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 efectivo para optimizar el procesamiento de los test vocacionales, los cuales tradicionalmente se analizaban y calificaban de forma manual, generando demoras y limitaciones en la orientación estudiantil. Esta situación afectaba la calidad del servicio de orientación vocacional, dificultando la emisión de los resultados a los estudiantes. Utilizando el test vocacional IEPPO, se procesaron las respuestas aplicando la metodología KDD. Los modelos fueron desarrollados y evaluados en Python, utilizando indicadores de rendimiento como exactitud, precisión, recall y F1 score. Los resultados mostraron que el modelo de Árbol de Decisiones fue el más eficaz en la clasificación, alcanzando una exactitud de 86.00%, una precisión del 86.23%, un recall del 86.00% y un F1 score de 86.11%, superando a las Redes Neuronales y la Regresión Logística. Estas conclusiones evidencian el potencial del Árbol de Decisiones para automatizar la orientación vocacional, brindando un apoyo confiable a la Oficina de SOVIO al facilitar el procesamiento inmediato y automatizado de los resultados, permitiendo clasificar los tipos vocacionales según las respuestas de los estudiantes.

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

Abstract— This research paper compared three predictive machine learning models: Decision Tree, Neural Networks, and Logistic Regression, with the aim of identifying which was the most effective in optimizing the processing of vocational tests, which were traditionally analyzed and graded manually, generating delays and limitations in student guidance. This situation affected the quality of the vocational guidance service, making it difficult to issue results to students. Using the IEPPO vocational test, the responses were processed applying the KDD methodology. The models were developed and evaluated in Python, using metrics such as accuracy, precision, recall, and F1 score. The results demonstrated that the Decision Tree model was the most effective in classification, achieving an accuracy of 86.00%, a precision of 86.23%, a recall of 86.00%, and an F1 score of 86.11%, outperforming Neural Networks and Logistic Regression. These findings demonstrate the potential of the Decision Tree to automate career guidance, providing reliable support to the SOVIO Office by facilitating immediate and automated processing of results and allowing for the classification of career types based on student responses.

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

C&T Riqchary

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

Vol. 6 Núm. 2 (2024) - Publicado: 22/03/24 - Indexaciones
Número: doi.org/10.57166/riqchary/V6.n2.2024
Páginas: 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 Apurimac, 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 adquirido una creciente relevancia dentro de los sistemas educativos a nivel global. Este proceso permite a los estudiantes explorar sus intereses, capacidades e intereses personales orientados a tomar decisiones fundamentadas sobre su trayectoria académica y laboralnal [1]. Sin embargo, persiste una notoria carencia de servicios de orientación vocacional que sean eficientes, accesibles y adaptados a las necesidades individuales de los estudiantes, lo cual repercute negativamente en su desarrollo académico y laboral. Diversos estudios resaltan que una orientación inadecuada puede derivar en una mala elección de carrera, desmotivación e incluso abandono escolar [2].

En el contexto latinoamericano, y específicamente en Perú, esta problemática se agrava por la limitada incorporación de herramientas tecnológicas en los programas de orientación vocacional. A pesar de que existen iniciativas destinadas a este fin, muchas de ellas no disponen de recursos técnicos adecuados que permitan optimizar su funcionamiento frente a una demanda estudiantil cada vez mayor [3]. En la región del Cusco, por ejemplo, el Servicio de Orientación Vocacional e Información Ocupacional (SOVIO), a cargo de la Gerencia Regional de Trabajo y Promoción del Empleo (GRTPE), sigue utilizando procesos manuales basados en papel, Microsoft Excel y Word, lo cual genera demoras significativas en el procesamiento, análisis y entrega de resultados [4].

Frente a esta situación, surge la necesidad de aplicar soluciones tecnológicas que permitan automatizar el procesamiento de los resultados de los test vocacionales y mejorar la precisión de las recomendaciones generadas. En este contexto, el Machine Learning (ML), como subdisciplina de la inteligencia artificial, ofrece un enfoque prometedor al permitir el análisis eficiente de grandes volúmenes de datos y la generación de recomendaciones personalizadas basadas en patrones identificados [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, aplicados al análisis automatizado de datos del Inventario de Estilos Personales y Preferencias Ocupacionales (IEPPO). La comparación se realizará en función de métricas clave como la exactitud, la precisión, el recall y F1 Score, con el propósito de determinar cuál de estos modelos ofrece mejores resultados para apoyar el proceso de orientación vocacional en la GRTPE Cusco. La incorporación de estas tecnologías permitirá optimizar el procesamiento de datos y reducir tiempos de respuesta.

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, valido 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 evaluaron diversos modelos de machine learning para determinar su efectividad en la predicción de la deserción en el ámbito universitario, 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 ha sido desarrollado por un modelo de predicción para los resultados académicos de los estudiantes universitarios, utilizando métodos KDD y evaluando la 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

a) 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).
b) Equipos de cómputo: Computadoras con sistema operativo Windows 64 bits, procesador Core i5 o superior, 8GB de RAM, y 512GB de almacenamiento.
c) Software:
- Lenguaje y librerías para análisis y entrenamiento:
  Python, empleando librerías como pandas y numpy para tratar los datos, matplotlib y seaborn para generar gráficos, y scikit-learn para construir y evaluar los modelos.
- Entorno de desarrollo y pruebas:
  Google Colab, para la programación en la nube, entrenamiento de modelos y visualización de resultados.
  Microsoft Excel, para el procesamiento de datos.
- Frameworks y herramientas para el desarrollo web:
  Frontend: React, con herramientas complementarias.
  Backend (servidores y procesamiento):
  NestJS y FastAPI.
  Base de datos: PostgreSQL.

d) Conexión a internet: Para trabajar en la nube (Google Colab), desplegar servidores locales y consumir las APIs entre backend y frontend.
e) Cuestionario IEPPO: Instrumento para obtener dato 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 distingue por centrarse en la resolución de problemas concretos y aplicados, haciendo uso de conocimientos y teorías ya establecidas para generar resultados con aplicaciones directas en contextos reales, y descriptivo porque medirá y recogera 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 año de secundaria de 38 colegios tanto de instituciones públicas y privadas de la región Cusco, así también la muestra incluye el total de la población.

2.5 Proceso KDD

En la fig.1 se muestran las etapas del Proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD) que se utilizó para la investigación.

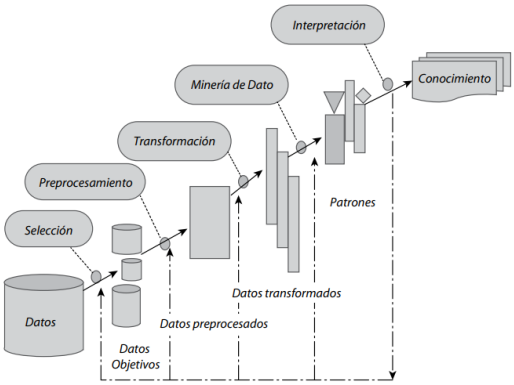

Fig. 1. Etapas del proceso KDD

2.6 Procedimiento

Primero se realizó la recopilación de datos de 3,770 estudiantes de 4° y 5° año de secundaria de instituciones públicas y privadas de región Cusco mediante la aplicación del test vocacional IEPPO, realizado entre agosto y diciembre de 2023, de forma presencial y virtual. La recolección fue realizada por la Gerencia Regional de Trabajo y Promoción del Empleo Cusco, a través de la Oficina SOVIO, conforme a la Directiva SOVIO N.° 001-2012-MTPE y la Resolución Ministerial N.° 116-2009-TR, que respaldan este servicio público gratuito dirigido a instituciones que no cuentan con profesionales en psicología. Los datos, de naturaleza sensible y privada, fueron almacenados en archivos Excel por colegio para su posterior clasificación en los tipos vocacionales.

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:

a) No se parece a mí.
b) Se parece a mí.
c) Me interesa.
d) Me interesa poco o nada.
e) Soy hábil.
f) No soy hábil.

Los datos fueron transformados y preparados para su análisis, lo que incluyó la consolidación y estructuración de las variables que fueron utilizadas.

Tercero, se realizó el procesamiento de las respuestas del test vocacional, obteniendo su clasificación según los 7 tipos vocacionales que son: liderazgo, técnico mecánico, social, organizado, artístico, emprendimiento e investigación. Se

realizó el proceso de 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 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.

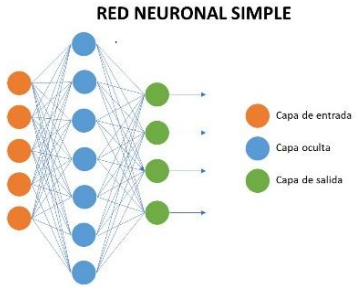

Fig. 2. Arquitectura utilizada para la Red Neuronal (Multilayer Perceptron, MLP) [16]

a) Capa de Entrada (Input Layer): El número de neuronas en la capa de entrada se definió según la cantidad de características en el conjunto de datos, las cuales fueron 118, representando cada una de las preguntas respondidas por los estudiantes.
b) Capa Oculta (Hidden Layer): Se configuró una capa oculta con 5 neuronas mediante el parámetro hidden_layer_sizes, buscando un modelo más simple que reduzca el riesgo de sobreajuste y optimice el tiempo de entrenamiento. Cada neurona aplica una transformación no lineal a los datos de entrada, permitiendo que la red capte relaciones complejas entre las variables.
c) Capa de Salida (Output Layer): Esta capa contiene 7 neuronas, correspondiente a los 7 tipos vocacionales identificados. Cada neurona representa una clase, y el modelo predice la clase con la mayor probabilidad para cada estudiante.
d) Se utilizó max_iter con valor de 1000 para asegurar que el algoritmo tenga suficientes iteraciones para converger, Alpha con valor de 1.0 como término de regularización para evitar el sobreajuste, y random_state=42 para garantizar la reproducibilidad del entrenamiento. Estos valores fueron seleccionados para mantener un balance entre la precisión del modelo y el uso eficiente de los recursos computacionales.

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

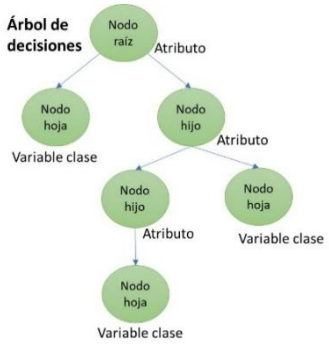

Fig. 3. Arquitectura utilizada para Árbol de decisiones

a) Cada nodo interno representa una decisión basada en una característica de los datos (un nodo a una pregunta).
b) Los nodos hoja representan las decisiones finales, que en este caso son los tipos vocacionales.
c) Cada rama del árbol indica el resultado de una decisión, lo que lleva a otra pregunta.
d) 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.

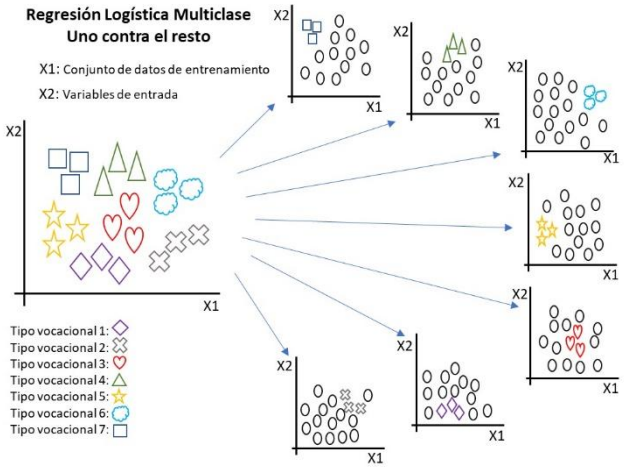

Fig. 4. Arquitectura utilizada para Regresión logística.

a) Capa de entrada: Cada columna del test vocacional se convirtió en una característica para el modelo, los datos categóricos fueron convertidos en numéricos.
b) Procesamiento – Uno contra el resto: Para cada tipo 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.
c) 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 indicadores de rendimiento para evaluar la eficacia de nuestros modelos predictivos, incluyendo la exactitud (accuracy), la precisión (precision), el recall (sensibilidad o recuperación) y F1 Score [17]. Las ecuaciones empleadas para calcular cada métrica son las siguientes:

- Exactitud (Accuracy): Esta métrica mide la proporción total de aciertos del modelo, incluye tanto las predicciones positivas correctas como las negativas correctas 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): Esta métrica refleja la habilidad del modelo para reconocer la totalidad de los casos positivos.

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

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

- F1 Score: Mide el balance entre la precisión y el recall, esta métrica utilizamos por contener datos desbalanceados.

$$F1\ Score = 2 * \frac{Precisión * Recall}{Precisión + Recall}$$

Estas métricas fueron fundamentales para la evaluación del rendimiento de los modelos predictivos como: Árbol de Decisión, Redes Neuronales y Regresión Logística, ya que cada uno de estos algoritmos tuvo un comportamiento diferente en términos de exactitud, precisión, recall y F1 score. 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 las respuestas de los estudiantes de secundaria en los 7 tipos vocacionales. Para ello, se utilizó una base de datos de 2751 evaluaciones, se fue procesando 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 muestran los resultados obtenidos para cada uno de ellos.

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 divide 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 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 el conjunto de entrenamiento como en de validación.

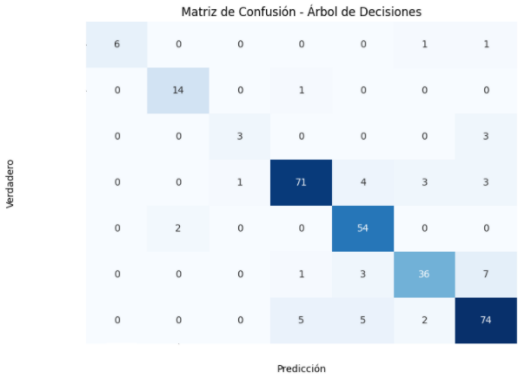

Fig. 5. Matriz de confusión de modelo de predicción de Árbol de Decisiones.

a) Exactitud: Con un resultado porcentual de 86.00%, el Árbol de Decisiones logró clasificar correctamente los casos en el conjunto de prueba, lo que significa que la mayoría de los estudiantes fueron clasificados a su tipo vocacional correcto. Este nivel de exactitud indica que el modelo pudo identificar patrones muy claros y consistentes entre las respuestas y los tipos vocacionales.

b) Precisión: Obteniendo un resultado porcentual de 86.23%, 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).

c) Recall: Con un resultado porcentual de 86.00%, esta métrica midió la capacidad del modelo para identificar correctamente a los estudiantes en su tipo vocacional, sin dejar de clasificar correctamente a los que realmente pertenecen a cada categoría.

C&T Riqchary
Revista de investigación en ciencia y tecnología
Vol. 6 Núm. 2 (2024) - publicado:22/09/2024
DOI https://doi.org/10.57166/riqchary.v6.n2.2024.128

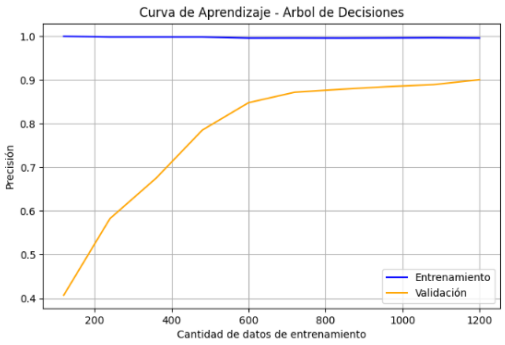

Fig. 6. Curva de aprendizaje del modelo de predicción Árbol de decisiones.

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.

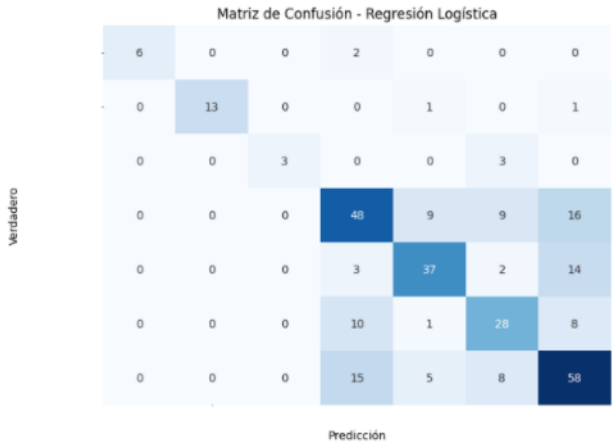

Fig. 7. Matriz de confusión de modelo de predicción de Regresión Logística.

a) Exactitud: Se obtuvo un resultado porcentual de 64.33%, 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.
b) Precisión: Con un resultado porcentual de 65.43% muestra que el modelo produjo un número de falsos positivos, lo que indica que a menudo clasificó erróneamente a los estudiantes en sus tipos vocacionales incorrectos. Este valor también refleja la limitación del modelo para ajustar los datos, ya que no puede manejar de manera efectiva patrones no lineales.
c) Recall: Se obtuvo un resultado porcentual de 64.33%, el cual nos indica que el modelo tuvo problemas para identificar correctamente todas las instancias de cada tipo 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 a 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 cercano al modelo de Regresión Logística, su matriz de confusión se puede ver en la fig.8.

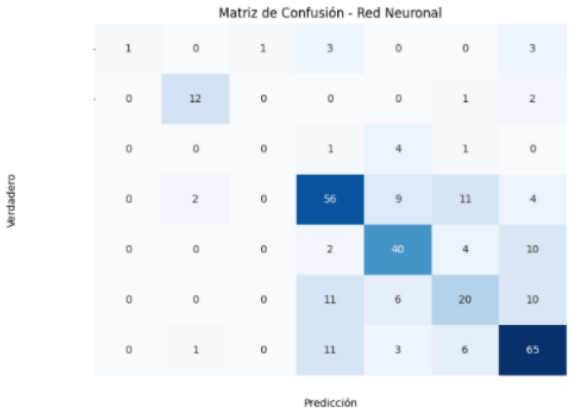

Fig. 8. Matriz de confusión de modelo de predicción de Red Neuronal.

a) Exactitud: Con un resultado porcentual de 64.67%, mejorando en un pequeño porcentaje a la Regresión Logística, pero aún lejos del Árbol de Decisiones. Esto sugiere que la Red Neuronal fue capaz de identificar algunas interacciones complejas en los datos, pero no de manera óptima.
b) Precisión: Obteniendo un resultado porcentual de 64.04%, la Red Neuronal mostró una menor capacidad para reducir los falsos positivos en comparación con la Regresión Logística. Hubo errores en la clasificación, lo que implica que, sigue siendo menos eficiente que el Árbol de Decisiones.
c) Recall: Con un resultado porcentual de 64.67%, indica que la Red Neuronal fue capaz de recuperar un 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, recall y F1 Score:

TABLA 1
Resultados comparativos de la validación en exactitud, precisión, recall, F1 Score para los tres modelos predictivos.

Modelo	Exactitud	Precisión	Recall	F1 Score
Árbol de Decisones	0.86	0.86	0.86	0.861
Regresión Logistica	0.64	0.65	0.64	0.649
Redes Neuronales	0.64	0.64	0.64	0.644

El Árbol de Decisiones mostró un rendimiento alto, 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 efectiva para este tipo de clasificación y, dado su rendimiento, podría ser directamente implementado en procesos de orientación vocacional con un buen 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 del tipo vocacional.

Las Redes Neuronales demostraron un rendimiento cercano 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 86% sugiere que el sobreajuste no fue significativo, puesto que se dividió en conjunto de prueba, de entrenamiento y testeo.

Se realizó una página web donde se integró el modelo de árbol de decisiones entrenado, permitiendo a los estudiantes ingresar sus respuestas como en la fig.9 y obtener su tipo vocacional de forma inmediata como se muestra en la fig.10. Esta solución tecnológica representa una mejora significativa frente al procesamiento tradicional de datos, que solía ser manual, lento y propenso a errores. Gracias a la automatización, se agiliza el proceso de orientación vocacional, se reduce el tiempo de análisis y se garantiza mayor precisión en los resultados, beneficiando así a la comunidad estudiantil con una herramienta moderna, eficiente y fácil de usar.

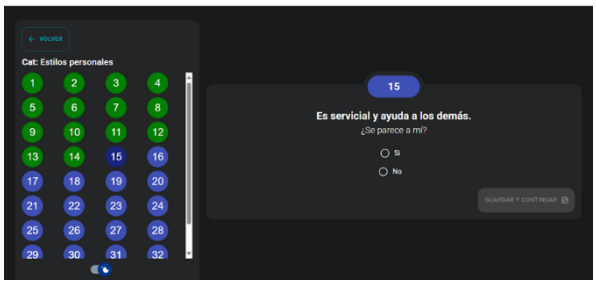

Fig. 9. Ingreso de respuestas del test IEPPO en la página web.

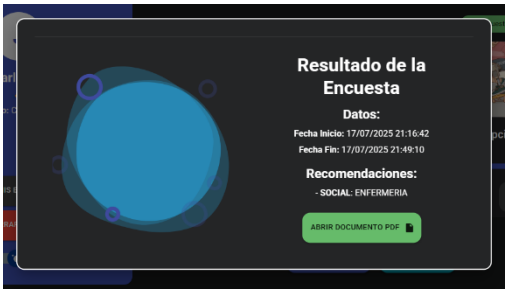

Fig. 10. Resultado de encuesta, clasificando su tipo vocacional.

4 CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones

Como resultado de esta investigación, se concluyó que el modelo predictivo de Árbol de Decisiones fue el más efectivo para mejorar el servicio de orientación vocacional dirigido a los estudiantes de secundaria en la GRTPE Cusco, superando a las Redes Neuronales y a la Regresión Logística, lo que demuestra su capacidad para clasificar con mayor precisión los tipos vocacionales a partir de las respuestas del test IEPPO, optimizando así el procesamiento y análisis de los datos recolectados.

El análisis de la precisión mostró que el Árbol de decisiones es el modelo más preciso, con un valor porcentual es su validación de 86.23%. 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 86.00%; las Redes neuronales alcanzaron una exactitud de 64.67%, mientras que la Regresión logística obtuvo 64.33%.

En relación con el recall, el Árbol de Decisiones demostró nuevamente su superioridad con un valor porcentual en su validación de 86.00%, midiendo correctamente a los estudiantes que pertenecen a cada tipo vocacional.

4.2 Recomendaciones

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.

Si bien el estudio se centró en tres modelos de predicción, resulta recomendable analizar otros enfoques de machine learning, como SVM y Random Forest, para determinar si ofrecen mayor precisión o estabilidad frente a diversas 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, nº 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, nº 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/dirgen-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, nº 2, p. 27, 2023. https://doi.org/10.15381/rinvp.v26i2.25325

[7] N. Luna Reyes, «Implementación de un sistema de predicción del rendimiento acaddémico de los estudiantes de Ingenieria 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. Hancco 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, nº 2, p. 12, 2018.

[17] Y. Pérez Vera, A. E. Aco Tito y B. O. Hancco Condori, «Análisis comparativo de Técnicas de Machine Learning para la predicción de casos de deserción universitaria,» Revista Ibérica de Sistemas y Tecnologias de Informacion, nº 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.