Project-1

Experiment-4:

Iris data:

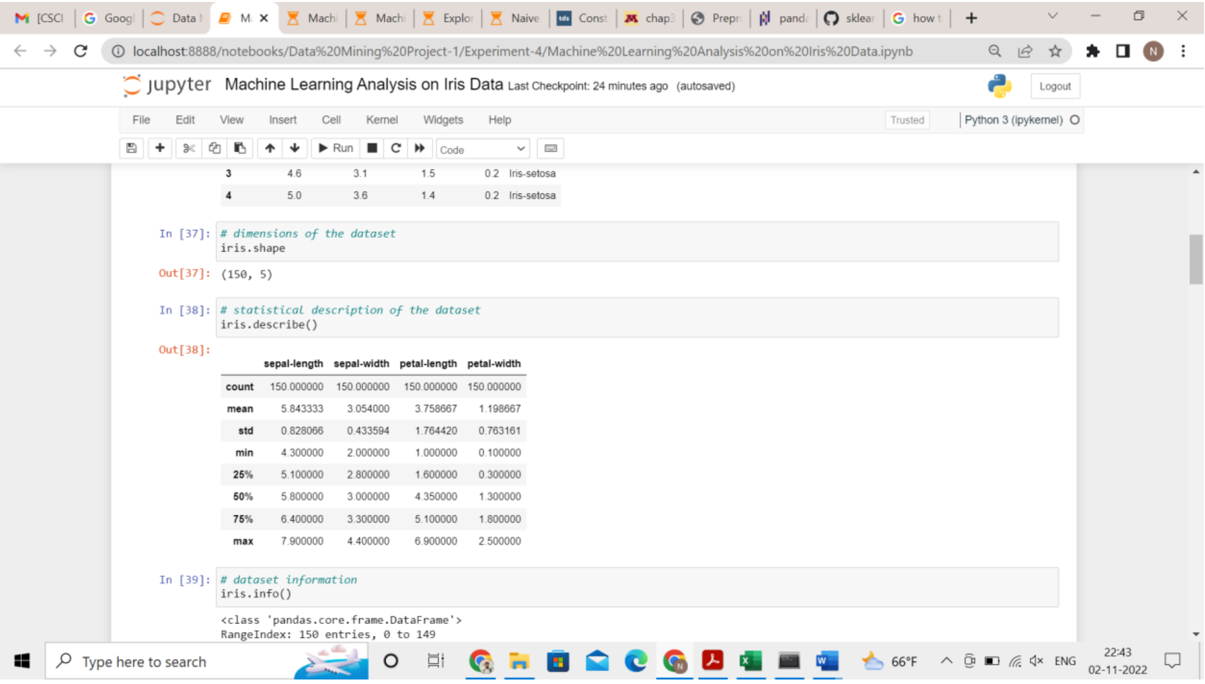
Context: The iris dataset is about finding the presence of diabetes in a person depending upon the attributes such as {sepal length, sepal width, petal length, petal width}.

Content: The input dataset has 150 records with 5 columns where first 4 columns are the features of the dataset and the last column has the label. The output label is a categorical datatype and the dataset has outcome with 50 Iris-setosa, 50 Iris-versicolor and 50 Iris-virginica. All features are of continuous data type.

Missing values or null data points: There is no missing or null data in the given dataset.

Feature: From the histogram data grouped between individual feature and outcome, all features have variations in their plots compared to other features, hence, all these features have significant impact while training the model.

Feature Description:


localhost:8888/notebooks/Data%20Mining%20Project-1/Experiment-4/Machine%20Learning%20Analysis%20on%20Iris%20Data.ipynb
jupyter Machine Learning Analysis on Iris Data Last Checkpoint: 24 minutes ago (autosaved)
Logout
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Trusted
Python 3 (ipykernel)
Run
Code
Type here to search
66°F
ENG
22:43
02-11-2022

3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

```
In [37]: # dimensions of the dataset
         iris.shape

Out[37]: (150, 5)

In [38]: # statistical description of the dataset
         iris.describe()

Out[38]:
```

	sepal-length	sepal-width	petal-length	petal-width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

```
In [39]: # dataset information
         iris.info()

         <class 'pandas.core.frame.DataFrame'>
         RangeIndex: 150 entries, 0 to 149
```

Testing Strategy: Stratified K Fold with 10 splits is the testing strategy deployed.

Classification accuracy of algorithms:

Using Train/ Test split:

Model	Accuracy
KNN	1.000000
DB	0.973684
GNB	0.315789
BNB	0.973684

Using K-Fold cross validation:

Model	Accuracy

Project-1

KNN	0.966667
DB	0.953333
GNB	0.333333
BNB	0.953333

KNN has the highest accuracy in both the train/ test split and K-Fold validation. Hence, KNN would be a better model in the case of Iris data.