Informe del Proyecto: Análisis de Diabetes

Estudiante

Juan Daniel Díaz Pinzón

Programa

Ciencia de Datos

Institución

Fundación Universitaria Compensar

Asignatura

Programación para ciencia de Datos II

Docente

Sebastián Rodríguez Muñoz

Bogotá D.C 29/09/2024

Introducción

El presente informe documenta el análisis de un conjunto de datos relacionados con la diabetes, con el objetivo de identificar patrones y relaciones entre diversas variables clínicas. El proyecto se centra en la predicción de los niveles de insulina y glucosa en sangre, utilizando técnicas estadísticas y modelos de aprendizaje automático. Se implementó un dashboard interactivo con **Dash** para visualizar los resultados de este análisis.

Plataforma y Herramientas Utilizadas

Plataforma: Google Colab, que permite la ejecución de código Python en un entorno en línea, facilitando la colaboración y el acceso a bibliotecas de datos.

Lenguaje de Programación: Python.

Bibliotecas:

Pandas: Para manipulación y análisis de datos.

NumPy: Para operaciones matemáticas y manejo de arrays.

Plotly y Dash: Para la creación de visualizaciones interactivas y dashboards.

Scikit-Learn: Para la implementación de modelos de regresión lineal y logística.

Matplotlib y Seaborn: Para visualizaciones estáticas.

Descripción del Conjunto de Datos

El conjunto de datos se obtuvo de la plataforma Kaggle. Este conjunto contiene **70,000 entradas y 34 columnas** que incluyen tanto variables cuantitativas como cualitativas, tales como:

Variables Cuantitativas:

Niveles de Insulina
Edad
Índice de Masa Corporal (BMI)
Presión Arterial
Niveles de Colesterol
Niveles de Glucosa en Sangre

Variables Cualitativas:

Marcadores Genéticos
Historia Familiar
Hábitos Dietéticos

Análisis Descriptivo

Se realizó un análisis descriptivo de las variables para entender su distribución. Las estadísticas descriptivas revelaron información sobre la media, mediana y desviación estándar de las variables cuantitativas, proporcionando una visión general del estado de salud de los pacientes en el conjunto de datos.

Visualizaciones

Se generaron gráficos como diagramas de dispersión y boxplots para explorar visualmente las relaciones entre variables, revelando que los niveles de glucosa tienden a aumentar con la edad y el IMC.

Transformaciones de Variables

Para mejorar la capacidad predictiva de los modelos, se llevaron a cabo varias transformaciones de variables, incluyendo:

Interacciones: Creación de nuevas variables como Age_BMI_Interaction.
Transformaciones Logarítmicas: Se aplicó una transformación logarítmica a los niveles de insulina para manejar la asimetría de los datos.

Pruebas de Hipótesis

Se realizaron pruebas t para evaluar si existen diferencias significativas en los niveles de glucosa entre diferentes grupos de edad:

Hipótesis: No hay diferencias significativas en los niveles de glucosa entre los grupos de edad.
Resultados: Se obtuvo una estadística t de -45.87 y un p-valor de 0.0, indicando diferencias significativas.

Modelos Predictivos

Regresión Logística

Objetivo: Predecir la probabilidad de que un paciente tenga niveles elevados de insulina.
Variables: Se utilizaron Log_Insulin_Levels, Age, y BMI.
Resultados:
Precisión: 0.98

La regresión logística se eligió debido a su capacidad para manejar variables dependientes categóricas, en este caso, para clasificar si los niveles de insulina son altos o no.

Regresión Lineal

Se implementó una regresión lineal para modelar la relación entre Blood Glucose Levels y otras variables independientes como Insulin Levels, Age, y BMI. Aunque inicialmente se usó para predecir los niveles de glucosa, se decidió no incluir esta parte en el dashboard por problemas de rendimiento.

Resultados Preliminares:
Error Cuadrático Medio (MSE): 64.02
R^2 Score: 0.45

La regresión lineal se utilizó para predecir una variable continua (niveles de glucosa) basándose en otras variables independientes.

Conclusiones Generales

Se concluye que las transformaciones y el modelado adecuado mejoran la capacidad predictiva del análisis.

Las pruebas de hipótesis ofrecieron información valiosa sobre diferencias significativas entre grupos.

Se identificaron variables relevantes que impactan los niveles de insulina y glucosa, lo que puede ser útil para futuras investigaciones y recomendaciones en salud.

Recomendaciones

Continuar recopilando datos adicionales para reforzar los modelos.

Considerar técnicas de regularización en los modelos de regresión para evitar el sobreajuste.

Explorar la inclusión de nuevas variables que puedan contribuir al análisis.

Bibliografía

scikit-learn. (2023, August 29). LinearRegression. https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

scikit-learn. (2023, August 29). LogisticRegression. https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

SciPy. (2023, August 29). Hypothesis tests and related functions.

https://docs.scipy.org/doc/scipy/reference/stats.html#hypothesis-tests-and-relatedfunctions

Quarto. (2023, August 29). Getting Started with Jupyter. Quarto Documentation. https://quarto.org/docs/get-started/computations/jupyter.html

statsmodels. (2023, August 29). ttest_ind Documentation.

https://www.statsmodels.org/