Predictor de nivel de anemia en niños mediante métodos de clasificacion

Cristobal Medina-Meza[1], Cristobal Estrada-Salinas[1] and Michel Emiliano Bureau-Romo[1]

[1] *Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias*

Resumen—El aprendizaje computacional consta de métodos analíticos para prueba de hipótesis, potenciados por un software que ejecute los métodos para diferentes casos. Las aplicaciones de estos métodos pueden variar, dependiendo del contexto donde se prueba su funcionalidad. Para esta situación problema, se busca demostrar si dichos métodos pueden ser usados en el campo de la medicina y las implicaciones que ello tendría para el estudio del mismo.

Palabras clave—Anemia, KNN, SVC, RandomForest

I. Introducción a la problemática

a. Aprendizaje supervisado en salud

El aprendizaje supervisado en el ámbito de la salud es una técnica de inteligencia artificial en la que se entrena un modelo utilizando un conjunto de datos. Los datos se refieren a ejemplos donde se conoce la respuesta correcta, como diagnósticos médicos o resultados de tratamientos.

En el aprendizaje supervisado en salud, el modelo analiza patrones y características presentes en los datos de entrenamiento para aprender a realizar predicciones o clasificaciones en nuevos datos no etiquetados. Por ejemplo, se puede entrenar un modelo para diagnosticar enfermedades basándose en datos clínicos, imágenes médicas o resultados de pruebas.

b. Descripción del dataset y su preprocesamiento

La base de datos utilizada para este proyecto es un conjunto que explora algunas características de personas que padecen anemia. Características puntuales que describen al paciente, tales como la edad, u otras más personales como el tipo de residencia. Con ello, se quiere demostrar si alguna de estas características puede ser un factor de riesgo para contraer anemia. La hipótesis es que la anemia se puede contraer por medios que se relacionan con dichas cualidades.
Para comenzar el preprocesamiento de los datos, se eliminaron todos los registros que tuvieran por lo menos un valor vacío con ayuda de la función drop.na(), posteriormente verificamos los tipos de datos de cada variable, y, reemplazamos las variables de tipo "object" por variables numéricas, para 'Age in 5-year groups' la reemplazamos con la media del intervalo, 'Type of place of residence' con 1 para urbano y 0 para rural, 'Highest educational level' con los años que requiere cursar cada nivel, 'Wealth index combined' con 0 para 'Poorest' hasta 4 para 'Richest', 'Anemia level' desde 0 para 'Not anemic' hasta 3 para 'Severe' y reemplazando los valores de las variables restantes con 1 para 'Yes' y 0 para 'No'. Posteriormente, se eliminaron las columnas: 'Current marital status', 'Currently residing with husband/partner', 'When child put to breast','Anemia level.1','Taking iron pills, sprinkles or syrup' ya que se consideraron irrelevantes para la variable a predecir. Gracias a esto, nos quedamos con variables solo de tipo "int64" y "float64".
Posteriormente, aplicamos la función "RandomOverSampler()" para una mejor división de nuestros datos, y los dividimos en 2 conjuntos con ayuda de la función "train_test_split", prueba y entrenamiento. Teniendo así un total de 7392 datos de entrenamiento y 2464 de prueba.

c. Análisis exploratorio con gráficos

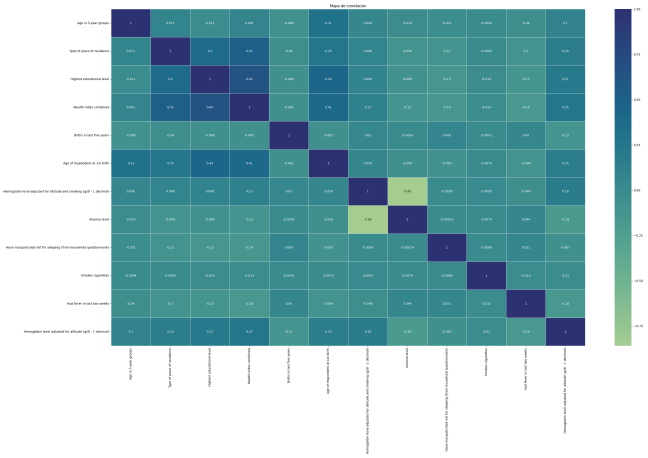

Fig. 1: Mapa de correlación

Analizando el Mapa de correlación nos damos cuenta de que es poco probable que nuestras variables sean linealmente dependientes las unas de las otras, a excepción del 'Hemoglobin level adjusted for altitude and smoking (g/dl - 1 decimal)' con el nivel de anemia, pero no será eliminada debido a que el nivel de anemia es nuestra variable a predecir.

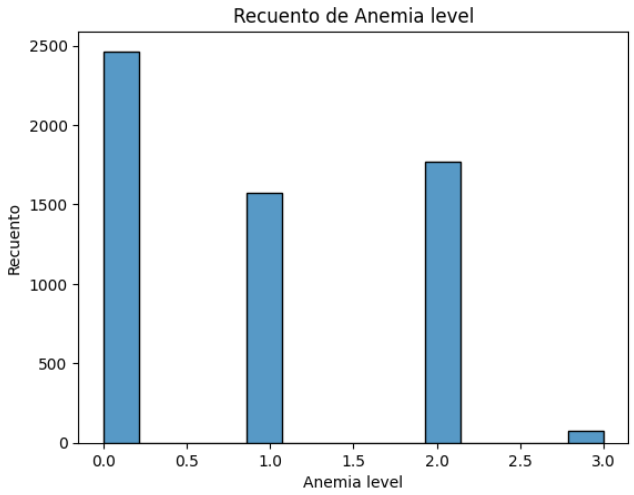

Fig. 2: Recuento de la anemia

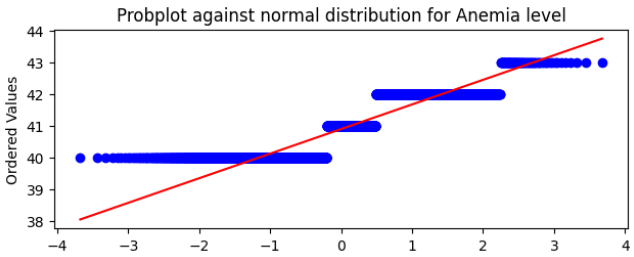

Fig. 3: Probplot anemia

Por último, analizando el histograma y el Probplot del nivel de anemia es poco probable que este siga una distribución conocida.

II. Metodología

La prueba con la que se realizó el experimento consta de tres modelos, “SCV”, “KNN” y "RFC":

SVC: El Support Vector Classification (SVC) es un algoritmo de aprendizaje supervisado diseñado para problemas de clasificación. Su objetivo es encontrar un hiperplano óptimo en un espacio de características que maximice el margen entre clases. Utiliza vectores de soporte, puntos cercanos al hiperplano, para definir este margen. La técnica del kernel le permite manejar datos no lineales al mapearlos a un espacio de características de mayor dimensión. El parámetro de regularización (C) equilibra la maximización del margen y la correcta clasificación de los puntos de entrenamiento. El SVC es versátil y aplicable en casos lineales y no lineales, siendo eficaz en problemas de clasificación donde la separación clara entre clases es fundamental. Este modelo obtuvo buenos resultados con precision, recall y f1-score de 0.92, teniendo solo 61 falsos positivos y 9 falsos negativos.

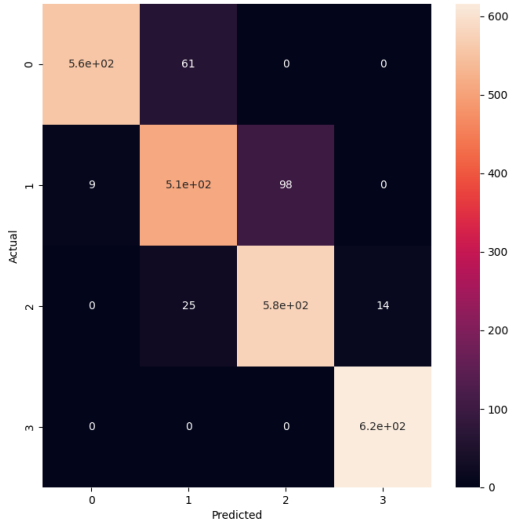

Fig. 4: Matriz de Confusión SVC

KNN (K-Nearest-Neighbors): El segundo método consiste en tomar muestras de un conjunto una vez que este tiene definidas sus características. Dichas muestras tendrán cierto tamaño inferior al tamaño de la población, pero la característica principal es que sus componentes no tendrán mucha diferencia entre ellos. Se trata de encontrar similitudes entre los datos y agruparlos en diferentes grupos, cada uno con componentes con características semejantes. Dichos conjuntos más cercanos son conocidos como “K-Nearest-Neighbors” en inglés, de modo que el subconjunto tendrá K componentes que tengan una fuerte relación entre sí.
Este modelo obtuvo buenos resultados con precision, recall y f1-score de 0.92, teniendo solo 71 falsos positivos y 38 falsos negativos.

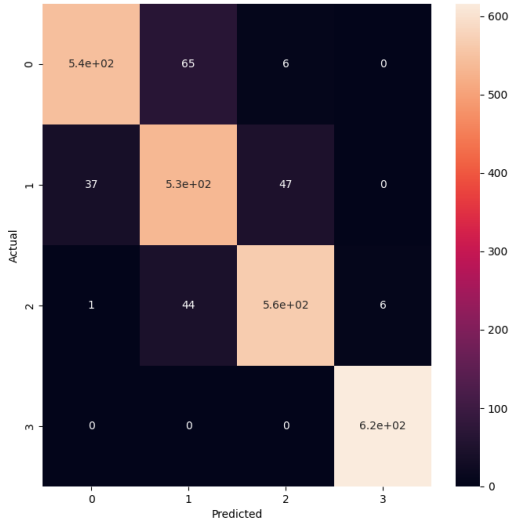

Fig. 5: Matriz de Confusión Knn

Random Forest Classifier: Este es un algoritmo de ensamble, fue implementado utilizando RandomizedSearchCV para encontrar el mejor número de estimadores (N) entre 50, 75, 100, 150, 200 y 300. El modelo resultante, con el mejor parámetro encontrado N=100, demostró un rendimiento sólido

en la clasificación de cuatro clases. El informe de clasificación revela altas precisiones y recall para cada clase, indicando una capacidad confiable para predecir correctamente las categorías. La matriz de confusión visualizada mediante un mapa de calor muestra una buena concordancia entre las predicciones y los valores reales. La métrica de precisión general del modelo alcanzó un 95%, destacando la efectividad del Random Forest en este problema específico. Este enfoque de ajuste de hiperparámetros proporciona una configuración óptima para el modelo, brindando una herramienta robusta y precisa para la clasificación multiclase en este contexto particular. Este modelo también obtuvo buenos resultados teniendo solo 58 falsos positivos y 2 falsos negativos.

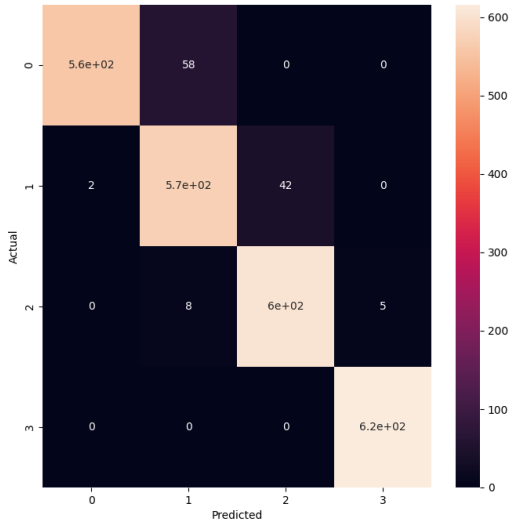

Fig. 6: Matriz de Confusión RandomForest

III. Resultados

Comparativa de métodos

Modelo	Precision	Recall	F1-score	Support
SVC	0.92	0.92	0.92	2464
KNN	0.92	0.92	0.92	2464
RFC	0.95	0.95	0.95	2464

Table 1: Resultados de los modelos SVC, KNN y RFC.

Como se puede observar en la tabla, el mejor modelo es el RandomForestClassifier, pero tiene como desventaja que es un modelo de aprendizaje no supervisado, por lo que no podemos observar como funciona, como posible mejora, se podría probar con un mayor número de estimadores para tener una mayor presicion.