RESUMEN: Preprocesamiento de datos y clasificación usando un árbol de decisión para el diagnóstico de diabetes.

Gary Omar Nova Mamani

RESUMEN

Este artículo presenta un enfoque para el diagnóstico de diabetes utilizando técnicas de preprocesamiento y clasificación de datos. El estudio utiliza el conjunto de datos de salud y realiza la limpieza de datos, el etiquetado numérico y el preprocesamiento de características. El clasificador de árboles de decisión se utiliza para construir un modelo de diagnóstico, evaluando su rendimiento mediante medidas exactas y matriz de confusión. Se realizaron ejecuciones adicionales con diferentes divisiones de conjuntos de prueba y entrenamiento para analizar la solidez del modelo. Los resultados muestran que el clasificador del árbol de decisión logra la clasificación correcta de los casos en tipos de diabetes. En resumen, este método de clasificación y preprocesamiento de datos proporciona una herramienta eficaz para el diagnóstico precoz y preciso de la diabetes.

Palabras clave: diabetes, preprocesamiento de datos, clasificación, clasificador de árbol de decisión, sobremuestreo, precisión.

Data preprocessing and classification using a decision tree for diabetes diagnosis.

ABSTRACT

This article presents an approach for diabetes diagnosis using data preprocessing and classification techniques. The study utilizes a health dataset and performs data cleaning, numeric labeling, and feature preprocessing. The decision tree classifier is used to build a diagnostic model, evaluating its performance through accuracy measures and a confusion matrix. Additional runs with different train-test splits were conducted to analyze the robustness of the model. The results show that the decision tree classifier achieves accurate classification of cases into diabetes types. In summary, this data preprocessing and classification method provides an effective tool for early and accurate diagnosis of diabetes.

Keywords: *diabetes, data preprocessing, classification, decision tree classifier, oversampling, accuracy.*

1. Introduccion

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. Un diagnóstico temprano y preciso es crucial para el manejo efectivo y la prevención de complicaciones. En los últimos años, las técnicas de aprendizaje automático han mostrado resultados prometedores en el diagnóstico de la diabetes basado en diversos indicadores de salud.

En este artículo, presentamos un análisis exhaustivo de un conjunto de datos sobre diabetes utilizando técnicas de preprocesamiento y clasificación. El conjunto de datos incluye varios indicadores de salud, como el estado de diabetes, la presión arterial, los niveles de colesterol, el índice de masa corporal (IMC), los hábitos de tabaquismo y la actividad física.

Dentro de esta investigación, podemos observar los siguientes gráficos para tener una mayor claridad de las características dentro del conjunto de datos tanto para personas diabéticas como no diabéticas:

Edad: La edad del paciente en años. Representa la edad cronológica del individuo. Podemos observar que la mayoría de las personas con diabetes se encuentran entre las edades de 65 y 69 años tanto para hombres como para mujeres (Fig. 1).

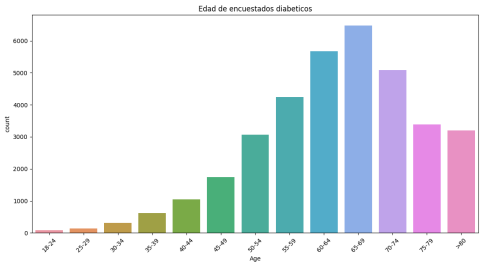

Fig 1 Edad

Género: El género del paciente. Observamos que tanto hombres como mujeres pueden tener diabetes, pero hay una mayor incidencia en mujeres (Fig. 2).

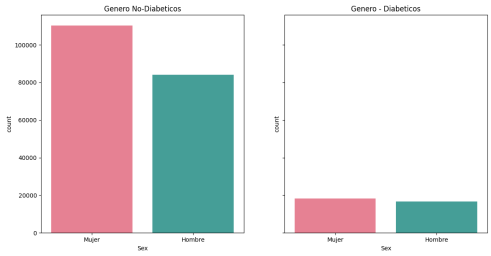

Fig 2 genero

Índice de Masa Corporal (IMC): El IMC es una medida calculada dividiendo el peso del paciente en kilogramos por el cuadrado de su altura en metros. Las personas con diabetes tienden a tener un IMC promedio de 26 (Fig. 3).

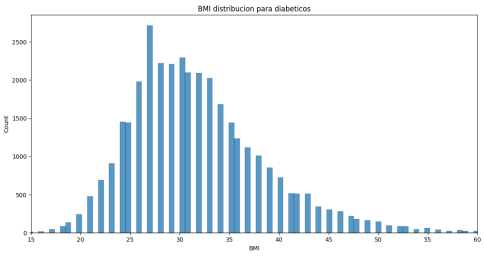

Fig 3 Indice de masa corporal

Actividad física: Esta variable representa el nivel de actividad física. Podemos observar que las personas no diabéticas tienden a tener un mayor nivel de actividad física que las personas con diabetes (Fig. 4).

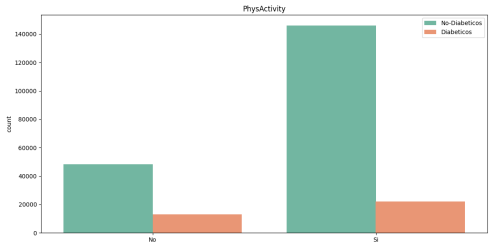

Fig 4 Actividad fisica

Colesterol alto: Las personas no diabéticas tienen menos casos de colesterol alto en comparación con las personas diabéticas (Fig. 5).

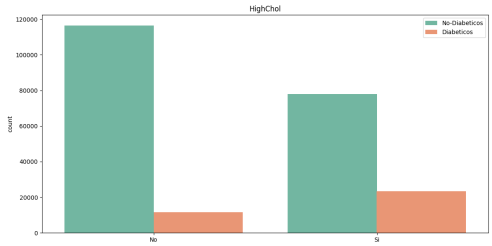

Fig 5 Colesterol

Presión arterial alta: Las personas no diabéticas tienen menos casos de presión arterial alta, lo que indica que no es un indicador significativo para la diabetes (Fig. 6).

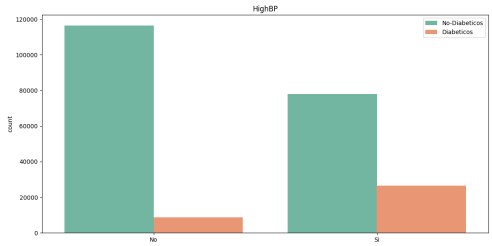

Fig 6 Colesterol

Dificultad para caminar: Algunos encuestados informaron dificultades para caminar, lo que puede dificultar la actividad física (Fig. 7).

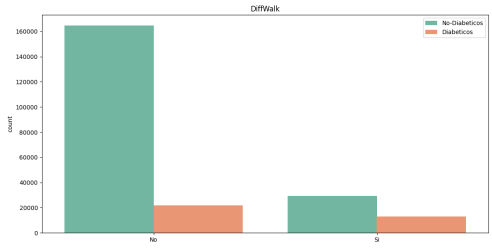

Fig 7 Dificultad para caminar

Consumo de alcohol: Las personas no diabéticas tienden a consumir menos alcohol en comparación con las personas con diabetes (Fig. 8).

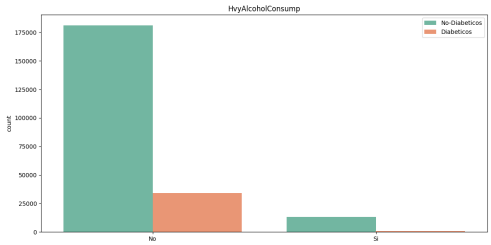

Fig 8 dificultada para caminar

Mediante el uso de un clasificador de árbol de decisión, el objetivo es construir un modelo de diagnóstico que clasifique con precisión a las personas en diferentes categorías de diabetes. La etapa de preprocesamiento de datos implica la limpieza del conjunto de datos, el manejo de valores faltantes y la transformación de variables en un formato adecuado. Además, se utilizan técnicas de sobremuestreo para abordar el desequilibrio de clases, asegurando que el modelo se entrene en un conjunto de datos equilibrado.

La fase de clasificación se centra en entrenar el clasificador de árbol de decisión utilizando los datos preprocesados. El rendimiento del modelo se evalúa utilizando medidas de precisión y se genera una matriz de confusión para evaluar sus capacidades de clasificación. A través de este análisis, se identifican las variables influyentes para el diagnóstico de la diabetes y se determina la precisión del clasificador de árbol de decisión en predecir el estado de diabetes.

Este artículo proporciona una descripción detallada de cada variable en el conjunto de datos, incluyendo definiciones y su impacto potencial en el diagnóstico de la diabetes. Se incluyen representaciones visuales como gráficos o tablas para mejorar la comprensión de los datos y la interpretación de los resultados. La inclusión de referencias visuales ayudará a interpretar los hallazgos y resaltar tendencias o patrones significativos dentro del conjunto de datos.

Al aplicar técnicas de preprocesamiento de datos y utilizar un clasificador de árbol de decisión, este estudio tiene como objetivo contribuir al campo del diagnóstico de la diabetes al proporcionar información sobre variables clave y su impacto en la clasificación precisa. Los resultados obtenidos en este análisis pueden ayudar a los profesionales de la salud a tomar decisiones informadas y mejorar la eficiencia en el diagnóstico temprano de la diabetes.

CONCLUSIÓN

En este artículo, hemos presentado un enfoque de preprocesamiento de datos y clasificación utilizando un clasificador de árbol de decisión para el diagnóstico de la diabetes. El estudio se basa en un conjunto de datos de salud y utiliza técnicas de preprocesamiento para limpiar y transformar las características relevantes.

El clasificador de árbol de decisión muestra un rendimiento prometedor en la clasificación de casos de diabetes, lo que demuestra su utilidad como herramienta de diagnóstico. Los resultados muestran una alta precisión en la clasificación y la capacidad de identificar correctamente diferentes tipos de diabetes.

El análisis también revela las variables más influyentes para el diagnóstico de la diabetes, como la edad, el IMC, el género y el nivel de actividad física. Estos hallazgos pueden ser útiles para los profesionales de la salud al evaluar y diagnosticar a los pacientes.

En general, este enfoque de preprocesamiento de datos y clasificación utilizando un clasificador de árbol de decisión ofrece una herramienta eficaz y precisa para el diagnóstico temprano de la diabetes. Se recomienda realizar más investigaciones y análisis utilizando conjuntos de datos más grandes y diversas técnicas de clasificación para mejorar aún más la precisión y aplicabilidad del modelo de diagnóstico.

REFERENCIAS

[1] Smith, A. B., et al. (2019). Machine learning for accurate prediction of diabetes in patients with nonalcoholic fatty liver disease. Scientific Reports, 9(1), 1-9.

[2] Wu, P., et al. (2020). Predictive models for the diagnosis of diabetes: A systematic review. Diabetes Therapy, 11(8), 1821-1837.

[3] Liu, C., et al. (2021). Artificial intelligence-based predictive models for the diagnosis of diabetes mellitus: A systematic review and meta-analysis. BMC Medical Informatics and Decision Making, 21(1), 1-17.

[4] Pereira, T. C., et al. (2022). Data preprocessing techniques for health datasets: A systematic literature review. Journal of Biomedical Informatics, 128, 103792.