CAT 2
19MAM67
BIG DATA ANALYTICS LAB

TEAM MEMBERS:
Chezhian T N (1934005)
Varsha B(1934055)

Dataset: IRIS dataset - Classification of flowers species
Dataset link: https://www.kaggle.com/datasets/uciml/iris

Database Connection:

```
import findspark
findspark.init()

import pyspark # only run after findspark.init()
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

dataframe_mysql = spark.read.format("jdbc").options(
    url="jdbc:mysql://localhost:3306/irisdatabase",
    driver = "com.mysql.jdbc.Driver",
    dbtable = "iris",
    user="root",
    password="1234").load()
```

Importing Libraries:

```
import findspark
import pyspark
```

```
from pyspark.ml.feature import VectorAssembler
# from pyspark.ml.classification import LogisticRegression,RandomForestClassifier, DecisionTreeClassifier
import pandas as pd
from sklearn.metrics import classification_report, confusion_matrix
```

Pre-Processing:

```
from pyspark.ml.feature import StringIndexer

indexer = StringIndexer(inputCol=dataframe_mysql.columns[-1],outputCol="Species_num").fit(dataframe_mysql)
```

```
index_df = indexer.transform(dataframe_mysql)
```

```
index_df.show(5)
```

```
+-------------+------------+-------------+------------+-----------+-----------+
|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|Species_num|
+-------------+------------+-------------+------------+-----------+-----------+
|          5.1|         3.5|          1.4|         0.2|Iris-setosa|        0.0|
|          4.9|         3.0|          1.4|         0.2|Iris-setosa|        0.0|
|          4.7|         3.2|          1.3|         0.2|Iris-setosa|        0.0|
|          4.6|         3.1|          1.5|         0.2|Iris-setosa|        0.0|
|          5.0|         3.6|          1.4|         0.2|Iris-setosa|        0.0|
+-------------+------------+-------------+------------+-----------+-----------+
only showing top 5 rows
```

```
index_df = index_df.drop("Species")
```

```
assembler = VectorAssembler(inputCols=index_df.columns[:-1],
                            outputCol='features',
                            handleInvalid='skip')
index_df= assembler.transform(index_df)
index_df.show(5)
```

```
+-------------+------------+-------------+------------+-----------+-----------------+
|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|Species_num|         features|
+-------------+------------+-------------+------------+-----------+-----------------+
|          5.1|         3.5|          1.4|         0.2|        0.0|[5.1,3.5,1.4,0.2]|
|          4.9|         3.0|          1.4|         0.2|        0.0|[4.9,3.0,1.4,0.2]|
|          4.7|         3.2|          1.3|         0.2|        0.0|[4.7,3.2,1.3,0.2]|
|          4.6|         3.1|          1.5|         0.2|        0.0|[4.6,3.1,1.5,0.2]|
|          5.0|         3.6|          1.4|         0.2|        0.0|[5.0,3.6,1.4,0.2]|
+-------------+------------+-------------+------------+-----------+-----------------+
only showing top 5 rows
```

```
index_df.columns
```

```
['SepalLengthCm',
 'SepalWidthCm',
 'PetalLengthCm',
 'PetalWidthCm',
 'Species_num',
 'features']
```

Testing Training:

```
train_df , test_df = index_df.select('features','Species_num').randomSplit([0.8,0.2], seed = 123)
```

Logistic Regression and Prediction:

```
lr = LogisticRegression(featuresCol = 'features', labelCol ='Species_num' )
lr_model = lr.fit(train_df)
```

```
predictionslr = lr_model.transform(test_df)
```

```
predictionslr.show(5)
```

```
+-----------------+-----------+--------------------+--------------------+----------+
|         features|Species_num|       rawPrediction|         probability|prediction|
+-----------------+-----------+--------------------+--------------------+----------+
|[4.4,3.0,1.3,0.2]|        0.0|[1061.07123699051...|[1.0,1.3970847698...|       0.0|
|[4.6,3.2,1.4,0.2]|        0.0|[1056.64773662478...|[1.0,8.6216195032...|       0.0|
|[4.8,3.0,1.4,0.3]|        0.0|[922.338833462604...|[1.0,1.1302410426...|       0.0|
|[4.8,3.1,1.6,0.2]|        0.0|[952.505287430215...|[1.0,5.1675450394...|       0.0|
|[4.9,3.0,1.4,0.2]|        0.0|[938.768278594419...|[1.0,6.8824386417...|       0.0|
+-----------------+-----------+--------------------+--------------------+----------+
only showing top 5 rows
```

Model Evaluation:

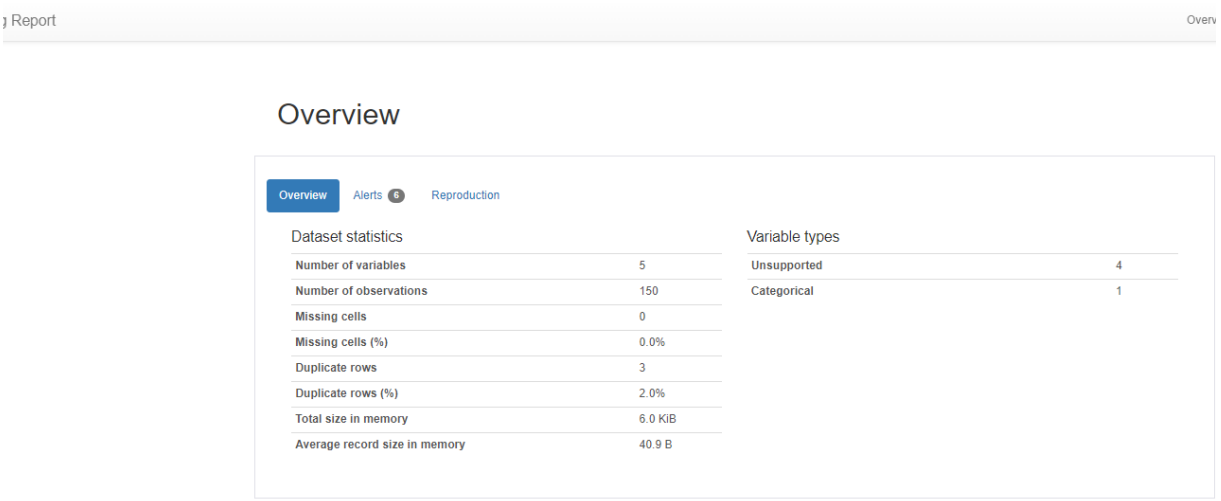
```
result = predictionslr.toPandas()
true_labels=(test_df.select("Species_num")).toPandas()
predicted_labels=result["prediction"]

print("-- Logistic Regression --")
print("------------------------------------------------------------------------")
print("Classification Report\n",classification_report(true_labels, predicted_labels))
print("------------------------------------------------------------------------")
print("Confusion matrix\n",confusion_matrix(true_labels,predicted_labels),"\n\n")
LR=confusion_matrix(true_labels,predicted_labels)
```

```
-- Logistic Regression --
------------------------------------------------------------------------
Classification Report
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        13
         1.0       1.00      0.86      0.92         7
         2.0       0.90      1.00      0.95         9

    accuracy                           0.97        29
   macro avg       0.97      0.95      0.96        29
weighted avg       0.97      0.97      0.97        29

------------------------------------------------------------------------
Confusion matrix
 [[13  0  0]
 [ 0  6  1]
 [ 0  0  9]]
```

Visualisation:

```
from pandas_profiling import ProfileReport
report = ProfileReport(dataframe_mysql.toPandas())
report
```

```
Summarize dataset: 100%|██████████| 18/18 [00:00<00:00, 37.27it/s, Completed]
Generate report structure: 100%|██████████| 1/1 [00:00<00:00,  3.12it/s]
Render HTML: 100%|██████████| 1/1 [00:00<00:00,  9.62it/s]
```



Variables

~~SepalLengthCm~~
~~Unsupported~~

REJECTED
UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB

~~SepalWidthCm~~
~~Unsupported~~

REJECTED
UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB

~~PetalLengthCm~~
~~Unsupported~~

REJECTED
UNSUPPORTED

Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB

~~PetalWidthCm~~
~~Unsupported~~

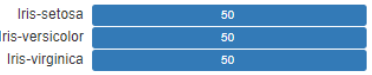
REJECTED

Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB

Species
Categorical

UNIFORM

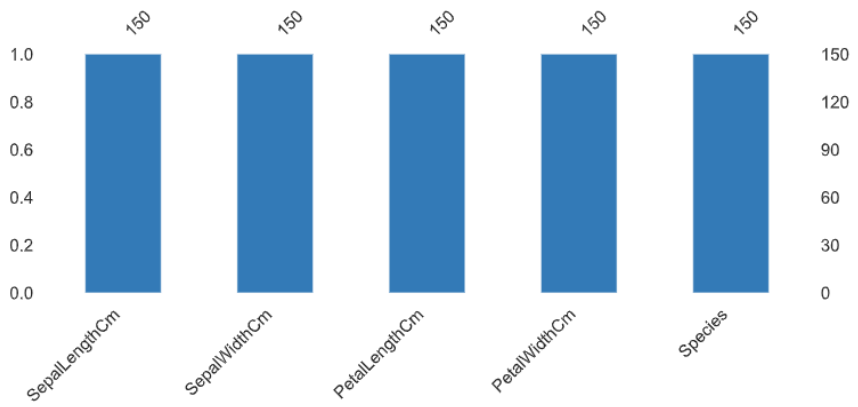
Distinct	3
Distinct (%)	2.0%
Missing	0
Missing (%)	0.0%
Memory size	1.3 KiB



Toggle details

Missing values

Count | Matrix



A simple visualization of nullity by column.

Sample

First rows

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa

Last rows

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
140	6.7	3.1	5.6	2.4	Iris-virginica
141	6.9	3.1	5.1	2.3	Iris-virginica
142	5.8	2.7	5.1	1.9	Iris-virginica
143	6.8	3.2	5.9	2.3	Iris-virginica
144	6.7	3.3	5.7	2.5	Iris-virginica
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

Duplicate rows

Most frequently occurring

	Species	# duplicates
0	Iris-setosa	50
1	Iris-versicolor	50
2	Iris-virginica	50