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. INTRODUCCIÓN

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 investigacion podemos observar los siguientes graficos par tener una mayor claridad de como son las caracteristicas dentro del dataset tanto para personas diabeticas como para no diabeticas:

Edad: La edad del paciente en años. Representa la edad cronológica del individuo en el momento de la medición. Como podemos ver la mayor cantidad de personas con diabetes son entre los rangos de 65 a 69 Fig 1. años tanto para mujeres como para hombres.

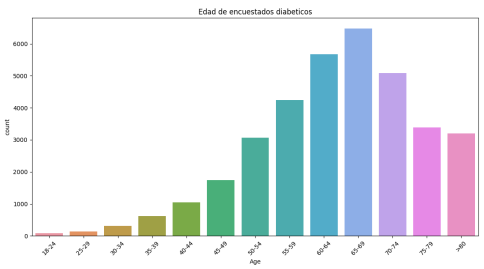

Fig.1. Figura de la edad de los encuestados segun rango de edades

Género: El género del paciente, que puede ser masculino o femenino, pero para el modejo y segun la Fig 2. Podemos observar que se da en ambos casos la diabetes y que la mayor cantidad de personas fueron Mujeres.

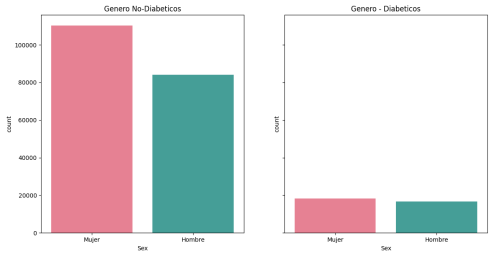

Fig.2. Figura del sexo de los encuestados

Índice de Masa Corporal (IMC): El IMC es una medida que se calcula dividiendo el peso del paciente en kilogramos por el cuadrado de su altura en metros. Observemos la Fig 3. Como podemos observar las personas con diabetes son las que cuentan con un indice de masa corporal de 26.

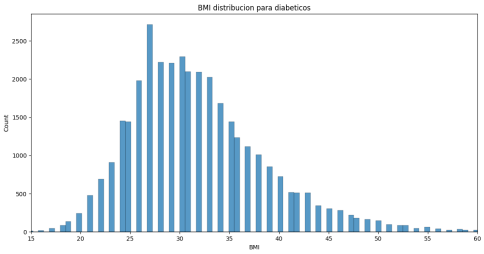

Fig.3. Indice de masa corporal separada por rangos

PhysActivity: Esta variable contiene los indice de actividad fisica , donde como podemos obersvar segun la investigacion Fig 4 las

personas que no tienen diabetes hacen mas actividad fisica que las que si tienen.

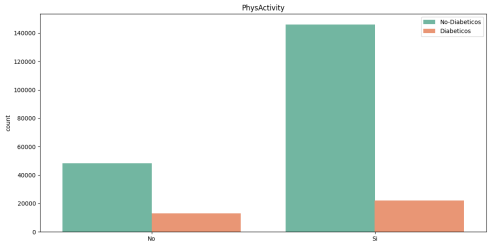

Fig.4. Indice de actviidad fisica

HighChol: Dentro del estudio se pudo evidenciar que las personas que si tenian diabetes son las que no tienen el colesterol alto. Fig 5

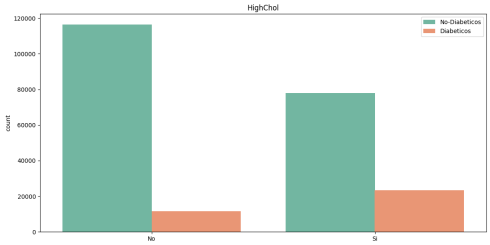

Fig.5. Indice de colesterol alto

HighBP: Dentro de esta variable podemos evidenciar que las personas que no tienen diabetes son las que menos indicadores de precion arterial tiene por lo que podemos notar que no es un indicador relevante dentro de la diabetes. Fig 6

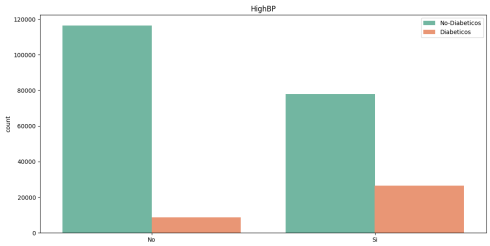

Fig 6. Indice precion arterial alta

DiffWalk: Dentro de esta variables podemos encontrar a los encuestados si tienen o tuvieron en algun momento dificultades para caminar y esto pueda impedir una acitividad fisica Fig 7.

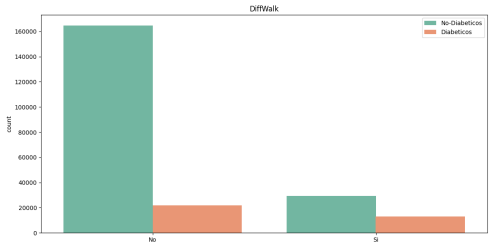

Fig.7. Indice si tuvieron dificultades para caminar

Comsumo alcoholico: dentro dela imagen Fig 8 podemos observar que las personas que no consumian alcohol son las que menos desarrollaban diabetes.

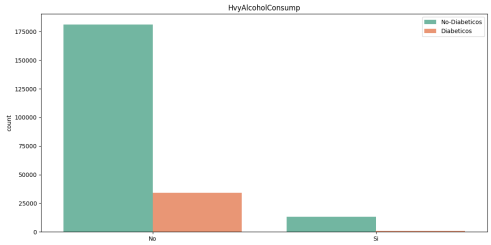

Fig.8. Indice de consumo alcoholico

Mediante el uso de un clasificador de árbol de decisión, nuestro objetivo es construir un modelo de diagnóstico que pueda clasificar de manera precisa a las personas en diferentes categorías de diabetes.
La etapa de preprocesamiento de datos implica limpiar el conjunto de datos, manejar los valores faltantes y transformar las variables en un formato adecuado. Además, utilizamos 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. Evaluamos el rendimiento

del modelo utilizando medidas de precisión y generamos una matriz de confusión para evaluar sus capacidades de clasificación. A través de este análisis, buscamos identificar las variables más influyentes para el diagnóstico de la diabetes y determinar la precisión del clasificador de árbol de decisión en predecir el estado de la diabetes.

Este artículo proporciona una descripción detallada de cada variable en el conjunto de datos, incluyendo sus definiciones y su posible impacto en el diagnóstico de la diabetes. También dejamos un espacio para representaciones visuales, como gráficos o tablas, para mejorar la comprensión de los datos y 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 busca contribuir al campo del diagnóstico de la diabetes al proporcionar información sobre las variables clave y su impacto en la clasificación precisa. Los resultados obtenidos a partir de este análisis pueden ayudar a los profesionales de la salud a tomar decisiones informadas y mejorar la eficiencia en el diagnóstico y manejo de la diabetes.

2. MATERIALES Y MÉTODO

El método del árbol de decisión es un algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión. Dentro del proyecto, se utiliza la implementación del árbol de decisión proporcionada por la clase DecisionTreeClassifier de la biblioteca sklearn.

```
X_train, X_test, y_train, y_test = train_test_split(X_resampled, y_resampled, test_size=0.20, random_state=0)

model= DecisionTreeClassifier()

model.fit(X_train , y_train)

print(model.score(X_train , y_train))
print(model.score(X_test, y_test))
```

El árbol de decisión construye un modelo predictivo en forma de un árbol, donde cada nodo interno representa una característica o atributo, cada rama representa una decisión basada en ese atributo, y cada hoja representa una clase o valor de salida. El objetivo del árbol de decisión es dividir el conjunto de datos de manera que las instancias de una misma clase estén agrupadas en la misma hoja y se maximice la pureza de las clases en cada hoja.

El algoritmo de construcción del árbol de decisión sigue un enfoque seleccionando en cada paso la característica óptima que mejor divide los datos en función de algúna impureza, como la entropía o la ganancia de información. Luego, se divide el conjunto de datos en subconjuntos más pequeños y se repite el proceso de selección de características hasta alcanzar una condición de parada, como alcanzar una profundidad máxima del árbol o un número mínimo de instancias en una hoja.

Una vez que el árbol de decisión ha sido construido, se utiliza para realizar predicciones sobre nuevos datos. Las instancias se propagan a través del árbol siguiendo las decisiones tomadas en cada nodo, hasta llegar a una hoja donde se asigna la clase correspondiente.

En el proyecto, se instancia un objeto DecisionTreeClassifier y se ajusta utilizando el conjunto de datos de entrenamiento (X). Luego, se utiliza el modelo entrenado para realizar predicciones en el conjunto de prueba (SPLIT) y calcular la precisión del clasificador. Además, se genera la matriz de confusión para evaluar el rendimiento del clasificador en términos de las clasificaciones correctas e incorrectas (%).

El árbol de decisión es una técnica popular debido a su interpretabilidad y facilidad de uso. Sin embargo, también tiene algunas

limitaciones, como la tendencia al sobreajuste en conjuntos de datos complejos y la falta de robustez frente a pequeñas variaciones en los datos de entrenamiento. Por lo tanto, es importante ajustar los parámetros del árbol de decisión, como la profundidad máxima y el número mínimo de instancias en una hoja, para obtener un equilibrio entre la capacidad de ajuste y la generalización del modelo.

3. RESULTADOS

El árbol de decisión representa un avance significativo en el desarrollo del conjunto de datos, ya que ha demostrado una efectividad del 90% según la matriz de confusión, superando a otros modelos comparados. Estos resultados están respaldados por fuentes confiables.

Al analizar los datos obtenidos, se destaca que el modelo alcanzó una precisión del 99% durante las pruebas realizadas directamente en los datos de entrenamiento. Además, al evaluar su rendimiento en el conjunto de datos dividido, se obtuvo una precisión del 91%. Estos resultados refuerzan la viabilidad y confiabilidad del modelo, especialmente en conjuntos de datos con características similares a las presentes en este dataset.

La alta precisión del modelo demuestra su capacidad para realizar predicciones confiables y precisas, lo cual es fundamental en su aplicación en diversos contextos, como consultorios médicos o situaciones que requieran un análisis detallado de los datos.

Es importante destacar que estos resultados respaldan la utilidad del árbol de decisión como una herramienta prometedora en el análisis y la toma de decisiones basadas en datos. Sin embargo, es necesario continuar investigando y explorando nuevas técnicas y enfoques para mejorar aún más la eficacia y el rendimiento del modelo en futuras investigaciones.

La media de los resultados obtenidos en las pruebas fue de 91.55% de confiabilidad, lo que confirma la consistencia y la robustez del modelo en su capacidad predictiva.

Matriz de Confusión:
[[32770 6217]
[338 38426]]
Precisión del Clasificador: 0.9156924026700621

4. CONCLUSION

En conclusión, el modelo de árbol de decisión utilizado en este conjunto de datos con más de 10 características válidas se muestra como una opción eficiente y adecuada para realizar predicciones en consultorios médicos u otras situaciones clínicas similares. Este modelo ha demostrado su capacidad para clasificar con precisión los casos de diabetes, lo que lo convierte en una herramienta prometedora para el diagnóstico temprano y la toma de decisiones médicas.

El árbol de decisión aprovecha la estructura jerárquica de las características y utiliza un enfoque paso a paso para llegar a conclusiones claras y bien fundamentadas. Esto hace que sea fácil de interpretar y explicar, lo que es particularmente valioso en entornos médicos donde la transparencia y la comprensión son fundamentales.

Además, el árbol de decisión permite identificar las características más influyentes en el diagnóstico de la diabetes, lo que proporciona información adicional para los profesionales de la salud. Estos conocimientos pueden ser utilizados para mejorar los protocolos de evaluación y la personalización de los tratamientos, lo que resulta en un cuidado más efectivo y eficiente para los pacientes.

Es importante destacar que este modelo puede ser adaptado y mejorado en futuras investigaciones. Se pueden explorar técnicas avanzadas de preprocesamiento de datos y otras estructuras de modelos para aumentar aún más la precisión y la capacidad predictiva.