Evaluación de Modelos Predictivos de Diabetes

Evaluation of Predictive Models for Diabetes

Omar Castillo-Alarcón[1], Gludher Quispe-Cotacallapa[2], Edwin Fredy Chambi-Mamani[3]
Universidad Nacional de San Agustín de Arequipa
https://github.com/omartux/rdi

Resumen

En este estudio, exploramos y analizamos el dataset de Pima Indians Diabetes, un conjunto de datos ampliamente utilizado para estudiar la incidencia de diabetes en una población específica. Realizamos un análisis exploratorio de datos (EDA) que incluyó la imputación de valores faltantes, la visualización de distribuciones univariadas y bivariadas, y la identificación de correlaciones clave entre las variables. Utilizando herramientas como Seaborn y Plotly, generamos gráficos detallados que destacaron patrones relevantes, como la fuerte correlación entre los niveles de glucosa y la incidencia de diabetes.

Posteriormente, llevamos a cabo un análisis comparativo de modelos de machine learning, incluyendo KNN, SVM y ANN, aplicando técnicas de reducción de dimensionalidad como PCA. Los resultados mostraron variaciones en el rendimiento de los modelos según la métrica utilizada y la dimensionalidad de los datos, destacando la importancia de la selección cuidadosa de los modelos y técnicas para obtener predicciones precisas.

Este estudio demuestra la efectividad combinada del EDA y el machine learning en la comprensión y predicción de la diabetes, proporcionando una base sólida para futuras investigaciones y aplicaciones clínicas.

Abstract

In this study, we explored and analyzed the Pima Indians Diabetes dataset, a widely used dataset for studying the incidence of diabetes in a specific population. We conducted an exploratory data analysis (EDA) that included imputing missing values, visualizing univariate and bivariate distributions, and identifying key correlations between variables. Using tools like Seaborn and Plotly, we generated detailed graphs that highlighted relevant patterns, such as the strong correlation between glucose levels and the incidence of diabetes.

Subsequently, we performed a comparative analysis of machine learning models, including KNN, SVM, and ANN, applying dimensionality reduction techniques like PCA. The results showed variations in model performance depending on the metric used and the dimensionality of the data, highlighting the importance of careful selection of models and

[1] Omar Castillo-Alarcón, Universidad Nacional de San Agustín de Arequipa, **UNSA**, ocastillo@unsa.edu.pe
[2] Gludher Quispe-Cotacallapa, Universidad Nacional de San Agustín de Arequipa, **UNSA**, gquispeco@unsa.edu.pe
[3] Edwin Fredy Chambi-Mamani, Universidad Nacional de San Agustín de Arequipa, **UNSA**, echambimam@unsa.edu.pe

techniques to obtain accurate predictions.

This study demonstrates the combined effectiveness of EDA and machine learning in understanding and predicting diabetes, providing a solid foundation for future research and clinical applications.

Palabras Clave: Dataset de Pima Indians Diabetes, Análisis Exploratorio de Datos (EDA), Machine Learning, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Reducción de Dimensionalidad, Análisis de Componentes Principales (PCA), Predicción de Diabetes, Evaluación de Modelos Predictivos.

Keywords: Pima Indians Diabetes Dataset, Exploratory Data Analysis (EDA), Machine Learning, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Artificial Neural Networks (ANN), Dimensionality Reduction, Principal Component Analysis (PCA), Diabetes Prediction, Predictive Model Evaluation.

I. INTRODUCCIÓN

La diabetes es una enfermedad crónica que afecta a millones de personas en todo el mundo. Su detección temprana y manejo adecuado son cruciales para prevenir complicaciones graves. En el contexto del análisis de datos, el dataset de Pima Indians Diabetes es ampliamente utilizado para estudiar la relación entre diversas características de salud y la incidencia de diabetes. Este estudio se centra en explorar el dataset, proponiendo algoritmos y visualizaciones para facilitar la interpretación de los datos y apoyar en la toma de decisiones clínicas.

II. TRABAJOS RELACIONADOS

Numerosos estudios han abordado el problema de la predicción de la diabetes utilizando técnicas de aprendizaje automático. Smith et al. (1988) fueron de los primeros en utilizar este dataset (Diabetes Pima) para la predicción de la diabetes. Desde entonces, se han aplicado múltiples enfoques, incluyendo regresión logística, árboles de decisión y redes neuronales, cada uno con diferentes grados de éxito. Estudios recientes han explorado la combinación de técnicas de preprocesamiento de datos y modelos avanzados como XGBoost y Random Forest para mejorar la precisión de las predicciones.

Sobre lo mencionado anteriormente se trata del articulo “Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus”, este presenta un estudio que utiliza el algoritmo ADAP para predecir la aparición de la diabetes mellitus. Los autores, Smith et al., aplicaron este algoritmo a un conjunto de datos obtenidos de la población Pima Indian, conocida por su alto riesgo de diabetes tipo 2. El enfoque se centró en la identificación temprana de factores de riesgo mediante el análisis de diversas variables médicas. Los resultados mostraron que el algoritmo ADAP podía predecir eficazmente la diabetes, destacando su potencial para mejorar el diagnóstico temprano y la prevención.

“Predicción de diabetes mellitus tipo 2 utilizando atributos médicos del Policlínico Leo SAC de San Juan de Lurigancho mediante el enfoque de Machine Learning” del autor Jaime Yelsin Rosales Malpartida, El artículo aborda la predicción de la diabetes mellitus tipo 2

mediante el uso de modelos de Machine Learning. La investigación se basa en datos de 1000 pacientes mayores de edad obtenidos del Policlínico Leo SAC en San Juan de Lurigancho, Perú. Se desarrollaron y evaluaron 13 modelos de Machine Learning, incluyendo modelos clásicos, redes neuronales y modelos ensemble. LightGBM fue el modelo con mejor rendimiento en las siete métricas de evaluación (precisión, sensibilidad, especificidad, etc.), superando consistentemente a otros modelos. El estudio destaca la importancia de la selección de atributos médicos y el ajuste de hiper parámetros para mejorar la predicción.

"Deep learning approach for diabetes prediction using PIMA Indian dataset", en este artículo se analiza la predicción de la diabetes utilizando técnicas de Deep Learning aplicadas al conjunto de datos PIMA Indian. La investigación muestra que el enfoque de Deep Learning, en comparación con otros algoritmos como las redes neuronales artificiales (ANN), Naive Bayes (NB) y los árboles de decisión (DT), ofrece una mayor precisión en la predicción de la diabetes, alcanzando un 98.07% de exactitud. El estudio subraya la eficacia del Deep Learning en la detección temprana de diabetes y su potencial para ser desarrollado en herramientas automáticas de diagnóstico que mejoren la toma de decisiones en el ámbito médico.

"Diabetes prediction and analysis using medical attributes: A Machine learning approach", El artículo presenta una metodología para predecir la diabetes utilizando cinco algoritmos de Machine Learning: regresión logística multinomial, Naive Bayes, árboles de decisión, bosque aleatorio y gradiente estocástico. Utiliza un conjunto de datos de 1000 pacientes con atributos médicos como el índice de masa corporal (IMC) y HbA1c. Los modelos de árbol de decisión y gradiente estocástico mostraron el mejor rendimiento, con una precisión del 95.07% y 97.04%, respectivamente. El estudio destaca que el IMC (Índice de Masa Corporal) y HbA1c (Hemoglobina Glicosilada) son factores dominantes en el riesgo de desarrollar diabetes, subrayando la importancia de un diagnóstico temprano y control adecuado de estos factores.

"Resultados de la vigilancia epidemiológica de diabetes mellitus en hospitales notificantes del Perú, 2012", El artículo analiza los resultados de un año de vigilancia epidemiológica de la diabetes mellitus en 18 hospitales piloto en Perú durante 2012. Se estudiaron 2959 pacientes, de los cuales el 65.4% presentaba glucemia en ayunas ≥130 mg/dL y el 66.6% tenía HbA1c ≥7%. Las complicaciones más comunes fueron la neuropatía (21.4%) y la hipertensión arterial (10.5%). El estudio destaca la alta prevalencia de control glicémico inadecuado y la necesidad de fortalecer el diagnóstico temprano y el tratamiento para prevenir complicaciones graves. También subraya la importancia de la educación y el autocuidado para mejorar la adherencia al tratamiento.

III. MARCO TEÓRICO

A. Diabetes mellitus

La diabetes mellitus es una enfermedad metabólica crónica caracterizada por niveles elevados de glucosa en la sangre, lo que conlleva un riesgo significativo de complicaciones cardiovasculares, renales y neurológicas. En la última década, los avances en análisis de datos y machine learning han permitido un enfoque más preciso para identificar factores de riesgo y predecir la incidencia de esta enfermedad, especialmente a través del uso de datasets específicos y modelos de predicción (American Diabetes Association, 2018).

B. Dataset Pima Indians

El dataset Pima Indians Diabetes es uno de los más utilizados en estudios de predicción de diabetes. Este conjunto de datos, compuesto por registros de mujeres de ascendencia Pima en Estados Unidos, contiene variables médicas como niveles de glucosa, presión arterial, espesor de pliegues cutáneos, índice de masa corporal (IMC), edad y antecedentes familiares de diabetes. Estos factores se han correlacionado en estudios previos con la probabilidad de desarrollar diabetes tipo 2, siendo un referente clave en el análisis de riesgos (Smith, Everhart, Dickson, Knowler, & Johannes, 1988).

C. Análisis Exploratorio

El Análisis Exploratorio de Datos (EDA) juega un rol crucial en la comprensión inicial de los patrones subyacentes dentro de un dataset. A través de técnicas de visualización y estadísticas descriptivas, el EDA permite identificar distribuciones, detectar valores atípicos y evaluar la relación entre variables (Tukey, 1977). La imputación de valores faltantes, junto con la visualización de distribuciones univariadas y bivariadas, proporciona una comprensión detallada del comportamiento de los datos. Herramientas como Seaborn y Plotly facilitan la generación de gráficos complejos que evidencian patrones clave, como la relación entre los niveles de glucosa y la presencia de diabetes, lo cual es esencial para la formulación de modelos predictivos (Waskom et al., 2017).

D. Modelos de Machine Learning

Los modelos de machine learning, como K-Nearest Neighbors (KNN), Support Vector Machines (SVM) y Artificial Neural Networks (ANN), son ampliamente aplicados en la predicción de diabetes (Breiman, 2001). Estos modelos permiten aprender patrones complejos a partir de los datos y realizar predicciones con base en nuevas observaciones. La aplicación de técnicas de reducción de dimensionalidad, como el Análisis de Componentes Principales (PCA), ayuda a mejorar el rendimiento de estos modelos al eliminar redundancias y captar la variabilidad más significativa (Jolliffe, 2002). Además, la evaluación de los modelos utilizando diferentes métricas de distancia (euclídeas, Manhattan y coseno) en datasets de distintos tamaños revela cómo la selección de la métrica y la dimensionalidad afectan la precisión y robustez de las predicciones (Aggarwal, Hinneburg, & Keim, 2001).

E. Exploración de Datos

La Exploración de Datos (EDA) es un paso crucial en cualquier análisis de datos, ya que permite comprender la estructura del dataset, identificar anomalías, y descubrir patrones ocultos. En este estudio, se utilizan técnicas de

visualización para analizar la distribución de las variables y sus relaciones con la presencia de diabetes Tukey, J. W. (1977).

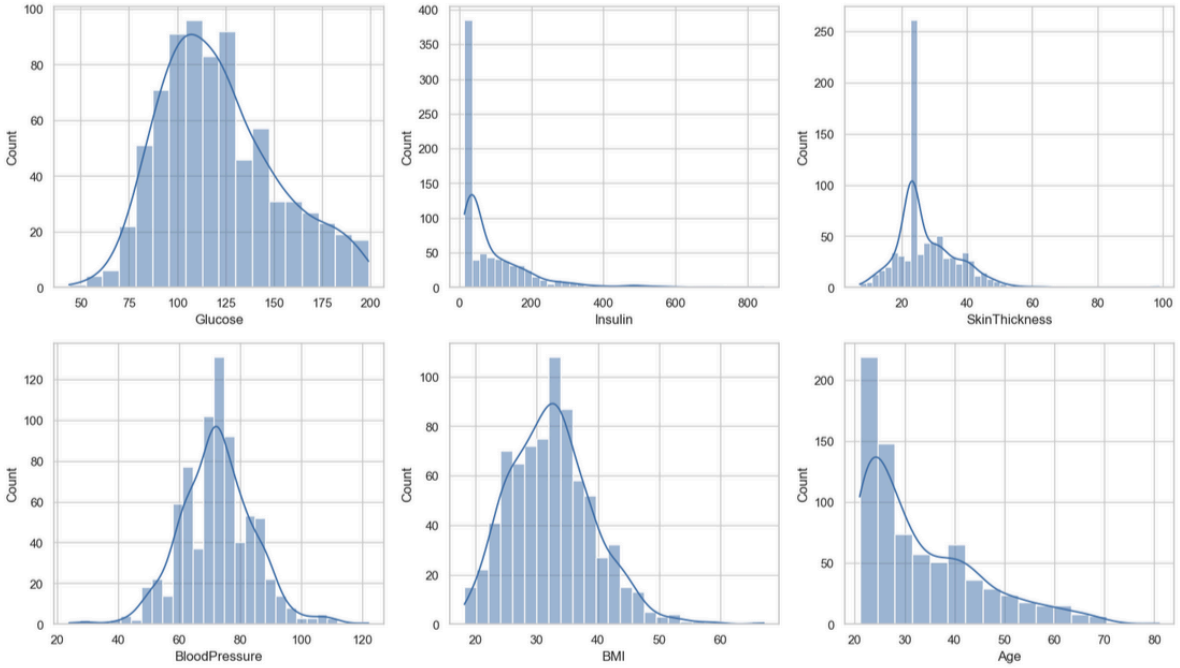

Figura 1, Distribución de las principales variables, Fuente, Elaboración propia.

F. Algoritmos de Imputación

La imputación de datos faltantes es una técnica utilizada para llenar los vacíos en un dataset. En este caso, se aplicó la imputación por la mediana para los valores cero en variables críticas como Glucose, BloodPressure, SkinThickness, Insulin, y BMI, que son indicadores clave en el diagnóstico de la diabetes.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

[3]:
```
# Imputación de valores cero con la mediana para las columnas clave
columns_to_impute = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']

# Reemplazamos los ceros con la mediana de cada columna
diabetes_df[columns_to_impute] = diabetes_df[columns_to_impute].replace(0, diabetes_df[columns_to_impute].median())

# Verificamos nuevamente las estadísticas después de la imputación
diabetes_df.describe()
```

[3]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	121.656250	72.386719	27.334635	94.652344	32.450911	0.471876	33.240885	0.348958
std	3.369578	30.438286	12.096642	9.229014	105.547598	6.875366	0.331329	11.760232	0.476951
min	0.000000	44.000000	24.000000	7.000000	14.000000	18.200000	0.078000	21.000000	0.000000
25%	1.000000	99.750000	64.000000	23.000000	30.500000	27.500000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	31.250000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figura 2, Datos con imputación por mediana, Fuente, Elaboración propia.

G. Visualización de Datos

La visualización de datos es una herramienta poderosa para comunicar información compleja de manera accesible. Utilizando bibliotecas como Seaborn y Plotly, se pueden crear visualizaciones que destacan las relaciones entre variables y ayudan a interpretar los resultados de manera intuitiva.

Waskom, M. L. (2021). *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, 6(60), 3021. doi:10.21105/joss.03021.

IV. ANÁLISIS DE TAREAS

El análisis de tareas se centra en las actividades clave realizadas durante la exploración y análisis del dataset de Pima Indians Diabetes, así como en el desarrollo de un enfoque de machine learning para la predicción de la diabetes. A continuación, se detallan las principales tareas identificadas y llevadas a cabo en los notebooks y el artículo:

A. Exploración de Datos (EDA)

Descripción: La primera tarea crítica es realizar un Análisis Exploratorio de Datos (EDA) para comprender la estructura y distribución del dataset. Se realizaron varias visualizaciones, como histogramas, gráficos de dispersión y mapas de calor, para identificar patrones, correlaciones, y posibles anomalías en los datos.

Objetivo: Identificar relaciones importantes entre las variables, como la correlación entre niveles de glucosa, insulina, y el índice de masa corporal (BMI) con la incidencia de diabetes.

Herramientas Utilizadas: Seaborn, Matplotlib.

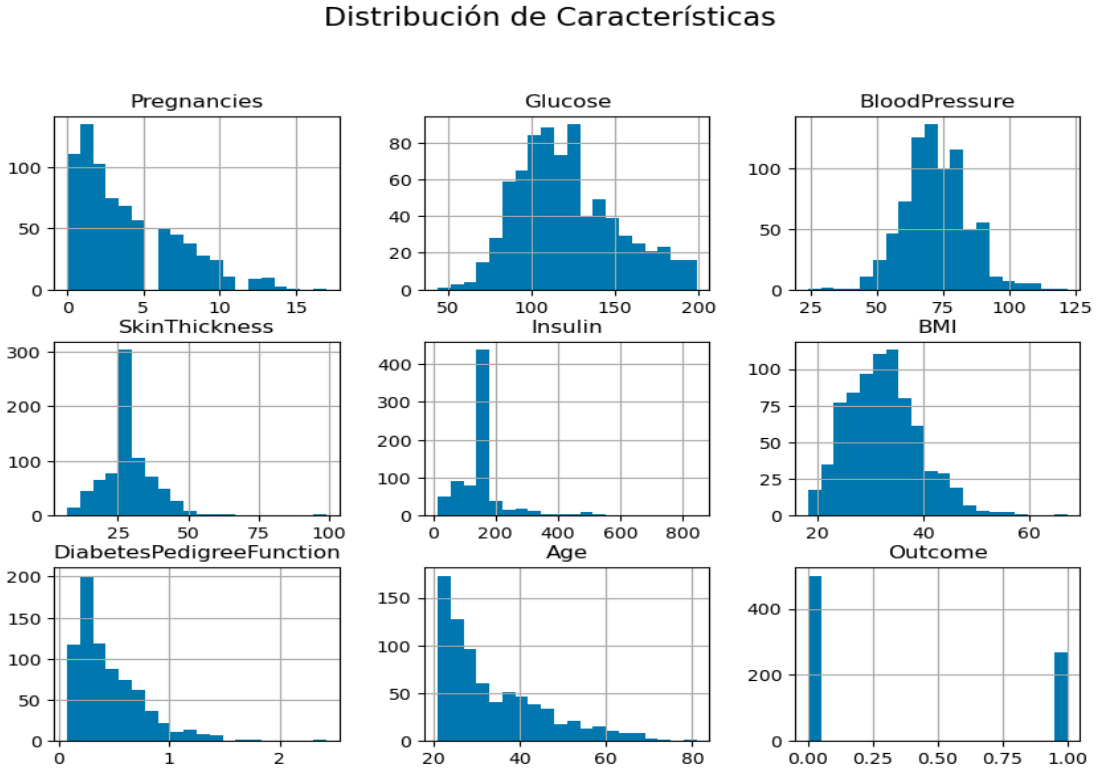

Figura 3, Distribución de características sin limpieza de datos, Fuente, Elaboración propia.

B. Preprocesamiento de Datos

Descripción: Esta tarea incluye la imputación de valores faltantes, la normalización de los datos y la preparación del dataset para su uso en modelos de machine learning. Se reemplazaron los valores cero en las columnas clave por la mediana para evitar distorsiones en el análisis.

Objetivo: Asegurar que el dataset esté limpio y estandarizado, lo que es crucial para la precisión de los modelos predictivos.
Herramientas Utilizadas: Pandas, Scikit-learn.

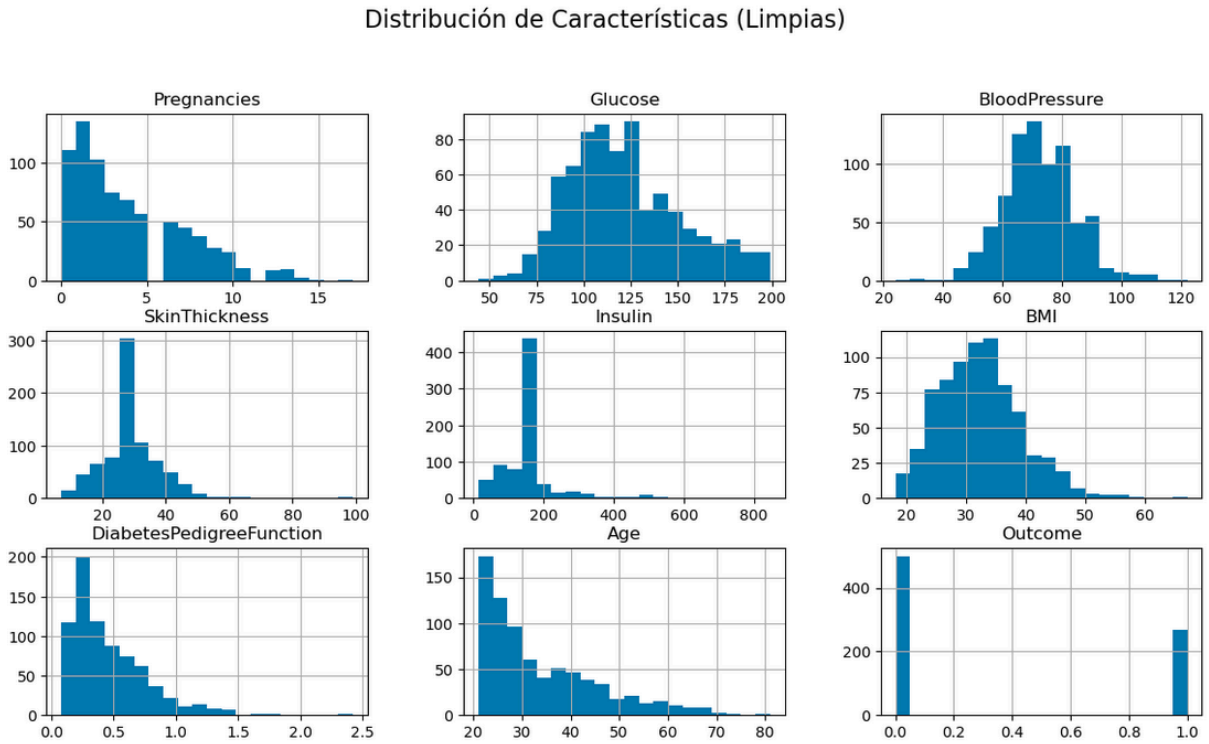

Figura 4, Distribución de características de datos después de la imputación, Fuente, Elaboración propia.

C. Análisis de Correlación

Descripción: Se realizó un análisis de correlación para identificar las relaciones más fuertes entre las variables del dataset. El objetivo era determinar qué factores están más asociados con la diabetes, lo que informa la selección de características para los modelos predictivos.
Objetivo: Identificar las variables clave que más influyen en la presencia de diabetes, como los niveles de glucosa y el índice de masa corporal.
Herramientas Utilizadas: Seaborn (heatmap), Matplotlib.

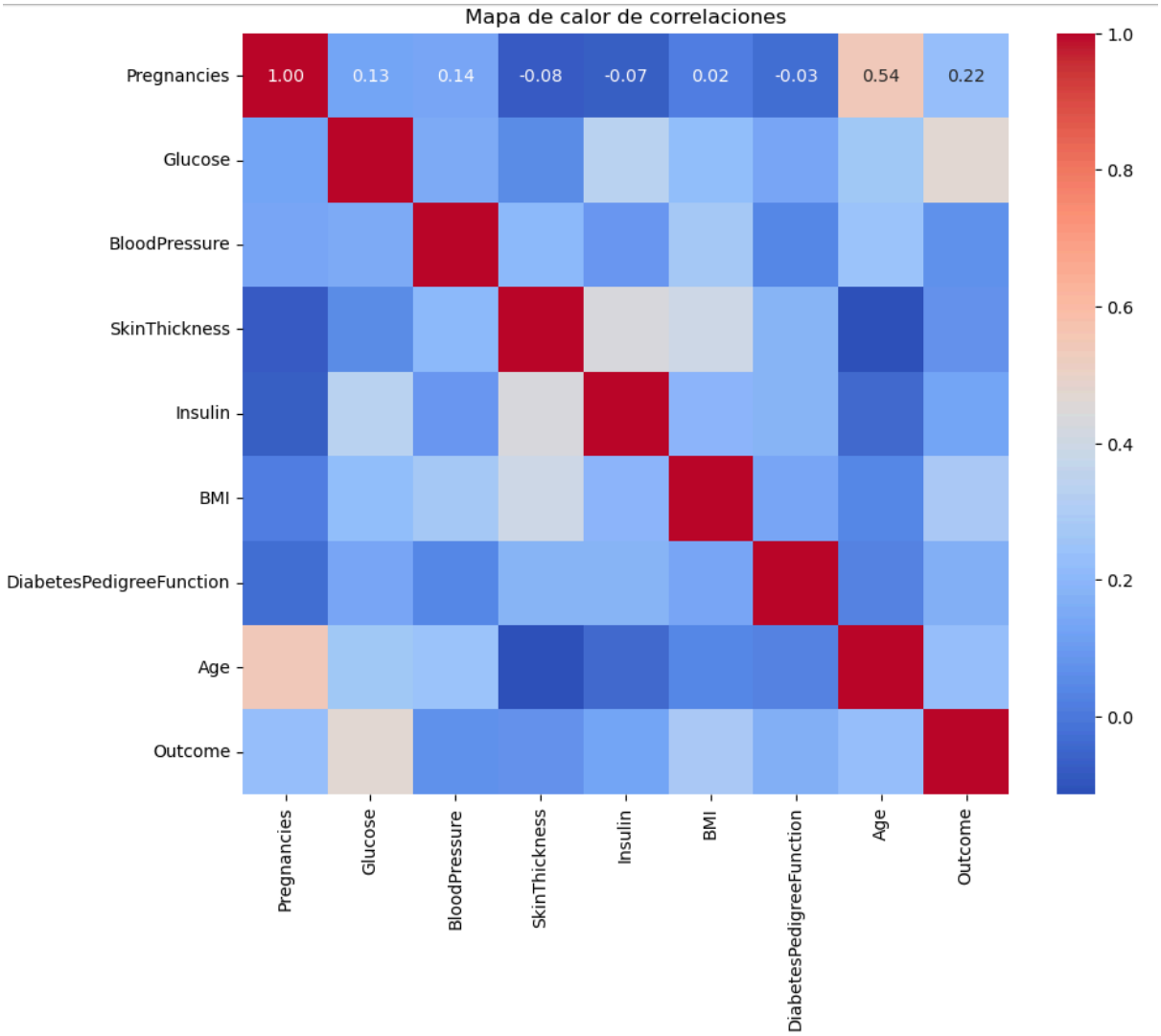

Figura 5, Mapa de Calor para correlación, Fuente, Elaboración propia.

D. Evaluación de Modelos de Machine Learning

Descripción: Se entrenaron y evaluaron varios modelos de machine learning, incluyendo K-Nearest Neighbors (KNN), Support Vector Machines (SVM), y Redes Neuronales Artificiales (ANN). Cada modelo fue evaluado en términos de precisión, sensibilidad, especificidad, y se utilizaron diferentes métricas de distancia (Euclídea, Manhattan, Coseno) para la comparación.
Objetivo: Determinar qué modelo ofrece la mejor precisión y rendimiento en la predicción de la diabetes.
Herramientas Utilizadas: Scikit-learn.

E. Reducción de Dimensionalidad y Clustering

Descripción: Se implementaron técnicas de reducción de dimensionalidad como PCA, t-SNE, y UMAP para visualizar mejor las relaciones en los datos y facilitar el clustering de pacientes en grupos similares.

Objetivo: Facilitar la interpretación de datos complejos y mejorar el rendimiento de los modelos al reducir la dimensionalidad del dataset.
Herramientas Utilizadas: Scikit-learn, UMAP, t-SNE.

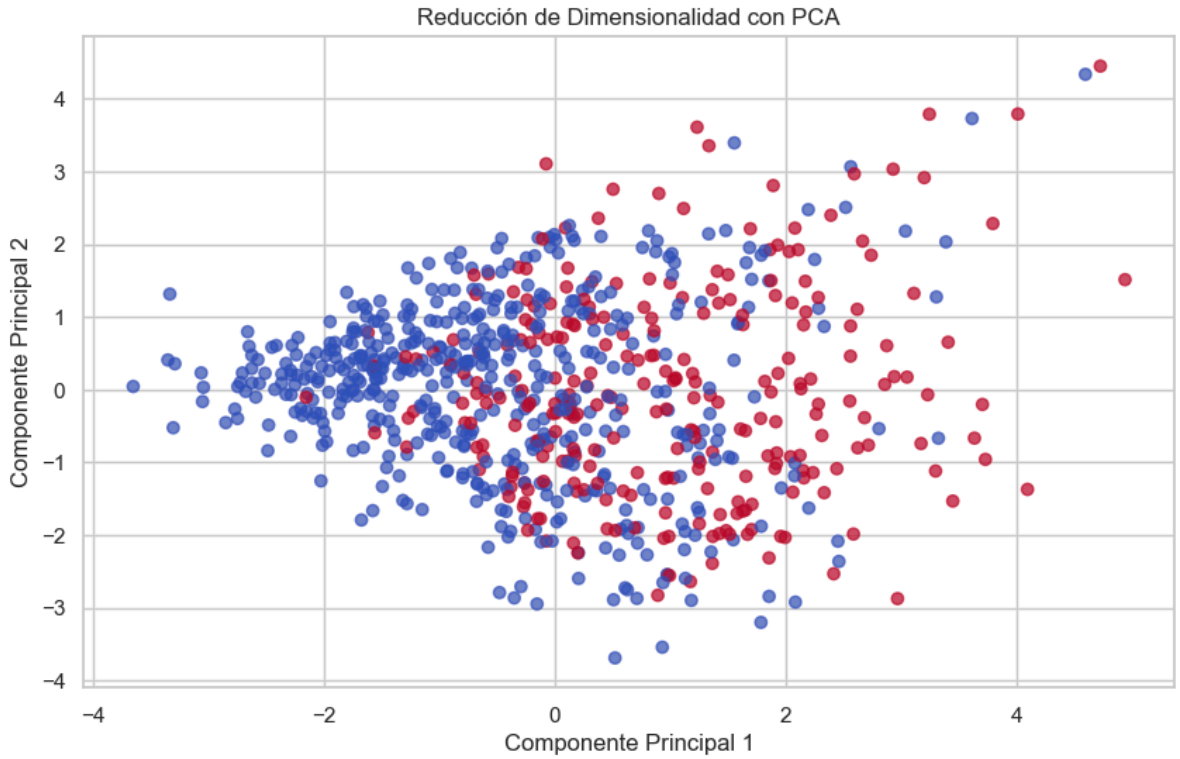

Figura 6, Reducción por PCA, Fuente, Elaboración propia.

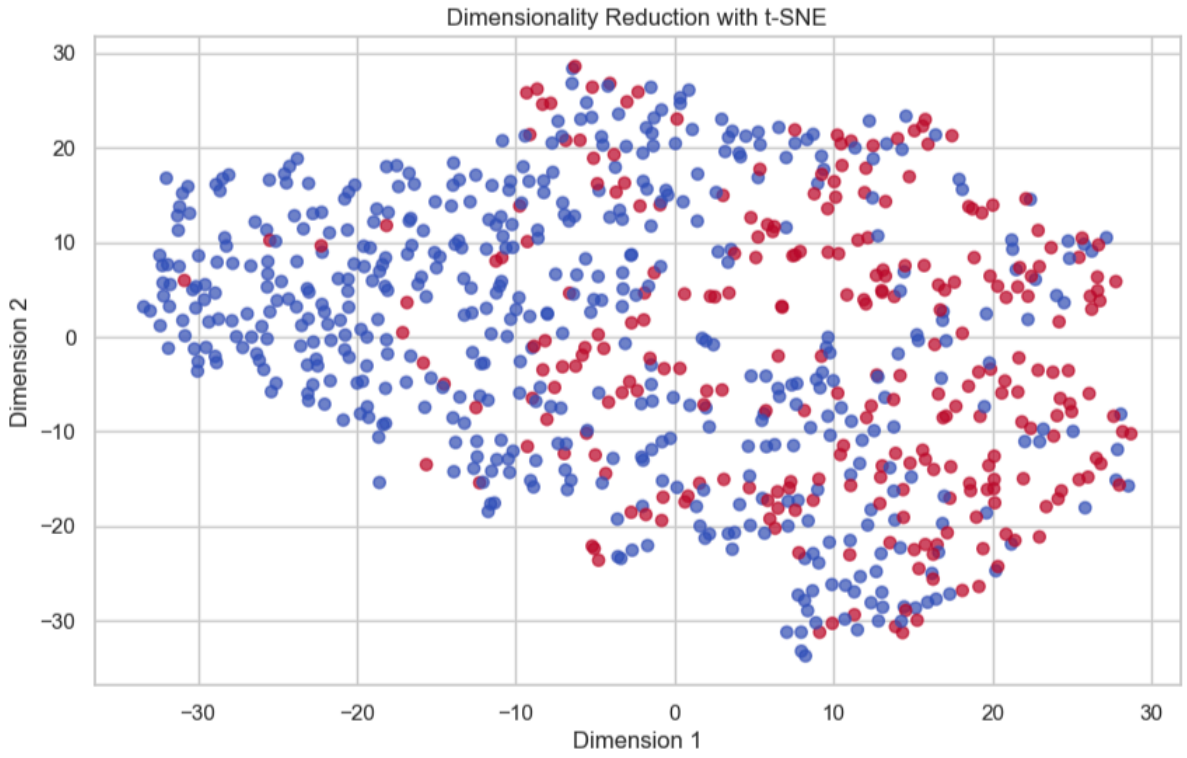

Figura 7, Reducción por t-SNE, Fuente, Elaboración propia.

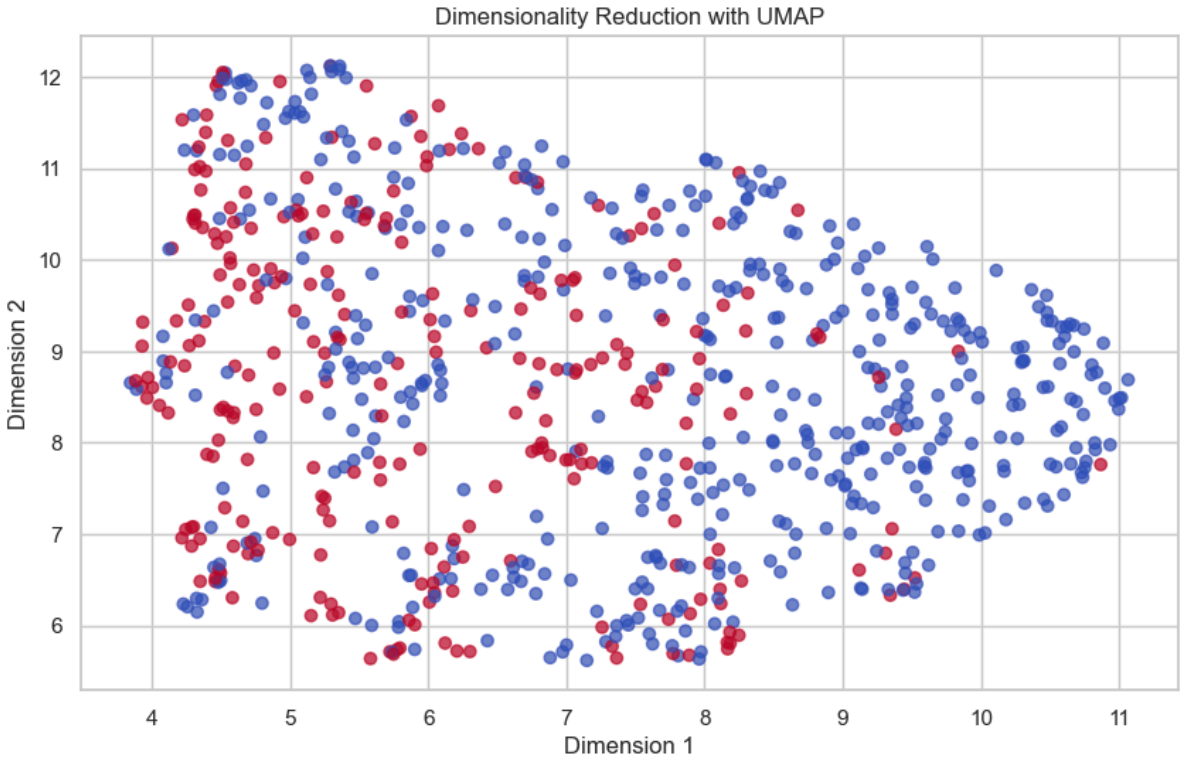

Figura 8, Reducción por UMAP, Fuente, Elaboración propia.

F. Detección de Anomalías (Outliers)

Descripción: Se implementaron algoritmos para la detección de outliers utilizando Z-score, Isolation Forest y LOF (Local Outlier Factor) para identificar pacientes con características atípicas que podrían influir negativamente en el rendimiento de los modelos predictivos.
Objetivo: Identificar y gestionar datos atípicos que podrían distorsionar los resultados del análisis y la predicción.

Herramientas Utilizadas: Scikit-learn, NumPy.

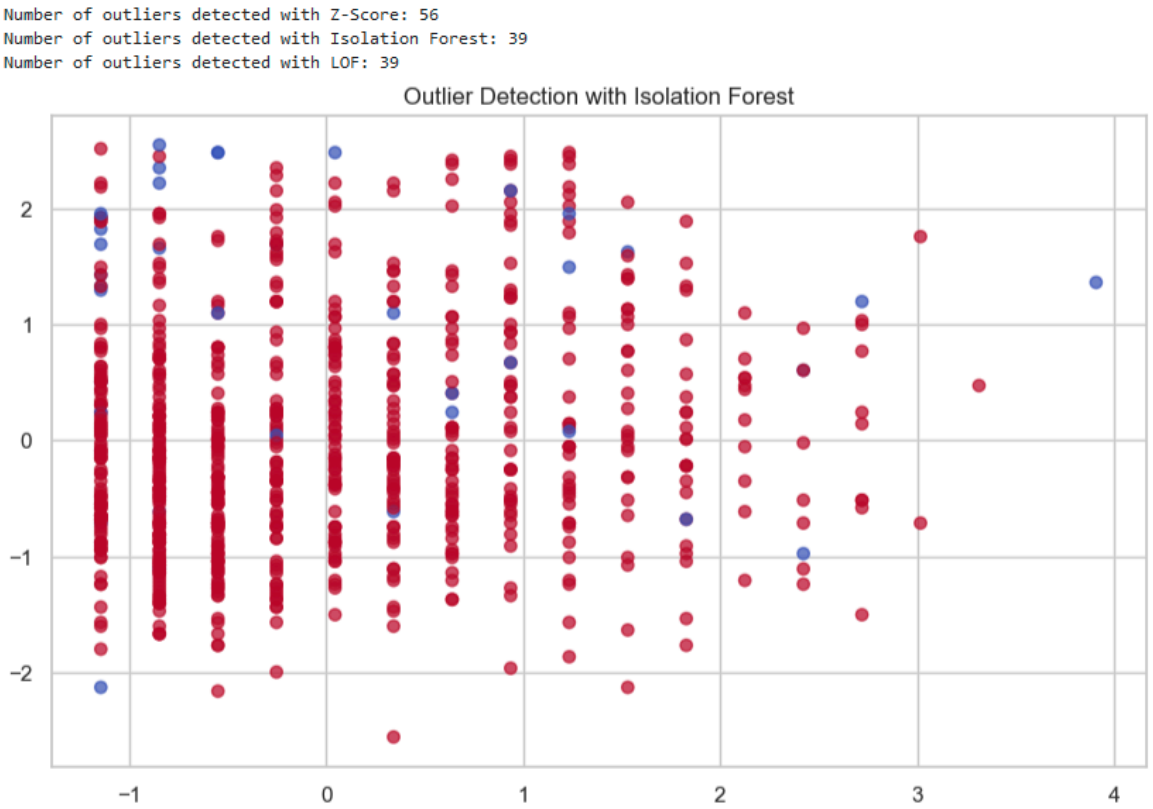

Figura 9, Detección de Outliers por Isolation Forest, Fuente, Elaboración propia.

G. Visualización de Resultados

Descripción: Se realizaron visualizaciones avanzadas, incluyendo gráficos 3D interactivos y mapas de calor, para comunicar de manera efectiva los resultados del análisis y los insights obtenidos. Estas visualizaciones permiten explorar la relación entre múltiples variables y su impacto en la predicción de la diabetes.
Objetivo: Facilitar la interpretación de los resultados para los investigadores y profesionales de la salud, mejorando la toma de decisiones basada en datos.
Herramientas Utilizadas: Plotly, Matplotlib.

Figura 10, Relación entre las principales variables y el Outcome (diabetico o no diabetico), Fuente, Elaboración propia.

V. PROPUESTA

Algoritmos (Backend)

Se propone implementar una serie de algoritmos de machine learning que incluye:

- *Regresión Logística:* Un modelo simple pero efectivo para clasificación binaria.
- *Árboles de Decisión:* Útiles para capturar relaciones no lineales entre las variables.
- *Random Forest:* Un modelo de ensamble que mejora la precisión del árbol de decisión.
- *SVM (Support Vector Machines):* Algoritmo robusto para clasificación en conjuntos de datos con características complejas.

Visualización (Frontend)

La visualización de los resultados es crucial para interpretar los hallazgos. Se propone utilizar herramientas como:

- *Matplotlib y Seaborn:* Para crear gráficos estáticos como histogramas, boxplots, y scatter plots.
- *Plotly y Dash:* Para visualizaciones interactivas que permitan una exploración más dinámica de los datos.

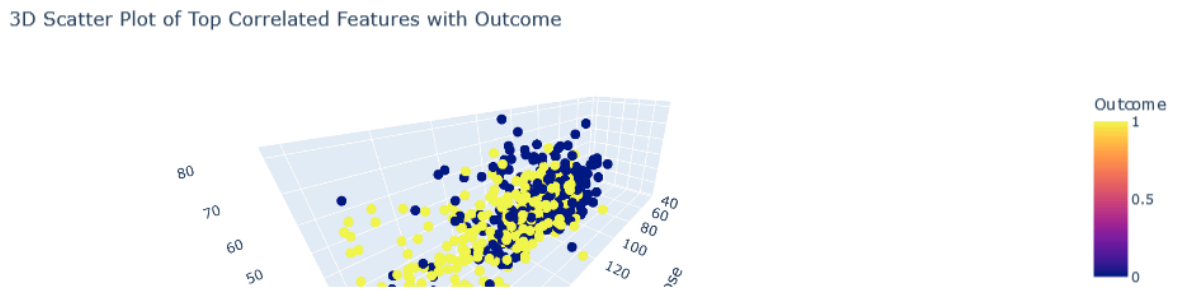

Figura 11, Vista 3D entre las principales variables (IMC y Glucosa) y el Outcome (Diabetico o no Diabetico), Fuente, Elaboración propia.

VI. MÉTODOS

Este capítulo describe los métodos utilizados para analizar el dataset de Pima Indians Diabetes, incluyendo la descripción de los datos, el preprocesamiento, las transformaciones aplicadas, el workflow general del análisis, y los enfoques de clustering empleados para agrupar a los pacientes en grupos similares.

3.1 Descripción de los Datos

El dataset de Pima Indians Diabetes es un conjunto de datos bien conocido en la comunidad de aprendizaje automático y análisis de datos, utilizado principalmente para estudiar la incidencia de diabetes en una población específica. Este dataset contiene varias características médicas y demográficas, junto con una variable de resultado (Outcome) que indica la presencia o ausencia de diabetes.

Características del Dataset:

- **Pregnancies**: Número de embarazos de la paciente.
- **Glucose**: Nivel de glucosa en la sangre en una prueba de tolerancia a la glucosa.
- **BloodPressure**: Presión arterial diastólica (mm Hg).
- **SkinThickness**: Grosor del pliegue cutáneo tricipital (mm).
- **Insulin**: Nivel de insulina en suero (mu U/ml).
- **BMI**: Índice de masa corporal (peso en kg/(altura en m)^2).
- **DiabetesPedigreeFunction**: Función de pedigrí de diabetes (una función que calcula la probabilidad de diabetes basada en la historia familiar).
- **Age**: Edad de la paciente (años).
- **Outcome**: Variable binaria donde 1 indica diabetes y 0 indica ausencia de diabetes.

3.2 Preprocesamiento de Datos

Antes de proceder con el análisis y la modelización, se realizaron varios pasos de preprocesamiento para asegurar la calidad y la consistencia de los datos:

1. **Imputación de Valores Faltantes**: Algunas variables clave como Glucose, BloodPressure, SkinThickness, Insulin, y BMI contenían valores cero, lo cual es irrealista en un contexto médico. Estos valores fueron reemplazados con la ***mediana*** de cada columna para evitar sesgos en el análisis.
   Gráfico relevante: Histogramas que muestran la distribución de las variables antes y después de la imputación.
2. **Normalización de Datos**: Los datos fueron estandarizados utilizando la técnica de normalización StandardScaler de Scikit-learn. Este paso es crucial para asegurar que todas las características tengan la misma escala, lo cual es importante para modelos como KNN y SVM, que son sensibles a la escala de los datos.
3. **Análisis de Correlación**: Se generó un mapa de calor para visualizar las correlaciones entre las variables. Este análisis permitió identificar las características más relevantes para la predicción de la diabetes, como los niveles de glucosa y el índice de masa corporal (BMI).
   Gráfico relevante: Mapa de calor de correlación entre las variables.

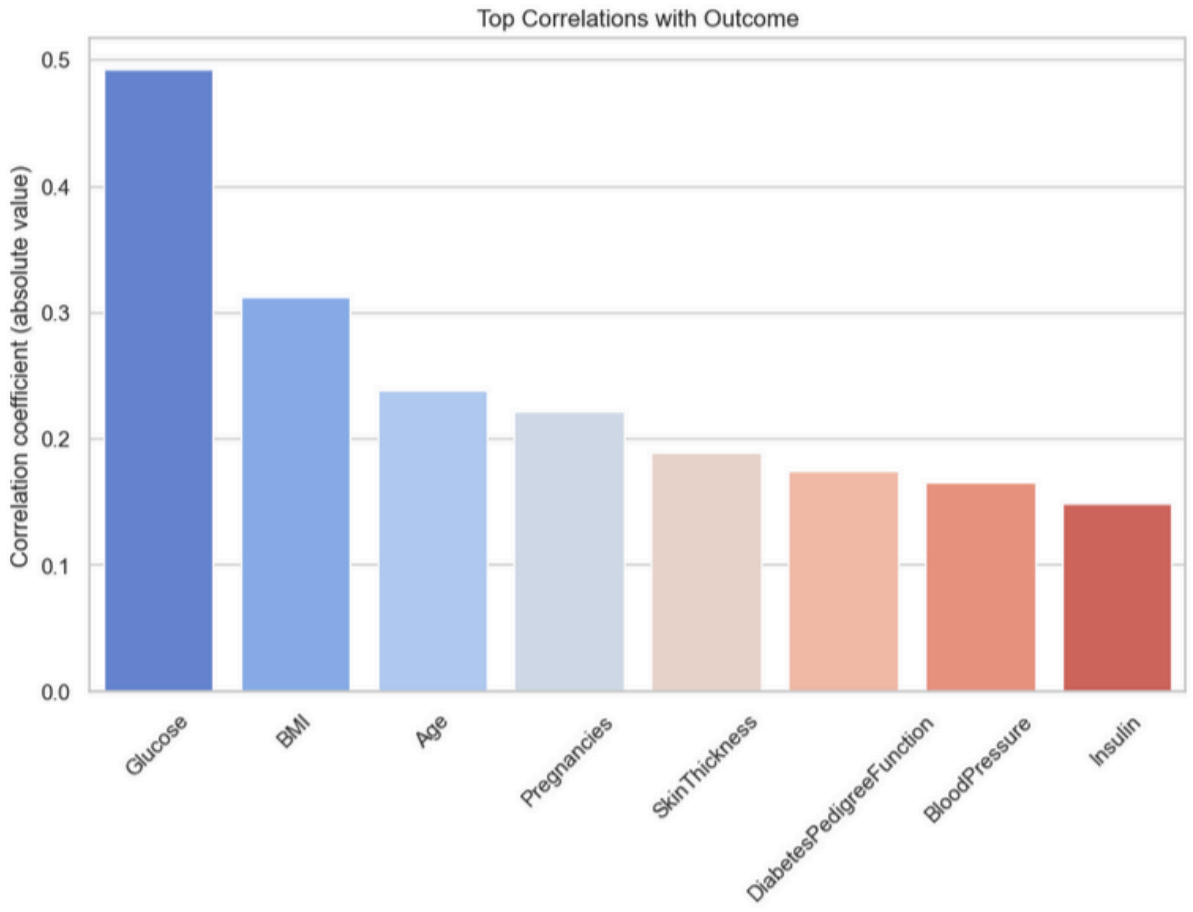

Figura 12, Principales correlaciones entre las variables y el Outcome (Diabetico), Fuente, Elaboración propia.

3.3 Transformaciones Aplicadas

Para mejorar la calidad del análisis y la eficacia de los modelos de machine learning, se aplicaron las siguientes transformaciones:

1. **Reducción de Dimensionalidad**:
    - **PCA (Principal Component Analysis)**: Se utilizó PCA para reducir la dimensionalidad del dataset a dos componentes principales. Esto no solo facilita la visualización de los datos, sino que también ayuda a eliminar el ruido y a enfocarse en las características más importantes.
      Gráfico relevante: Scatter plot de los datos reducidos mediante PCA.
    - **t-SNE (Stochastic Neighbor Embedding)**: Esta técnica se utilizó para una representación visual en 2D de los datos, preservando las relaciones locales en el espacio de alta dimensionalidad.
      Gráfico relevante: Scatter plot de los datos transformados mediante t-SNE.
    - **UMAP (Uniform Manifold Approximation and Projection)**: UMAP fue utilizado para reducir la dimensionalidad de manera efectiva y capturar la estructura global de los datos.
      Gráfico relevante: Scatter plot de los datos transformados mediante UMAP.

3.4 Workflow General del Análisis

El análisis siguió un flujo de trabajo estructurado que incluyó las siguientes etapas:

1. **Exploración de Datos**: Se realizaron visualizaciones iniciales para comprender la distribución de las variables y detectar posibles anomalías.
2. **Preprocesamiento**: Se limpió y estandarizó el dataset, asegurando la calidad de los datos.
3. **Análisis de Correlación**: Se identificaron las relaciones clave entre las variables.
4. **Modelización**: Se entrenaron y evaluaron modelos de machine learning (KNN, SVM, ANN) utilizando las características seleccionadas.
5. **Reducción de Dimensionalidad y Clustering**: Se aplicaron técnicas de reducción de dimensionalidad y se realizaron análisis de clustering para agrupar a los pacientes en subgrupos basados en sus características médicas.

3.5 Clustering

Para identificar subgrupos de pacientes con características similares, se utilizaron técnicas de clustering, apoyadas en la reducción de dimensionalidad:

1. **Clustering Basado en PCA**: Los datos reducidos mediante PCA fueron utilizados para realizar clustering, identificando patrones en los grupos de pacientes que podrían no ser evidentes en el espacio original de alta dimensionalidad.
   Gráfico relevante: Gráfico de clustering basado en PCA.
2. **Clustering Basado en t-SNE y UMAP**: Además de PCA, t-SNE y UMAP se utilizaron para explorar cómo se agrupan los pacientes en un espacio reducido. Estos métodos permiten una mejor visualización de las relaciones locales entre los pacientes.
   Gráfico relevante: Gráficos de clustering basados en t-SNE y UMAP.

VII. DISEÑO VISUAL

El diseño visual se centrará en la claridad y la simplicidad, utilizando una combinación de gráficos estáticos e interactivos para comunicar los hallazgos de manera efectiva. Las visualizaciones se organizarán en torno a:

- *Distribuciones de variables:* Histogramas, boxplots.
- *Correlaciones:* Heatmaps y scatter plots para mostrar relaciones entre variables.
- *Modelos predictivos:* Visualización de la precisión y otras métricas de rendimiento de los modelos de machine learning.

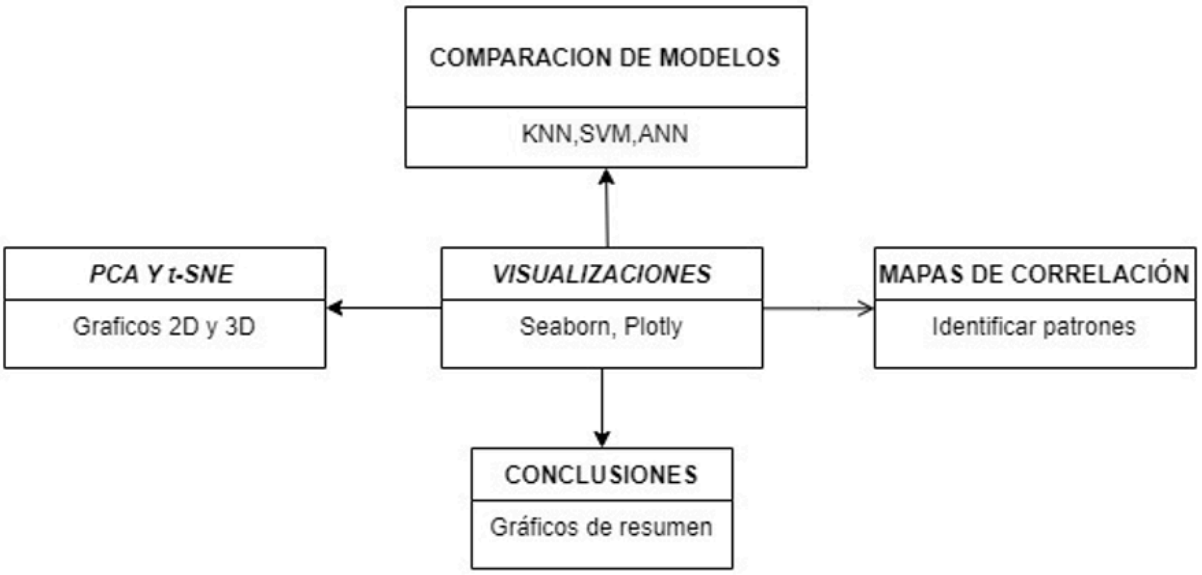

Figura 13, Diagrama de clases para el diseño visual, Fuente, Elaboración propia.

VIII. RESULTADOS: CASOS DE ESTUDIO

En esta sección, se presentan los resultados obtenidos a partir del análisis del dataset, incluyendo:

- *Casos de estudio de individuos con y sin diabetes:* Comparación de sus características principales.
- *Evaluación de modelos:* Rendimiento de los modelos predictivos propuestos en términos de precisión, recall, F1-score, y AUC-ROC.

IX. DISCUSIÓN

Los resultados obtenidos serán discutidos en relación con los trabajos previos y las hipótesis planteadas. Se explorará cómo las características identificadas influyen en la presencia de diabetes y se analizarán las limitaciones del estudio, así como las posibles mejoras en futuros análisis.

X. CONCLUSIONES

Identificación de Factores Clave en la Predicción de Diabetes: A través del análisis exploratorio de datos (EDA), se identificaron variables críticas como los niveles de glucosa y el índice de masa corporal (BMI) que muestran una fuerte correlación con la presencia de diabetes. Esto destaca la relevancia de dichas variables para la predicción y diagnóstico temprano.

Eficacia de la Imputación de Datos Faltantes: La imputación de valores faltantes utilizando la mediana en variables como Glucose, BloodPressure, y BMI permitió mejorar la calidad del dataset, facilitando análisis más precisos y la generación de visualizaciones confiables que capturan los patrones significativos del conjunto de datos.

Comparación de Modelos de Machine Learning: Los modelos de machine learning evaluados, incluyendo KNN, SVM y ANN, mostraron diferencias en su rendimiento dependiendo de la métrica utilizada (distancia euclidiana, Manhattan, coseno) y la dimensionalidad del dataset. El uso de técnicas de reducción de dimensionalidad como PCA permitió una mejor interpretación y, en algunos casos, una mejora en la precisión de los modelos.

Importancia de la Selección de Técnicas y Modelos: Los resultados subrayan la importancia de seleccionar cuidadosamente tanto las técnicas de preprocesamiento de datos como los modelos de machine learning en función de la naturaleza del problema y las características específicas del dataset. Esta selección influye significativamente en la robustez y precisión de las predicciones.

Aplicabilidad Clínica y Futura Investigación: Este estudio demuestra cómo la combinación del EDA y los modelos de machine learning puede servir como una herramienta valiosa en la identificación temprana de diabetes, ofreciendo una base sólida para futuras investigaciones y potenciales aplicaciones clínicas en el diagnóstico y monitoreo de pacientes.

REFERENCIAS

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Springer.

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. International Conference on Database Theory, Springer.

American Diabetes Association. (2018). Classification and diagnosis of diabetes: Standards of Medical Care in Diabetes—2018. Diabetes Care, 41(Supplement 1), S13-S27. https://doi.org/10.2337/dc18-S002

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care. IEEE Computer Society Press.

Tukey, J. W. (1977). Exploratory Data Analysis. Addison-Wesley.

Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Wiley.