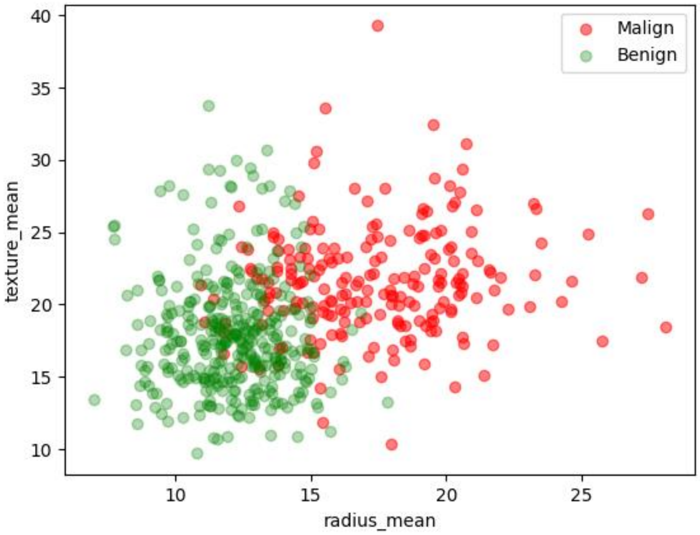
Introduction:

This project aims to train 3 Machine Learning models on a breast cancer diagnostic data set to predict if a person has Malignant (threatening) or Benign (non-threatening) breast cancer.

The data set has 30 real valued input features and the Predicting Field where the data is represented as M for malignant or B for Benign and has 569 instances. The 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass which describe the characteristics of the cell nuclei present in the image. The distribution of the class is 357 benign, 212 malignant.

The features of the data set are standardized for the training set and test set individually so the information of the test data remains unseen.



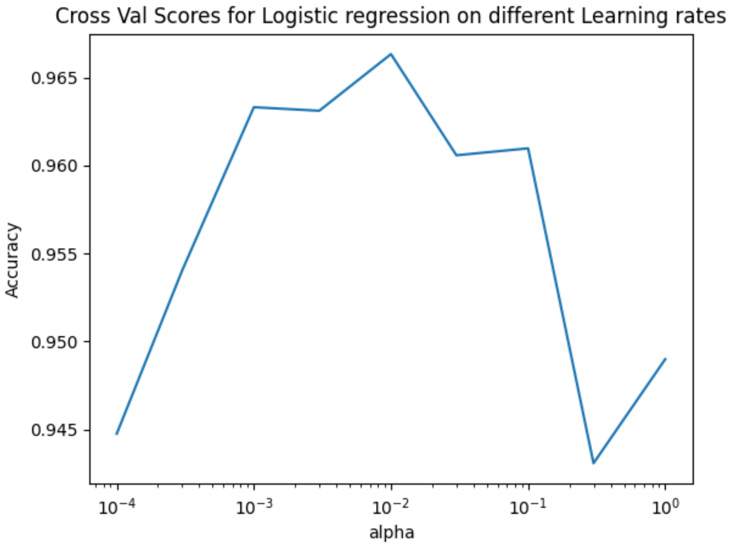
Model Selection:

The trivial model for this task is random guess and the three implemented models are logistic regression (LR), K- nearest neighbours (KNN) and Multi layer perceptron (MLP) or Neural Networks. The three models are implemented using Scikit Sklearn machine learning library.

Logistic Regression Model:

Logistic Regression is implemented through Stochastic Gradient decent and the sigmoid activation function to predict the given data set. LR has learning rate α as its hyperparameter which is the step size to update the weights.

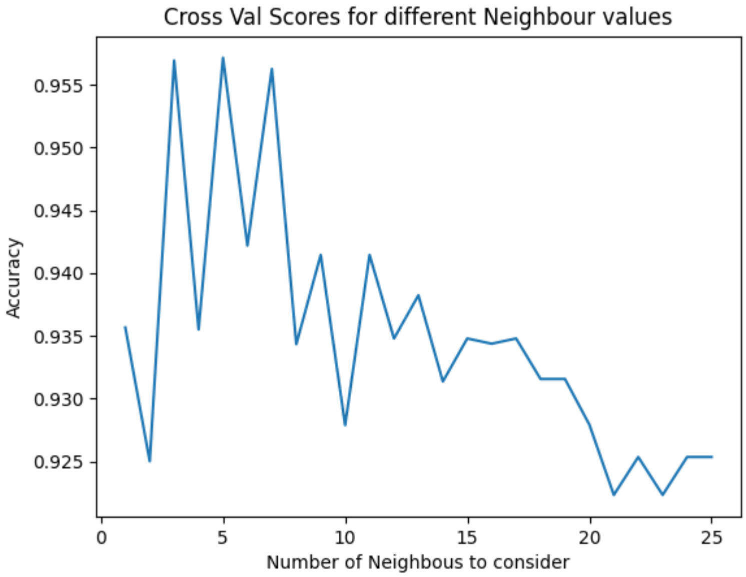
The α was tuned from this list [1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001,0.0003, 0.0001] using K-fold cross validation on the training data set. Cross validation of the training data set yielded an F-1 score and the mean of which was taken, this was repeated for all different α values stated above, from this the best α was chosen using the best mean of K-fold scores.



K- Nearest Neighbours:

This model predicts the data using the K-nearest neighbour's vote. The main hyperparameter for this model is the n-neighbours which is the number of neighbours the model uses to classify the data.

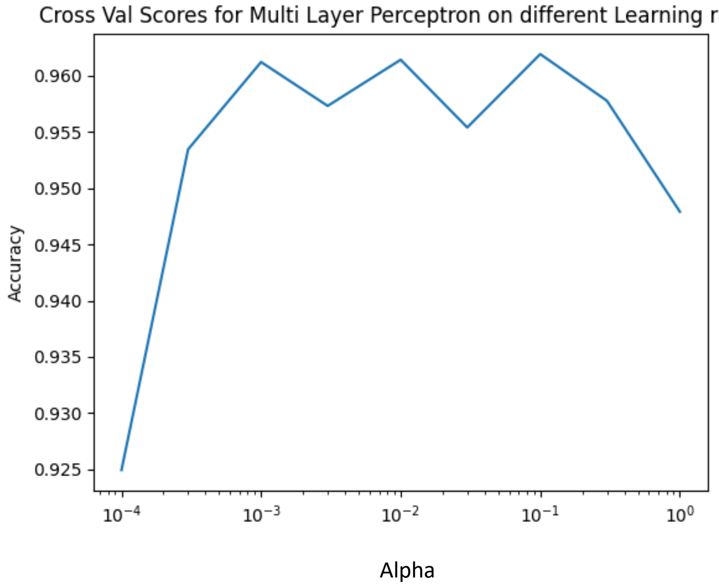
The n-neighbours hyperparameter is tuned from the range of [1, 25] using K-fold cross validation on the training data set. Cross validation of the training data set yielded an F-1 score and the mean is taken, this was repeated for all values of n-neighbours from the range state above, from this the best n-neighbours was chosen using the best mean of K-fold scores



Multi Layer Perceptron:

Multi Layer perceptron is a neural network model with multiple layers containing multiple perceptron or nodes. Each node is essentially an input for all nodes in the next layer and is called a fell forward neural network. This model has 1 hidden layer with 50 perceptron or nodes in each layer.

The hyperparameter α was tuned from this list [1, 0.3, 0.1, 0.03, 0.01, 0.003, 0.001,0.0003, 0.0001] using K-fold cross validation on the training data set. Cross validation of the training data set yielded an F-1 score and the mean of which was taken, this was repeated for all different α values stated above, from this the best α was chosen using the best mean of K-fold scores.

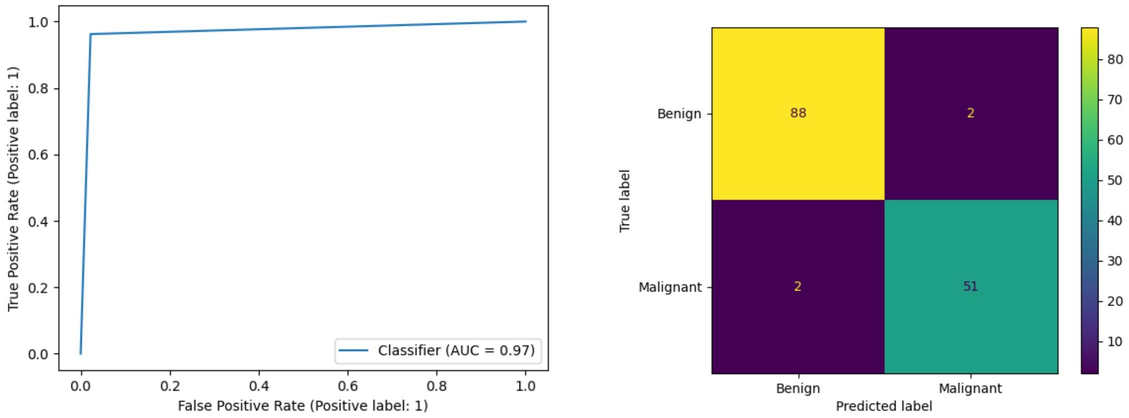


Evaluation metric:

Each model was evaluated based on F-1 Score of the model on the test data set. We use the F-1 score as the metric rather than accuracy (the number of correct predicted samples on the total number of samples) because of the real cost of predicting a false negative, since predicting Benign when the actual data is Malignant is much worse than the vice-versa. Therefore, the F-1 score helps weigh the consequences of the model choosing a false negative into the accuracy giving us a much better idea of the true accuracy of the model.

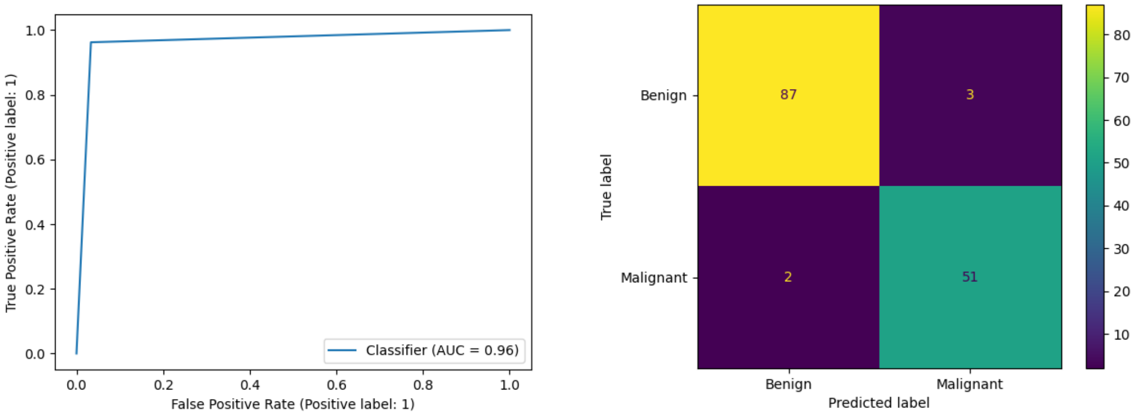
Results / Observations:

The accuracy of the trivial model is 0.5 or 50% whereas for Logistic Regression we can see that the F-1 score for is 0.96 or 96% on the test data set with the F-1 score of 0.97 on the training data set which shows the model is generalizable.



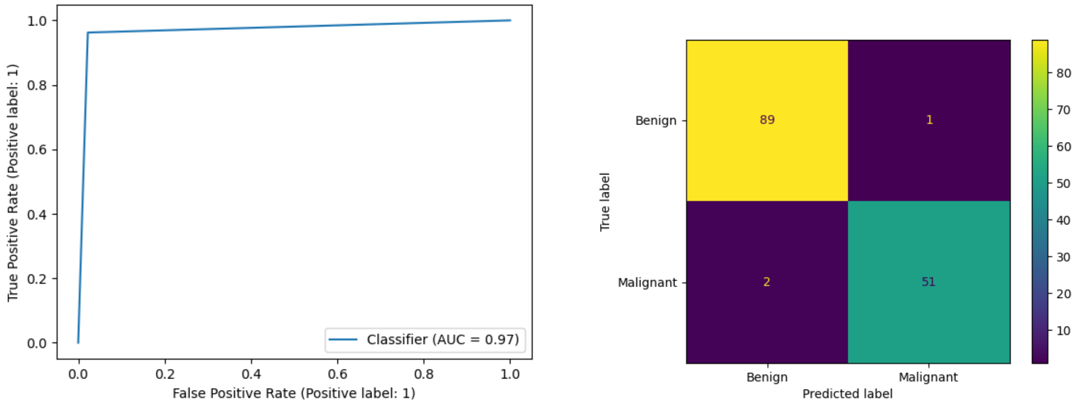
ROC and Confusion Matrix for Logistic regression

Furthermore, we can see that the number of false negatives is considerably low than what could be for the trivial problem. Similarly, the number of predicted false positives are also low making Logistical regression a good model for this data set. The same Can be said for KNN where the F-1 score for the test data set is 0.95 or 95% and an F-1 score of 0.96 on the training data set, though the accuracy for KNN is considerably better than the trivial model it is still lacking behind logistic regression and having a slight increase in the number of predicted false positives.



ROC and Confusion Matrix for KNN

We see MLP is a very well learnt model on this data set with an F-1 score of 0.97 or 97% and an F-1 score of 0.98 on the training data set with lower number of false positives but does not provide a considerable advantage over the other two models.



ROC and Confusion Matrix for MLP

Conclusion:

The 3 learnt models of Logistic regression, K-Nearest Neighbours and Multi layer perceptron are a great at predicting the breast cancer diagnostics data with high generalizability and minimized number of false negatives which makes the predictions more reliable compared to the trivial model.