Análisis de datos académicos de estudiantes

Proyecto Final - Minería de datos

noviembre de 2020

Andrés Leonardo Moreno Romero
andres.moreno-r@mail.escuelaing.edu.co

Esteban Guerrero Russi
esteban.guerrrero-r@mail.escuelaing.edu.co

1. Introducción

Se dice que el desempeño académico de un estudiante durante la formación básica y media puede relacionarse con aspectos demográficos, académicos, familiares y comportamentales de la persona. Se cree que estos últimos pueden tener una estrecha relación con el desempeño académico de un estudiante, sin embargo, estos no son estudiados con mucha frecuencia en la actualidad. Este trabajo tiene como objetivo identificar parámetros y descubrir información relevante de un dataset con información académica recolectada de estudiantes de diferentes nacionalidades, buscando con esto predecir el desempeño de un estudiante en determinada asignatura mediante la implementación de modelos de clasificación supervisada como Logistic regression, Artificial Neural Networks, K-Nearest Neighbors y clasificación no supervisada como clustering. Estos modelos son desarrollados luego de haber ejecutado una serie de pasos de análisis y preprocesamiento de datos.
Al identificar relaciones y patrones entre las variables del dataset, también se podrían desarrollar a futuro estrategias apropiadas que permitan a los estudiantes obtener mejores resultados académicos. Este trabajo tiene como resultado modelos de clasificación con precisiones superiores al 70% y dentro de sus hallazgos se encuentra que existe una estrecha relación entre algunos parámetros como cantidad de participaciones y fuentes visitadas, con el desempeño académico de un estudiante en determinada asignatura.

2. Trabajo relacionado

En los últimos años, se han desarrollado varios estudios en el campo de minería de datos con el fin de mejorar métodos de enseñanza y tomar mejores decisiones a nivel educacional mediante el estudio de diferentes factores que pueden afectar el desempeño de un estudiante. Los autores en sus dos publicaciones estudian y analizan la información académica del dataset educacional recopilado del sistema de e-learning, Kallboard 360, mediante un X-API,[1] con el fin de predecir el desempeño de un estudiante en una asignatura teniendo como base otra información relacionada con su origen, hábitos y comportamientos en clase. Mediante técnicas de clasificación como Naïve Bayes, Decision Tree y Artificial Neural Network los autores logran desarrollar modelos con una precisión de hasta el 73.9% que predicen el desempeño del estudiante teniendo en cuenta aspectos comportamentales y de hasta 56.2% de precisión para modelos que no consideran aspectos comportamentales [2]. Posterior al desarrollo de estos modelos, los autores aplican algoritmos de conjunto como Bagging Boosting y Random Forest para mejorar el rendimiento de los modelos iniciales. Dentro de los hallazgos más importantes de estos trabajos se tiene que existe una estrecha relación entre algunas variables comportamentales como la cantidad de veces que un estudiante participa en clase o los recursos visitados, con el desempeño final en la asignatura.

3. Definición del problema y algoritmo

3.1 Problema

La documentación y recolección de datos académicos ha crecido considerablemente en los últimos años, por lo que descubrir conocimiento en el campo educacional se ha convertido en un gran reto de minería de datos. En el dataset que se estudia para este trabajo se tiene información demográfica, académica, y comportamental de estudiantes (entradas) y mediante la implementación de técnicas y algoritmos de minería de datos se quiere encontrar relaciones entre parámetros, predecir el desempeño de un estudiante en determinada asignatura y plantear estrategias que podrían mejorar los resultados académicos de los estudiantes en el futuro (salidas).

3.2 Algoritmo

K - Nearest Neighbor (KNN)

El clasificador de vecino más cercano, conocido por sus siglas en inglés KNN, es uno de los métodos de clasificación más simples, pero más usados en la actualidad debido a sus numerosas aplicaciones. Cuando a este algoritmo se le asigna un conjunto de datos para clasificar, busca el elemento más similar (más cercano) dentro del conjunto de datos de entrenamiento y le asigna la etiqueta de su clase. La siguiente ilustración expresa este funcionamiento.

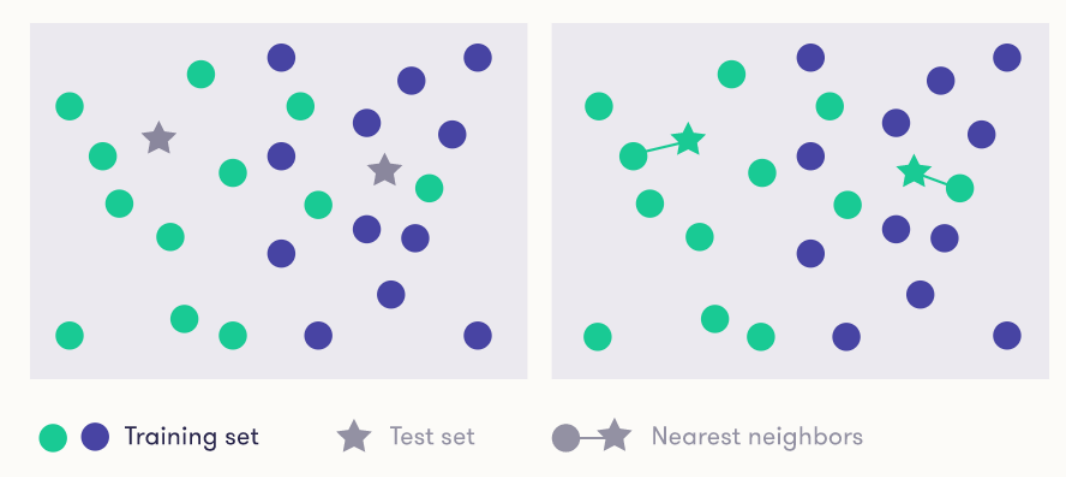


Ilustración 1. Vecino más cercano

En este ejemplo, se tienen elementos de un conjunto de datos de entrenamiento que pertenecen a una clase (verde o azul) y dos elementos de datos de prueba que son representados por estrellas. Los dos elementos de prueba se clasifican en la clase "verde" porque sus vecinos más cercanos son ambos verdes.

Para cuantificar que tan cerca está un elemento de otro se requiere usar una métrica de distancia y según el caso en estudio, el método de cálculo puede variar. En muchas ocasiones, cuando se representan los elementos del dataset en un plano, se calcula la distancia geométrica entre estos con el fin de asignar la

clase, esta distancia se conoce como euclidiana. Sin embargo, este método puede variar según se requiera; para este trabajo se usaron tres distancias diferentes (Euclidiana, Manhattan y Chebyshev) con el fin de evaluar posibles diferencias en los resultados y elegir aquel método que mejore la exactitud del modelo.

Otro elemento que es importante considerar y que puede afectar los resultados del modelo es la cantidad de vecinos (K) que se usan para determinar la clase de un elemento utilizando el algoritmo. No existe un número de vecinos específico que sea ideal, K podría variar dependiendo de los datos en estudio, de la distancia utilizada, entre otros factores.

Debido a que el algoritmo tiene como base el cálculo de distancias, es necesario que todos los elementos dentro del dataset sean de tipo número para poder realizar dicho calculo.

Logistic Regression

Este método de clasificación surge de adaptar el método de regresión lineal para obtener resultados representados en clases y no en valores numéricos. Tanto la regresión lineal como logística, calculan una función lineal de las entradas, la diferencia radica en que en la regresión logística las salidas se calculan pasando por una función sigmoide, cuya función es convertir cualquier número en un valor entre 0 y 1.

$$S(z) = \frac{1}{1 + e^{(-z)}}$$

Fórmula matemática función sigmoide

Además de predecir la clase, esta técnica permite obtener un valor de incertidumbre de la predicción. Dado que el valor de salida es un número entre 0 y 1, se puede interpretar como una probabilidad. Esto es útil porque se puede usar la regresión logística para predecir la probabilidad de que algo suceda, en lugar de usar un valor numérico.

En educación, al recopilar datos de estudiantes se puede tener disponible información básica como nombre, ID, edad, etc., e información relacionada con comportamientos como horas de estudio o cantidad de participaciones durante las clases. En un grupo de estudiantes se podrían analizar la cantidad de horas estudiadas frente al resultado final del examen de la asignatura (Aprobar/No abrobar). Teniendo suficientes datos, sería posible predecir el resultado en el examen de un estudiante sabiendo la cantidad de horas que dedicó a estudiar o de igual manera, encontrar el número de horas que se requieren para poder aprobar el examen.

Para el data set utilizado en este trabajo, el método de regresión logística tiene variación, en lugar de hacer clasificación en dos grupos (0 y 1) lo hace en tres debido al número de clases que hay en este problema. También, al tratarse de un método que utiliza una fórmula matemática para hallar la función adecuada, es necesario que todas las variables del dataset sean de tipo numérico.

Artificial Neural Networks (ANN)

Los algoritmos de redes neuronales no son mucho más complejos que otros, ya que se mantiene la idea de combinar una función lineal, con una no lineal, como la sigmoide. Lo que las redes neuronales permiten, es conectar múltiples modelos para poder formar una red. La parte no lineal del modelo se denomina función de activación y cada uno de esos modelos es conocido como una neurona. Las neuronas están conectadas entre sí permitiendo que la salida de una neurona sea la entrada de otra.

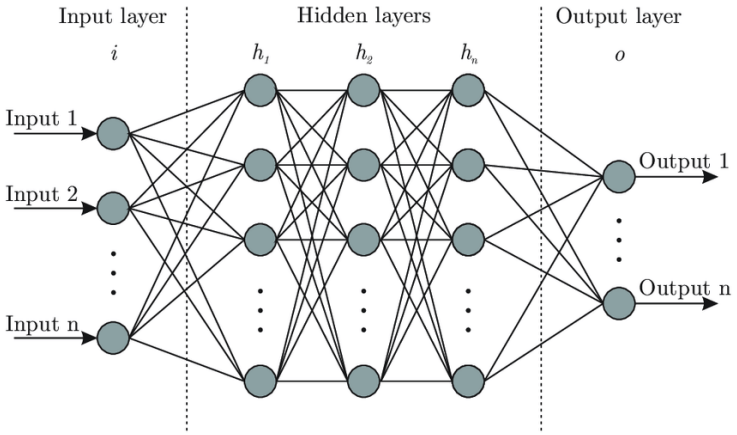


Ilustración 2. Estructura de una red neuronal

Observando la red neuronal anterior se tienen tres nodos de entrada *i* formando la capa de entrada o input layer, los nodos *h* en este caso forman las capas ocultas o hidden layers, y finalmente se tienen dos nodos de salida *o* que conforman la capa de salida u output layer.

La capa de entrada consta de neuronas que obtienen sus entradas directamente de los datos. Las capas ocultas usan las salidas de otras neuronas como entrada, y sus salidas se usan como entradas para otras capas de neuronas, este proceso se da gracias a las funciones de activación, las cuales determinan si se genera la salida de una neurona o no. Finalmente, la capa de salida produce la salida de toda la red, y de manera probabilística se elige una de estas neuronas como el resultado de la predicción de la clase. Todas las neuronas de una capa determinada obtienen entradas de las neuronas de la capa anterior y alimentan con sus salidas a la siguiente.

El hacer uso de redes neuronales tiene una ventaja respecto a los demás algoritmos, estas redes son capaces de aprender durante el proceso y van mejorando su rendimiento con el propósito de minimizar el error.

4. Evaluación Experimental

4.1 Datos

El dataset contiene atributos de los estudiantes que se pueden dividir en cuatro categorías: demográficos, académicos, participación de los padres y comportamiento. A continuación, se muestran todos los atributos y su respectiva descripción.

Categoría	Atributo	Descripción
Demográficos	Nationality (Nacionalidad)	Nacionalidad del estudiante
	Gender (Genero)	Genero del estudiante (Masculino o femenino)
	Place of Birth (Lugar de nacimiento)	Lugar de nacimiento del estudiante
	Parent responsible for student (Acudiente)	Si el acudiente es el padre o la madre
Académicos	Educational Stages (Grado de educación)	Nivel de educación al que pertenece el estudiante (primaria, media vocacional o bachillerato)
	Grade Levels (Grado)	Grado al que pertenece el estudiante va desde primero hasta grado 12
	Section ID (Curso)	Curso al que pertenece el estudiante (A, B o C)
	Semester (Semestre)	Semestre del año (Primero o segundo)
	Topic (Materia)	Materia en la que fue tomado el desempeño del estudiante.
	Student Absence Days (Ausencias)	Ausencias del estudiante a clase (mayor a 7 veces o menor a 7 veces)
Participación de los padres	Parent Answering Survey (Encuesta)	Si los padres respondieron a la encuesta enviada por la institución o no
	Parent School Satisfaction (Satisfacción de los padres)	Grado de satisfacción de los padres con la institución
Comportamiento en clase	Discussion groups (Foros)	Numero de veces que realizó esta actividad, medido en la plataforma Kalboard 360 e-learning.
	Visited resources (Recursos visitados)	
	Raised hand on class (Levantar la mano)	
	Viewing announcements (Revisar anuncios)	

A continuación, se presenta el análisis gráfico y estadístico de algunas variables del dataset.

Género:

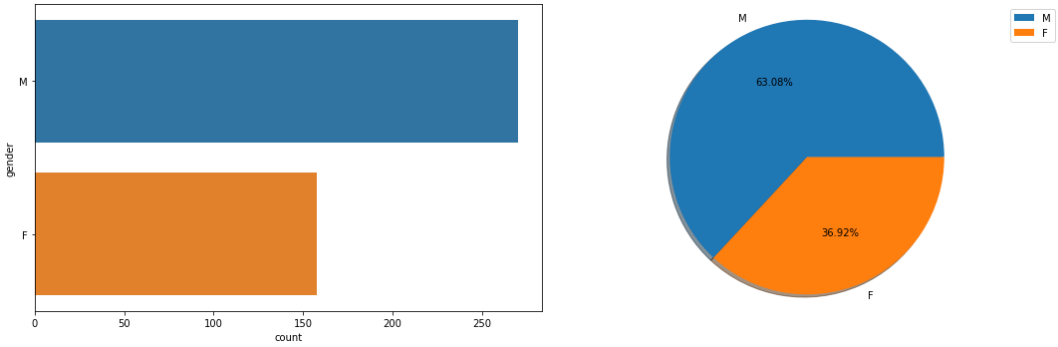


Ilustración 3. Género

Se puede observar que pedominan la cantidad de hombres en los registros del dataset, que representan en total el 63.08% del total, frente a 36.92% que corresponden a mujeres.

Grado:

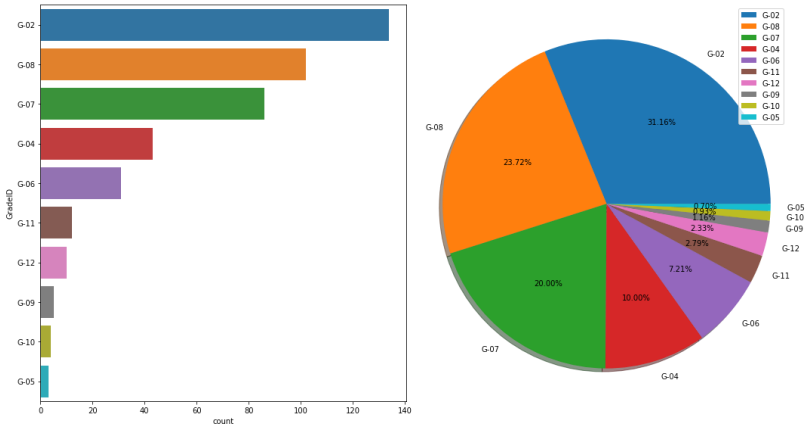


Ilustración 4. Grado

En este caso, se puede observar que predominan los registros de estudiantes de segundo grado (31.16%) seguido de estudiantes de octavo y séptimo grado (23.72% y 20% respectivamente).

Topic:

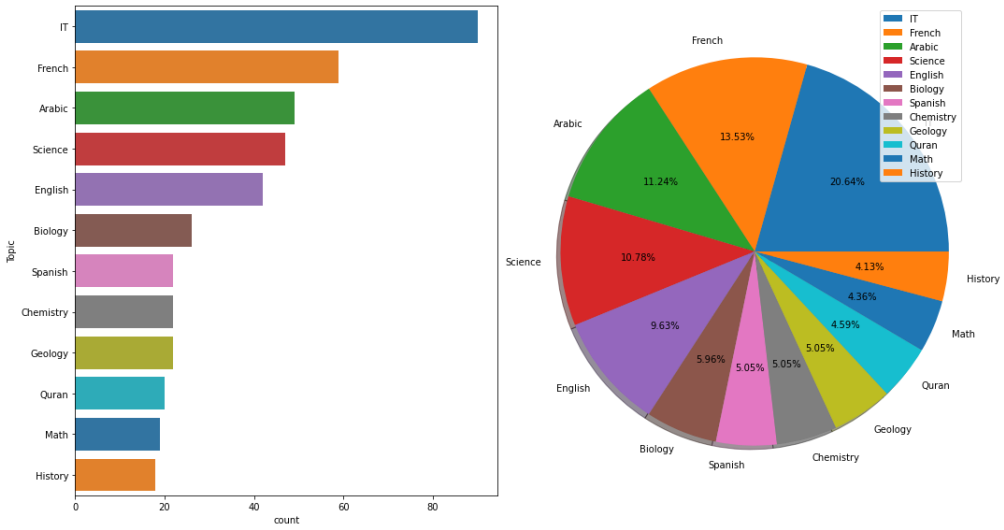


Ilustración 5. Topic

Topic hace referencia al nombre o temática de la clase a la que corresponde el registro del rendimiento del estudiante. Existen 12 cursos, donde el mayor número de registros se ven para la asignatura IT, seguido del cuso de francés y en tercer lugar el curso de árabe. El curso de IT supera los 80 registros (20.64%). Por su parte, los cursos de español, biología, química, geología, Quran, matemáticas e historia redondean los 20 registros, es decir entre 4% y 5% del total aproximadamente.

Participaciones (levantar la mano):

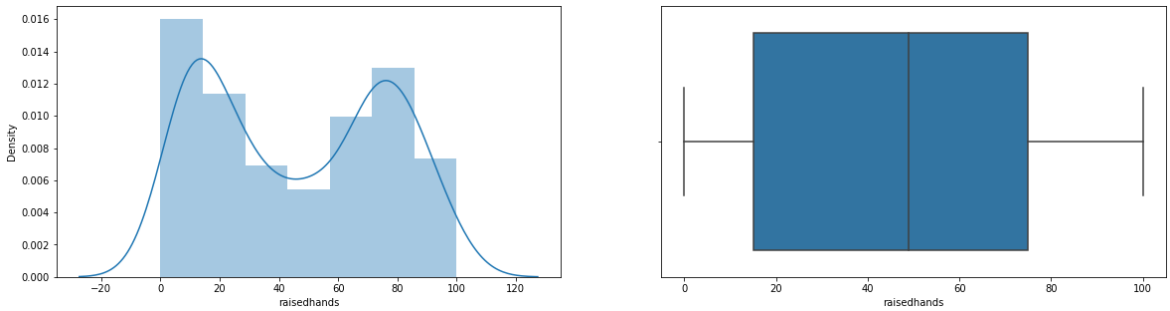


Ilustración 6. Participaciones

Este atributo hace referencia a la cantidad de veces que un estudiante alzó la mano para participar en el total de las clases. Se encuentra en un rango entre 0 y 100, con una media 46.2 veces.

Revisión de nuevos anuncios:

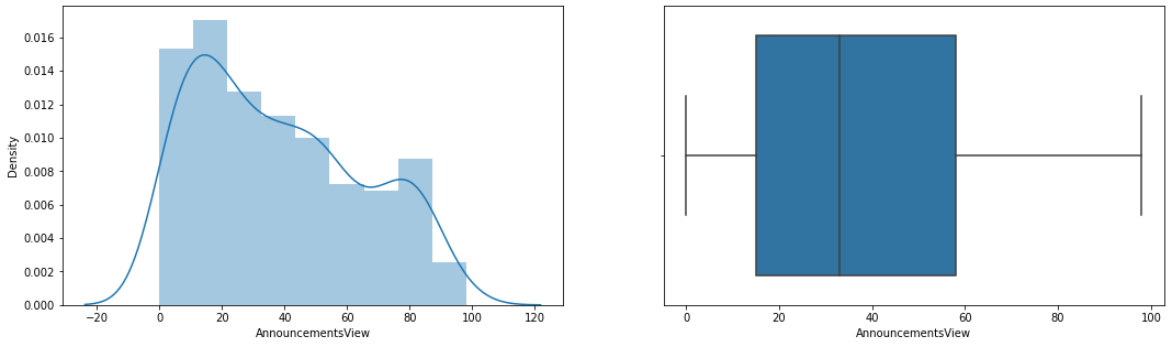


Ilustración 7. Revisión de nuevos anuncios

Esta variable registra la cantidad de veces que un estudiante revisa los anuncios nuevos que se hacen. Los valores se encuentran en un rango entre 0 y 98, con una media de 38.23. La mayor densidad de registros está entre 15 y 58 veces.

Días de ausencia estudiantil:

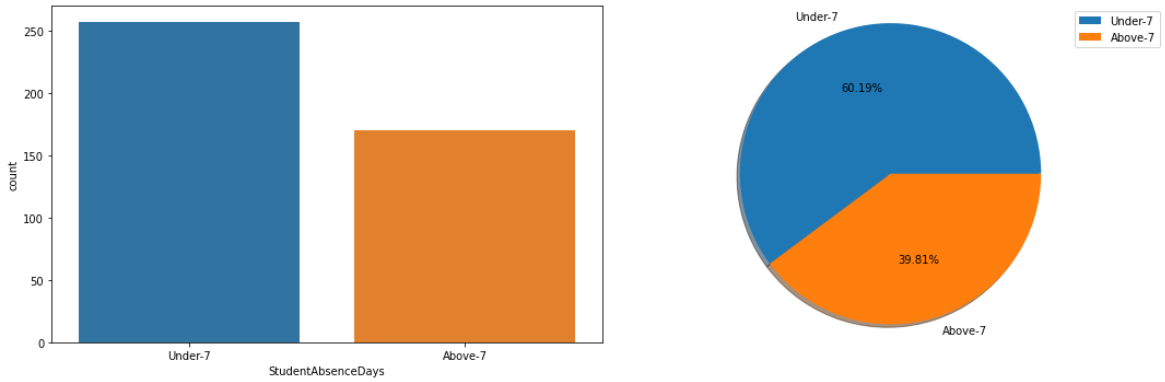


Ilustración 8. Días de ausencia estudiantil

Esta variable muestra sin un estudiante estuvo ausente durante más de 7 días o no. El 60.19% de los estudiantes indica no haber estado ausente por más de 7 dias y el 39.81% indica haber estado ausente por más de 7 días.

Clase:

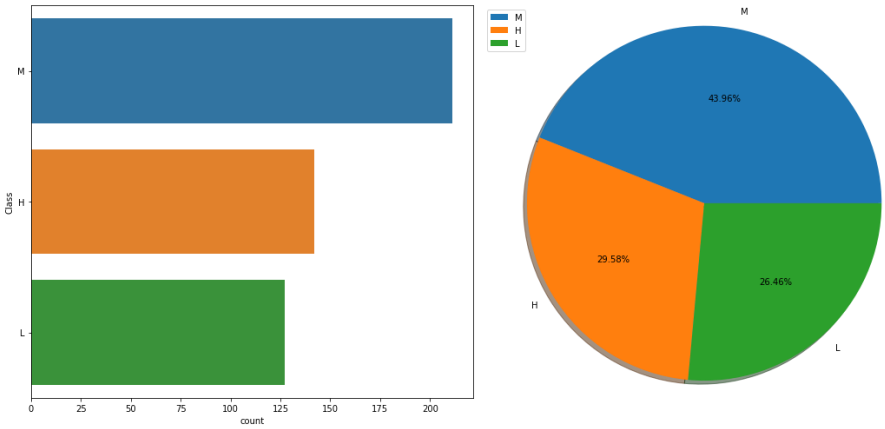


Ilustración 9. Clase

Finalmente se analiza la clase, que corresponde al rendimiento del estudiante en la asignatura, donde L(low) es rendimiento bajo, (M) Medium es rendimiento promedio y H(High) es rendimiento alto. Se puede observar que el 43.96% tiene un rendimiento promedio, el 29.58% un rendimiento alto y el 26.46% un rendimiento bajo.

Reducción dimensionalidad:

Al hacer un análisis de los datos, se encuentra que las variables “Nacionalidad” y “Lugar de Nacimiento” se comportan de manera similar, por lo que se decide eliminar una de ellas (Lugar de Nacimiento) para reducir dimensionalidad.

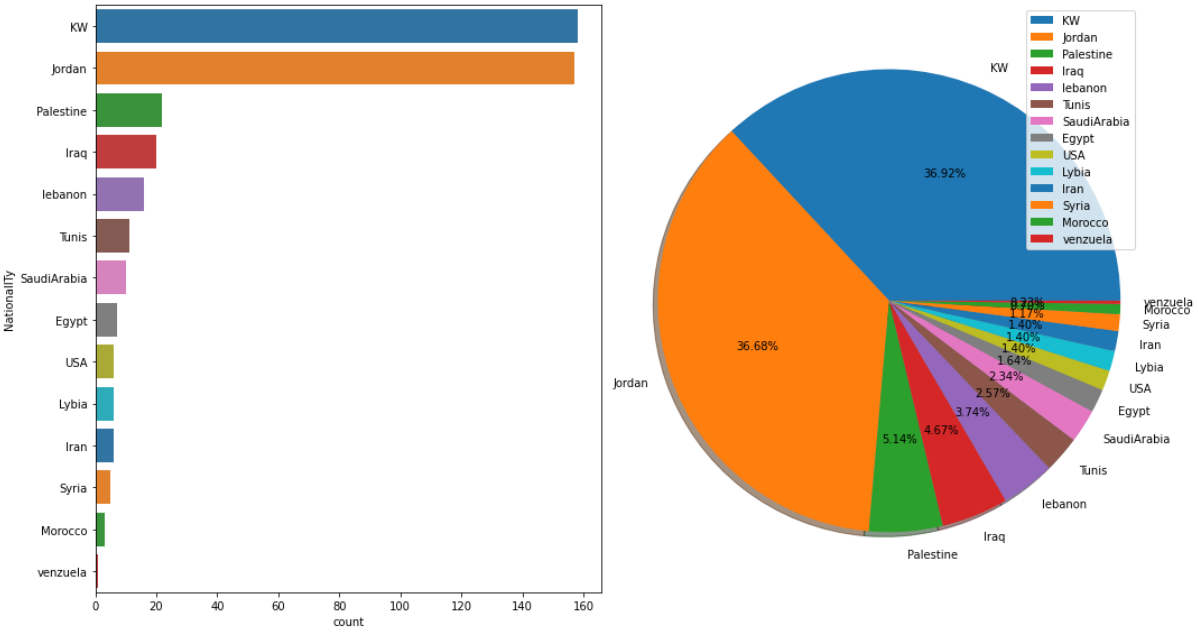


Ilustración 10. Lugar de nacimiento

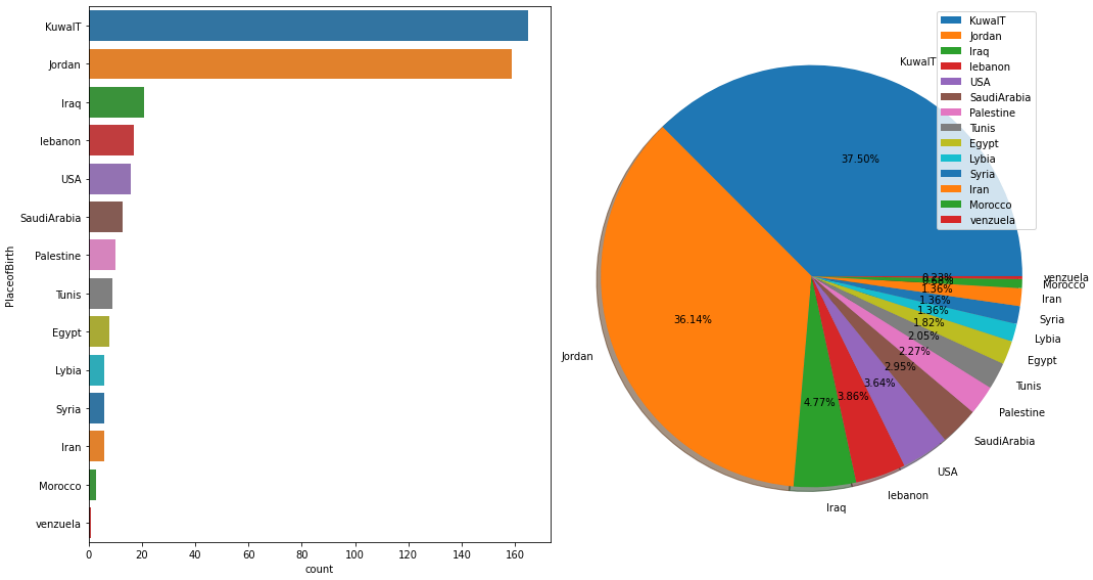


Ilustración 11. Nacionalidad

Imputación

Debido a que el dataset contiene valores faltantes, se realizó la imputación de dichos valores. Para hacer este proceso se utilizaron las técnicas: Imputación por la moda e imputación con el algoritmo KNN con K = 5 y 10.

Imputación por la moda

La técnica de imputación por la moda consiste en reemplazar los valores faltantes por el dato que más se repite dependiendo de la columna en la que se encuentre.

Imputación KNN

La técnica de imputación K - Nearest Neighbor funciona de manera similar a la técnica de clasificación con el mismo nombre (explicada anteriormente). En este caso, el algoritmo mide la distancia entre todos los atributos y reemplaza el valor nulo por el valor más “cercano” teniendo en cuenta el número de K.

A continuación, se encuentra la evaluación de los tres métodos.

Dataset	Exactitud
Moda	93.932%
KNN5	93.333%
KNN10	93.216%

Tabla 1. Evaluación métodos de imputación.

Se puede evidenciar que el mejor método fue el de imputación por la moda.

Normalización

Como se puedo evidenciar en el análisis de los datos, todos tienen unidades de medidas diferentes y escalas diferentes, por este motivo se deciden normalizar todos los datos para que todos los atributos estén en una misma escala y una única dimensión. Al hacer esto, se elimina la posibilidad de que los algoritmos le den más peso a las variables que tienen una escala mayor.

La técnica de normalización consiste en dejar todos los valores en un rango entre 0 y 1 haciendo uso de los valores máximos y mínimos de cada columna. La fórmula que utiliza el método es la siguiente:

$$X_{normalizado} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4.2 Metodología

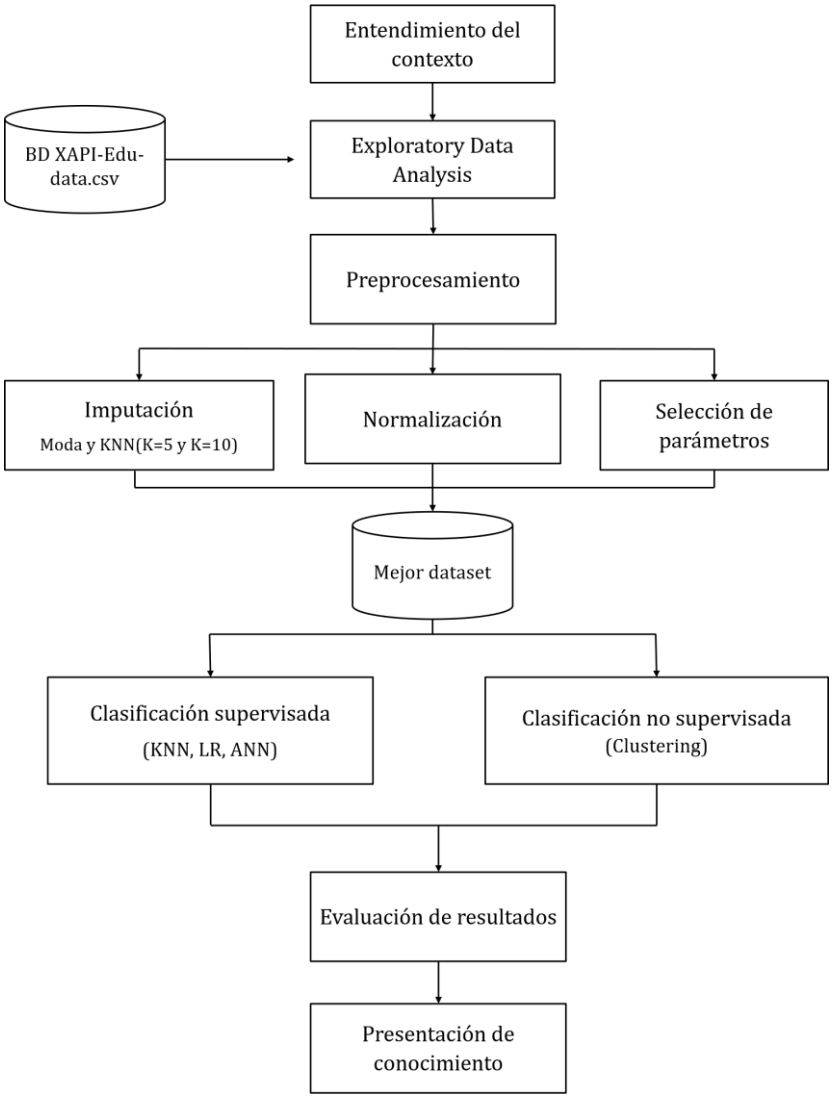


Ilustración 12. Flujo de trabajo

Para el desarrollo de este trabajo se siguieron los pasos de la metodología que se muestra en el flujo de trabajo de la *ilustración 12*, cuyos pasos principales se basan en la metodología CRISP-DM. Inicialmente se hace un entendimiento del entorno y del contexto en el que se documentan los datos, de donde se sabe que los datos son capturados del sistema Kalboard 360 E-learning usando la API (XAPI). Luego de esto se procede a entender la base de datos mediante EDA, donde se estudia cada uno de los atributos del dataset *XAPI-Edu-data.csv* mediante la visualización de gráficos y análisis estadístico. Posteriormente se procede a realizar el preprocesamiento de datos, donde se aplican técnicas de imputación, normalización y donde seleccionan los parámetros que lleva el dataset con los que se desarrolla el trabajo.

Para evaluar los métodos de imputación y elegir el mejor dataset se hace una comparación de cada uno con el conjunto de datos original y se calcula una medida de exactitud para cada método, lo que se usa como criterio para elegir el mejor método entre estos. En cuanto a la selección de datos, se analizan atributos innecesarios o repetidos con el fin de omitirlos y luego de esto obtener el 'mejor dataset'.

Teniendo el dataset generado luego del preprocesamiento se procede a aplicar las técnicas de clasificación supervisada (K-nearest neighbour, Logistic regression y Artificial neural networks) y no supervisada (k-means clustering y hierarchical clustering).

Para las técnicas de clasificación supervisada se dividen los datos con la siguiente proporción: 80% datos de entrenamiento y 20% datos de prueba. Con el fin de comparar estos métodos, se realiza una predicción de las clases, este resultado es comparado con los datos de prueba y se calcula precisión de cada modelo con el fin de determinar el mejor para este caso. Esta evaluación es acompañada por matrices de confusión, las cuales muestran la cantidad de aciertos y desaciertos de cada modelo al predecir la clase mediante la predicción.

Para clasificación no supervisada se aplican dos técnicas, k-means clustering y Hierarchical clustering. En este paso se busca hallar patrones y comportamientos similares entre los datos con el fin de agrupar los registros en grupos y determinar el número de clusters que sería más apropiado para analizar los datos del problema. Dentro de este proceso se analiza cada atributo del dataset de manera independientemente y su comportamiento dentro de cada cluster.

Luego de desarrollar y evaluar los modelos de clasificación, se procede a validar las hipótesis del problema. Las afirmaciones que se quieren validar con este trabajo se muestran a continuación:

- La participación en clase, cantidad de recursos visitados, participación en foros, cantidad de revisiones de anuncios y ausencias a clase, son aspectos que tienen un alto impacto en el desempeño académico de los estudiantes.
- Estudiantes con una cantidad de ausencias a clase superior a 7, tienden a obtener un desempeño bajo en la asignatura.
- No existe una relación relevante entre el país de origen y el desempeño final del estudiante en la asignatura.

Finalmente se usa el conocimiento obtenido para plantear algunas propuestas basadas en los resultados que podrían mejorar el rendimiento académico de los estudiantes en el futuro.

4.3 Resultados

Clasificación supervisada

Con la implementación de los algoritmos y siguiendo la metodología que se explicó anteriormente, se lograron desarrollar modelos de clasificación supervisada que logran predecir el desempeño académico de un estudiante en determinada asignatura.

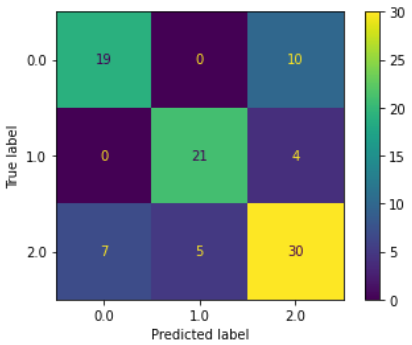
Los modelos de clasificación supervisada fueron evaluados inicialmente con la precisión de cada uno, métrica que en minería de datos se conoce como *accuracy*.

El cálculo de esta métrica se obtiene de las matrices de confusión, las cuales también fueron útiles para observar los resultados numéricos del modelo al realizar una predicción con datos de prueba y de esta manera comparar la predicción con los valores actuales de la clase que correspondían a estos datos. La matriz de confusión aplicada a cada modelo sigue la estructura que se muestra en la figura a continuación.

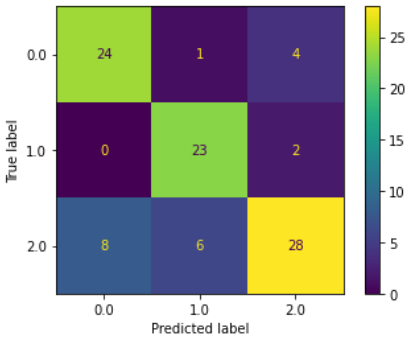
		Predicción	
		Positivo	**Negativo**
Actual	**Positivo**	Verdadero positivo (VP)	Falso Negativo (FN)
	Negativo	Falso positivo (FP)	Verdadero negativo (VN)

Tabla 2. Estructura básica matriz de confusión

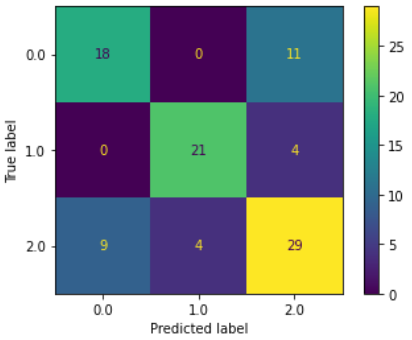
A continuación, se muestran las matrices de confusión de los tres métodos.



Logistic Regression.



KNN



Artificial Neural Networks

Se puede evidenciar que en el método KNN existen menos falsos negativos y falsos positivos.

De esta manera, se dice que la precisión es la proporción del número total de predicciones que fueron calculadas correctamente y se calcula de la siguiente manera.

$$Accuracy = \frac{VP + VN}{VP + FN + FP + VN}$$

Los resultados de esta métrica para cada método de clasificación se presentan en el siguiente gráfico.

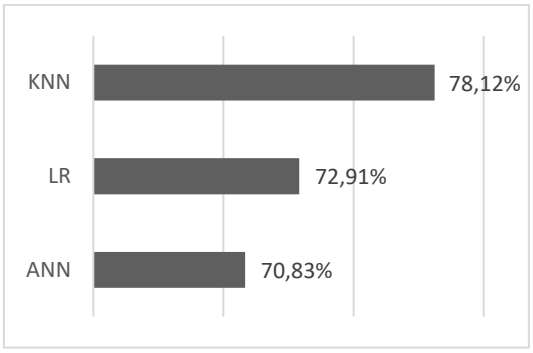


Ilustración 13. Precisión de modelos clasificación supervisada

Clasificación no supervisada

Los métodos de clasificación no supervisada fueron evaluados mediante un método gráfico, buscando observar los clusters en el mismo plano para observar el distanciamiento y diferenciación entre cada grupo.

Para el método K-means inicialmente se usa el grafico de codo con el fin de determinar el número de clusters más apropiado mediante la identificación del punto de inflexión.

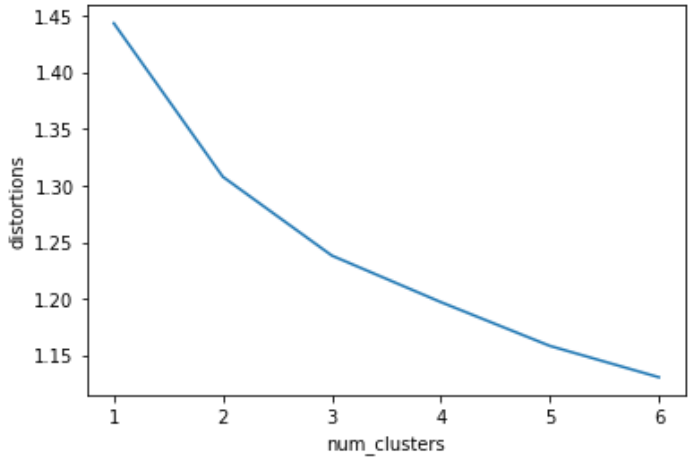


Ilustración 14. Gráfico de codo

De aquí no se logra visualizar un punto que se diferencie significativamente, pero se observan cambios de pendientes significativos en K=2 y K=3 por lo que se decide aplicar el método con ambos valores.

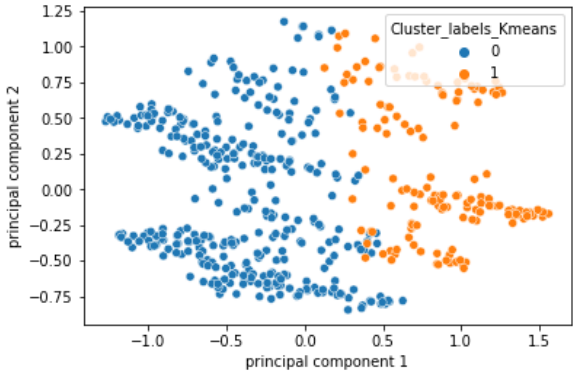


Ilustración 15. K-means con K=2

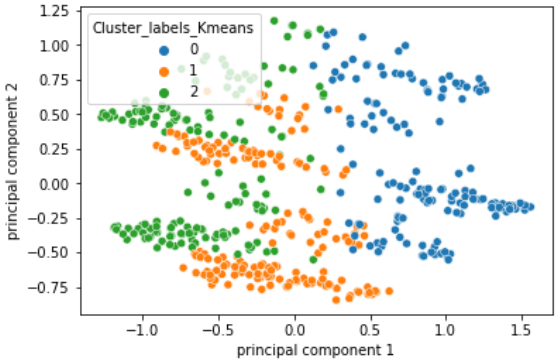


Ilustración 16. K-means con K=3

Comparando este resultado, se observa que hay una mayor diferenciación entre los clusters cuando K=2. Al tener tres clusters, uno de estos se interpone sobre los otros, lo que se visualiza con los puntos en conflicto que existen entre clusters diferentes. Dado este análisis, del método K-means se concluye que el número de clusters más adecuado es dos.

Para el método de hierarchical clustering se desarrolla un dendograma donde se puede evidenciar la jerarquización y mediante colores, el número de clases más recomendado.

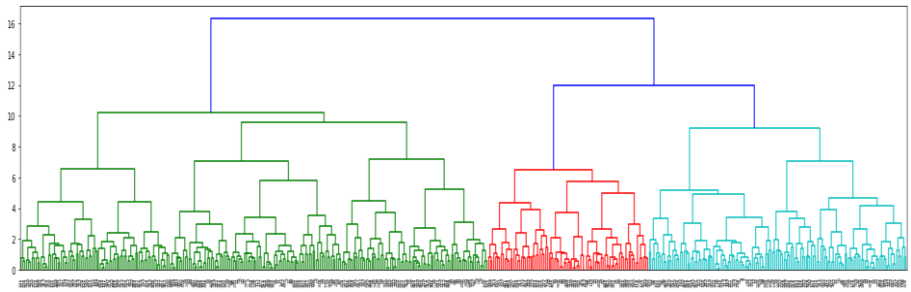


Ilustración 17. Dendograma para hierarchical clustering

De acuerdo con el dendograma, el número de clusters más adecuado es tres, sin embargo, dado el análisis realizado en K-means, se decide aplicar el método de hierarchical clustering usando dos y tres clusters.

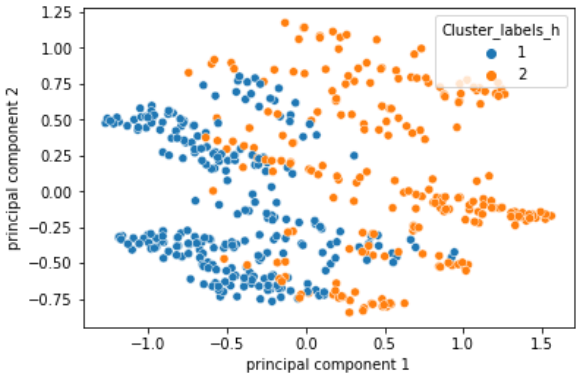


Ilustración 18. Hierarchical clustering – 2 clusters

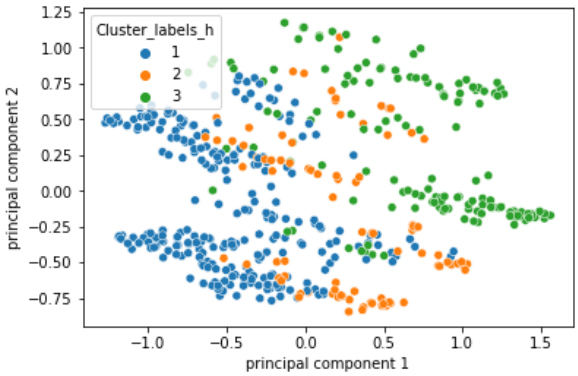


Ilustración 19. Hierarchical clustering – 3 clusters

Para ambos casos se observa que existe superposición entre algunos puntos, aunque también se sigue evidenciando una mayor diferenciación entre clusters cuando estos son dos, en lugar de tres. Sin embargo, sigue siendo el algoritmo de K-means con K=2 aquel que muestra una distribución más diferenciada, siendo este el método de clasificación no supervisada más apropiado para este caso de estudio.

Analizando el comportamiento de los dos grupos encontrados con el método de clasificación de k-means se encuentra que existen diferencias relevantes entre algunos atributos, las cuales se considera importante analizar.

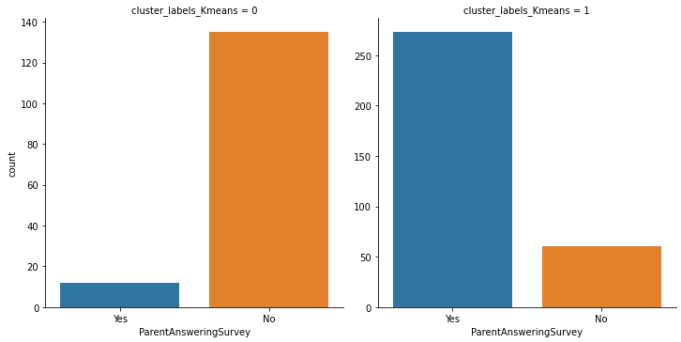


Ilustración 20. Padres que respondieron a la encuesta

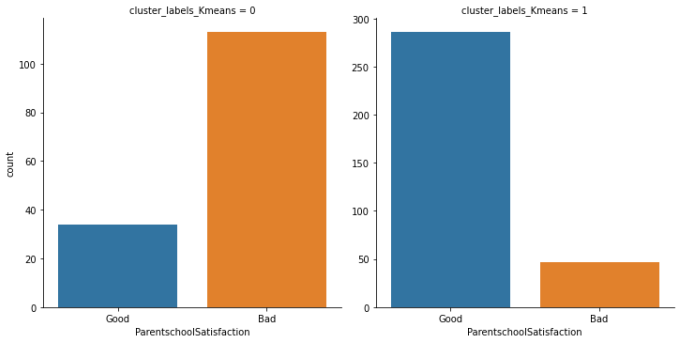


Ilustración 21. Satisfacción de los padres con la escuela para ambos clusters

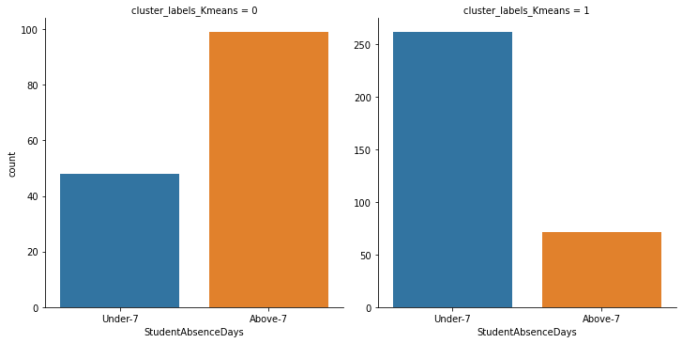


Ilustración 22. Ausencias a clase para los 2 clusters

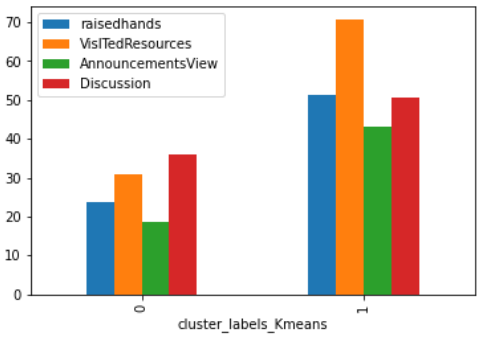


Ilustración 23. Comportamiento en clase para los 2 clusters

Las variables relacionadas con la participación de los padres muestran una notoria diferencia entre los dos grupos. En el primero se puede observar que la mayor parte no respondieron la encuesta y además no se encuentran satisfechos con la escuela, por su parte, el segundo grupo evidencia que la gran mayoría de padres respondieron la encuesta y se encuentran satisfechos con la institución. En cuanto aspectos comportamentales y hábitos de los estudiantes, también se evidencian algunas diferencias relevantes. El primer grupo muestra que para la mayoría de los estudiantes sus ausencias a clase superan una cantidad de siete, frente al segundo grupo, donde la mayoría de los estudiantes no supera esta cantidad. Finalmente, las variables de comportamiento en clase muestran que también existen diferencias entre ambos grupos, siendo el grupo 2 aquel que refleja mayor participación en estos aspectos.

4.4 Discusión

Para comprobar los hallazgos y evidenciar si existe alguna relación determinante entre las variables que diferencian los dos grupos analizados en clasificación no supervisada, y la clase, se decide analizar cada que variable de manera independiente, estudiando su verdadera correlación con la clase.

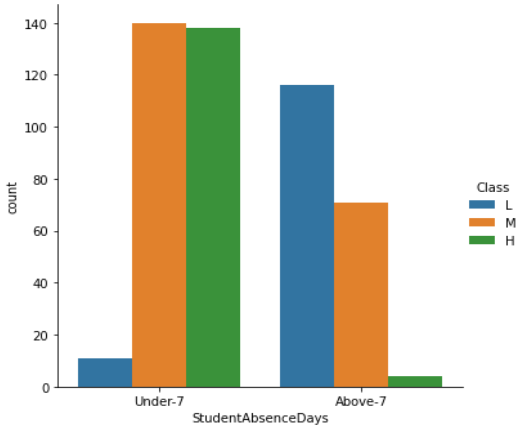


Ilustración 24. Ausencias a clase en comparación con la clase

Se puede evidenciar que los estudiantes que tienen ausencias menores a 7 días tienen mejor desempeño que los estudiantes que tienen ausencias mayores a 7 días, por lo que se evidencia una relación entre el ausentismo y las malas calificaciones.

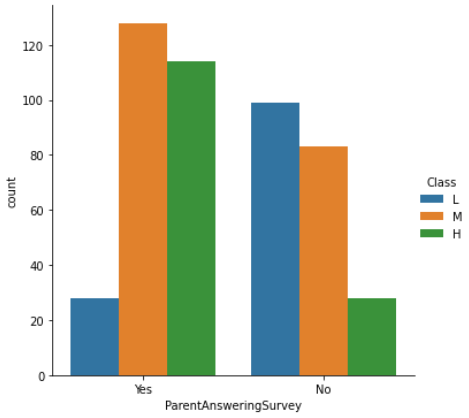


Ilustración 25. Padres que responden la encuesta en comparación con la clase

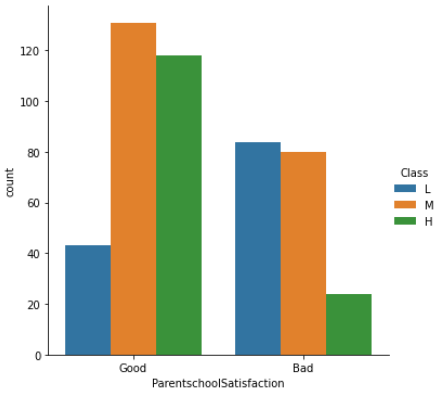


Ilustración 26. Satisfacción de los padres en comparación con la clase

Las dos graficas anteriores hacen referencia a dos atributos relacionados con los padres del estudiante, en estas graficas se puede observar que los hijos de padres que tienen una buena relación con la escuela y respondieron a la encuesta realizada por la institución tienen un mejor desempeño que los hijos de padres que no están satisfechos con la escuela y tampoco respondieron a la encuesta.

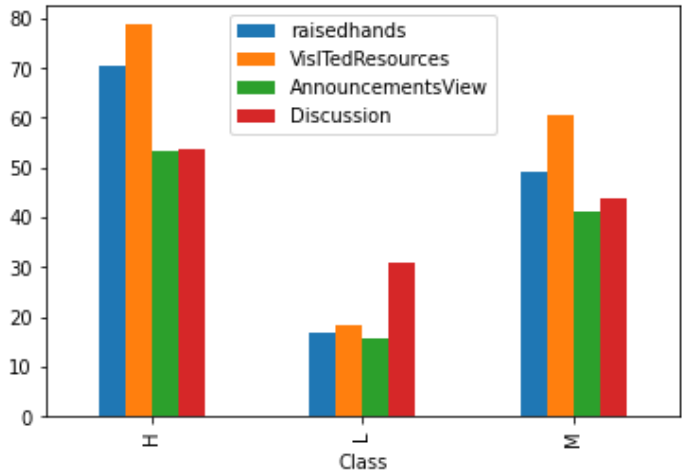


Ilustración 27. Variables de comportamiento en clase en comparación con la clase

Finalmente, se puede observar que los estudiantes que fueron más activos en la plataforma y en las clases, tienen un mejor desempeño que los que no lo fueron. Es decir, existe una relación entre la participación del estudiante y sus resultados académicos.

Otra manera de comprobar la relación de los atributos con la clase es haciendo uso del "Feature importance" generado por los métodos de clasificación. En los métodos aplicados anteriormente (KNN, ANN y LR) no existe una manera de calcular este parámetro, o si la existe, pierde interpretabilidad al tratarse de más de dos clases. Sin embargo, haciendo uso del algoritmo de Random Forest, se puede obtener esta medida. A continuación, se muestra la importancia de cada atributo en un Random Forest de 100 árboles.

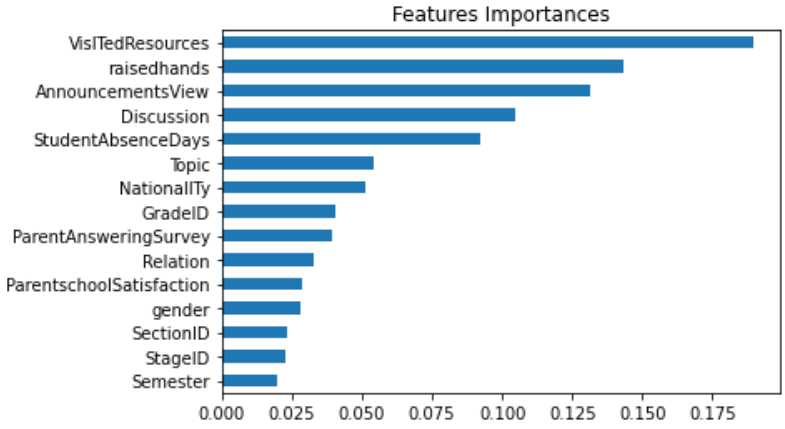


Ilustración 28. Features importance

Con el gráfico anterior se pueden confirmar las relaciones declaradas anteriormente ya que los atributos más importantes para determinar la clase son: Recursos visitados, levantar la mano, vistas de anuncios y discusión (referentes a la participación del estudiante) y la ausencia del estudiante.

Teniendo estos hallazgos como base, se plantean las siguientes recomendaciones para que las instituciones educativas puedan mejorar el desempeño académico de sus alumnos.

1. Incentivar la participación en clase mediante bonificaciones extra, actividades dinámicas y llamados a participación repentinos y aleatorios.
2. Capacitar a los estudiantes en temas relacionados con uso de la plataforma, búsqueda de recursos en línea y foros educativos.
3. Reforzar las medidas de control relacionadas con asistencia a clase.

5. Conclusiones

Las bases de datos académicas ofrecen enormes cantidades de información y de conocimiento oculto que es importante que sea descubierto. En este trabajo se implementaron técnicas de minería de datos con el fin de analizar registros recopilados de un sistema de e-learning llamado Kallboard 360 que usa la experiencia API (XAPI). Implementando algoritmos de clasificación supervisada y no supervisada se encontró que las variables comportamentales como participación en clase, recursos visitados, participación en foros, revisión de anuncios y ausencias a clase, evidencian un fuerte impacto en el desempeño académico de un estudiante. Para predecir el desempeño de un estudiante teniendo la información requerida, se desarrollaron modelos usando técnicas como K-nearest neighbour, Logistic Regression y Artificial Neural Networks, los cuales obtuvieron como resultado una precisión del 78.12%, 72.91% y 70.83% respectivamente. En cuanto al conjunto de datos estudiado, las técnicas de clasificación no supervisada que se implementaron (K-means clustering y Hierarchical clustering) permiten concluir que los estudiantes se distribuyen en dos grupos (o clusters), cuyos elementos individuales comparten características similares pero hacen que ambos grupos se diferencien entre sí en aspectos principalmente relacionados por las variables comportamentales mencionadas anteriormente, junto con otras, como la participación y satisfacción de los padres con la escuela.
El trabajo futuro relacionado con este proyecto puede realizarse con bases de datos de mayor tamaño y complejidad con el fin de obtener algoritmos que muestren un mejor rendimiento. Además de esto, se pueden estudiar otros aspectos relevantes que podrían tener un alto impacto en el desempeño de un estudiante, como las horas y métodos de estudio, o incluso, las técnicas de enseñanza implementadas por los docentes.

6. Bibliografía

[1]. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.
[2]. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.
[3]. 2020. [Online]. Available: https://mc.ai/how-to-implement-a-neural-network-with-only-1-and-1/. [Accessed: 22- Nov- 2020].