INSTITUTO POLITÉCNICO NACIONAL

**ESCUELA SUPERIOR DE CÓMPUTO**

Minería de Datos

Practica  
ARBOL DE CLASIFICACIÓN

**Integrantes:**

* **Sánchez Zarazúa Jesua Antonio**

**Profesora: Ocampo Botello Fabiola**

**Grupo: 3CM9**

**Fecha de entrega: 03/04/2020**

Practica. Arboles de Clasificación

INTRODUCCION

Los árboles de decisión son un de la familia de modelos de aprendizaje automático más utilizados. Se pueden utilizar tanto para resolver problemas de clasificación como de regresión. Una de sus principales ventajas es la facilidad con la que se puede interpretar los resultados en base a reglas. Permitiendo no solo obtener un resultado, sino que inspeccionar los motivos por los que se llega a una predicción dada. Por ejemplo, en un modelo que permita predecir la aparición de fallos en una maquinaria se puede explorar el proceso lógico del algoritmo, identificando los valores de las características que llevar a una conclusión dada. Esto permite nos solo utilizar el modelo para predecir un valor, sino actuar sobre las causas para evitar la aparición de resultados no deseados. Siendo la visualización de árboles de decisión una forma de facilitar esto.

Esta practica se realizo usando el lenguaje de programación Python. Además de las librerías para machine learning Scikit Learn y otras librerias especificas como Pandas para la lectura del dataset .CSV y Graphviz para construir una imagen del modelo del árbol resultado del aprendizaje

También se ha de recalcar que se uso Jupyter Notebook como herramienta de desarrollo, por lo que el resultado es un Notebook que posee cada línea de código y resultado obtenido, visible sin la necesidad de compilar el archivo.

INTENCION RECOPILACION DE DATOS

Este conjunto de datos es originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es predecir el diagnóstico sobre si un paciente tiene diabetes o no, basándose en ciertas mediciones de diagnóstico incluidas en el conjunto de datos. Se colocaron varias restricciones en el proceso de selección de las instancias que formaban parte de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de herencia indígena Pima.

Fuente:

<https://www.kaggle.com/uciml/pima-indians-diabetes-database>

Kaggle Pima Indians Diabetes

Diccionario de Datos

Captura de pantalla de un celular

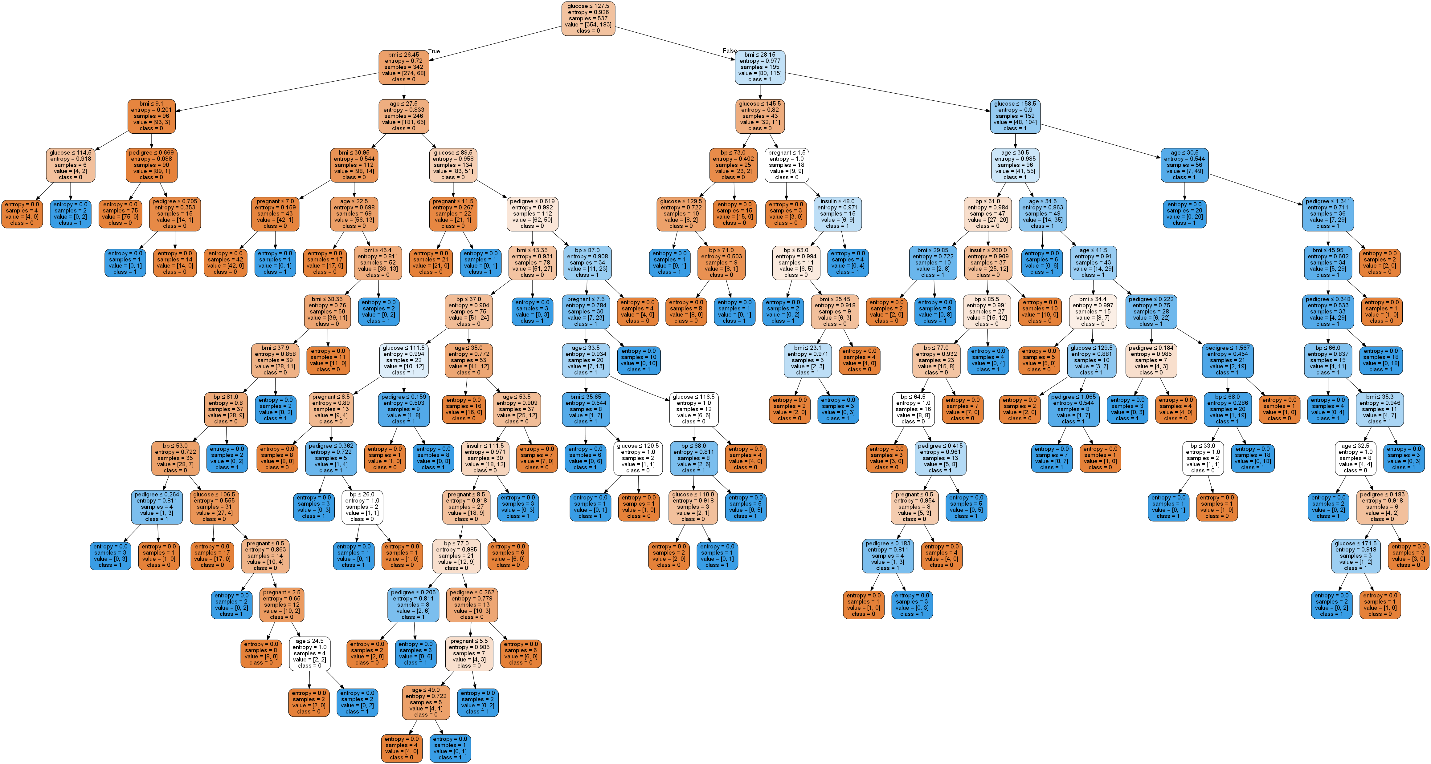
Descripción generada automáticamente

* Numero de Datos: 768 filas
* Numero de Columnas: 9 Columnas
* Tipos de Datos contenidos: Numericos(int)

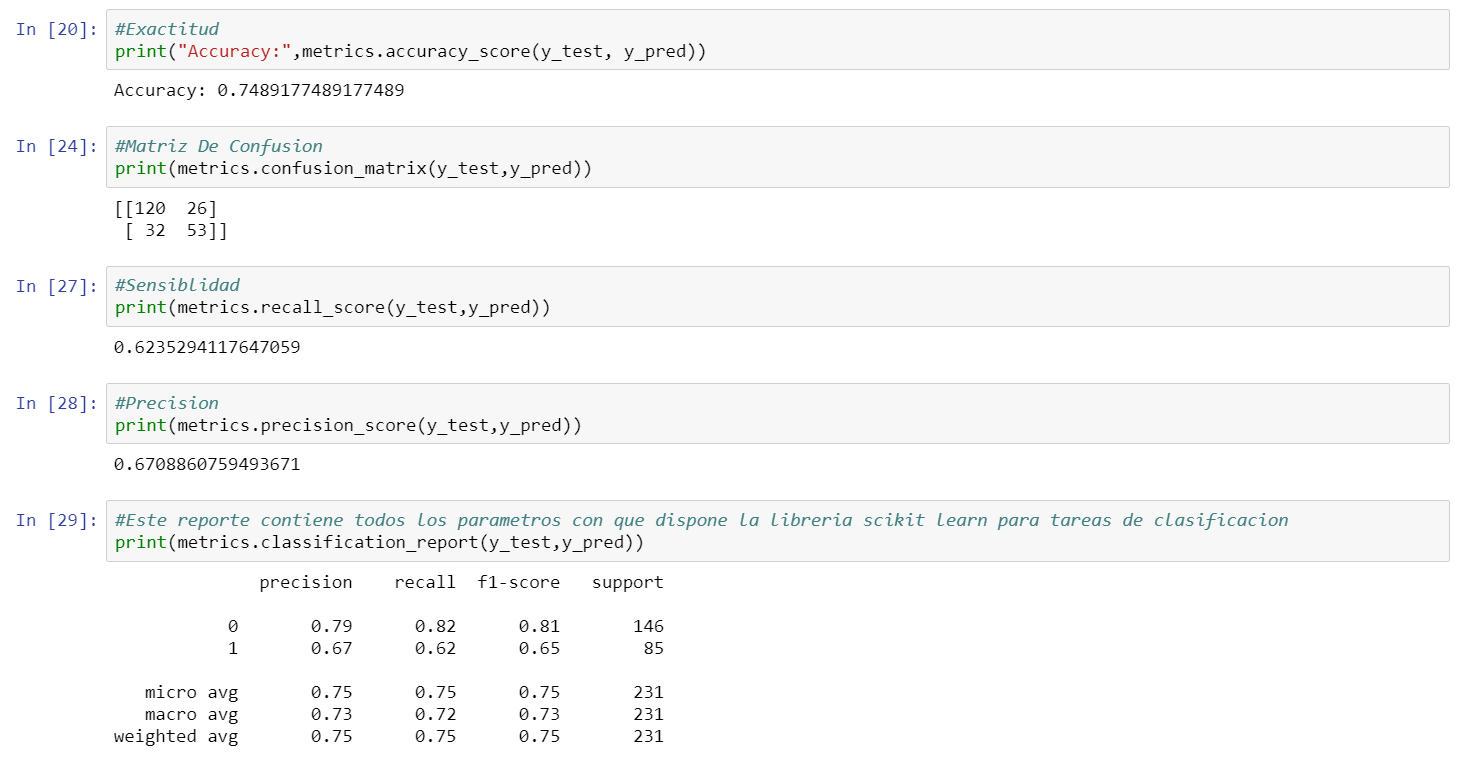
OBJETIVO DE ANALISIS DE DATOS

Se entiende que los valores contenidos son indispensables para clasificar casos futuros como pasientes que padecen de diabetes o no, todos las columnas son descritas con tipos de datos numericos y no se encuentran datos no valios dentro del dataset por lo que se puede aceptar el uso de estos en su estado actual para el aprenizaje del arbol, la unica tarea es dividir los datos en dos cunjuntos para el entrenamiento y para el prueba.

WORKFLOW ARBOL



**MEDIDAS**



**ANEXO: Cuaderno Completo de Jupyter Notebook**

Captura de pantalla de un celular

Descripción generada automáticamente