Machine Learning para la Predicción Clínica y el Análisis Educativo: Casos de Insuficiencia Cardíaca y Coursera

Machine Learning para la predicción Clínica y el Análisis Educativo: Casos de Insuficiencia Cardíaca y Coursera

Alejandro Garcia Jorge Luis, García Mendoza Isaías, Trinidad Medina Kevin Yahir

agjo220210@upemor.edu.mx, gmio220060@upemor.edu.mx, tmko220776@upemor.edu.mx

RESUMEN

En esta investigación se aplicaron técnicas de aprendizaje automático para analizar dos conjuntos de datos distintos.

Primero, se utilizó el algoritmo de agrupamiento : k-means en un conjunto de datos de cursos de Coursera. El análisis reveló que la mayoría de los cursos poseen calificaciones muy altas, concentradas entre 4.5 y 4.9. Se descubrió que los cursos de nivel principiante son los más populares en términos de inscripciones y que no existe una correlación fuerte entre la calificación de un curso y su número de estudiantes.

Segundo, se compararon los algoritmos : Naive Bayes y k-NN para predecir la mortalidad a partir de un conjunto de datos sobre insuficiencia cardíaca. Los resultados mostraron una clara superioridad del modelo k-NN (con k=3) , el cual alcanzó una exactitud del 91.87%, en contraste con el 65.50% de Naive Bayes. Aunque Naive Bayes tuvo una alta sensibilidad para detectar fallecimientos, su precisión fue muy baja (37.95%), haciéndolo poco fiable. El modelo k-NN demostró un mejor equilibrio entre métricas, consolidándose como una herramienta más robusta y confiable para esta tarea de clasificación.

Palabras claves: *K-Means, K-NN, Naive Bayes, Aprendizaje Automático, Clasificación, Agrupamiento, Análisis Predictivo, Insuficiencia Cardíaca.*

ABSTRACT

In this research, machine learning techniques were applied to analyze two different datasets.

First, the k-means clustering algorithm was used on a dataset of Coursera courses. The analysis revealed that most courses have very high ratings, concentrated between 4.5 and 4.9. It was found that beginner-level courses are the most popular in terms of enrollment and that there is no strong correlation between a course's rating and its number of students.

Second, the Naive Bayes and k-NN algorithms were compared to predict mortality from a dataset on heart failure. The results showed a clear superiority of the k-NN model (with k=3), which achieved an accuracy of 91.87%, in contrast to 65.50% for Naive Bayes. Although Naive Bayes had high sensitivity for detecting deaths, its accuracy was very low (37.95%), making it unreliable. The k-NN model demonstrated a better balance between metrics, establishing itself as a more robust and reliable tool for this classification task.

Keywords: K-Means, K-NN, Naive Bayes, Machine Learning, Classification, Clustering, Predictive Analysis, Heart Failure.

1. INTRODUCCIÓN

En el documento se presenta un análisis de datos apoyado en la aplicación de tres algoritmos de aprendizaje automático: k-means, Naive Bayes y k-NN. Para ello, se emplean dos conjuntos de datos con objetivos distintos. El primero, Coursera Course Dataset, se utiliza con el algoritmo de agrupamiento no supervisado k-means, con la finalidad de identificar grupos naturales a partir de variables como la calificación promedio, etc. El segundo, Heart Failure Prediction Dataset, contiene registros médicos de pacientes con enfermedades cardiovasculares y se emplea en un problema de clasificación, en el cual se aplican y comparan los algoritmos supervisados Naive Bayes y k-NN para predecir la mortalidad de los pacientes.

2. TRABAJOS RELACIONADOS

Se colocan las referencias relacionadas a la clasificación de datos utilizando k-nn, naive bayes y k-means.

Insuficiencia cardiaca:

1. "Implementation of Machine Learning Model to Predict Heart Failure Disease". Este artículo aborda sobre la dificultad de predecir la insuficiencia cardíaca a tiempo, y como representa un reto para cardiólogos y cirujanos. Se comprobó que el uso de modelos de aprendizaje automático mejora la comprensión de los datos médicos y la predicción de riesgos. El estudio logró aumentar la precisión en la predicción de insuficiencia cardíaca en comparación con trabajos previos [1].
2. "A Machine Learning Approach to Management of Heart Failure Populations". El estudio habla sobre cómo la insuficiencia cardíaca es una enfermedad frecuente y costosa, lo que exige nuevas estrategias para gestionar eficazmente poblaciones de pacientes. Los modelos de machine learning entrenados con datos clínicos masivos lograron predecir el riesgo de mortalidad a 1 año con alta precisión. La priorización basada en la reducción del riesgo de mortalidad superó otros métodos, demostrando la posibilidad de optimizar la gestión poblacional de pacientes [2].

Coursera

3. "Best Combination of Machine Learning Algorithms for Course Recommendation System in E-learning". Este estudio aborda la necesidad de mejorar los sistemas de recomendación de cursos en entornos de aprendizaje en línea, utilizando datos de inscripción para identificar el comportamiento de los estudiantes. Se exploraron cinco métodos que combinan algoritmos de minería de datos para determinar cuál ofrece mejores recomendaciones. Concluyendo que el enfoque combinado superó al método basado únicamente en reglas de asociación. [3].

3. DESCRIPCIÓN DE LA TÉCNICA

A continuación se describen los algoritmos k-means, K-NN y Naive Bayes.

Conjunto de datos del curso Coursera

La agrupación de medias K es un algoritmo de aprendizaje no supervisado utilizado para la agrupación en clústeres de datos, que agrupa puntos de datos no etiquetados en grupos o clústeres [4]. Su funcionamiento es iterativo: primero se eligen los centroides de forma aleatoria; después, cada dato se asigna al clúster más próximo, y posteriormente los centroides se recalculan como la media de los puntos de su grupo. Este proceso de asignación y actualización se repite hasta que las agrupaciones se estabilizan y ya no se producen cambios relevantes.

Para este análisis se utilizó el conjunto de datos Coursera Course Dataset [5], que contiene información de alrededor de 900 cursos de la plataforma Coursera. Entre sus variables, se consideraron especialmente la calificación promedio de cada curso (course_rating) y el número de estudiantes inscritos (course_students_enrolled). Con base en estos datos, el objetivo fue identificar posibles segmentos naturales de cursos mediante técnicas de agrupamiento.

El proceso para segmentar los cursos del dataset utilizando k-Means fue el siguiente:

1. Preparación y Limpieza de Datos: El primer paso fue transformar la columna course_students_enrolled, que estaba en formato de texto, a un formato numérico. Esto se logró eliminando la letra "k" y multiplicando los valores por 1,000 para obtener el número real de inscritos.
2. Selección de Características: Se seleccionaron únicamente las dos columnas numéricas relevantes para el clustering: la course_rating y la nueva columna numérica de estudiantes inscritos (course_students_enrolled).
3. Normalización: Debido a que la escala de las dos variables era muy diferente (calificaciones de 1-5 vs. inscritos en miles), se aplicó una normalización (Z-transform) para que ambas tuvieran el mismo peso en el cálculo de las distancias y así evitar que el número de inscritos dominara el análisis.
4. Determinación del Número Óptimo de Clusters (k): Para tomar una decisión informada sobre el valor de k, se utilizó la Técnica del Codo. Se ejecutó el algoritmo k-Means para un rango de valores de k (de 2 a 17) para cada uno, siendo k=5 es más óptimo. El "codo" en el gráfico, que representa el punto donde añadir más clusters deja de aportar una mejora significativa, sugirió el número óptimo de segmentos a utilizar. Véase Tabla 1.

Tabla 1 valores de K.

Valor de K	Avg. within centroid distance
2	-1.340
3	-0.792
5	-0.476
8	-0.284
17	-0.114

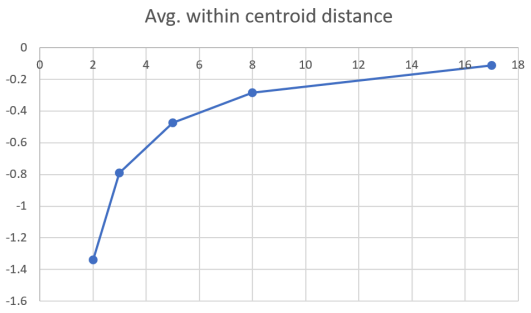


Imagen 1. Técnica del codo.

5. Ejecución y Análisis de los Clusters: Finalmente, se ejecutó el algoritmo k-Means con el valor de k óptimo. El resultado fue la segmentación de los cursos en distintos grupos. El análisis de los centroides de cada clúster permitió interpretar el significado de cada segmento.

Conjunto de datos de predicción de insuficiencia cardíaca (Bayes y k-NN)

Bayes: El Teorema de Bayes es una proposición matemática ampliamente utilizada en estadística y cálculo de probabilidad. Se utilizan los algoritmos bayesianos para calcular la probabilidad de que ocurra B si ha ocurrido A, o de saber si ha ocurrido A cuando sabemos que ha ocurrido B, con amplios conjuntos de datos y probabilidades condicionales [6].

Para este análisis se utilizó el conjunto de datos Heart Failure Prediction Dataset [7] este conjunto de datos contiene registros médicos de pacientes con enfermedades cardiovasculares. El objetivo es clasificar si un paciente sobrevivirá o no a un evento de insuficiencia cardíaca.

Variable Objetivo (Label): Mortality. Esta es la variable que el modelo intentará predecir.

El modelo de clasificación en RapidMiner se desarrolló iniciando con la carga del archivo FIC.Full CSV.csv y la definición de la variable Mortality como etiqueta de predicción. Posteriormente, se aplicó una validación cruzada de 10 particiones para garantizar la fiabilidad de los resultados. Durante la fase de entrenamiento, se utilizó el algoritmo Naive Bayes, que aprendió las relaciones de probabilidad entre las variables médicas y la mortalidad. En la fase de prueba, el modelo generó predicciones sobre los subconjuntos reservados y su desempeño se evaluó mediante métricas como exactitud, precisión y la matriz de confusión, promediadas en las diez iteraciones del proceso.

k-NN: El algoritmo de k-nearest neighbors (KNN) es un clasificador de aprendizaje supervisado no paramétrico, que emplea la proximidad para realizar clasificaciones o predicciones sobre la agrupación de un punto de datos individual [5].

Para este análisis se ocupó el mismo conjunto que para el bayes y la misma variable objetivo. La clasificación de un nuevo dato mediante el algoritmo k-NN se fundamenta en la idea de que un punto suele pertenecer a la misma clase que sus vecinos más cercanos. Para ello, primero se define un valor de k, que indica cuántos vecinos se tendrán en cuenta; luego, se calcula la distancia entre el nuevo punto y todos los datos de entrenamiento, identificando los k más próximos. Finalmente, el nuevo dato se clasifica según la mayoría de votos de esos vecinos. En este análisis, se empleó un valor de k = 3.

La construcción y validación del clasificador k-NN comenzó con la importación del conjunto de datos FIC.Full CSV.csv y la definición de la variable Mortality como etiqueta de predicción. Para garantizar la fiabilidad de los resultados, se aplicó una validación cruzada de 10 particiones. En la fase de entrenamiento se configuró el algoritmo con k = 3, almacenando los datos de referencia. Durante la fase de prueba, cada registro del conjunto de validación fue clasificado en función de sus tres vecinos más cercanos, asignándole la clase mayoritaria. Finalmente, el desempeño del modelo se evaluó mediante métricas como la exactitud y otras medidas de calidad predictiva.

4. RESULTADOS

A continuación se muestra como término visualmente la parte de descripción de la técnica. Véase Imagen 1, Imagen 2 e Imagen 3.

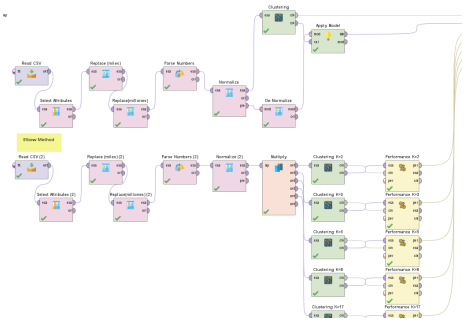


Imagen 2. k-MEANS.

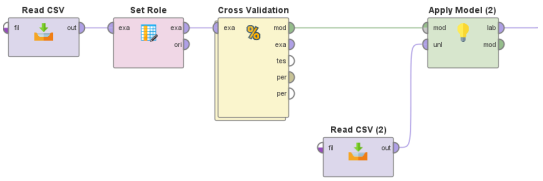


Imagen 3. BAYES

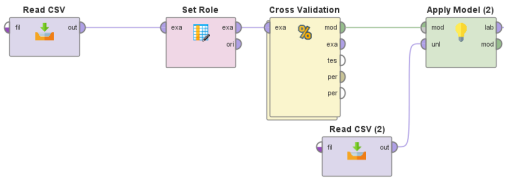


Imagen 4. k-NN.

Para el k-MEANS, se encontraron 5 hallazgos:

Hallazgo 1: La mayoría de los cursos tienen una calificación muy alta. Parece haber una tendencia a que los cursos en la plataforma estén muy bien calificados, lo que sugiere una alta satisfacción general de los usuarios. Véase Imagen 4.

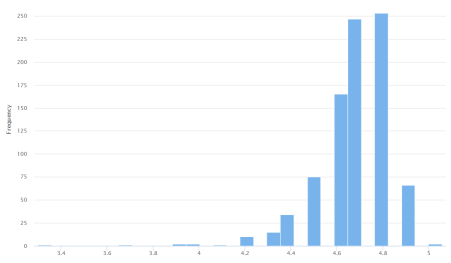


Imagen 5. Hallazgo 1.

En la gráfica (Imagen 4) se muestra una gran concentración de cursos con calificaciones entre 4.5 y 4.9, confirmando que las calificaciones bajas son raras.

Hallazgo 2: La Universidad de Pensilvania y la Universidad de Michigan son los proveedores de contenido más prolíficos. Ciertas organizaciones dominan la plataforma, ofreciendo una cantidad significativamente mayor de cursos que otras. Esto puede indicar alianzas estratégicas o un enfoque en ciertos campos del conocimiento. Véase la imagen 5.

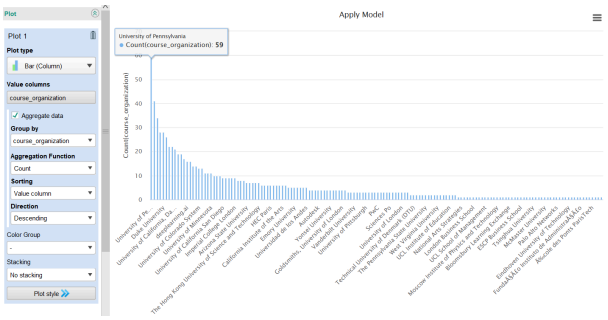


Imagen 6. Hallazgo 2.

Hallazgo 3: Los cursos para principiantes son, por lejos, los más populares en términos de inscripción. La plataforma atrae a una gran cantidad de estudiantes que buscan iniciar en un nuevo tema. Los cursos de nivel "Beginner" tienen un número de inscritos considerablemente mayor que los de nivel intermedio o avanzado. Véase Imagen 6.

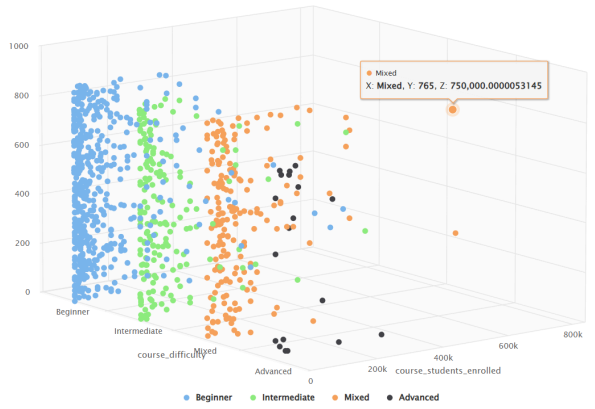


Imagen 7. Hallazgo 3.

Hallazgo 4: No hay una correlación fuerte entre la calificación de un curso y el número de estudiantes inscritos. Podrías esperar que los cursos mejor calificados atraigan a más estudiantes, pero el

análisis muestra que muchos cursos con calificaciones perfectas (4.9 o 5.0) tienen pocos estudiantes, mientras que otros con calificaciones ligeramente menores (4.6-4.7) son masivos. La popularidad parece depender más del tema o del instructor. Véase Imagen 7.

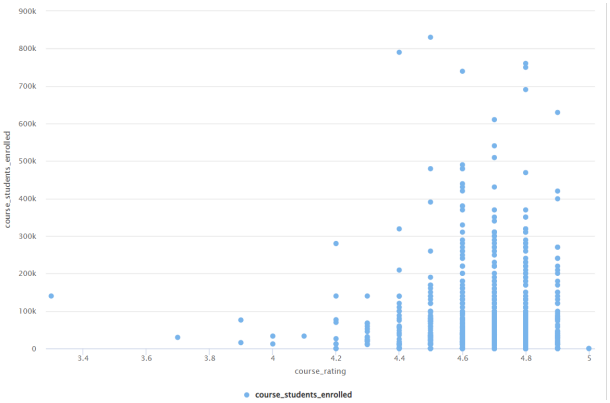


Imagen 8. Hallazgo 4.

Hallazgo 5: Las "Especializaciones" no necesariamente tienen más inscritos que los Cursos individuales. Los cursos individuales más exitosos pueden superar en número de inscritos a especializaciones completas. Esto sugiere que los estudiantes a menudo buscan conocimientos específicos en lugar de un programa completo. Véase imagen 8.

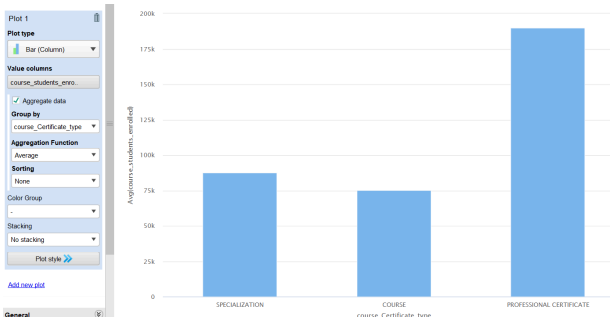


Imagen 9. Hallazgo 5.

Para la clasificación, veremos primero la matriz de Confusión para cada modelo. Véase Tabla 2.

Tabla 2. matriz de confusión.

	BAYES		k-NN	
True	0	1	0	1
0	167	6	269	11
1	121	74	19	69

Podemos ver que 0->No muere y 1->Falleció. Tomando esto en cuenta, vemos que el modelo k-NN tiene mejores predicciones. Bayes tiene un problema muy serio para identificar a los pacientes en riesgo. Aunque es bueno para predecir quién sobrevivirá, falla estrepitosamente en detectar a los que fallecieron

Ahora veamos la tabla comparativa de métricas. Véase Tabla 3.

Tabla 3. Métricas.

Métrica	Algoritmo K-NN (k=3)	Algoritmo Naive Bayes
Accuracy	91.87%	65.50%
Precision	78.41%	37.95%
Recall (Sensibilidad)	86.25%	92.5%

La exactitud (Accuracy) se calculó como la proporción de predicciones correctas respecto al total de casos, usando la fórmula de la Imagen 9.

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

Imagen 10. Accuracy.

La precisión (Precisión) se empleó para conocer qué tan confiable es el modelo al predecir casos positivos (mortalidad). Se obtiene usando la fórmula de la Imagen 10.

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

Imagen 11. Precisión.

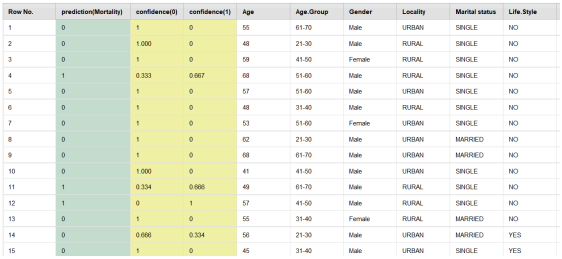
La sensibilidad o recall (Recall) permitió evaluar la capacidad del modelo para identificar correctamente los casos positivos reales. Usando la fórmula de la Imagen 11.

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

Imagen 12. Precisión.

La conclusión es que k-NN supera claramente a Naive Bayes: tiene mucha mayor exactitud (91.9% vs 65.5%), mejor precisión (78.4% vs 37.9%) y buen recall (86.2% vs 92.5%). Naive Bayes detecta más fallecimientos (alto recall) pero se equivoca mucho al predecirlos (baja precisión), mientras que k-NN logra un balance mucho mejor entre precisión y sensibilidad, siendo el modelo más confiable en general.

Podemos observar la salida de KNN. Vease imagen 12.



Row No.	prediction(Mortality)	confidence(0)	confidence(1)	Age	Age.Group	Gender	Locality	Marital status	Life.Style
1	0	1	0	55	61-70	Male	URBAN	SINGLE	NO
2	0	1.000	0	48	21-30	Male	RURAL	SINGLE	NO
3	0	1	0	59	41-50	Female	RURAL	SINGLE	NO
4	1	0.333	0.667	68	51-60	Male	RURAL	SINGLE	NO
5	0	1	0	57	51-60	Male	URBAN	SINGLE	NO
6	0	1	0	48	31-40	Male	RURAL	SINGLE	NO
7	0	1	0	53	51-60	Female	URBAN	SINGLE	NO
8	0	1	0	62	21-30	Male	URBAN	MARRIED	NO
9	0	1	0	68	61-70	Male	URBAN	MARRIED	NO
10	0	1.000	0	41	41-50	Male	URBAN	SINGLE	NO
11	1	0.334	0.666	49	61-70	Male	RURAL	SINGLE	NO
12	1	0	1	57	41-50	Male	RURAL	SINGLE	NO
13	0	1	0	55	31-40	Female	RURAL	MARRIED	NO
14	0	0.666	0.334	56	21-30	Male	URBAN	MARRIED	YES
15	0	1	0	45	31-40	Male	URBAN	SINGLE	YES

Imagen 13. Resultado de Bayes.

Podemos observar la salida de Bayes. Vease imagen 13.

5. CONCLUSIONES Y TRABAJOS FUTUROS

El estudio confirma la utilidad de los algoritmos de aprendizaje automático para obtener información relevante a partir de distintos tipos de datos. En el análisis de cursos en línea mediante k-means, se observó que la popularidad de un curso, medida por la cantidad de inscripciones, no está directamente relacionada con su calificación. Los cursos para principiantes destacan como los más demandados. Por otro lado, en la predicción de insuficiencia cardíaca, se compararon los algoritmos Naive Bayes y k-NN, siendo este último, con k=3, el que obtuvo mejores resultados. Con una exactitud del 91.87%, k-NN se mostró como un clasificador más confiable y equilibrado, logrando identificar con precisión a los pacientes en riesgo.

REFERENCIAS

[1] ALOTAIBI, Fahd Saleh. Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications*, 2019, vol. 10, no 6.

[2] JING, Linyuan, et al. A machine learning approach to management of heart failure populations. *Heart Failure*, 2020, vol. 8, no 7.

[3] AHER, Sunita B.; LOBO, L. M. R. J. Best combination of machine learning algorithms for course recommendation system in e-learning. *International Journal of Computer Applications*, 2012, vol. 41, no 6

[4] E. Kavlakoglu y V. Winland, "¿Qué es la agrupación en clústeres k-means?", Think – IBM.https://www.ibm.com/mx-es/think/topics/k-means-clustering.

[5] S. M., "Coursera Course Dataset," Kaggle, 2019. https://www.kaggle.com/datasets/siddharthm1698/coursera-course-dataset.

[6] Kraz Team, "Qué es el Teorema de Bayes y qué importancia tiene en Machine Learning", Blog de Kraz | Data Solutions, 14-jun-2022. https://blog.kraz.ai/marketing/que-es-el-teorema-de-bayes-y-que-importancia-tiene-en-machine-learning.