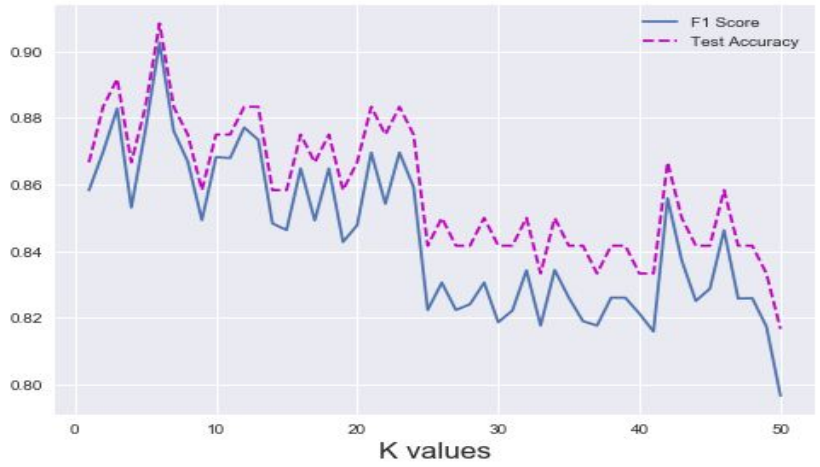
Machine Learning Project (416044)

Introduction

In the given Machine Learning task, various Supervised Machine Learning techniques have been used to implement the ML models for prediction on the Iris dataset. It is a multiclass problem that concerns the prediction of classes from one the three Setosa,Virginica or Versicolor class of flowers. Features included in predictions are petal width, petal length, sepal width and sepal length. A wide variety of ML classifiers have been used in the implementation ranging from Linear Logistic regression model to assemble Random Forests. Below is the summary of findings after successful implementation of models.

K Nearest Neighbours

- Different K- values produced a very different result for each value of K. Best values of K found out to be was at K=6




```
grid.best_params_
```

```
{'n_neighbors': 6}
```

it is evident that f1 and accuracy are maximum somewhere near k=6

Confusion Matrix, F1 and Accuracy score the best model are described below.

```
[24]: model.show_prediction_results(X_test,y_test,conf_matrix=True)
```

```
Accuracy of model is 0.9

F1 score of model is 0.9038901601830664

Confusion Matrix:

[[ 9  0  0]
 [ 0  8  0]
 [ 0  3 10]]

Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       0.73      1.00      0.84         8
           2       1.00      0.77      0.87        13

    accuracy                           0.90        30
   macro avg       0.91      0.92      0.90        30
weighted avg       0.93      0.90      0.90        30
```

Logistic Regression with ElasticNet

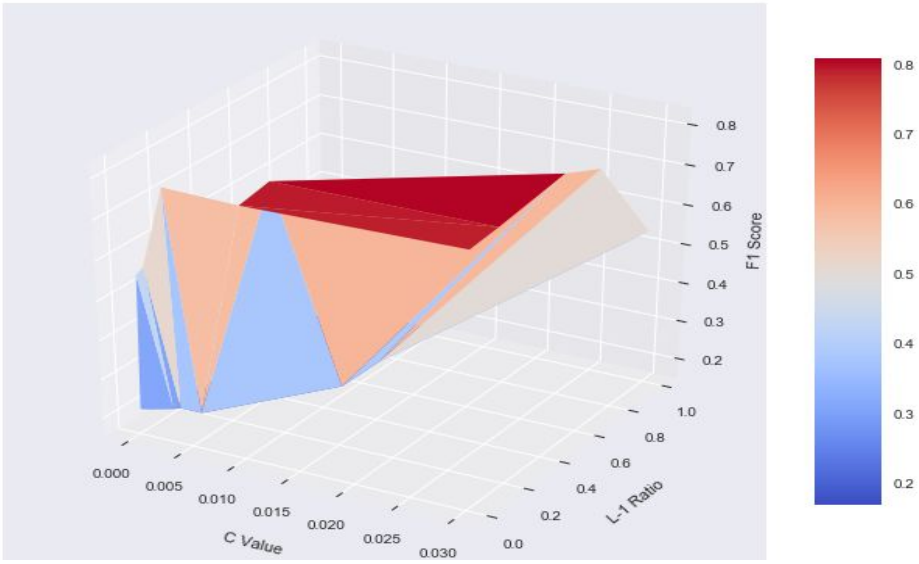
- ElasticNet implemented Logistic Regression produced the worse scores with 0.8 F1 and Accuracy with the best model whose alpha and l1-ratio are 0.03, 0.5 respectively. Surface plot and the metrics for the same are given below.

```
grid.best_params_
```

```
{'C': 0.03, 'l1_ratio': 0.5}
```

```
model.show_prediction_results(X_test,y_test,conf_matrix=True)
```

```
Accuracy of model is 0.8

F1 score of model is 0.8090909090909091

Confusion Matrix:

[[9 0 0]
 [0 8 0]
 [0 6 7]]

Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       0.57      1.00      0.73         8
           2       1.00      0.54      0.70        13

    accuracy                           0.80        30
   macro avg       0.86      0.85      0.81        30
weighted avg       0.89      0.80      0.80        30
```



Support Vector Machine with RBF Kernel

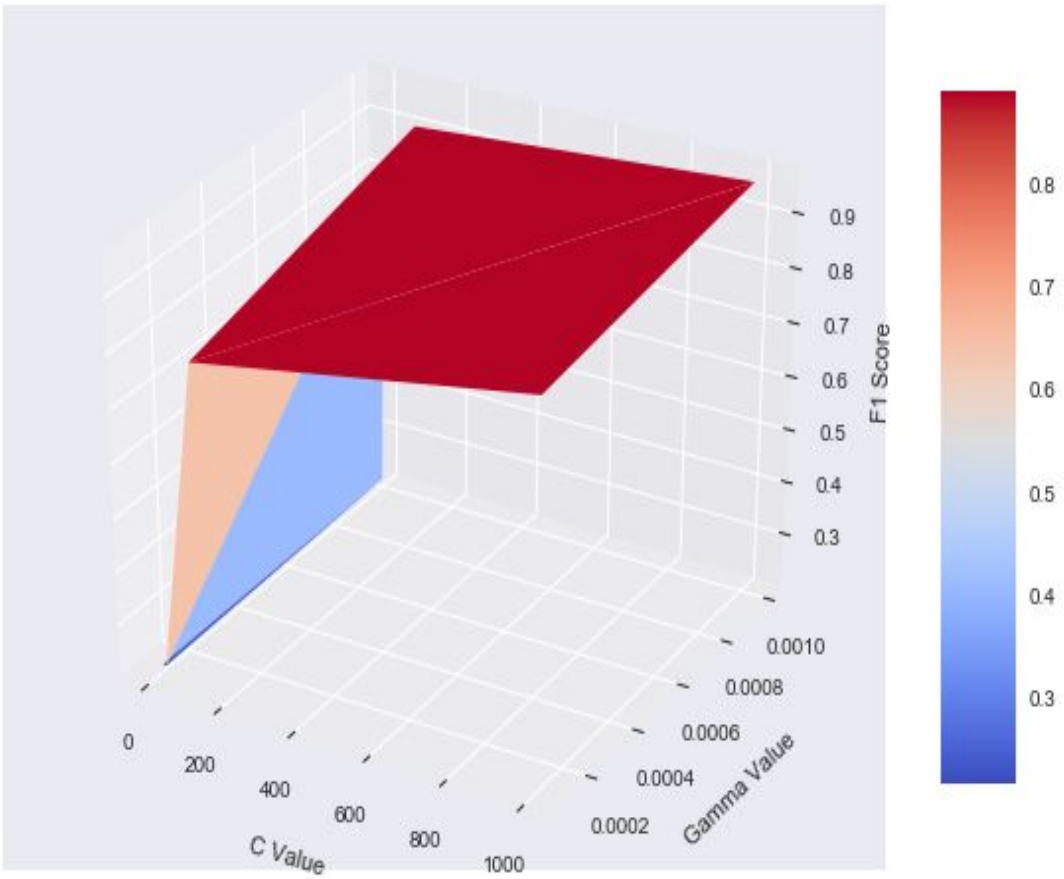
- SVC produced the best result on the given data with more than 0.95 accuracy as well as F1 score with values of C = 1000 and gamma= 0.001. Surface plot of F-1 scores versus gamma and C values are depicted below.

```
mean_test_acc = result['mean_test_accuracy']
grid.best_params_
```

```
{'C': 0.03, 'l1_ratio': 0.5}
```

```
model.show_prediction_results(X_test,y_test,conf_matrix=True)
```

```
Accuracy of model is 0.8

F1 score of model is 0.8090909090909091

Confusion Matrix:

[[9 0 0]
 [0 8 0]
 [0 6 7]]

Classification Report:

              precision    recall  f1-score   support

           0       1.00      1.00      1.00         9
           1       0.57      1.00      0.73         8
           2       1.00      0.54      0.70        13

    accuracy                           0.80        30
   macro avg       0.86      0.85      0.81        30
weighted avg       0.89      0.80      0.80        30
```



Random Forest

- Random Forest classifiers are used with different varieties of numbers of classifiers and the depth of the trees. With a best depth of 300 and using 700 Decision Trees, the model performed quite well with the accuracy and F 1 score of 0.9. Final metric for the same is given below.

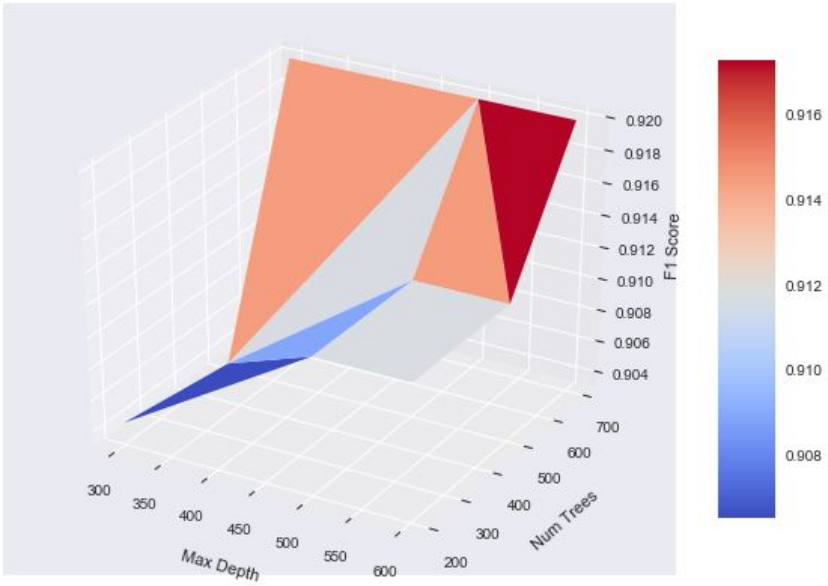
```
grid.best_params_
```

```
{'max_depth': 300, 'n_estimators': 700}
```

```
model.show_prediction_results(X_test,y_test,conf_matrix=True)
```

```
Accuracy of model is 0.9

F1 score of model is 0.9038901601830664

Confusion Matrix:

[[ 9  0  0]
 [ 0  8  0]
 [ 0  3 10]]

Classification Report:
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	9
1	0.73	1.00	0.84	8
2	1.00	0.77	0.87	13
accuracy			0.90	30
macro avg	0.91	0.92	0.90	30
weighted avg	0.93	0.90	0.90	30



Discussion

Out of all the given classifiers, Support Vector Classifier (SVC) with the penalty of C = 1000 and gamma = 0.001 has produced the best results for our dataset under the given constraints with accuracy of 0.966 and Micro F1 score of 0.967 on test data. This is due to the fact that Iris dataset is fully separable by the support vectors and would have produced a score of more than the current if more data preprocessing and post pata processing like SMOTE would have been applied to the classes. Even though the classes were in proportion but due to the split and high CV (10), the classes were absent in either training or test set. After applying over sampling techniques like SMOTE, we can easily achieve a better score than this.

Logistic Regression with ElasticNet turns out to be the worst model in terms of prediction because of many reasons like the given constraints on the penalty (alpha) the ratio between l-1 and l-2 terms. Other than that, it turns out to be a very simple model and we can say it has a higher bias than expected because every other model performed well over 0.9 given the chances of random guessing are 0.33 for finding the right classes.

Future Improvement

There is a huge room for improvement for the model by using feature engineering and trying different parameters for the model. Generating new features like ratio between sepal, petal width and length or finding a new function for feature generation could help. By using different oversampling techniques like SMOTE or for the simplest part, by simply stratifying the target label *y*, we can increase the chances of getting higher scores. ***As the dimensions of the attribute (p) is far less than the number of samples (N), we can easily use any of the Forward as well as backward elimination or Recursive methods to select features***. Using mutual scores and other statistical tests or changing the threshold value for correlation elimination, we can improve our model.