**Module 3 – Data Analytics Report**

**Name: Maviya Shaikh**

**Student number: 100766785**

**Date: April 5th,2020**

Table of Contents

[1. Introduction 3](#_Toc36980172)

[2. Key Questions 3](#_Toc36980173)

[3. Dataset Summary 3](#_Toc36980174)

[4. Model Analysis 4](#_Toc36980175)

[i. Standard Models 4](#_Toc36980176)

[ii. Learning curve and Optimized model of Random Forest Algorithm 6](#_Toc36980177)

[5. Results 7](#_Toc36980178)

[6. Conclusion 7](#_Toc36980179)

[7. Appendices 7](#_Toc36980180)

# Introduction

The purpose of this project is to develop the solution for Princess Margret Hospital to identify whether the patient has stage 1 cancer or not. The objective of this project is to alert the patient at potential cancer risk so that they can start recovering at early stage of the cancer.

# Key Questions

This report will provide answer to the below three major questions of the project:

1. **To identify patient who has stage 1 cancer**

The solution develop should be answerable to detect all the patient with cancer as misclassification of this patient is not tolerated at any cost because the patient might lose their life.

1. **To identify patient who do not have cancer**

It is also very important to detect these patients because they should not get wrong treatment which is not required.

1. **To understand the key factors that affect the cancer**

It helps to know the characteristics of the features that led cancer in the patient. So, it is necessary to understand the dataset thoroughly.

# Dataset Summary

The dataset provided has 10 variables that includes 9 independent variables and 1 dependent variable. The independent variables (V1, V2, V3, V4, V5, V6, V7, V8, V9) are cell structures and one dependent variable (Class) identifies cancer on the basis of the cell structure. Class 0 represent no cancer and Class 1 represent cancer in the patient.

The result of Exploratory Data Analysis is as follows:

* The basic statistics like mean, median, mode and standard deviation is calculated using describe method to provide the brief overview of the dataset
* From pandas profiling, it is identified that dataset has 10 missing rows which is removed to avoid discrepancy in the final deliverable
* From histogram and box plot, it is determined that data is not normally distributed
* From heat-map, it is determined that the variables are highly co-related with each other
* From the Tukey outlier detection test 120 rows are detected as outliers, but this is the original data of the patient, so in this project outliers are not removed, and the model selected to detect the cancer are not sensitive to outliers
* The dataset is not balanced, therefore Synthetic Minority Over-sampling Technique (SMOTE) is used to balance the dataset by manually creating the data for minority class
* The data is scaled using Standard scalar method to normalize the values of the independent continuous variable within specific range

# Model Analysis

For the model analysis, the following task are completed in order to build the model:

* The training and test size of the model is decided as 80% to train the model and 20% to test the model
* The train and test data are scaled using Standard Scalar method and after that trained data is balanced using SMOTE technique to remove the biased in the model
* In the prediction of the model, Class 0 represent *‘No Cancer’* and Class 1 represent ‘*Cancer’*

## Standard Models

The three standard models are created to obtain the objective of the project. The [confusion matrix](#_Appendices) and [classification report](#_Appendices) of each model is given below to analyse the performance of the model and all the terms are explained in appendix.

1. **K-Nearest Neighbour**: The reason for choosing this model is that it is a simple machine learning model and it does not make any assumption about distribution of the data. It classifies the data points analysing nearest neighbour from the training set.

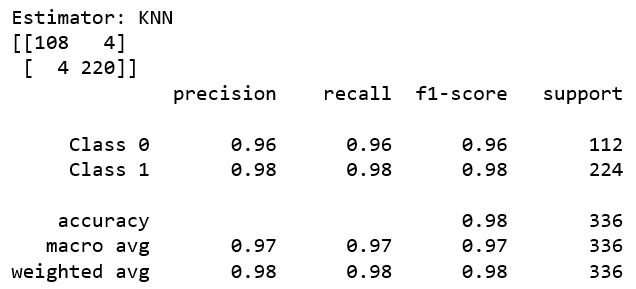


Figure 1: Confusion matrix and classification report of K-NN

Confusion matrix:

* No cancer – KNN model correctly identified 108 patients and misclassify 4 patients
* Cancer – KNN model correctly identified 220 patients and misclassify 4 patients

Classification report:

* Overall, the model got 98% predictions correct and identified 98% of the patient correctly in detecting the cancer and no cancer

1. **Random Forest:** It is a classification algorithm consist of multiple decision trees. The reason for choosing this model is that it operates as ensemble i.e. class with the most votes is selected for prediction.

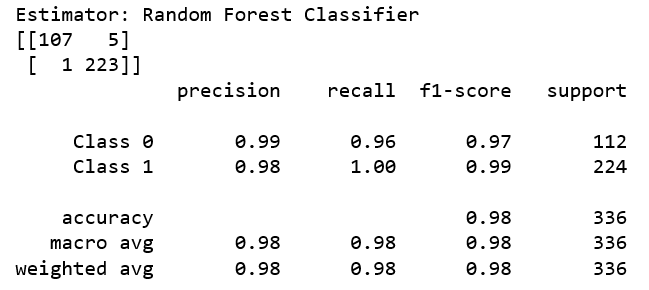


Figure 2: Confusion matrix and classification report of Random Forest

Confusion matrix:

* No cancer – Random Forest model correctly identified 107 patients and misclassify 5 patients
* Cancer – Random Forest model correctly identified 223 patients and misclassify only 1 patient

Classification report:

* Overall, the model got 98% predictions correct and identified 98% of the patient correctly in detecting the cancer and no cancer

1. **Neural Network:** The reason for choosing this model is that it has the ability to perform complex problems and is good at handling large amount of data. It tries to mimic the brain by running the data points through various hidden layers and provides the final output in terms of prediction.

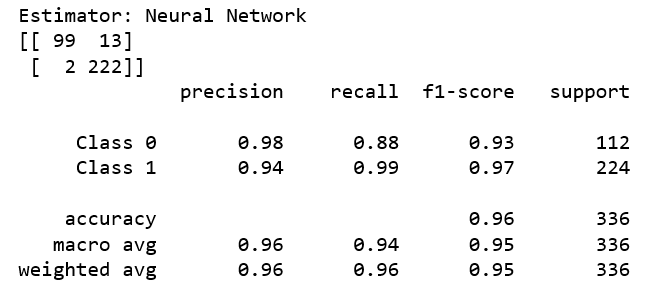


Figure 3: Confusion matrix and classification report of Neural Network

Confusion matrix:

* No cancer – Neural Network model correctly identified 99 patients and misclassify 13 patients
* Cancer – Neural Network model correctly identified 222 patients and misclassify 2 patients

Classification report:

* Overall, the model got 96% predictions correct and identified 96% of the patient correctly in detecting the cancer and no cancer

**Among all the models, Random Forest algorithm and K-NN algorithm has highest accuracy of 98%. However, the f1-score of Random Forest is 1% more than K-NN. Thus, the Random Forest algorithm is chosen as the best model.**

## Learning curve and Optimized model of Random Forest Algorithm

The learning curve for Random forest algorithm is created to evaluate the performance of the model on training and validation data. The below figure indicates number of training sample on X-axis and recall on Y-axis for analysis. It is analysed that model correctly predicts on training data as the model is trained on it. The purpose of this learning curve is to see the prediction on validation or test data, as the instances of dataset increases the model tries to fit perfectly. So, this model seems a right fit for this dataset.

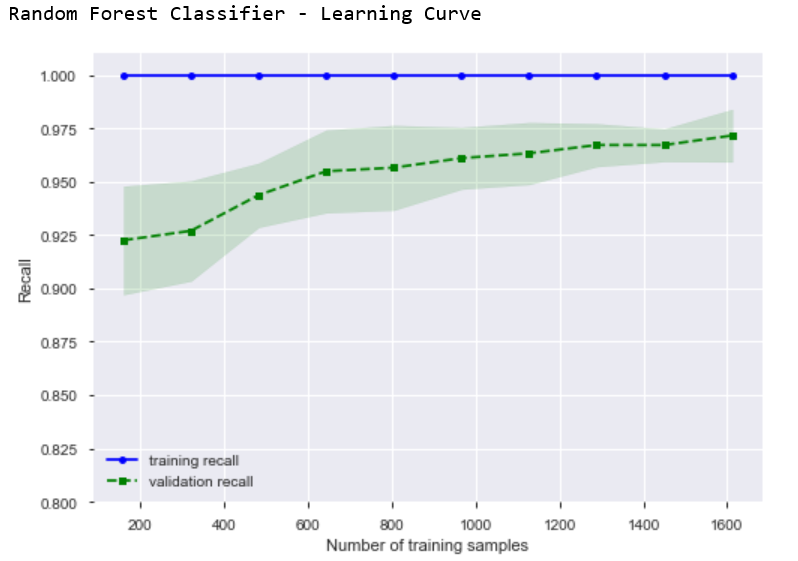


Figure 4: Learning Curve of Random Forest

The optimized model is created using [cross validation](#_Appendices) and [Grid Search](#_Appendices) Optimization technique. However, the optimized model gives the same result for confusion matrix and classification report as the standard model. Instead, nested cross validation gives less accuracy than the standard model.

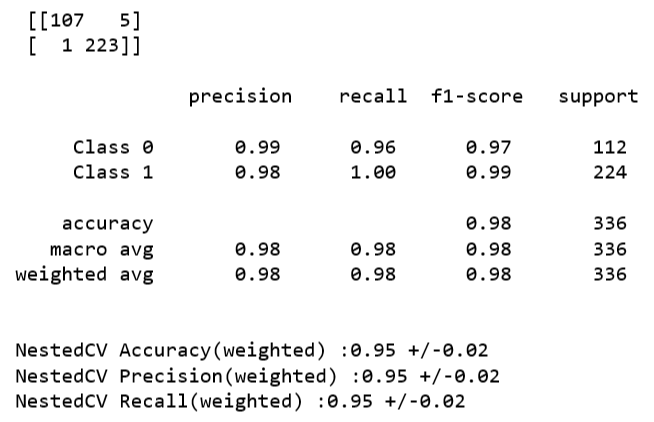


Figure 5: Confusion matrix and classification of optimized Random Forest model

Thus, the standard model outperformed the optimized model and the computational cost of the standard model is also less as it eliminates computational time of optimization techniques.

# Results

The insights from the model analysis is as follows:

* All the models give better prediction for patient who has cancer (Class 1) compare to patient who do not have cancer (Class 2)
* The optimized learning model of Random Forest does not give better performance than the standard model
* From the learning curve, it is derived that as the data points are added the validation recall will improve. So, if more data is collected the performance of the model is improved
* The K-NN is simple algorithm while Neural Network is complex algorithm. So, Random Forest algorithm suits the dataset of this project better than other algorithms because it is more accurate and prevents overfitting
* Our goal of reaching more than 95% of accuracy, precision, recall and f1-score is accomplished with the development of Random Forest model

# Conclusion

The solution developed provides the answers to three key questions of the project. Firstly, the Random Forest Classifier is successful in detecting all the cancer patients except one patient. Secondly, the Random Forest Classifier is 96% accurate in detecting patient who do not have cancer. Finally, the features do not have normal distribution of the data and are highly co-related with each other that depicts that the more co-related the independent variables are, it results into better prediction of the cancer because intuitively the cell structures of an individual’s patient body are similar.

# Appendices

**Confusion matrix**: It is a table that is used to evaluate the performance of a model on the 20% test data based on the 80% train data.

**Classification report**: It demonstrates the results of key metrics for the model. The explanation of classification report metrices are as follows:

1. **Precision**: It tells us how often the model is correct when it makes the prediction
2. **Recall**: It is the ratio of correctly classified the positive instances from the total number of positive instances or correctly classified negative instances from the total number of negative instances i.e. It identifies all the relevant instances from retrieved instances
3. **F1-Score**: It is the harmonic mean of precision and recall
4. **Accuracy**: It is the ratio of the number of correct predictions to the total number of predictions made i.e. It represents the percentage of predictions that model got right

**Cross-Validation:** It measure the performance of the model on new test data by dividing the dataset into training set and validation or testing set.

**Grid Search:** It trains the algorithm for all combinations by using learning rate and number of layers. It calculates the best parameters of the model using cross validation technique.