

## Revista Científica y Tecnológica UPSE

### Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela ‘Lic. Angélica Villón L.’

Application of data mining techniques to predict the academic performance of the students of the ‘Lic. Angélica Villón L.’



(1,2)Walter Armando Orozco Iguasnia\* <https://orcid.org/0000-0002-8224-1152>, (2)Alex Joao Villao Balón <https://orcid.org/0000-0002-1214-7304>, (2)Jaime Benjamín Orozco Iguasnia <https://orcid.org/0000-0002-6929-1411>, (3)Mercedes de Jesús Villarroel Sánchez <https://orcid.org/0000-0001-8373-5747>.

(1) Universidad Rey Juan Carlos, España.

(2) Universidad Península de Santa Elena, Ecuador.

(3) Unidad Educativa Salinas. (24h00272)

#### Resumen

Una de las metas que tiene la escuela ‘Lic. Angélica Villón L.’ es el mejoramiento del nivel académico de sus estudiantes, para ello, contar con herramientas que permitan disponer de información académica a tiempo para la toma de decisiones resulta básico. Para este fin, se priorizan elementos como la población estudiantil, las calificaciones que alcanzan en cada nivel, el soporte familiar, entre otros; todo esto catalogado como el rendimiento académico del estudiante. No obstante, la evaluación de este indicador se encuentra limitada debido a la influencia de varios factores que requieren ser procesados de forma integrada en función de su nivel de influencia. La investigación realizada es de tipo observacional, nivel exploratorio; muestra el empleo de herramientas de inteligencia de negocios como soporte para la toma de decisiones; se crea un almacén de datos como repositorio unificado a través de procesos ETL; se entrena modelos de aprendizaje supervisado como máquina de soporte de vectores, redes neuronales y árboles de decisión de regresión para predecir el rendimiento académico. Los datos históricos de los estudiantes son la fuente para la aplicación de los modelos. Finalmente, se identifica el modelo que mejor precisión tiene a través de métricas válidas en el contexto de análisis de regresiones.

#### Abstract

One of the goals of the ‘Lic. Angélica Villón L.’ school is to improve the academic level of its students. For this, having tools that allow the availability of academic information for decision-making is essential. For this purpose, elements such as the student population, the grades achieved at each level, family support, among others, are prioritized, all cataloged as the student's academic performance. However, the evaluation of this indicator is limited due to the influence of several factors that need to be processed in an integrated way depending on their level of influence. The research carried out is observational, exploratory level; shows the use of business intelligence tools as support for making decisions; a data warehouse is created as a unified repository through ETL processes; Supervised learning models such as vector support machines, neural networks and regression decision trees are trained to predict academic performance. Historical student data is the source for the application of the models. Finally, the model with the best precision is identified through valid metrics in the context of regression analysis.

#### Palabras clave:

Minería de datos, inteligencia de negocios, educación.

#### Keywords:

Data mining, business intelligence, education

**Recibido:** noviembre 4/2021    **Aceptado:** noviembre 20/2021    **Publicado:** diciembre 28/2021

**Forma de citar:** Orozco Iguasnia, W.; Villao Balón, A.; Orozco Iguasnia, J.; Villarroel Sánchez, M. (2021). Aplicación de técnicas de minería de datos para predecir el desempeño académico de los estudiantes de la escuela ‘Lic. Angélica Villón L.’. Revista Científica y Tecnológica UPSE, 8 (2) pág. 68-75. DOI: 10.26423/rctu.v8i2.637.

\* *Autor para correspondencia:* worozco@upse.edu.ec

## 1. Introducción

Actualmente, las organizaciones están experimentando importantes cambios que aportan al fortalecimiento de su gestión con base en la aplicación de diferentes métodos para fundamentar la toma de decisiones [1]. En particular, las organizaciones fundamentan su accionar en los datos y su significado como soporte para el trabajo diario. Esta área de estudio se la conoce como Inteligencia de Negocios (BI), que, entre otros aspectos, trae consigo un sinnúmero de beneficios de los que se destacan:

- Permite una visión del pasado, el presente, y el futuro al que puede aspirar una empresa.
- Se acompaña estrictamente del monitoreo con reglas del negocio o métricas que permiten mantener el control de las metas fundamentales de la empresa.
- Aporta información actualizada.

Bajo este contexto, las estrategias en BI se pueden interpretar como la coordinación de forma efectiva de las tecnologías para el análisis adecuado de los datos, cuyo fin es alinearse a las metas y objetivos de una organización.

El desarrollo de las tecnologías de la información ha generado una gran cantidad de bases de datos y enormes datos en diversas áreas [2]. La investigación en bases de datos y tecnología de la información ha dado lugar a un enfoque orientado al almacenamiento y manipulación de los datos como soporte para la toma de decisiones. Este enfoque es el que se conoce como la minería de datos, se centra en acciones como el descubrimiento de conocimientos por medio de la extracción y el análisis de datos.

En cuanto a metodologías para la aplicación de minería de datos se distinguen varias. Según Gironés en el libro “Minería de datos: Modelos y algoritmos”, se destaca la metodología CRISP-DM compuesta por las fases: comprensión de negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, y finalmente, despliegue [3].

Por otro lado, Hernández en el libro “Introducción a la Minería de Datos”, resalta la metodología Descubrimiento de Conocimiento en Bases de Datos (KDD- Knowledge Discovery in Databases), la cual, se compone de cinco fases: integración y recopilación, selección, limpieza y transformación, minería de datos, evaluación e interpretación, y finalmente, difusión y uso [4].

Para un progreso óptimo, las instituciones buscan priorizar y establecer estrategias efectivas que solventen sus necesidades. En este aspecto, la información toma un papel primordial, y es, a través de la misma, que las instituciones desarrollan una mentalidad basada en la mejora continua, donde el análisis de los datos se vuelve una característica fundamental. Para el presente estudio,

se toma como caso de estudio la escuela “Lic. Angélica Villón Lindao” ubicada en el cantón de Santa Elena de la provincia del mismo nombre.

Esta institución, con el propósito de evaluar su desarrollo académico con base en los datos académicos históricos que dispone, tiene la necesidad de implementar una herramienta que apoye la toma de decisiones al finalizar un período escolar. Por ende, la aplicación de metodologías de minería de datos llega a facilitar el análisis extenso del flujo de información.

Una de las debilidades detectadas en la institución, es la falta de información organizada que permita a través de un adecuado proceso, predecir cómo serán los patrones futuros del rendimiento académico, para poder tomar las decisiones adecuadas que salvaguarden las metas institucionales. Trabajos anteriores muestran que existen al menos dos limitantes en este proceso: la cantidad de estudiantes evaluados (población) y la variedad de técnicas empleadas para realizar una correcta minería de datos. El presente artículo muestra cómo la aplicación de metodologías de minería de datos favorece, no solo la toma de decisiones, sino la comunicación y la administración de datos.

Mediante el empleo de la metodología KDD se evidencia la recopilación e integración de los datos, la creación de un almacén de datos o data warehouse, y, la aplicación de tres técnicas de minería de datos: árboles de decisión, redes neuronales y vectores de soporte de regresión, una variante de Máquinas de Vectores de Soporte (SVM-Support Vector Machines). Se trata de seleccionar el mejor modelo que se ajuste a los datos de rendimiento académico que se dispone.

Siendo modelos de regresión, para la evaluación de las técnicas se emplean tres métricas: error absoluto medio (MAE-Mean Absolute Error), error cuadrático medio (MSE-Mean Square Error) y la raíz del error cuadrático medio (RMSE-Root Mean Square Error), además, el coeficiente de determinación  $R^2$  [5].

## 2. Trabajos relacionados

Del trabajo “Técnicas de minería de datos para mejorar la precisión de las predicciones del rendimiento académico de los estudiantes: un estudio de caso con Xorro-Q” desarrollado por Gomathy Suganya [6], se destaca el uso de modelos clasificadores, sin embargo, existen restricciones como la limitación a un solo curso o el desequilibrio de los datos durante su evaluación. Otro aporte, “Uso de Aprendizaje Supervisado para predecir el rendimiento estudiantil” desarrollado por Murat Pojon [7], tiene su enfoque en un caso de estudio con fuentes de datos públicas, lo que es una característica importante ya que se puede disponer de datos para validar la eficiencia de los modelos.

En el artículo “Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de

ingeniería, mediante aprendizaje automático” resalta el uso de algoritmos como K vecinos más cercanos, árboles de decisión y perceptrón multicapa para mediante algoritmos de clasificación, estimar el rendimiento académico de estudiantes de ingeniería industrial [8].

Otro artículo importante es “Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN” donde se automatizó un modelo predictivo de estudiantes pertenecientes al Instituto Politécnico Nacional (IPN). Las predicciones bordearon una precisión aproximada del 73% [9].

También, destaca el aporte “Predicción del rendimiento académico aplicando técnicas de minería de datos” donde el análisis se centra en estudiantes de Estadística General de la UNALM. En este, se aplican redes bayesianas, regresión logística, entre otras, estableciendo modelos de clasificación y evaluando mediante técnicas como la matriz de confusión para generar los resultados [10].

### 3. Materiales y métodos

La investigación está orientada a la aplicación de modelos para análisis predictivo del rendimiento académico estudiantil, en ese contexto, el fundamento metodológico tiene como base la minería de datos y en particular, se empleará la metodología de Descubrimiento de Conocimiento en Bases de Datos (KDD) [4], la cual, se compone de cinco etapas:

La primera etapa, consiste en la extracción de datos a través de fuentes como hojas de cálculos en Excel, además de una base de datos estudiantil realizada en Microsoft Access que posee información entre los años 2015 y 2019. Esta base de datos inicial contiene información elemental de los alumnos como: nombres, apellidos, dirección, entre otros. Por otro lado, en las hojas de cálculo se encuentran datos más específicos del estudiante, además, información con respecto a los padres. De esta manera, se logra construir el dataset con variables de interés que permitan obtener información valiosa luego del procesamiento.

La segunda etapa, tiene como objetivo la creación de un data warehouse con el propósito de disponer de un repositorio de datos único, que integre las variables disponibles desde las fuentes de datos identificadas. De esta manera, se obtendrá un conjunto de datos objetivo, fuente para el procesamiento. Para la creación de esta estructura se apoya en procesos ETL (Extracción, transformación y carga), mediante la herramienta Pentaho Data Integration.

La tercera etapa, se trata de la aplicación de diferentes modelos de minería de datos con la finalidad de realizar el análisis predictivo sobre los datos de rendimiento académico de los estudiantes. Como actividad previa necesaria, se realiza el análisis exploratorio de datos para

determinar qué variables tienen influencia en la problemática planteada. Para minimizar los errores de procesamiento, en función del tipo de dato de cada variable, es necesario normalizar sus valores para disminuir la distancia entre el valor real y el valor calculado.

El entorno de ejecución de Jupyter Notebook, es la herramienta de soporte para los experimentos. El lenguaje de programación aplicado es Python incluyendo librerías como: pandas, numpy, matplotlib seaborn y scikit-learn. Se emplearán las siguientes técnicas de procesamiento supervisadas de aprendizaje automático:

- a) Árboles de decisión.
- b) Redes neuronales.
- c) Máquinas de vector de soporte (SVM).

En la cuarta etapa, se analizan los resultados obtenidos para determinar cuál es el modelo que posee mejor funcionamiento. Los modelos entrenados corresponden a modelos de regresión, por lo tanto, los métodos de evaluación incluyen las siguientes métricas: error absoluto medio (MAE), error cuadrático medio (MSE) y raíz del error cuadrático medio (RMSE). Cada métrica representa la valoración de la eficiencia del modelo en función de los resultados reales y el resultado de la predicción. [5]. Es de interés identificar el modelo que posee menor error. Además, se evaluará el coeficiente de determinación para conocer el ajuste de la predicción. El proceso de entrenamiento será iterativo en busca de lograr mejores resultados.

La quinta etapa, se centra en cómo la información obtenida se transforma en conocimiento. De esta manera, con los resultados obtenidos, se transmite el conocimiento a los administradores de la escuela mediante el empleo de una capacitación, así, existirá conocimiento a disposición que servirá de base a los administradores para tomar decisiones que posibiliten mejoras en el rendimiento académico de los estudiantes.

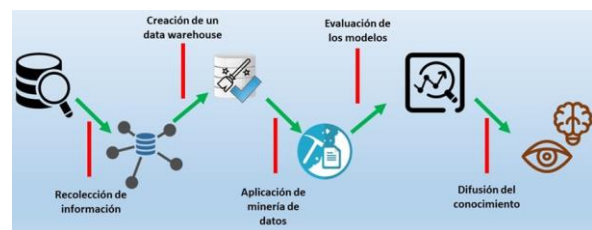


Figura 1. Metodología KDD.

#### 3.1. Pentaho Business Intelligence

Es un conjunto de herramientas para procesos ETL (extracción, transformación y carga de datos), dispone de capacidades de generación de informes y cuadros de mando [11]. Dentro de su estructura sobresalen las transformaciones que poseen múltiples formatos para las entradas y salidas en la integración de los datos [11].

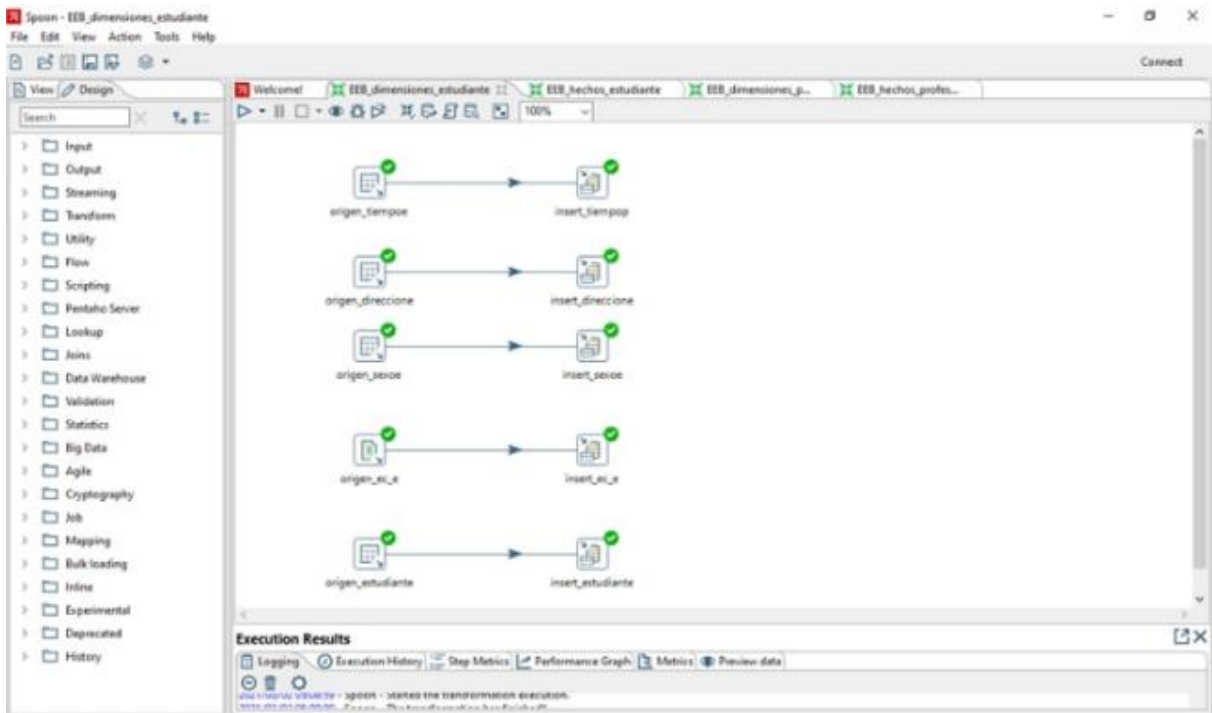


Figura 2. Proceso ETL.

Existen dos enfoques para crear un almacén de datos o “data warehouse”. El primero, enfoque Inmon, se caracteriza por la creación de un almacén de datos general para el posterior establecimiento de datamarts, que centralizan la información a un departamento en específico, y, el enfoque Kimball que describe un proceso contrario, partiendo de la creación de datamarts para luego generar un almacén de datos [12].

Para este proyecto, se escogió el segundo enfoque, Kimball, creándose dos datamarts, uno relacionado a los estudiantes y otro a los profesores. En la imagen anterior se puede apreciar una de las transformaciones realizadas para cargar el almacén de datos.

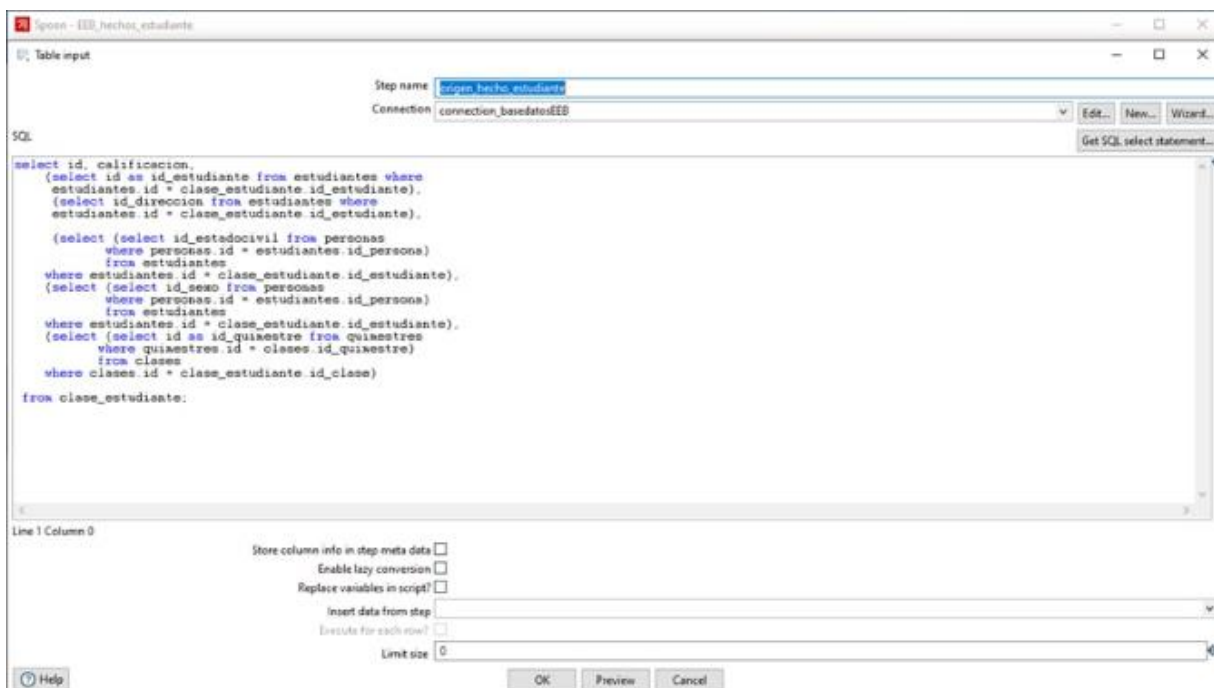


Figura 3. Proceso de tabla de hechos en el almacén de datos.

### 3.2. Jupyter Notebook

Es una aplicación web de código abierto que le permite crear y compartir documentos que contienen código en vivo, ecuaciones, visualizaciones y texto narrativo. Los usos incluyen: limpieza y transformación de datos, aprendizaje automático, visualización de datos, entre otros [13].

### 3.3. Python

Es un lenguaje de scripting potente, interpretado, de código abierto, de uso general y gratuito para aplicaciones web. Es un lenguaje de programación fácil pero poderoso que proporciona estructura y soporte para aplicaciones grandes [14].

### 3.4. Scikit-learn

Es una biblioteca de código abierto de populares algoritmos de aprendizaje automático que permite construir este tipo de sistemas [15]. Además, se compone de herramientas sencillas y eficientes para el análisis predictivo de datos, accesibles para todos y reutilizables en diversos contextos [16].

### 3.5. Matplotlib

Es un paquete de Python para trazado 2D que genera gráficos con calidad de producción. Admite el trazado interactivo y no interactivo, y puede guardar imágenes en varios formatos de salida (PNG, PS y otros). Puede utilizar varios conjuntos de herramientas de ventana (GTK +, wxWidgets, Qt, etc.) y proporciona una amplia variedad de tipos de gráficos (líneas, barras, gráficos circulares, histogramas y muchos más) [17].

### 3.6. Seaborn

Es una librería construida sobre matplotlib e integrada con pandas, permite realizar visualizaciones con un enfoque técnico y estético. Entre los gráficos que permite hacer destacan diagramas de caja, los cuales, sirven para determinar y analizar la presencia de valores atípicos dentro del conjunto de datos [18].

### 3.7. Inteligencia de Negocios

Los datos se producen tan rápido y en volúmenes extensos que es imposible analizarlos y usarlos de manera efectiva cuando se utilizan métodos manuales tradicionales como hojas de cálculo. Bajo este concepto, surge la Inteligencia de Negocios, que, reúne datos en forma utilizable para su análisis pertinente. La Inteligencia de Negocios apoya la toma de decisiones basada en hechos utilizando datos históricos en lugar de suposiciones carentes de objetividad [19].

### 3.8. Árboles de decisión

Es un algoritmo cuya finalidad es reconocer la existencia de relaciones en un determinado conjunto de datos por medio de procesos que imitan el funcionamiento del cerebro humano [20].

En este trabajo, se entrenó el modelo “árboles de decisión de regresión”. Así, se realizó la importación de “Decision Tree Regressor” para establecer el regresor. Como paso base se determinan los parámetros para la construcción del modelo de regresión, de los cuales dependerá el rendimiento del modelo. Como todos los modelos de aprendizaje supervisados, la posibilidad de un sobreajuste existe por lo que es necesario entrenar con diferentes configuraciones. Entre los parámetros establecidos para el modelo, están: la profundidad del árbol y el número mínimo de muestras necesarias para la división de cada nodo interno.

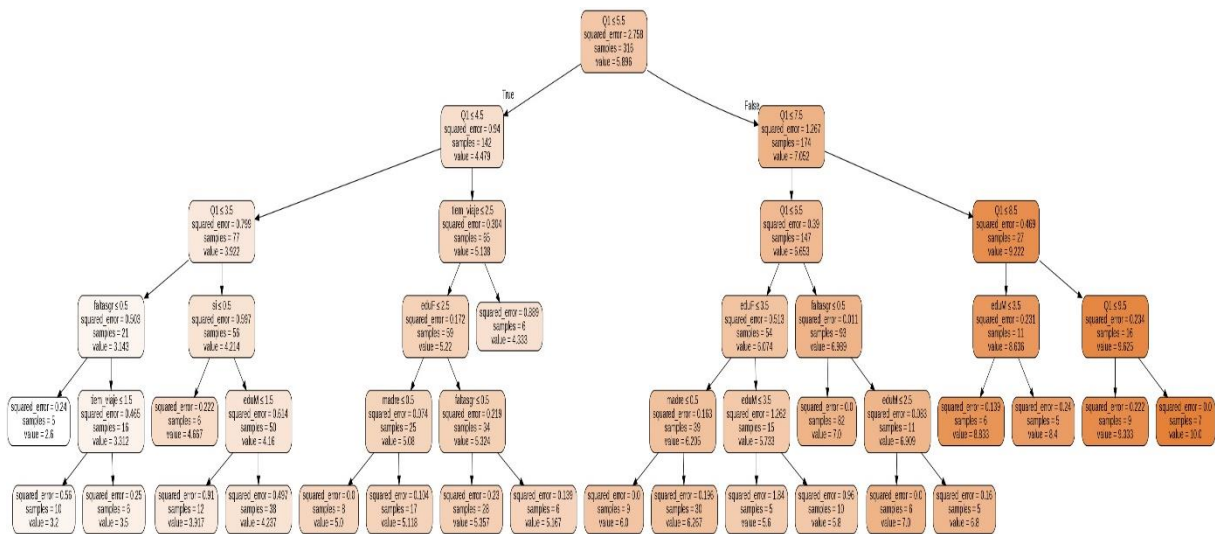


Figura 4. Estructura del árbol de decisión de regresión.



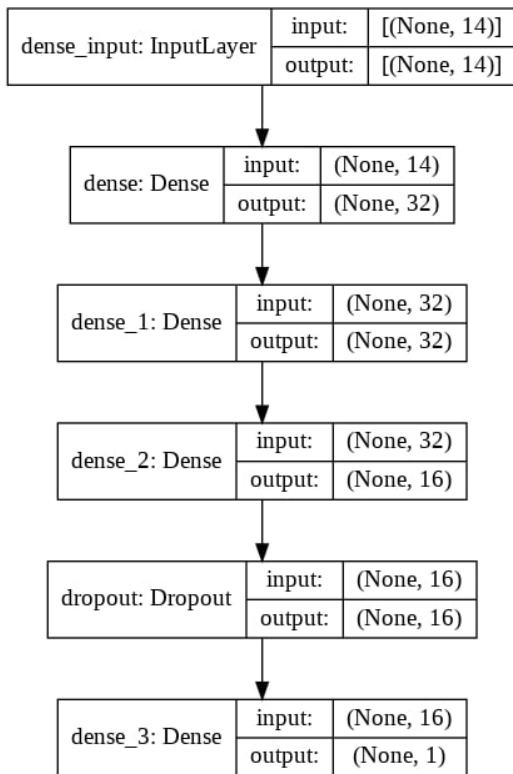
Los valores de las métricas de rendimiento fueron los siguientes:

**Tabla 1:** Métricas árbol de decisión.

Métrica	Valor
MAE	0.41
MSE	0.38
RMSE	0.62
R <sup>2</sup>	0.89

### 3.9. Redes neuronales

Es un algoritmo cuya finalidad es reconocer la existencia de relaciones en un determinado conjunto de datos por medio de procesos que imitan el funcionamiento del cerebro humano [20]. La ventaja de la red neuronal es que tiene el potencial de detectar todas las interacciones posibles entre las variables predictoras. La red neuronal también podría hacer una detección completa sin tener ninguna duda incluso en relaciones complejas no lineales entre variables dependientes e independientes.



**Figura 5.** Arquitectura de la red neuronal.

La figura 5 muestra la arquitectura de la red neuronal entrenada con 14 entradas. Su arquitectura consta de una capa de entrada, dos capas ocultas y una capa de

salida. En cuanto a función de activación se empleó la función Unidad Lineal Rectificada conocida como “ReLU”, mientras que el optimizador para la compilación del modelo fue “Adam” relacionado al momento lineal (momentum) y varianza de la tasa de aprendizaje.

Las métricas obtenidas fueron:

**Tabla 2:** Métricas redes neuronales.

Métrica	Valor
MAE	0.58
MSE	0.56
RMSE	0.75
R <sup>2</sup>	0.84

### 3.10. Máquina de vectores de soporte

Es un algoritmo que se lleva a cabo mediante la búsqueda de un hiperplano que separa entre un conjunto de objetos que tienen diferentes clases. Este hiperplano se elige maximizando el margen entre las dos clases para reducir el ruido y aumentar la precisión de los resultados [21].

Con el objetivo de emplear un modelo enfocado en la regresión, se empleó la variante de la Máquina de Vectores de Soporte (SVM), siendo esta, los Vectores de Soporte de Regresión (SVR). Entre los parámetros que se establecieron están: kernel, constante o parámetro de regularización C, valor gamma y un valor para épsilon.

Para este modelo se obtuvieron los siguientes resultados:

**Tabla 3:** Métricas Vectores de Soporte Regresión.

Métrica	Valor
MAE	0.62
MSE	0.78
RMSE	0.88
R <sup>2</sup>	0.78

## 4. Discusión y resultados

Predecir el rendimiento académico de los estudiantes, implica experimentar con diferentes modelos que sean más eficientes y tengan una mejor precisión. Los modelos citados en la sección estudios relacionados, utilizan sus propios datos y entrenan modelos que se ajustan sus condiciones; de manera similar, la presente investigación aplica los modelos DT, NN y SVR para

un dataset caso de estudio con el propósito de tener un mejor rendimiento y un error más bajo.

En función de las métricas válidas para modelos de regresión: MAE, MSE, RMSE [16], la tabla 4 muestra un resumen consolidado de la ejecución de cada modelo. Se añade el coeficiente de determinación como un elemento adicional a considerar.

**Tabla 4.** Resultados de los modelos.

Técnicas de minería de datos	Métricas de rendimiento			R <sup>2</sup>
	MAE	MSE	RMSE	
Árboles de decisión	0.41	0.38	0.62	0.89
Redes neuronales	0.58	0.56	0.75	0.84
SVR	0.62	0.78	0.88	0.78

Según los resultados obtenidos, el modelo que mejor rendimiento posee, debido a un menor valor del error en sus métricas, es el modelo de árboles de decisión de regresión, el cual posee un MAE de 0.41, un MSE de 0.38, y, un RMSE de 0.62. Además, al evaluar el coeficiente de determinación (R<sup>2</sup>), se obtuvo un valor de 0.89 puedo constatar que el modelo obtenido era óptimo.

El árbol de decisión de regresión permite determinar cuáles son los patrones determinantes para que los estudiantes de la institución posean un rendimiento adecuado.

- Si la calificación del estudiante durante el primer ciclo (Q1) es mayor a 9.5, el estudiante aprobará con un promedio mayor a 9.
- Si la calificación del estudiante durante el primer ciclo (Q1) es menor o igual a 8.5, el estudiante aprobará con un promedio mayor a 8. No obstante, este patrón se encuentra relacionado al nivel de educación de la madre, pues, si la ponderación equivale a un valor menor o igual a 3.5, el estudiante tendrá un promedio aproximado de 8.8, caso contrario, no superará el 8.4.
- Si la calificación del estudiante durante el primer ciclo (Q1) es menor o igual a 7.5 pero mayor a 6.5, el factor determinante serán las faltas graves que haya tenido durante su periodo escolar, pues, si estas son mínimos o nulas, podrá aprobar con un promedio de 7.0. No obstante, si estas faltas son mayores el patrón vuelve a estar influenciado por el nivel de educación de la madre, ya que, si este tiene una ponderación menor o igual a 2.5, el estudiante también podrá aprobar con un promedio de 7.0

## 5. Conclusiones

La generación del almacén de datos desde fuentes de datos heterogéneas permitió realizar satisfactoriamente el proceso de minería de datos. Las técnicas de minería empleados son válidas en el contexto planteado, sin embargo, los resultados asociados a cada técnica van a depender principalmente de los datos que se dispongan y del ajuste de parámetros de entrenamiento.

El método que mejor rendimiento tuvo corresponde a los árboles de decisión. Los errores obtenidos mediante sus métricas fueron de: un MAE de 0.41, un MSE de 0.38, un RMSE de 0.62, además, al evaluar el coeficiente de determinación, este valor fue de 0.89, lo que indicia que el modelo generado fue óptimo al realizar comparaciones entre los valores reales y valores de la predicción.

Los patrones obtenidos permitieron conocer que factores como la educación de la madre y las faltas graves, además de una alta nota del primer ciclo, son determinantes para que el estudiante pueda aprobar al final de su respectivo periodo académico.

## 6. Referencias bibliográficas

Bobadilla Sancho, J. (2020). Machine Learning y Deep Learning: Usando Python, Scikit y Keras (Primera ed.). Bogotá, Cundinamarca, Colombia: Ra-Ma.

Contreras, L., Fuentes, H., & Rodríguez, J. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *SciELO*, 13(5), 233-246.

Curto Díaz, J. (2010). Introducción al Business Intelligence (Primera ed.). Barcelona, España: UOC.

Fernández Fernández, J., & Flórez López, R. (2008). Las Redes Neuronales Artificiales: Fundamentos teóricos y aplicaciones prácticas (Primera ed.). La Coruña, España: Netbiblo.

Gamarra Gómez, F. (2019). MODELO BASADO EN MACHINE LEARNING PARA EL NEURORRENDIMIENTO ACADÉMICO DE ESTUDIANTES UNIVERSITARIOS. *Revista Ciencia y Tecnología para el Desarrollo-UJCM*, 5(9), 10-18. Obtenido de <https://revistas.ujcm.edu.pe/index.php/rctd/article/view/137/121>

Gironés Roig, J., Casas Roma, J., Minguillón Alfonso, J., & Caihuellas Quiles, R. (2017). Minería de datos: Modelos y algoritmos (Primera ed.). Barcelona, España: UOC.



- Gupta, R. (2002). *Making use of Python*. (B. Ryan, Ed.) New York, United States: Wiley Publishing. Recuperado el 17 de Diciembre de 2020
- Hernández Orallo, J., Ramírez Quintana, M., & Ferri Ramírez, C. (2005). *Introducción a la Minería de Datos*. (D. Fayerman Aragón, Ed.) Madrid, España: Pearson. Recuperado el 17 de Diciembre de 2020
- Jupyter. (15 de Diciembre de 2020). Jupyter. Recuperado el 17 de Diciembre de 2020, de <https://jupyter.org/>
- Moncecchi, G., & Garreta, R. (2013). *Learning scikit-learn: Machine Learning in Python*. Birmingham, Midlands Occidentales, Reino Unido: Packt Publishing. Recuperado el 11 de Febrero de 2021
- Nevpro. (23 de Septiembre de 2020). Pentaho Business Intelligence Tool. Recuperado el 17 de Diciembre de 2020, de <https://www.nevprobusinesssolutions.com/pentaho-business-intelligence/>
- Pardillo Vela, J., Trujillo Mondejar, J., & Mazón López, N. (2011). *Diseño y explotación de almacenes de datos: Conceptos básicos de modelado multidimensional* (Primera ed.). San Vicente, Alicante, España: ECU.
- Pojon, M. (Junio de 2017). *Using Machine Learning to Predict Student Performance*. Recuperado el 10 de Diciembre de 2020, de <https://trepo.tuni.fi/bitstream/handle/10024/101646/GRADU-1498472565.pdf?sequence=1>
- Predicción del rendimiento académico aplicando técnicas de minería de datos. (2017). *Anales Científicos*, 78(1), 26-33. Obtenido de <https://dialnet.unirioja.es/servlet/articulo?codigo=6171237>
- Pulido Romero, E., Escobar Domínguez, Ó., & Núñez Pérez, J. (2019). *Base de datos* (Primera ed.). Ciudad de México, México, México: Patria.
- Rico Páez, A., & Sánchez Guzmán, D. (2018). *Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN*. *SciELO*, 8(16).
- Rodríguez Parrilla, J. (2014). *Cómo hacer inteligente su negocio: Business Intelligence a su alcance* (Primera ed.). Ciudad de México, México, México: Patria.
- Scikit-learn. (28 de Junio de 2021). Scikit-learn. Recuperado el 19 de Julio de 2021, de <https://scikit-learn.org/stable/>
- Seaborn. (15 de Agosto de 2021). Seaborn. Recuperado el 17 de Agosto de 2021, de <https://seaborn.pydata.org/>
- Suganya, G. (2018). *Data Mining Techniques to Improve Predictions Accuracy of Students' Academic Performance: A Case Study with Xorro-Q*. Recuperado el 10 de Diciembre de 2020, de [https://mro.massey.ac.nz/bitstream/handle/10179/14655/02\\_whole.pdf?sequence=2&isAllowed=y](https://mro.massey.ac.nz/bitstream/handle/10179/14655/02_whole.pdf?sequence=2&isAllowed=y)
- Tosi, S. (2009). *Matplotlib for Python Developers*. Birmingham, Midlands Occidentales, Reino Unido: Packt Publishing. Recuperado el 11 de Febrero de 2021