Assignment 4

Introduction

BAM-1043-01 Assignment
C#: c0932089_Assignment4
Name: Prabesh Rai
Date: 5 April, 2024

Question:
Task Description – Using the Databricks community edition notebook and Apache Spark, calculate the Correlation, and explain the relationship between different variables;

Please write a detailed description of the following along with the screenshots from your notebook;

1. Data can be accessed from the following URL:
https://raw.githubusercontent.com/selva86/datasets/master/Iris.csv.
2. Calculate the correlation Using DataFrame API.
3. Calculate the correlation matrix Using RDD.
4. Correlation heat map using Correlation matrix.

I have downloaded the data in a local file store: /FileStore/tables/Iris_dataset.csv

Cmd 1

Iris Dataset Location

```
1  /FileStore/tables/Iris_dataset.csv
```

Fig: a screenshot of file location

Cmd 2

Initializing Spark session and Loading the Data

```
1  from pyspark.sql import SparkSession
2
3  # Initialize Spark session
4  spark = SparkSession.builder \
5      .appName("Iris Correlation Analysis") \
6      .getOrCreate()
7
8  # Read data from URL into DataFrame
9  file_path = "/FileStore/tables/Iris_dataset.csv"
10 iris_df = spark.read.csv(file_path, header=True, inferSchema=True)
11
12 # Show the DataFrame schema and first few rows
13 iris_df.printSchema()
14 display(iris_df)
```

```
root
 |-- Id: integer (nullable = true)
 |-- SepalLengthCm: double (nullable = true)
 |-- SepalWidthCm: double (nullable = true)
 |-- PetalLengthCm: double (nullable = true)
 |-- PetalWidthCm: double (nullable = true)
 |-- Species: string (nullable = true)
```

Table

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	1	5.1	3.5	1.4	0.2	Iris-setosa
2	2	4.9	3	1.4	0.2	Iris-setosa
3	3	4.7	3.2	1.3	0.2	Iris-setosa
4	4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	5	3.6	1.4	0.2	Iris-setosa
6	6	5.4	3.9	1.7	0.4	Iris-setosa
7	7	4.6	3.4	1.4	0.3	Iris-setosa

150 rows | 2.57 seconds runtime

Fig: a screenshot of loading the data for calculation

2 Ans:

Cmd 3

Calculating Correlation Using DataFrame API

```
1  from pyspark.sql.functions import corr
2
3  # Calculate correlation using DataFrame API
4  correlation = iris_df.corr("SepalLengthCm", "PetalLengthCm")
5
6  print("Correlation between SepalLengthCm and PetalLengthCm :", correlation)
7  display(correlation)
```

▸ (2) Spark Jobs

```
Correlation between SepalLengthCm and PetalLengthCm : 0.8717541573048717
0.8717541573048717
```

Command took 0.62 seconds -- by raiprabesh789@gmail.com at 4/5/2024, 2:51:42 PM on My Cluster

Fig: a screenshot of calculating correlation Using Datframe API

The correlation analysis using DataFrame API shows a strong positive relationship between the SepalLengthCm and PetalLengthCm variables in the Iris dataset. With a correlation coefficient of approximately 0.87, we observe a robust linear association between these two attributes. This correlation coefficient indicates that as the length of the sepals increases, there is a corresponding increase in the size of the petals, and vice versa. The scatter plot visualization further reinforces this finding, as the data points cluster around a positively sloped line. This strong positive correlation implies that changes in SepalLengthCm are closely related to changes in PetalLengthCm, providing valuable insights into the interdependence of these two characteristics within the Iris dataset.

3 Ans:

Cmd 4

Calculating Correlation matrix Using Resilient Distributed Dataset (RDD)

```
1  from pyspark.mllib.stat import Statistics
2  from pyspark.mllib.linalg import Vectors
3  import pandas as pd
4
5  # Select only numeric columns
6  columns = ["SepalLengthCm",  "SepalWidthCm", "PetalLengthCm", "PetalWidthCm"]
7  data = iris_df.select(columns)
8
9  # Convert the DataFrame into an RDD of Vectors
10 rdd_vectors = data.rdd.map(lambda row: Vectors.dense(row))
11
12 # Calculate the Pearson correlation matrix using the RDD of Vectors
13 correlation_matrix = Statistics.corr(rdd_vectors, method="pearson")
14
15 correlation_df = pd.DataFrame(correlation_matrix, columns=columns, index=columns)
16 print("Correlation matrix:")
17 print(correlation_df)
18 display(correlation_df)
```

▸ (4) Spark Jobs

▸ data: pyspark.sql.dataframe.DataFrame = [SepalLengthCm: double, SepalWidthCm: double ... 2 more fields]

```
Correlation matrix:
               SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm
SepalLengthCm       1.000000     -0.109369       0.871754      0.817954
SepalWidthCm       -0.109369      1.000000      -0.420516     -0.356544
PetalLengthCm       0.871754     -0.420516       1.000000      0.962757
PetalWidthCm        0.817954     -0.356544       0.962757      1.000000
```

Table +

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
1	1	-0.10936924995064932	0.8717541573048723	0.8179536333691638
2	-0.10936924995064932	1	-0.4205160964011551	-0.3565440896138061
3	0.8717541573048723	-0.4205160964011551	1	0.9627570970509673
4	0.8179536333691638	-0.3565440896138061	0.9627570970509673	1

Fig: a screenshot of calculating correlation matrix Using RDD and its result

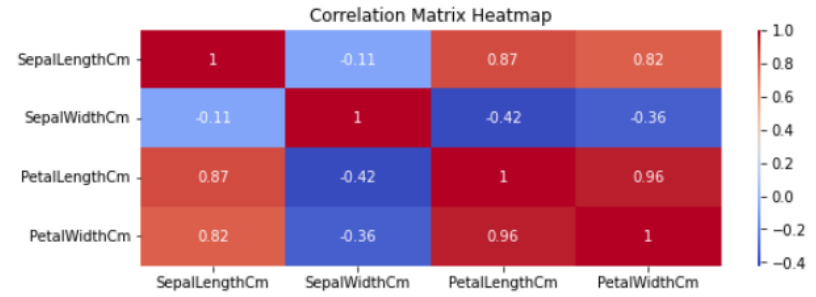
From the above the correlation matrix presents several key relationships among the variables in the Iris dataset. Firstly, SepalLengthCm exhibits a strong positive correlation with PetalLengthCm (correlation coefficient = 0.871754) and PetalWidthCm (correlation coefficient = 0.817954). This shows that as the length of sepals increases, there is a corresponding increase in the length and width of petals. Conversely, SepalWidthCm shows a weak negative correlation with both SepalLengthCm (correlation coefficient = -0.109369) and PetalLengthCm (correlation coefficient = -0.420516), implying a slight decrease in sepal width as sepal and petal lengths increase. Secondly, SepalWidthCm and PetalWidthCm demonstrate a moderate negative correlation (correlation coefficient = -0.356544), indicating that as the width of sepals increases, the width of petals tends to decrease. Finally, the correlation between PetalLengthCm and PetalWidthCm is notably strong (correlation coefficient = 0.962757), highlighting a robust positive relationship where an increase in petal length corresponds closely with an increase in petal width. These correlation insights offer a valuable understanding of the interdependencies and associations between different attributes of the Iris dataset.

4 Ans:

The heatmap displays the correlation coefficients between pairs of variables: SepalLengthCm, SepalWidthCm, PetalLengthCm, and PetalWidthCm.

Correlation Heat Map Using Correlation Matrix

```
1   import seaborn as sns
2   import matplotlib.pyplot as plt
3   import pandas as pd
4
5   # Convert the correlation matrix to a Pandas DataFrame
6   correlation_df = pd.DataFrame(correlation_matrix, columns=columns, index=columns)
7
8   # Create the heatmap using Seaborn
9   plt.figure(figsize=(9, 3))
10  sns.heatmap(correlation_df, annot=True, cmap="coolwarm", cbar_kws={"aspect": 60})
11  plt.title("Correlation Matrix Heatmap")
12  plt.show()
```



Command took 0.47 seconds -- by raiprabesh789@gmail.com at 4/5/2024, 2:56:08 PM on My Cluster

Fig: a screenshot of calculating correlation Heat Map Using Correlation Matrix and its result

From the above figure, correlation values range from -0.4 to 1, with red indicating positive correlations and blue indicating negative correlations.