación, se muestran los resultados acumulativos adjuntados en una tabla:

Característica	*Orange*	*Rapidminer*	*Knime*
Nombre de la herramienta	Orange	AI Studio	Knime
Característica principal	Plataforma de visualización y análisis de datos con widgets para tareas de minería de datos, visualización y machine learning.	Plataforma de ciencia de datos con flujo de trabajo visual, soporte para machine learning y deep learning, análisis predictivo.	Plataforma de análisis de datos con flujo de trabajo visual, integración de múltiples herramientas de minería de datos y machine learning.
Posibilidades de Uso	Análisis exploratorio de datos, visualización interactiva, machine learning educativo y simple.	Análisis predictivo, modelado de datos, text mining, análisis de series temporales.	Análisis predictivo, procesamiento de datos, integración de datos, análisis de datos complejos.
Facilidad de Uso	Alta, con una interfaz basada en widgets que facilita la construcción de flujos de trabajo.	Alta, con interfaz de usuario intuitiva y soporte para creación de modelos sin código.	Media, interfaz visual con nodos, pero requiere mayor aprendizaje para uso avanzado.
Escalabilidad	Media, adecuado para proyectos de tamaño pequeño a mediano.	Alta, soporta grandes volúmenes de datos y puede integrarse con sistemas distribuidos como Hadoop y Spark.	Alta, puede manejar grandes volúmenes de datos y se integra bien con tecnologías de big data como Apache Hadoop y Spark.
Integraciones	Varias bibliotecas de Python, R, y otras herramientas de machine learning.	Bases de datos SQL, NoSQL, Hadoop, Spark, servicios en la nube como AWS, Google Cloud y Azure.	Bases de datos SQL, NoSQL, sistemas big data (Hadoop, Spark), herramientas de BI y servicios en la nube.

Soporte	Buena, con documentación, comunidad activa y tutoriales disponibles.	Excelente, con documentación extensa, comunidad activa y soporte profesional disponible.	Excelente, con documentación extensa, una comunidad activa y soporte profesional.
Costo	Gratuito y de código abierto.	Pago, con diferentes planes de suscripción y una versión gratuita con funcionalidades limitadas.	Gratuito y de código abierto, con versiones comerciales disponibles para características adicionales y soporte profesional.
Calificación en base a su uso	4.3/5 - Valorada por su simplicidad y efectividad para tareas educativas y de visualización.	4.8/5 - Destacado por su facilidad de uso y potente capacidad de análisis predictivo.	4.5/5 - Elogiado por su flexibilidad, capacidad de integración y robustez en el análisis de datos.

Tabla 3.Comparación de Resultados

4.2. VENTAJAS Y DESVENTAJAS

Herramienta	*Ventajas*	*Desventajas*
Orange	Facilidad de Uso: Interfaz de arrastrar y soltar muy intuitiva, ideal para principiantes y prototipos rápidos. Visualización Interactiva: Excelente para exploración y análisis visuales, con gráficos interactivos y fáciles de interpretar.	Capacidades Limitadas de Preprocesamiento: Herramientas de preprocesamiento básicas, menos adecuadas para tareas complejas. Soporte para Algoritmos Avanzados: Menor variedad de algoritmos avanzados y menos opciones de personalización en comparación con otras herramientas.
Rapid Miner	Capacidades de Preprocesamiento: Amplias herramientas para limpieza y transformación de datos, adecuado para	Curva de Aprendizaje Media: La interfaz ofrece muchas opciones que pueden ser confusas para usuarios

	proyectos que requieren preprocesamiento intensivo. Soporte para Algoritmos: Extenso soporte para algoritmos de aprendizaje automático y minería de datos, fácil integración de nuevos algoritmos.	nuevos, con una curva de aprendizaje algo empinada. Costos: Versiones avanzadas y soporte pueden ser costosos, lo que puede ser una barrera para individuos o pequeñas organizaciones.
Knime	Flexibilidad y Potencia: Amplia variedad de nodos para tareas complejas de análisis de datos y capacidad de manejar flujos de trabajo robustos. Extensibilidad: Fuerte soporte para extensiones y plugins, integración con Python, R, y Big Data.	Curva de Aprendizaje: Mayor complejidad en la configuración y uso, lo que puede ser abrumador para principiantes. Consumo de Recursos: Puede requerir más recursos computacionales, especialmente con flujos de trabajo grandes o complejos.

Tabla 4: Ventajas y Desventajas

4.3. ALGORITMOS Y TÉCNICAS PRINCIPALES EN MINERÍA DE DATOS

Según (Cañadas, 2022) en AbDatum, la minería de datos es un campo de la estadística que aplica diferentes métodos y estrategias con el objetivo de encontrar patrones en grandes cantidades de datos. Para conseguirlo hace uso de metodologías de la estadística, computación, ciencia de datos o programación. En este artículo, se detallan 9 de los algoritmos y técnicas más usados en el data mining o minería de datos para encontrar la información relevante que se esconde dentro del dato:

- Limpieza de datos: Proceso de eliminar y corregir errores en los datos.
- Feature engineering: Crea nuevas características para mejorar modelos.
- Árboles de decisión: Modelos jerárquicos para tomar decisiones.
- Random Forest: Conjunto de árboles para mejorar precisión.
- Máquinas de vectores de soporte (SVM): Divide datos en espacios separados para clasificación.
- Técnicas de clusterización: Agrupa datos similares en clústeres.
- K Nearest Neighbors (KNN): Clasifica basado en vecinos más cercanos.

- Redes neuronales: Algoritmos inspirados en la estructura y función del cerebro humano, utilizados para detectar patrones complejos en los datos.
- Naive Bayes: Clasifica usando el teorema de Bayes con suposición ingenua de independencia.

A continuación, se presenta una tabla que resume la disponibilidad de cada algoritmo en las herramientas Orange, KNIME y RapidMiner:

Algoritmo	Orange	KNIME	Rapid Miner
Limpieza de datos	Si	Si	Si
Feature engineering	Si	Si	Si
Árboles de decisión	Si	Si	Si
Random Forest	Si	Si	Si
Máquinas de vectores de soporte (SVM)	Si	Si	Si
Técnicas de clusterización	Si	Si	Si
K Nearest Neighbors (KNN)	Si	Si	Si
Redes neuronales	Si	Si	Si
Naive Bayes	Si	Si	Si

Tabla 5: Check List de algoritmos comunes en las herramientas

4.4. EVALUACIÓN TÉCNICA DE LAS HERRAMIENTAS

El análisis de las herramientas Orange, KNIME y RapidMiner ha revelado que cada una de ellas ofrece ventajas específicas que pueden ser aprovechadas dependiendo del contexto del proyecto de minería de datos.. Orange se destaca por su interfaz intuitiva y visualizaciones interactivas, lo cual facilita la comprensión de conceptos complejos para quienes están comenzando en minería de datos. Esta accesibilidad hace que Orange sea especialmente útil para estudiantes y académicos que desean aprender y aplicar técnicas de manera práctica sin necesidad de tener una experiencia previa en programación.

Por otro lado, KNIME sobresale por su capacidad para manejar flujos de trabajo complejos y grandes volúmenes de datos. La flexibilidad y la integración de múltiples herramientas lo hacen ideal para proyectos que requieren análisis detallados y personalización en cada fase del proceso.

La comunidad activa y el soporte técnico de KNIME también son recursos valiosos para resolver problemas y avanzar en el dominio de técnicas avanzadas de minería de datos.

RapidMiner ofrece una robusta automatización de procesos y una amplia gama de modelos avanzados, lo que lo hace ideal para proyectos que necesitan análisis predictivo y optimización de modelos de manera eficiente. Aunque puede tener una curva de aprendizaje más pronunciada, las capacidades analíticas de RapidMiner son poderosas para investigaciones y aplicaciones prácticas tanto en entornos académicos como empresariales.

En conclusión, la elección entre Orange, KNIME y RapidMiner depende de las necesidades específicas del proyecto y del nivel de experiencia en minería de datos. Orange destaca como la opción preferida para quienes se están iniciando en este campo, proporcionando una plataforma accesible y educativa para aprender y experimentar con técnicas avanzadas de análisis de datos.

4.5. CONCLUSIONES

- Las herramientas de minería de datos Orange, RapidMiner, y KNIME ofrecen una variada gama de capacidades, con diferencias notables en términos de flexibilidad y escalabilidad. RapidMiner y KNIME destacan en la capacidad de manejar flujos de trabajo complejos y grandes volúmenes de datos, gracias a sus robustas opciones de preprocesamiento y extensibilidad. Orange, con su interfaz amigable e intuitiva, es especialmente efectiva para la enseñanza, la exploración de datos y la creación rápida de prototipos, aunque su escalabilidad y capacidades de preprocesamiento son más limitadas.

- La facilidad de uso varía considerablemente entre estas herramientas, afectando la curva de aprendizaje para usuarios nuevos. Orange se presenta como la herramienta más accesible para principiantes debido a su diseño de arrastrar y soltar y su enfoque en la visualización interactiva. RapidMiner, aunque tiene una interfaz similar basada en flujo, requiere un tiempo considerable para dominar debido a su extenso menú de opciones y configuraciones. KNIME, mientras tanto, ofrece gran flexibilidad y potencia, pero su complejidad y la necesidad de configuraciones detalladas pueden resultar abrumadoras para usuarios sin experiencia previa.

- En términos de preprocesamiento y soporte para algoritmos avanzados, RapidMiner y KNIME ofrecen una cobertura más amplia, facilitando tareas complejas de limpieza, transformación de datos, y análisis predictivo. Estas herramientas son adecuadas para usuarios que requieren un análisis avanzado y detallado, así como integración con otros sistemas y tecnologías. Orange, aunque eficaz para tareas básicas y de exploración, ofrece menos opciones para preprocesamiento complejo y algoritmos avanzados, limitando su uso en aplicaciones de análisis de datos más sofisticadas.

- Por su facilidad de uso y su capacidad para simplificar conceptos complejos a través de una interfaz intuitiva y widgets interactivos, Orange se destaca como la mejor opción para apoyar el aprendizaje educativo y la rápida introducción a la minería de datos. Su diseño facilita a los estudiantes y principiantes la comprensión de los procesos de análisis de datos y el desarrollo de habilidades prácticas sin la necesidad de una curva de aprendizaje pronunciada.

4.6. RECOMENDACIONES

- Al seleccionar una herramienta de minería de datos, es fundamental considerar la complejidad del proyecto y las capacidades requeridas. Para proyectos que demandan un análisis avanzado, integración con múltiples fuentes de datos, y flujos de trabajo complejos, RapidMiner y KNIME son opciones recomendables debido a su flexibilidad y potencia. Sin embargo, para proyectos que requieren prototipos rápidos, enseñanza o análisis exploratorio, Orange es una excelente opción por su simplicidad y facilidad de uso.

- Dada la variabilidad en la curva de aprendizaje de estas herramientas, se recomienda invertir en capacitación adecuada para el personal que utilizará las herramientas. Para organizaciones que optan por RapidMiner y KNIME, considerar la contratación de soporte técnico o la adquisición de materiales educativos puede acelerar la curva de aprendizaje y mejorar la eficiencia en el uso de estas herramientas complejas. En el caso de Orange, aunque más intuitivo, la capacitación en el uso de widgets y módulos específicos puede optimizar su aplicación.

- Es importante realizar una evaluación continua de las necesidades del proyecto y las capacidades ofrecidas por la herramienta seleccionada. A medida que los proyectos evolucionan y las necesidades de análisis se vuelven más complejas, puede ser necesario

migrar a una herramienta más robusta o integrar múltiples herramientas para aprovechar las fortalezas de cada una. En este contexto, combinar Orange para la fase inicial de exploración de datos con KNIME o RapidMiner para el análisis avanzado y la integración de datos puede proporcionar una solución integral y eficiente.

- Para elegir la herramienta adecuada, se recomienda usar Orange en contextos educativos y de enseñanza, donde la simplicidad y la visualización intuitiva son cruciales. RapidMiner es ideal para usuarios que requieren capacidades avanzadas de preprocesamiento y modelado predictivo, mientras que KNIME es preferible para proyectos empresariales complejos que requieren integración de datos desde diversas fuentes y una mayor flexibilidad en la construcción de flujos de trabajo.

5. GLOSARIO DE TÉRMINOS

- A
  - Algoritmo
    - Procedimiento o fórmula para resolver un problema. En el contexto de minería de datos, se refiere a métodos específicos para realizar tareas de análisis de datos.
  - Árbol de decisión
    - Herramienta de soporte de decisiones que utiliza un modelo en forma de árbol para representar decisiones y sus posibles consecuencias.
- B
  - Base de datos
    - Colección organizada de datos que permite su acceso, gestión y actualización.
- C
  - Clasificación
    - Técnica de aprendizaje supervisado que asigna una etiqueta a un conjunto de datos basado en las características de los datos.
  - Clústeres
    - Grupos de datos similares formados a partir de un conjunto de datos más grande, utilizados en técnicas de aprendizaje no supervisado.

- D
  - Data mining
    - Proceso de descubrir patrones en grandes conjuntos de datos utilizando técnicas de aprendizaje automático, estadística y sistemas de bases de datos.

  - Dataset
    - Conjunto de datos utilizados para análisis y procesamiento en minería de datos.

- E
  - Evaluación
    - Proceso de medir la efectividad de un modelo o técnica en minería de datos.

  - Extracción de conocimiento
    - Proceso de identificar información útil y comprensible a partir de grandes volúmenes de datos.

- I
  - Indicadores de desempeño
    - Métricas utilizadas para evaluar la efectividad de modelos de minería de datos.

- K
  - K-means
    - Algoritmo de agrupamiento que divide un conjunto de datos en K clústeres o grupos basados en características similares.

  - KNIME
    - Plataforma de análisis de datos que permite la integración de herramientas de minería de datos y aprendizaje automático.

- M
  - Minería de datos
    - Proceso de analizar grandes conjuntos de datos para encontrar patrones y relaciones que pueden ser útiles para la toma de decisiones.

- Método
    - Procedimiento o técnica utilizada en minería de datos para analizar y procesar datos.

- O
    - Orange
        - Herramienta de software para la minería de datos y aprendizaje automático que proporciona una interfaz visual para análisis de datos.

- P
    - Preprocesado
        - Etapa de preparación de datos para el análisis que incluye la limpieza y transformación de datos brutos.

- R
    - RapidMiner
        - Plataforma de software que proporciona herramientas para el análisis predictivo y la minería de datos.

- S
    - Selección de características
        - Proceso de identificar las características más relevantes para el análisis de datos.

- T
    - Técnicas de aprendizaje no supervisado
        - Métodos de aprendizaje automático que identifican patrones en datos sin etiquetas predeterminadas.

    - Técnicas de aprendizaje supervisado
        - Métodos de aprendizaje automático que utilizan datos etiquetados para entrenar modelos predictivos.

- V
    - Validación cruzada
        - Técnica de evaluación que divide los datos en partes para entrenar y probar un modelo en diferentes subconjuntos de datos.

- Visualización de datos
    - Representación gráfica de datos para facilitar la comprensión de patrones y tendencias.

Las definiciones fueron tomadas y generadas en base a Real Academia Española. (s.f.). Diccionario de la lengua española. Recuperado el 7 de junio, 2024, de https://www.rae.es/

6. REFERENCIAS BIBLIOGRÁFICAS

Aggarwal, C. (2015). *Data Mining: The Textbook.* . Springer.

Agrawal, R., Imielinski, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases.* . ACM SIGMOD Record.

Beltrán González, B. (2011). *Notas de Minería de Datos. Facultad de Ciencias de la Computación, BUAP.* Obtenido de https://www.cs.buap.mx/~bbeltran/NotasMD.pd

Breiman, L. (2001). *Random forests.* . Machine Learning.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and Regression Trees.* . Wadsworth & Brooks/Cole. .

Cañadas, R. (12 de febrero de 2022). *Abdatum*. Obtenido de https://abdatum.com/tecnologia/algoritmos-mineria-datos

Chaudhuri, S., Dayal, U., & Narsayya, V. (2011). *An overview of business intelligence technology.* . Communications of the ACM.

Chen, M., Mao, S., & Liu, Y. (2014). *Big Data: A Survey.* . Mobile Networks and Applications, .

Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification.* . IEEE Transactions on Information Theory.

Davenport, T. H., & Patil, D. J. (2012). *Data Scientist: The Sexiest Job of the 21st Century.* . Harvard Business Review.

Demsar, J., Zupan, B., Leban, G., & Curk, T. (2013). *Orange: Data Mining Toolbox in Python.* . Journal of Machine Learning Research.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From data mining to knowledge discovery in databases.* . AI Magazine.

Gantz, J., & Reinsel, D. (2012). *Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC.* . Obtenido de https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

Garcia, J. (2015). *Minería de datos y sus aplicaciones.* . Editorial X.

García, P. (2018). *Métricas de evaluación en modelos de clasificación: precisión, exactitud, sensibilidad, especificidad y valor F. Revista de Minería de Datos y Análisis Predictivo*. Obtenido de https://abdatum.com/tecnologia/algoritmos-mineria-datos

Gartner. (2022). *Magic Quadrant for Data Science and Machine Learning Platforms.* . Obtenido de https://www.gartner.com/doc/reprints?id=1-25TML95C&ct=221227&st=sb

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* . Elsevier.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining.* . MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* . Springer.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications.* . Chapman & Hall/CRC.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression.* . Wiley.

IONOS. (23 de junio de 2024). *Software de Data Mining: Las mejores herramientas.* Obtenido de https://www.ionos.com/es-us/digitalguide/online-marketing/analisis-web/software-de-data-mining-las-mejores-herramientas/

Knime. (2023). *Why KNIME? https://www.knime.com/why-knime* . Obtenido de https://www.knime.com/why-knime

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis (5th ed.).* . Wiley.

Páez, S. (2016). *Análisis comparativo de herramientas open source para data mining sobre datos públicos del ministerio de educación de la República del ecuador.* PUCE. Quito: PUCE.

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach (4th ed.).* . Pearson.

Tan, P., Steinbach, M., & Kumar, V. (2018). *Introduction to Data Mining (2nd ed.).* . Pearson.

Terán, E. (06 de julio de 2022). *Implementación de minería de datos en la gestión académica de las instituciones de educación superior.* Obtenido de http://pucedspace.puce.edu.ec/handle/23000/4660

Van Rijin, J., Vanschgren, J., Torgo, L., & Brazdil, P. (2018). *A perspective on open science in machine learning. Springer.* Obtenido de https://link.springer.com/article/10.1007/s00521-018-3551-3

Witten , I., Frank, E., Hall, M., & Pal, C. (2016). *Data Mining: Practical Machine Learning Tools and Techniques (4th ed.).* . Morgan Kaufmann.

Zaki, M., & Wagner, M. (2017). *Data Mining and Machine Learning: Fundamental Concepts and Algorithms.* . Cambridge University Press.