Pima Indians Diabetic Dataset

Mireya Sánchez Pinzón

CES Juan Pablo II

Índice

CONTEXTO DEL DATASET

CONTENIDO DEL DATASET

PLANTEAMIENTO DEL PROBLEMA

PROPUESTA DE SOLUCIÓN

1. Contexto del dataset

01

Parte de un estudio del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales.

02

Se estudian pacientes son mujeres de al menos 21 años de edad de ascendencia india Pima.

03

Particularmente significativo debido al alto riesgo de diabetes dentro de la población de los indios Pima.

2. Contenido del dataset

Contiene 768 observaciones

Compuesto de 8 variables de entrada y 1 variable de salida

No es un conjunto de datos equilibrado y los valores faltantes se reemplazan por 0

2. Contenido del dataset (variables)

Pregnancies: número de embarazos.

Glucose: concentración a las 2 horas en una prueba de tolerancia oral a la glucosa.

BloodPressure: presión arterial diastólica (mm Hg).

SkinThickness: grosor del pliegue cutáneo del tríceps (mm).

Insulin: insulina sérica de 2 horas (mu U/ml).

BMI: índice de masa corporal (peso en kg/(altura en m^2)).

DiabetesPedigreeFunction: determina el riesgo de diabetes tipo 2 según los antecedentes familiares, cuanto mayor es la función, mayor es el riesgo.

Age: edad del paciente.

Outcome: variable objetivo. Indica si el paciente es diabético o no.

3. Planteamiento del problema

El objetivo es predecir si un paciente tiene diabetes o no, basándose en determinadas mediciones diagnósticas.

4. Propuesta de solución
Preprocesamiento de Datos

1. **Limpieza**: Eliminar valores nulos o inconsistentes.
2. **Normalización**: Estandarizar las variables para que tengan la misma escala.
3. **Detección de Outliers**: Identificar y eliminar valores atípicos que puedan afectar el modelo.
4. **Análisis de Correlación**: Evaluar la relación entre las variables para identificar posibles colinealidades.
5. **Selección de Características**: Elegir las variables más relevantes para la predicción.

4. Propuesta de solución
Modelos de Aprendizaje Automático

Regresión logística:

Un modelo clásico para problemas de clasificación binaria.

Redes neuronales:

Ofrecen mayor flexibilidad y capacidad de aprendizaje.

Árboles de decisión:

Permiten una mejor interpretación del modelo.

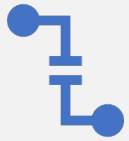
4. Propuesta de solución
Evaluación del Modelo

Validación cruzada: Dividir el conjunto de datos en subconjuntos para entrenar y evaluar el modelo.

Métricas de evaluación: Utilizar métricas como precisión, sensibilidad, especificidad y AUC para medir el rendimiento del modelo.

Comparación de modelos: Evaluar diferentes modelos y seleccionar el que tenga mejor rendimiento.

4. Propuesta de solución
Optimización del Modelo



Ajuste de hiperparámetros: Ajustar los parámetros del modelo para obtener un mejor rendimiento.

Comparar diferentes modelos para obtener los mejores resultados.

Pima Indians Diabetic Dataset

Mireya Sánchez Pinzón

CES Juan Pablo II