risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Revista Ibérica de Sistemas y Tecnologías de Información

Recebido/Submission: 22/03/2023
Aceitação/Acceptance: 03/07/2023

Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria

Anthony Edwin Aco Tito[1], Bryan Orlando Hancco Condori[1], Yasiel Pérez Vera[1]

aacot@unsa.edu.pe; bhanccoco@unsa.edu.pe; yperezv@unsa.edu.pe

[1] Universidad Nacional de San Agustín de Arequipa, Santa Catalina 117, 04001, Arequipa, Perú.

DOI: 10.17013/risti.51.84–98

Resumen: La deserción universitaria afecta negativamente a muchos estudiantes, este suceso puede estar relacionado con problemas personales, cuestiones económicas, entre otros. Ante tal situación surge la importancia de desarrollar una forma de predecir estos casos, para esto se propuso el uso de técnicas de Machine Learning, las utilizadas fueron Regresión Logística, Naive Bayes, Red Neuronal Perceptrón Multicapa, Árbol de Decisión, Support Vector Machine y Random Forest; se seleccionó un *Dataset*, que pasó por una limpieza de datos, se corrigieron los datos faltantes y los valores atípicos; luego se eliminaron los registros cuya variable de salida era Matriculado, centrándose en los tipos Abandono y Graduado. Cada modelo fue entrenado y probado mediante validación cruzada con pliegues, finalmente, se compararon en función de métricas de precisión, exactitud y exhaustividad, donde se concluyó que la Regresión Logística es la técnica que mejores resultados proporciona para predecir la deserción universitaria en el dataset considerado.

Palabras-clave: Análisis comparativo; Deserción Universitaria; Machine Learning; Predicción; Regresión Logística.

Comparative analysis of Machine Learning Techniques for the prediction of cases of university dropout

Abstract: University dropout has a detrimental impact on numerous students; this phenomenon may be associated with personal issues, economic constraints, and other factors. Given this situation, the importance of developing a predictive model for such cases arises. To achieve this, Machine Learning techniques were proposed and employed, including Logistic Regression, Naive Bayes, Multilayer Perceptron Neural Network, Decision Tree, Support Vector Machine, and Random Forest. A dataset was selected and underwent data cleaning, addressing missing values and outliers. Subsequently, records with the 'Enrolled' outcome variable were removed, focusing solely on 'Dropout' and 'Graduate' categories. Each model was trained and tested using cross-validation with folds. Ultimately, they were compared based

on accuracy, precision, and recall metrics, leading to the conclusion that Logistic Regression is the technique that yields the best results for predicting university dropout in the considered dataset.

Keywords: Comparative analysis, University dropout, Logistic Regression, Machine Learning, Prediction.

1. Introducción

En la actualidad, los estudiantes universitarios se enfrentan a múltiples dificultades y obstáculos, que varían de acuerdo al contexto de cada país, en el momento de cursar estudios universitarios. Uno de estos problemas es el abandono universitario, en el que los estudiantes pueden abandonar o retrasar sus estudios. Esto se debe a muchos factores. Por ejemplo, según Echeverry (2020), las principales razones por las que muchos estudiantes consideran el abandono como una opción son identificar trayectorias profesionales alternativas que creen que pueden ofrecer mejores resultados, resolver sus problemas económicos y sus dificultades personales. Del mismo modo, según Miño de Gauto (2021) se reconocen factores internos, propios de los propios estudiantes, como el estado civil, las expectativas no cumplidas, la falta de orientación profesional, las dificultades emocionales, la autoestima, etc. Por otro lado, están los factores externos, que se refieren a factores sociales e institucionales. A partir de aquí, podemos observar que los factores internos son muy similares, lo que también ocurre en otros países. Esto puede intensificarse o disminuirse en función de los denominados factores externos.

Este problema es de gran importancia porque la deserción universitaria, según María del Carmen Parrino (2014) tiene consecuencias no sólo a nivel personal, como provocar sentimientos de frustración o fracaso, sino que también representa un importante desperdicio de recursos para las familias afectadas y la sociedad; esto sumado al hecho de que muy probablemente no consigan los puestos de trabajo deseados, provocando subempleo y, en consecuencia, un ingreso mucho menor de lo esperado (Viale Tudela, 2014); observando las cifras presentes en varios países notamos que estamos ante un problema en el que son varios los agentes que están fallando en la misión de garantizar un acceso a la educación y así entonces a una calidad de vida más alta.

A nivel internacional y según los datos de Lenin C. Guerra (2022) podemos ver el caso de los EEUU, donde la deserción universitaria ronda al 40%, siendo una de las más altas, mientras que en Canadá ronda el 31%. Además de ello también se visualiza un patrón internacional, que indica que el 77% de la deserción universitaria se realiza al final del 2do año.

Analizando ahora los datos a nivel de Latinoamérica, los datos obtenidos por el Banco Mundial (Ferreyra, et. al., 2017) muestran que la tasa de deserción en países como Bolivia, Colombia y Panamá están por encima del 30%, mientras que para países como Argentina, Chile, Uruguay ronda el 10%. El caso del Perú también es analizado, y en contraste con el promedio regional, tenía niveles bajos de deserción, rondando el 10%.

Otro factor a considerar son los problemas causados por la pandemia del COVID-19, pues según Ferreyra (2017), la deserción de muchos estudiantes se debió a motivos tales como

los escasos recursos tecnológicos, problemas socioeconómicos, y en el peor de los casos problemas de salud; toda esta situación dejó secuelas que aún se reflejan a día de hoy.

Ante tal situación, se propone un conjunto de técnicas que pueden facilitar la predicción o clasificación de los estudiantes en base a sus perfiles, nos referimos a las técnicas de Machine Learning, ya que dicho campo cuenta actualmente con diversas aplicaciones en distintos ámbitos, como en la medicina, la educación, entre otros.

El presente artículo contiene las siguientes secciones, comenzando con un análisis de los Trabajos Relacionados, seguidamente se describen los Materiales y Métodos requeridos para esta investigación, posteriormente realizamos la sección de Resultados y Discusión, donde nos encargamos de analizar los resultados obtenidos por cada una de las técnicas de Machine Learning para así encontrar el método que nos dé mejores resultados en términos de exactitud, precisión y exhaustividad; y finalizamos con la sección de Conclusiones.

2. Trabajos Relacionados

En la presente sección se presentan algunos antecedentes de trabajos en los que se aplicaron técnicas de Machine Learning para la predicción de la deserción universitaria, entre ellos tenemos el trabajo de Martin et al. (2021) en la que se propuso modelos estándar como Regresión Logística, Support Vector Machine, Árbol de Decisión, Random Forest y modelos Boosting; los autores concluyen que los modelos Boosting son mejores modelos, sin embargo estos tampoco logran clasificar correctamente las clases minoritarias.

En otro caso el estudio realizado por Tocto, Huamaní y Zuloaga (2023) se aplicó una Red Neuronal para la predicción de estudiantes que dejarán los estudios obteniendo una precisión de 66.30%, es por esto que los mismos autores sugieren que se deben tener en cuenta más atributos que estén relacionados con la deserción universitaria para poder mejorar el modelo.

Por otro lado, en el trabajo realizado por Solis, et al. (2018) se analiza y se selecciona al mejor algoritmo de Machine Learning con el objetivo de predecir la deserción de estudiantes universitarios, para ello se evaluaron a los algoritmos de Random Forest, Redes Neuronales, Support Vector Machine y Regresión Logística, obteniendo sus mejores resultados una precisión del 82% y una sensibilidad del 71% con el algoritmo de Random Forest. Cabe destacar que los mismos autores sugieren la necesidad de agregar en sus datos a variables que en los últimos años han tenido mayor correlación con el problema de la deserción académica.

Así mismo, el estudio realizado por Contreras, Fuentes y Rodríguez (2020) aplicaron varios algoritmos de Machine Learning entre los cuales, los autores concluyen que Support Vector Machine y Red Neuronal Perceptrón son los de mayor precisión, sin embargo, al igual que el trabajo anterior dicho valor (66.4%) no es ideal por lo que también se recomienda agregar como variables a más factores que influyen en el rendimiento académico.

De igual forma, en el trabajo de Ayala, Valenzuela y Espinosa (2020), se plantearon modelos para la predicción de la deserción universitaria, aplicando técnicas de Machine Learning, entre las cuales la regresión logística es la que obtuvo el mejor desempeño,

sin embargo, los autores resaltan la importancia de la falta de un modelo con mayor interpretabilidad como los árboles de decisión.

Finalmente, en el trabajo de Bedregal, Aruquipa y Cornejo (2020) se aplicó el modelo de árbol de decisión CHAID para predecir la posible deserción universitaria en los alumnos de la escuela profesional de Ingeniería de Sistemas de la Universidad Nacional de San Agustín, en la que se logró una correcta identificación del 90.24% de los casos de deserción. Concluyen además que de acuerdo a la preparación de los datos se puede realizar también un análisis de rendimiento académico, con el fin de predecir el comportamiento de los estudiantes destacados.

3. Materiales y Métodos

En la presente sección se detallan los conceptos previos, los modelos empleados y las herramientas aplicadas en la sección de análisis de datos.

3.1. Marco Teórico

En la presente sección se detallan los principales referentes teóricos de Machine Learning, además de los conceptos de sus principales algoritmos y las métricas de evaluación de estos algoritmos.

3.1.1. Machine Learning

Denominado también como Aprendizaje Automático, es una vertiente de la Inteligencia Artificial, López y Aguirre (2019) mencionan que su objetivo es el planteamiento de técnicas que posibilitan que las computadoras aprendan y sean capaces de generalizar en base a conjuntos de datos que les fueron proporcionados anteriormente, dichas técnicas son utilizadas cuando se tiene un gran volumen de datos, cosa que a un ser humano le resultaría complicado analizar para poder sacar conclusiones al respecto.

3.1.2. Regresión Logística

Es una técnica estadística de varias variables, que facilita el análisis de la relación que se encuentra entre un grupo de variables independientes y una variable dependiente, esta última es dicotómica, es decir que solo puede tomar dos valores, que generalmente son cero y uno, donde cero es ausencia y uno presencia, es decir la variable dependiente es discreta, a diferencia del grupo de variables independientes la cuales pueden ser tanto cuantitativas como cualitativas; además la función de partida es exponencial, lo que ayuda a representar mejor ciertos fenómenos (Lizares, 2017).

3.1.3. Árboles de decisión

Este modelo basado en árboles, permite clasificar predecir el valor de una variable de entrada en base a un grupo de variables independientes, proporciona una forma más gráfica de representar las alternativas o eventos que surgen a partir de la elección de una opción, en la que finalmente se nos proporciona la más acertada probabilísticamente; todo esto mediante un proceso que puede ser observado visualmente; en dicho proceso

se asumen que los conjuntos que participan son disjuntos, es decir que todos los objetos de entrada son direccionados a un solo grupo de pertenencia (Lizares, 2017) .

3.1.4. Support Vector Machine

Las máquinas de Vectores Soporte (SVM en inglés) pertenecen a la clase de los clasificadores de tipo lineal, que utilizan hiperplanos para crear una separación en el espacio original si los casos de entrada son linealmente separables o en un espacio transformado. Se diferencian de otros métodos de aprendizaje automático para generar separación o clasificación porque se centran en minimizar el riesgo estructural. El objetivo es encontrar un hiperplano que sea equidistante con los ejemplos más cercanos pertenecientes a las clases separadas, donde los elementos de entrenamiento ubicados en la frontera de la separación son los llamados vectores de soporte. Suelen tener gran capacidad de generalización y además evita los sobreentrenamientos (Carmona, 2016).

3.1.5. Random Forest

Es un método de Machine Learning conjunto que permite la clasificación y la regresión, el cual está conformado por un conjunto de árboles de decisión, construidos mediante *bagging*, cuya variación se controla considerando a los datos de entrenamiento. En el proceso de clasificación, cada árbol emite su voto de forma individual sobre la clase predicha, en la que finalmente se considera aquella clase con votación mayoritaria (Faw Agregar, Gaber y Elyan, 2014).

3.1.6. Redes Neuronales - Perceptrón Multicapa

Es una red neuronal que puede tener una o varias capas ocultas entre las capas de salida y de entrada, en donde todas la conexiones mantienen la dirección de ir de una capa inferior a una superior, las neuronas ubicadas en una misma capa no están conectadas, además la suma total de neuronas en la capa de entrada es la misma que a la totalidad de características para el patrón problema, en cuanto al total de neuronas en la capa de salida, esta es la misma que el total clases. Para la elección del número de capas y del número de neuronas para cada capa, priorizamos la tarea de optimizar la red hasta que sea adecuada con los suficientes parámetros, y además tenga una buena generalización para la tarea de clasificación o regresión (Ramchoun, 2016).

3.1.7. Naive Bayes

El algoritmo de Naive Bayes es un clasificador de tipo probabilístico, que determina un grupo de valores probabilísticos en base a la frecuencia de los valores en un determinado conjunto de datos de entrenamiento, tiene como base al Teorema de Bayes, y se asume una independencia entre las variables de entrada (Fawagreh, Gaber y Elyan, 2014).

3.1.8. Métricas para la evaluación de algoritmos de Machine Learning

Existen varias métricas para la evaluación de algoritmos de Machine Learning, entre ellas podemos encontrar a la precisión, que es la relación entre el número de casos positivos correctamente clasificados y el total de casos positivos predichos.

También existe la métrica de exactitud o accuracy, que se obtiene a partir de la división entre la suma de clasificaciones correctas y el total de los casos clasificados. Borja et al. (2020) indica que, a pesar de su cálculo y compresión, su mayor desventaja es la de producir menos valores discriminatorios para el caso de multiclase desbalanceado.

Otra métrica importante es la exhaustividad, también denominada como Recall o sensibilidad, que es la relación entre el número de casos positivos correctamente clasificados y el total de casos positivos reales.

3.2. Herramientas

3.2.1. Python

Python es un lenguaje de programación con una gran facilidad de uso y de alto nivel, Upasani y Virendra (2020) detallan que cuenta con una gran variedad de librerías útiles para la carga, procesamiento y visualización de datos, además de librerías capaces de ejecutar algoritmos de machine learning para la clasificación, predicción, clustering, etc.

3.2.2. Google Collab

Google Collab es un entorno de Jupyter Notebook basado en la nube que se ejecuta en diversos navegadores, Colab nos permite escribir y ejecutar código de Python o de R para aplicaciones en Inteligencia Artificial, garantizando por ejemplo el desarrollo y entrenamiento de redes neuronales, entre otros algoritmos. También permite compartir el código con múltiples usuarios proporcionando edición simultánea del código.

3.2.3. Scikit Learn

Scikit-Learn es una biblioteca de tipo código abierto en constante desarrollo y mejora. En Upasani y Virendra (2020) mencionan que esta librería proporciona múltiples algoritmos de Machine Learning, para aprendizaje supervisado y no supervisado, y otras múltiples herramientas científicas que la han permitido llegar a ser un instrumento ideal para el desarrollo de aplicaciones de Machine Learning.

4. Resultados y Discusión

En la presente sección se detalla el análisis de los datos, la limpieza de datos, la aplicación de los algoritmos y se finaliza con la comparación de resultados.

4.1. Análisis de datos

El dataset analizado para la clasificación de los estudiantes que completarían sus estudios y los que los abandonarían antes de finalizarlos fue el de *Predict students' dropout and academic success*, obtenido de la plataforma Zenodo (Realinho, 2021). Este dataset fue reconectado con el auspicio del proyecto SATDAP - Capacitação da Administração Pública under grant POCI-05-5762-FSE-000191, Portugal. Contiene datos de los programas de estudios: agronomía, diseño, educación, enfermería, periodismo, administración, servicios sociales, y tecnología, obtenidos del Instituto Politécnico de Portalegre (Martins et al., 2021).

El dataset consta de 35 columnas y 4424 registros, como se indica en Zenodo (Realinho, 2021). Las columnas correspondientes a las variables de entrada son las que se observan en la Tabla 1. Las variables de salida se corresponden con el campo Target de tipo categórico, estas son Dropout, Enrolled y Graduate cuyas traducciones son Abandono, Matriculado y Graduado.

Variables	**Tipos**
Edad al momento de la inscripción	Numérico
Modalidad de postulación	Categórico
Orden de postulación	Numérico
Curso	Categórico
Unidades curriculares 1er semestre (matriculado)	Numérico
Unidades curriculares 1er semestre (aprobado)	Numérico
Unidades curriculares 1er semestre (calificación)	Numérico
Unidades curriculares 1er semestre (acreditadas)	Numérico
Unidades curriculares 1er semestre (evaluaciones))	Numérico
Unidades Curriculares 1er semestre (sin evaluaciones)	Numérico
Unidades Curriculares 2do semestre (matriculado)	Numérico
Unidades Curriculares 2do semestre (aprobado)	Numérico
Unidades Curriculares 2do semestre (calificación)	Numérico
Unidades Curriculares 2do semestre (acreditadas)	Numérico
Unidades Curriculares 2do semestre (evaluaciones)	Numérico
Unidades curriculares 2do semestre (sin evaluaciones)	Numérico
Desplazados	Categórico
Asistencia diurna/nocturna	Categórico
Deudor	Categórico
Ocupación de la madre	Categórico
Ocupación del padre	Categórico
PIB	Numérico
Necesidades educativas especiales	Categórico
Género	Categórico
Tasa de inflación	Numérico
Internacional	Categórico
Estado civil	Categórico
Calificación de la madre	Categórico
Calificación del padre	Categórico
Calificación previa	Categórico
Nacionalidad	Categórico
Matrículas al día	Categórico

Variables	Tipos
Becario	Categórico
Tasa de desempleo	Numérico

Tabla 1 – Las columnas correspondientes a las variables de entrada con sus tipos.

4.2. Limpieza de datos

Una vez cargados los datos correspondientes, utilizando librerías de Python, comprobamos las variables con menor relevancia y que, por tanto, contribuyen en menor grado a los modelos. Para ello, verificamos a través de su correlación con la variable de salida *Target*, como se muestra en la Figura 1.

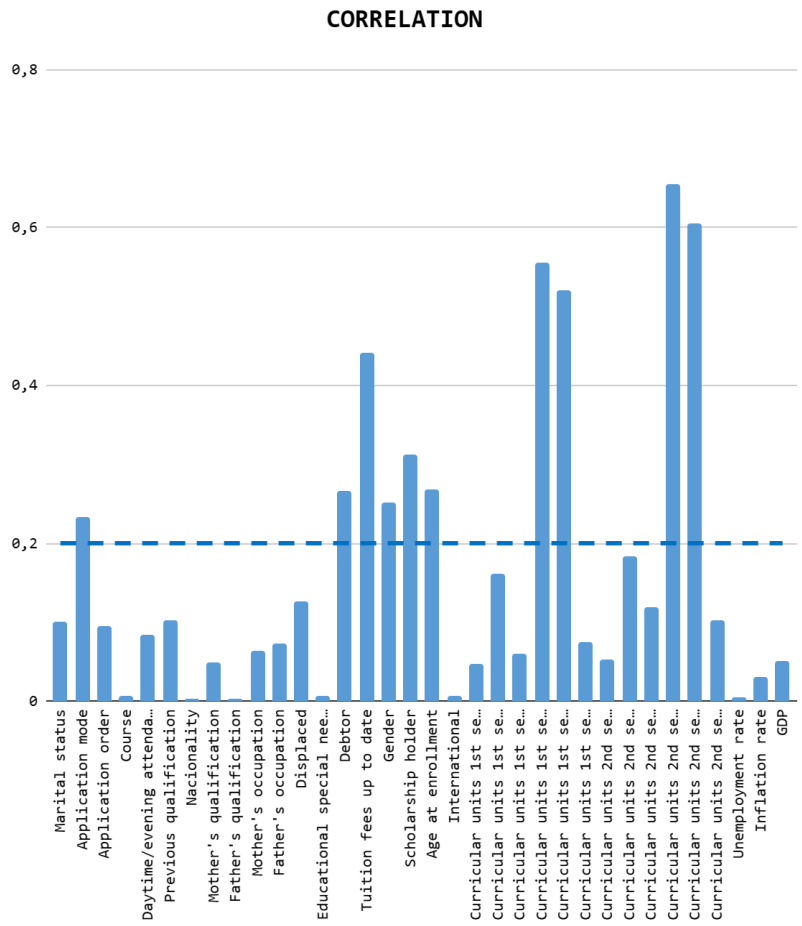

Figura 1 – Correlación entre las variables de entrada y el objetivo.

Este gráfico muestra la correlación de cada variable de entrada con la variable objetivo, así por ejemplo, variables como Unidades Curriculares 2do semestre (aprobado) y Unidades Curriculares 2do semestre (calificación) tienen el mayor grado de correlación con nuestra variable objetivo, por el contrario, variables como Nacionalidad o Internacional se encuentran entre las de menor correlación; para esto consideramos como variables importantes aquellas cuyo valor absoluto del resultado es mayor o igual a 0,2, reduciendo así las 35 variables iniciales a sólo 11.

Continuamos con la verificación y el tratamiento de los *Data Missing* y los *Outliers*; en nuestro conjunto de datos no se encontraron datos vacíos, y para el tratamiento de los *Outliers* utilizamos un método paramétrico que valida los datos anómalos mediante el rango Inter cuartil y el diagrama de caja, excluyendo así los datos extremadamente atípicos, excepto los campos categóricos, como se muestra en la figura 2, después del ajuste, pasamos de tener inicialmente un total de 4424 filas a 3072 filas.

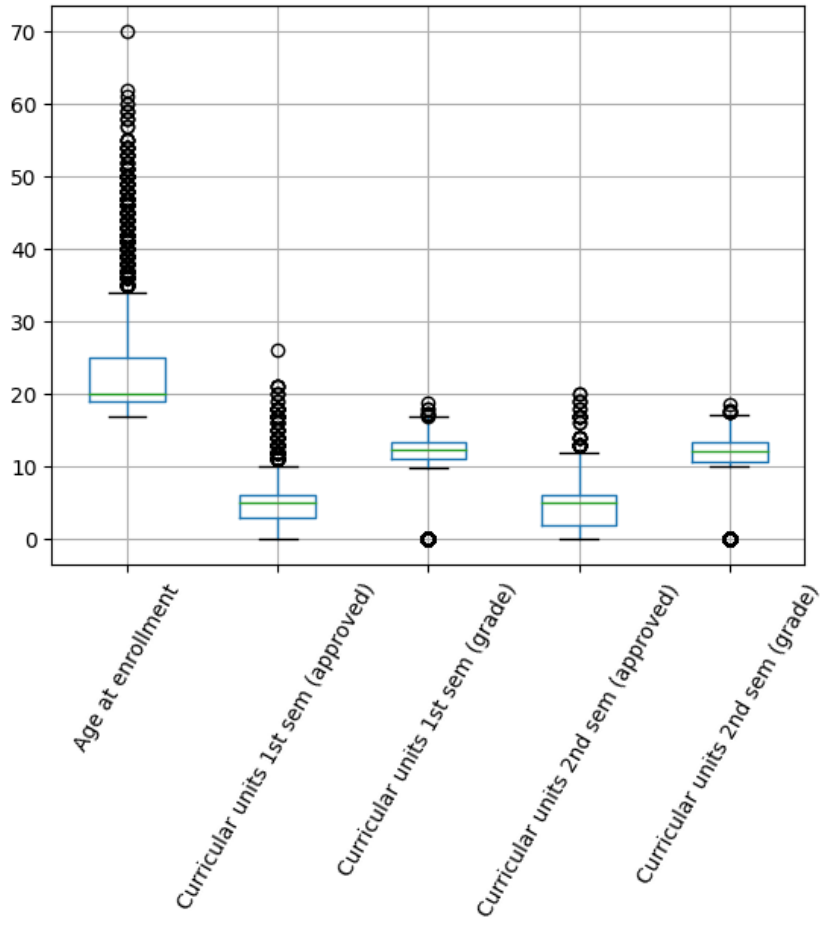

Figura 2 – Gráfico de cajas para la det6cción de valores atípicos.

Así también, debido a que nuestra variable de salida *Target* tiene tres resultados posibles, es necesario descartar los resultados de tipo Matriculado, ya que sólo nos interesan los resultados Abandono y Graduado, resultando una reducción del número de registros de 3072 filas a 2422 filas.

Por último, la fase de transformación de datos, en la que se transformó la variable de salida *Target* a valores numéricos, Abandono en 0 y Graduado en 1.

4.3. Aplicación de los algoritmos

Para el análisis comparativo se evaluaron los algoritmos que provee la librería Scikit-Learn, junto con los parámetros asignados, como se muestra en la Tabla 2.

Algoritmo	Parámetros
Logistic Regression	penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='newton-cholesky', max_iter=100, multi_class='auto', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None
Decision Trees	criterion='entropy', splitter='best', max_depth=5, min_samples_split=150, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, class_ weight=None, ccp_alpha=0.0
Support Vector Machine (SVM)	C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False, random_state=None
Random Forest	n_estimators=6, criterion='gini', max_depth=29, min_samples_split=2, min_ samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='sqrt', max_leaf_ nodes=None, min_impurity_decrease=0.0, bootstrap=True, oob_score=False, n_ jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None, ccp_alpha=0.0, max_samples=None
Neural Network - Multilayer Perceptron	hidden_layer_sizes=(100), activation='logistic', *, solver='adam', alpha=1e-5, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False, warm_ start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2=0.999, epsilon=1e-08, n_iter_no_ change=10, max_fun=15000
Naive Bayes	priors=None, var_smoothing= 1e-9

Tabla 2 – Parámetros asignados a cada algoritmo

Luego, se llevó a cabo un proceso de entrenamiento con validación cruzada, donde se consideraron diez pliegues. Nueve de ellos se utilizan en esta etapa y uno en la validación, repitiendo el proceso diez veces, donde se cambia el pliegue utilizado en la validación.

4.4. Comparación de resultados

Los resultados obtenidos al evaluar los diez pliegues en los seis modelos en las métricas de precisión, exactitud y exhaustividad fueron los siguientes:

En el gráfico de precisión, Figura 3, se presenta el rendimiento de los modelos en esta métrica, siendo el modelo de Random Forest el máximo en el pliegue tres. En cambio, el modelo de árbol de decisión obtiene el mínimo en el séptimo pliegue.

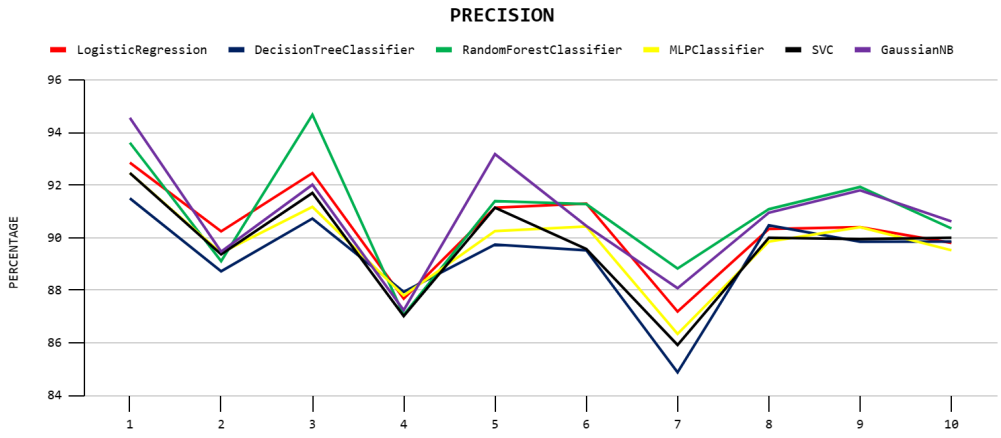

Figura 3 – Gráfico de la métrica de Precisión en los diez pliegues.

En el gráfico de precisión, Figura 4, se muestran los rendimientos de los modelos en esta métrica, alcanzando el modelo SVM y Perceptrón Multicapa el máximo en el pliegue uno. Por otro lado, el modelo Naive Bayes obtuvo el mínimo en el pliegue cuatro.

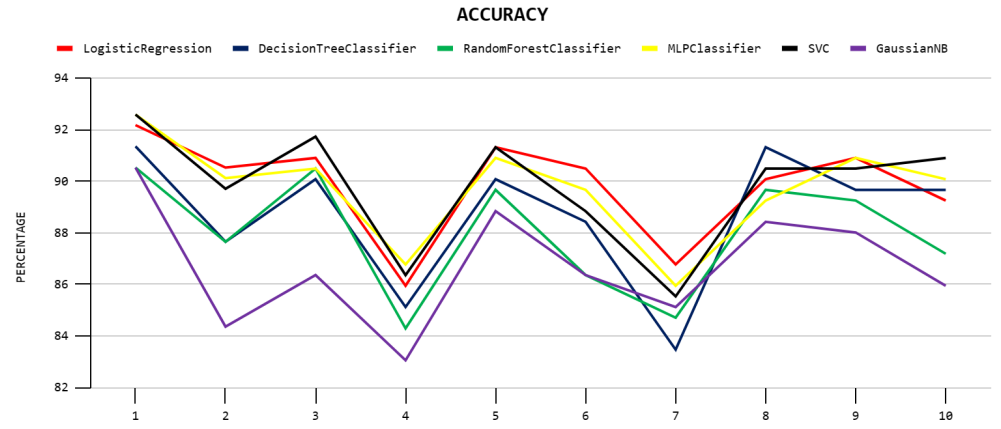

Figura 4 – Gráfico de la métrica de Exactitud en los diez pliegues.

En el gráfico de exhaustividad, Figura 5, se muestran los rendimientos de los modelos en esta métrica, siendo el modelo de Árbol de decisión el máximo en el pliegue ocho y el

modelo SVM en el pliegue diez. Por otro lado, el modelo Naive Bayes obtuvo el mínimo en el pliegue dos.

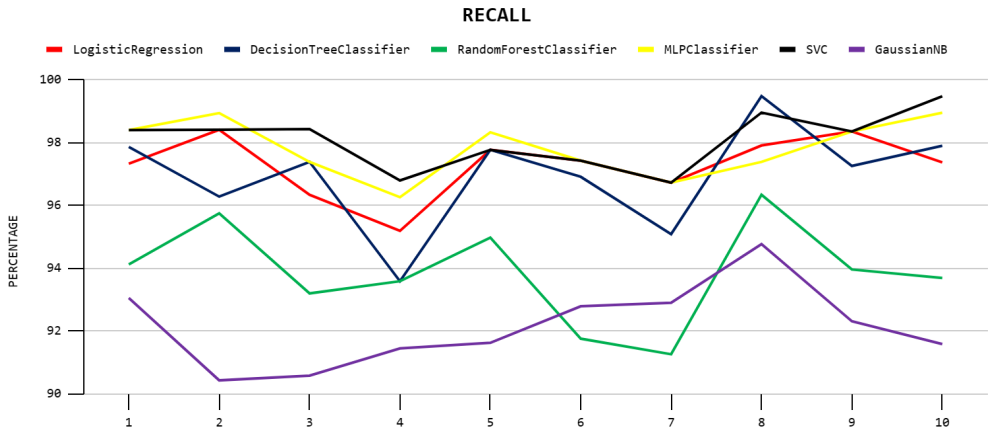

Figura 5 – Gráfico de la métrica de Exhaustividad en los diez pliegues.

A los resultados obtenidos en cada métrica se les empleó la prueba estadística de Shapiro-Wilk, en la que se verificó la normalidad para cada caso, seguidamente, se realizó la prueba de Levene para confirmar que los datos tengan iguales varianzas, en ambos casos, se reflejó una tendencia a superar las pruebas en la mayor parte de los casos. Cumplidas estas dos condiciones, aplicamos la prueba ANOVA para Muestras Repetidas, se encontró que existe una diferencia significativa en al menos dos medias dentro del grupo de resultados para las tres métricas.

Sin embargo, aún es necesario conocer dónde están estas diferencias. Por lo tanto, aplicamos la prueba T de Student para muestras relacionadas, considerando un intervalo de confianza del 99%, comparando de par en par los modelos y agrupando las técnicas que no presentan diferencias significativas. Así se forman grupos que van desde los mejores resultados (Grupo 1) a los peores, como se muestra en la Tabla 3.

Técnica/Métrica	**Precisión**		**Exactitud**		**Exhaustividad**	
	Grupo	**Promedio**	**Grupo**	**Promedio**	**Grupo**	**Promedio**
Regresión Logística	1	90.34	1	89.84	1	97.28
Árboles de Decisión	2	89.32	2	88.69	2	96.95
SVM	2	89.72	1	89.80	1	98.07
Random Forest	1	90.94	3	87.98	3	93.86
Perceptrón Multicapa	2	89.77	1	89.68	1	97.81
Naive Bayes	1	90.84	3	86.70	3	92.14

Tabla 3 – Descripción de los grupos y medias correspondientes a cada técnica/métrica correspondiente.

Para la métrica Precisión, el grupo que obtuvo los mejores resultados está compuesto por Naive Bayes, Random Forest y Regresión Logística. En lo que corresponde a la métrica de Exactitud, este está formado por Regresión Logística, SVM y Perceptrón Multicapa. Por último, para la métrica Exhaustividad, el grupo está formado por SVM, Perceptrón multicapa y Regresión logística.

A partir de estos resultados, se observa que todos los modelos obtienen buenos resultados, sin embargo, al realizar la comparación, la Regresión Logística aparece continuamente en el grupo de los mejores resultados para las tres métricas, lo que la convierte en el algoritmo que proporciona el mejor modelo en este dataset. Le siguen Support Vector Machine y Perceptrón Multicapa, que sólo aparecen en el primer grupo para las métricas Exactitud y Exhaustividad. Esto coincide con lo encontrado en (Ayala, Valenzuela y Espinosa, 2020), donde la Regresión Logística obtuvo las mayores puntuaciones entre otras técnicas propuestas para predecir la deserción universitaria, también coincide con los resultados obtenidos por Contreras, Fuentes y Rodríguez (2020) donde las técnicas de SVM y Red Neuronal Perceptrón obtuvieron buenos resultados, además en el presente estudio se superó por mucho el desempeño de los mismos modelos aplicados.

Por otra parte, comparando con los resultados presentados en el trabajo realizado por Martin et al. (2021), en el cual se utilizó el mismo dataset, se puede observar una gran mejora en los desempeños de los modelos de Regresión Logística, SVM, Árbol de Decisión y Random Forest, superando incluso el desempeño de los modelos Boosting.

5. Conclusiones

En el presente estudio se hizo un análisis comparativo de lo obtenido por los seis modelos propuestos utilizando las métricas de precisión, exactitud y exhaustividad, aplicando diferentes métodos estadísticos para encontrar los mejores modelos para cada una de las métricas analizadas. Se llegó a las siguientes conclusiones:

- Las técnicas de Machine Learning son herramientas importantes en diversos campos, entre ellos la deserción universitaria.
- Las técnicas de Machine Learning propuestas realizan la clasificación o regresión de la deserción universitaria con buenos resultados.
- A pesar de la eficacia de las técnicas de aprendizaje automático, algunas de ellas arrojaron resultados inferiores en comparación con otras. Este es el caso de las técnicas Árbol de decisión y Naive Bayes, que obtuvieron los mejores resultados en términos de precisión, pero los peores en términos de precisión y exhaustividad.
- La técnica de Regresión Logística es la que proporciona los mejores resultados con un intervalo de confianza del 99%, en comparación con las demás técnicas consideradas en el análisis.
- Dados los resultados obtenidos con distintas técnicas de Machine Learning, en futuros trabajos pretendemos realizar estudios que profundicen en otras ramas de la Inteligencia Artificial, como el Deep Learning.

Referencias

Ayala-Yaguara, H. Y., Valenzuela-Sabogal, G. M., & Espinosa-García, A. (2020). Obtención de un modelo de minería de datos aplicado a la deserción universitaria del programa de Ingeniería de Sistemas de la Universidad de Cundinamarca. *Revista Ontare*, *7*, 134–150. https://doi.org/10.21158/23823399.v7.n0.2019.2676

Bedregal Alpaca, N., Aruquipa Velazco, D., & Cornejo Aparicio, V. (2020). Técnicas de data mining para extraer perfiles comportamiento académico y predecir la deserción universitaria. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *(E30)*, 592–604. https://www.proquest.com/scholarly-journals/técnicas-de-data-mining-para-extraer-perfiles/docview/2385757429/se-2

Borja-Robalino, R., Monleon-Getino, A., Monleón-Getino, A., & Rodellar, J. (2020). *Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning*. RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao. (E30), 184–196. https://www.risti.xyz/issues/ristie30.pdf

Carmona, E. J. (2016). *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. https://www.researchgate.net/publication/263817587

Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, *13*(5), 233–246. https://doi.org/10.4067/S0718-50062020000500233

Dimitoglou, G., Adams, J. A., & Jim, C. M. (2012). Comparison of the C4.5 and a Naive Bayes Classifier for the Prediction of Lung Cancer Survivability Index Terms-Data mining, mining methods and algorithms, text mining. *Journal of Computing*, *4*(8). https://doi.org/10.48550/arXiv.1206.1121

Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: from early developments to recent advancements. *Systems Science & Control Engineering*, *2*(1), 602–609. https://doi.org/10.1080/21642583.2014.956265

Ferreyra, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., & Urzúa, S. (2017). *At a Crossroads: Higher Education in Latin America and the Caribbean*. World Bank. https://doi.org/10.1596/978-1-4648-1014-5

Lening, C. G. (2022). *Non-completion in Postsecondary education: Why are so many students not finishing their courses?* Centre for the Study of Science and Innovation Policy. https://www.schoolofpublicpolicy.sk.ca/csip/publications/making-waves/non-completion-in-postsecondary-education-why-are-so-many-students-not-finishing-their-courses.php

Lizares Castillo, M. (2017). *Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico*. https://hdl.handle.net/20.500.12672/7122

López Martínez, J. G., & Méndez Aguirre, Ó. A. (2019). *Técnicas de Machine learning para la predicción de desempeño académico en el desarrollo del espacio proyectivo del pensamiento espacial*. http://repository.pedagogica.edu.co/handle/20.500.12209/11451

Lovón Cueva, M. A., & Cisneros Terrones, S. A. (2020). Repercusiones de las clases virtuales en los estudiantes universitarios en el contexto de la cuarentena por COVID-19: El caso de la PUCP. *Propósitos y Representaciones*, *8*(SPE3). https://doi.org/10.20511/pyr2020.v8nSPE3.588

Martins, M. V, Tolledo, D., Machado, J., Baptista, L. M. T., & Realinho, V. (2021). Early Prediction of student's Performance in Higher Education: A Case Study. In Á. Rocha, H. Adeli, G. Dzemyda, F. Moreira, & A. M. Ramalho Correia (Eds.), *Trends and Applications in Information Systems and Technologies* (pp. 166–175). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-030-72657-7_16

Mirna, E. M. G.. (2021). Factores condicionantes de la deserción universitaria. *Ciencia Latina Revista Científica Multidisciplinar*, *5*(4), 5316–5328. https://doi.org/10.37811/cl_rcm.v5i4.691

Parrino, M. C. (2014). Factores intervinientes en el Fenómeno de la Deserción Universitaria. *Revista Argentina de Educación Superior*, *8*. https://dialnet.unirioja.es/servlet/articulo?codigo=4753784

Ramchoun, H., Amine, M., Idrissi, J., Ghanou, Y., & Ettaouil, M. (2016). Multilayer Perceptron: Architecture Optimization and Training. *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(1), 26. https://doi.org/10.9781/ijimai.2016.415

Ruiz Echeverry, P. (2020). Consideración de deserción universitaria en estudiantes de Comunicación Social. Un estudio de caso. *Revista Nexus Comunicación*, 1–25. https://doi.org/10.25100/nc.v0i28.10643

Solís, M., Moreira, T., González, R., Fernández, T., & Hernández, M. (2018). Perspectives to Predict Dropout in University Students with Machine Learning. *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, (pp. 1–6). https://doi.org/doi:10.1109/iwobi.2018.8464191

Tocto, P., Huamaní, G. T., & Zuloaga, L. (2023). Aplicación aprendizaje automático en la gestión universitaria: Modelo de clasificación de la deserción de los estudiantes en Ingeniería en el Perú. *Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology*, *2023-July*. https://doi.org/10.18687/LACCEI2023.1.1.1332

Upasani, D. E., & Virendra, V. S. (2020). *Machine Learning with Python*. IPH.

Valentim, R., Machado, J., Baptista, L., & Martins, M. V. (2021). *Predict students' dropout and academic success*. Zenodo. https://zenodo.org/record/5777340#.Y7FJotJBwUE

Viale Tudela, H. E. (2014). Una Aproximación Teórica A La Deserción Estudiantil Universitaria. *Revista Digital de Investigación En Docencia Universitaria*, *8*(1), 59–76. https://doi.org/10.19083/ridu.8.366