UNIVERSIDAD BERNARDO O'HIGGINS
FACULTAD DE INGENIERÍA, CIENCIAS Y TECNOLOGÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

Magíster en Ingeniería Informática

Metodología de la Investigación en Informática

Construcción Modelo de Clasificación, usando Pyspark y MLlib

Profesor: Sebastián Ulloa
Estudiante: Carlos Alfredo Castillo Rodríguez
Enero - 2023

Introducción

En este informe se presenta un script de Python que se utiliza para preparar y analizar un conjunto de datos de un archivo CSV utilizando Apache Spark. El objetivo del script es entrenar y evaluar un modelo de clasificación utilizando el conjunto de datos procesado.

Para lograr este objetivo, el script realiza varias tareas, como descargar e instalar Apache Spark y Java Development Kit (JDK), integrarse con Google Drive para tener acceso al archivo CSV, inicializar Spark y crear una sesión de Spark, leer el archivo CSV en un DataFrame de Spark, explorar y analizar el DataFrame, seleccionar un subconjunto de columnas del DataFrame original y codificar las columnas categóricas como valores numéricos, crear un VectorAssembler de Spark ML para combinar todas las características en un vector único, dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, y finalmente entrenar y evaluar un modelo de clasificación utilizando el conjunto de entrenamiento y el conjunto de prueba.

En las siguientes secciones se describe en detalle cada uno de los pasos realizados por el script y se presentan los resultados obtenidos.

Desarrollo

Este script realiza varias tareas en Python para preparar y analizar un conjunto de datos de un archivo CSV. Algunos de los pasos clave que realiza el script incluyen:

Descargar e instala Apache Spark y Java Development Kit (JDK).

Integrar el script con Google Drive para tener acceso al archivo CSV.

Inicializar Spark y crear una sesión de Spark.

Leer el archivo CSV en un DataFrame de Spark.

Explorar y analiza el DataFrame, mostrando algunos de los primeros registros y verificando el esquema del DataFrame.

Utilizar la librería de visualización Pandas para dibujar un diagrama de dispersión múltiple para ver las relaciones entre las variables numéricas del conjunto de datos.

Seleccionar un subconjunto de columnas del DataFrame original y define algunas de ellas como categóricas.

Utilizar un StringIndexer de Spark ML para codificar las columnas categóricas como valores numéricos. Crea un VectorAssembler de Spark ML para combinar todas las características en un vector único.

Dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba. Entrena un modelo de clasificación utilizando el conjunto de entrenamiento.

Evaluar el modelo entrenado utilizando el conjunto de prueba y muestra algunas métricas de rendimiento.

Conclusiones

El script de Python que se presenta en este informe es capaz de preparar y analizar un conjunto de datos de un archivo CSV utilizando Apache Spark.

Se han realizado varias tareas para lograr este objetivo, tales como descargar e instalar Apache Spark y JDK, integrarse con Google Drive, leer el archivo CSV en un DataFrame de Spark, explorar y analizar el DataFrame, seleccionar un subconjunto de columnas, codificar las columnas categóricas como valores numéricos, crear un VectorAssembler de Spark ML, dividir el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba, y entrenar y evaluar un modelo de clasificación.