

Modelo predictivo de machine learning para el servicio de orientación vocacional hacia los estudiantes de secundaria en la Gerencia Regional de Trabajo y Promoción del Empleo Cusco 2024

Machine learning predictive model for vocational guidance service for high school students in the Regional Management of Labor and Employment Promotion Cusco 2024

Malu Beatriz Silva-Zarate ^A, Mario Aquino-Cruz ^B

Resumen— Este trabajo de investigación comparó tres modelos predictivos de Machine Learning: Árbol de Decisiones, Redes Neuronales y Regresión Logística, con el objetivo de identificar cuál fue el más adecuado para mejorar el servicio de orientación vocacional dirigido a estudiantes de secundaria. La orientación vocacional ha sido un desafío global, ya que muchos jóvenes no tenían claridad sobre qué carrera seguir, lo que a menudo resultaba en decisiones equivocadas y arrepentimientos posteriores. A través del test vocacional IEPPPO, se evaluaron las respuestas de los estudiantes utilizando la metodología KDD. Los modelos fueron entrenados y evaluados utilizando Python, utilizando métricas como exactitud, precisión y recall. Los resultados indicaron que el modelo Árbol de Decisiones fue el más efectivo, alcanzando una exactitud del 99.82%, una precisión del 99.83% y un recall del 99.82%, superando a las Redes Neuronales y la Regresión Logística. Estas conclusiones resaltaron la capacidad del Árbol de Decisiones para mejorar el proceso de orientación vocacional, proporcionando una herramienta precisa y confiable para guiar a los estudiantes en su elección de carrera.

Palabras clave: Árbol de decisiones, Orientación vocacional, Redes neuronales, Regresión logística.

Abstract— This research work compared three predictive Machine Learning models: Decision Tree, Neural Networks and Logistic Regression, with the objective of identifying which one was the most suitable to improve the vocational guidance service aimed at high school students. Vocational guidance has been a global challenge, as many young people were unclear about what career to pursue, often resulting in wrong decisions and later regrets. Through the IEPPPO vocational test, students' responses were assessed using KDD methodology. The models were trained and evaluated using Python, using metrics such as accuracy, precision, and recall. The results indicated that the Decision Tree model was the most effective, achieving an accuracy of 99.82%, precision of 99.83% and recall of 99.82%, outperforming Neural Networks and Logistic Regression. These findings highlighted the ability of the Decision Tree to improve the vocational guidance process, providing an accurate and reliable tool to guide students in their career choice.

Keywords: Decision tree, Vocational orientation, Neural networks, Logistic regression.



Revista de Investigación en Ciencia y Tecnología
ISSN: 2810-8124 (en línea) / ISSN: 2706-543x
Universidad Nacional Micaela Bastidas de Apurímac – Perú

Vol. 6 Núm. 2 (2024) - Publicado: 13/08/24 - [Indexaciones](#)
Número: doi.org/10.57166/riqchary/V6.n2.2024
Pág. 41- 48 | Recibido 12/10/2024 ; Aceptado 04/11/2024



doi.org/10.57166/riqchary.v6.n2.2024.128

Autores:

- A. **ORCID iD** <https://orcid.org/0009-0002-2456-150X>
Malu Beatriz Silva-Zarate, Bachiller de la Universidad Nacional Micaela Bastidas de Apurímac, malusilvaz2000@gmail.com.
- B. **ORCID iD** <https://orcid.org/0000-0002-2552-5669>
Mario Aquino-Cruz, Docente en la Universidad Nacional Micaela Bastidas de Apurímac, maquino@unamba.edu.pe

1 INTRODUCCIÓN

En los últimos años, la orientación vocacional ha cobrado una creciente relevancia en los sistemas educativos alrededor del mundo. Este proceso permite a los estudiantes explorar sus intereses, habilidades y aspiraciones, con el objetivo de tomar decisiones informadas sobre su futuro académico y profesional [1]. Sin embargo, a nivel global, se observa una notable carencia de servicios de orientación que sean eficientes y accesibles, lo que afecta negativamente el desarrollo personal y profesional de los jóvenes. Diversos estudios subrayan la importancia de mejorar la calidad y la cobertura de estos servicios, ya que una orientación inadecuada puede conducir a una elección errónea de carrera, desmotivación, y abandono escolar [2].

En el contexto latinoamericano, y específicamente en Perú, los retos en el ámbito de la orientación vocacional son evidentes. A pesar de la existencia de programas dedicados a este fin, muchos de ellos carecen de los recursos tecnológicos necesarios para optimizar su funcionamiento y ofrecer un servicio adecuado a la creciente demanda estudiantil [3]. Esta situación es particularmente evidente en la región de Cusco, donde los mecanismos tradicionales de orientación no logran cubrir eficientemente las necesidades de los estudiantes. El Servicio de Orientación Vocacional e Información Ocupacional (SOVIO) es uno de los pocos programas que intentan proporcionar este apoyo a nivel regional, pero su proceso, basado en herramientas manuales y tradicionales, presenta numerosas limitaciones [4].

El Machine learning (ML), como subdisciplina de la inteligencia artificial, tiene un impacto significativo en diversas áreas, desempeña un papel crucial al ayudar a las personas a identificar y desarrollar sus habilidades y áreas de interés de manera más eficiente. El ML analiza grandes volúmenes de datos y a partir de esta información genera recomendaciones personalizadas [5].

La presente investigación abordó esta problemática con el objetivo de evaluar y comparar tres modelos predictivos de Machine learning, Árbol de decisiones, Redes neuronales y Regresión logística, mediante las métricas de exactitud, precisión y recall, e identificó el más efectivo para mejorar la toma de decisiones vocacionales en el servicio de orientación vocacional de la Gerencia Regional de Trabajo y Promoción del Empleo Cusco (GRTPE). El uso de estas tecnologías permitirá procesar y analizar automáticamente los datos obtenidos del Inventario de Estilos Personales y Preferencias Ocupacionales (IEPPO), logrando así una mayor precisión y rapidez en los resultados.

1.1 Trabajos Relacionados

En diversos estudios, se ha explorado la utilización de modelos predictivos y técnicas de Machine learning para optimizar la toma de decisiones académicas y vocacionales. Un estudio desarrolló un modelo predictivo para la elección de una carrera profesional para estudiantes de una universidad privada en Arequipa, en el cual, validó un modelo de regresión logística, como una herramienta de ayuda para, también revelando que factores como género y la edad de los estudiantes influyen en las respuestas y elección de su carrera [6]. Otro trabajo realizó la implementación de un sistema de predicción del rendimiento académico de estudiantes de ingeniería, en el cual utilizaron técnicas de minería de datos, utilizando la metodología KDD y algoritmos de clasificación como REPTree, Vote, SMOreg, RandomForest LWL, Bagging, Kstar, ZeroR, M5P, RandomTree e IBK, siendo el algoritmo Kstar el más eficiente [7].

Se realizó una comparación de técnicas de machine learning para predecir la deserción universitaria, teniendo como métricas la medición de la precisión, la exactitud y exhaustividad, identificando que la Regresión Logística fue la técnica más efectiva en esta tarea [8]. Otro estudio elaboró un modelo predictivo para el rendimiento académico de los estudiantes universitarios, aplicando la metodología KDD y evaluó los indicadores de precisión, sensibilidad y especificidad y aplicando el uso de algoritmos como SVM y KNN mostró que SVM fue el más eficaz, logrando altos niveles de precisión y mayor asertividad [9] [10]. Así también el estudio que desarrolló un modelo predictivo que apoye el seguimiento académico de estudiantes, realizó la comparación de diversos modelos de redes neuronales alcanzando una precisión de 98.97%, demostrando ser una herramienta eficaz para su objetivo [11].

A nivel internacional, un estudio desarrolló un agente inteligente basado en machine learning para predecir la elección de carrera de los estudiantes, los resultados mostraron que la técnica de k-vecinos cercanos fue la más eficaz para la predicción [12]. Además, en otro trabajo se compararon algoritmos de aprendizaje supervisado y no supervisado para sugerencias vocacionales, concluyendo que KNN y SVM fueron los más efectivos en la clasificación de vocaciones, alcanzando tasas de precisión del 90% [13]. Finalmente, en una investigación se implementó un sistema web utilizando técnicas de machine learning para mejorar la orientación vocacional de estudiantes de secundaria, utilizando la metodología KDD con resultados satisfactorios en la recomendación de carreras profesionales [14].

2 MATERIALES Y METODOLOGÍA

2.1 Materiales

- Datos de estudiantes: 3770 evaluaciones de estudiantes de 4° y 5° de secundaria de colegios públicos y privados en la Región Cusco, obtenidos mediante el Inventario de Estilos Personales y Preferencias Ocupacionales (IEPPO).
- Equipos de cómputo: Computadoras con sistema operativo Windows 64 bits, procesador Core i5 o superior, 8GB de RAM, y 512GB de almacenamiento.
- Software:
 - Python (librerías: pandas, numpy, matplotlib, scikit-learn)
 - Google Colab para desarrollo y evaluación de modelos de machine learning.
 - Excel para procesamiento y organización de datos.
- Conexión a internet: Para el uso de Google Colab y almacenamiento en la nube.
- Cuestionario IEPPO: Instrumento para recopilar información sobre las preferencias vocacionales de los estudiantes.

2.2 Tipo y nivel de investigación

Este estudio se clasifica como una investigación aplicada con nivel descriptivo, para Hernandez Sampieri y otros este se caracteriza por su enfoque en la solución de problemas prácticos y específicos, utilizando conocimientos y teorías existentes para generar resultados con aplicaciones directas en contextos reales, y descriptivo porque medirá y recogerá información de manera independiente o conjunta sobre los conceptos o las variables a las que se refieren [15].

2.3 Diseño de la investigación

El diseño es no experimental de corte transversal, observando fenómenos en su entorno natural sin manipular variables; en este estudio no se manipularon variables de forma directa, los datos se recopilaron tal como existen en la realidad, el enfoque fue analizar esos datos utilizando modelos predictivos, sin alterar el entorno ni aplicar intervenciones experimentales.

2.4 Población y muestra

La población objetivo incluye 3770 evaluaciones de estudiantes de cuarto y quinto de secundaria de 38 colegios entre públicos y privados de la región Cusco, así también la muestra incluye el total de la población.

2.5 Proceso KDD

En la fig.1 se muestra las etapas del procedimiento KDD que se utilizó para la investigación.

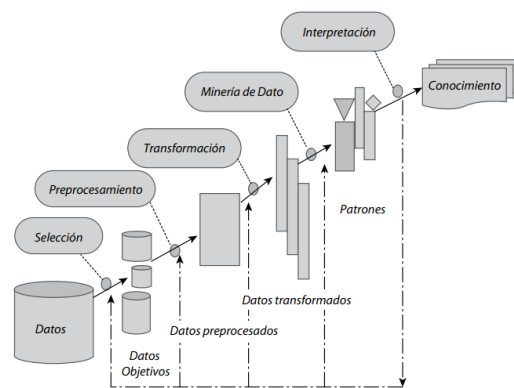


Fig. 1. Etapas del proceso KDD

2.6 Procedimiento

Primero se realizó la recopilación de los datos de estudiantes de 4to y 5to de secundaria de colegios públicos y privados en las 13 provincias de la Región Cusco, a través de la Gerencia Regional de Trabajo y Promoción del Empleo Cusco, los datos específicos provienen de la aplicación del test vocacional IEPPO que son de naturaleza sensible y privada; consta de 118 preguntas, realizados a los estudiantes entre agosto y diciembre de 2023, tanto de manera presencial y también mediante la aplicación de Google formulario. Se obtuvo un conjunto de datos de 3770 evaluaciones, para su posterior clasificación en los tipos vocacionales, los datos fueron almacenados en documentos Excel por cada colegio.

Segundo, las respuestas de los test vocacionales IEPPO están divididas en tres partes: 33 preguntas en estilos personales, 47 preguntas en actividades de preferencia y 38 preguntas en percepción de habilidad, cada división fue respondida con respuesta de:

- No se parece a mí.
- Se parece a mí.
- Me interesa.
- Me interesa poco o nada.
- Soy hábil.
- No soy hábil.

Los datos fueron transformados y preparados para su análisis, lo que incluyó la normalización y estandarización de las variables que fueron utilizadas

Tercero, teniendo ya los datos clasificados según los 7 tipos vocacionales como liderazgo, técnico mecánico, social, organizado, artístico, emprendimiento e investigación, se agregó una variable nueva en la base de datos, con un dato que no incluye el test IEPPO, esta variable trata sobre la carrera profesional universitaria, técnica u ocupacional que le gustaría al estudiante ejercer, puesto que el profesional del área, la psicóloga mencionó que esta información es muy importante para la toma de decisión del tipo vocacional a sugerir. Se realizó la eliminación de datos puesto que los estudiantes no completaron este campo; posteriormente se procedió a realizar la selección y limpieza de la base de datos, garantizando la calidad; este proceso incluyó también la eliminación

de datos duplicados, eliminación de test incompletos y corrección de errores, quedando un total de 2751 datos.

Cuarto, los datos finalmente obtenidos fueron normalizadas y estandarizadas para facilitar su uso en los modelos predictivos, se dividieron en tres segmentos, 70% para el entrenamiento, obteniendo un total de 1926 test vocacionales, 20% para la validación, obteniendo un total de 550 test vocacionales y 10% para el testeo, obteniendo 275 test vocacionales.

Quinto, se realizó el entrenamiento con los datos, utilizando la herramienta Python; se realizó en una computadora de 64 bit con sistema operativo Windows, se desarrolló con los tres modelos predictivos de machine learning (árboles de decisión, redes neuronales y regresión logística), los modelos fueron entrenados con el conjunto de entrenamiento y validados con conjunto de validación.

La arquitectura utilizada para el modelo de predicción basado en Redes neuronales se puede observar en la fig.2.

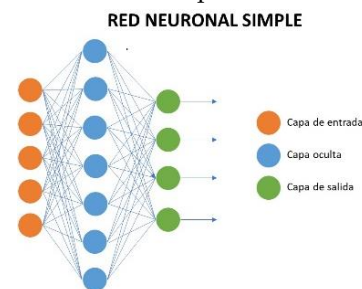


Fig. 2. Arquitectura utilizada para la Red Neuronal (Multilayer Perceptron, MLP) [16]

- Capa de Entrada (Input Layer): El número de neuronas en esta capa está determinado por el número de características (columnas) de los datos de entrada que son 119 neuronas.
- Capa Oculta (Hidden Layer): Por defecto, se utilizó una capa oculta con 100 neuronas. Las neuronas en esta capa se conectaron a todas las neuronas de la capa de entrada y aplicaron la transformación no lineal definida por la función de activación, esta convirtió las entradas negativas en 0 y dejó pasar los valores positivos.
- Capa de Salida (Output Layer): El número de neuronas en la capa de salida se determinó por el número de clases, en este caso se tiene 7 tipos vocacionales, por lo tanto, la capa de salida tendrá 7 neuronas, permitiendo clasificar cada estudiante en una de las categorías.
- En la codificación realizada se utilizó parámetros para lograr un balance entre rendimiento y tiempo de computación, con el hiperparámetro `hidden_layer_sizes` configurado su valor a 100, lo que implica una sola capa oculta con 100 neuronas, `random_state` establecido en 42 para garantizar la reproducibilidad del proceso, `max_iter` el cual se incrementó a 9000 iteraciones para asegurar la convergencia del conjunto de datos.

En la fig.3 se muestra la arquitectura que se utilizó para el modelo de predicción de Árbol de decisiones.

Malu Beatriz Silva-Zarate, Mario Aquino-Cruz

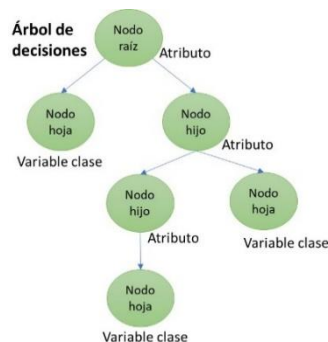


Fig. 3. Arquitectura utilizada para Árbol de decisiones

- Cada nodo interno representa una decisión basada en una característica de los datos (un nodo a una pregunta).
- Los nodos hoja representan las decisiones finales, que en este caso son los tipos vocacionales.
- Cada rama del árbol indica el resultado de una decisión, lo que lleva a otra pregunta.
- En la codificación realizada, el árbol creció automáticamente al entrenar el modelo, con el parámetro `random_state=42` para asegurar la reproducibilidad del resultado, se utilizó el criterio "gini" (impureza de Gini) para decidir las divisiones en el árbol.

En la fig.4 se muestra la arquitectura que se utilizó para el modelo de predicción de Regresión logística multiclase.

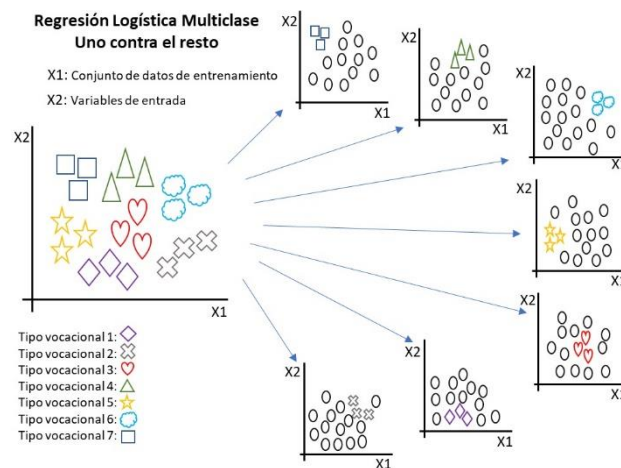


Fig. 4. Arquitectura utilizada para Regresión logística.

- Capa de entrada: Cada columna del test vocacional se convirtió en una característica para el modelo, los datos categóricos fueron convertidos en numéricos.
- Procesamiento - Uno contra el resto: Para cada clase vocacional, se generó un clasificador que intentó predecir si una observación pertenece a esa clase o no; se realizó la aplicación de la función logística y ajuste de pesos.
- Capa de salida: Se seleccionó la clase final con la mayor probabilidad de pertenecer a una de los 7 tipos vocacionales entre todas las predicciones de los clasificadores.

- d) Para la codificación se aseguró la convergencia y facilidad para reproducir los resultados, utilizando `max_iter`: incrementado a 1000 para permitir la convergencia del modelo teniendo en cuenta en tamaño del conjunto de datos y también el `random_state` configurado con un valor de 42 para la reproducibilidad.

2.7 Medición de desempeño

Para esta sección se ha utilizado tres métricas de rendimiento para evaluar la eficacia de nuestros modelos predictivos, incluyendo la exactitud (accuracy), la precisión (precision), y el recall (sensibilidad o recuperación) [17]. Las ecuaciones empleadas para calcular cada métrica son las siguientes:

- Exactitud (Accuracy): Esta métrica mide la proporción total de predicciones correctas (verdaderos positivos y verdaderos negativos) con respecto al número total de predicciones.

$$Exactitud = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precisión (Precision): La precisión refleja cuántas de las predicciones positivas realizadas por el modelo fueron realmente correctas.

$$Precision = \frac{TP}{TP + FP}$$

- Recall (Sensibilidad): El recall mide la capacidad del modelo para identificar correctamente todas las muestras positivas.

$$Recall = \frac{TP}{TP + FN}$$

En estas ecuaciones, TP, TN, FP y FN representan las cantidades de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos, respectivamente.

Estas métricas fueron fundamentales para la evaluación del rendimiento de los modelos predictivos como: árboles de decisión, redes neuronales y regresión logística, ya que cada uno de estos algoritmos tuvo un comportamiento diferente en términos de exactitud, precisión y recall. El análisis conjunto de estas métricas nos permitieron seleccionar el modelo que mejor se ajusta a nuestros objetivos.

3 RESULTADOS

3.1 Evaluación de Modelos Predictivos

En este estudio se ha llevado a cabo un proceso de entrenamiento y evaluación de tres modelos predictivos: Árbol de Decisiones, Regresión Logística y Redes Neuronales, con el objetivo de determinar cuál es más efectivo para clasificar a los estudiantes de secundaria en siete tipos vocacionales basados en los resultados de un test vocacional. Para ello, se utilizó una base de datos de 2751 evaluaciones, se fue preprocesando las respuestas textuales mediante codificación para convertirlas en datos numéricos, lo que permitió su posterior análisis con los modelos de clasificación; se codificaron tanto

las respuestas a las preguntas del test como la variable objetivo (el tipo vocacional), utilizando LabelEncoder para convertir las respuestas en cadenas a valores numéricos. A continuación, se presentan los resultados obtenidos para cada uno de ellos en términos de exactitud, precisión y recall.

3.1.1 Resultados del Árbol de Decisiones

El modelo de Árbol de Decisiones es un modelo basado en la creación de reglas lógicas [18], este modelo dividió los datos en subconjuntos más pequeños, lo que le permitió manejar las relaciones complejas y no lineales. Los resultados obtenidos para la validación de este modelo fueron extremadamente altos, los detalles de su desempeño se observan en la siguiente matriz de confusión en la fig.5 y su curva de aprendizaje del modelo se puede visualizar en la fig.6 en el cual se aprecia cómo evolucionaron las métricas de precisión tanto en los datos de entrenamiento como en los datos de validación.

Matriz de Confusión - Árbol de Decisiones

	0	1	2	3	4	5	6
0	48	0	1	0	0	0	0
1	0	40	0	0	0	0	0
2	0	0	21	0	0	0	0
3	0	0	0	132	0	0	0
4	0	0	0	0	92	0	0
5	0	0	0	0	0	86	0
6	0	0	0	0	0	0	131
Verdadero	0	1	2	3	4	5	6
Predicción	0	1	2	3	4	5	6

Fig. 5. Matriz de confusión de modelo de predicción de Árbol de Decisiones.

- Exactitud: Con un resultado porcentual de 99.82%, el Árbol de Decisiones logró clasificar correctamente los casos en el conjunto de prueba, lo que significa que casi todos los estudiantes fueron asignados a su tipo vocacional correcto. Este alto nivel de exactitud indica que el modelo pudo identificar patrones muy claros y consistentes entre las respuestas y los tipos vocacionales.
- Precisión: Obteniendo un resultado porcentual de 99.83%, esta métrica midió qué tan efectivo es el modelo al predecir cada tipo vocacional sin producir demasiados falsos positivos (errores de clasificación en los que el modelo asigna incorrectamente una clase).
- Recall: Con un resultado porcentual de 99.82%, esta métrica midió la capacidad del modelo para identificar correctamente a los estudiantes de cada tipo vocacional, sin dejar de clasificar correctamente a los que realmente pertenecen a cada categoría.

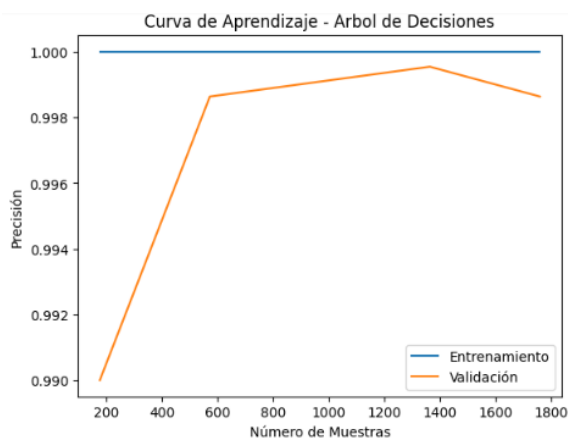


Fig. 6. Curva de aprendizaje del modelo de predicción Árbol de decisiones.

3.1.2 Resultados de la Regresión Logística

La Regresión Logística es un modelo lineal que intentó encontrar la mejor línea divisoria entre las clases en función de las características de entrada [19]. A pesar de ser un modelo sencillo y eficiente, los resultados de validación no alcanzaron el rendimiento del Árbol de Decisiones, su desempeño se observa en la siguiente matriz de confusión en la fig.7.

Matriz de Confusión - Regresión Logística

	0	1	2	3	4	5	6
0	21	1	0	18	9	0	0
1	0	6	1	0	0	19	14
2	0	0	0	15	5	1	0
3	14	2	0	84	24	5	3
4	7	0	1	50	32	0	2
5	0	7	0	2	1	67	9
6	0	10	5	5	3	5	103
Predicción	0	1	2	3	4	5	6

Fig. 7. Matriz de confusión de modelo de predicción de Regresión Logística.

- Exactitud: Se obtuvo un resultado porcentual de 56.81%, el modelo de Regresión Logística tuvo un desempeño significativamente inferior en comparación con el Árbol de Decisiones. Este resultado sugiere que las relaciones entre las respuestas del test vocacional y los tipos vocacionales no son completamente lineales, lo que dificulta a la Regresión Logística capturar esas relaciones.
- Precisión: Con un resultado porcentual de 54.38% muestra que el modelo produjo un número considerable de falsos positivos, lo que indica que a menudo clasificó erróneamente a los estudiantes en categorías incorrectas. Este valor relativamente bajo también refleja la limitación del modelo para ajustar los datos, ya que no puede manejar de manera efectiva patrones no lineales.
- Recall: Se obtuvo un resultado porcentual de 56.81%, el

cual nos indica que el modelo tuvo problemas para identificar correctamente todas las instancias de cada clase vocacional. Esto se traduce en que muchas categorías no fueron completamente recuperadas, lo que significa que el modelo no fue capaz de predecir correctamente una parte significativa de los estudiantes para algunos tipos vocacionales.

3.1.3 Resultados de las Redes Neuronales

Las Redes Neuronales son modelos altamente flexibles [20], su función fue capturar relaciones no lineales y complejas entre los datos. En este caso, el rendimiento en el resultado de validación fue intermedio, ubicándose entre la Regresión Logística y el Árbol de Decisiones, su matriz de confusión se puede ver en la fig.8.

Matriz de Confusión - Red Neuronal

	0	1	2	3	4	5	6
0	13	0	0	26	8	1	1
1	0	19	0	4	0	5	12
2	0	0	9	6	3	3	0
3	10	1	0	82	29	7	3
4	4	0	2	32	54	0	0
5	0	4	3	13	0	66	0
6	0	6	0	1	0	1	123
Predicción	0	1	2	3	4	5	6

Fig. 8. Matriz de confusión de modelo de predicción de Red Neuronal.

- Exactitud: Con un resultado porcentual de 66.42%, mejorando notablemente el desempeño de la Regresión Logística, pero aún lejos del Árbol de Decisiones. Esto sugiere que la Red Neuronal fue capaz de identificar algunos patrones complejos en los datos, pero no de manera óptima.
- Precisión: Obteniendo un resultado porcentual de 66.35%, la Red Neuronal mostró una mayor capacidad para reducir los falsos positivos en comparación con la Regresión Logística. Sin embargo, aún hubo errores en la clasificación, lo que implica que, aunque capturó más detalles que la Regresión Logística, sigue siendo menos eficiente que el Árbol de Decisiones.
- Recall: Con un resultado porcentual de 66.42%, indica que la Red Neuronal fue capaz de recuperar un mayor número de instancias correctas en comparación con la Regresión Logística, pero tampoco logró alcanzar la perfección del Árbol de Decisiones.

3.1.4 Comparación de los Modelos

En la siguiente Tabla 1 se presentan los resultados de validación comparativos de los tres modelos en cuanto a exactitud, precisión y recall:

TABLA 1

Resultados comparativos de la validación en exactitud, precisión y recall para los tres modelos predictivos.

Modelo	Exactitud	Precisión	Recall
Árbol de Decisiones	0.99	0.99	0.99
Regresión Logística	0.56	0.54	0.56
Redes Neuronales	0.66	0.66	0.66

El Árbol de Decisiones mostró un rendimiento casi perfecto, lo que sugiere que las relaciones entre las respuestas del test vocacional y los tipos vocacionales son altamente discernibles por un modelo que utiliza reglas de decisión. Este resultado indica que el Árbol de Decisiones es una herramienta extremadamente efectiva para este tipo de clasificación y, dado su rendimiento, podría ser directamente implementado en procesos de orientación vocacional con un alto nivel de confianza.

El desempeño de la Regresión Logística revela que este tipo de modelo no es el más adecuado para este problema, debido a la naturaleza compleja y no lineal de los datos. Aunque es un algoritmo fácil de implementar y eficiente computacionalmente, no fue capaz de capturar la riqueza de las relaciones entre las variables predictoras y la clasificación vocacional.

Las Redes Neuronales demostraron un rendimiento superior a la Regresión Logística debido a su capacidad para capturar relaciones no lineales. Sin embargo, su desempeño aún fue considerablemente inferior al del Árbol de Decisiones.

En el caso del Árbol de Decisión, su alto rendimiento tanto en entrenamiento como en prueba cercano al 99% sugiere que el sobreajuste no fue significativo, puesto que se dividió en conjunto de prueba, de entrenamiento y testeo. Sin embargo, modelos más complejos, como las Redes Neuronales, podrían ser más susceptibles, lo cual explica su menor rendimiento (66.42%).

4 CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

El análisis de la precisión mostró que el Árbol de decisiones es el modelo más preciso, con un valor porcentual en su validación de 99.83%. Esto implica que el modelo comete muy pocos errores al predecir el tipo vocacional de los estudiantes, reduciendo los falsos positivos y proporcionando predicciones confiables.

En términos de exactitud, que mide el porcentaje total de predicciones correctas, el Árbol de decisiones también superó a los demás modelos, con un valor porcentual en la validación de 99.82%; las Redes neuronales alcanzaron una exactitud de 66.42%, mientras que la Regresión logística obtuvo 56.81%.

En relación con el recall, el Árbol de Decisiones demostró nuevamente su superioridad con un valor porcentual en su

validación de 99.82%, midiendo correctamente a los estudiantes que pertenecen a cada tipo vocacional.

4.2 Recomendaciones

Dado que el modelo de Árbol de Decisiones ha demostrado un rendimiento casi perfecto en términos de exactitud, precisión y recall, se recomienda implementar este modelo en un sistema web de orientación vocacional de la GRTPE Cusco. Este modelo puede ser utilizado para clasificar de manera eficiente a todos los estudiantes en sus respectivos tipos vocacionales, proporcionando una herramienta, con resultados al instante y evitando la calificación manual por parte del personal de la institución, así también resultados confiables y precisos mejorando la toma de decisiones en el proceso de orientación.

Aunque el Árbol de Decisiones ha mostrado un rendimiento excelente en este estudio, se recomienda realizar un monitoreo continuo de su desempeño, puesto que los datos y las necesidades de los estudiantes pueden evolucionar, por lo que es importante que el modelo sea evaluado periódicamente para garantizar que continúe proporcionando resultados precisos y útiles.

Este estudio se centró en tres modelos predictivos, sería recomendable evaluar otros algoritmos de machine learning, como Support Vector Machines (SVM) o Random Forest, para verificar si otros enfoques podrían ofrecer un rendimiento aún mejor o más robusto en distintas condiciones.

AGRADECIMIENTOS

Agradezco a la Gerencia Regional de Trabajo y Promoción del Empleo de Cusco (GRTPE Cusco) por su valioso apoyo en la recolección de datos a través de la aplicación de test vocacionales a los estudiantes de secundaria de la región Cusco, su colaboración fue fundamental para la conformación de la base de datos utilizada en el presente estudio, la cual permitió llevar a cabo el entrenamiento y evaluación de los tres modelos predictivos de Machine learning.

REFERENCIAS

- [1] T. De León Mendoza y R. Rodríguez Martínez, «El efecto de la orientación vocacional en la elección de carrera,» *Revista Mexicana de Orientación Educativa*, vol. V, n° 13, p. 7, 2008.
- [2] X. F. Erazo Guerra y E. d. R. Rosero Morales, «Orientación vocacional y su influencia en la deserción universitaria,» *Horizontes Revista de Investigación en Ciencias de la Educación*, vol. V, n° 18, p. 16, 2021. <https://doi.org/10.33996/revistahorizontes.v5i18.198>
- [3] OCDE, «Education at a Glance 2019 OECD Indicators,» Paris, 2019.
- [4] Ministerio de Trabajo y Promoción del Empleo Cusco, «Ministerio de Trabajo y Promoción del Empleo Cusco,» 21 Noviembre 2016. [En línea]. Available:

<https://www2.trabajo.gob.pe/el-ministerio-2/sector-empleo/dir-gen-form-cap-lab/sobre-el-sovio/>. [Último acceso: 10 Octubre 2024].

- [5] C. OVALLE PAULINO, «Modelo predictivo basado en Machine Learning para la Cadena de Suministro y su influencia en la gestión logística de una empresa de venta de autos,» *Revista de la ACM*, p. 15, Abril 2022.
- [6] E. Franco Delgado y M. Polanco Valenzuela, «Elección de la carrera profesional: modelo predictivo en estudiantes de una universidad privada de Arequipa (Perú),» *Revista de Investigación en Psicología*, vol. XXVI, n° 2, p. 27, 2023.
<https://doi.org/10.15381/rinvp.v26i2.25325>
- [7] N. Luna Reyes, «Implementación de un sistema de predicción del rendimiento académico de los estudiantes de Ingeniería de Sistemas de la Universidad Jose Maria Arguedas utilizando tecnicas de mineria de datos para la adecuada toma de decisiones.,» UNAJMA, Andahuaylas, 2020.
- [8] A. E. Aco Tito, B. O. Hanco Condori y Y. Pérez Vera, «Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria,» *Revista Ibérica de Sistemas y Tecnologías de Información*, Arequipa, 2023.
- [9] J. D. Garcia Dionisio, «Machine learning para predecir el rendimiento académico de los estudiantes universitarios,» Lima, 2021.
- [10] D. I. Candia Oviedo, «Predicción del rendimiento académico de los estudiantes de la UNSAAC a partir de sus datos de ingreso utilizando algoritmos de aprendizaje automático,» Cusco, 2019.
- [11] H. Caselli Gismondi, «Modelo predictivo basado en Machine Learning como soporte para el seguimiento académico del estudiante universitario,» Chimbote, 2021.
- [12] D. Quimbayo, «Agente Inteligente para predecir la elección de carrera de estudiantes de bachiller,» Colombia, 2024.
- [13] B. E. Cisneros Bravo, «Algoritmo de Machine Learning y uso de propiedades semánticas para la identificación y sugerencia vocacional,» Estado de Mexico, 2023.
- [14] R. A. Leyva Osorio y K. A. Medina Arango, «Sistema de recomendación para vocacion profesional, aplicado a la carrera de ingeniería de sistemas ofrecida en la Unversidad de Cundimarca, extensión facatativa,» Cundinamarca, 2019.
- [15] R. Hernández Sampieri, C. Fernández Collado y M. d. P. Baptista Lucio, *Metodología de la Investigación*, vol. VI, Mexico: INTERAMERICANA EDITORES, 2014.
- [16] S. Savalia y V. Emamian, «Clasificación de arritmias cardíacas mediante perceptrones multicapa y redes neuronales convolucionales,» *Bioingeniería*, vol. V, n° 2, p. 12, 2018.
- [17] Y. Pérez Vera, A. E. Aco Tito y B. O. Hanco Condori, «Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria,» *Revista Ibérica de Sistemas y Tecnologías de Información*, n° 51, p. 15, 30 Setiembre 2023.
- [18] C. Arana, «MODELOS DE APRENDIZAJE AUTOMÁTICO MEDIANTE ÁRBOLES DE DECISIÓN,» Buenos Aires, 2021.
- [19] M. Domínguez, «Regresión Logística y Técnicas de Aprendizaje. Aplicaciones,» España, 2021.
- [20] C. Ruiz y M. Basualdo, «Redes Neuronales: Conceptos Básicos y Aplicaciones,» Buenos Aires, 2001.

BIOGRAFÍA

Malu Beatriz Silva Zarate, bachiller en Ingeniería Informática y Sistemas de la Universidad Nacional Micaela Bastidas de Apurímac.

Mario Aquino Cruz, Docente en la Universidad Nacional Micaela Bastidas de Apurímac Perú, MSc. en Informática, investigador en las áreas de informática educativa, IoT, inteligencia artificial y ciberseguridad.