SVM Classification on iris dataset

Prepared By: Erum Akmal

1. Problem Statement for Project:

In this project, I have used iris dataset and worked on support vector machines (SVM).

2. Data Sources and Description:

For data sources I used the data set for iris shared by the instructor for the assignment. (https://archive.ics.uci.edu/dataset/53/iris)

The dataset had 150 samples of iris flowers, with features such as sepal length, sepal width, petal length, and petal width.

3. Data Mining Operations:

-Data wrangling:

After loading and describing the data set, I proceeded with preprocessing of data using label encoder. After that data was normalized using min max scaler. In the next step, splitting of data was done.

SVM Classification was done after that and different kernel functions in SVM were added to see which one performs best.

Classification report for iris dataset and precision was generated after that. Accuracy was analyzed for different models.

-Libraries Used: I have used multiple libraries in this project including pandas, NumPy, matplotlib, seaborn, scikit-learn.

-Data Modeling: I have performed the analysis below in my project:

- Preprocessing of data
- Data normalization using Min Max Scaler
- Modeling with SVM Classification with different kernel functions
- Classification Report
- ROC Curve for different kernel functions

4. Model Outputs:

Below is my evaluation of the model performance for SVM:

```
Classification (SVM/SVC) Report for IRIS dataset: linear
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
ROC AUC: 1.00
-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.
Classification (SVM/SVC) Report for IRIS dataset: poly
Accuracy: 0.93
Precision: 0.93
Recall: 0.93
F1 Score: 0.93
ROC AUC: 0.99
-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.
Classification (SVM/SVC) Report for IRIS dataset: sigmoid
Accuracy: 0.37
Precision: 0.30
Recall: 0.37
F1 Score: 0.33
ROC AUC: 0.71
-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.-.
Classification (SVM/SVC) Report for IRIS dataset: rbf
Accuracy: 1.00
Precision: 1.00
Recall: 1.00
F1 Score: 1.00
ROC AUC: 1.00
```

In the above model, it is observed that SVM Classification was done and different kernel functions were added. Based on the above model, accuracy is best for linear (100%) and rbf (100%).

Model accuracy is also good for poly (93%) here and sigmoid has lease accuracy (37%).

ROC Curve was generated for this project
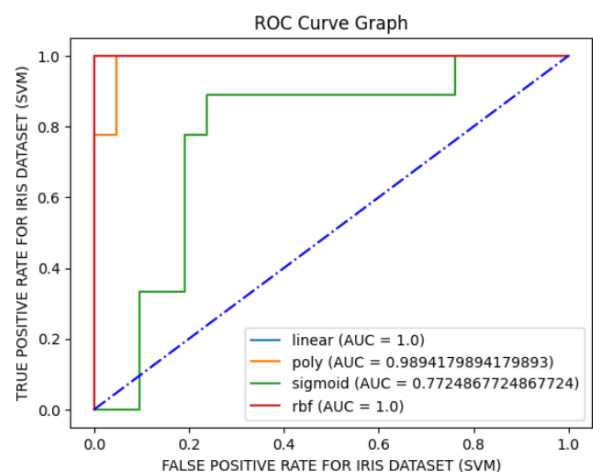
5. Limitations and Challenges:

Some of the limitations of this study are that I have only used a few variables in this study. There could be multiple other factors that could have been added, and more models could have applied.

In the above model, it is observed that SVM Classification was done, and different kernel functions were added. Based on the above model, accuracy is best for linear (100%) and rbf (100%).

6. Conclusion:

I was able to solve the problem partially. Based on the above model, accuracy is best for linear and rbf. SVM Classification with linear and RBF kernel has worked the best with 100% accuracy. With poly kernel, accuracy was 93% and with sigmoid kernel, accuracy was 37%.