BREAST CANCER CLASSIFICATION

By: Utpal Mishra (20207425)

Under the supervision of

Dr. Brendan Murphy

DATA AND COMPUTATION SCIENCE

UNIVERSITY COLLEGE DUBLIN

BELFIELD, DUBLIN 4, IRELAND

BREAST CANCER CLASSIFICATION

By: Utpal Mishra (20207425)

Under the supervision of Dr. Brendan Murphy

Submitted to the Department of Mathematics and Statistics in partial fulfilment of the requirements for the degree of Master's in Data and Computation Science.

DATA AND COMPUTATION SCIENCE

UNIVERSITY COLLEGE DUBLIN

BELFIELD, DUBLIN 4, IRELAND

CERTIFICATE

This is to certify that Project Report entitled “Breast Cancer Classification” which is submitted by **Utpal Mishra,** in partial fulfilment of the requirements for the award of master’s degree in Data and Computational Science at University College Dublin, is a record of the candidate's own work carried out by him under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Supervisor: Dr. Brendan Murphy

Date: 29th December 2020

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the master's Project undertaken during the Autumn Trimester. I owe a special debt of gratitude towards ***Dr. Brendan Murphy*** *School of Mathematics and Statistics, University College Dublin, Dublin, Ireland for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant effort that our endeavours have seen the light of the day.*

I also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project.

Name: Utpal Mishra
Student Number: 20207425
Date: 29th December 2020

TABLE OF CONTENTS

CONTENT PAGE NO.

PROJECT OBJECTIVE

The general objective of the project is to design a system and automated the process of breast cancer detection using the techniques of machine learning and provide some assistance in diagnosis and treatment to the radiologists and physicians. To achieve this, the following are some specific objectives:

1. To analyse previous related works on breast cancer detection and classification to select relevant methods and techniques.

2. To opt for befitting methods for segmentation, data pre-processing, feature extraction and classification.

3. To deliver a prototype for the proposed approach.

4. To assess the performance of the project in terms of f-score, precision, recall, error and classification accuracy.

5. To deploy a time - efficient and reliable prototype that could assist pathologists.

ABSTRACT

According to **Cancer Prevention and Control** (CPC) breast cancer is the second most common type of cancer after lung cancer and the fifth most common cause of cancer death. Among all the form of cancer in women, breast cancer contributes to over 20 % and has resulted with more in millions new cases each year (over 5 lakhs recorded in 2012). With lack in diagnosis and treatment in India, a woman is detected with breast cancer in every 4 minutes along-with the highest death cases worldwide (2012) i.e., in every 8 minutes.

Automating the process will be beneficial for the people suffering from cancer, being can be cost - effective with elevated accuracy. And with this project, I would like to present multiple models from the traditional to neural networks and to average/ weighted and stacked ensemble models, evaluating, analysing, understanding, summarising, and interpreting different features and element of the models from data pre-processing to model tuning to obtain some efficient model.

In the project, the best fitted model with and without feature selection and with under-sampling/ under-sampling are the traditional and average/ weighted average / stacked ensemble models of Random Forest, Support Vector Machine, K-Nearest Neighbours and Neural Network with more than 98% accuracy. Moreover, in the case of feature selection and under-sampling, Decision Tree + Random Forest and SVM + Naïve Bayes are also found to be suitable models with high performance (both with 100% accuracy).

CHAPTER 1
INTRODUCTION

According to **Cancer Prevention and Control** (CPC) breast cancer is the second most common type of cancer after lung cancer and the fifth most common cause of cancer death. The National **Breast Cancer Foundation** has valued around 200000 new breast cancer cases and 40000 deaths every year in women while in men, these statistics are 1700 and 450, respectively.

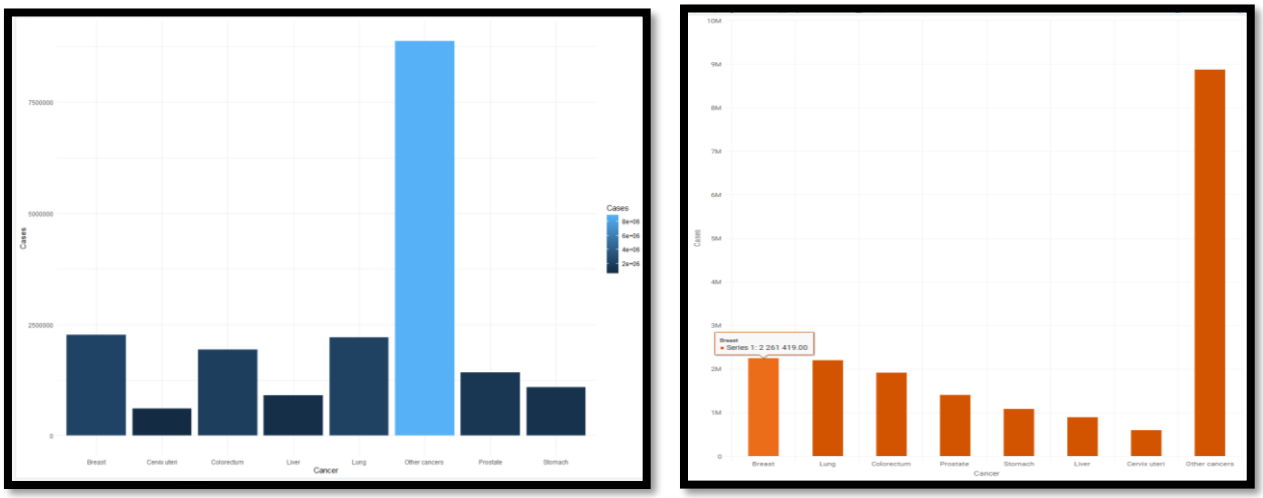

Figure 1: Frequency Plots for Cancer Cases

Among these, cancer is a term for a class of diseases characterized by abnormal cell (which lose the ability of dividing) that grow and invade healthy cells in the body. Breast cancer starts in the cells of the breast as a group of cancer cells that can spread to the surrounding tissues and to other areas of the body. It occurs in both men and women, even if male breast cancer is rare. Recent statistics show that breast cancer is a serious disease with high incidence rate and one of the leading causes of early mortality of women. Given such conditions, early diagnosis of breast cancer is considered vigorous, because statistics have shown a five-year survival rate of 96% for those whose cancer was discovered in the early stages. At the early stage, detection can be performed using the following two strategies:

A. Early Diagnosis: Using effective treatments to reduce the risk of cancer by escalating the identification proportion at the early stage.

B. Screening: Rectifying the presence of cancerous cells before the symptoms can showcase using various tools such as:

i. Breast Self - Exam (BSE)
ii. Clinical Breast Exam (CBE)
iii. Mammography

In terms of medical diagnosis and screening techniques, X-ray mammography is currently the most common technique used in clinical practice due to its low cost and accessibility. To improve the accuracy and efficiency of mammogram examination, computer-aided detection is focused on the identification of the location of suspect regions while computer-aided diagnosis is targeted to characterization (i.e., malignancy versus benignity).

Invasive ductal carcinoma (IDC) is the most common form of breast cancer. About 80% of all breast cancers are invasive ductal carcinomas. Doctors often do the biopsy or a scan if they detect signs of IDC. The cost of testing for breast cancer sets one back with $5000, which is a very big amount for poor families and manual identification of presence and extent of breast cancer by a pathologist is critical. Therefore, automation of detection of breast cancer reduce cost and time as well as improve the accuracy of the test. This research project aims to probe into the possibility of detecting and classifying breast cancer from the Breast Cancer Wisconsin (Diagnostic) Data.

CHAPTER 2
EXPLORATORY DATA ANALYSIS

ABOUT DATA

i. The Breast Cancer Classification data has been extracted from Wisconsin Dataset.
ii. The data contain "diagnosis" as the dependent variable and rest 31 numerical features as independent ones.
iii. Within the attribute "diagnosis", category 1 represents "Malignant" (Abnormal) while 0 reflects "Benign" (Normal).
iv. The dimensions of the data are 567x32 which has been divided (0.80:0.20) into training and testing data with shapes (455, 32) and (112, 32), respectively.

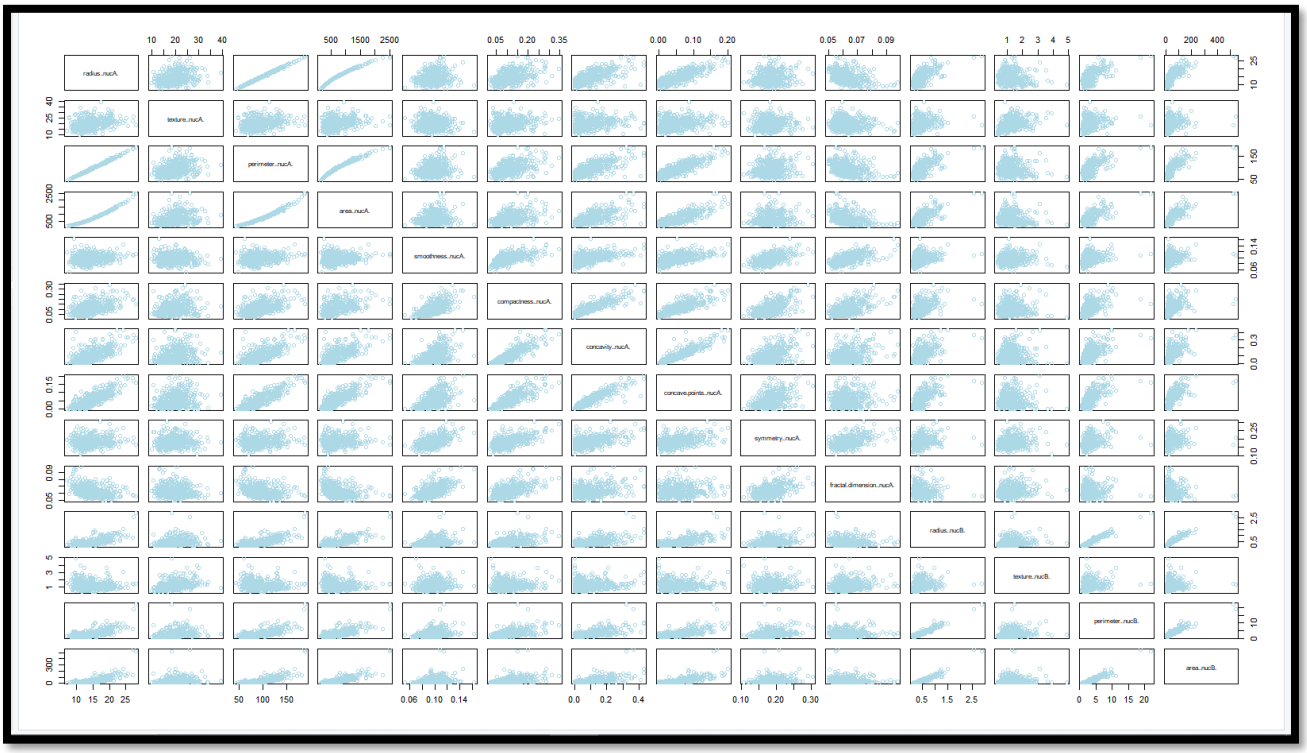

Figure 2: Data scatter plot

DATA PRE-PROCESSING

Outliers

The points with expected value of response variable for predictor variables is known as an Outlier. Since outliers show an expected statistic, they tend to disrupt the measurement errors due to significant deviations. Thus, it becomes to remove these points for better model fitting. Using a boxplot, the outliers can be easily be examined as can be seen from the plot with features nucA, nucB and nucC containing high number of outliers as shown below.

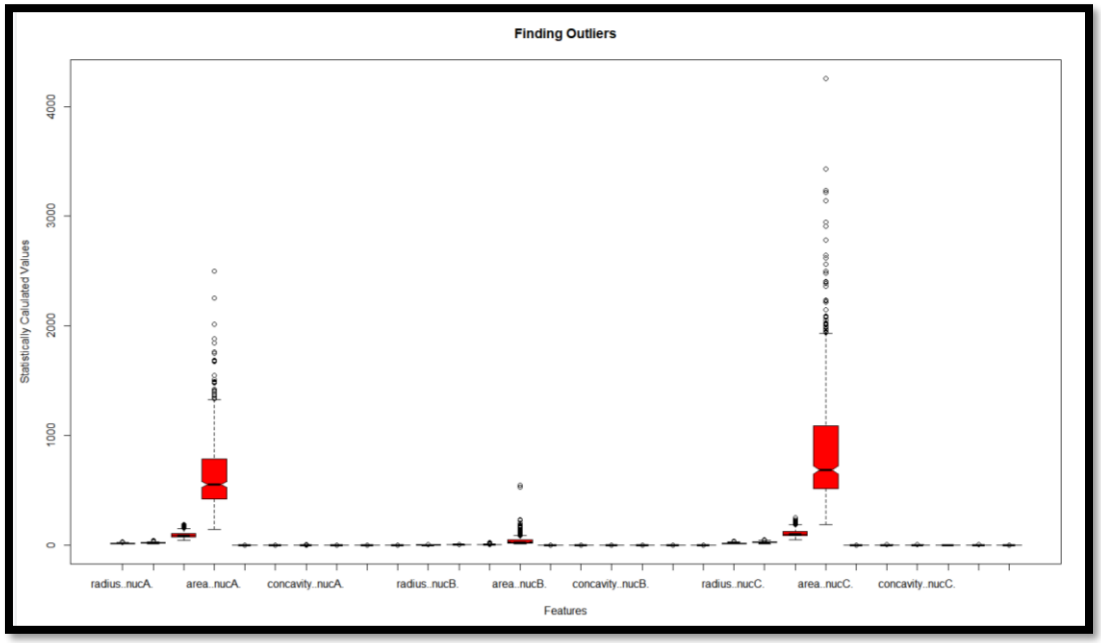

Figure 3: Visualize outliers in the data using boxplot

Using functions in RStudio, the extreme values are removed from the dataset as can be witnessed with reduced plotting scale in below boxplot.

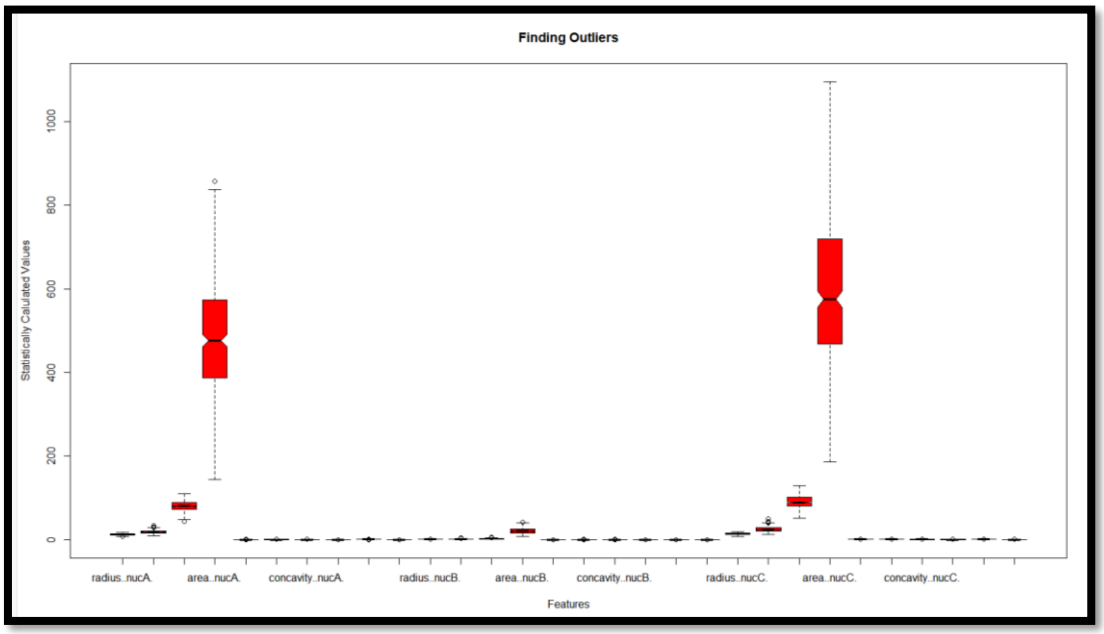

Figure 4: Visualize data using boxplot after removing outliers

Using an additional function in the notebooks (3/ 3.1/ 3.2) with feature selection, the outliers have been also been removed correspondingly.

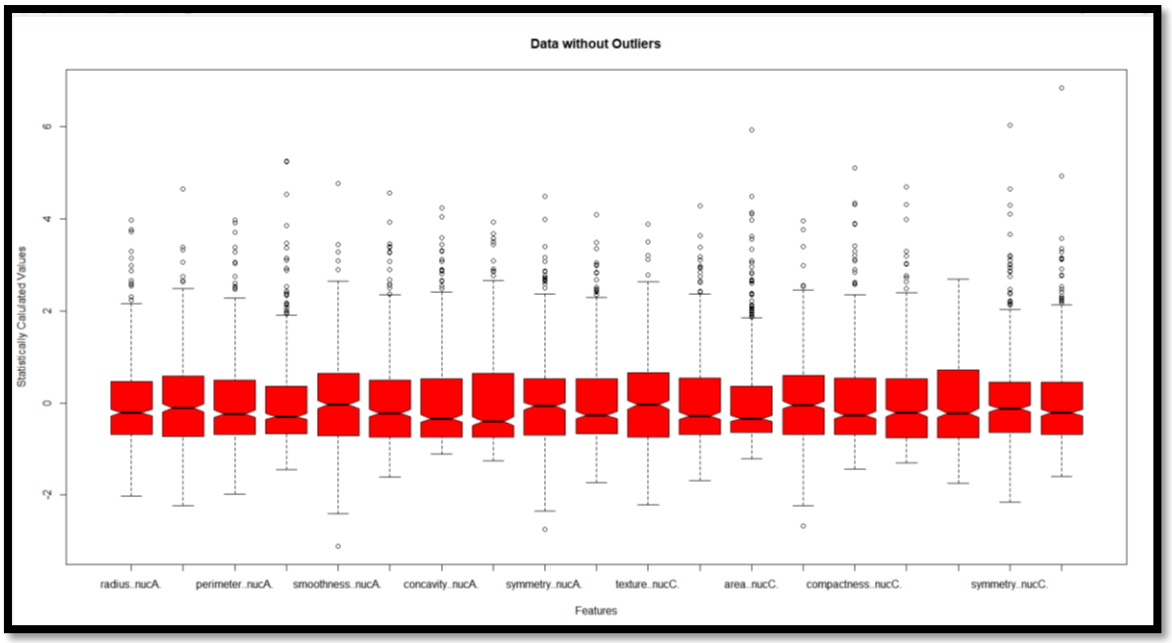

Figure 5: Boxplot for feature selected model

Standardization

Statistically, rescaling the data with mean and standard deviation to retain the internal data consistency is termed as Standardizing data. The approach is subtracting the data of respective features with their mean upon the column standard deviation.

Most of the time, we will encounter data with a wider range of fluctuating values which get distorted more when we perform statistical evaluation. Thus, the benefit is to rescale or compress the data relative to mean and make the data more related. Z-score is known to be a common approach to show standardization and has been used in the project.

```
### Standardizing the Data

```{r}
tail(data[c(-1)])
```
```

	radius..nucA. <dbl>	texture..nucA. <dbl>	perimeter..nucA. <dbl>	area..nucA. <dbl>	smoothness..nucA. <dbl>	compactness..nucA. <dbl>
559	14.59	22.68	96.39	657.1	0.08473	0.13300
560	11.51	23.93	74.52	403.5	0.09261	0.10210
561	14.05	27.15	91.38	600.4	0.09929	0.11260
562	11.20	29.37	70.67	386.0	0.07449	0.03558
563	15.22	30.62	103.40	716.9	0.10480	0.20870
569	7.76	24.54	47.92	181.0	0.05263	0.04362

6 rows | 1-7 of 30 columns

```
```{r}
data[c(-1)] = as.data.frame(scale(data[c(-1)]))
tail(data[c(-1)])
```
```

	radius..nucA. <dbl>	texture..nucA. <dbl>	perimeter..nucA. <dbl>	area..nucA. <dbl>	smoothness..nucA. <dbl>	compactness..nucA. <dbl>
559	1.2074246	1.044616	1.3320394	1.2613611	-0.66118207	1.0532034
560	-0.4776464	1.350437	-0.4341414	-0.5578477	-0.09888665	0.3240996
561	0.9119901	2.138231	0.9274411	0.8546216	0.37778004	0.5718533
562	-0.6472477	2.681368	-0.7450602	-0.6833846	-1.39188070	-1.2454793
563	1.5520982	2.987189	1.8981540	1.6903386	0.77095869	2.8393898
569	-2.5292750	1.499677	-2.5823081	-2.1539596	-2.95175101	-1.0557707

6 rows | 1-7 of 30 columns

Figure 6: Performing standardization of data

Resampling

Due to less intellectual about the variance of the data, it becomes difficult to estimate the parameters for the population. Thus, estimating the parameters multiples times is the approach with resampling. Statistically, estimating the class distribution can be used to find variance of the estimated parameters accurately with resampling.

There are 2 commonly known methods of resampling:

i. Bootstrap: building a balanced class distribution by either decreasing the majority class, increasing (duplicating) the minority class or by integrating both the approaches.

ii. k-fold Cross-Validation: to estimate test error, (specially with small data), cross validation makes k groups where the training is done a group of features while the testing is performed on the remaining portion.

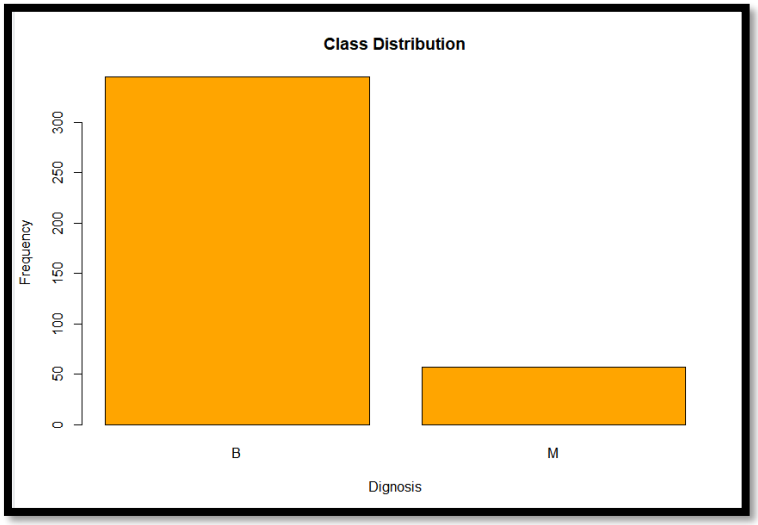

Figure 7: Imbalanced Class Distribution

From the above class distribution, we can see that the ratio of class is unequal which means when we will train the model with this unsampled data, the model will be biased towards benign class with less generalization. Thus, to generalize the model with unbiased class distribution we must perform resampling.

For instance, we will consider the following methods, ‘under’, ‘over’ or ‘both’ sampling:

i. Over-sampling: a statistical technique designed an unbiased class distribution from an imbalanced class distribution with classification dataset. This technique removes examples from the training dataset that belong to the minority class to a better balance the class distribution of ~50.9% Malign and ~49.1% Benign.

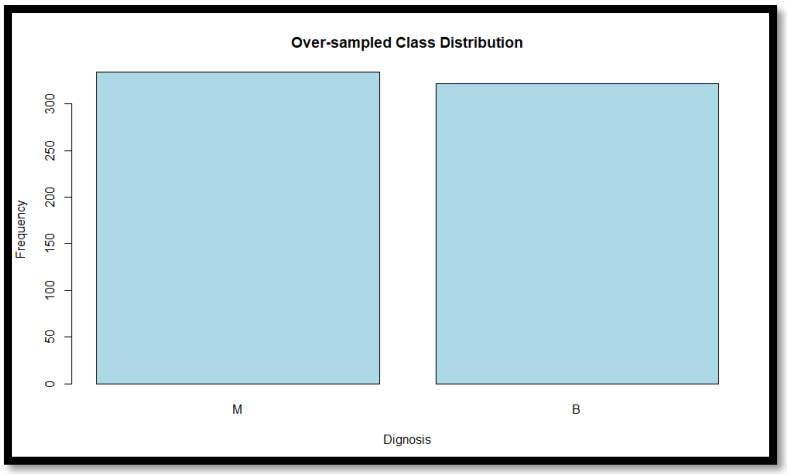

Figure 8: Over-sampled Class Distribution

ii. Under-sampling: a statistical technique designed an unbiased class distribution from a skewed class distribution with classification dataset. Under-sampling techniques remove examples from the training dataset that belong to the majority class to a better balance the class distribution, such as reducing the skew from a 1:100 to a 1:10, 1:2, or even a 1:1 class distribution. This is different from oversampling that involves adding examples to the minority class to reduce the skew in the class distribution.

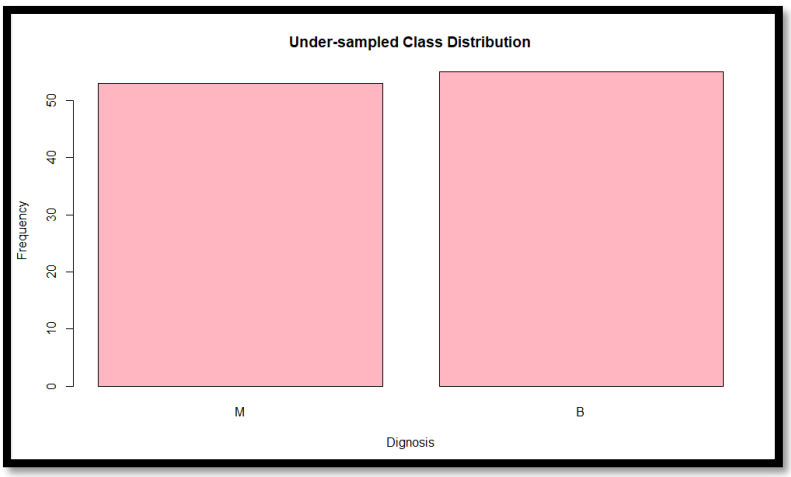

Figure 9: Under-sampled Class Distribution

iii. Both-sampling: this technique is based on the blend of under and over-sampling and was found to contain ~50.9% Malign and ~49.1% Benign.

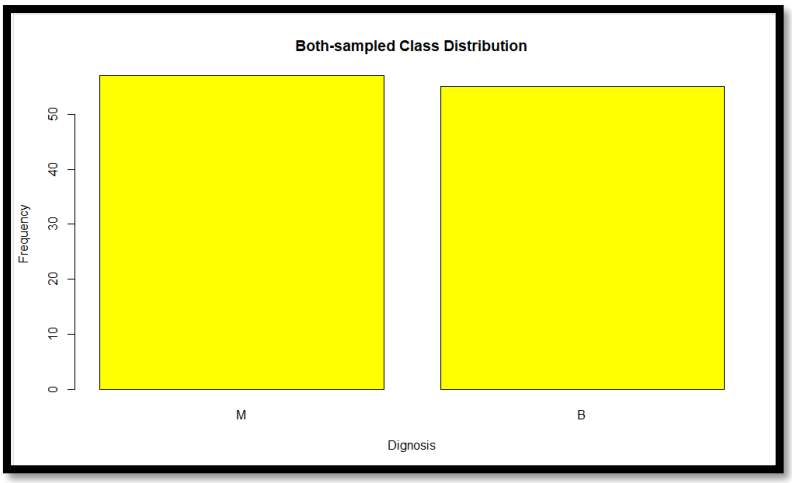

Figure 10: Under-sampled and Over-sampled Class Distribution

From the modelling, it was found that under-sampling and over-sampling have found to be an optimized option with different models thus, being a crucial component, both have been discussed in the project.

Features Selection

Finding correlation values between the features of the data helps to understand the degree of correlation between the data attributes. For the model, in this approach only highly correlated features have been selected to be trained and modelled.

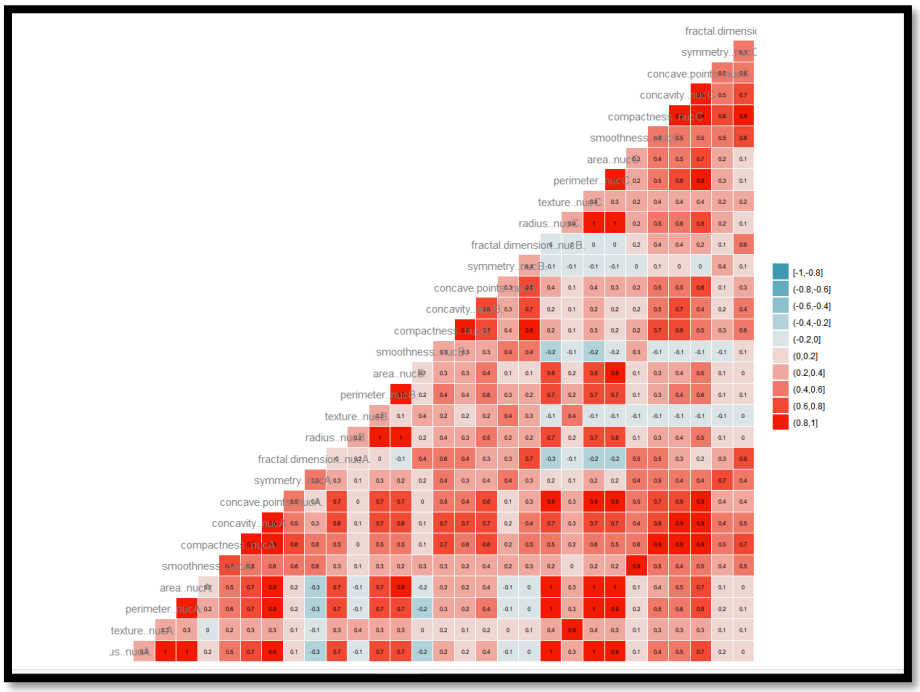

Figure 11: Data correlation plot

A strong correlation i.e. [0.8, 1] is shown by dark red blocks while as we move to dark sky-blue blocks (lowest correlation), the strength of relationship between the data attributes

decreases. This correlation is also useful to fetch out only highly correlated features and build a classification model.

CHAPTER 3
LITERATURE SURVEY

ALGORITHMS

Over the course of the project, the following machine learning traditional and hybrid/ ensemble classification algorithms have been used to decipher and evaluate the model on different scenarios with an attempt to have best fitted models:

1. Naive Bayes (NB)

Naive Bayes is a probabilistic classifier which applies Bayes' theorem with strong independent assumptions. In this model, all properties are separately taken into consideration to find any relationship between them. It assumes that predictive attributes are conditionally independent given the class. Additionally, the values of the numeric attributes are distributed within each class. Naive Bayes is fast and performs well even with a small dataset. However, it is hard to find independent properties in real life. Researchers have deployed NB classifiers for breast cancer detection and achieved the maximum accuracy with only five dominants.

2. Decision Tree (DT)

Decision Tree is a data mining technique used for early detection of breast cancer. It is a model that presents classification or regression as a tree. In this model, the dataset is split into small sub - data, then into smaller ones. As a result, the tree is developed and at the last level reveals the result. In a tree structure, the leaves characterise the class labels while the branches characterise conjunctions of features which lead to the class labels. Hence, DT is not sensitive to noise. Now, if we are dealing with a larger dataset, greater are the splits and such huge trees result in elevating the complexity and leads to overfitting. A simple approach to limit the splitting is by fixing the minimum training dataset at the input and the maximum depth of the decision-making tree. Another approach is by using the method of Pruning.

An effective technique to optimize the performance of Decision Tree is by simply getting rid of the attributes with lesser information or higher entropy. There are two approaches to execute this technique to either start from root or leaves. The bottom - up method in which child attributes having lesser significance are removed without altering the accuracy is known as Reduce Error Pruning.

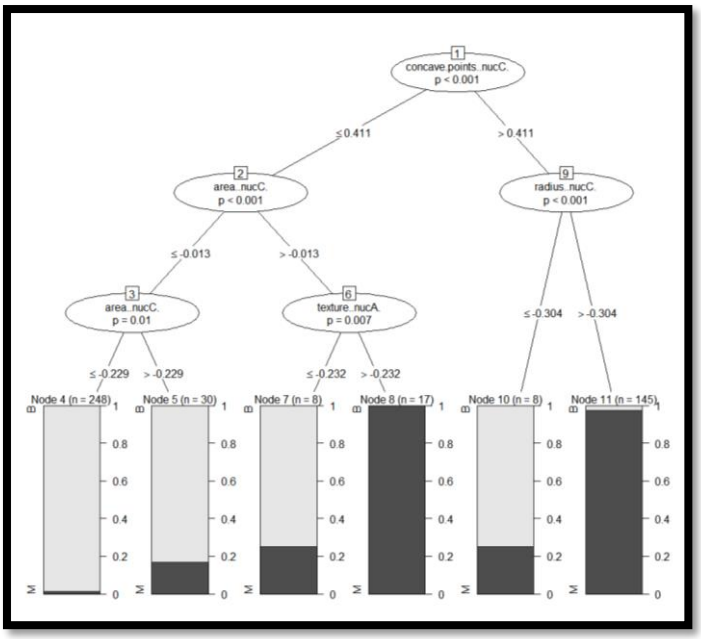

Figure 12: Decision Tree for a fitted model

3. Random Forest (RF)

This algorithm is an extension of Decision Tree, as being the building block. The name "Random Forest", is self - explanatory where "Forest" signifies averaging the prediction of the tree and "Random" explains two concepts:

i. For tree building: random sampling of training data points for building the tree.

ii. For splitting nodes: after sample data is fetched, a random subset of features is chosen for analysis to split the nodes, as the tree learns from the random sample and their features. For building a tree, multiple samples are used multiple times which is known as Bootstrapping.

It can also be defined as a random sampling of observational data with replacement. While training each tree, different samples used for building trees might have a higher variance, but the cumulative variance of the forest must be low and not at the cost of an increase in bias. The prediction is made by averaging the prediction of each Decision Tree and is termed as Bootstrap Aggregating or Bagging. It can also be said as a different randomized (bootstrapped) subset of data for making up the individual trees that are trained and then averaged for prediction. The contrast between Decision Tree and Random Forest is that, instead of shortlisting the result from a single tree, multiple trees are taken into consideration and evaluated for prediction. As in epitome, predicting a product in the market by the cumulative reviews for E - commerce websites, Google, and YouTube.

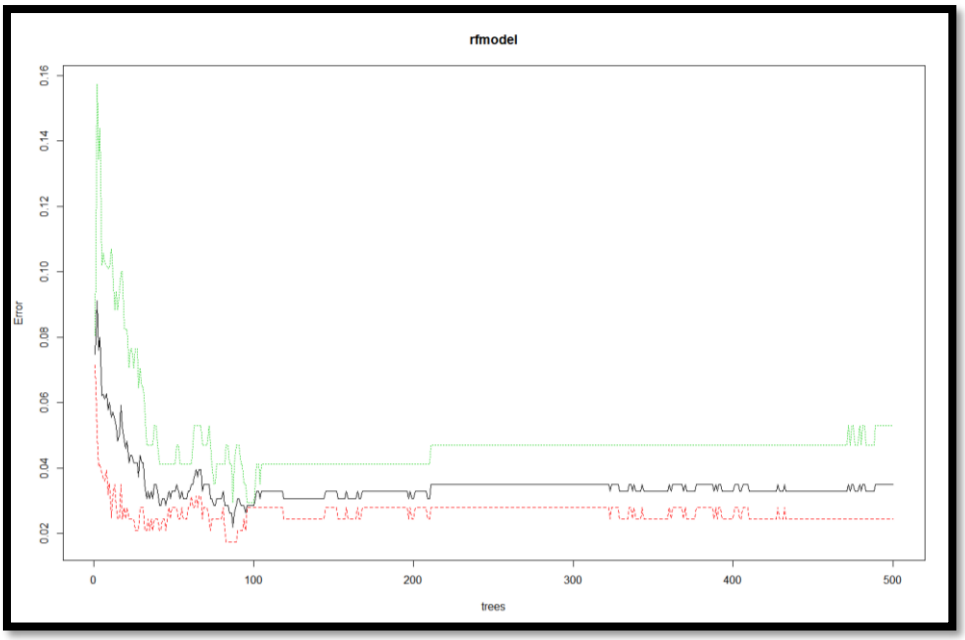

Figure 13: Error vs random tree plot for a fitted random tree model

4. K - Nearest Neighbours (KNN)

KNN is a supervised learning method. It can be used for diagnosing and classifying cancer. In this method, the model is trained in a specific field and new data is given to it. Furthermore, similar data is used by the machine for detecting "K". Therefore, the machine finds KNN for the unknown data. It is recommended to choose a large dataset for training also K value must be an odd number determining closeness.

For calculating the closeness, the distance between data observations is estimated and is known as Minkowski Distance. The distance calculated in a normed vector space i.e., space where the distance is observed in vectors is called a Minkowski Distance. "Normed " signifies that vectors have length, and no vector can have negative length.

5. Support Vector Machines (SVM)

SVM is a supervised machine learning algorithm that is a pattern classification model. It is used as a training algorithm for learning classification and regression rules from collected data. The motive of this method is to segregate data until a hyperplane with high minimum distance is found. Using SVM, we can classify two or more data types. SVM is an efficient method for diagnosing breast cancer. Its accuracy can increase when combined with feature selection. This accuracy can be obtained on the WBCD dataset, considering five features. SVM is the most efficient way of statistical learning. It can easily identify the decision boundaries between different classes of breast cancer. The input features are selected after calculating the F - score of each input feature. The significance of each feature is evaluated by it' s F - score. The SVM parameters are optimised by grid search.

It is called a Discriminative Classifier and learns about classification from given statistics depending on the observed data. It can also be called as a Discriminative Model or a Conditional Model for statistical classification in (Supervised) Machine Learning defined by

separating hyperplanes. In a multi - dimensional space for separating classes, labelled training data are used to form optimal hyperplanes (2 D) while 3 D hyperplanes are used for kernel transformations.

6. Neural Networks (NN)

ANN is a machine learning model like the human brain' s nerve system having a large number of interconnected nodes. Each node has two states: 0 represents the inactive state and 1 represents the active state. Furthermore, each node has either a positive or a negative weight that tunes the strength of that node and can activate or deactivate it. The machine is trained on samples of data using ANN. The trained machine is then used to detect the pattern of hidden data. It can search for patterns between patients' healthcare and personal records to identify high - risk lesions.

The advantages of using ANN are - They can learn and model non - linear and complex relationships. They can generalise. No restrictions are imposed on the input variables (like how they should be distributed). Its region growing - based segmentation method is improved by using the extracted intensity features from ROIs and applying the ANN to generate an adaptive threshold.

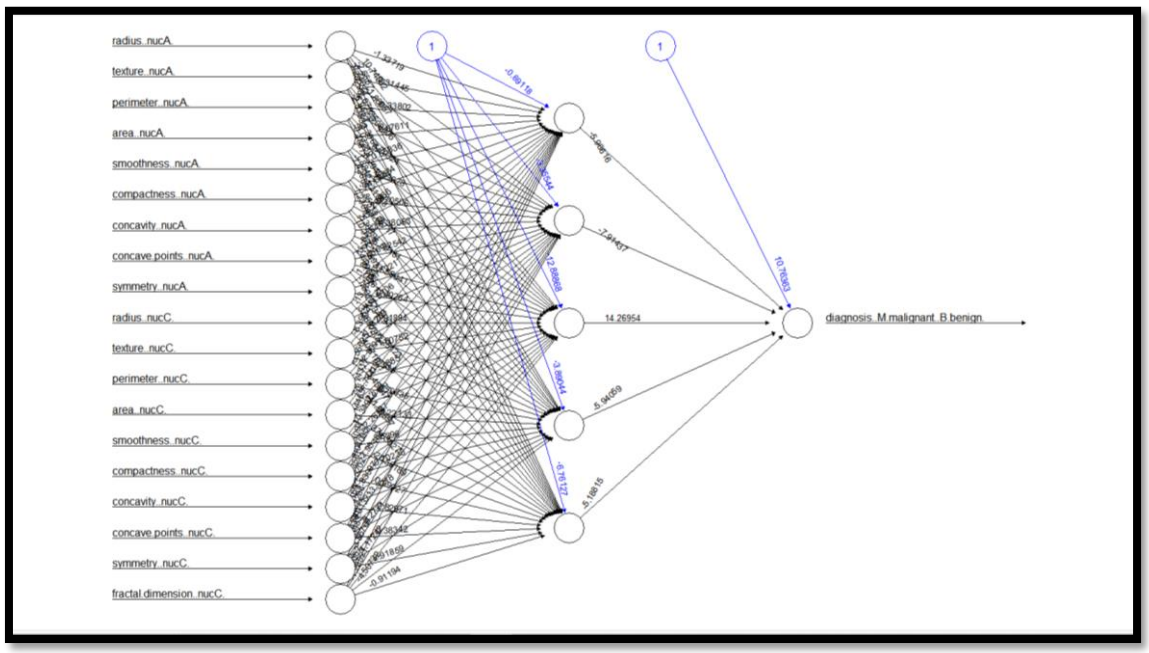

Figure 14: Neural network model 1 (used) architecture

7. Ensemble Models

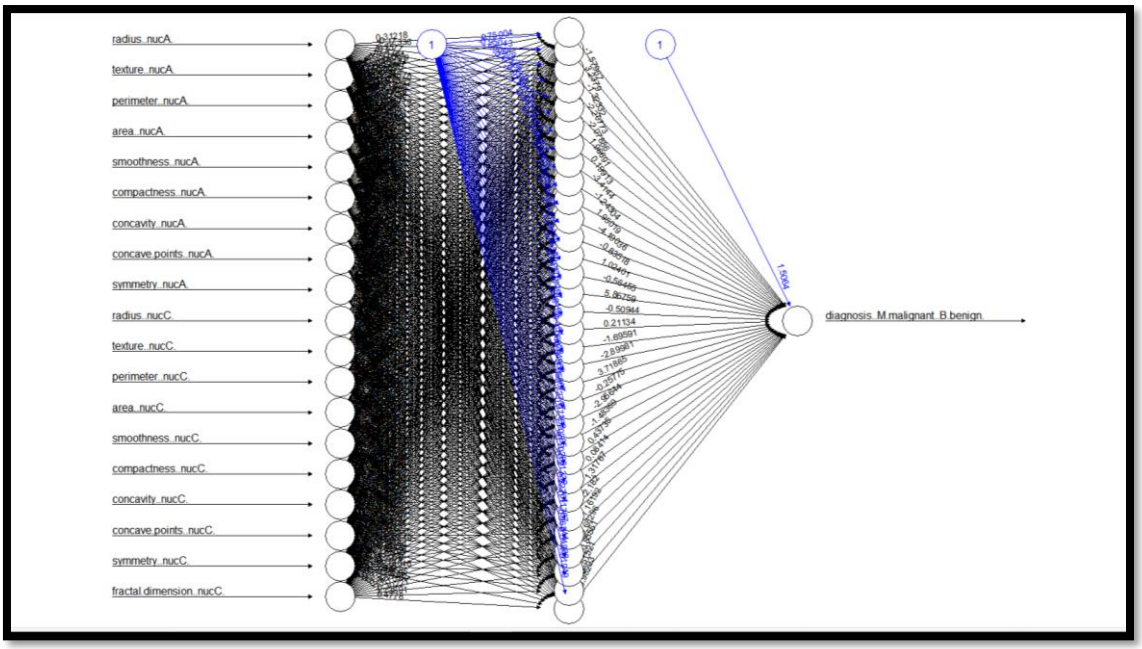

Figure 15: Neural network model 2 (used) architecture

To make the model more accurate, less biased and more robust, we integrate multiple algorithms with different in a approach to build an ensemble model with the technique of ensembling. Some types of ensemble modelling are as follow:

i. Averaging: taking average of the prediction from the considered models.

ii. Majority Voting: taking majority vote from the prediction of multiple the models.

iii. Weighted Average: taking weighted average (assigning weight to the predictions of various models) of the predictions from multiple models.

iv. Bagging: to reduce the variance, ‘n’ rows from the observation is taken and replace with an alternative row values which is then followed by using averaging or majority voting to build a tree model further.

v. Boosting: in this method, we try to reduce bias and overfitting by boosting the performance of the model. For this, we train the entire data with an initial algorithm and then later models are trained on the fitted residuals of the previous models. This technique helps in boosting the performance, specially used with multiple weak models to be making them share their specific well-trained portion of the data and gives a better statistic at the end.

CHAPTER 4
MODEL EVALUATIONS

MODEL EVALUATIONS WITHOUT RESAMPLING DATA

Traditional Models and Neural Networks

ALGORITHMS	WITHOUT RESAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	93.75%	97.06%	75.00%
Random Forest	97.50%	97.18%	100.00%
Support Vector Machine	98.75%	100.00%	91.67%
Naïve Bayes	96.25%	100.00%	78.57%
K-Nearest Neighbour	96.25%	95.83%	100.00%
Neural Network: Model 1	95.00%	98.51%	76.92%
Neural Network: Model 2	96.52%	100.00%	78.57%

Table 1: Traditional models statistics without resampling

Hybrid Models

HYBRID MODELS	WITHOUT RESAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	96.25%	100.00%	72.73%
DT + SVM	96.25%	98.55%	81.82%
RF + SVM	97.50%	100.00%	81.82%
RF + NB	97.50%	100.00%	81.82%
RF + KNN	97.50%	100.00%	81.82%
RF + NN	97.50%	100.00%	81.82%
SVM + NB	96.25%	98.55%	81.82%
SVM + KNN	98.75%	98.55%	100.00%
SVM + NN	98.75	98.55%	100.00%
NB + NN	96.25%	95.65%	100.00%
RF + SVM + KNN	98.75%	98.55%	100.00%
RF + SVM + NN	98.75%	98.55%	100.00%
Stacked Ensemble: RF + SVM -> NN	97.50%	100.00%	81.82%

Table 2: Hybrid/ Ensemble models statistics without resampling

MODEL EVALUATIONS WITH RESAMPLING (Over-Sampling) DATA

Traditional Models and Neural Networks

ALGORITHMS	OVER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	94.12%	96.92%	91.55%
Random Forest	98.53%	100.00%	95.71%
Support Vector Machine	95.59%	94.37%	96.92%
Naïve Bayes	90.44%	91.18%	89.71%
K-Nearest Neighbour	98.75%	98.75%	100.00%
Neural Network: Model 1	95.59%	97.01%	94.20%
Neural Network: Model 2	97.06%	97.10%	97.01%

Table 3: Traditional models statistics with over-sampling

Hybrid Models

HYBRID MODELS	OVER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	96.32%	95.65%	97.01%
DT + SVM	96.32%	98.55%	94.03%
RF + SVM	96.32%	98.55%	94.03%
RF + NB	94.12%	97.10%	91.04%
RF + KNN	98.53%	95.65%	100.00%
RF + NN	97.79%	95.65%	100.00%
SVM + NB	92.65%	94.20%	91.04%
SVM + KNN	95.59%	97.10%	94.03%
SVM + NN	95.59%	97.10%	94.03%
NB + NN	90.44%	89.86%	91.04%
RF + SVM + KNN	98.53%	94.02%	100.00%
RF + SVM + NN	97.06%	94.02%	100.00%
Stacked Ensemble: RF + SVM -> NN	98.53%	100.00%	95.52%

Table 4: Hybrid/ Ensemble model statistics with over-sampling

MODEL EVALUATIONS WITH RESAMPLING (Under-Sampling) DATA

Traditional Models and Neural Networks

ALGORITHMS	UNDER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	92.50%	94.37%	77.78%
Random Forest	96.25%	95.83%	100.00%
Support Vector Machine	98.75%	98.57%	100.00%
Naïve Bayes	95.00%	100.00%	73.33%
K-Nearest Neighbour	95.59%	100.00%	91.78%
Neural Network: Model 1	97.50%	100.00%	84.62%
Neural Network: Model 2	96.25%	100.00%	78.57%

Table 5: Traditional models statistics with over-sampling

Hybrid Models

HYBRID MODELS	UNDER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	95.00%	100.00%	63.64%
DT + SVM	95.00%	100.00%	63.64%
RF + SVM	96.25%	100.00%	72.73%
RF + NB	96.25%	100.00%	72.73%
RF + NN	96.25%	100.00%	72.73%
SVM + NB	95.00%	100.00%	63.64%
SVM + NN	98.75%	100.00%	90.91%
NB + NN	95.00%	94.20%	100.00%
RF + SVM + NN	98.75%	100.00%	90.91%
Stacked Ensemble: RF + SVM -> NN	98.53%	100.00%	61.54%

Table 6: Hybrid/ Ensemble model statistics with under-sampling

WITH FEATURE SECLECTION

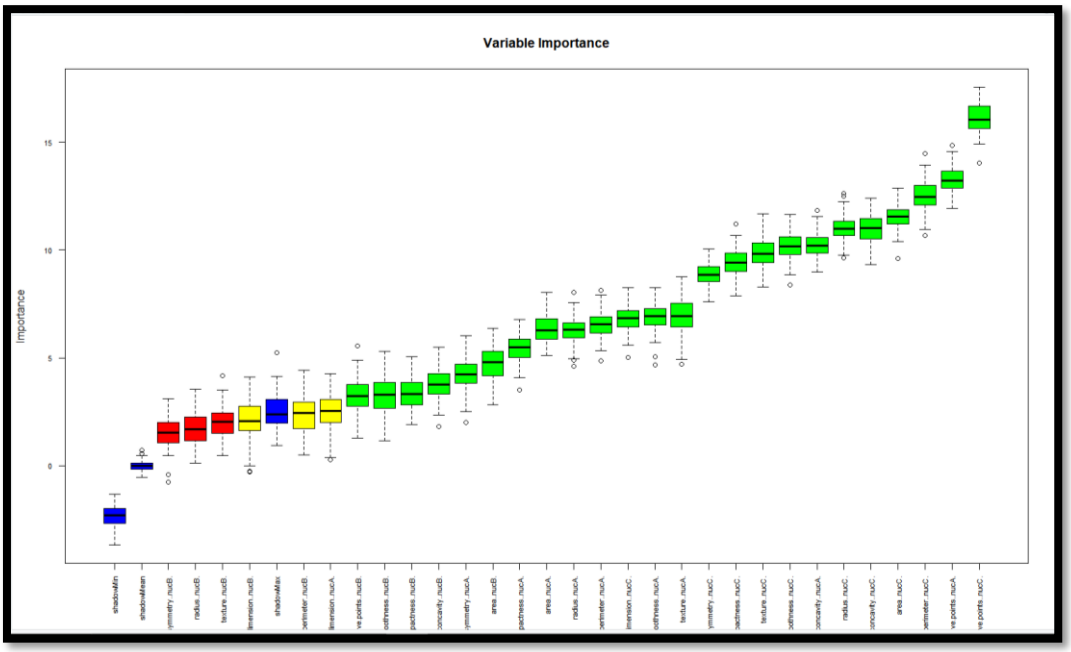

Figure 16: Feature selection significance Plot

In machine learning, the basic approach is to train the model from the available data with maximum generalization so that the model can be used to work on tons of unsampled and untrained data as can be seen in Notebook 2. But if, we can train the model with lesser features as available from the data, we opt to train the model from those selected features which have similar or higher generalization compared to a complete set of data variables. Thus, feature selection plays an important role in machine learning as can be seen in Notebook3, 3.1 and 3.2.

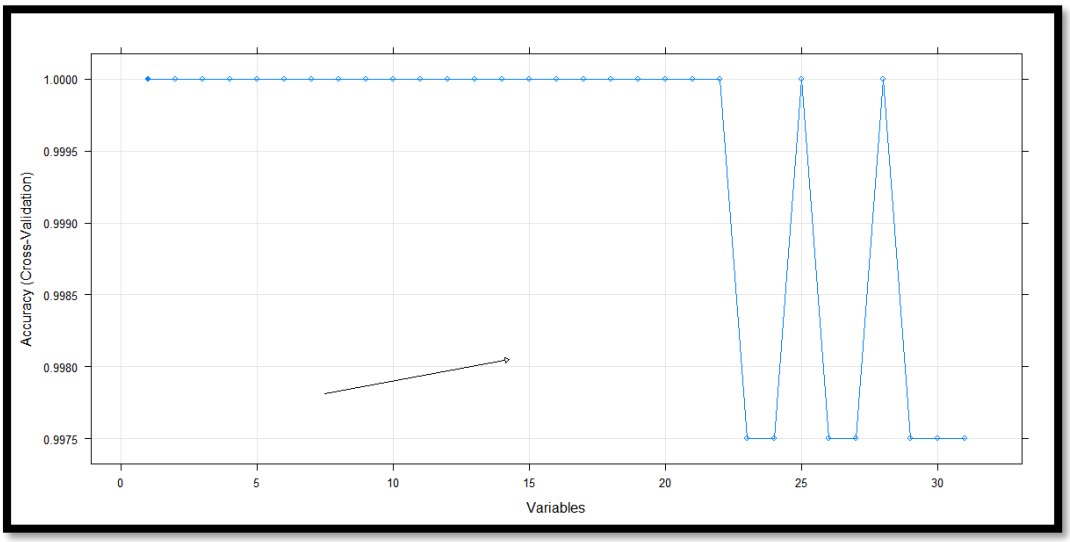

Figure 17: Selected features cross-validation plot

As can be seen the below plot, since these features have higher correlation compared all the available features, a classification model is prepared and compared with the others prepared model.

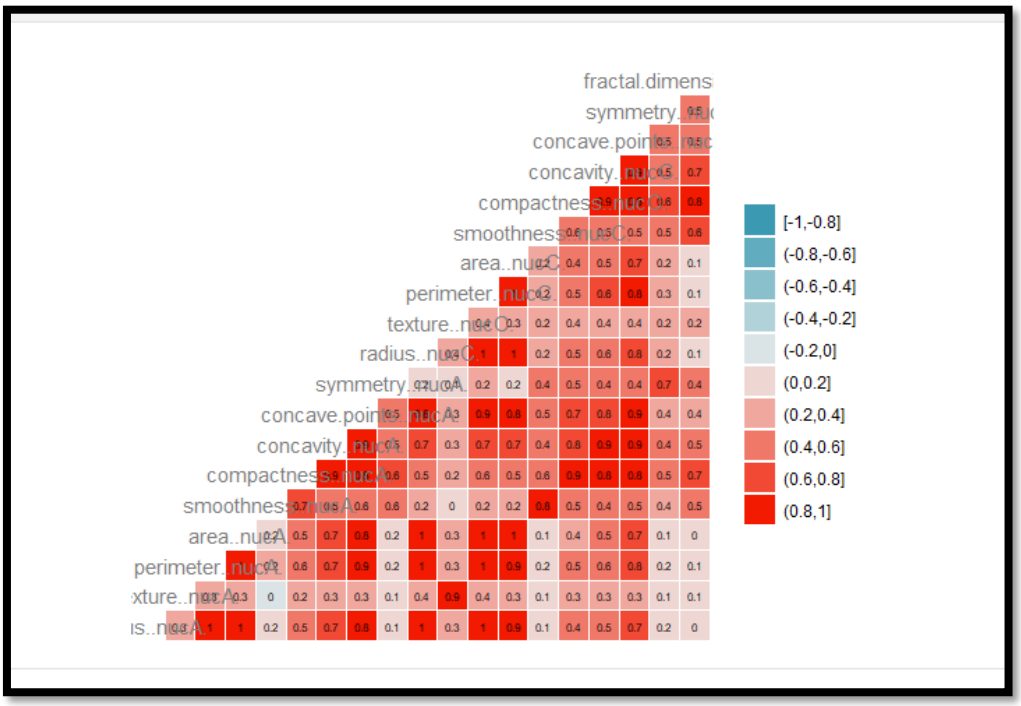

Figure 18: Data correlated plot with selected features

MODEL EVALUATIONS WITHOUT RESAMPLING DATA

Traditional Models and Neural Networks

ALGORITHMS	FEATURE SELECTION WITHOUT RESAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	92.04%	91.89%	92.31%
Random Forest	97.35%	97.22%	97.56%
Support Vector Machine	98.23%	97.26%	100.00%
Naïve Bayes	95.58%	94.59%	97.44%
K-Nearest Neighbour	97.50%	97.18%	100.00%
Neural Network: Model 1	96.46%	100.00%	91.32%
Neural Network: Model 2	97.35%	97.22%	97.56%

Table 7: Traditional models statistics with feature selection without resampling

Hybrid Models

HYBRID MODELS	FEATURE SELECTION WITHOUT RESAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	94.69%	100.00%	85.71%
DT + SVM	92.92%	100.00%	80.95%
RF + SVM	96.46%	100.00%	90.48%
RF + NB	94.69%	98.59%	88.10%
RF + NN	97.37%	98.59%	95.24%
SVM + NB	92.04%	100.00%	78.75%
SVM + NN	98.23%	100.00%	95.24%
NB + NN	95.58%	98.59%	90.48%
RF + SVM + NN	99.12%	98.59%	100.00%
Stacked Ensemble: RF + SVM -> NN	98.53%	100.00%	95.12%

Table 8: Hybrid/ Ensemble models statistics with feature selection without resampling

MODEL EVALUATIONS WITH RESAMPLING (Under-Sampling) DATA

Traditional Models and Neural Networks

ALGORITHMS	FEATURE SELECTION WITH UNDER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	98.78%	100.00%	97.67%
Random Forest	100.00%	100.00%	100.00%
Support Vector Machine	98.08%	100.00%	96.43%
Naïve Bayes	97.56%	100.00%	95.45%
K-Nearest Neighbour	99.12%	98.16%	100.00%
Neural Network: Model 1	98.08%	100.00%	96.43%
Neural Network: Model 2	99.12%	100.00%	98.18%

Table 9: Traditional models statistics with feature selection and under-sampling

Hybrid Models

HYBRID MODELS	FEATURE SELECTION WITH UNDER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	100.00%	100.00%	100.00%
DT + SVM	98.78%	97.56%	100.00%
RF + SVM	97.56%	97.50%	97.62%
RF + NB	98.78%	97.50%	100.00%
RF + NN	98.78%	97.50%	100.00%
SVM + NB	100.00%	100.00%	100.00%
SVM + NN	98.23%	97.53%	97.62%
NB + NN	97.56%	97.50%	100.00%
RF + SVM + NN	97.56%	97.50%	100.00%
Stacked Ensemble: RF + SVM -> NN	100.00%	100.00%	100.00%
Stacked Ensemble: RF + DT -> NN	100.00%	100.00%	100.00%

Table 10: Hybrid/ Ensemble models statistics with feature selection and under-sampling

MODEL EVALUATIONS WITH RESAMPLING (Over-Sampling) DATA

Traditional Models and Neural Networks

ALGORITHMS	FEATURE SELECTION WITH OVER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
Decision Tree	94.33%	95.32%	96.67%
Random Forest	98.58%	98.59%	98.57%
Support Vector Machine	99.29%	100.00%	98.59%
Naïve Bayes	92.20%	91.67%	92.75%
K-Nearest Neighbour	99.12%	98.16%	100.00%
Neural Network: Model 1	97.87%	100.00%	95.89%
Neural Network: Model 2	98.58%	100.00%	97.22%

Table 11: Traditional models statistics with feature selection and over-sampling

Hybrid Models

HYBRID MODELS	FEATURE SELECTION WITH OVER-SAMPLING		
	ACCURACY	SENSITIVITY	SPECIFICITY
DT + RF	96.45%	98.59%	94.29%
DT + SVM	96.45%	98.59%	94.29%
RF + SVM	99.29%	100.00%	98.57%
RF + NB	98.58%	98.59%	98.57%
RF + NN	98.87%	98.59%	97.14%
SVM + NB	99.29%	98.59%	100.00%
SVM + NN	97.16%	97.18%	97.14%
NB + NN	92.20%	91.55%	92.86%
RF + SVM + NN	98.58%	100.00%	97.18%
Stacked Ensemble: RF + SVM -> NN	98.58%	100.00%	97.18%

Table 12: Hybrid/ Ensemble models statistics with feature selection and over-sampling

CHAPTER 5
RESULTS AND CONCLUSION

From an extensive model's evaluations, it has been found that without feature selection, over-sampling has given statistically better traditional and hybrid/ ensemble models compared to under-sampling models. Within this domain, Random Forest (RF) has 97.79%, K-Nearest Neighbours (KNN) has 98.75%, Neural Networks (NN) has 97.06% model accuracy, and their hybrid/ ensemble models has also performed with impressive evaluations. And with under-sampling, SVM has performed exceptionally well as a traditional model with 98.75% accuracy and a hybrid model of Random Forest and Neural Network with 98.75% accuracy.

Moreover, when the models were built with feature selection i.e., the models with only highly correlated features, SVM with over-sampling has performed with 99.29% accuracy while the all the under-sampling models has performed the best relative to all the models created. Within traditional model with feature selection and under-sampling, Random Forest has 100%, Neural Network and KNN has 99.12%, Decision Tree has 98.78% and SVM has 98.08% model accuracy. Furthermore, the averaged/ weighted average ensemble models with feature selection and under-sampling i.e., the model's DT + RF, SVM + NB, DT + SVM, RF + NB and RF + NN have shown a great performance while within the stacked ensemble models, both RF + DT -> NN (100.00% accuracy) and RF + SVM -> NN (98.78% accuracy) has shown an impressive statistic.

In conclusion, the project clearly explains the importance of removing the outliers, standardizing/ normalizing, resampling the data with imbalance class distribution and feature selection. From the complete set of models built, all the traditional models and most of the ensemble models (i.e., average, weighted average and stacked) with over-sampling have shown the best model evaluations.

Moreover, Image classification can be more accurate and effective in case of Breast Cancer Classification when integrated with the Data Resampling, Data Augmentation, Transfer Learning and building hybrid models.

REFERENCES

[1]. Myers ER, Moorman P, Gierisch JM, et al. Benefits and harms of breast cancer screening: a systematic review. JAMA. 2015;314:1615\[Dash]1634. [PubMed] [Google Scholar]

[2]. Group DES. Systematic review of cancer screening literature for updating American Cancer Society breast cancer screening guidelines. Duke Clinical Research Institute, Durham, NC: Guidelines Development Group. 2014 [Google Scholar]

[3]. Swedish Council on Health Technology Assessment. Computer-Aided Detection (CAD) in mammography screening. Stockholm: 2011. SBU systematic review summaries. [Google Scholar]

[4]. Mehdi Habibzadeh Motlagh, Mahboobeh Jannesari, HamidReza Aboulkheyr, Pegah Khosravi, Olivier Elemento, Mehdi Totonchi, and Iman Hajirasouliha

[5]. Babak Ehteshami Bejnordi, Guido Zuidhof, Maschenka Balkenhol, Meyke Hermsen, Peter Bult, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak

[6]. Alexander Rakhlin, Alexey Shvets, Vladimir Iglovikov and Alexandr A. Kalinin

[7]. Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrz¸ebsk, Thibault Févrb, Joe Katsnelson, Eric Kim, Stacey Wolfson, Ujas Parikha, Sushma Gaddama, Leng Leng Young Lina, Kara Hoj* , Joshua D. Weinsteina, Beatriu Reig, Yiming Gao, Hildegard Totha, Kristine Pysarenkoa, Alana Lewina, Jiyon Leea, Krystal Airolaa, Eralda Memaa, Stephanie Chung, Esther Hwang, Naziya Samreena, S. Gene Kima, , Laura Heacocka, Linda Moya, Kyunghyun Chob, and Krzysztof J. Gerasa

[8]. Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S. Gene Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, Kyunghyun Cho

[9]. Quinlan JR, Rivest RL. Inferring decision trees using the minimum description length principle. Information and Computation. 1989 March 1;80(3):227\[Dash]48.

[10]. Quinlan JR. Simplifying decision trees. International Journal of Man-Machine Studies. 1987 September 1;27(3):221\[Dash]34.

[11]. Mehta M, Rissanen J, Agrawal R. MDL-based decision tree pruning. KDD. 1995 August 20;21(2):216\[Dash]221.

[12]. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu and Philip S. Yu, et al., "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14, No. 1, 1-37, DOI: 10.1007/s10115-007- 0114-2.

[13]. Hang Yang, Fong, S, "Optimized very fast decision tree with balanced classification accuracy and compact tree size," Data Mining and Intelligent Information Technology Applications (ICMiA), 2011 3rd International Conference on, Pp.57-64, 24-26 Oct. 2011.

[14]. Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017). "A framework for sensitivity analysis of decision trees". Central European Journal of Operations Research. 26 (1): 135\[Dash]159. doi:10.1007/s10100-017-0479-6. PMC 5767274. PMID 29375266.

[15]. ^ Quinlan, J. R. (1987). "Simplifying decision trees". International Journal of Man-Machine Studies. 27 (3): 221\[Dash]234. CiteSeerX 10.1.1.18.4267. doi:10.1016/S0020-7373(87)80053-6.

[16]. Ho, Tin Kam (1995). Random Decision Forests (PDF). Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14\[Dash]16 August 1995. pp. 278\[Dash]282. Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.

[17]. ^ Jump up to:a b c d Ho TK (1998). "The Random Subspace Method for Constructing Decision Forests" (PDF). IEEE Transactions on Pattern Analysis and Machine Intelligence. 20 (8): 832\[Dash]844. doi:10.1109/34.709601.

[18]. https://www.analyticsvidhya.com/blog/2017/02/introduction-to-ensembling-along-with-implementation-in-r/

[19]. https://gco.iarc.fr/#cancer-overtime

[20]. https://www.who.int/health-topics/cancer#tab=tab_1

[21]. https://machinelearningmastery.com

[22]. https://www.analyticsvidhya.com/blog/2017/02/introduction-to-ensembling-along-with-implementation-in-r/