SVM – Assignment 4
ID: eo9232
Name: Md Reza
IE7860 – Winter 2022

Project Report

Background and Methods:

IRIS Data Classification: The IRIS database has 50 samples distributed among three different species of IRIS. These samples have specific characteristics to classify into *Iris Setosa, Iris Virginica, and Iris Versicolor*. This study proposes an SVM C-Support Vector Classification with a *LINEAR* transfer function to separate and classify the IRIS data.

Diabetes Prediction: This study proposes a machine learning model SVM C-Support Vector Classification with *RBF* transfer function to identify diabetes patients.

Apple Stock Price Time Series Forecasting: This study includes time series forecast to predict Apple stock prices using SVM C Epsilon-Support Vector Regression with *LINEAR* transfer function, Mean Squared Error (MSE) loss, and Regressor Attributes.

Pre-processing:

IRIS Data Classification	**1.** Check for missing values **2.** One-hot encoding of categorical data **3.** Scaled the dataset
Diabetes Prediction	**1.** Check for missing values **2.** Impute missing values with mean values **3.** Scaled the dataset
Apple Stock Price Time Series Forecasting	**1.** Check for missing values **2.** Determine the best regressor input **3.** Create regressor attributes **4.** Scaled the dataset

Justification of Choosing Activation and Loss Function:

MLP Model	**Transfer Function**	**Justification**
IRIS Data Classification	*LINEAR*	The linear kernel performs well for classification or regression problems. In general, the linear kernel performs well when a linear decision boundary or a linear fit into the data.
Diabetes Prediction	*RBF*	RBF kernel creates non-linear combinations of selected features to uplift the samples into a higher-dimensional space that would allow a linear decision boundary to separate the classes.
Apple Stock Price Time Series Forecasting	*LINEAR*	The linear kernel performs well for classification or regression problems. In general, the linear kernel performs well when a linear decision boundary or a linear fit into the data.

Hyper-Parameter Tuning:

MLP Model	Grid Search with Cross-Validation	Justification
IRIS Data Classification	GridSearchCV (From Scikit-learn)	To find the best parameters.
Diabetes Prediction	GridSearchCV (From Scikit-learn)	To find the best parameters.
Apple Stock Price Time Series Forecasting		Not applicable

Model Performance:

MLP Model	Regularization Parameter	Justification
IRIS Data Classification	C Parameter	The C parameter trades off the correct classification of training examples against the margin of the decision function. Hence, for the more significant value of C, optimization will choose a smaller-margin hyperplane to classify all training points correctly. In contrast, for the smaller value of C, the optimization will choose a larger-margin hyperplane with a simple decision function that costs training accuracy. Specifically, the C parameter behaves like a regularization parameter in the SVM model.
Diabetes Prediction	C Parameter	The C parameter trades off the correct classification of training examples against the margin of the decision function. Hence, for the more significant value of C, optimization will choose a smaller-margin hyperplane to classify all training points correctly. In contrast, for the smaller value of C, the optimization will choose a larger-margin hyperplane with a simple decision function that costs training accuracy. Specifically, the C parameter behaves like a regularization parameter in the SVM model.
Apple Stock Price Time Series Forecasting	C Parameter	The C parameter trades off the correct classification of training examples against the margin of the decision function. Hence, for the more significant value of C, optimization will choose a smaller-margin hyperplane to classify all training points correctly. In contrast, for the smaller value of C, the optimization will choose a larger-margin hyperplane with a simple decision function that costs training accuracy. Specifically, the C parameter behaves like a regularization parameter in the SVM model.

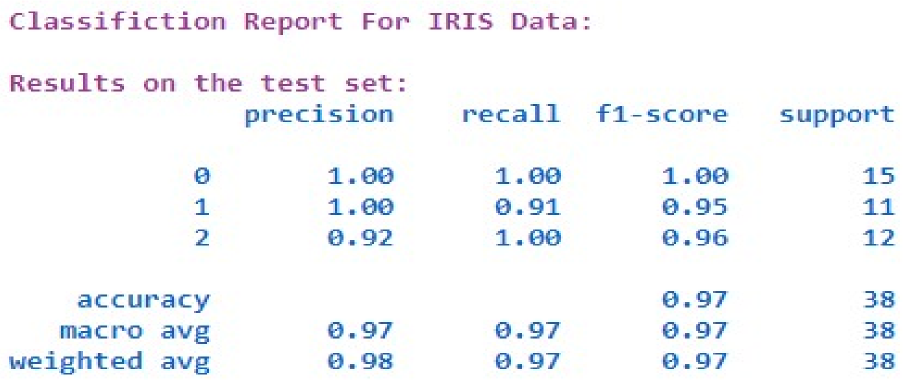
MLP Model Evaluation:

One way to measure models' performance on the training, validation, and test data is to compare the prediction values with the actual values that include Accuracy, Cross-Validation Score, Precision, Recall, F1-Score, AUC Score, and Confusion Matrix. Please refer to the **MLP Model Summary Table** in the appendix below for details on model performance.

Summary:

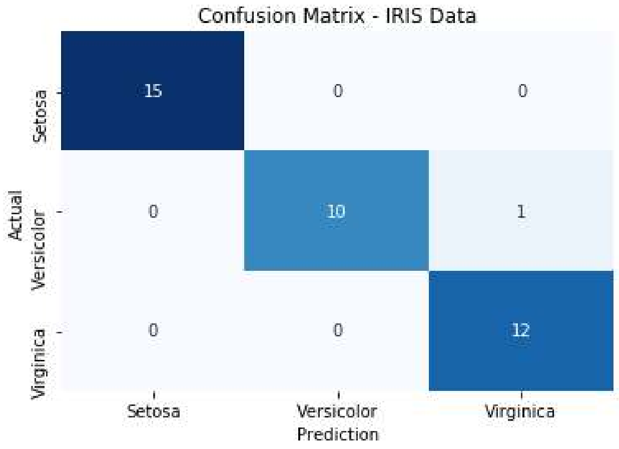
After comparing models’ output, it turned out that Precision, Recall, F1-Score, Cross-Validation Score, Accuracy, and R Squared (or r2) scores and the score for training and test data play a crucial role in evaluating the model performance. The goals were to build MLP models to demonstrate the power of the SVM Kernels in terms of classification, prediction, and time series forecasting. The observations-based models’ outputs are as follows:

MLP IRIS species classification model with C-Support Vector Classification performed well to correctly classify all training points with smaller-margin hyperplane and high accuracy, as demonstrated in the below classification report. Grid-search with cross-validation is used for hyper-parameters tuning to find the best parameters.



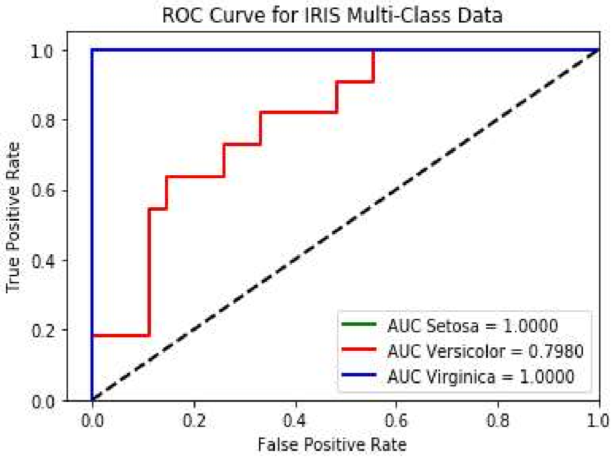
```
Classifiction Report For IRIS Data:

Results on the test set:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        15
           1       1.00      0.91      0.95        11
           2       0.92      1.00      0.96        12

    accuracy                           0.97        38
   macro avg       0.97      0.97      0.97        38
weighted avg       0.98      0.97      0.97        38
```

In addition to the precision, recall, and f1-score, the performance of this multi-class classification model is measured with confusion matrix and area under the AUC curve, and it seems the model performance is promising in classifying IRIS species as demonstrated below.



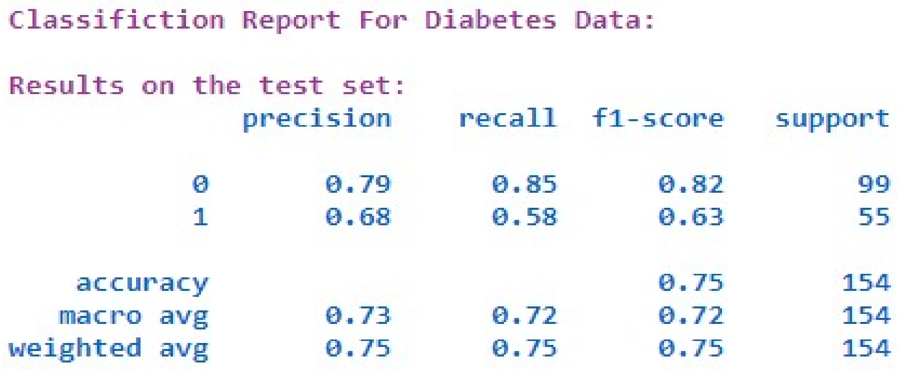
Confusion Matrix - IRIS Data

Actual \ Prediction	Setosa	Versicolor	Virginica
Setosa	15	0	0
Versicolor	0	10	1
Virginica	0	0	12



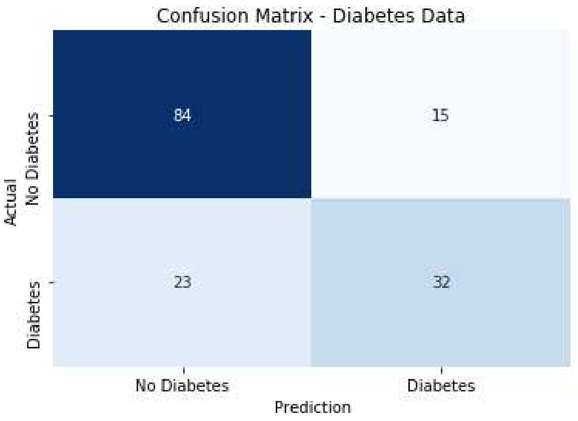
ROC Curve for IRIS Multi-Class Data

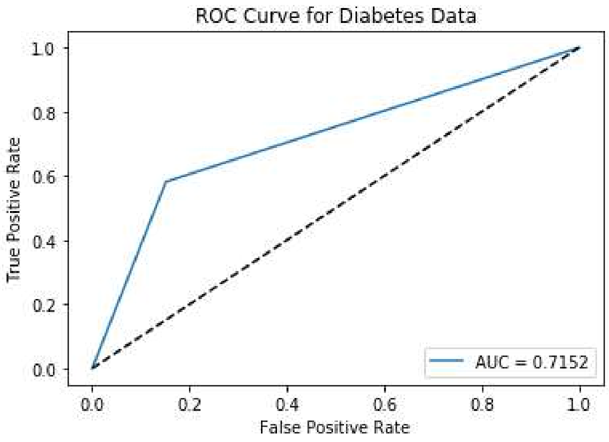
MLP Diabetes prediction model with C-Support Vector Classification has about 80% accuracy that could improve with the model parameters optimizations. Grid-search with cross-validation is also leveraged for hyper-parameters tuning to find the best parameters.



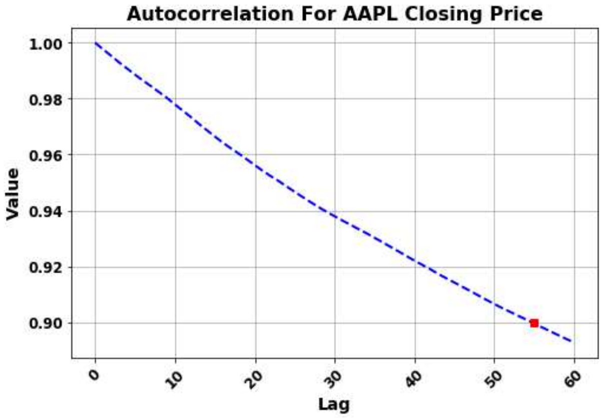
```
Classifiction Report For Diabetes Data:

Results on the test set:
              precision    recall  f1-score   support

           0       0.79      0.85      0.82        99
           1       0.68      0.58      0.63        55

    accuracy                           0.75       154
   macro avg       0.73      0.72      0.72       154
weighted avg       0.75      0.75      0.75       154
```

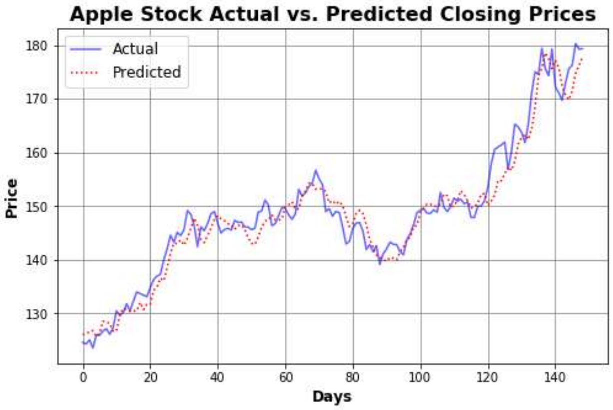
However, with about 80% precision, 85% recall, and 72% AUC score, the Confusion Matrix could detect many possible diabetes cases.





MLP Apple stock price time-series forecasting model with mean squared error and regressor inputs from the autocorrelation chart below *(i.e., 55)* this time-series forecasting model performed well in predicting the closing price at least for a given day on the forecasting horizon.





In addition, the R-Squared Score, Mean Squared Log Error (MSLE), and Mean Absolute Percentage Error (MAPE) measured the distance between the actual values and values lying on the predictor hyperplane also seems promising in predicting the closing price reliably with this time-series forecasting model.

APPENDIX:

Model Summary:

Model Summary			Training	Test	Training Cross-Validation		Test Cross-Validation		R-2 Score, MSLE, & MAPE
MLP Model	**Dataset**	**Transfer Function**	**Accuracy**	**Accuracy**	**Mean Score**	**Standard Deviation**	**Mean Score**	**Standard Deviation**	**N/A**
IRIS Data Classification	IRIS Data	*LINEAR*	**98.21**	**97.37**	**94.62**	**8.3**	**97.50**	**7.50**	**N/A**
Diabetes Prediction	Diabetes Data	*RBF*	**83.06**	**75.32**	**76.54**	**3.97**	**79.37**	**7.06**	**N/A**
Apple Stock Price Time Series Forecasting	AAPL Stock Data	*LINEAR*	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**94.98%** **0.0003** **1.51%**