



Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Maestría en Ciencias de la Computación

Técnicas de aprendizaje automático para predecir el desempeño de los estudiantes de nivel básico en pruebas estandarizadas

Tesis presentada para obtener el grado de Maestro en Ciencias de la Computación

Presenta:

Lic. Mario Sánchez González

Directores de tesis:

Dra. Meliza Contreras González

Dr. Roberto Contreras Juárez

Puebla, Puebla

Noviembre de 2025

Dedicatoria

Dedico este trabajo a mi padre, mi madre y mi hermana, por ser quienes han depositado su fe en mí con el único propósito de ayudarme a superar nuevos límites, me han guiado a lo largo de estos años en esta cruzada y han sido mi inspiración en la búsqueda del éxito para un mejor futuro.

Agradecimientos

Agradezco a mis asesores, la Dra. Meliza Contreras González y el Dr. Roberto Contreras Juárez, por su guía y enseñanzas durante estos años en la maestría y de su apoyo en la realización de este trabajo.

Agradecimiento especial ente a la SECIHTI por brindarme apoyo a través de la beca nacional para estudios de posgrado con número de apoyo 4012819, la cual me ayudó en mi formación a lo largo de este programa de maestría.

Índice general

Dedicatoria	1
Agradecimientos	2
1. Capítulo 1. Introducción	6
1.1 Planteamiento del problema	6
1.2 Objetivos.....	7
1.3 Antecedentes	7
1.4 Justificación	8
1.5 Distribución de la tesis	9
2. Capítulo 2. Marco Teórico	10
2.1 Aprendizaje automático	10
2.1.1 Terminología.....	11
2.1.2 Entrenamiento y prueba del modelo de aprendizaje.....	12
2.2 Aprendizaje supervisado.....	12
2.2.1 Regresión Lineal.....	12
2.2.2 Random Forest.....	13
2.2.3 K-NN	13
2.3 Aprendizaje no supervisado.....	14
2.3.1 K-Means	15
2.3.2 DBSCAN.....	15
2.4 Métricas de evaluación	15
2.4.1 Coeficiente de determinación R^2	15
2.4.2 Error cuadrático medio	16
2.4.3 Interpretación conjunta de R^2 y RMSE	17
2.4.4 Índice Kappa de Cohen.....	17
3. Capítulo 3. Estado del Arte.....	19
3.1 Aprendizaje supervisado.....	19
3.2 Aprendizaje no supervisado.....	23
4. Capítulo 4. Propuesta de Solución.	27

4.1 Recolección de datos	27
4.2 Preprocesamiento de datos	30
4.3 Selección de técnicas de minería de datos.....	30
4.4 Modelado y entrenamiento	31
4.4.1 Implementación de Regresión Lineal	32
4.4.2 Implementación de Random Forest.....	33
4.4.3 Implementación de K-Nearest Neighbors	33
4.4.4 Implementación de K-Means.....	34
4.4.5 Implementación de DBSCAN	35
4.5 Evaluación e interpretación del modelo	36
5. Capítulo 5. Resultados experimentales.....	37
5.1 Aprendizaje no supervisado.....	37
5.1.1 Resultados de K-Means	37
5.1.2 Resultados de DBSCAN.....	44
5.1.3 Consideraciones del aprendizaje no supervisado	48
5.2 Aprendizaje supervisado.....	48
5.2.1 Resultados no ideales.....	50
5.2.2 Resultados al 20% de prueba.....	51
5.2.3 Resultados al 40% de prueba.....	55
5.2.4 Resultados al 60% de prueba.....	58
5.2.5 Resultados al 80% de prueba.....	61
5.2.6 Resultados al 99% de prueba.....	65
5.3 Discusión de resultados	68
5.3.1 Interpretación de los resultados de regresión lineal.....	68
5.3.2 Interpretación de los resultados de Random Forest.....	69
5.3.3 Interpretación de los resultados de K-NN	69
5.3.4 Interpretación conjunta de los modelos desarrollados.....	69
6. Capítulo 6. Conclusiones.....	71
7. Capítulo 7. Trabajos futuros.....	72
Referencias	73

Capítulo 1. Introducción

Este capítulo está dedicado a la presentación de los enfoques que han dado forma a este trabajo de tesis, empezando por las cuestiones que motivaron el desarrollo de esta investigación, las cuales se explican en el planteamiento del problema junto a los antecedentes y justificación correspondientes a este trabajo. Además de exponer los objetivos, tanto general como específicos, que se desean completar a través de la implementación de la propuesta de solución mostrada en este documento.

1.1 Planteamiento del problema

Los exámenes de ingreso administrados por las instituciones educativas son uno de los principales desafíos que enfrentan los estudiantes que desean continuar su educación. Estos exámenes son pruebas estandarizadas que miden las habilidades y conocimientos de los estudiantes, lo que ayuda a seleccionar estudiantes que tienen más probabilidades de tener éxito en sus estudios futuros.

Las pruebas estandarizadas se utilizan para tratar de medir en los estudiantes lo que saben o lo que pueden hacer. Una prueba de matemáticas, por ejemplo, debe aportar información sobre lo que el sustentante sabe de determinado tema o la habilidad que posee para resolver problemas matemáticos. Para ello, se le hacen varias preguntas y si acierta o si falla, podremos saber un poco más sobre su capacidad y habilidad.

Los modelos de la Teoría de Respuesta al Ítem (TRI) postulan que la relación entre el rendimiento de un sustentante en determinado reactivo y la variable latente que subyace a dicho rendimiento, puede describirse por una función monótona creciente. Esto es, para cada nivel de habilidad, generalmente denominado por la letra griega theta, θ , existe una probabilidad, $P(X_i = 1|\theta)$, de que un examinado con dicha habilidad conteste correctamente el reactivo X_i . Se esperaría que esta probabilidad crezca mientras la habilidad se incrementa.

La función matemática que describe la relación entre los diferentes niveles de habilidad y la probabilidad de contestar correctamente determinado reactivo es conocida como Curva Característica del Ítem (CCI) y constituye la base de la TRI.

El propósito de este trabajo de tesis es utilizar técnicas de minería de datos para analizar las calificaciones promedio de Español, Matemáticas, Ciencias Naturales y Ciencias Sociales de estudiantes egresados de secundaria con el fin de encontrar patrones que ayuden a predecir su desempeño en el examen de ingreso al bachillerato. El desempeño se mide con

base en el nivel de habilidad/conocimiento demostrado en una prueba estandarizada representado por θ .

La propuesta es realizar técnicas de predicción, clasificación y agrupamiento incluidas en la minería de datos sobre las calificaciones promedio de estudiantes procedentes de diferentes escuelas que solicitaron su ingreso al nivel medio superior del estado de Puebla.

1.2 Objetivos

Objetivo General:

- Desarrollar un modelo que estime el valor de la habilidad académica real de los estudiantes en una prueba estandarizada para su admisión al nivel medio superior, basándose en su desempeño en diversas áreas de conocimiento.

Objetivos Específicos:

- Analizar datos relacionados a las características necesarias para cubrir la participación de estos en un examen de admisión.
- Evaluar los datos obtenidos mediante una metodología de minería de datos.
- Implementar diferentes algoritmos de aprendizaje automático supervisado y no supervisado.
- Comparar los algoritmos implementados respecto a sus resultados considerando los valores de R^2 , MSE y $Kappa$ sobre los modelos utilizados.

1.3 Antecedentes

La demanda de los jóvenes en la actualidad con relación a la búsqueda de una oportunidad para desarrollarse en el ámbito profesional es uno de los aspectos más interesantes en la actualidad; es interesante como la cantidad de jóvenes aspirantes a matriculas educativas en el nivel medio-superior en los últimos años ha incrementado considerablemente.

No obstante, con el paso del tiempo es posible apreciar como el número de estudiantes aceptados decrece considerablemente al final de una generación escolar, lo que genera muchas interrogantes relacionadas a las posibles causas que conllevan a este comportamiento en la actualidad.

La minería de datos es el descubrimiento del conocimiento en las bases de datos y básicamente consiste en la extracción de información que se oculta de manera implícita en

los datos [21]. La minería de datos puede considerarse entonces como una metodología de análisis de datos utilizada para identificar patrones, inicialmente desconocidos, que se ocultan en los conjuntos de datos.

La minería de datos se ha utilizado en muchas áreas, especialmente en el área comercial, pero en los últimos años ha cobrado un gran interés en el ámbito educativo. La minería de datos aplicada a la educación se conoce como minería de datos educativos (“EDM”, por sus siglas en inglés) y es la encargada de extraer información útil de las bases de datos de las instituciones educativas aplicando herramientas y técnicas de minería de datos para una mejor comprensión o con el fin de obtener mayores conocimientos [22].

Como es sabido, la educación es uno de los factores más importantes para el desarrollo de un país, esto obliga a que las instituciones educativas tengan como objetivo primordial ofrecer programas educativos de calidad y una forma de cumplir con esta calidad es a través de la evaluación del rendimiento o desempeño académico de los estudiantes. Este proceso de evaluación es bastante complejo ya que no puede ser limitado a una simple calificación, pues existen muchos factores que pueden afectar los buenos o malos resultados y conocerlos ayuda a los responsables de la enseñanza-aprendizaje a planificar y personalizar sus programas educativos en función de la información recibida.

La “EDM”, es uno de los enfoques de la minería de datos que puede proporcionar una ayuda eficaz para revelar las complejas relaciones que hay detrás de las calificaciones y el diagrama de la Figura 1.1 explica las diferentes etapas de la metodología para la extracción de información sobre datos educativos para la toma de decisiones académicas [23].

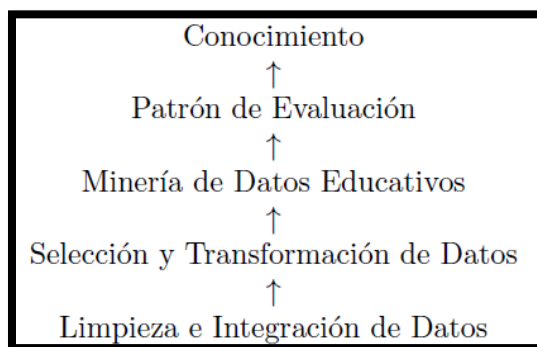


Figura 1.1. Etapas de la EDM

En resumen, la **EDM** ayuda a extraer la información que previamente era desconocida y que resulta útil para algún proceso.

1.4 Justificación

La investigación realizada en este trabajo de tesis está enfocada en el estudio de diferentes modelos de minería de datos y aprendizaje automático aplicados a la predicción

del desempeño de estudiantes que buscan ingresar a una educación de nivel medio superior, así como observar el impacto que pueden tener aspectos socioeconómicos en el resultado obtenido por dichos estudiantes, los cuales van más allá del ambiente académico.

1.5 Distribución de la tesis

La distribución de esta tesis será de la siguiente forma:

- **Marco teórico:** en esta sección se exponen los conceptos y teoría necesarios para comprender los fundamentos de esta tesis. Aquí se presentan los temas relacionados a aprendizaje automático, métricas de evaluación, minería de datos, entre otros.
- **Estado del arte:** esta sección está enfocada a la exposición de trabajos relacionados, artículos de investigación que han servido de inspiración para el desarrollo de la metodología de solución implementada en este trabajo de tesis.
- **Propuesta de solución:** en esta sección se describe la metodología de minería de datos con sus respectivos pasos y como fueron implementados para cumplir con los objetivos planteados en esta tesis.
- **Resultados:** esta sección muestra los resultados obtenidos tras la implementación de los pasos correspondientes a la metodología de solución explicada en la sección anterior.
- **Conclusiones:** la última sección de este trabajo plasma las conclusiones de este trabajo de tesis, así como el trabajo futuro que puede derivar de esta investigación.

Capítulo 2. Marco Teórico

Los conocimientos empleados en la realización de esta tesis se explican en este capítulo. Dentro de estas secciones se muestran los conceptos relacionados al aprendizaje automático que se implementaron en la propuesta de solución, elementos que forman parte del aprendizaje supervisado como *regresión lineal*, “*Random Forest*” y “*K-NN*”; los cuales nos servirán para la predicción de datos cuyos resultados son evaluados a través de métricas de evaluación como R^2 , MSE y $Kappa$ para determinar el desempeño de las técnicas implementadas. Por parte del aprendizaje no supervisado se exponen las explicaciones de los métodos de “*K-Means*” y “*DBSCAN*”, que serán empleados para obtener una evaluación general del comportamiento de las variables a trabajar a través de clústeres y la visualización gráfica.

2.1 Aprendizaje automático

El aprendizaje automático es definido como la ciencia (y arte) de programar a las computadoras para aprender a partir de los datos, un campo de estudio que da a las computadoras la capacidad de aprender sin ser programadas explícitamente [20].

Como se explica en [1], los algoritmos de aprendizaje han demostrado ser útiles en diversos dominios de aplicación, como en la minería de datos en grandes bases de datos que contienen regularidades implícitas, las cuales pueden ser descubiertas en forma automatizada, en dominios que son poco entendidos y donde los humanos no poseen el conocimiento necesario para desarrollar algoritmos efectivos, y en dominios donde los programas se deben adaptar en forma dinámica para responder a condiciones cambiantes en el ambiente. No todos los problemas pueden solucionarse de la misma forma, por lo que se debe considerar el tipo de técnicas de aprendizaje a utilizar, pueden ser supervisadas o no supervisadas según corresponda a la estructura del conjunto de datos y del tipo de problemas [17], algo que puede identificar en la Figura 2.1.

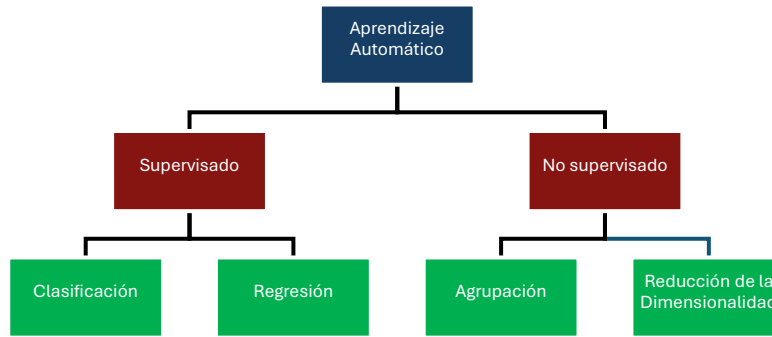


Figura 2.1. Diagrama del aprendizaje automático (elaboración propia)

2.1.1 Terminología

Antes de presentar los conceptos relacionados a los métodos de aprendizaje automático que se estudiarán en este trabajo es importante exponer la definición de conjunto de datos, concepto que será pronunciado a lo largo de este documento. Los datos se definen como representaciones numéricas de fenómenos reales observables, ejemplos de dichos fenómenos se encuentran en el número de miembros de una familia, el precio de un producto en el mercado, la velocidad de descarga de internet, entre otros.

Sin embargo, al tratar con grandes cantidades de datos es necesario el uso de estructuras matemáticas (como vectores, matrices, etc.) para el almacenamiento de estos y un manejo simple [38]. Cuando se relaciona con el aprendizaje automático, la estructura utilizada para almacenar los datos se denomina conjunto de datos, la cual se denota con la letra F [24]. Dicho conjunto F maneja una arquitectura conformada por m filas (registros) y n columnas (atributos o características).

Tabla 2.1. Ejemplo de conjunto de datos

X_1	X_2	...	X_n
$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$
$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$
$\vdots$	$\vdots$		$\vdots$
$x_{m,1}$	$x_{m,2}$	...	$x_{m,n}$

Como se puede observar en la Tabla 2.1, un conjunto de datos se conforma por dos partes, la primera es el vector de atributos x' , representado por todas las columnas del conjunto de datos y que serán referenciados como atributos a lo largo de este documento, y la matriz de datos X en el cual se almacenan los datos dentro de sus filas y columnas. Si bien es cierto que un conjunto de datos puede incorporar muchos atributos, es importante

considerar las diferentes naturalezas o tipos que estos atributos puedan presentar, en caso de tipos numéricos sean estos decimales o enteros, atributos de tipo texto o cadena, tipo binario, nominales y demás.

2.1.2 Entrenamiento y prueba del modelo de aprendizaje

El propósito del aprendizaje automático radica en la creación de modelos que permitan la resolución de problemas de gran complejidad, para esto se sustrae una cantidad $m * n$ de datos históricos provenientes del conjunto F para generar dichos modelos de aprendizaje, lo cual se conoce comúnmente como proceso de entrenamiento de un algoritmo de aprendizaje.

Para poder entrenar un modelo de aprendizaje es necesario partir el conjunto de datos en dos partes esenciales, la primera se utiliza para el entrenamiento del modelo de aprendizaje, la cual se conoce como conjunto de entrenamiento. Y el segundo es empleado para probar el modelo desarrollado, llamado conjunto de prueba. Esta partición dependerá de cuantos datos se deseen emplear, ya sean 80% y 20% o 60% y 40%, entre muchos escenarios de entrenamiento.

2.2 Aprendizaje supervisado

La meta del *aprendizaje supervisado* según [1] es aprender el mapeo de las respuestas correctas para los datos de entrada que le son proporcionados; para esto, se utiliza un conjunto de datos de entrenamiento constituidos por pares que consisten en patrones de entrada y salida correcta [16]. De esta forma, el sistema aprende el mapeo de la salida correcta para cada patrón de entrada que se le presenta. Algunos ejemplos de técnicas de aprendizaje de máquina supervisado son:

- Árboles de decisión
- Máxima entropía
- “*Naïve Bayes*”
- Máquinas de soporte vectorial
- Redes neuronales

2.2.1 Regresión Lineal

En [2], la *regresión lineal* es explicada como una técnica de análisis de datos que predice el valor de datos desconocidos mediante el uso de otro valor de datos relacionado y

conocido. Modela matemáticamente la variable desconocida o dependiente y la variable conocida o independiente como una ecuación lineal, en la Figura 2.2 se observa un ejemplo del comportamiento gráfico de esta técnica aplicado a un conjunto de datos que representan las horas de publicidad en televisión y el número de ventas realizadas.

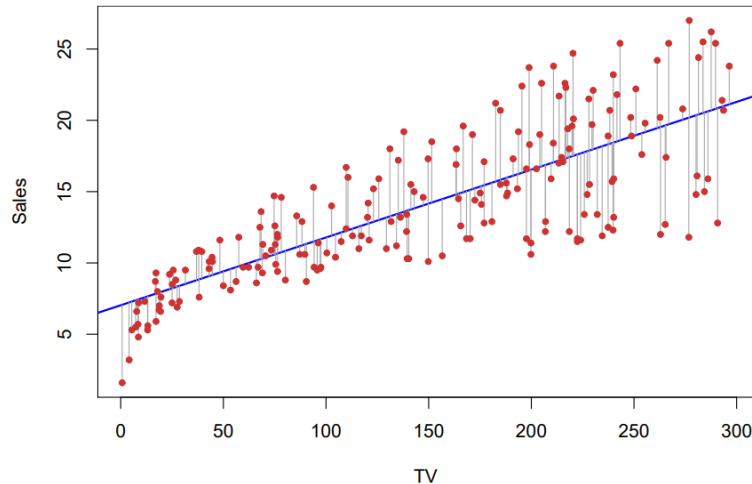


Figura 2.2. Ejemplo de regresión lineal aplicado a publicidad y ventas en TV [17]

2.2.2 Random Forest

Según [3], los “*Random Forest*” (bosques aleatorios por su traducción) es un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. Este método combina múltiples árboles de decisión para mejorar la precisión y evitar el sobreajuste.

Utiliza el método de “*bagging*” (ensamble) para crear cada árbol de decisión en el bosque. Esto implica tomar múltiples muestras aleatorias con reemplazo del conjunto de datos original y entrenar un árbol de decisión en cada una de estas muestras. Para la regresión, el resultado final es el promedio de las predicciones de todos los árboles. Mientras que, para la clasificación, el resultado final es la clase más votada por los árboles.

2.2.3 K-NN

Según lo expuesto en [3], “*K-NN*” (K-Vecinos más cercanos) es un algoritmo de aprendizaje supervisado utilizado tanto para clasificación como para regresión. Es uno de los algoritmos más simples y efectivos, basado en la idea de que objetos similares estarán cerca en el espacio de características.

“*K-NN*” clasifica un punto de datos en función de los K puntos de datos más cercanos en el espacio de características. Para regresión, el valor predicho es el promedio de los valores de estos K vecinos más cercanos.

2.3 Aprendizaje no supervisado

El *aprendizaje no supervisado* se usa para extraer información de ejemplos que no poseen una clase definida, es decir, no se necesita un valor objetivo que se relacione con las demás características, ahora se relacionarán una con otra para generar grupos [19]. Estos grupos están formados por datos puntuales que comparten características similares a los centroides de los grupos, estos definen a los grupos y suelen estar alejados uno del otro, lo cual nos permite definir el problema de agrupamiento, también conocido como clustering [18].

Como se explica en [1], en las técnicas de aprendizaje no supervisado no se requiere la intervención de humanos para elaborar un conjunto de datos previamente categorizados para ser presentado al algoritmo de aprendizaje. La meta del aprendizaje no supervisado es encontrar patrones interesantes considerando la distribución y composición de los datos que le son presentados, ejemplos de las técnicas de aprendizaje no supervisado son las técnicas de agrupamiento, las cuales pueden verse ejemplificadas en la Figura 2.3.

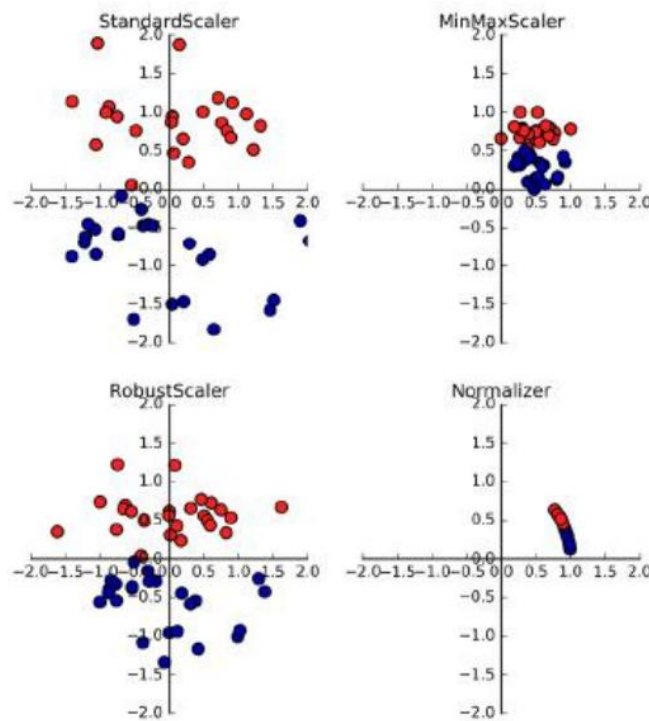


Figura 2.3. Ejemplos de agrupamiento de datos [16]

2.3.1 K-Means

El método “*K-Means*” (*K-Medias*), según lo explicado por [2], es un algoritmo de clustering no supervisado que agrupa un conjunto de datos en K clústeres o grupos. El objetivo es minimizar la variación dentro de cada clúster y maximizar la variación entre clústeres.

“*K-Means*” selecciona K centroides iniciales, que pueden ser puntos seleccionados aleatoriamente o utilizando algún método como “*K-Means++*”. Cada punto de datos se asigna al clúster cuyo centroide esté más cercano, según una medida de distancia, típicamente la distancia euclidiana. Luego los centroides de los clústeres se recalculan como el promedio de todos los puntos asignados a cada clúster. Esto se repite hasta que los centroides ya no cambian significativamente o se alcanza un número máximo de iteraciones.

2.3.2 DBSCAN

En lo explicado por [4], “*DBSCAN*” (Agrupamiento Espacial Basado en la Densidad de Aplicaciones con Ruido) es un algoritmo de clustering no supervisado que agrupa puntos de datos que están densamente empaquetados juntos, marcando como ruido aquellos puntos que están en áreas de baja densidad.

“*DBSCAN*” agrupa puntos en clústeres en función de la densidad de los puntos, los clústeres se forman donde hay una alta densidad de puntos. Los parámetros principales del algoritmo son ϵ , que se define como el radio de vecindad de un punto, y *MinPts* significando el número mínimo de puntos necesarios para formar un clúster.

2.4 Métricas de evaluación

El rendimiento de los modelos de aprendizaje supervisado es evaluado con respecto a sus resultados generados con base a la predicción de los datos, en esta sección se expondrán las métricas empleadas para la valoración de los modelos desarrollados para este trabajo.

2.4.1 Coeficiente de determinación R^2

El *coeficiente de determinación R^2 (R-squared)* mide la proporción de la varianza de la variable dependiente (por ejemplo, calificaciones) que puede explicarse por las variables independientes (como asistencia, hábitos de estudio, etc.) y su valor oscila entre 0 y 1.

- Si $R^2 = 1$, el modelo explica perfectamente la variabilidad.

- Si $R^2 = 0$: el modelo no explica nada.

La interpretación en modelos educativos es muy importante y

- un **R^2 alto** (por ejemplo, ≥ 0.85) indica que los factores incluidos en el modelo explican bien el rendimiento académico.

Más específicamente, si por ejemplo que $R^2 = 0.85$, esto significa que las variables predictoras explican alrededor del 85% de la varianza de la variable de respuesta.

- un **R^2 bajo** sugiere que hay variables importantes no consideradas (como factores emocionales, familiares o institucionales).

En el contexto educativo, **R^2 también ayuda a justificar la inclusión de variables socioeconómicas o de comportamiento**, mostrando su impacto en el desempeño.

Para calcular R^2 se emplea la expresión matemática

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.1)$$

donde x e y son los valores pronosticados, en este caso, $R^2 = r_{xy}^2$.

2.4.2 Error cuadrático medio

El **error cuadrático medio** o **MSE (Mean Squared Error)** es el promedio de los cuadrados de las diferencias entre los valores reales y los predichos. Siempre es un número ≥ 0 y cuanto más pequeño sea el **MSE**, mejor es el modelo.

Consideremos los valores de pronóstico

$$\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$$

así como los valores reales

$$y = \{y_1, y_2, \dots, y_n\}$$

entonces, el **MSE**, se puede calcular con la expresión

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.2)$$

Su rango es $[0, +\infty)$ y cuando el valor predicho es completamente consistente con el valor real, es igual a 0, lo cual corresponde a un modelo perfecto. Cuanto mayor es el error, mayor es su valor.

En modelos educativos, un ***MSE* bajo** indica que las predicciones del modelo están cerca de las calificaciones reales. También es útil para **comparar modelos**, por ejemplo, si el modelo *A* tiene menor *MSE* que el modelo *B*, entonces *A* es más preciso.

En educación, puede ayudar a **identificar estudiantes en riesgo** si el modelo predice consistentemente por debajo del rendimiento real.

Finalmente, el *RMSE* (“Root Mean Square Error”) es en realidad la raíz cuadrada del *MSE* es más intuitivo en orden de magnitud. Por ejemplo, $RMSE = 10$, se puede considerar que el efecto de regresión es 10 diferente del valor real en promedio.

2.4.3 Interpretación conjunta de R^2 y *RMSE*

Para comprender mejor las dos métricas en discusión, consideremos los siguientes casos.

- *RMSE* bajo, R^2 alto. Este es el mejor caso ya que las predicciones son precisas (*RMSE* bajo) y el predictor determina principalmente el valor observado (R^2 alto). Un *RMSE* bajo significa que los residuos están ajustados alrededor de 0, en relación con la escala de la variable de respuesta.
- *RMSE* bajo, R^2 bajo. Aquí es posible generar buenas predicciones (*RMSE* bajo), pero no gracias al predictor. En cambio, los valores observados se encuentran en su mayoría dentro de un rango restringido y, por lo tanto, son fáciles de adivinar incluso sin un predictor.
- *RMSE* alto, R^2 alto. En este caso, el modelo hace predicciones deficientes (*RMSE* alto), pero se equivocan sistemáticamente al tener un sesgo más o menos constante (R^2 alto). Entonces, aunque las predicciones son malas, todavía hay algo de esperanza ya que el predictor determina fuertemente el valor observado, solo que el modelo no capitaliza correctamente.
- *RMSE* alto, R^2 bajo. Finalmente, tenemos el peor de los casos, el modelo hace predicciones malas (*RMSE* alto), y el predictor nos da poca o ninguna información sobre las observaciones reales (R^2 bajo).

2.4.4 Índice Kappa de Cohen

El *índice Kappa de Cohen* mide el grado de acuerdo entre dos clasificadores (por ejemplo, predicción vs realidad), ajustado por el acuerdo esperado por azar, su valor va de **-1 a 1** y

- si ***Kappa* = 1**: acuerdo perfecto.

- sí ***Kappa*** = 0: acuerdo igual al azar.
- sí ***Kappa*** < 0: peor que el azar.

En los modelos educativos se usa cuando el rendimiento académico se clasifica en categorías (por ejemplo, "alto", "medio", "bajo").

- Un ***Kappa*** alto indica que el modelo clasifica correctamente a los estudiantes en sus niveles de rendimiento.
- Es especialmente útil en **sistemas de alerta temprana**, donde se busca identificar estudiantes que podrían reprobado o abandonar.

Capítulo 3. Estado del Arte

En este capítulo se abarcan los artículos e investigaciones relacionados al tema investigado en esta tesis. Los trabajos recopilados en estas secciones se seleccionaron con respecto a las tendencias que comparten una similitud con este trabajo, priorizando el ámbito de la investigación sobre la educación a través de los métodos de aprendizaje automático, desde la implementación de diferentes modelos de aprendizaje supervisado para la evaluación de aptitudes académicas, así como el monitoreo de diferentes programas educativos o técnicas de enseñanza a través del “*clustering*” con el uso de modelos de aprendizaje no supervisado.

3.1 Aprendizaje supervisado

En el campo de la educación se han implementado en diversas ocasiones las técnicas y conocimientos orientados a la minería de datos, tal es el caso de [5], donde se enfocan estas técnicas para el diseño de un sistema de aprendizaje asistido para la enseñanza del inglés, utilizando técnicas avanzadas de inteligencia artificial (aprendizaje profundo o “*deep learning*”) y análisis de datos. Este trabajo se enfoca en abstraer características para evaluar y analizar las habilidades de los estudiantes en áreas como vocabulario, pronunciación y gramática. Además, utiliza análisis de datos para monitorear y registrar la información de aprendizaje, lo que ayuda a optimizar las estrategias de enseñanza.

El desempeño de las instituciones universitarias es otro de los aspectos investigados a través del uso del análisis de datos, como se demuestra por parte de [6], estas técnicas son usadas para abordar el desarrollo de un modelo para evaluar el nivel educativo de universidades, utilizando algoritmos avanzados de inteligencia artificial (aprendizaje profundo). Para esto, se introduce un modelo denominado “*AGA-BP*” (que parece ser una combinación de algoritmos genéticos y redes neuronales “*backpropagation*”). El concepto de “nivel educativo” se estructura en torno a indicadores que consideran tanto los recursos invertidos como los resultados obtenidos. También se diseñan cuestionarios para recopilar datos que alimenten el modelo de evaluación.

La enseñanza del inglés es uno de los pilares fundamentales en la creación de oportunidades profesionales en la actualidad, si bien [5] se enfoca en el sistema de aprendizaje de este, también se puede aprovechar para analizar grandes volúmenes de datos relacionados con la educación del inglés, proporcionando oportunidades para mejorar la enseñanza y enfrentando los desafíos que surgen con la adopción de estas tecnologías avanzadas, algo que se retoma y se expone por parte de [7]. También aborda los problemas y

objetivos de esta minería de datos educativa y propone sugerencias para mejorar la aplicación de la tecnología de aprendizaje profundo en este ámbito.

Los métodos de enseñanza durante la cuarentena iniciada en el año 2020 también son objeto de investigación y aplicación para la minería de datos, como expone [8], estos conocimientos se pueden emplear para analizar los problemas prácticos del uso de los “*MOOC*” (cursos en línea masivos y abiertos) en la educación tras su popularización, refiriéndose a esto como la “*era post-MOOC*”. Como solución, presenta un modelo de aprendizaje profundo basado en “*SPOC*” para aplicarlo en la educación de la salud psicológica de estudiantes universitarios. En este modelo, los estudiantes acceden a contenidos antes de clase mediante videos en la plataforma “*MOOC*”, luego interactúan en discusiones y actividades prácticas en el aula con la orientación del profesor. La evaluación del curso combina el rendimiento en línea y presencial. Este modelo busca optimizar la enseñanza de salud mental al combinar los amplios recursos de los “*MOOCs*” con la interacción y oportunidad del aula tradicional.

El área de los deportes en un nivel universitario también pueden ser un campo de investigación interesante, como se expone en [9], la minería de datos se puede emplear para analizar datos relacionados con el deporte en instituciones educativas, como colegios y universidades. Este trabajo explica cómo la tecnología de minería de datos, que es un método avanzado de análisis de datos, se puede utilizar para extraer información útil de grandes volúmenes de datos. Se describe la definición y los principios de la tecnología de minería de datos, así como su proceso de aplicación en el análisis de datos relacionados con el deporte en el ámbito universitario. Además, se detallan ejemplos específicos de cómo se utiliza esta tecnología para mejorar la gestión y desarrollo de la información deportiva en estas instituciones.

La aplicación de técnicas de minería de datos ya ha sido implementada en algunos casos de estudio, como se expresa en [10], donde este trabajo aborda la importancia de predecir el rendimiento académico de los estudiantes en instituciones de educación superior. Para ello, se utilizan técnicas de minería de datos (TMD) como regresión logística, árboles de decisión, redes bayesianas y redes neuronales. El objetivo es identificar los factores más influyentes en el aprendizaje de los estudiantes y ayudar a los docentes a mejorar el proceso de enseñanza mediante intervenciones pedagógicas más eficientes.

En otro trabajo [11] se aborda la predicción del rendimiento escolar de los estudiantes basándose en tres factores: sus aptitudes, el autoconcepto académico (cómo se perciben a sí mismos en el ámbito académico) y las atribuciones causales (las razones que atribuyen a sus éxitos o fracasos). Se realizaron análisis correlacionales y de regresión jerárquica para estudiar cómo estas variables afectan el rendimiento en las áreas de Lengua y Matemáticas. Los resultados indican que, aunque las aptitudes son el factor con mayor poder predictivo, tanto el autoconcepto académico como las atribuciones causales también contribuyen de

manera significativa a la explicación del rendimiento en ambas áreas. Los hallazgos tienen implicaciones teóricas y prácticas para mejorar el proceso de enseñanza y aprendizaje.

Un ejemplo más del uso de la minería de datos en la educación está en [12], un trabajo que se centra en el uso de técnicas de análisis de datos, específicamente la minería de reglas de asociación y clasificación, para analizar datos educativos. Con el objetivo de ayudar a mejorar la planificación administrativa y los servicios estudiantiles. Se aplicaron dos técnicas principales: la minería de reglas de asociación y los árboles de decisión. La primera técnica se utilizó para analizar los datos de admisión de estudiantes con el fin de descubrir patrones útiles que apoyen la planificación del proceso de admisión. La segunda técnica se aplicó a los datos de calificaciones de cursos y empleos de los estudiantes graduados, con el objetivo de predecir las posibilidades de empleo tras la graduación.

El uso de técnicas de minería de datos eficientes para ayudar a mejorar el rendimiento académico de los estudiantes es observado en el trabajo realizado por [13], donde se aborda el desafío de predecir el rendimiento académico de los estudiantes debido a la gran cantidad de datos almacenados en las bases de datos de las instituciones educativas. Las técnicas de minería de datos se han implementado en varias instituciones para predecir el rendimiento académico de los estudiantes. El propósito de predecir dicho rendimiento es ayudar a los profesores e instituciones a desarrollar estrategias para mejorar el aprendizaje de los estudiantes con dificultades, y también definir estrategias específicas para los estudiantes con buen desempeño, para que puedan rendir aún mejor.

El trabajo realizado por [14] es un ejemplo más del enfoque de predecir y analizar el rendimiento académico de los estudiantes utilizando técnicas de minería de datos. Esta investigación propone un marco para predecir el rendimiento académico de estudiantes de primer año de licenciatura en un curso de informática. Para esto, se aplican tres métodos de clasificación: Árbol de Decisión, “*Naïve Bayes*” y Perceptrón Multicapa, utilizando la herramienta de minería de datos WEKA. El objetivo es determinar cuál de estos modelos ofrece la mejor predicción del rendimiento académico de los estudiantes. Se realizan experimentos para comparar la precisión de los modelos y el conocimiento extraído se utiliza para identificar el nivel de éxito de los estudiantes en su primer semestre.

Un ejemplo de la predicción del desempeño académico es observado en [15], donde se analiza cómo los datos generados por las instituciones de educación superior pueden revelar aspectos interesantes sobre el comportamiento de aprendizaje de los estudiantes. Este estudio compara el desempeño de varias técnicas de clasificación aplicadas a datos de comportamiento estudiantil, con el objetivo de encontrar un modelo cualitativo que prediga el rendimiento académico. También se explora el reequilibrio de datos para comprobar si esto mejora los modelos de clasificación.

La implementación de árboles de clasificación para medir el desempeño académico puede observarse en [25], donde se implementan herramientas como el árbol de decisión J48,

“Naïve Bayes”, “Random Forest” y Perceptrón Multicapa para la evaluación de cinco áreas académicas con el propósito de predecir el desempeño que tendrán en pruebas para su ingreso al bachillerato, todo esto limitando los parámetros de investigación únicamente a datos académicos, excluyendo todo tipo de datos que no se relacionen a estos campos.

La Tabla 3.1 muestra un abstracto de todos los trabajos mencionados anteriormente, listando el año de publicación, objetivo de la investigación y las técnicas o conocimientos implementados.

Tabla 3.1. Trabajos relacionados con el aprendizaje supervisado

	Año	Objetivo	Algoritmos
[5]	2023	Desarrollo de un sistema de aprendizaje asistido del inglés que personalice la enseñanza y aumente el tiempo de estudio usando deep learning y análisis de datos	▪ Deep Learning ▪ Técnicas de análisis de datos
[6]	2023	Proposición de un modelo de evaluación del nivel educativo universitario usando aprendizaje profundo (AGA-BP)	▪ Modelo AGA-BP (Algoritmos Genéticos + Backpropagation)
[7]	2020	Exploración de la relevancia y aplicación del aprendizaje profundo en la minería de datos educativos del inglés	▪ Aprendizaje profundo
[8]	2019	Mejora de la enseñanza de salud mental en estudiantes universitarios usando un modelo basado en SPOC y deep learning	▪ Modelo SPOC con aprendizaje profundo
[9]	2021	Aplicación de la minería de datos para analizar información deportiva en instituciones educativas	▪ Técnicas generales de minería de datos
[10]	2017	Predicción el rendimiento académico en base a datos académicos en un curso específico	▪ Regresión logística ▪ Árboles de decisión ▪ Redes bayesianas ▪ Redes neuronales
[11]	2008	Determinar si variables motivacionales explican varianza adicional en el rendimiento escolar, además de las aptitudes	▪ Regresión jerárquica
[12]	2019	Aplicación de minería de reglas de asociación y	▪ Reglas de asociación ▪ Árboles de decisión

		clasificación para mejorar la planificación de admisiones y predecir empleos	
[13]	2021	Estudio de técnicas de minería de datos para predecir el rendimiento académico y definir estrategias pedagógicas	▪ Técnicas de minería de datos
[14]	2018	Predicción del rendimiento de estudiantes de informática de primer año y comparar técnicas de clasificación	▪ Árbol de Decisión ▪ Naïve Bayes ▪ Perceptrón Multicapa (MLP) usando WEKA
[15]	2019	Comparación de técnicas de clasificación sobre datos de comportamiento estudiantil para mejorar modelos predictivos	▪ Varias técnicas de clasificación (no detalladas) ▪ Reequilibrado de datos
[25]	2024	Predicción del desempeño de estudiantes pruebas para su ingreso al bachillerato limitado a parámetros de investigación únicamente académicos	▪ Árbol de decisión J48 ▪ Naïve Bayes ▪ Random Forest ▪ Perceptrón Multicapa

3.2 Aprendizaje no supervisado

Con los métodos de aprendizaje no supervisado se ha planteado la evaluación de las propiedades cualitativas desde una perspectiva diferente a lo observado en el aprendizaje supervisado. En el ámbito de la educación no se muestra una excepción, como se expone en [26], el concepto de analizar cómo los clústeres educativos (conjuntos de instituciones, empresas y actores sociales) pueden aplicarse como herramienta estratégica en el desarrollo de la educación nacional (en este caso, en Ucrania). Todo esto con el objetivo de que los clústeres mejoren la innovación, el intercambio de información, la cooperación y la competitividad en el sistema educativo.

Este tipo de observaciones se pueden observar incluso en [27], donde se plantea como los clústeres de innovación (que integran investigación, educación y producción) pueden mejorar la competitividad en distintos sectores de la economía. El estudio busca demostrar como la interacción entre ciencia, educación e industria en un ciclo de innovación genera sinergias y beneficios mutuos.

La observación de clústeres a gran escala en distintos entornos de desarrollo tecnológico para el beneficio de la educación es un tema que se abarca en [28], en el cual se propone un modelo de clúster educativo en el área de TI (Tecnologías de la Información) en

Ucrania, integrando la industria tecnológica, el Estado, el sistema educativo y el ámbito científico para fortalecer la formación masiva y competitiva de profesionales en informática.

Por otra parte, también se recalca en trabajos como [29] el uso de distintos métodos de “*machine learning*” (aprendizaje automático) para clasificar y agrupar (es decir “*clustering*”) universidades e instituciones de educación superior en Polonia, tomando en cuenta atributos como los campos de estudio, la orientación de programas y otras características. Esto con el propósito de comparar el desempeño de estos en diferentes enfoques.

La implementación de estos métodos de aprendizaje no supervisado en la educación también ha producido frutos en la educación relacionados a la investigación, algo que se plasma en [30] con respecto al avance de diferentes áreas investigadas como “*MOOCs*”, procesamiento de lenguaje natural, entre otros. Algo que se examina en este trabajo en un periodo considerado entre 1979 y 2023, para identificar tendencias, patrones de publicación y principales áreas de investigación.

El uso de técnicas de aprendizaje no supervisado para la identificación de patrones de comportamiento se ha enfocado en el trabajo de [31], el cual se enfoca en una población de estudiantes de nivel superior, con el propósito de formar clústeres con características comunes entre los estudiantes. Esto se realizó con el uso de modelos de “*K-Means*” y máquinas de soporte vectorial.

En el caso presentado por [32], se muestra como el uso de modelos de “*K-Means*” mejorados y redes neuronales convolucionales pueden generar un modelo de evaluación de la educación estética que sea más objetivo y eficaz que los métodos tradicionales. Incluso comentan resultados óptimos en implementaciones con $k = 10$, lo cual demuestra mejoras significativas en calidad de clustering, eficiencia computacional y uso de memoria.

La información expuesta por el trabajo de [33] visualiza como objetivos el mejoramiento de los sistemas de recomendación en educación en línea y ofrecer recomendaciones más precisas y personalizadas de cursos en plataformas educativas. Para ello utilizaron clustering, memoria bidireccional a largo plazo (*BiLSTM*) y perceptrones multicapa para la elaboración de un modelo que alcanzó una precisión del 96% en las recomendaciones, superando métodos tradicionales.

Los objetivos del trabajo de [34] se enfocaron en la evaluación del nivel de informatización educativa en universidades, así como la proposición de un método automatizado con técnicas de IA (Inteligencia Artificial) para seleccionar indicadores relevantes y mejorar la precisión en la evaluación. Esto fue desarrollado con clustering, máquinas de soporte vectorial junto a aprendizaje con refuerzo, logrando una precisión promedio de 93.64%.

Un trabajo en particular visualiza la exploración para aplicar inteligencia artificial y algoritmos de clustering en el análisis de datos de estudiantes en plataformas de educación en línea, así como identificar patrones de comportamiento estudiantil que puedan usarse para personalizar la enseñanza y las recomendaciones de recursos educativos. Esto se examina en la investigación de [35], donde se analizó el desempeño de las técnicas “*K-Means*” y clustering en cuadrícula, con los cuales se propuso un modelo que combina los métodos anteriormente mencionados para un mejor rendimiento en conjunto.

Los datos de la investigación de [36] muestran un enfoque basado en clustering no supervisado para clasificar el rendimiento académico de estudiantes del grupo socioeconómico B40 (40% con menores ingresos en Malasia), esto se realizó con el objetivo de entender los patrones de desempeño de estos estudiantes para ayudar a reducir tasas de deserción, aumentar la graduación y mejorar su situación socioeconómica. Para ello se emplearon modelos de “*K-Means*”, “*BIRCH*” (Reducción y Agrupamiento Iterativo Balanceado mediante Jerarquías) y “*DBSCAN*” de los cuales el modelo de “*K-Means*” presentó un desempeño óptimo para esta situación.

El trabajo realizado por [37] muestra un interés en identificar patrones de aprendizaje y agrupar estudiantes según su progreso, con esta información se planeaba detectar estudiantes con bajo o alto rendimiento y apoyar el desarrollo de estrategias de enseñanza adaptativa. Para esta investigación se empleó un modelo de “*K-Means*” que reveló patrones claros de aprendizaje y permitió identificar a los estudiantes con mejor o peor desempeño, demostrando utilidad práctica en la educación superior.

En la Tabla 3.2 se muestra la información abstraída de los trabajos mencionados previamente, en la cual se enlistan el año de publicación, objetivo del trabajo en cuestión y métodos o conceptos empleados en estos documentos.

Tabla 3.2. Trabajos relacionados con el aprendizaje no supervisado

	Año	Objetivo	Algoritmos
[26]	2017	Análisis de clústeres educativos para su aplicación como herramienta estratégica en el desarrollo de la educación nacional	▪ Teoría de clustering
[27]	2019	Estudio de clústeres de innovación para la mejora de la competitividad en distintos sectores de la economía	
[28]	2020	Propuesta y análisis de un modelo de clúster educativo para fortalecer la formación masiva y competitiva de profesionales en informática	

[29]	2022	Evaluación de distintos métodos de machine learning para clasificar y agrupar universidades e instituciones de educación superior en Polonia	▪ Métodos de clasificación ▪ Clustering
[30]	2023	Análisis bibliométrico y sistemático de los artículos sobre machine learning aplicados a la educación	▪ Métodos bibliométricos
[31]	2023	Desarrollo y validación de un modelo de predicción de rendimiento académico e identificación patrones de comportamiento	▪ K-Means ▪ Máquinas de soporte vectorial
[32]	2025	Creación de un modelo de evaluación de la educación estética	▪ K-Means ▪ Red neuronal convolucional (ResNet-101)
[33]	2024	Mejora de los sistemas de recomendación en educación en línea	▪ Clustering ▪ BiLSTM ▪ Perceptrón multicapa
[34]	2024	Elaboración de un método automatizado con técnicas de IA para seleccionar indicadores relevantes y mejorar la precisión en la evaluación	▪ Clustering ▪ Máquinas de soporte vectorial ▪ Aprendizaje por refuerzo
[35]	2024	Diseño de un modelo de plataforma educativa en línea basada en IA incorporando algoritmos de clustering	▪ K-Means ▪ Grid clustering ▪ GBKM Clustering (Grid + Bayesian K-Means)
[36]	2022	Desarrollo de un enfoque basado en clustering no supervisado para la clasificación del rendimiento académico de estudiantes	▪ K-Means ▪ BIRCH ▪ DBSCAN
[37]	2022	Proposición de un modelo para explotar datos de plataformas digitales de aprendizaje en educación superior	▪ K-Means

Capítulo 4. Propuesta de Solución.

La realización de este trabajo de tesis requerirá de un proceso conformado por diferentes etapas, además de una base de datos conformada por información de diferentes estudiantes de nivel básico que han aplicado para pruebas estandarizadas con el propósito de ingresar al nivel medio-superior. Para la elaboración de este trabajo se han considerado los siguientes puntos a tomar en cuenta para obtener resultados satisfactorios [24].

1. Recolección de datos
2. Preprocesamiento de datos
3. Selección de técnicas de minería de datos
4. Modelado y entrenamiento
5. Evaluación del modelo
6. Interpretación del modelo

A lo largo de esta sección explicaremos los pasos y acciones a realizar en cada uno de estos puntos.

4.1 Recolección de datos

Los datos han sido tomados de una población de estudiantes de nivel básico, estos alumnos proceden de diferentes instituciones educativas y que han presentado una prueba estandarizada como requisito para ingresar al nivel medio superior.

Con la finalidad de proteger las identidades de los estudiantes, los datos considerados no contienen nombres ni identificadores personales, únicamente se han tomado los datos estrictamente académicos de cada estudiante como son promedio de calificaciones en diferentes áreas de conocimiento y el nivel de habilidad mostrado en la prueba. Estos datos se encuentran almacenados en un archivo .csv para un manejo óptimo y práctico.

Estos datos han sido extraídos de la aplicación de una prueba estandarizada aplicada en el año 2022 a alumnos de nivel básico por parte de la Benemérita Universidad Autónoma de Puebla, esto con el propósito de evaluar las aptitudes académicas para su admisión a los planes de estudios de educación media-superior ofertados por la universidad. Estos datos constituyen no solo el resultado de áreas académicas evaluadas en este proceso, sino también algunos aspectos mencionados anteriormente generado por un registro socioeconómico generado por 17334 estudiantes. Los resultados de estas áreas académicas evaluadas, junto con los datos recopilados por cada estudiante mediante el registro socioeconómico serán tomados como variables para el entrenamiento de los modelos a desarrollar en este proyecto, los cuales se plasman de forma general en la Tabla 4.1.

Tabla 4.1. Estructura de datos del dataset utilizado en este proyecto

Nombre de Variable	Tipo de Variable	Rango de Variable	Contenido de Variable
ID	Cuantitativa	1-17334	Identificador del estudiante
HP	Cuantitativa	200-1000	Puntajes de Habilidades del Pensamiento
Español	Cuantitativa	200-1000	Puntajes de Español
Mate	Cuantitativa	200-1000	Puntajes de Matemáticas
CN	Cuantitativa	200-1000	Puntajes de Ciencias Naturales
CS	Cuantitativa	200-1000	Puntajes de Ciencias Sociales
Promedio	Cuantitativa	200-1000	Puntajes promediados de las cinco áreas evaluadas
Aceptado	Cuantitativa	0-1	Estado de aceptación para ingreso al área solicitada
Género (Masculino, Femenino)	Cualitativa	0-1	Tipo de genero del estudiante
NIVEL SOCIOECONOMICO	Cualitativa	1-6	Rango socioeconómico del estudiante
FAMILIA	Cuantitativa	1-14	Número de miembros familiares viviendo con el estudiante
L_INDIG	Cualitativa	0-1	Indica si el estudiante habla una lengua indígena
LENGUA INDIGENA	Cualitativa	0-12	Tipo de lengua indígena que habla el estudiante
NO_ESP	Cualitativa	0-1	Indica si el estudiante tiene problemas para entender el español
PAIS	Cualitativa	1-18	País de nacimiento del estudiante

CP	Cuantitativa	0-859	Código Postal de residencia del estudiante
MUNICIPIO	Cualitativa	0-280	Municipio de residencia del estudiante
ESTADO	Cualitativa	0-18	Estado de residencia del estudiante
CLAVE_ESC	Cualitativa	1-1616	Clave del centro de trabajo
NOMBRE_ESCUELA	Cualitativa	1-1201	Nombre de la escuela de procedencia del estudiante
SECTOR	Cualitativa	0-2	Tipo de escuela de procedencia del estudiante
ESTADO_ESC	Cualitativa	1-27	Estado donde se ubica la escuela de procedencia
GRADO_MADRE	Cualitativa	0-13	Grado máximo de estudios de la madre o tutora
GRADO_PADRE	Cualitativa	0-13	Grado máximo de estudios del padre o tutor
RANGO_CAL	Cuantitativa	1-4	Rango donde se ubica el promedio parcial o final del estudiante
INTERVALO_CAL	Cuantitativa	1-11	Intervalo del promedio parcial o final del estudiante
DISCAPACIDAD	Cualitativa	0-1	Indica si el estudiante tiene una discapacidad
APOYO	Cualitativa	0-9	Tipo de discapacidad que tiene el estudiante
APOYO_EGA	Cualitativa	0-2	Indica si el estudiante necesita algún apoyo para realizar su examen
TIPO_APOYO	Cualitativa	0-2	Tipo de apoyo requerido por el estudiante

4.2 Preprocesamiento de datos

Una vez obtenido los datos es importante verificar la integridad de estos con respecto a lo requerido para la realización de este trabajo, por lo que será necesario la depuración de todos los datos que no cumplan con las características necesarias. Será necesario eliminar todos aquellos datos que estén incompletos, que no sean consistentes comparados con la información requerida y transformarlos en caso de requerirlo, ya que en ocasiones es necesario ajustar el tipo de dato que se introducirá a algunos algoritmos o para modelarlos de forma distinta. Esto implica la eliminación de algunas variables cuyo impacto no es significativo para su implementación dentro de los modelos a desarrollar, con el propósito de encontrar la correlación entre diferentes variables con respecto al comportamiento que presentarán al momento de ser procesados por los modelos desarrollados.

En primer lugar, se considera una configuración utilizando las variables académicas, esto con la finalidad de contrastar el comportamiento de las predicciones con el uso de otras variables no académicas, consideradas por la literatura como fundamentales en el proceso de enseñanza-aprendizaje. Dichas variables y relaciones contempladas para la implementación en los modelos desarrollados se plasman en la Tabla 4.2.

Tabla 4.2. Configuraciones de variables seleccionadas para la implementación.

HP', 'Español', 'Mate', 'CN', 'CS'
'HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'
'HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'
'HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'
'HP', 'Español', 'Mate', 'CN', 'CS', 'L INDIG', 'LENGUA INDIGENA', 'NO ESP'
'HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO MADRE', 'GRADO PADRE'
'HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO ESC', 'INTERVALO CAL'

4.3 Selección de técnicas de minería de datos

Con los datos preparados para su uso se procederá a identificar las técnicas y algoritmos que serán de utilidad para las diversas tareas a realizar en este trabajo, tanto para la predicción de valores como para la identificación de patrones. En el caso de la predicción y clasificación de valores es necesario incorporar algunas técnicas relacionadas al aprendizaje supervisado, las cuales nos permitirán procesar de diferentes formas las configuraciones mencionadas anteriormente en la sección 4.2. En este caso se eligieron tres técnicas para la obtención de estos resultados.

- *Regresión Lineal*
- *“Random Forest”*
- *“K-Nearest Neighbors”*

La aproximación del método de *regresión lineal* para la predicción de valores, que en este caso será el promedio del puntaje que los estudiantes obtendrán en sus pruebas

estandarizadas, presenta escenarios interesantes incluso con la intervención de variables que representan aspectos no académicos al momento de entrenar este modelo; lo cual es un aspecto importante que se observa al utilizar árboles de decisión como se presenta en la técnica de “*Random Forest*” e incluso en el uso de la cercanía de los datos como es el caso de “*K-NN*”, permitiendo la comparación de estas técnicas para una mejor perspectiva para su uso en escenarios similares a este estudio.

No obstante, un punto que resalta mejor la perspectiva sobre la interacción de estas variables entre sí es la agrupación de estos datos a través de la complejidad de la configuración a procesar, es decir, como se visualizan estos grupos de datos al aplicar diferentes variables y obtener una visualización general de cómo se conforman estos parámetros, por lo que se han elegido dos técnicas para la observación de estas características.

- “*K-Means*”
- “*DBSCAN*”

4.4 Modelado y entrenamiento

Una vez se ha definido que tipo de algoritmos serán empleados se hará la implementación en código de los conocimientos seleccionados, dicha codificación se hará en Python, un lenguaje de programación con diferentes alternativas para analizar y procesar datos. Para este trabajo de tesis se implementaron las librerías de *panda*, *numpy*, *keras*, *scikit-learn*, *matplotlib* y *time* para la implementación de las técnicas mencionadas en la sección 4.3, estas librerías son herramientas comunes en la aplicación de estas técnicas, con basta documentación para su uso y cuya importación puede apreciarse en el listado 4.1.

```
# =====  
# Tratamiento de datos  
import pandas as pd  
import numpy as np  
  
# =====  
# Gráficas  
import matplotlib.pyplot as plt  
  
# =====  
# Aprendizaje supervisado  
from sklearn.model_selection import train_test_split  
from sklearn.preprocessing import StandardScaler  
from sklearn.ensemble import RandomForestRegressor  
from sklearn.neighbors import KNeighborsRegressor  
from keras.models import Sequential  
from keras.layers import Dense
```

```

from keras.optimizers import Adam

# =====
# Aprendizaje no supervisado
from sklearn.cluster import KMeans
from sklearn.cluster import DBSCAN
from sklearn.decomposition import PCA

# =====
# Métricas
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.metrics import cohen_kappa_score

# =====
# Utilidades
import time

```

Listado 4.1. Librerías de Python implementadas

Seguidamente presentaremos las implementaciones de los algoritmos de aprendizaje supervisado y no supervisado desarrollados en esta tesis.

4.4.1 Implementación de Regresión Lineal

Para el método de *regresión lineal* se optó por el uso de la librería *keras*, ya que esta herramienta permite el control del entrenamiento del modelo a desarrollar a través de un número de épocas necesarias para la finalización del proceso de aprendizaje, las cuales pueden modificarse de acuerdo con el rendimiento que este modelo muestre con respecto a la tasa de prueba configurada y las variables independientes introducidas, lo cual se puede apreciar en el listado 4.2 a continuación.

```

# =====
# Dividir los datos en conjuntos de entrenamiento (40%) y prueba (60%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.6,
random_state=42)

# Normalizar las características (tanto de entrenamiento como de prueba)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Crear el modelo de regresión lineal simple con Keras
model = Sequential()
model.add(Dense(1, input_dim=len(features), activation='linear')) # Capa
Lineal (regresión lineal)

```

```
# Compilar el modelo
model.compile(loss='mean_squared_error', optimizer=Adam(learning_rate=0.01))

# Entrenar el modelo al 60% con 300 Épocas EXITO
model.fit(X_train_scaled, y_train, epochs=300, verbose=1,
validation_data=(X_test_scaled, y_test))

# Hacer predicciones con el conjunto de prueba
y_pred = model.predict(X_test_scaled).flatten()
```

Listado 4.2. Modelo de regresión lineal en Python

4.4.2 Implementación de Random Forest

Con respecto a la técnica de “*Random Forest*”, se optó por el uso de la librería *scikit-learn* para el desarrollo de este modelo, este tipo de modelos muestran un desempeño que varía según el número de árboles generados para el proceso de prueba del modelo, esta configuración nos permite observar el comportamiento del modelo en conjunto a la tasa de prueba y variables independientes configuradas, lo cual puede apreciarse en el listado 4.3.

```
# =====
# Dividir los datos en conjuntos de entrenamiento (40%) y prueba (60%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.6,
random_state=42)

# Crear el modelo de Random Forest
model = RandomForestRegressor(
    n_estimators=300, # Número de árboles en el bosque
    max_depth=None,   # Sin límite de profundidad
    random_state=42   # Fijar la semilla para reproducibilidad
)

# Entrenar el modelo con el conjunto de entrenamiento
model.fit(X_train, y_train)

# Hacer predicciones con el conjunto de prueba
y_pred = model.predict(X_test)
```

Listado 4.3. Modelo de Random Forest en Python

4.4.3 Implementación de K-Nearest Neighbors

En el caso de la técnica de “*K-NN*” también se ha optado por el uso de la librería *scikit-learn* para la implementación de herramientas que nos permitan desarrollar este

modelo, con esta librería no solo podemos configurar el número de vecinos a considerar para el aprendizaje del modelo en cuestión, sino también que parámetros de distancia afectaran a la selección de cercanía entre puntos de datos, lo cual se muestra en el listado 4.4.

```
# =====  
# Dividir los datos en conjuntos de entrenamiento (40%) y prueba (60%)  
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.6,  
random_state=42)  
  
# Normalizar las características (tanto de entrenamiento como de prueba)  
scaler = StandardScaler()  
X_train_scaled = scaler.fit_transform(X_train)  
X_test_scaled = scaler.transform(X_test)  
  
# Crear el modelo KNN con k vecinos  
model = KNeighborsRegressor(  
    n_neighbors=50, # Número de vecinos a considerar  
    metric='minkowski', # Métrica de distancia (Euclídea por defecto)  
    p=2 # Parámetro de la distancia Minkowski (p=2 -> Euclídea)  
)  
  
# Entrenar el modelo (esto simplemente almacena los datos de entrenamiento)  
model.fit(X_train_scaled, y_train)  
  
# Hacer predicciones con el conjunto de prueba  
y_pred = model.predict(X_test_scaled)
```

Listado 4.4. Modelo de K-NN en Python

4.4.4 Implementación de K-Means

En relación con la técnica de “*K-Means*” se decidió que el desarrollo de este modelo se efectuaría con la librería *scikit-learn*, en este caso es posible seleccionar el número de clústeres o grupos de datos que se generarán, por lo que para este caso se definen, con base en los atributos académicos a procesar, es decir, se definen cinco clústeres a resaltar en los datos y se visualizan haciendo uso de PCA (Análisis de Componentes Principales) realizando una reducción de dimensionalidad para representarlos en un gráfico 2D, esto se expone en el listado 4.5.

```
# =====  
# Normalizar las características  
scaler = StandardScaler()  
X_scaled = scaler.fit_transform(X)  
  
# Configurar y entrenar el modelo K-Means
```

```

n_clusters = 5 # Define el número de clústeres
kmeans = KMeans(n_clusters=n_clusters, random_state=42)
kmeans.fit(X_scaled)

# Obtener los resultados del modelo
labels = kmeans.labels_ # Etiquetas asignadas a cada punto
centroids = kmeans.cluster_centers_ # Coordenadas de los centroides
inertia = kmeans.inertia_ # Inercia (calidad del ajuste)

# Reducir la dimensionalidad a 2D para visualización (usando PCA)
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

```

Listado 4.5. Modelo de K-Means en Python

4.4.5 Implementación de DBSCAN

El modelo de la técnica de “*DBSCAN*” se implementó a través de la librería *scikit-learn* al igual que el modelo de “*K-Means*”; no obstante, a diferencia de este último, maneja la vecindad de los puntos para generar grupos de distintos tamaños sin la necesidad de un centroide, lo que puede darnos una perspectiva diferente a lo observable con “*K-Means*”, este modelo se plasma en el listado 4.6. Al igual que en “*K-Means*” se hace uso del PCA para reducir la dimensión de los datos y poder observarlos en 2D.

```

# =====
# Normalizar las características
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Configurar y entrenar el modelo DBSCAN
eps = 0.5 # Radio de vecindad
min_samples = 5 # Número mínimo de muestras para formar un clúster
dbscan = DBSCAN(eps=eps, min_samples=min_samples)
labels = dbscan.fit_predict(X_scaled)

# Calcular el número de clústeres encontrados (excluyendo ruido)
n_clusters = len(set(labels)) - (1 if -1 in labels else 0)
n_noise = list(labels).count(-1)

# Reducir la dimensionalidad a 2D para visualización (usando PCA)
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_scaled)

```

Listado 4.6. Modelo de DBSCAN en Python

4.5 Evaluación e interpretación del modelo

Una vez que los modelos de aprendizaje sean implementados y entrenados será momento de comprobar la calidad de los resultados que estos devuelvan, esto se hará considerando métricas de evaluación como MSE , R^2 y $Kappa$ para los modelos de aprendizaje supervisado. Además, un aspecto a observar, que aunque no es relevante en este trabajo, es el tiempo que estos modelos tardan en ejecutarse para devolver los resultados correspondientes, por lo que para este fin se ha decidido medir el tiempo de inicio de entrenamiento del modelo correspondiente hasta que este finaliza, midiendo estos tiempos en términos de milisegundos como se observa en la ecuación 4.1.

$$Tiempo\ de\ ejecución = \frac{(tiempo\ final - tiempo\ inicial)}{1000} \quad (4.1)$$

La consideración de medir los tiempos de ejecución está enfocada en datos meramente informativos, sobre todo porque estos modelos fueron ejecutados en un equipo convencional con las siguientes características.

- Procesador 12th Gen Intel(R) Core(TM) i7-12700H (2.30 GHz)
- 16GB Memoria RAM
- Tarjeta gráfica NVIDIA GeForce RTX 4070 16GB

Con los resultados experimentales generados por los algoritmos de aprendizaje supervisado, además de los patrones observados en los clústeres hechos por los algoritmos de aprendizaje no supervisado, se podrán comparar las diferencias entre las distintas métricas, las fortalezas y debilidades de los modelos generados y las posibles aplicaciones a futuro que pueden generarse con base a lo registrado en esta tesis.

Capítulo 5. Resultados experimentales.

La propuesta de solución implementada en este trabajo de tesis ha plasmado diversos resultados que han sido de utilidad para la interpretación de los datos procesados por los modelos desarrollados, las cuales son explicadas a lo largo de este capítulo. Estos resultados involucran una observación general del comportamiento de las configuraciones contempladas en la Tabla 4.2 dentro de los modelos de aprendizaje no supervisado para una perspectiva clara de las características estudiadas. Así como el análisis de las predicciones hechas por los modelos de aprendizaje supervisado, en el que se hizo uso de los resultados generados con las métricas de evaluación para examinar el rendimiento de los modelos implementados a través de diferentes etapas de experimentación, las cuales ejercen diferentes niveles de estrés para el entrenamiento de los modelos desarrollados y una inspección general del tiempo de ejecución para la generación de resultados.

5.1 Aprendizaje no supervisado

Los resultados obtenidos por los modelos de aprendizaje no supervisado han arrojado gráficos interesantes con respecto a las configuraciones mostradas en la Tabla 4.2 del capítulo anterior; diversos clústeres generados por estos modelos permiten tener un panorama general del comportamiento de las variables entre sí, sobre todo por las diferentes características que representan sobre los estudiantes que aspiran por un lugar para su educación media superior, de este modo observamos que exista una correlación entre las variables académicas y las variables socioeconómicas al observar cambios en la agrupación de los datos.

5.1.1 Resultados de K-Means

Una de las formas en las que se puede generar un clúster es a través del uso de centroides para manejar diferentes tamaños y formas de clústeres para la observación de patrones o la relación de las variables entre sí; por tanto, en este caso observaremos en una primera instancia la correlación de las variables independientes académicas que se toman como base en la primera configuración de la Tabla 4.2, lo cual se puede examinar en las figuras mostradas en esta sección.

Estas gráficas son visualizaciones poderosas de sintetizar patrones latentes en datos educativos complejos.

Los ejes están etiquetados como:

- Componente 1: "Agrupación basada en características educativas (Componente 1)"

- Componente 2: "Agrupación basada en características educativas (Componente 2)"

En la construcción de los ejes se aplicó una técnica de reducción de dimensionalidad como Análisis de Componentes Principales (PCA) después de aplicar “*K-Means*”. PCA transforma las variables originales (habilidades, español, matemáticas, ciencias naturales, ciencias sociales y el puntaje estandarizado) en componentes no correlacionados que capturan la mayor varianza posible en los datos.

En este caso, el componente 1 (eje X) representa una combinación lineal de las variables originales que explica la mayor parte de la variabilidad en los datos. Por su parte, el componente 2 (eje Y) representa la segunda combinación lineal más importante que es ortogonal a la primera. Es importante mencionar que al normalizar la información esto ocurre alrededor del 0 (media normal), por lo que los ejes abarcan valores negativos y positivos.

Aunque no se puede interpretar directamente como “matemáticas” o “español”, sí se puede decir que cada componente refleja patrones latentes en el rendimiento académico. Cada punto representa a un estudiante y su ubicación en el plano indica cómo se posiciona ese estudiante respecto a los patrones latentes extraídos por PCA.

- Puntos cercanos entre sí: estudiantes con perfiles académicos similares.
- Puntos alejados: estudiantes con diferencias significativas en sus calificaciones y puntaje estandarizado.
- Color del punto: indica el clúster al que pertenece según “*K-Means*”. Cada color representa un grupo de estudiantes con características académicas similares.

A pesar de que el puntaje estandarizado obtenido por el estudiante en la prueba fue incluido como variable dependiente en el modelo original, en PCA y “*K-Means*” todos los datos se tratan como insumos para encontrar patrones. Por lo tanto,

- Si el puntaje estandarizado tiene alta correlación con alguno de los componentes, los clústeres reflejan niveles de desempeño académico.

Es posible validar esto analizando la media del puntaje estandarizado dentro de cada clúster. Si hay diferencias significativas, los clústeres pueden representar niveles de riesgo o excelencia académica. Sin embargo, este análisis se escapa del objetivo de este trabajo, por lo que se puede considerar como un trabajo futuro.

Para la observación del comportamiento de las variables aplicadas en estos modelos se ha optado por implementar en primera instancia las variables académicas del conjunto de datos, para posteriormente agregar distintas variables no académicas que permitan generar configuraciones más complejas.

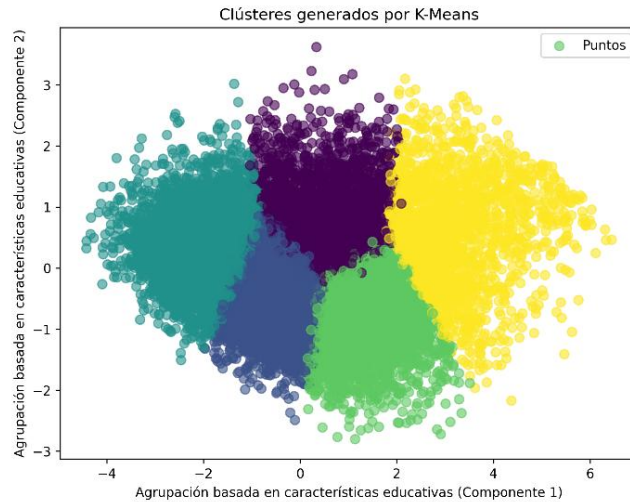


Figura 5.1. Clústeres generados por K-Means

En la Figura 5.1 se observan 5 grupos diferentes destacados de la población general, como estamos tomando la base que son las características académicas podemos visualizar como dichas variables se reflejan en el conjunto de datos, donde cada grupo suele destacar de forma distinta a los demás, que puede interpretarse como la aptitud de los estudiantes sobre ciertas áreas académicas en comparación a otras, lo cual se destaca en los diferentes colores.

Como se ha dicho, la Figura 5.1 está construida utilizando exclusivamente datos académicos. Cada color representa un grupo de estudiantes que comparten similitudes académicas, similitudes que son determinadas por variación en sus datos académicos.

Si al análisis “*K-Means*” se le introducen otros factores como las variables no académicas, entonces es posible observar cambios en los grupos generados por este modelo. La Figura 5.2 muestra los grupos que se forman al considerar la variable *Género* con los niveles, “*masculino*” y “*femenino*”.

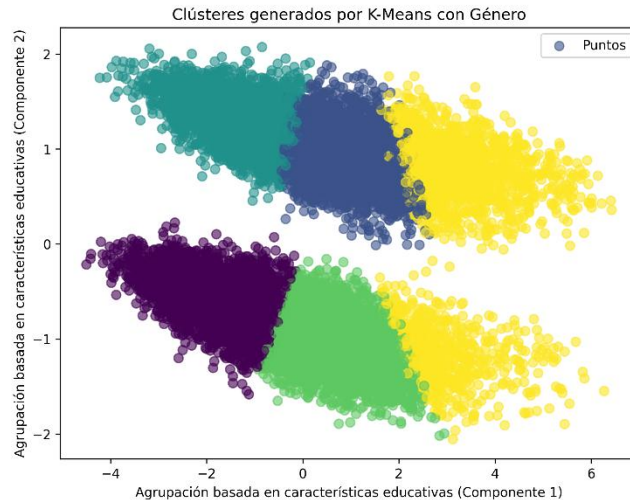


Figura 5.2. Clústeres generados por K-Means aplicando *Género*

Como se observa en la Figura 5.2, la aplicación de la variable de género ha dividido la población de datos en dos grupos. Esto es atribuible a la naturaleza binaria de la variable *Género* cuyos valores “*masculino*” y “*femenino*” se interpretan en valores de 0 y 1, respectivamente, lo que causa una segmentación en la que se puede examinar como cada grupo conserva ciertas aptitudes académicas destacadas, ya que cada clúster refleja una combinación de colores diferentes en relación con las aptitudes académicas en donde ambos grupos tienen un mejor desempeño. Es interesante observar como el grupo superior, correspondiente al género “*femenino*”, muestra mejores resultados al poseer una variabilidad positiva respecto al componente 2 (eje Y).

Una pregunta que surge, al tratarse de una prueba estandarizada que se utiliza como examen de admisión, hace referencia al agrupamiento de los estudiantes respecto a la variable cuyos valores son “*Aceptado*” y “*No aceptado*”. El agrupamiento utilizando “*K-Means*” puede observarse en la Figura 5.3.

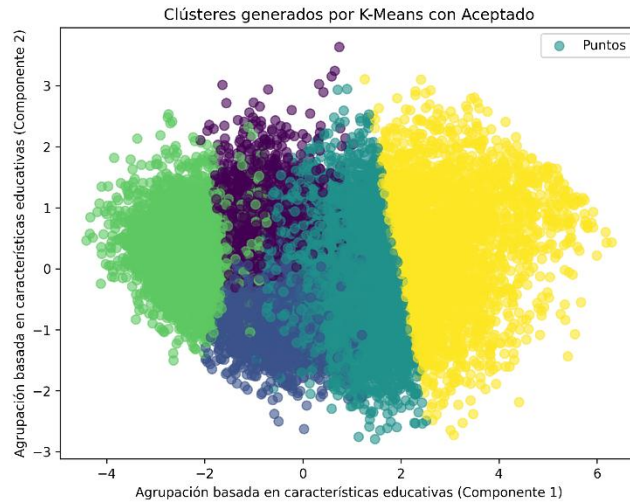


Figura 5.3. Clústeres generados por K-Means aplicando *Aceptado*

Si bien esta variable no académica puede resultar interesante, se trata de un factor que depende de decisiones administrativas, ya que la aceptación de un estudiante está en función del cupo definido por la instancia a la que el estudiante desea ingresar. Tal como se observa en la Figura 5.3, estudiantes con desempeño bajo (color verde) generalmente se agrupan como no aceptados y estudiantes con desempeño muy alto (color amarillo) son agrupados como aceptados.

Sin embargo, también se puede observar que existen grupos que se superponen y esto se atribuye al hecho que hay estudiantes con desempeño aceptable que son aceptados y estudiantes con buen desempeño que no son aceptados, lo que sin duda hace que esta variable pierda interés. Este mismo comportamiento se observa al considerar la variable “*nivel socioeconómico*”, tal como se muestra en la figura 5.4.

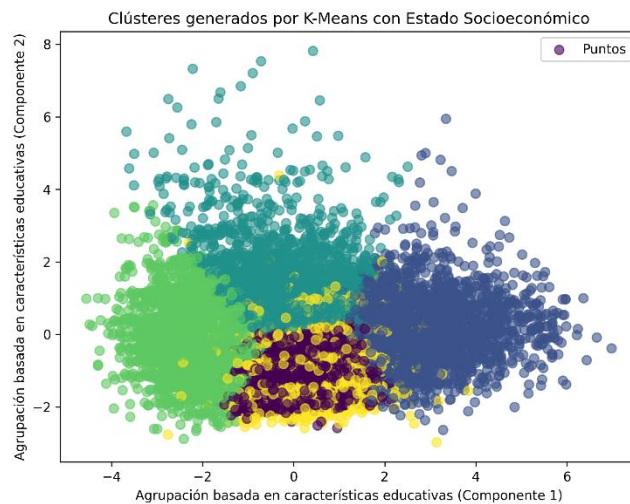


Figura 5.4. Clústeres generados por K-Means aplicando variables socioeconómicas

La Figura 5.4 muestra clústeres generados por la implementación de las áreas académicas y las variables correspondientes al estado socioeconómico del estudiante y el número de integrantes familiares del mismo. Como se observa en la figura anterior, el comportamiento de estas variables cambia considerablemente en comparación con lo observado en la Figura 5.1, donde podemos ver como estas características han generado un clúster con diferentes zonas coloreadas que colindan e incluso se sobreponen, esto recae en las propiedades de ambas variables y su contexto fuera del ámbito académico.

De acuerdo con la Figura 5.3 y Figura 5.4, las variables de aceptación y nivel socioeconómico no muestran una gran relevancia sobre el desempeño académicos de los estudiantes. Hay estudiante que académicamente son muy buenos pero que no son aceptados y hay estudiantes con nivel socioeconómico bajo que tienen un desempeño académico alto.

Otras variables que resultan interesantes son las asociadas a las lenguas indígenas, que determinan si un estudiante tiene como lengua materna algún dialecto o lengua indígena. La Figura 5.5 muestra la agrupación que “*K-Means*” realiza al conjuntar esta variable con las variables académicas.

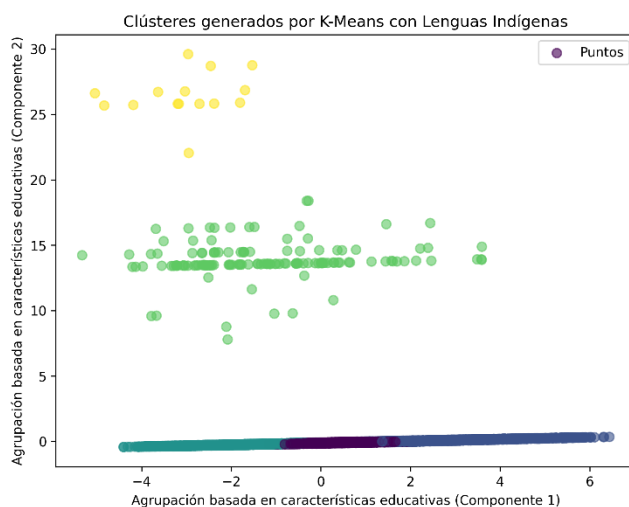


Figura 5.5. Clústeres generados por K-Means aplicando variables de lenguas indígenas

En el caso de la Figura 5.5 se puede examinar un resultado que varía completamente en comparación a las figuras mostradas anteriormente, en esta ocasión se observa como la gran mayoría de los datos se mantienen en la parte inferior de la gráfica, un indicativo de aquellos estudiantes que no manejan alguna lengua indígena, mientras que los estudiantes que hacen uso de lenguas indígenas se mantienen en grupos semi dispersos en diferentes intervalos.

Al hacer un análisis más detallado se puede observar cómo los estudiantes que no hacen uso de algún dialecto o lengua indígena muestran una mejor agrupación en referencia a la variabilidad del componente 1. Esto no ocurre con los estudiantes cuya lengua materna no es el español, sus datos agrupados muestra una mayor dispersión de sus datos académicos,

incluyo mostrando variaciones negativas. Esto es signo de que los estudiantes de lengua materna diferente al español presentan desempeños más pobres que aquellos que su lengua materna si es el español.

El mismo comportamiento se observa al considerar las variables relacionadas al lugar de procedencia, que indica la ubicación de la escuela de la cual proviene el estudiante. Esta variable resulta particularmente importante porque indica los lugares donde se ubican las escuelas, dando pie a considerar si la ubicación es de alta o baja marginación. La Figura 5.6 muestra el agrupamiento.

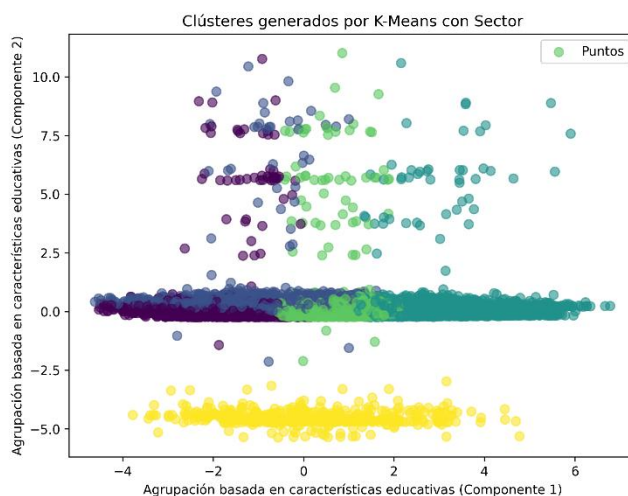


Figura 5.6. Clústeres generados por K-Means aplicando variables de estado de procedencia

La última instancia implementada dentro de este modelo de “*K-Means*” corresponde a las variables académicas junto con la variable de “*estado*”, la cual indica el estado de donde proviene el estudiante. Los resultados son interesantes y pueden examinarse en la Figura 5.7.

El comportamiento que se observa es muy similar a considerar únicamente variables académicas (Figura 5.1). Esto muestra que el desempeño académico de los estudiantes no está influenciado por el estado del cual provienen. Desde luego, podría ser mucho más interesante considerar una perspectiva más profunda sobre la destreza de los estudiantes considerando poblaciones locales y categorizando en alta y baja marginación en lugar de los estados.

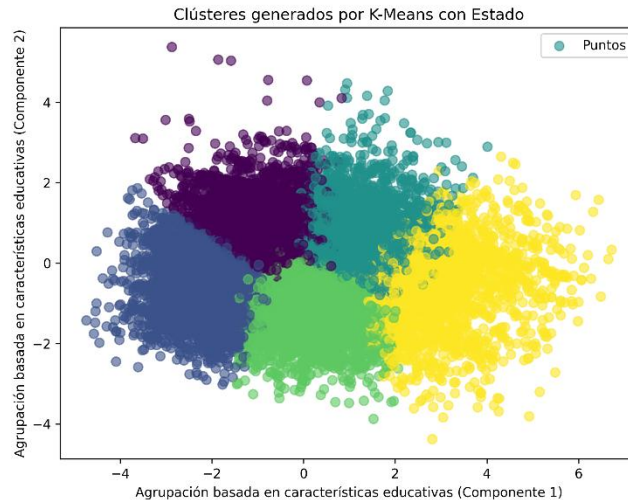


Figura 5.7. Clústeres generados por K-Means aplicando variables de estado de residencia

5.1.2 Resultados de DBSCAN

Para la generación de clústeres con un modelo de “*DBSCAN*” se hace uso de la distancia como radio para considerar puntos de datos como vecinos cercanos y un número mínimo de puntos dentro de ese radio para formar un clúster. Estas condiciones dan como resultado un número indefinido de clústeres con los cuales obtendremos una perspectiva diferente a lo obtenido con el modelo de “*K-Means*”, lo cual plasmaremos a continuación.

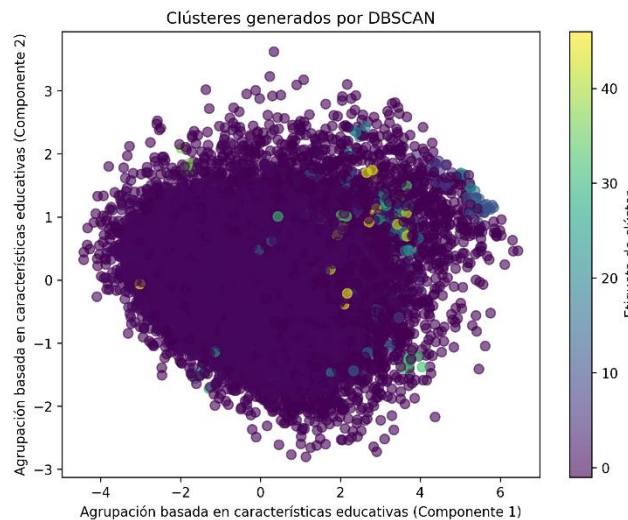


Figura 5.8. Clústeres generados por DBSCAN

Como se puede observar en la Figura 5.8, los grupos que se generan dentro del conjunto producido por variables académicas son muy diferentes a lo visto en las figuras de la sección 5.1.1, donde los grupos tenían un tamaño más grande y destacaba más el coloreado de estos grupos, en este caso podemos identificar como los grupos en este caso son más dispersos en esta ocasión. Esto puede interpretarse como aquellos estudiantes que han tenido

un desempeño similar a sus cercanos, en donde aquellos puntos de datos de color púrpura son los estudiantes que obtuvieron un desempeño promedio, y mientras más cambian estos puntos en la escala de color significa que han tenido un mejor desempeño donde el color amarillo son aquellos que han mostrado un alto rendimiento sobre la mayoría de los datos.

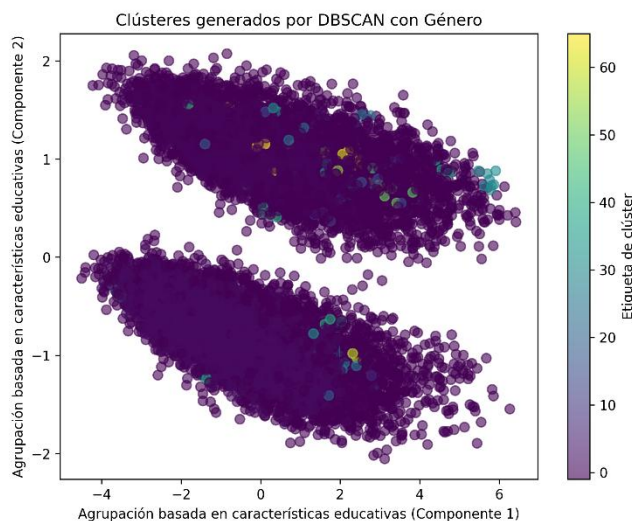


Figura 5.9. Clústeres generados por DBSCAN aplicando *Género*

En la Figura 5.9 se plasma la interacción de las variables académicas con la variable de *Género*, donde la segmentación entre hombres y mujeres demuestra que ambos lados han presentado candidatos con resultados destacados por encima del promedio en sus pruebas estandarizadas.

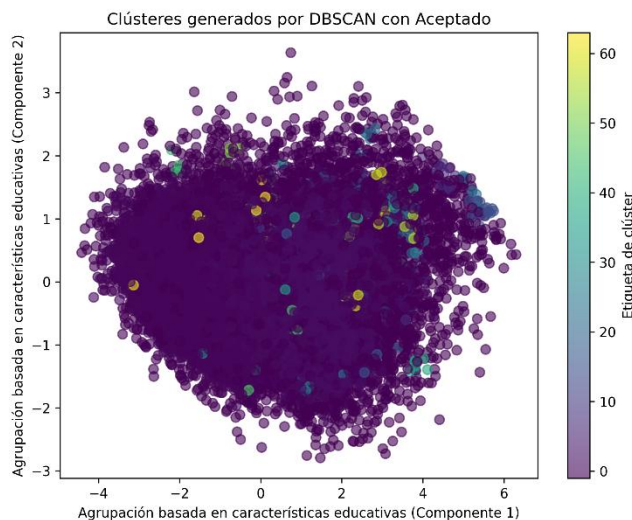


Figura 5.10. Clústeres generados por DBSCAN aplicando *Aceptado*

Al aplicar la variable *Aceptado* a las variables académicas, como se observa en la Figura 5.10, los alumnos que fueron aceptados y rechazados muestran casos excepcionales que generan algunas interrogantes en aquellos casos donde un desempeño notable no aseguró

el éxito del estudiante sobre su prueba, lo cual se observa en aquellos puntos de diferente color.

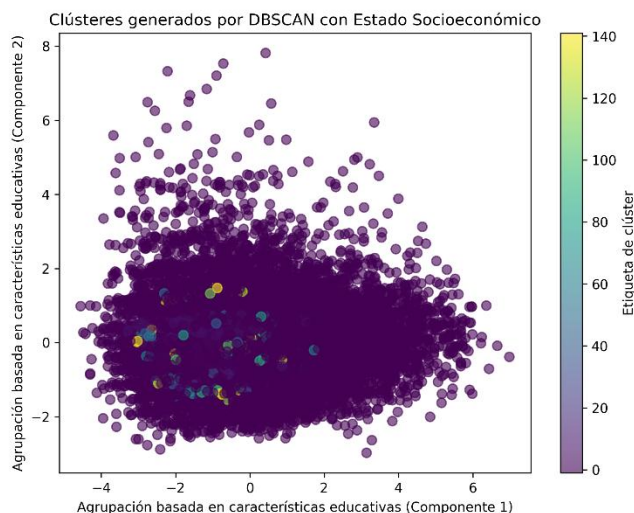


Figura 5.11. Clústeres generados por DBSCAN aplicando variables socioeconómicas

En el caso de variables académicas relacionándose con variables socioeconómicas, el cual se aprecia en la Figura 5.11, se puede observar como la estructura resultante concentra a los casos sobresalientes se concentran en algunas zonas dentro de la población general, un aspecto que presenta cuestiones interesantes con respecto al impacto que estos aspectos del estudiante pueden afectar el desempeño académico de estos.

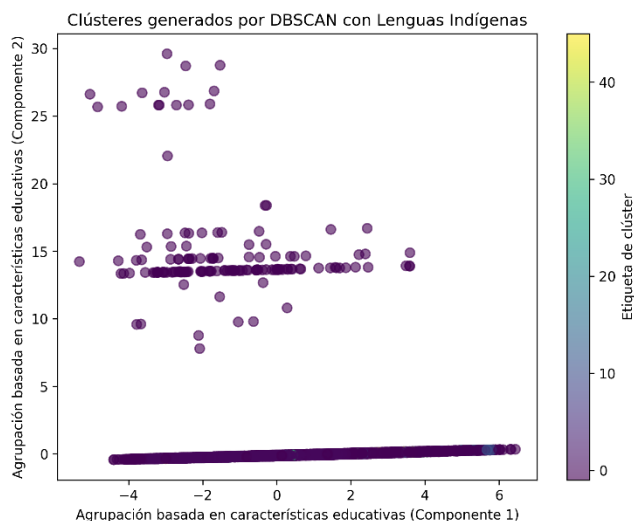


Figura 5.12. Clústeres generados por DBSCAN aplicando variables de lenguas indígenas

La Figura 5.12 muestra un panorama interesante con respecto a los resultados observables, ya que en esta ocasión no se aprecian comportamientos sobresalientes entre aquellos estudiantes, lo que puede indicar que, a pesar de sobresalir en algunas áreas académicas como se plasma en la Figura 5.5, los estudiantes hablantes de lenguas indígenas pueden mejorar en sus aptitudes académicas para un mejor desempeño.

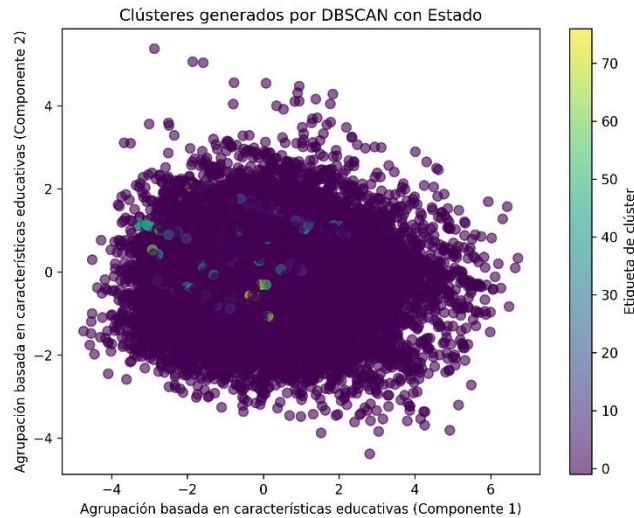


Figura 5.13. Clústeres generados por DBSCAN aplicando variables de estado de residencia

Los resultados vistos en la Figura 5.13 plasman un escenario parecido a lo visto en la Figura 5.11, en el cual los datos que presentan un comportamiento destacado que sobresalga de los demás se aglomeran en algunas partes dentro del clúster más grande, con la excepción de que se logra apreciar como estos datos están más concentrados y su distinción es apenas observable en comparación a otras figuras dentro de esta sección.

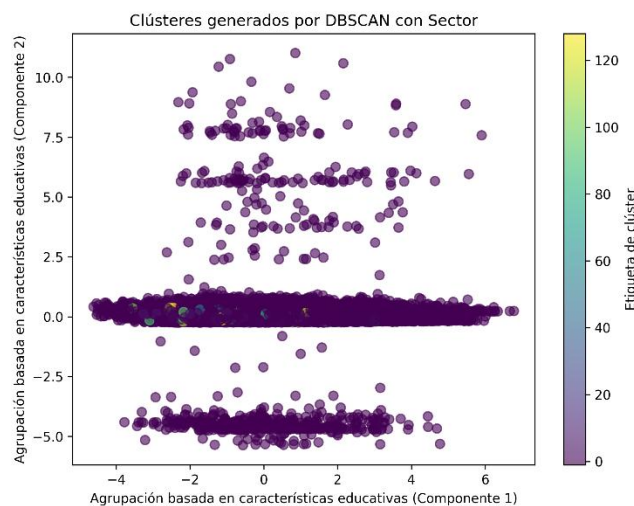


Figura 5.14. Clústeres generados por DBSCAN aplicando variables de estado de procedencia

Por último, tenemos la implementación de variables académicas relacionadas con variables de estado de procedencia, cuyos resultados se aglomeran en un grupo en particular, lo cual genera interrogantes para futuras investigaciones sobre el desempeño de los estudiantes foráneos que aplican para estas pruebas estandarizadas.

5.1.3 Consideraciones del aprendizaje no supervisado

Es importante hacer hincapié en el objetivo de este trabajo de tesis. El estudio busca encontrar patrones que conlleven a la predicción del desempeño académico de los estudiantes que presentan una prueba estandarizada a partir de sus calificaciones históricas del nivel académico inmediato anterior.

Buscar la predicción del desempeño académico de los estudiantes en una prueba estandarizada, requiere la utilización de modelos que regresen como resultado el valor numérico que representa categóricamente el puntaje que el estudiante podría obtener al presentar la prueba.

Por desgracia, los modelos de aprendizaje no supervisado que fueron analizados en este trabajo no permiten la obtención de este valor numérico, sin embargo, si proporcionan información sobre el comportamiento de los datos al considerar las variables académicas en conjunto con otras variables, lo que sin duda es un aliciente para la aplicación de los modelos de aprendizaje supervisado.

No obstante, los modelos de aprendizaje no supervisado ofrecen una gran área de oportunidad, ya que abren nuevos caminos que pueden ser explorados en trabajos futuros, por ejemplo, al considerar en forma más detallada el lugar de procedencia de los estudiantes y generar los clústeres para posteriormente hacer un análisis de aprendizaje supervisado por clúster para determinar estudiantes en zona de riesgo o de alto rendimiento.

5.2 Aprendizaje supervisado

Los modelos de aprendizaje supervisado desarrollados en este trabajo de tesis han revelado resultados interesantes entre sí. El primer resultado mostrado por cada uno de los modelos en cuestión son los datos predichos con base en las variables configuradas en la Tabla 4.2, donde la predicción de este estudio se enfocó en la variable *Promedio*, la cual se designó como variable dependiente debido a su rasgo cuantitativo y el rango que esta posee para una presentación de resultados.

Para una mejor experimentación de los modelos desarrollados se decidió incrementar gradualmente la tasa de prueba para estos modelos en etapas de 20, 40, 60, 80 y 99 por ciento, con el propósito de evaluar las técnicas implementadas a través de distintos márgenes de estrés. Además, cada uno de los modelos tuvo ajustes independientes para cada uno de sus funcionamientos, el modelo de *regresión lineal* necesitó del ajuste del número de épocas por cada tasa de prueba implementada. El modelo de “*Random Forest*” fue probado con cuatro configuraciones de número de árboles, las cuales fueron 50, 100, 150 y 300 árboles. Por último, el modelo de “*K-NN*” fue ajustado a través de cuatro configuraciones de números de vecinos a considerar para la predicción, estos fueron en 5, 10, 25 y 50 vecinos.

No obstante, para una mejor observación de los resultados obtenidos se optó por una comparación de los datos reales que forman parte del dataset original con los datos predichos obtenidos por los modelos de aprendizaje supervisado, tal y como se muestra en la Tabla 5.1.

Tabla 5.1. Muestra de datos predichos

Promedio Real	Promedio Predicho
425.9841	423.9704
616.83604	614.82587
547.63934	545.6396
542.50036	540.50085
408.85206	406.837
539.03268	537.0081
646.30388	644.3191
541.68166	539.672
510.51372	508.4995
534.1572	532.14984
579.9499	577.9498

Con estos datos obtenidos se procede a graficarlos para una mejor observación de los datos a comparar, puesto que es una forma más eficaz de revisar el comportamiento de estos que revisar cada uno de estos individualmente, esto puede apreciarse en la Figura 5.15 en la cual se graficaron los datos reales y predichos en el modelo de *regresión lineal* con base a variables académicas al 40% de tasa de prueba.

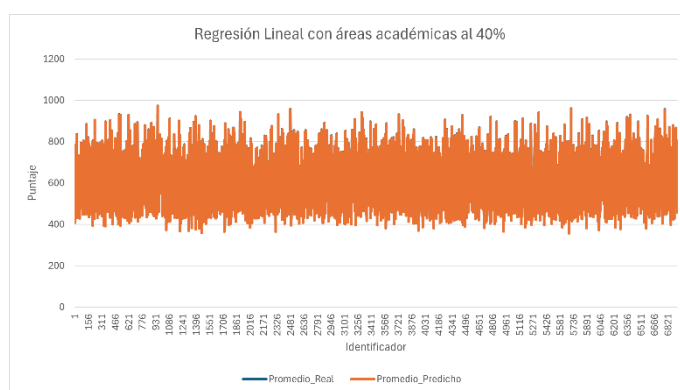


Figura 5.15. Ejemplo de los datos de áreas académicas por Regresión lineal

Con la predicción de los datos graficada es posible tener un panorama más claro del rendimiento que estos modelos de aprendizaje supervisado presentan según las configuraciones asignadas. Sin embargo, para una observación más objetiva de estos modelos es necesario hacer una evaluación precisa del desempeño con base a sus resultados, para ello se han evaluado los modelos con tres métricas de evaluación para aprendizaje supervisado, R^2 , MSE y $Kappa$, en la Tabla 5.2 se encuentra como ejemplo los valores de las métricas obtenidas por el modelo de *regresión lineal* con base a variables académicas al 40% de tasa de prueba.

Tabla 5.2. Muestra de las métricas de evaluación

200 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999829066	1.695453554	0.981833042

Por último, se tomó en consideración el tiempo de ejecución que los modelos de aprendizaje supervisado necesitaron para el procesamiento y predicción de los datos, esta característica nos permite diferenciar la eficacia de cada modelo en términos de tiempo, el cual se ha evaluado en milisegundos para tener valores más precisos sobre dichos modelos. En el caso del modelo de *regresión lineal* con base a variables académicas al 40% de tasa de prueba se obtuvo un tiempo de 65065.62 ms., un tiempo notable pero no por ello el más bajo de todos los escenarios implementados en este trabajo.

Dentro de este apartado mostraremos los resultados óptimos pertenecientes a los modelos de aprendizaje supervisado desarrollados en este trabajo de tesis, cada uno según las configuraciones implementadas tanto en sus variables independientes, número de épocas usadas en *regresión lineal*, número de árboles en “*Random Forest*” o número de vecinos a considerar en “*K-NN*” y la tasa de prueba asignada para cada uno de los puntos anteriormente mencionados.

5.2.1 Resultados no ideales

Un aspecto importante que aclarar sobre estos modelos de aprendizaje supervisado es la calidad de los resultados debido al ajuste de la tasa de aprendizaje y prueba con relación a las configuraciones necesarias para cada uno de estos (épocas, número de árboles o vecinos a considerar). Ya que si estos parámetros no son ajustados correctamente es posible obtener valores de tipo negativos en métricas de evaluación como R^2 o *Kappa*, y valores demasiado grandes en *MSE*, lo cual es contraproducente y cuya causa puede deberse al sobre aprendizaje del modelo o a la falta de aprendizaje por parte de este, lo cual se muestra en la Tabla 5.3 donde se entrenó el modelo de *regresión lineal* con una tasa de prueba del 40% y un número de épocas de 150.

Tabla 5.3. Métricas de evaluación de Regresión lineal no óptimas

150 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	-0.639210584	16258.90797	-0.084773066
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	-0.639243455	16259.23401	-0.084788765
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	-0.639341074	16260.20227	-0.084770265

HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	-0.639187699	16258.68098	-0.084785963
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	-0.639390015	16260.6877	-0.084785963
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	-0.639157019	16258.37668	-0.084825591
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	-0.639235138	16259.15152	-0.084773066

Un punto interesante al examinar los resultados de la Tabla 5.3 es el gran margen de error que puede interpretarse como un desempeño deplorable que refleja la poca o nula información generada por esta implementación, lo cual se debe a la falta de entrenamiento del modelo de *regresión lineal*, esto se observa en los valores que proporcionar tanto por R^2 , MSE y $Kappa$, un punto detallado en la sección 2.4.3. Esta perspectiva incluso puede verse reflejada en la Figura 5.16, donde se puede observar lo diferentes que son los datos reales de los datos predichos por el modelo de *regresión lineal* con una configuración no óptima para su implementación.

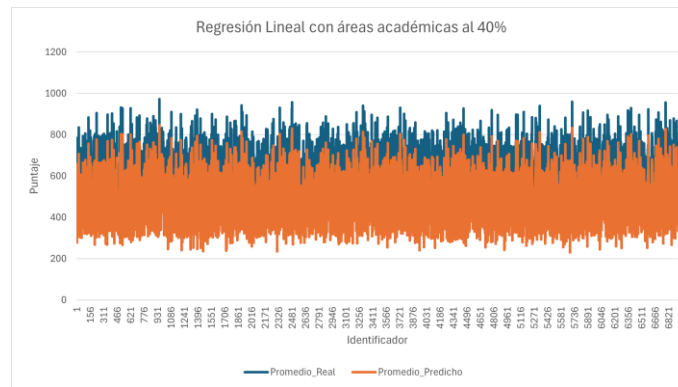


Figura 5.16. Datos no óptimos de áreas académicas por Regresión lineal

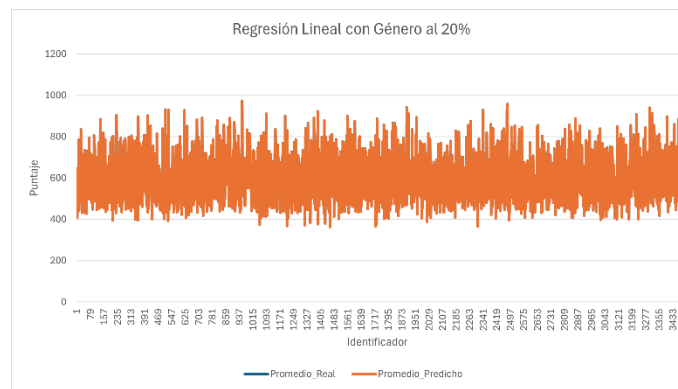
5.2.2 Resultados al 20% de prueba

El modelo de *regresión lineal* ajustado al 20% de tasa de prueba con un total de 150 épocas para entrenamiento ha devuelto resultados interesantes para las instancias implementadas, lo cual puede observarse en la Tabla 5.4, donde podemos destacar un rendimiento notable en la configuración con géneros la cual tuvo un tiempo de ejecución de 62576.85 ms.

Tabla 5.4. Métricas de evaluación de **regresión lineal** al 20% de prueba

150 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999878728	1.187714867	0.982732256
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.999878527	1.189680017	0.982734324
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.999879005	1.185000407	0.982732256
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.999879044	1.18461686	0.982732256
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.99987919	1.18318981	0.982332273
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO MADRE', 'GRADO PADRE'	0.99987916	1.183483841	0.982732256
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.999878736	1.18763659	0.982332273

Los valores de R^2 y MSE de la Tabla 5.4 se mantienen en un rango de valores aceptables, un desempeño que también se ve representado en la medida $Kappa$, un indicio de la calidad de los datos predichos y la fiabilidad de estos, este argumento se ve reflejado gráficamente en la Figura 5.17.

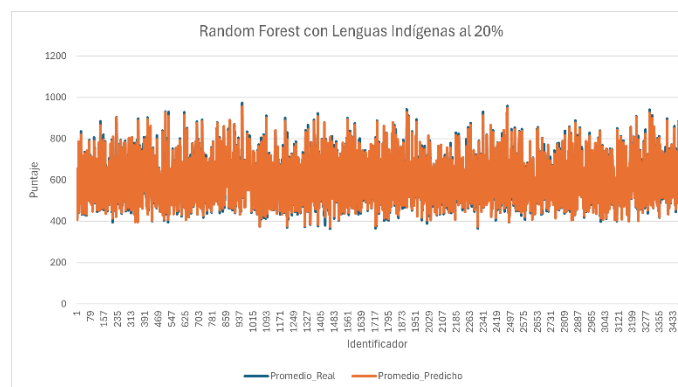
**Figura 5.17.** Regresión lineal con variables académicas y *Género* al 20%

En otra instancia, el modelo de “*Random Forest*” ajustado al 20% de prueba con un total de 300 árboles fue el escenario con mejores resultados para esta etapa de entrenamiento, en el cual la configuración con variables de lenguas indígenas ha presentado resultados óptimos con respecto a su tiempo de ejecución, el cual fue de 15673.36 ms.

Tabla 5.5. Métricas de evaluación de Random Forest al 20% de prueba

300 Árboles			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.994524768	53.62320628	0.929053613
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.994500351	53.86234064	0.929068474
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.993614829	62.53494977	0.924983214
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.994257483	56.24094069	0.929875402
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.994539205	53.48181621	0.929469534
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.994173326	57.06515988	0.925442238
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.994295963	55.86407769	0.927459387

Las instancias vistas en la Tabla 5.5 han retornado valores de evaluación que reflejan un desempeño eficiente con base a los datos de R^2 y MSE , así como su evaluación en $Kappa$ en los cuales dicho panorama muestra un comportamiento que se asimila a lo visto por parte del modelo de *regresión lineal*. Esto puede observarse en la Figura 5.18, en la cual estos argumentos muestran la fiabilidad de los datos predichos en comparación con los datos reales.

**Figura 5.18.** Random Forest con variables académicas y lenguas indígenas al 20%

La última implementación de esta etapa, el modelo de “ K -NN” ajustado al 20% de prueba registró métricas interesantes al ser probado con un total de 25 vecinos cercanos, esto puede observarse claramente en la Tabla 5.6 en la cual se observa un margen de resultados

destacable en la configuración de variables académicas y su tiempo de ejecución de 23.58 ms.

Tabla 5.6. Métricas de evaluación de K-NN al 20% de prueba

25 Vecinos			
Configuración implementada	R ²	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.99735356	25.91864567	0.953243502
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.99602738	38.90696532	0.937367796
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.994770939	51.2122647	0.932119432
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.991407701	84.15107516	0.917240295
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.994045934	58.31280073	0.949187331
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.986341343	133.769864	0.911539757
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO CAL'	0.988871183	108.9931708	0.913558251

El desempeño de este modelo se mantiene en condiciones estables que se respaldan con los valores de la Tabla 5.6, lo cual marca una perspectiva que se mantiene a la par de los modelos de *regresión lineal* y “*Random Forest*”. Esto puede observarse en la Figura 5.19 en la cual se puede ver cómo la predicción se asemeja lo suficiente a los valores predichos.

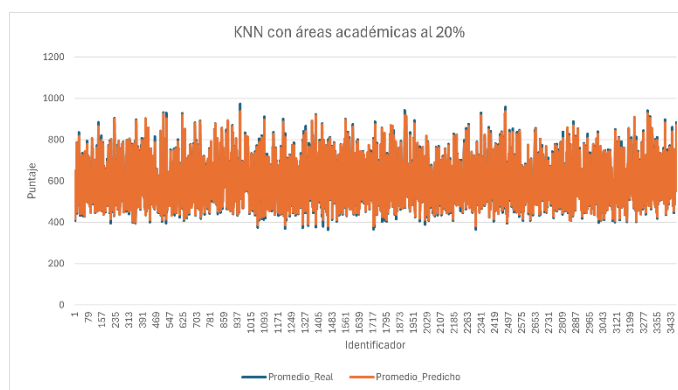


Figura 5.19. K-NN con variables académicas al 20%

5.2.3 Resultados al 40% de prueba

La segunda etapa de entrenamiento involucró el incremento de la tasa de prueba para los tres modelos implementados, empezando por el modelo de *regresión lineal*, el cual mostró resultados apreciables con el cambio de la tasa de prueba en un 40%, en este caso se presentaron métricas de evaluación óptimas al procesar las variables académicas y de estado de residencia en un número de 200 épocas junto a un tiempo de ejecución de 78705.41 ms.

Tabla 5.7. Métricas de evaluación de regresión lineal al 40% de prueba

200 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999829066	1.695453554	0.981833042
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.99982938	1.692340245	0.981834162
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.999829281	1.693315857	0.981833042
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.99982892	1.696900497	0.981833087
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.999829334	1.692794166	0.981833087
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.999828885	1.697243352	0.982232998
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO CAL'	0.982232998	1.693287617	0.982033002

Es posible apreciar como los resultados de R^2 , MSE y $Kappa$ en la Tabla 5.7 se mantienen en un margen de valores similares a los vistos en la implementación realizada por este mismo modelo que se registran en la Tabla 5.4 en la sección 5.2.2. Lo cual se puede observar incluso en la Figura 5.20 donde los valores reales y predichos se asemejan bastante.

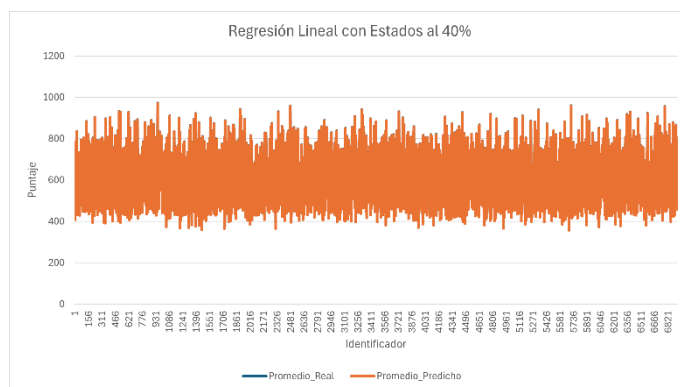


Figura 5.20. Regresión lineal con variables académicas y de estado de residencia al 40%

Por parte del modelo de “*Random Forest*” se observan resultados óptimos con la tasa de prueba al 40%, donde los resultados más destacados se contemplan en la configuración de variables académicas junto a lenguas indígenas, la cual tuvo un tiempo de ejecución de 11568.32 ms.

Tabla 5.8. Métricas de evaluación de Random Forest al 40% de prueba

300 Árboles			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.993954363	59.96511489	0.924430561
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.993866004	60.84152582	0.924231119
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.993117876	68.26202039	0.921977762
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.993668541	62.80011569	0.92341898
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.993941113	60.09654173	0.925427686
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.993536058	64.11418174	0.922809881
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO CAL'	0.993700244	62.48565815	0.924445343

Los resultados de la Tabla 5.8 se asemejan a los vistos en la Tabla 5.5 de la sección 5.2.2, donde si bien es comparable el incremento en los valores del *MSE*, los valores de R^2 y *Kappa* se mantienen en un rango óptimo, lo cual se plasma en la Figura 5.21 con la comparación de los valores predichos y reales.

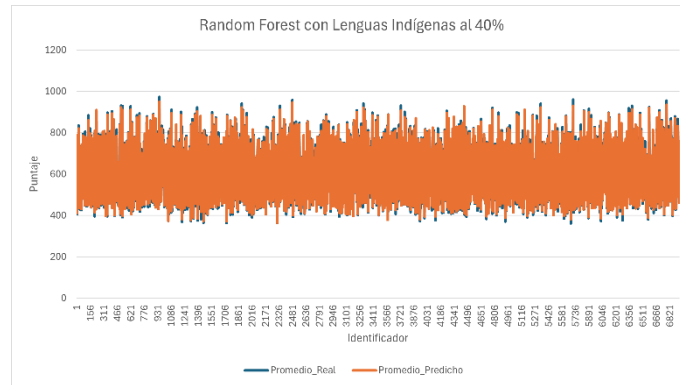


Figura 5.21. Random Forest con variables académicas y lenguas indígenas al 40%

El último modelo implementado, el modelo de “*K-NN*” reveló resultados notables cuando fue probado con la tasa de prueba al 40%, siendo los resultados óptimos encontrados en la configuración de variables académicas con un número de 25 vecinos cercanos, donde las métricas de evaluación mostradas en la Tabla 5.9, junto a la Figura 5.22 demuestran un desempeño relevante a través de esta configuración que se refleja incluso con su tiempo de ejecución de 19.77 ms.

Tabla 5.9. Métricas de evaluación de K-NN al 40% de prueba

25 Vecinos			
Configuración implementada	R ²	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.996921378	30.53605909	0.947429905
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.995317226	46.44723423	0.934492077
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.994299445	56.54234033	0.926639839
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.990292789	96.28332715	0.909601931
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO_ESP'	0.992972528	69.70368721	0.942357452
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.985979197	139.0687335	0.898370265
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.986997019	128.9732226	0.902503681

Estas instancias implementadas han dado resultados interesantes en la Tabla 5.9, la cual en comparación a los datos de la Tabla 5.6 de la sección 5.2.2 muestran datos que reflejan un panorama estable con respecto al rendimiento del modelo desarrollado, ya que los valores de R^2 y $Kappa$ se mantienen en un rango aceptable. Sin embargo, el MSE ha crecido en comparación a la iteración anterior pero aún se conserva en un intervalo tolerable, que se refleja en una predicción viable apreciable en la Figura 5.22.

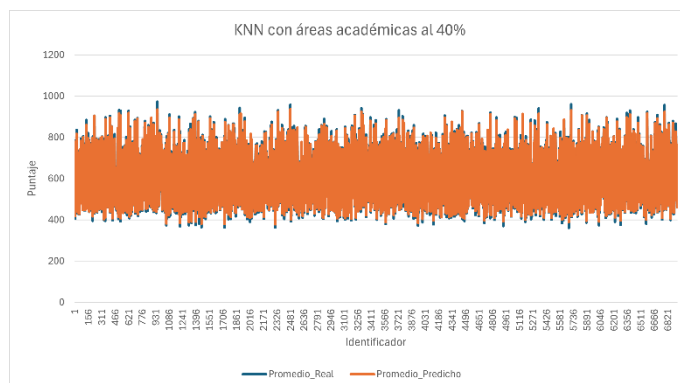


Figura 5.22. K-NN con variables académicas y lenguas indígenas al 40%

5.2.4 Resultados al 60% de prueba

A través de la tercera etapa de experimentación de estos modelos se incrementó nuevamente la tasa de prueba de estos modelos a un 60%, en el primer modelo implementado con estos parámetros fue el modelo de *regresión lineal* con un total de 300 épocas, en el cual se vio un desempeño notable en la configuración de variables académicas junto a la variable *Aceptado* al tener un tiempo de ejecución de 116782.85 ms.

Tabla 5.10. Métricas de evaluación de regresión lineal al 60% de prueba

300 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999862091	1.377608215	0.983355219
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.999862365	1.374879594	0.983355219
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.999861829	1.380229741	0.983620561
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.99986226	1.37592345	0.983353629
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO_ESP'	0.999862005	1.378466943	0.983487886
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.999862025	1.378269347	0.98322256

HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO CAL'	0.999861756	1.380958658	0.983355219
---	-------------	-------------	-------------

Similar a las secciones anteriores, los valores de R^2 , MSE y $Kappa$ se mantienen en un rango sobresaliente en el cual se muestra un desempeño sobresaliente con base a la predicción de los datos, esto se refuerza con una observación comparativa de dichos datos en la Figura 5.23 en la cual examina la similitud de los datos.

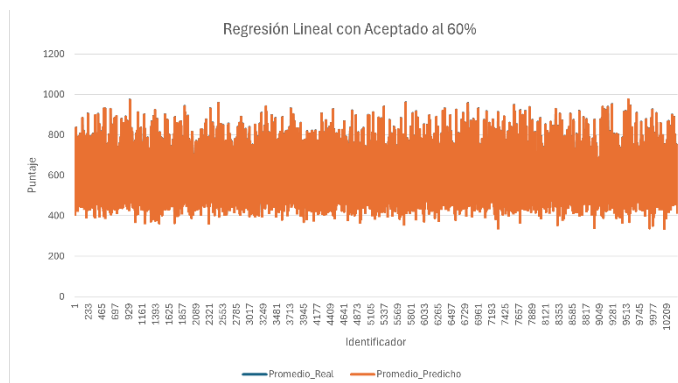


Figura 5.23. Regresión lineal con variables académicas y *Aceptado* al 60%

El modelo de “*Random Forest*” ha presentado un tiempo de ejecución de 7499.97 ms. al ser implementado con la tasa de prueba del 60% con un número de 300 árboles, con estos parámetros se han generado métricas de evaluación presentadas en la Tabla 5.11, en la cual se observa como la configuración de variables académicas en conjunto a lenguas indígenas han presentado una implementación óptima con los mejores valores de esta implementación según los parámetros dados.

Tabla 5.11. Métricas de evaluación de Random Forest al 60% de prueba

300 Árboles			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.992258731	77.32972977	0.918608632
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.992196339	77.95297837	0.918070853
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.99181179	81.79434665	0.917535324
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.991881345	81.0995403	0.915389618
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.992250381	77.41314309	0.918741528

HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO','GRADO_MADRE', 'GRADO_PADRE'	0.991762266	82.28905946	0.916062551
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.991951988	80.39387483	0.91605818

A pesar de presentar un cambio ligero entre los valores de R^2 , MSE y $Kappa$ en la Tabla 5.11 en comparación a lo obtenido en la Tabla 5.8, el desempeño del modelo de “*Random Forest*” se mantiene admisible, con variaciones notables comparadas a las vistas en *regresión lineal* pero mostrando seguridad al momento de predecir, lo que se plasma en la Figura 5.24.

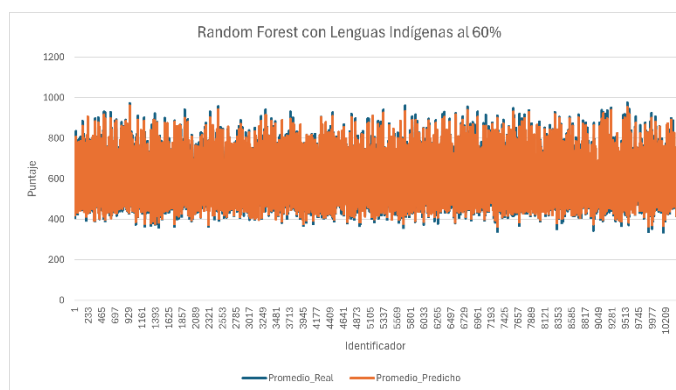


Figura 5.24. Random Forest con variables académicas y lenguas indígenas al 60%

En la implementación del 60% de la tasa de prueba al modelo de “*K-NN*” dio resultados interesantes en las configuraciones asignadas, pero la configuración de variables académicas junto a un ajuste de 25 vecinos ha sido la que ha presentado resultados óptimos para esta prueba, como es el caso de su tiempo de ejecución de 14.5 ms.

Tabla 5.12. Métricas de evaluación de K-NN al 60% de prueba

25 Vecinos			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.995996516	39.99193708	0.940917683
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.993630246	63.6292719	0.926353533
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.992726236	72.65968379	0.914688409
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.988348784	116.3872945	0.892486728

HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.99236498	76.26838046	0.936035005
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.982647659	173.3374505	0.881819476
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.984221691	157.6139922	0.883707998

Los resultados vistos en la Tabla 5.12 son considerablemente diferentes a lo visto en la Tabla 5.9 de la sección 5.2.3, donde los valores de MSE han crecido nuevamente. Esto es intrigante, ya que según lo explicado en la sección 2.4.3 aunque se mantengan los valores de R^2 y $Kappa$, indicando un desempeño admisible, un valor en incremento de parte del MSE puede indicar una deficiencia en la calidad de la predicción. No obstante, hasta ahora las predicciones se mantienen en un rango aceptable, lo cual se reitera en la Figura 5.25, con la cual se comparan dichos datos predichos y reales.

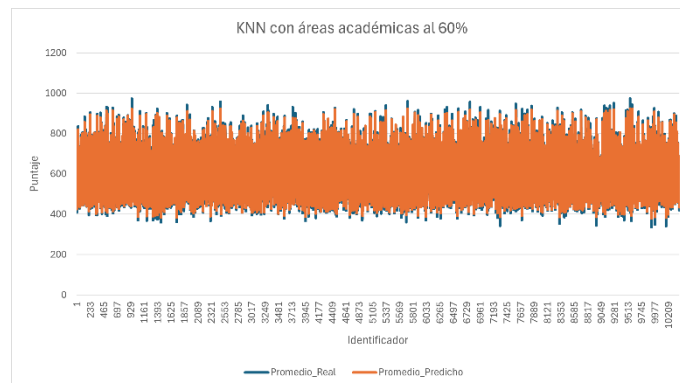


Figura 5.25. K-NN con variables académicas al 60%

5.2.5 Resultados al 80% de prueba

Esta penúltima etapa de entrenamiento se subió la tasa de prueba al 80%, con lo cual los modelos comenzaron a ser presionados a procesar con una menor cantidad de datos para entrenar y predecir, esto se refleja en el modelo de *regresión lineal* cuyos parámetros fueron ajustados a 600 épocas. Esto causó un tiempo de ejecución de 186092.81 ms.

Tabla 5.13. Métricas de evaluación de regresión lineal al 80% de prueba

600 Épocas de procesamiento			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999967524	0.327060291	0.992411615
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.999967496	0.327338616	0.992511595

HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.999967472	0.327583946	0.992511291
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.999967563	0.326660716	0.992511595
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.99996754	0.326896597	0.992411615
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.999967602	0.326273454	0.992511595
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.99996754	0.326899583	0.992511595

Si bien es cierto que el tiempo de ejecución ha aumentado considerablemente a través de las pruebas, los resultados vistos en la Tabla 5.13 se han mantenido en un margen admisible, los valores de las tres métricas de evaluación son óptimos y los resultados comparativos en la Figura 5.26 muestran un desempeño estable por parte de este modelo de aprendizaje, algo que ha divergido de los otros modelos.

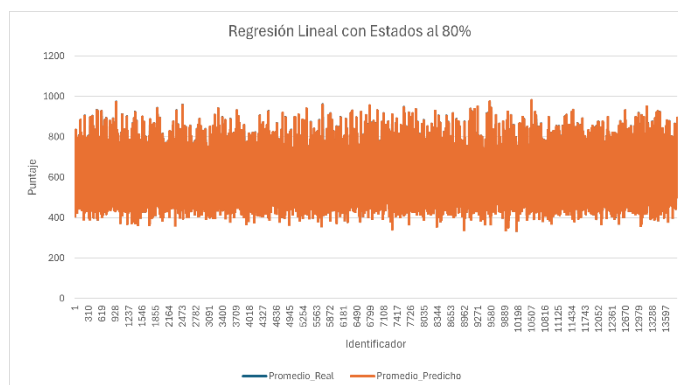


Figura 5.26. Regresión lineal con variables académicas y de estado de residencia al 80%

El modelo de “*Random Forest*” con la tasa de prueba al 80% demostró resultados considerables al procesar las configuraciones asignadas, en el que la configuración de variables con géneros presentó resultados considerables en sus métricas de evaluación en la Tabla 5.14 y un tiempo de ejecución de 3454.55 ms.

Tabla 5.14. Métricas de evaluación de Random Forest al 80% de prueba

300 Árboles			
Configuración implementada	R ²	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.988827829	112.5123716	0.904862535
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.988735432	113.4428791	0.905357659

HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.987545314	125.4282829	0.897372159
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.988283507	117.9941169	0.902426424
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.988795287	112.8400959	0.90465565
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.98806353	120.2094554	0.901010113
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.988356072	117.2633242	0.903334184

Entre los números generados por la evaluación de este modelo que se observan en la Tabla 5.14 se pueden observar valores aceptables en las tres métricas. Sin embargo, se puede hacer una comparación con respecto a las iteraciones anteriores, el valor de *Kappa* ha decrecido gradualmente en cada una de las etapas de prueba, y aunque esta métrica ha bajado en cierto grado, esto se mantiene en un comportamiento tolerable con predicciones que discrepan en cierto grado con los datos reales pero que se asemejan de igual manera, algo que se examina en la Figura 5.27.

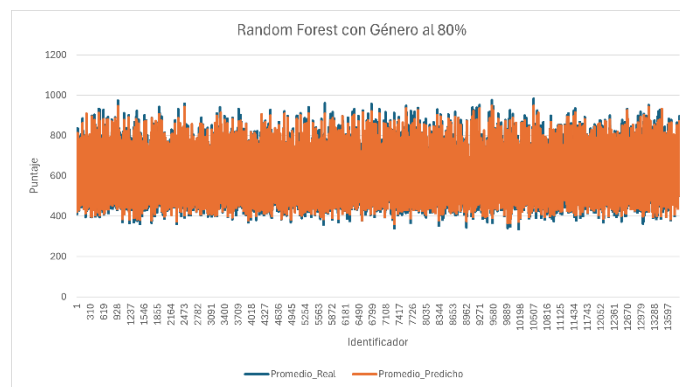


Figura 5.27. Random Forest con variables académicas y *Género* al 80%

Los resultados obtenidos del modelo de “*K-NN*” con una tasa de prueba del 80% muestra resultados destacados en un número de 10 vecinos, esto se refleja en la configuración de variables académicas, con un tiempo de ejecución de 13.91 ms., sus métricas de evaluación y datos predichos, los cuales se plasman en la Tabla 5.15.

Tabla 5.15. Métricas de evaluación de K-NN al 80% de prueba

10 Vecinos			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.994489322	55.49676998	0.928961031
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.991838945	82.18811638	0.91290657
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.991192052	88.70282073	0.907309111
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.985353979	147.4967189	0.881223745
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.991023877	90.39647882	0.925620227
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO PADRE'	0.980602829	195.3444625	0.873498438
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.983100818	170.1877861	0.881520042

Los datos de la Tabla 5.15 muestran condiciones parecidas al modelo de “*Random Forest*”, donde los valores de R^2 y MSE se mantienen en un lapso admisible, pero en el caso de $Kappa$ se ha visto una disminución del valor en este. Esto no parece afectar aún a las predicciones, algo identificable en la Figura 5.28, pero intrigante si se presiona aún más el modelo con una tasa de entrenamiento y prueba más limitada, lo cual se comprobará más adelante.

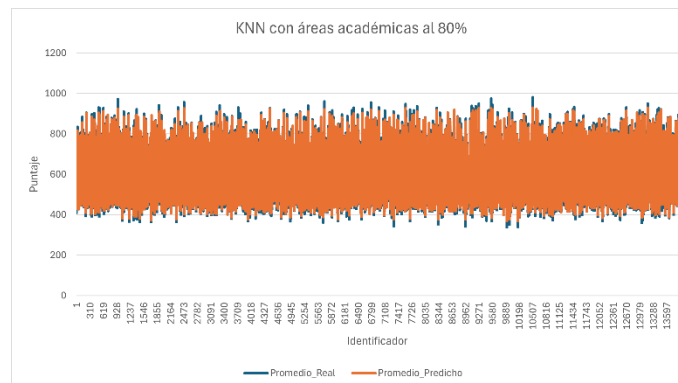


Figura 5.28. K-NN con variables académicas al 80%

5.2.6 Resultados al 99% de prueba

Para la finalización del entrenamiento de los modelos de aprendizaje supervisado se subió la tasa de prueba hasta un punto de exigencia lo suficientemente grande para probar los límites de los tres modelos, lo cual llevó a configurar este parámetro al 99%. Esto ha devuelto resultados sobresalientes, como se puede observar en la Tabla 5.16 relacionada al modelo de *regresión lineal* configurado con 10750 épocas, las cuales mantienen valores óptimos con respecto a sus métricas de evaluación y predicción de datos reflejados en la configuración de variables académicas y de estado de residencia, todo con un tiempo de ejecución de 3537700 ms.

Tabla 5.16. Métricas de evaluación de regresión lineal al 99% de prueba

10750 Épocas de procesamiento			
Configuración implementada	R ²	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.999603536	3.996099683	0.972904492
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.99960386	3.992833195	0.972823664
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.999604527	3.986113799	0.972823664
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.999604225	3.989153611	0.972905251
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.999605066	3.980680108	0.973146701
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.999604097	3.990450684	0.973308226
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO CAL'	0.999605209	3.979236949	0.972904696

En este punto de prueba, el modelo ha logrado mantener un comportamiento sobresaliente al tener métricas de evaluación satisfactorias a pesar de contar con parámetros de prueba estresantes, la Figura 5.29 muestra como la predicción de estos datos no dista tanto de la realidad, lo que también puede observarse en la Tabla 5.1, dando un panorama favorable sobre el mismo.

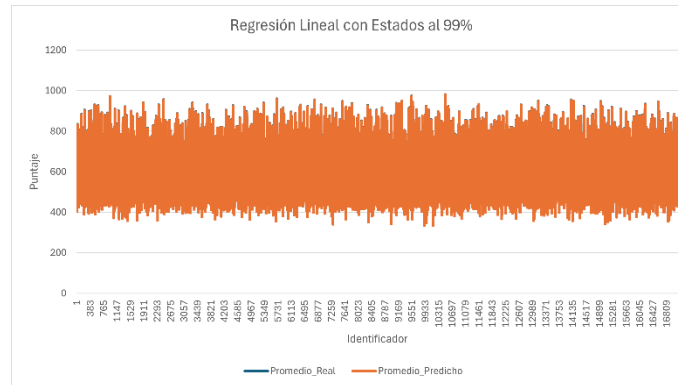


Figura 5.29. Regresión lineal con variables académicas y de estado de residencia al 99%

Los resultados del modelo de “*Random Forest*” a través de la tasa de prueba del 99% se mantienen en un rango aceptable con sus parámetros de 300 árboles pero con diferencias notables en comparación a otras etapas de entrenamiento, donde se presenta un escenario óptimo en la configuración de variables académicas y lenguas indígenas, con un tiempo de ejecución del 249.49 ms. y las métricas de evaluación presentadas en la Tabla 5.17.

Tabla 5.17. Métricas de evaluación de Random Forest al 99% de prueba

300 Árboles			
Configuración implementada	R ²	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.941013528	594.5458731	0.758935202
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.940757616	597.1253014	0.757295387
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.941280684	591.8531139	0.754032296
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.938381585	621.0741676	0.751408508
HP', 'Español', 'Mate', 'CN', 'CS', 'L INDIG', 'LENGUA INDIGENA', 'NO ESP'	0.94130089	591.6494483	0.758741305
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.938137531	623.5340829	0.751128936
HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.93913932	613.4366857	0.753260063

Los valores en este modelo son muy diferentes a los vistos en el modelo de *regresión lineal*, ya que mientras este último mantenía valores similares en todas sus iteraciones, para

“Random Forest” han cambiado en gran medida los valores de sus métricas de evaluación. Pues si bien los valores de R^2 se ubican en un rango admisible, el valor de MSE se ha disparado considerablemente a lo visto en la Tabla 5.14, algo que se aprecia incluso en el valor de $Kappa$, cuyo decrecimiento es muy notable en esta etapa y cuyas repercusiones a la predicción de este modelo se observan en la Figura 5.30, donde el margen de error es mucho más apreciable en comparación a instancias anteriores.

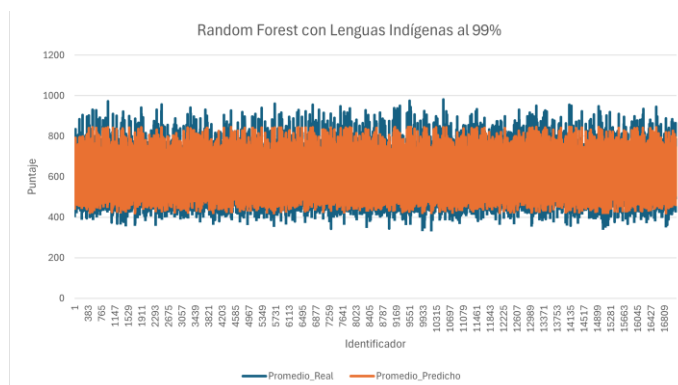


Figura 5.30. Random Forest con variables académicas y de lenguas indígenas al 99%

El último modelo entrenado con esta tasa de prueba, el “K-NN” presenta resultados destacados con sus parámetros a 5 vecinos cercanos, un número menor a las etapas anteriores, y sus resultados óptimos se presentan en la configuración de variables académicas la cual tuvo un tiempo de ejecución de 14.08 ms., sus métricas de evaluación y datos predichos pueden analizarse en la Tabla 5.18.

Tabla 5.18. Métricas de evaluación de K-NN al 99% de prueba

5 Vecinos			
Configuración implementada	R^2	MSE	Kappa
HP', 'Español', 'Mate', 'CN', 'CS'	0.958231966	420.9950364	0.806288957
HP', 'Español', 'Mate', 'CN', 'CS', 'Género (Masculino, Femenino)'	0.942818332	576.3546013	0.769575692
HP', 'Español', 'Mate', 'CN', 'CS', 'Aceptado'	0.940456787	600.1574662	0.753992508
HP', 'Español', 'Mate', 'CN', 'CS', 'NIVEL SOCIOECONOMICO', 'FAMILIA'	0.936663737	638.3889773	0.758378633
HP', 'Español', 'Mate', 'CN', 'CS', 'L_INDIG', 'LENGUA INDIGENA', 'NO_ESP'	0.948777695	516.288034	0.796269379
HP', 'Español', 'Mate', 'CN', 'CS', 'ESTADO', 'GRADO_MADRE', 'GRADO_PADRE'	0.926347716	742.3678714	0.732229502

HP', 'Español', 'Mate', 'CN', 'CS', 'SECTOR', 'ESTADO_ESC', 'INTERVALO_CAL'	0.924099349	765.0299704	0.741460498
--	-------------	-------------	-------------

Similar a lo visto anteriormente con “*Random Forest*”, en esta etapa de prueba el modelo de “*K-NN*” muestra un cambio considerable en sus métricas de evaluación, siendo la primera alerta detectada en *MSE*, con valores muy grandes comparados a los vistos en la Tabla 5.15, y en *Kappa* sucede lo mismo, una disminución de la calidad de los resultados predichos cuyo escenario se plasma en la Figura 5.31.

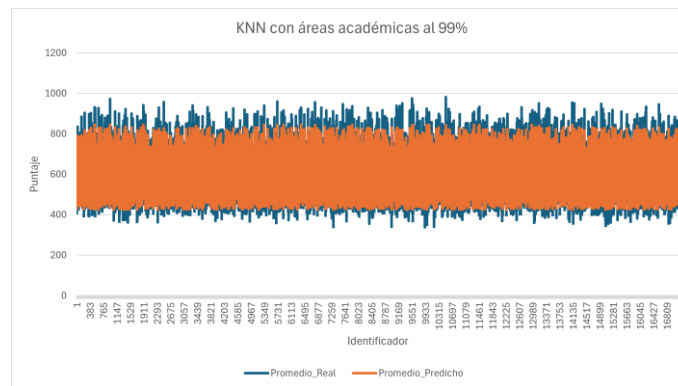


Figura 5.31. K-NN con variables académicas al 99%

5.3 Discusión de resultados

Los resultados mostrados anteriormente han planteado diferentes características de estos modelos en cuestión, los cuales explicaremos a continuación.

5.3.1 Interpretación de los resultados de regresión lineal

Los resultados del modelo de *regresión lineal* muestran una perspectiva clara del rendimiento de esta técnica, en la cual se puede examinar como los valores recuperados por las métricas de evaluación indican una eficacia relevante a través de las diferentes pruebas realizadas, un aspecto que se refleja considerablemente en las gráficas de los datos predichos por este modelo. Sin embargo, existe un dilema con respecto al desempeño, y ese es el tiempo de ejecución necesario para el entrenamiento y predicción de los datos incrementa considerablemente en cada una de las etapas, hasta llegar a un punto en que necesita demasiado tiempo y un gran número de épocas para poder hacer una predicción de todos los valores en total.

5.3.2 Interpretación de los resultados de Random Forest

Asimismo, los datos recopilados a través del modelo de “*Random Forest*” muestran un rendimiento notable en las configuraciones procesadas, las métricas de evaluación recopiladas muestran un comportamiento considerable en la mayoría de las etapas de entrenamiento y en un ajuste de parámetros consistente en el número de árboles, donde los valores óptimos registrados se obtuvieron con 300 árboles, demostrando que a mayor número de árboles ajustados la predicción mejora relativamente. No obstante, puede observarse un decremento en la eficacia del modelo cuando la tasa de prueba se incrementó a partir del 80%, donde los valores obtenidos bajaron de 0.90 en la métrica *Kappa* por primera vez en una de las pruebas en la configuración de variables académicas con *Aceptado*; posteriormente procedimos a observar de nuevo una caída en esta misma métrica y un incremento en los valores del *MSE*, demostrando que con un número bajo de datos para entrenar al modelo afectará negativamente a los resultados predichos por el mismo.

5.3.3 Interpretación de los resultados de K-NN

El rendimiento observado por el modelo de “*K-NN*” fue muy interesante, ya que este modelo logra mantener métricas de evaluación satisfactorias hasta una tasa de prueba del 80%. En cambio, cuando se prueban los límites del modelo “*K-NN*” al incrementar la tasa de prueba al 99% los valores en las métricas presentan un cambio radical al incrementar el *MSE* y caer notablemente la métrica *Kappa* y R^2 . Además de que durante las etapas de prueba del 20% al 60% se mantiene como ajuste para el modelo un total de 25 vecinos cercanos con resultados óptimos, pero esto cambia cuando la tasa sube al 80% donde el número de vecinos cercanos de 10 es el escenario con valores óptimos recibidos y posteriormente plasmar que al 99% de prueba, los valores óptimos se obtienen con un número de vecinos de 5. Lo que demuestra que en condiciones donde el número de datos de entrenamientos es bajo se necesita un número de vecinos bajo también.

5.3.4 Interpretación conjunta de los modelos desarrollados

Cabe recalcar que, así como los ajustes para el aprendizaje de cada uno ha afectado de diversas formas a los resultados producidos por los modelos de aprendizaje supervisado, tal como la tasa de prueba, las épocas en *regresión lineal*, el número de árboles en “*Random Forest*” y el número de vecinos en “*K-NN*”, un aspecto intrigante es como las variables independientes han presentado resultados óptimos para distintas implementaciones en ocasiones repetidas, el caso más claro se aprecia en el modelo de “*K-NN*”, cuyos resultados más sobresalientes se obtuvieron cuando las variables configuradas fueron sobre áreas académicas exclusivamente.

Por otra parte, los modelos de *regresión lineal* y “*Random Forest*” presentaron resultados óptimos en diversas configuraciones que involucraron a las áreas académicas con diferentes variables socioeconómicas, como las variables *Género*, *Aceptado*, las variables de lenguas indígenas y las de estado de residencia. Esto plantea un mejor desempeño por parte de estos modelos en comparación con el modelo de “*K-NN*”.

Esto se refuerza a través de los mapeos realizados con los modelos de aprendizaje no supervisado, que de forma general nos permiten identificar el comportamiento de las configuraciones probadas y determinar junto a la predicción de datos si existe una correlación de datos, ya que de saturar estos modelos con demasiadas variables independientes los resultados obtenidos no serían precisos.

Capítulo 6. Conclusiones

A través de este trabajo de tesis se exploraron los parámetros que formaron parte del desempeño de una población estudiantil para su ingreso a una educación de nivel medio-superior, si bien la mayoría de los trabajos que suelen explorar este tipo de situaciones suelen enfocarse en los aspectos académicos, en esta ocasión hemos considerado aspecto extraacadémicos que forman parte de la vida de estos estudiantes y que pueden afectar no solo su desempeño en la búsqueda de una mejor educación, sino en la búsqueda de un mejor futuro.

Los resultados obtenidos con “*K-Means*” y “*DBSCAN*” nos permitieron identificar como la población de estudiantes analizada presenta ciertas destrezas en algunas áreas académicas, como es el caso de los hombres y mujeres al comprobarlo con la variable de géneros. Sin embargo, estas aptitudes pueden verse reflejadas en los estudiantes locales y foráneos al procesar variables de lugar de procedencia, estado de residencia; las destrezas mostradas por los estudiantes que hacen uso de manejo de lenguas indígenas, incluso como los estudiantes que fallaron en sus pruebas estandarizadas presentaron algunos aciertos importantes.

En el caso de las técnicas de aprendizaje supervisado hemos visto como cada uno de los modelos desarrollados presenta destrezas y desventajas en diferentes ámbitos para generar un resultado satisfactorio, modelos como “*Random Forest*” y “*K-NN*” presentan una gran ventaja con respecto a las variables, sobre todo en la cuestión de los tiempos de ejecución para cantidades de datos de entrenamiento moderadas. No obstante, el modelo de *regresión lineal* ha presentado resultados extraordinarios incluso con una cantidad mínima de datos para entrenar, pero lo ha conseguido a un costo computacional muy alto al tardar más de una hora para el procesamiento y predicción completa de los datos, por lo que hemos concluido en que el uso de estos modelos debe ser planeado según las necesidades del usuario, un resultado inmediato y general producido por “*K-NN*” o “*Random Forest*”, siempre y cuando exista una cantidad suficiente de datos de entrenamiento; o un análisis y predicción más profundo y preciso que permita el uso mínimo de datos para entrenar como lo hace *regresión lineal*.

Capítulo 7. Trabajos futuros

Los planteamientos descubiertos en este trabajo de tesis pueden implementarse en trabajos futuros, como por ejemplo, identificar cuáles son las debilidades académicas de los aspirantes al nivel medio-superior en la actualidad y buscar la manera de apoyarlos para una mejora en su desempeño académico a través de la monitorización de estos con el desarrollo de un programa o aplicación web. Incluso este tipo de conocimientos pueden ser aplicados a otros niveles de la educación con otros datos o incluso más técnicas de aprendizaje automático con enfoques de solución diferentes a los planteados en este trabajo.

Como se mencionó en los resultados de “*K-Means*”, estas técnicas de aprendizaje no supervisado permiten tener un panorama más amplio que puede reforzarse a través de un análisis descriptivo de cada clúster generado, a través del cálculo de las medias y desviaciones estándar de cada grupo. A pesar de que en este trabajo no se profundizó en este aspecto, es una propuesta interesante a considerar en futuras investigaciones, en las cuales pueda implementarse el uso de etiquetas interpretativas (por ejemplo, “Alto rendimiento”, “Riesgo medio”, “Riesgo alto”, etc.). Incluso se puede pensar en el uso de “*boxplots*” (diagramas de caja) para una mejor observación de la varianza de las las calificaciones y el puntaje estandarizado entre clústeres, lo cual plantea argumentos y cuestiones intrigantes para aspectos completamente diferentes de un escenario.

Una de las políticas actuales es el cero rechazo por parte de las instituciones educativas y los modelos desarrollados e implementados en este trabajo permiten predecir el desempeño de los estudiantes utilizando datos académicos históricos, lo que ayuda a detectar aquellos estudiantes que traen un bajo rendimiento. Esto permite diseñar con previa antelación estrategias de apoyo que ayuden al estudiante a terminar satisfactoriamente sus estudios y continuar con su preparación profesional.

Referencias

- [1] Viera, Á. F. G. (2017). Técnicas de aprendizaje de máquina utilizadas para la minería de texto. *Investigación Bibliotecológica Archivonomía Bibliotecología E Información*, 31(71), 103-126. <https://doi.org/10.22201/iibi.0187358xp.2017.71.57812>
- [2] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. En Springer eBooks. <https://dl.acm.org/citation.cfm?id=1162264>
- [3] Hastie, T., Tibshirani, R. J., & Friedman, J. (2013). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. <http://catalog.lib.kyushu-u.ac.jp/ja/recordID/1416361>
- [4] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*. En Morgan Kaufmann Publishers Inc. eBooks. <https://dl.acm.org/citation.cfm?id=3086818>
- [5] Jing, Y. (2023). *Design of English Assisted Learning System Based on Deep Learning and Data Analysis Techniques*. IEEE. <https://doi.org/10.1109/csmis60634.2023.00063>
- [6] Wang, X., & Bai, X. (2023). *The Evaluation Model of University Education Level Based on Deep Learning Algorithms*. IEEE. <https://doi.org/10.1109/icaisc58445.2023.10200092>
- [7] Qian, W. (2020). *Analysis on the Application of Deep Learning Technology in the Field of English Education Big Data Mining*. IEEE. <https://doi.org/10.1109/icbdss51270.2020.00022>
- [8] Yang, H., & Wu, H. (2019). *The Application of SPOC-Based Deep Learning Model in Psychological Health Education of College Students in Post-MOOC Era*. IEEE. <https://doi.org/10.1109/itme.2019.00059>
- [9] Yu, K., Yuan, T., & Li, Y. (2021). *Application of Data Mining Technology in Sports Data Analysis in Colleges and Universities*. IEEE. <https://doi.org/10.1109/tcs52929.2021.00073>
- [10] Chiok, C. H. M. (2017). Predicción del rendimiento académico aplicando técnicas de minería de datos. *Anales Científicos*, 78(1), 26. <https://doi.org/10.21704/ac.v78i1.811>
- [11] Pérez, P. M., Vicente, M. P. C., & Costa, J. L. C. (2008). Predicción del rendimiento escolar de los alumnos a partir de las aptitudes, el autoconcepto

- académico y las atribuciones causales. *Horizontes Educativos*, 13(2), 11-23.
<https://dialnet.unirioja.es/descarga/articulo/3987524.pdf>
- [12] Rojanavas, P. (2019). Educational Data Analytics using Association Rule Mining and Classification. The 4th International Conference On Digital Arts, Media And Technology And 2nd ECTI Northern Section Conference On Electrical, Electronics, Computer And Telecommunications Engineering.
<https://doi.org/10.1109/ecti-ncon.2019.8692274>
- [13] Varade, R. V., & Thankanchan, B. (2021). Efficient Data Mining Techniques to Improve Academic Performance of Students. *SAMRIDDHI A Journal Of Physical Sciences Engineering And Technology*, 13(SUP 1), 81-84.
<https://doi.org/10.18090/samriddhi.v13is1.18>
- [14] M, R., F, N., & A, A. (2018). Predicting and Analysis of Students' Academic Performance using Data Mining Techniques. *International Journal Of Computer Applications*, 182(32), 1-6. <https://doi.org/10.5120/ijca2018918250>
- [15] Meghji, A. F., Mahoto, N. A., Unar, M. A., & Shaikh, M. A. (2019). Predicting Student Academic Performance using Data Generated in Higher Educational Institutes. *3C Tecnología_Glosas de Innovación Aplicadas A la Pyme*, 366-383.
<https://doi.org/10.17993/3ctecno.2019.specialissue2.366-383>
- [16] Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. <http://cds.cern.ch/record/2229831>
- [17] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical learning: with applications in Python. En Springer texts in statistics. <https://doi.org/10.1007/978-3-031-38747-0>
- [18] Géron, A. (s. f.). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. En O'Reilly Online Learning (2nd ed.). <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>
- [19] Aggarwal, C. C., & Reddy, C. K. (2018). Data Clustering: Algorithms and Applications. En Chapman and Hall/CRC eBooks (1st ed.). <https://doi.org/10.1201/9781315373515>
- [20] Géron, A. (s. f.). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. En O'Reilly Online Learning (2nd ed.). <https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>

- [21] Piatetsky-Shapiro, G., & Frawley, W. (1991). Knowledge Discovery in Databases. MIT Press.
- [22] Al-Razgan, M., Al-Khalifa, A. S., & Al-Khalifa, H. S. (2013). Educational Data Mining: A Systematic Review of the Published Literature 2006-2013. Lecture Notes In Electrical Engineering, 711-719. https://doi.org/10.1007/978-981-4585-18-7_80
- [23] Durairaj, M., & Vijitha, C. (2014). Educational Data mining for Prediction of Student Performance Using Clustering Algorithms. International Journal Of Computer Science And Information Technologies, 5(4), 5987-5991.
- [24] Zheng, A., & Casari, A. (2018). Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists. O'Reilly Media.
- [25] Juárez, R. C., Garnica, C. C., Sierra, E. A., & Torres, K. J. R. (2024). Técnicas de clasificación para predecir el desempeño de los estudiantes en pruebas estandarizadas. EDUCATECONCIENCIA, 32(03). <https://doi.org/10.58299/edutec.v32i2.793>
- [26] Gryshova, I., Zamlynskyi, V., & Shestakovska, T. (2017). IMPLEMENTING CLUSTER FORMS INTO NATIONAL EDUCATION DEVELOPMENT STRATEGY. Science And Education, 25(5), 56-61. <https://doi.org/10.24195/2414-4665-2017-5-11>
- [27] Barasheva, E., & Zedgenizova, I. (2019). Interaction of innovation clusters, science and education. □The □European Proceedings Of Social & Behavioural Sciences, 869-874. <https://doi.org/10.15405/epsbs.2019.03.02.100>
- [28] Kovaliuk, T., & Kobets, N. (2020). Cluster Model of IT Education in Ukraine. 2020 IEEE EUROPEAN TECHNOLOGY AND ENGINEERING MANAGEMENT SUMMIT (E-TEMS 2020), 1-6. <https://doi.org/10.1109/e-tems46250.2020.9111732>
- [29] Maślankowski, J., & Brzezicki, Ł. (2022). Evaluation of Machine Learning Methods for the Experimental Classification and Clustering of Higher Education Institutions. En Lecture notes in business information processing (pp. 31-45). https://doi.org/10.1007/978-3-030-95947-0_3
- [30] Pektaş, Ş. T. (2023). A Systematic Analysis of Machine Learning Studies in Education. PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON EDUCATION TECHNOLOGY AND COMPUTERS, ICETC 2023, 15, 451-455. <https://doi.org/10.1145/3629296.3629368>

- [31] Talib, N. I. M., Majid, N. A. A., & Sahran, S. (2023). Identification of Student Behavioral Patterns in Higher Education Using K-Means Clustering and Support Vector Machine. *Applied Sciences*, 13(5), 3267. <https://doi.org/10.3390/app13053267>
- [32] Li, M., Guo, S., & Chen, J. (2025). Construction of aesthetic education evaluation model based on K-means clustering analysis in the context of artificial intelligence. *Journal Of Computational Methods In Sciences And Engineering*. <https://doi.org/10.1177/14727978251321946>
- [33] Chinnadurai, J., Karthik, A., Ramesh, J. V. N., Banerjee, S., Rajlakshmi, P., Rao, K. V., Sudarvizhi, D., & Rajaram, A. (2024). Enhancing online education recommendations through clustering-driven deep learning. *Biomedical Signal Processing And Control*, 97, 106669. <https://doi.org/10.1016/j.bspc.2024.106669>
- [34] Shen, Y., & Lei, C. (2024). Research on evaluation of university education informatization level based on clustering technique. *Heliyon*, 10(4), e25215. <https://doi.org/10.1016/j.heliyon.2024.e25215>
- [35] Liu, B. (2024). Research on the Application of Clustering Algorithm for Online Education Students Based on Artificial Intelligence. *PROCEEDINGS OF INTERNATIONAL CONFERENCE ON ALGORITHMS, SOFTWARE ENGINEERING, AND NETWORK SECURITY, ASENS 2024*, 84-89. <https://doi.org/10.1145/3677182.3677199>
- [36] Nafuri, A. F. M., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., & Aliff, M. (2022). Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Applied Sciences*, 12(19), 9467. <https://doi.org/10.3390/app12199467>
- [37] Tuyishimire, E., Mabuto, W., Gatabazi, P., & Bayisingize, S. (2022). Detecting Learning Patterns in Tertiary Education Using K-Means Clustering. *Information*, 13(2), 94. <https://doi.org/10.3390/info13020094>
- [38] Harrington, P. (2012). *Machine Learning in Action*. Manning.
- [39] Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2020). Data mining for business analytics: Concepts, techniques, and applications in R. Wiley
- [40] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R
- [41] Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>