Revista Innovación y Software
Vol. 2, No. 2, Mes Septiembre-Febrero, 2021
ISSN: 2708-0935
Pág. 6-13
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos cortos
Temática: Inteligencia artificial
Recibido: 05/04/2021 | Aceptado: 08/06/2021 | Publicado: 30/09/2021

Identificadores persistentes:
ARK: ark:/42411/s6/a40
PURL: 42411/s6/a40

Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica

Predictive model for the early detection of students with high risk of academic dropout

Kevin Rivera Vergaray [1][0000-0001-5393-4382]*

[1] Universidad Nacional Mayor de San Marcos. Huaraz-Perú. kevin.rivera1@unmsm.edu.pe

* Autor para correspondencia: kevin.rivera1@unmsm.edu.pe

Resumen
Se comparan los resultados de 4 modelos predictivos, de regresión logística, árboles de decisión, KNN y una red neuronal para predecir la deserción académica de estudiantes en la Universidad Nacional Intercultural de la Amazonía, aplicado a un dataset extraído de la base de datos del sistema de gestión académica de la universidad, que contiene datos socioeconómicos y de rendimiento académico los cuales fueron procesados y formateados utilizando técnicas de onehotencoding para así poder aplicar los modelos predictivos ya mencionados. Para el procesamiento y formateo de datos se utilizó consultas Transac Sql y la aplicación de los modelos predictivos se hizo a través del Software Knime y utilizando Python a través de Google Colab. Los resultados obtenidos al aplicar 4 modelos predictivos son muy buenos ya que todos superaron el 80% de Accuracy, lo cual garantiza que puedan ser puestos en producción para el beneficio de la universidad y así pueda tomar mejores decisiones a la hora de abordar la deserción académica. Se concluye que aplicar un modelo predictivo en las universidades para la detección temprana de estudiantes con alto riesgo de deserción académica es viable y muy beneficioso para que las universidades a través de sus gestores académicos puedan aplicar estrategias mas focalizadas para reducir sus índices de deserción académica.

Palabras clave: Deserción académica, Modelo predictivo, Dataset.

Abstract
The results of 4 predictive models, logistic regression, decision trees, KNN and a neural network are compared to predict the academic dropout of students at the National Intercultural University of the Amazon, applied to a dataset extracted from the system's database. of academic management of the university, which contains socioeconomic and academic performance data which were processed and formatted using onehotencoding techniques in order to apply

the predictive models already mentioned. For data processing and formatting, Transac Sql queries were used and the application of predictive models was done through Knime Software and using Python through Google Colab. The results obtained by applying 4 predictive models are very good since they all exceeded 80% of Accuracy, which guarantees that they can be put into production for the benefit of the university and thus can make better decisions when addressing academic dropout. . It is concluded that applying a predictive model in universities for the early detection of students with high risk of academic dropout is viable and very beneficial so that universities, through their academic managers, can apply more focused strategies to reduce their academic dropout rates.
Keywords: *Academic dropout, Dataset, Predictive model*

Introducción

La deserción estudiantil universitaria no es un problema nuevo ni exclusivo del Perú. Este fenómeno se da en todo el mundo, es un viejo problema que tiene muchas variables y el cual no es preocupación exclusiva del mundo académico. La deserción estudiantil universitaria trae como consecuencia el aumento del número de alumnos con educación superior incompleta que se incorporan al mundo laboral y se convierten en sub empleados sin obtener los ingresos deseados; lo cual, perjudica al mismo estudiante, a sus familiares, al país y a la universidad pues esta ve afectado su presupuesto [1].

No importa el tipo de universidad o casa de estudio. Hay factores muy comunes que disminuyen las tasas de retención estudiantil en la educación superior. Pueden ser problemas individuales o una mezcla de factores. Por eso, las facultades deben trabajarlas de manera adecuada para reducir la deserción [2].

Actualmente el 40% de estudiantes de la UNIA son indígenas y el otro 60% de los estudiantes son mestizos (en su gran mayoría viven en el casco urbano) No se cuenta con información procesada de las deserciones por semestre académico. Se percibe la mayor cantidad de deserciones en las carreras de ingeniería, las principales causas, la dificultad de la carrera y la situación económica del estudiante. Es así que se pretende aplicar conceptos de machine learning y análisis de datos con la finalidad de predecir la deserción académica y así tomar las previsiones necesarias para reducirla. En el presente artículo se utilizó la base de datos del Sistema de Gestión Académica de la UNIA.

Materiales y métodos o Metodología computacional

En el presente trabajo de probaron con 4 modelos predictivos de machine learning para poder predecir el riesgo de deserción académica.

Deberá entenderse por deserción estudiantil o deserción académica, el abandono definitivo de las aulas de clase por diferentes razones y la no continuidad en la formación académica, que la sociedad quiere y desea en y para cada persona que inicia sus estudios universitarios [3].

- **Regresión logística:** La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo. Resulta útil su empleo cuando se tiene una variable dependiente dicotómica (un atributo cuya ausencia o presencia hemos puntuado con los valores cero y uno, respectivamente) y un conjunto de variables predictoras o independientes, que pueden ser cuantitativas (que se denominan covariables o covariadas) o categóricas. En este último caso, se requiere que sean transformadas en variables "dummy", es decir variables simuladas. El propósito del análisis consiste en: predecir la probabilidad de que a alguien le ocurra cierto "evento": por ejemplo, estar desempleado =1 o no estarlo = 0, ser pobre = 1 o no pobre = 0, recibirse de sociólogo =1 o no recibirse = 0). Determinar que variables pesan más para aumentar o disminuir la probabilidad de que a alguien le suceda el evento en cuestión Esta asignación de probabilidad de ocurrencia del evento a un cierto sujeto, así como la determinación del peso que cada una de las variables dependientes en esta probabilidad, se basan en las características que presentan los sujetos a los que, efectivamente, les ocurren o no estos sucesos [4].
- **Arboles de decisión:** Los árboles de decisión proveen de una herramienta de clasificación muy potente. Su uso en el manejo de datos la hace ganar en popularidad dadas las posibilidades que brinda y la facilidad con que son comprendidos sus resultados por cualquier usuario. El árbol en sí mismo, al ser obtenidos, determinan una regla de decisión. Esta técnica permite:

1. **Segmentación**: establecer que grupos son importantes para clasificar un cierto ítem.
2. **Clasificación**: asignar ítems a uno de los grupos en que está particionada una población.
3. **Predicción**: establecer reglas para hacer predicciones de ciertos eventos.
4. **Reducción de la dimensión de los datos**: Identificar que datos son los importantes para hacer modelos de un fenómeno.
5. **Identificación-interrelación**: identificar que variables y relaciones son importantes para ciertos grupos identificados a partir de analizar los datos.
6. **Recodificación**: discretizar variables o establecer criterios cualitativos perdiendo la menor cantidad posible de información relevante [5].

- **KNN:** K-Nearest-Neighbor es un algoritmo basado en instancia de tipo supervisado de Machine Learning. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos). Al ser un método sencillo, es ideal para introducirse en el mundo del Aprendizaje Automático. Sirve esencialmente para clasificar valores buscando los puntos de datos "más similares" (por cercanía) aprendidos en la etapa de entrenamiento y haciendo conjeturas de nuevos puntos basado en esa clasificación [6].
- **Red Neuronal:** Las redes neuronales Artificiales (RNAs) son modelos computacionales como un intento de conseguir formalizaciones matemáticas acerca de la estructura del cerebro. Las RNAs imitan la estructura hadware del sistema nervioso, centrándose Enel funcionamiento el cerebro humano, basado en el aprendizaje a través de la experiencia, con la consiguiente extracción de conocimiento a partir de la misma [7].

Para aplicar los modelos mencionados se utilizó un dataset, extraído a través de consultas SQL de la base de datos del sistema académico de la Universidad Nacional Intercultural de la Amazonía, dicho dataset este compuesto con datos registrados desde el 2005, en total se generaron 17 variables junto con el target "desercion". La tabla de datos contiene 18 columnas y 5803 filas, con un peso de 680.2 kb.

Tabla 1: Contenido de la data académica procesada

N°	Variable	Descripción
1	codigo	Código que identifica al estudiante.
2	sexo	Sexo del estudiante, 1 es masculino y 0 femenino.
3	mestizo	Si el estudiante es mestizo 1 y si no lo es 0.
4	indigena	Si el estudiante es indigena 1 y si no lo es 0.
5	pobre	Si el estudiante es pobre 1 y si no lo es 0.
6	pobre_extremo	Si el estudiante es pobre extremo 1 y si no lo es 0.
7	no_pobre	Si el estudiante es No pobre 1 y si no lo es 0.
11	educacion	Si el estudiante es de la facultad de educación 1 y si no lo es 0.
12	ingenieria	Si el estudiante es de la facultad de ingeniería 1 y si no lo es 1.
13	matriculas	Número de matriculas que tiene el estudiante durante su estadía en la universidad.
14	matriculas_aprobadas	Número de matriculas aprobadas que tiene el estudiante durante su estadía en la universidad.
15	matriculas_desaprobadas	Número de matriculas desaprobadas que tiene el estudiante durante su estadía en la universidad.
16	egresado	Si el estudiante es egresado 1 y si aun no lo es 0.
17	ponderado	Promedio ponderado acumulado del estudiante
18	semestres	Numero de semestres que se ha ,atriculado el estudiante.
19	desercion	1 si el estudiante deserto 0 si el estudiante no deserto.

Fuente: Elaboración propia.

Para aplicar los modelos se utilizó el software KNIME y a través de Python usando Google Colab.

Revista Innovación y Software
Vol. 1, No. 2, Mes Septiembre-Febrero, 2021
ISSN: 2708-0935
Pág. 6-13
https://revistas.ulasalle.edu.pe/innosoft

Resultados y discusión

Luego de aplicar y evaluar los modelos aplicados al conjunto de datos extraídos de la base de datos del sistema académico, los mejores resultados se obtuvieron con el KNN y el árbol de decisión. En el modelo KNN se obtuvo un Accuracy de 88,844% y un error de 11,156%, y en el modelo de árbol de decisión se obtuvo un Accuracy de 88,4% y un Error de 11,6%.

A continuación, se detallan los resultados obtenidos por cada modelo aplicado.

a) **Regresión logística:** Al aplicar regresión logística se obtuvo un Accuracy de 84,57% y un error de 15,43%.

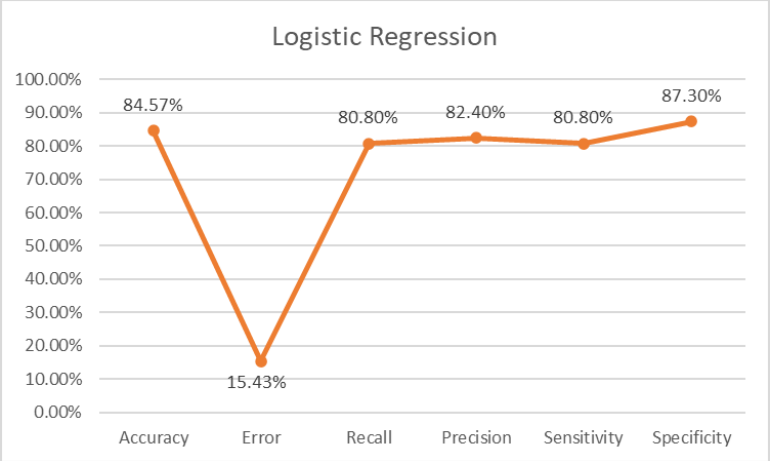

Figura 1: Resultados obtenidos aplicando regresión logística.

b) **Árbol de decisión:** Al aplicar arboles de decisión se obtuvo un Accuracy de 88,4% y un Error de 11,6%.

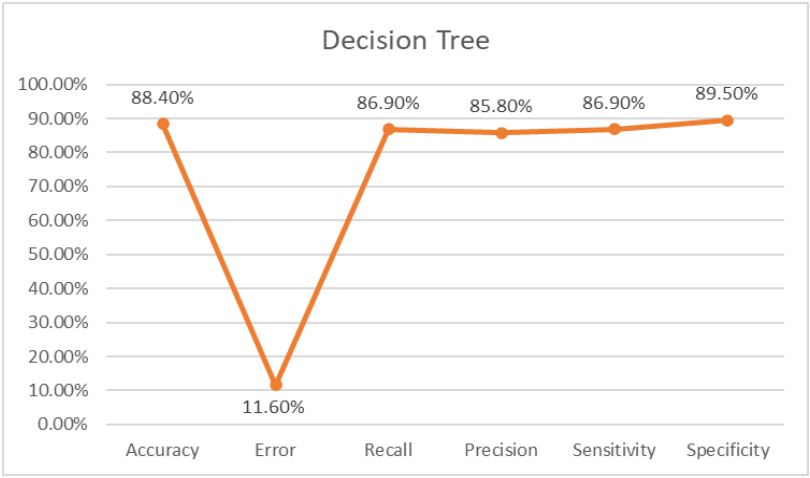

Figura 2: Resultados obtenidos aplicando arboles de decisión.

c) KNN: Al aplicar KNN se obtuvo un Accuracy de 88,4% y un Error de 11,6%.

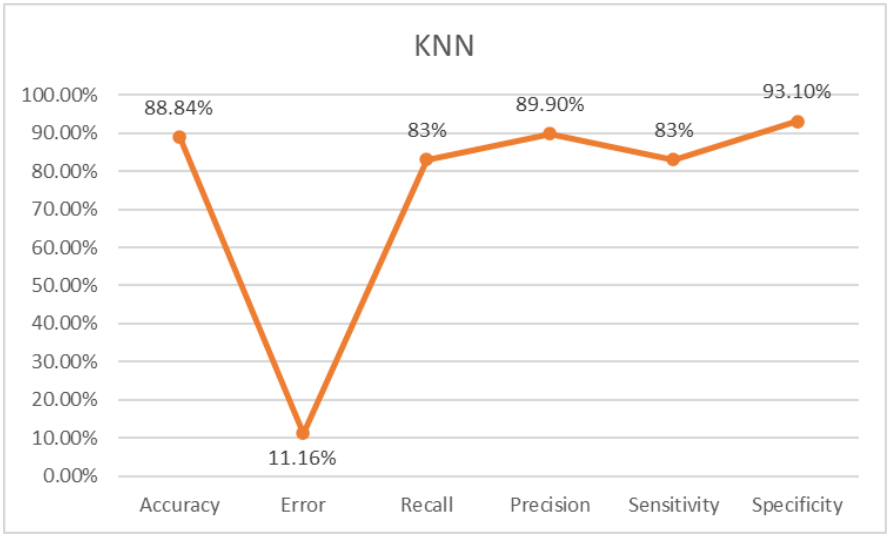

Figura 3: Resultados obtenidos aplicando KNN.

d) Red neuronal: Al aplicar la Red Neuronal se obtuvo un Accuracy de 82,33% y un Error de 17,67%.

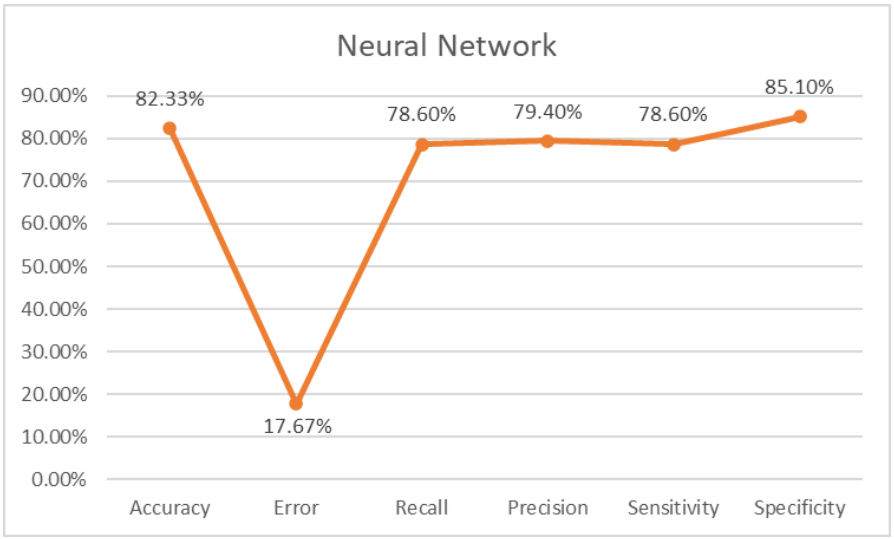

Figura 4: Resultados obtenidos aplicando una red neuronal.

Revista Innovación y Software
Vol. 1, No. 2, Mes Septiembre-Febrero, 2021
ISSN: 2708-0935
Pág. 6-13
https://revistas.ulasalle.edu.pe/innosoft

Se puede ver que en los 4 modelos predictivos aplicados se obtuvieron un Accuracy mayor al 80%, lo cual en gran parte es por el trabajo previo que se realizó para preparar la data utilizada, en el cual se aplicaron técnicas de onehotencoding entre otros. Finalmente presentamos una comparación de los resultados obtenidos con los 4 modelos predictivos utilizados.

Tabla 2: Cuadro comparativo de los resultados obtenidos.

Medida	Logistic Regression	Decision Tree	KNN	Neural Network
Accuracy	84.57%	88.40%	88.84%	82.33%
Error	15.43%	11.60%	11.16%	17.67%
Recall	80.80%	86.90%	83%	78.60%
Precision	82.40%	85.80%	89.90%	79.40%
Sensitivity	80.80%	86.90%	83%	78.60%
Specificity	87.30%	89.50%	93.10%	85.10%

Fuente: Elaboración propia.

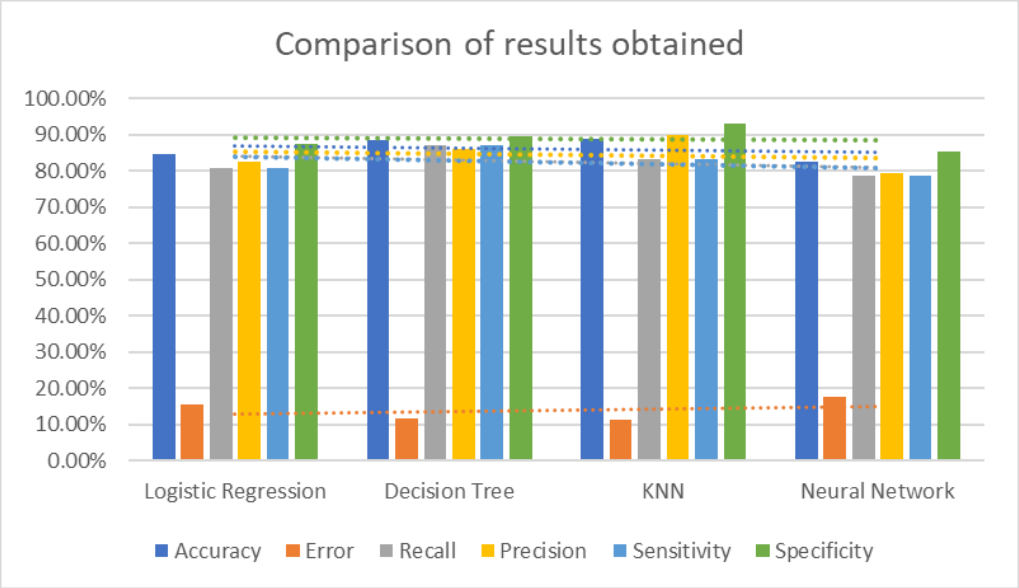

Figura 5: Comparación de resultados obtenidos.

Conclusiones

Se determina que el modelo de KNN y Arboles de decisión son los que describen un mayor ajuste de los datos analizados, estos modelos presentan un Accuracy bastante aceptables de 88.844 % y 88.4%.

Los modelos predictivos aplicados a datos académicos, como en nuestro caso pueden ayudar en la generación de mejores estrategias para los problemas que aquejan a las universidades, sobre todo públicas, como es la deserción académica.

El modelo debe analizar más datos referidos a la enseñanza no presencial implementada en las universidades a causa de la pandemia por el COVID-19, lo que posiblemente influya en los resultados obtenidos con la data utilizada.

Agradecimientos

Mi agradecimiento al curso de Tópicos Avanzados en Ingeniería de Software de la Maestría en Ingeniería de Sistemas e Informática con mención en Ingeniería de Software en UNMSM. Mi mayor agradecimiento a mi familia.

Referencias

[1] T. Viale, «Enfoque UPC,» 10 enero 2020. [En línea]. Available: https://enfoque.upc.edu.pe/mas-temas/educacion/desercion-estudiantil-universitaria-accionamos-o-reaccionamos/.

[2] uPlanner, «uPlanner,» 27 Marzo 2017. [En línea]. Available: https://uplanner.com/es/blog/8-causas-de-desercion-estudiantil-en-la-educacion-superior/.

[3] Gabriel, Jaime, Páramo, Arturo y Correa, Deserción Estudiantil Universitaria. Conceptualización, Revista Universidad Eafit, 1999.

[4] H. Chitarroni, La regresión logística, Buenos Aires: Instituto de Investigación en Ciencias Sociales, 2002.

[5] C. N. Bouza y A. Santiago, MODELACIÓN MATEMÁTICA DE FENÓMENOS DEL MEDIO AMBIENTE Y LA SALUD, Mexico: Universidad Autónoma de Guerrero, 2012.

[6] aprendemachinelearning, «Clasificar con K-Nearest-Neighbor ejemplo en Python,» July 2018. [En línea]. Available: https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/.

[7] R. F. López y J. M. F. Fernández, Las Redes Neuronales Artificiales - Fundamentos teoricos y aplicaciones prácticas., España: lorena bello, 2008.

Roles de Autoría

Kevin Rivera Vergaray: Conceptualización, Curación de datos, Investigación, Metodología, Software, Validación, Redacción - borrador original.