UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS
MAESTRÍA EN CIENCIA DE DATOS

APRENDIZAJE AUTOMÁTICO

Profesor: Jose Anastacio Hernandez Saldaña

Tarea 3: Modelo de Clasificación

Dora Alicia Guevara Villalpando
Matrícula: 1551003
Grupo: 003

21 de julio de 2024

UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Tarea 3: Modelo de Clasificación

Introducción.

El presente documento presenta los resultados obtenidos durante el proceso de realizar la Tarea 3 de Aprendizaje Automático. Esta tarea consiste en realizar el modelo de clasificación a una base de datos de nuestro interés.

Contexto.

Este conjunto de datos contiene detalles médicos de los pacientes, incluidas características como el nivel de glucosa, la presión arterial, el nivel de insulina, el IMC, la edad y más. La variable objetivo indica si un paciente tiene diabetes. El objetivo de este conjunto de datos es crear y evaluar varios modelos de aprendizaje automático o aprendizaje profundo para predecir la aparición de la diabetes.

Este archivo contiene los registros médicos de los pacientes, que incluyen diversas métricas relacionadas con la salud. El objetivo es utilizar estas características para predecir si un paciente tiene diabetes. A continuación, se incluye una descripción detallada de cada columna del conjunto de datos:

- **Embarazos**: Número de veces que la paciente ha estado embarazada.
- **Glucosa**: Concentración de glucosa plasmática a las 2 horas en una prueba de tolerancia a la glucosa oral.
- **Presión arterial**: Presión arterial diastólica (mm Hg).
- **Grosor de la piel:** Grosor del pliegue cutáneo del tríceps (mm).
- **Insulina**: Insulina sérica a las 2 horas (mu U/ml).
- **IMC**: Índice de masa corporal (peso en kg/(altura en m)^2).
- **DiabetesPedigreeFunction**: Función que puntúa la probabilidad de diabetes en función de los antecedentes familiares.
- **Edad**: Edad de la paciente (años).
- **Resultado**: Variable de clase (0 o 1), donde 1 representa la presencia de diabetes y 0 representa la ausencia de diabetes.

Desarrollo.

La base de datos cuenta con 768 filas y 9 columnas:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Árbol de decisión.

Los **árboles de decisión** son un modelo de aprendizaje automático supervisado que se utiliza tanto para la clasificación como para la regresión. Son ampliamente extendidos debido a su simplicidad, facilidad de interpretación y versatilidad en diversas aplicaciones.

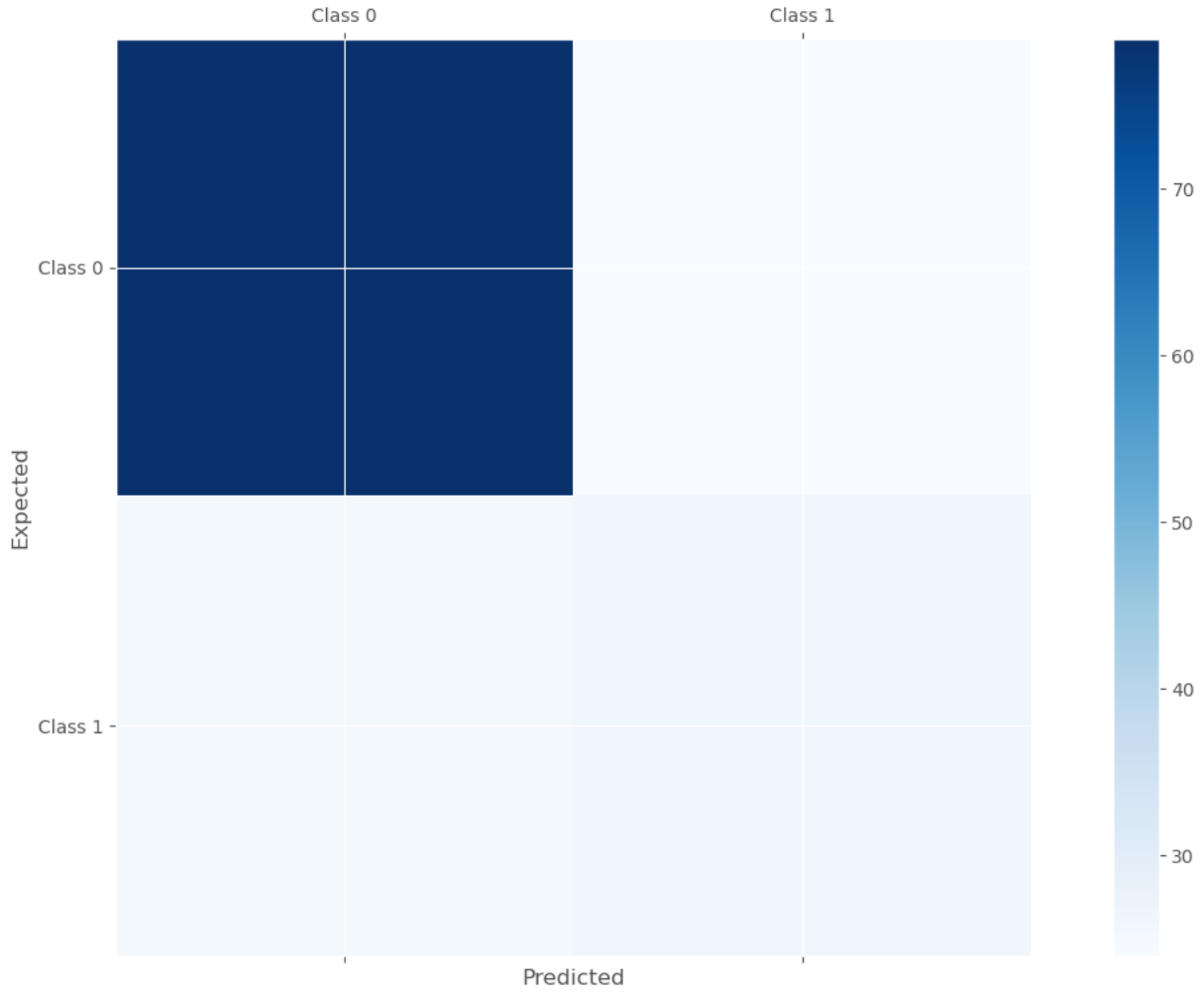
Los árboles de decisión aprenden de los datos generando reglas de tipo if-else y divisiones conocidas como nodos. Cada nodo representa una pregunta sobre los datos y cada rama del árbol representa una respuesta a esa pregunta. El proceso continúa hasta que se llega a una hoja del árbol, que representa la predicción final.

Dividimos la base de datos con el 80% para *Train* y el 20% para *Test*. Se obtuvieron los siguientes resultados:

```
Matriz de Confusión - DATOS ORIGINALES:
 [[79 24]
 [25 26]]
Métricas de Matriz de Confusión - DATOS ORIGINALES:
              precision    recall  f1-score   support

           0       0.76      0.77      0.76       103
           1       0.52      0.51      0.51        51

    accuracy                           0.68       154
   macro avg       0.64      0.64      0.64       154
weighted avg       0.68      0.68      0.68       154
```



Se tiene un accuracy del 68%.

Obtuvimos una tasa de clasificación del casi 70 %, lo que se considera una buena precisión.

Visualización del árbol de decisión