Contents

Acknowledgement

We would like to acknowledge and give our warmest thanks to our supervisor **Y N V Krishnya Chaitanya** who made this work possible. His guidance and advice carried us through all the stage of writing our project. We would like to thanks prof Atanu Kumar Ghosh for his invaluable support and contributions to this project. We would also like to thank few of our seniors and friends for their support in completion of this project. I would also like to thanks prof. Biswajit Roy, as our HOD, because of whome we got the opportunity to work on this project. Finally, we would like to thanks our parents and family as a whole for their continuous support and understanding when undertaking and writing my project work.

Abstract

Breast cancer is a prevalent form of cancer among women worldwide, and early detection significantly improves treatment outcomes. Multivariate analysis techniques offer a promising avenue for enhancing the accuracy and efficiency of breast cancer detection. This project aims to develop a robust breast cancer detection system leveraging multivariate analysis methods. The proposed system will integrate various multivariate analysis techniques such as Random Forest, Support Vector Machines (SVM) and K Nreaest Neighbour to analyze the data. The utilization of multivariate analysis will enable the identification of significant patterns and biomarkers associated with breast cancer across multiple data dimensions. The outcomes of this project are expected to contribute to the advancement of breast cancer detection methods, providing clinicians with a reliable tool for early diagnosis and personalized treatment strategies. Moreover, the insights gained from multivariate analysis may lead to the discovery of novel biomarkers and pathways associated with breast cancer progression, facilitating further research in this critical area of oncology.

Historical Background of Breast Cance

In **Egypt** It is found that the earliest known descriptions of breast cancer appear in ancient Egyptian medical texts. The Edwin Smith Papyrus (circa 1600 BCE), which is based on earlier texts, describes cases of tumors or ulcers of the breast. But there was no treatment for breast cancer. They treated it with a "fire drill".

We again found about breast cancer in this era. The Ebers Papyrus (1534 BCE) described breast cancer as the "swelling (tumor) of vessels".

The Greek physician **Hippocrates**, also called the "Father of Medicine," described breast cancer as a humoral disease. He believed it was caused by an imbalance of the body's four humors (blood, phlegm, yellow bile, and black bile), with an excess of black bile causing cancer. He believed that cancer should be left alone, because those who got treatment did not live as long as those who were untreated.

Galen proposed that breast tumor was a coagulum of black bile in what is known as the Galenic humoral theory.

During **Middle age , Rene** described the lymhatic theory of origin of breast cancer. Treatments were rudimentary and often involved cauterization, surgery without anesthesia, and herbal remedies. Superstition and religious explanations often dominated medical understanding.

John Hunter, a Scottish surgeon, told that palpable breast tumors were caused by coagulation of defective lymph. He told that proposed that some cancers might be cured by surgery if detected early. In 1713, Bernardino Ramazzini observed that nuns often had breast cancer; he blames lack of sexual intercourse as a cause of breast cancer.

In 1838, Johannes **Muller proposed** that cencer cells developed from the blastema in the normal tissues. In 1882, **William Halsted** developed the radical mastectomy, a procedure that removed the breast, underlying chest muscle, and lymph nodes. This became the standard treatment for nearly a century.

In 1626, **Janet Elizabeth Laneclaypon** led an unprecedented study that identified breast cancer risk factors. Then the development of radiation therapy and chemotherapy began to provide new treatment options. The introduction of the mammogram in the 1960s allowed for earlier detection of breast cancer. In 1962, **Robert Egan** reported the first cases of breast cancer detected using mammography. Breast cancer awareness in the '70s promoted by First Lady Betty Ford and journalist Rose Kushner. In 1976, Bernard Fisher showed that less invasive lumpectomy was as effective as disfiguring radical mastectomies. Then Tamoxifen approved for treatment of metastatic breast cancer in 1977. Susan G. Komen Breast Cancer Foundation was founded in 1982. In 1980s-90s the development of breast-conserving surgery (lumpectomy) combined with radiation therapy offered alternatives to radical mastectomy. The identification of genetic mutations (BRCA1 and BRCA2) linked to breast cancer revolutionized understanding and prevention.

In **21st Century,** advances in genomic medicine have led to personalized treatment plans based on the genetic makeup of tumors. Immunotherapy and other novel therapies continue to be developed. Breast cancer subtypes- HER+, ER+, basal etc were identified via gene expression profiling in 2000. The **CLEOPATRA** study started from 2010 in favor of combination therapy.

The history of breast cancer reflects the broader history of medicine, with significant advancements in understanding, detection, and treatment occurring particularly in the last century. Ongoing research continues to improve outcomes and offer hope for more effective treatments in the future.

Introduction

Cancer is a disease in which some of the body's cells grow uncontrollably and spread to other parts of the body. Cancer can start almost anywhere in the human body, which is made up of trillions of cells. Normally, human cells grow and multiply (through a process called cell division) to form new cells as the body needs them. When cells grow old or become damaged, they die, and new cells take their place.Sometimes this orderly process breaks down, and abnormal or damaged cells grow and multiply when they shouldn't. These cells may form tumors, which are lumps of tissue.

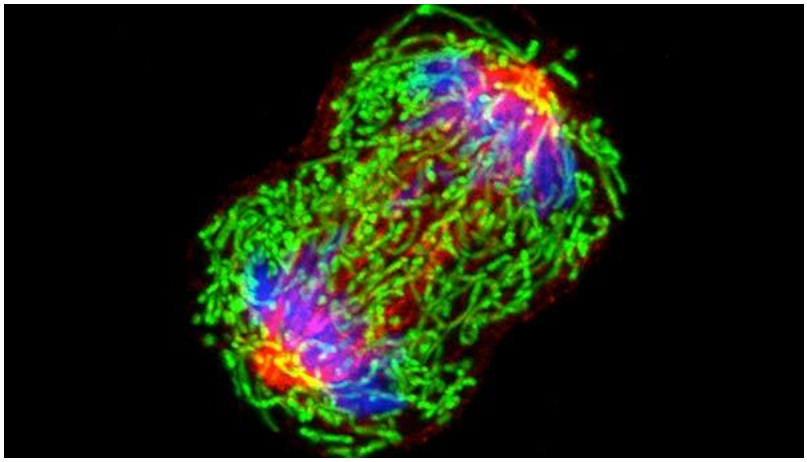


Figure 1: A Dividing Breast Cancer Cell

Breast cancer is the most commonly diagnosed cancer among women excluding non-melanoma of the skin.Cancer death is one of the major issues for the healthcare environment. It is one of the most significant reasons for women's death . It is the second leading cause of cancer death among women overall.

Data mining and machine learning techniques are straightforward and effective ways to understand and predict future data.Machine learning is one of the most popular models to easily train machines and create predictive models for successful decision-making. Machine learning helps with early diagnosis of breast cancer and determines the nature of the cancer by analysing the tumour size. Machine learning methods are the leading approaches to obtain favourable outcomes among classification and prediction problems.

Cancer detection using multivariate analysis involves the analysis of multiple variables simultaneously to identify patterns and relationships that can be indicative of the presence of cancer. Multivariate analysis can be a powerful tool for integrating and analyzing various data types to improve the accuracy of detection and diagnosis.

The objective of the study is –

- The classification goal is to find cancer at a early stage when it can be treated and may be cured on the basis of her symptoms.
- Comparison of the performance of the existing models.

Data Description

The data set of breast cancer patients was obtained from the 2017 November update of the SEER Program of the NCI, which provides information on population-based cancer statistics. The data set involved female patients with infiltrating duct and lobular carcinoma breast cancer (SEER primary cites recode NOS histology codes 8522/3) diagnosed in 2006-2010. Patients with unknown tumor size, examined regional LNs, regional positive LNs, and patients whose survival months were less than 1 month were excluded; thus, 4024 patients were ultimately included.

The internal structure of the data is given below:

```
  0    1
616 3408
'data.frame': 4024 obs. of  15 variables:
 $ Age                   : int  43 47 67 46 63 49 64 55 59 67 ...
 $ Race                  : Factor w/ 3 levels "Black","Other (American Indian/AK Native, Asian/Pacific Isla
 $ Marital.Status        : Factor w/ 5 levels "Divorced","Married (including common law)",..: 2 2 2 1 2 2 4
 $ T.Stage               : Factor w/ 4 levels "T1","T2","T3",..: 2 2 2 1 2 2 2 1 3 3 ...
 $ N.Stage               : Factor w/ 3 levels "N1","N2","N3": 3 2 1 1 2 3 1 1 1 2 ...
 $ X6th.Stage            : Factor w/ 5 levels "IIA","IIB","IIIA",..: 5 3 2 1 3 5 2 1 3 3 ...
 $ Grade                 : Factor w/ 4 levels "Moderately differentiated; Grade II",..: 1 1 2 1 1 1 1 1 1 2
 $ A.Stage               : Factor w/ 2 levels "Distant","Regional": 2 2 2 2 2 2 2 2 2 2 ...
 $ Tumor.Size            : int  40 45 25 19 35 32 22 15 70 55 ...
 $ Estrogen.Status       : Factor w/ 2 levels "Negative","Positive": 2 2 2 2 2 2 2 2 2 2 ...
 $ Progesterone.Status   : Factor w/ 2 levels "Negative","Positive": 2 2 2 2 2 2 2 2 2 2 ...
 $ Regional.Node.Examined: int  19 25 4 26 21 20 1 9 9 9 ...
 $ Reginol.Node.Positive : int  11 9 1 1 5 11 1 1 1 9 ...
 $ Survival.Months       : int  1 2 2 2 3 3 3 3 4 4 ...
 $ status                : int  1 1 0 0 0 1 0 1 0 0 ...
```

There are 15 features in the data set :

- **Age :-** It refers ages of women..
- **Race :-** It refers race of women ,It has 3 categories :'White', 'Black' and 'Other'.
- **Marital Status :-**It refers to the marital status of the women. It has five categories : Married, Divorced, Single, Widowed and Separated.
- **T Stage :-** T Stage refers to the size and extent of the primary tumor Category: 'T_1','T_2','T_3','T_4'. Using the TNA system, the "T" plus a letter or number (1 to 4) is used to describe the size and location of the tumor. Stage may also be divided into smaller groups that help describe the tumor in even more detail. Specific tumor stage information in listed below:
  - **T_1:** The tumor in the breast is 20 millimeters (mm) or smaller in size at its widest area. This is a little less than an inch.
  - **T_2:** The tumor is larger than 20 mm but not larger than 50 mm.
  - **T_3:** The tumor is larger than 50 mm.
  - **T_4:** The tumor falls into chest wall or skin or both.
- **N Stage :-** N Stage refers to the involvement of nearby lymph nodes Category: 'N_3','N_2','N_1'. The "N" in the TNA staging system stands for lymph nodes. These small, bean-shaped organs help fight infection. Lymph nodes near where the cancer started are called regional lymph nodes.

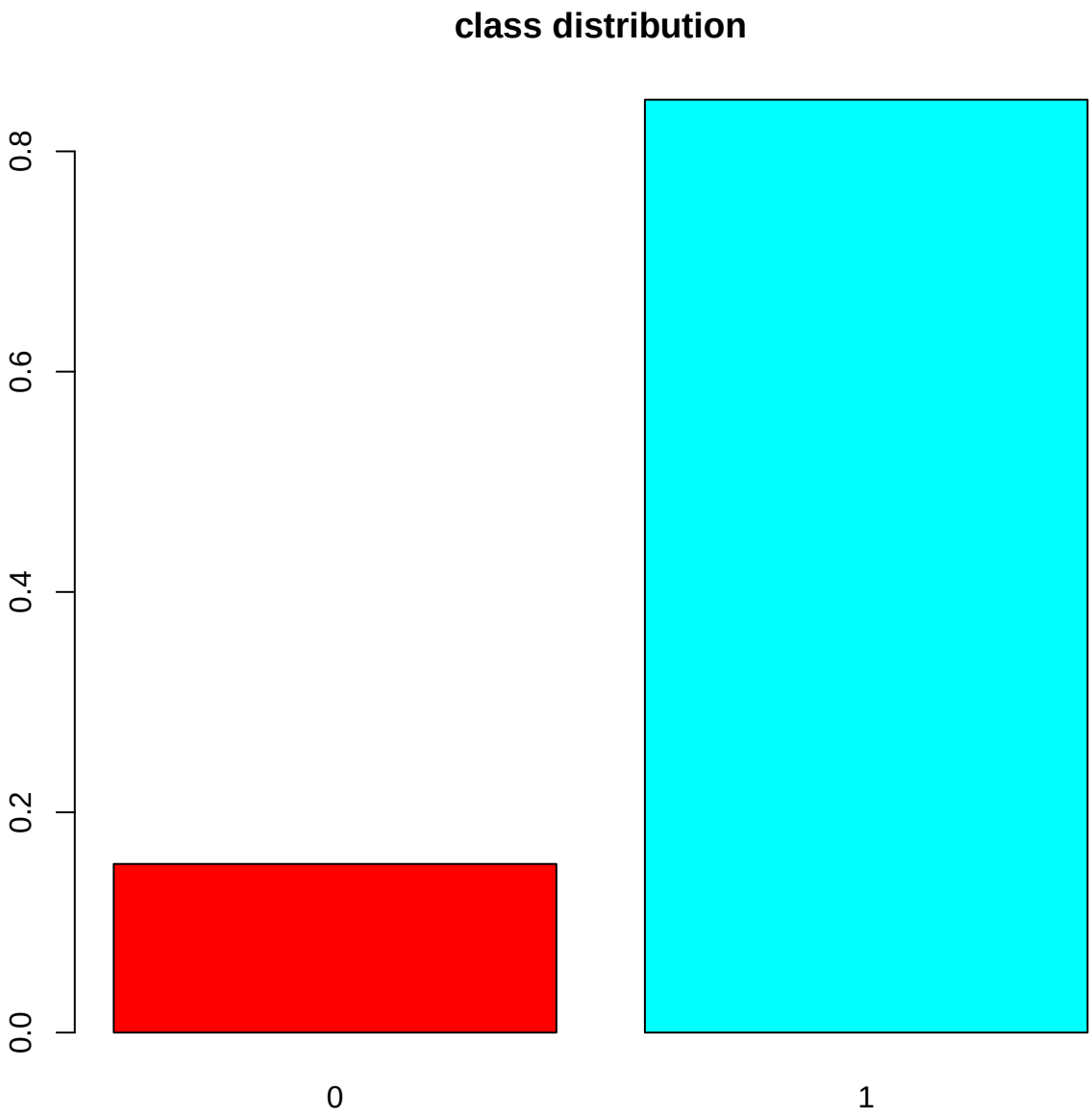
- **Regional lymph nodes :** Lymph nodes located under the arm, called the axillary lymph nodes. Lymph nodes located above and below the collarbone Lymph nodes located under the breastbone, called the internal mammary lymph nodes Lymph nodes in other parts of the body are called distant lymph nodes. The information below describes the staging.
- **N_0:** Cancer has not spread to nearby lymph nodes
- **N_1:** The cancer has spread to 1 to 3 axillary lymph nodes and/or the internal mammary lymph nodes. If the cancer in the lymph node is larger than 0.2 mm but 2 mm or smaller, it is called "micrometastatic".
- **N_2:** The cancer has spread to 4 to 9 axillary lymph nodes. Or, it has spread to the internal mammary lymph nodes, but not the axillary lymph nodes.
- **N_3:** The cancer has spread to 10 or more axillary lymph nodes, or it has spread to the lymph nodes located under the clavicle, or collarbone. It may have also spread to the internal mammary lymph nodes. Cancer that has spread to the lymph nodes above the clavicle, called the supraclavicular lymph nodes, is also described as N3.

- **X6th Stage :-** Doctors assign the stage of the cancer by combining the T, N, and A classifications , the tumor grade, and the results of ER/PR and HER2 testing. This information is used to help determine the prognosis . The simpler approach to explaining the stage of breast cancer is to use the T, N, and A classifications alone. This is the approach used below to describe the different stages. It has five categories: 'IIIC','IIIA','IIB','IIA' and 'IIIB'.
  - **Stage IIA:** Any 1 of these conditions: There is no evidence of a tumor in the breast, but the cancer has spread to 1 to 3 axillary lymph nodes. It has not spread to distant parts of the body (T_0, N_1, M_0). The tumor is 20 mm or smaller and has spread to 1 to 3 axillary lymph nodes (T_1, N_1, M_0). The tumor is larger than 20 mm but not larger than 50 mm and has not spread to the axillary lymph nodes (T_2, N_0, M_0).
  - **Stage IIB:** Either of these conditions: The tumor is larger than 20 mm but not larger than 50 mm and has spread to 1 to 3 axillary lymph nodes (T_2, N_1, M_0). The tumor is larger than 50 mm but has not spread to the axillary lymph nodes (T_3, N_0, M_0).
  - **Stage IIIA:** The tumor of any size has spread to 4 to 9 axillary lymph nodes or to internal mammary lymph nodes. It has not spread to other parts of the body (T_0, T_1, T_2, or T_3, N_2, M_0). Stage IIIA may also be a tumor larger than 50 mm that has spread to 1 to 3 axillary lymph nodes (T_3, N_1, M_0).
  - **Stage IIIB:** The tumor has spread to the chest wall or caused swelling or ulceration of the breast, or it is diagnosed as inflammatory breast cancer. It may or may not have spread to up to 9 axillary or internal mammary lymph nodes. It has not spread to other parts of the body (T_4; N_0, N_1, or N_2; M_0).
  - **Stage IIIC:** A tumor of any size that has spread to 10 or more axillary lymph nodes, the internal mammary lymph nodes, and/or the lymph nodes under the collarbone. It has not spread to other parts of the body (any T, N_3, M_0).

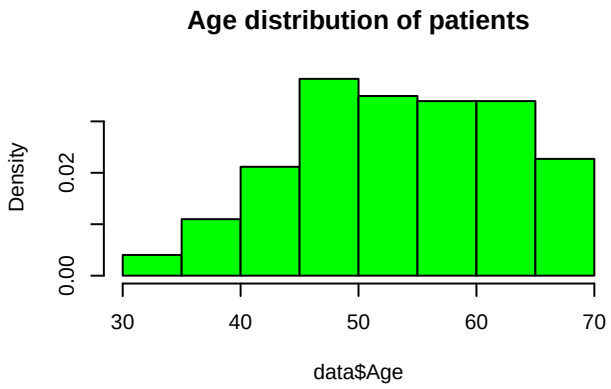
- **Differentiate :-** There are foue categories: 'Moderately differentiated; Grade II' , 'Poorly differentiated; Grade III' , 'Well differentiated; Grade I' and 'Undifferentiated; anaplastic; Grade IV'.
  - **Well differentiated :** Tumor cells and tissue looks most like healthy cells and tissue.The cells are slower-growing. These are called well-differentiated tumors and are considered low grade.
  - **Moderately differentiated:** The cells and tissue are somewhat abnormal and are called moderately differentiated. These are intermediate grade tumors.
  - **Poorly differentiated:** Cancer cells and tissue look very abnormal. These cancers are considered poorly differentiated, since they no longer have an architectural structure or pattern. These tumors are considered high grade.
  - **Undifferentiated:** These undifferentiated cancers have the most abnormal looking cells. These are the highest grade and typically grow and spread faster than lower grade tumors
- **Grade :-** The grade for differentiation Categories are '3', '2', '1', ' anaplastic; Grade IV'.

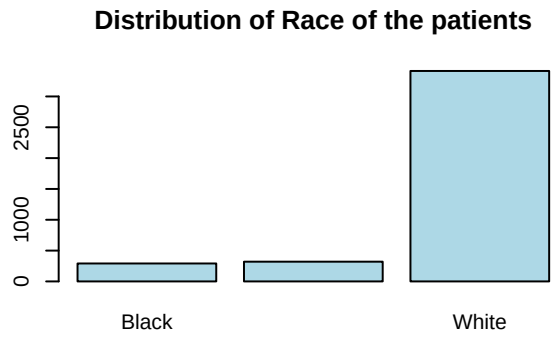
- **A Stage:-** It describes whether the cancer has spread to other parts of the body, called metastasis. This is no longer considered early-stage or locally advanced cancer. It has two categories : 'Regional' and 'Distant'.
  - **Regional:** Cancer has spread to nearby lymph nodes, tissues, or organs.
  - **Distant:** Cancer has spread to distant parts of the body.
- **Estrogen Status :-** It has two category: 'Positive' and 'Negative'.When a cell becomes cancerous, the number of hormone receptors increases on that cell. That's how it is marked as 'positive' and 'negetive'.
- **Progesterone Status :-** It also has two category: 'Positive' and 'Negative'.It is also categorized as estrogen status.
- **Regional Node Examined :-** It refers to the number of regional lymph nodes that were examined during the diagnostic process.
- **Regional Node Positive :-** It refers to the presence of cancer cells in the regional lymph nodes.
- **Survival Months :-** Number of months a patient survived.
- **Status :-** Status of the patient .There are two categories: 'Alive' and 'Dead'.

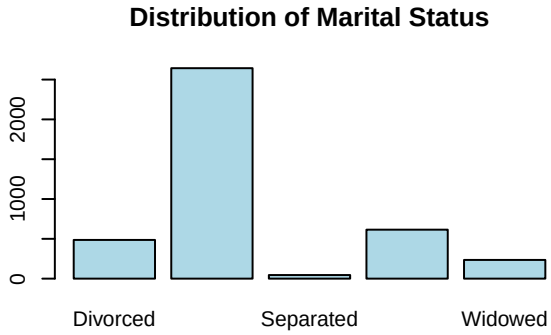
Visualization of the Data

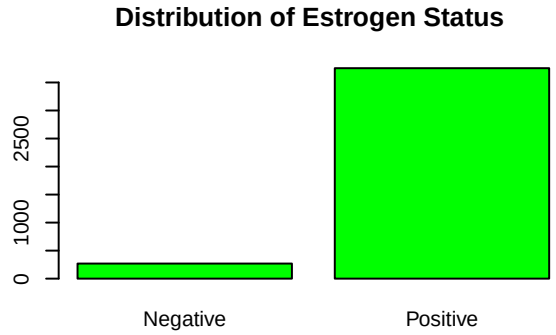


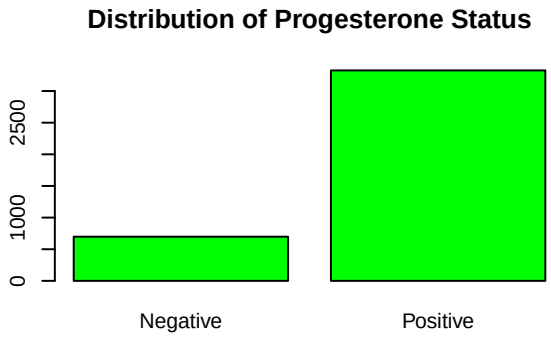
Analysis : In the diagram we can see that there are 3408 females are alive and 616 females are dead. So, there is an imbalance in the data.

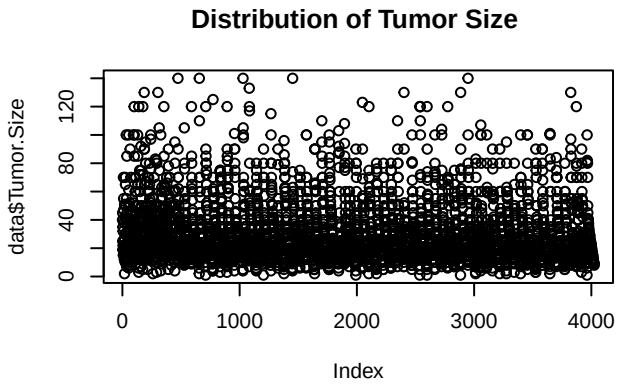






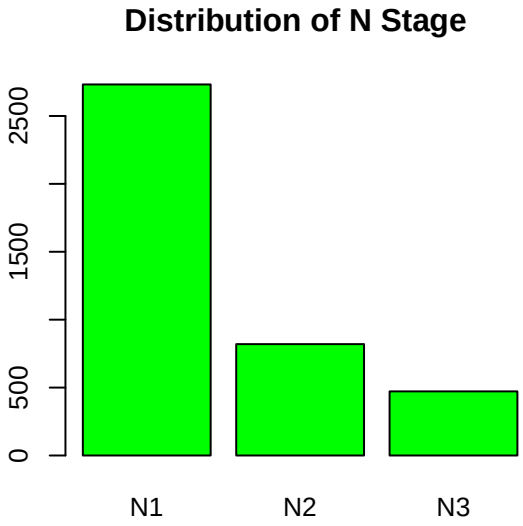


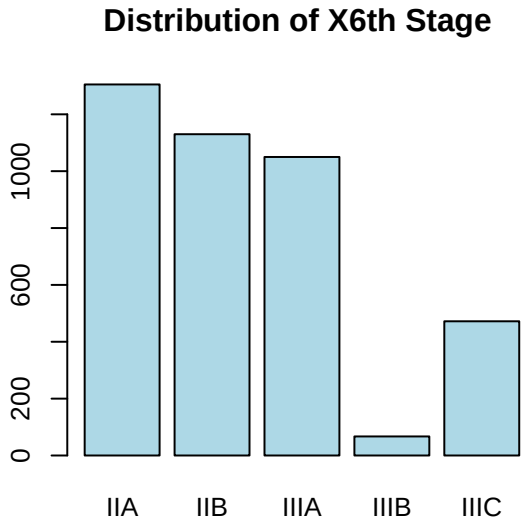


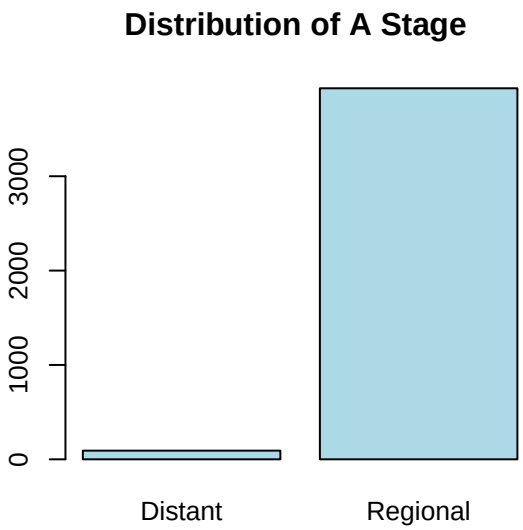


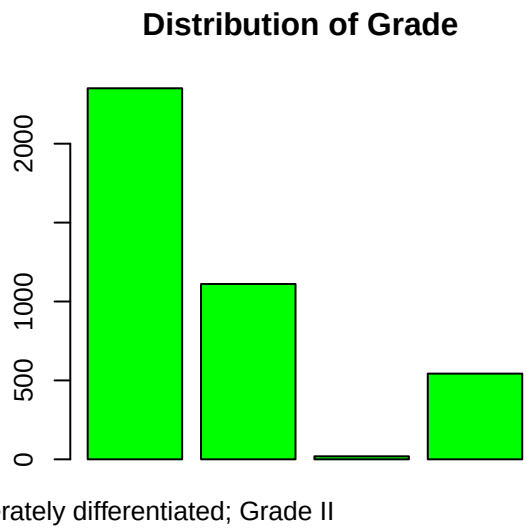
Analysis :

- From the graph of **Age** distribution of patients, we can see that most cancer patients are over 40.
- In the data, we see most of the people are white.
- From the data, it is clear that most of the patients are married.
- As we know, if a cell is cancerous it absorbs more hormones. The data also estublishes that.
- From the plot of the tumor size it can be concluded that most of the patients have tumor size upto 50 mm.



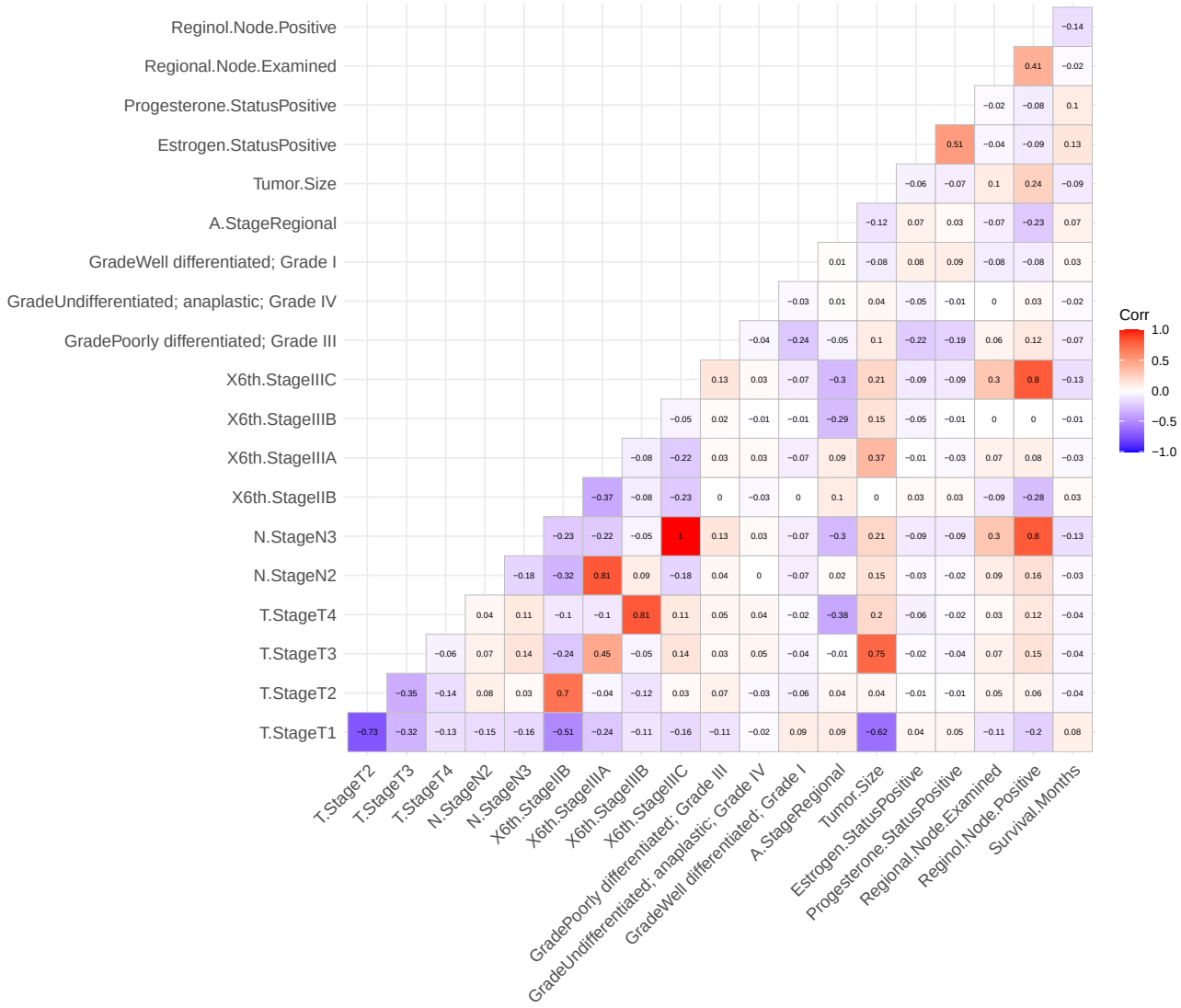






Analysis :

- From the distribution of **N Stage** ,we can see in the data most of the cancer has spread to 1 to 3 axillary lymph nodes and/or the internal mammary lymph nodes.
- We can see in the **X6th Stage** distribution that among the patients, any 1 of these conditions has happened.There is no evidence of a tumor in the breast, but the cancer has spread to 1 to 3 axillary lymph nodes. It has not spread to distant parts of the body (T0, N1, M0). The tumor is 20 mm or smaller and has spread to 1 to 3 axillary lymph nodes (T1, N1, M0). The tumor is larger than 20 mm but not larger than 50 mm and has not spread to the axillary lymph nodes (T2, N0, M0).
- From the distribution of **A Stage**, in most of the cases cancer has spread to nearby lymph nodes, tissues, or organs.
- From the distribution of **Grade** it is clear that most of the cancer cells are moderately differentiated.



Analysis : From the defination we know that, X6th stage is consist of N stage and T stage. We can see that from the correlation matrix also. The correlation between X6th stage and T stage and the correlation between X6th stage between N stage is higher than the others.

Methodology

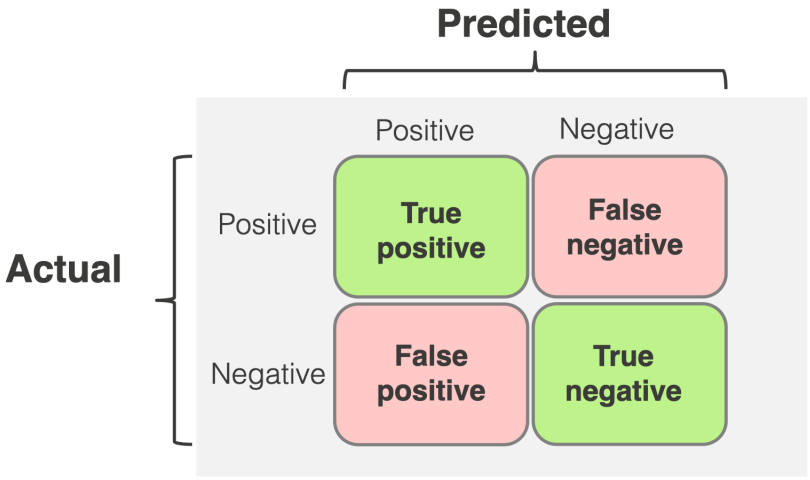
To fit models, we split the data set into 2 parts: 70% for train and 30% for test. And the data set is splitted into train set and test set for 100 times.

We also note the accuracy of the models (we display only the minimum accurecy,mean accurecy and maximum accurecy). And compare the models.

Here we work with 3 models named Random Forest model, Support Vector Machine and K-Nearest Neighbour. For Random Forest and SVM we use K-Fold corss validation ,where we take K=5.

Also we draw the ROC curves for the models by using Specificity and Sensitivity.

Here the confusion matrix is,



Specificity $= \frac{TN}{FP+TN}$

Sensitivity $= \frac{TP}{TP+FN}$

Accurecy $= \frac{TN+TP}{FP+TN+FN+TP}$

Random Forest Model

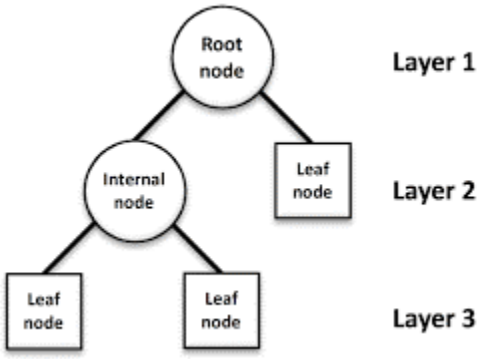
The decision tree is tree shape diagram or chart that helps determine a course of action or show a statistical probability. The chart is called a decision tree due to its resemblance to the namesake plant, usually outlined as an upright or a horizontal diagram that branches out. Beginning from the decision it (called a "node"); each "branch" of decision tree which represents possible decision or outcome, or reaction. The furthest branches on the tree represent the end results of a certain decision pathway and are called the "leaves".

The decision tree is graphical depiction of a decision and every potential outcome or result of making that decision. Individuals deploy decision tree in variety of situations, from something simple and personal to more complex industrial, scientific, or microeconomic undertakings.

By displaying a sequence of steps, a decision tree give people an effective and easy way to visualize and understands the potential effect of a decision and its range of possible outcomes. The decision tree also helps people

identify every potential option and weigh each course of action against the risk and reward so that each option can yield.

An organization can deploy decision tree as a kind of decision support machine. The structured model allows the reader of the chart to see how and why one would choice may lead to next, with the use of the branches indicating mutually exclusive options. The structure can allow users to take problem with multiple possible solutions and to display these solutions in simple, easy-to-understand format that also shows the relationship between different events and decisions.



Description of the above Figure:

- **Root Node:** Start at the root node, which contains the entire dataset.
- **Selecting the Best Attribute:** We use a metric like Gini impurity, entropy, or information gain, the best attribute to split the data.
- **Splitting:** Choose the best feature and split point to divide the dataset into subsets.
- **Recursive Splitting:** Apply the splitting process recursively to each subset.
- **Stopping Criteria:** Stop the splitting process when a stopping criterion is met (e.g., maximum depth, minimum samples per leaf).
- **Leaf Nodes:** Assign a class label (for classification) or a continuous value (for regression) to the leaf nodes.

Metrics for Splitting :

- **Gini Impurity:** Measures the likelihood of an incorrect classification of a new instance if it was randomly classified according to the distribution of classes in the dataset.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Where p_i is the probability of an instance being classified into i^{th} class.

- **Entropy:** Measures the amount of uncertainty or impurity in the dataset. Then the Entropy is given by

$$E = \sum_{i=1}^{n} -p_i log_2(p_i)$$

p_i is the probability of an instance being classified into ith class.
If entropy of a state is high, then we are very uncertain about the randomly picked point.

- **Information Gain:** Measures the reduction in entropy or Gini impurity after a dataset is split on an attribute. Then

$$IG(i) = E(Parent) - \sum_{i=1}^{n} w_i E(child_i)$$

The model compares every possible split and takes the one that maximizes information gain.

Advantages :

1. **Interpretability:** Decision trees are easy to understand and interpret. The model can be visualized, and the decision-making process is transparent.
2. **No Need for Data Normalization:** They do not require data normalization or scaling.
3. **Handling Non-Linear Relationships:** Decision trees can capture non-linear relationships between features and the target variable.
4. **Feature Importance:** They provide a measure of feature importance, which can be useful for feature selection. Disadvantages Overfitting: Decision trees can easily overfit the training data, especially if they are deep and have many leaves.

Disadvantages :

1. **Overfitting:** Decision trees can easily overfit the training data, especially if they are deep and have many leaves.
2. **Instability:** Small changes in the data can lead to different tree structures, making them sensitive to data variability.
3. **Bias:** They can be biased towards features with more levels (for categorical variables) or towards continuous variables with more possible splits.

IF WE ALREADY HAVE DECISION TREE, WHY DO WE NEED RANDOM FOREST?

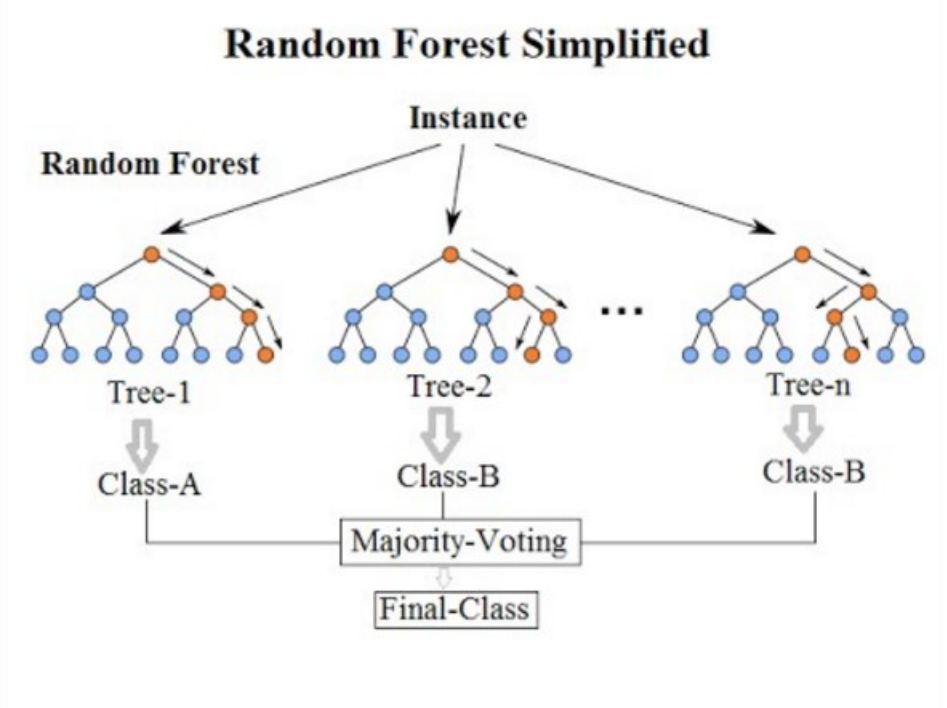
- Decision trees are highly sensitive to the training data which could result in high variance.
- Complex decision trees tend to overfit.

  A Random Forest is an ensemble learning algorithm used for classification, regression, and other tasks. It operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

The random Forest algorithm works in several steps which are discussed below:

1. **Ensemble of Decision Trees :** Random Forest belongs to the ensemble learning family of algorithms, that simply means it combines multiple models to improve overall performance. By aggregating the predictions of several models, it aims to reduce the risk of overfitting and improve generalization. Random Forest leverages the power of ensemble learning by constructing an army of Decision Trees.
2. **Random Feature Selection :** In addition to using different data subsets, Random Forest also introduces randomness in the feature selection process. At each split in a decision tree, a random subset of features is selected, and the best split is chosen from this subset. This further diversifies the trees and helps in reducing correlation among them.
3. **Bootstrap Aggregating (Bagging) :** The algorithm employs a technique called bootstrap aggregating, or bagging. This involves generating multiple subsets of the original training dataset by sampling with replacement. Each subset is used to train a different decision tree. This introduces diversity among the trees, as each tree sees a different set of data points.

4. **Decision Making and Voting :** When it comes to making predictions, each decision tree in the Random Forest casts its vote. For classification tasks, the final prediction is determined by the mode (most frequent prediction) across all the trees. In regression tasks, the average of the individual tree predictions is taken. This internal voting mechanism ensures a balanced and collective decision-making process.



Advantages :

- **High Accuracy:** By averaging multiple trees, Random Forest often achieves higher accuracy than a single decision tree.
- **Reduced Overfitting:** The combination of bagging and random feature selection helps in reducing overfitting, making the model more robust.
- **Feature Importance:** Provides insights into feature importance, which can be useful for understanding the underlying data.
- **Scalability:** Handles large datasets and high-dimensional data efficiently.

Disadvantages :

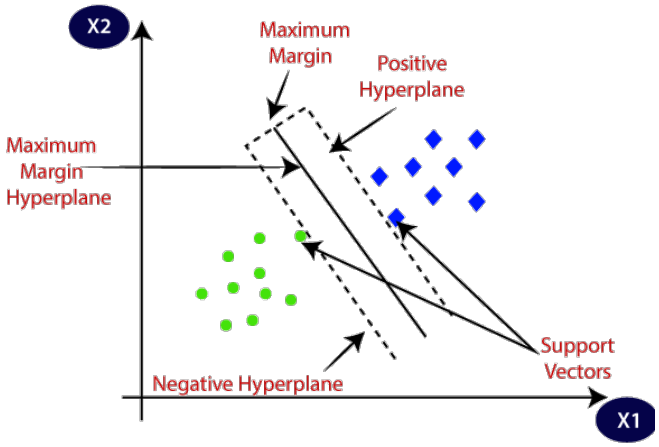
- **Complexity:** The model can be less interpretable than a single decision tree due to the combination of multiple trees.
- **Training Time and Resources:** Training multiple trees can be computationally intensive and requires more memory.
- **Prediction Speed:** Making predictions can be slower compared to simpler models, especially when the forest is large.

Support Vector Machine

Support Vector Machine(SVM) is an approach for classification.SVM is a generalization of a simple and intuitive classifier called the **Maximal Margin Classifier .**

WHAT IS MAXIMAL MARGIN CLASSIFIER ?

If our data can be perfectly separate by a hyperplane , then there will be infinitely many such hyperplanes.Now in order to construct a classifier based on separating hyperplane, a natural choice is maximal margin hyperplane.The maximal margin hyperplane is the separating hyperplane for which the margin is largest, that is , it is the hyperplane has the farthest minimum distance to the training observation.We can classify a test observation based on which side of maximal margin hyperplane it lies.This is nown as maximal margin classifier.



Support Vector Classifier :

In the above example if we add a new observation, then it might happen that the separating hyperplane changes dramatically.That is the resulting maximal hyperplane is not statisfactory.This is problematic because the distance of an observation from the hyperplane can be seen as a measure of our confidence that the observation was correctly classified.Also the fact of dramatic change in separating hyperplane may lead the overfitting in the training observation.

That means it could be worthwhile to misclassify a few training observation in order do better job in classifying the remaining observation. The support vector classifier does exactly the same.Rather than seeking the largest possible margin so that every observation is not only on the correct side of the hyperplane but also on the correct side of the margin,we instead allow some observation to be on the incorrect side of the margin or incorrect side of the hyperplane.

The support vector classifier is the solution of the optimization problem –

$$maximize_{\beta_1,\ldots,\beta_p,\epsilon_{1,\ldots,}\epsilon_n} M$$

Subect to

$$\sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \ldots\ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$$

Where,

M is the **width** of the Margin.

C is the tuning parameter.

$\epsilon_1, ..., \epsilon_n$ are the slack variables that allow individual observations to be on the wrong side of the margin or the hyperplane.

Role of the Tuning Parameter :

C bounds the sum of $\epsilon_i's$ and so it determines the number and severity of the violations to the margin (and to the hyperplane) that we will tolerate.We can think of C as a budget for the amount that the margin can be violated by n observations.

If **C=0 ,** then the budget for the margin can be violated by n observations is 0.That implies the maximal margin hyperplane is used if and only if the two classes are fully separable.

For **C>0**, not more than C observations can be on the wrong side of the hyperplane.

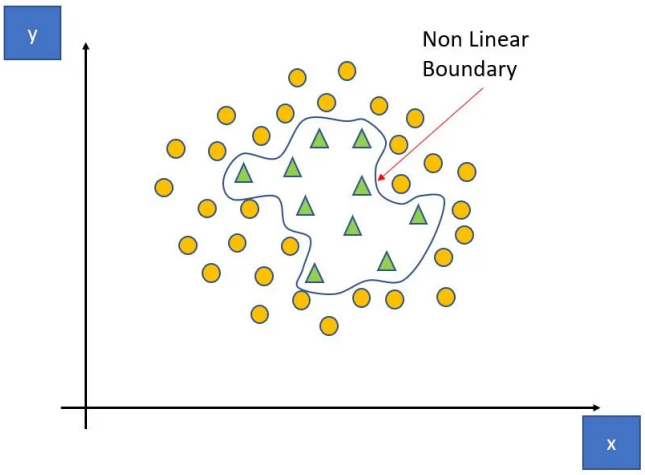
As C increases, we become more tolerant of violations of the margin and so the margin widen.Conversely as C decreases we become less tolerent of violations of the margin and so the margin narrows.

In practice C is treated as a tuning parameter that choosen via Cross-Validation.

Non-linear Decision Boundary :–

Till now we have discussed about the support vector classifier in terms of "**hyperplane**" , i.e, we are classifying using an linear decision boundary.

But in practice it may happen that we can't draw any linear boundary between the classes.Then we make a non-linear decision boundary. We can illustrate this by the following figure :–



Support Vector Machine : Support vector machine is an extension of support vector classifier that results from enlarging the feature space in a specific way, using **Karnels**.

Karnels : A Karnel is a function that quantifies the similarity of the observations.

Linear SVM : When we classify the classes using **linear Karnel** then it is called linear SVM.

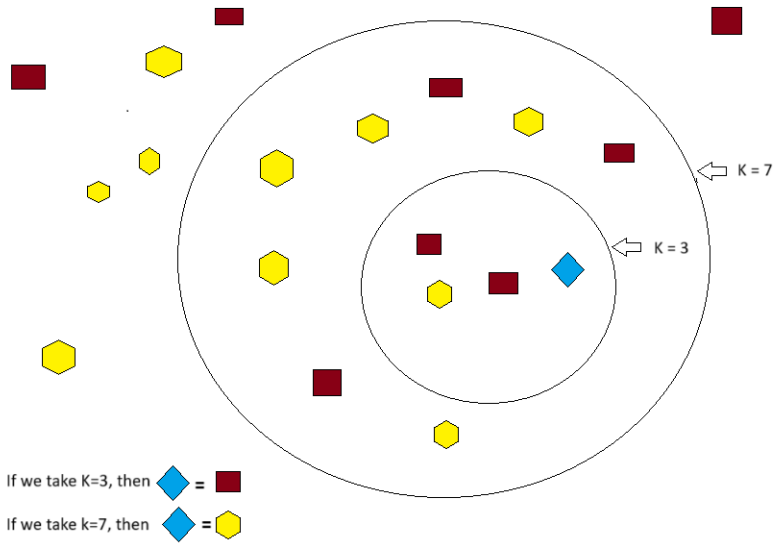
Non-linear SVM : When we classify the classes using **non-linaer Karnel** (such as, polynomial karnel,radial karnel etc.) then it is called non-linear SVM.

K-Nearest Neighbour

The k-nearest neighbors (K-NN) algorithm is a simple and versatile supervised machine learning algorithm used for both classification and regression tasks. It's a non-parametric method, meaning it doesn't make any assumptions about the underlying data distribution. Instead, it makes predictions based on the similarity of input data points. So, it classifies a data point based on how its neighbors are classified. It stores all available cases and classifies new cases based on a similarity measure.

HOW TO DEFINE "K"?

Consider the diagram below:



The graph above represents a data set consisting of two classes — red and yellow. A new data point has been introduced to the data set. This is shown by the blue point in the graph above. We'll then assign a value to **K** which denotes the number of neighbors to consider before classifying the new data entry. Let's assume the value of K is 3. Since the value of K is 3, the KNN algorithm will only consider the 3 nearest neighbors to the blue point (new entry). This is represented in the picture above. Out of the 3 nearest neighbors in the diagram above, the majority class is red so the new entry will be assigned to that class. The last data entry has been classified as red.

Let's assume the value of K is 7. Since the value of K is 7, the KNN algorithm will only consider the 11 nearest neighbors to the blue point (new entry). This is represented in the picture above. Out of the 11 nearest neighbors in the diagram above, the majority class is yellow so the new entry will be assigned to that class. The last data entry has been classified as yellow.

Generally "K" is chosen as "√n" , when n is total no of data points.

HOW TO DEFINE "NEAREST"?

To define "nearest" or "distance" we use distance metrics .

1. **Euclidean distance :**

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i + y_i)^2}$$

2. **Manhattan distance :**

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

3. **Minkowski distance :**

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

HOW TO DEFINE "NEIGHBORS"?

The K data points with the smallest distances to the target point are the nearest neighbors.

In classification problem, the class labels are determined by performing majority voting. In regression problem, the class label is calculated by taking average of the target values of K nearest neighbors.

Advantages

- It is simple to implement.
- It is robust to the noisy training data.
- No training is required before classification.

Disadvantages

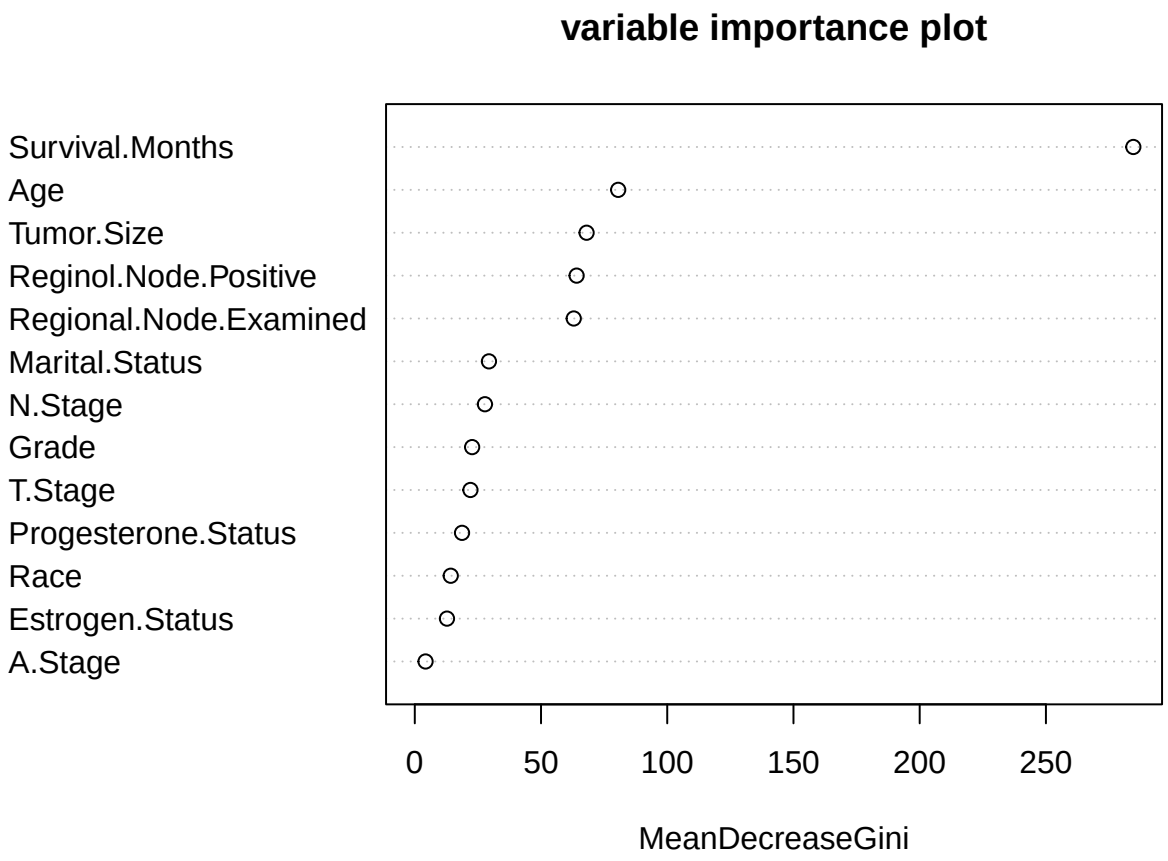
- Can be cost-intensive when working with a large data set.
- A lot of memory is required for processing large data sets.
- Choosing the right value of K can be tricky.

Results

i. Random Forest

```
[1] 92.0464
[1] 88.81524
[1] 90.60398
```

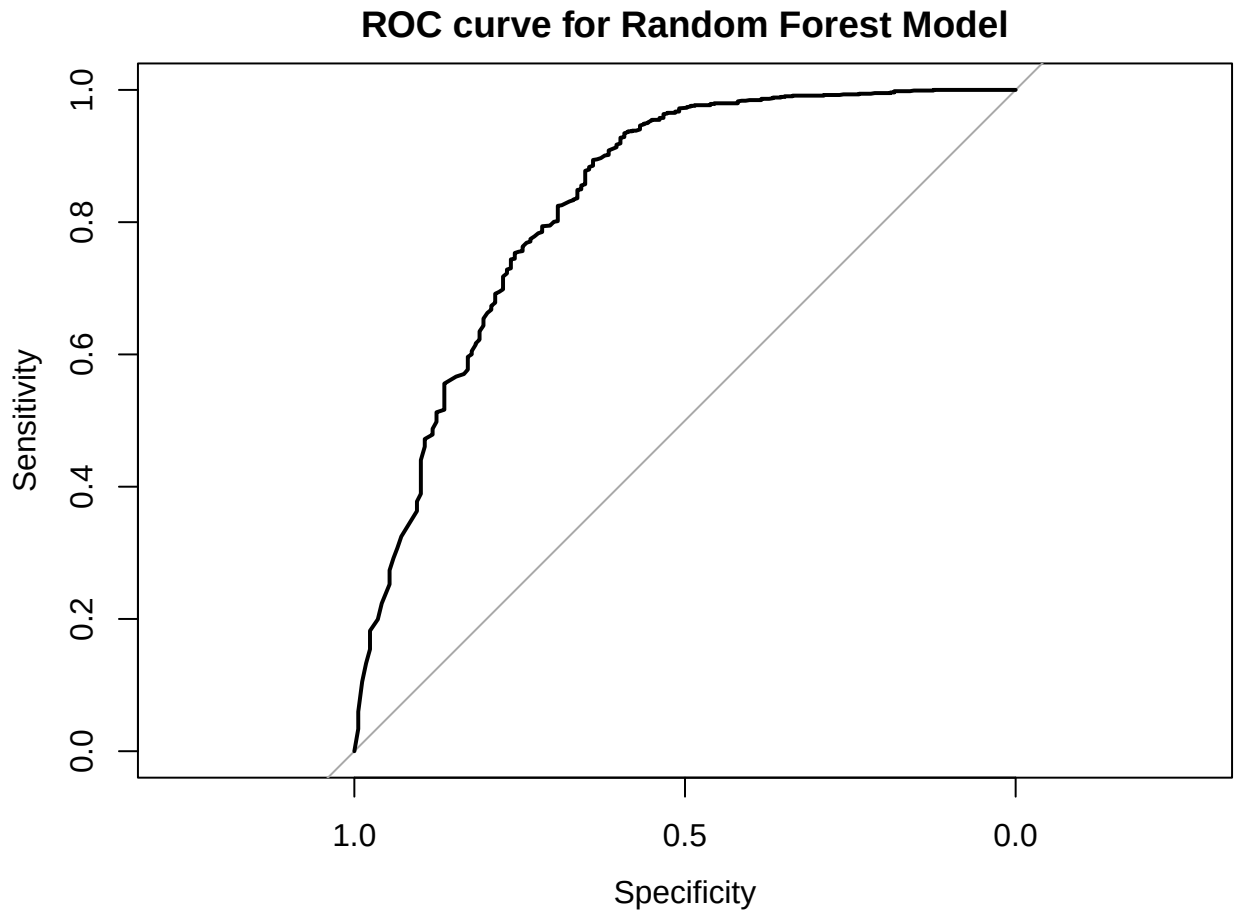
Here we can see that the maximum accurecy is around 92%, minimum accurecy is around 88.8 % and the mean accurecy is around 90.6% .



Analysis: In the above plot, the points represents the mean decrease Gini value,indicative of the importance of each variable.The higher value indicates the more importance of that variable in predicting the class (alive or dead).Removal of that variable causes the model to loose accurecy in prediction.

Here we can see that in the data, the variable "Survival Months" is most important to predict survival status.

```
Setting levels:  control = 0, case = 1
Setting direction:  controls < cases
```




```
Call:
roc.default(response = test_data$status, predictor = p1[, 2],     plot = TRUE, main = "ROC curve for Randor

Data: p1[, 2] in 169 controls (test_data$status 0) < 1038 cases (test_data$status 1).
Area under the curve: 0.8314
```

Therefore area under the curve (AUC) is 0.83.Thus we can say that Random forest model gives a good fit to the data.

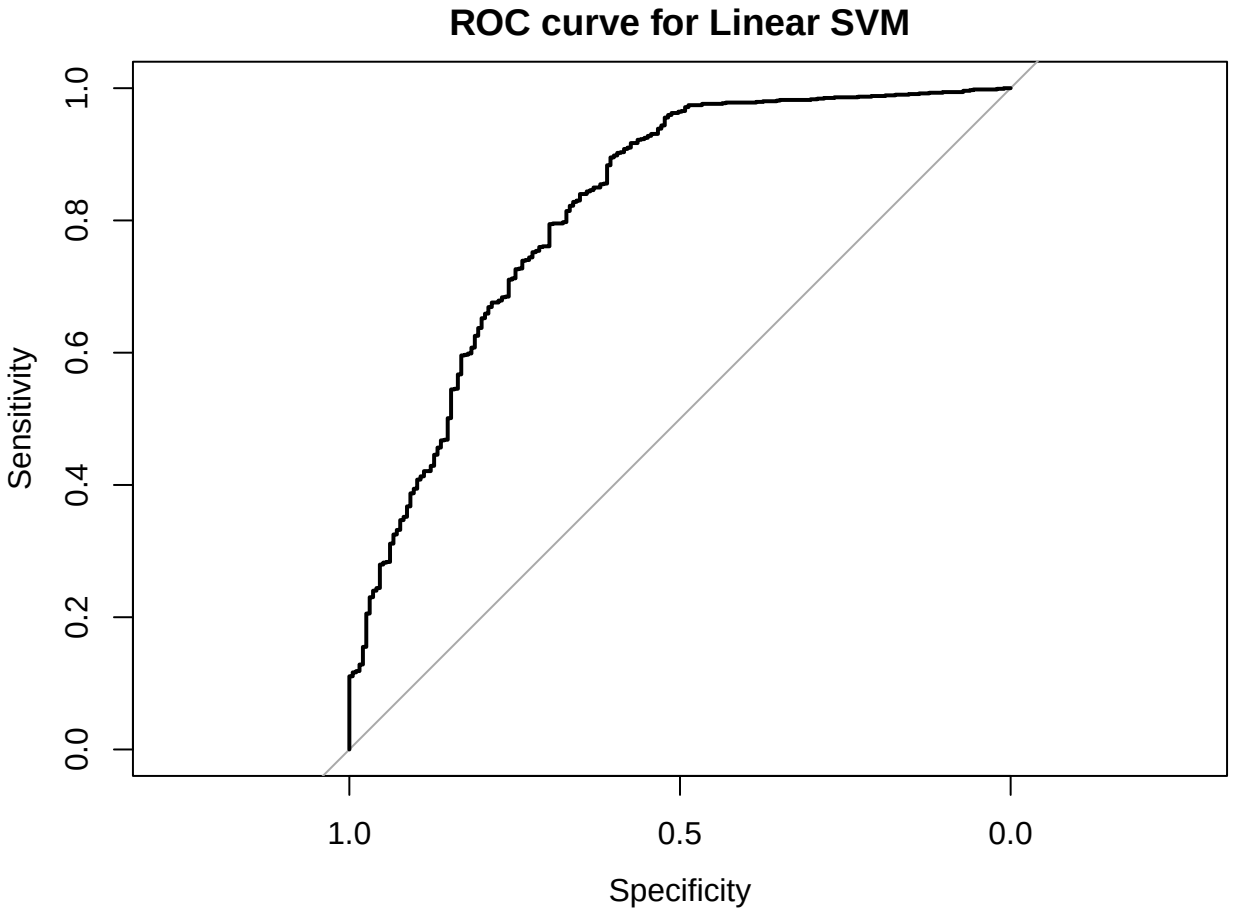
ii.Support Vector Machine (SVM):

Linear SVM :–

```
[1] 91.0522
[1] 87.48964
[1] 88.99006
```

Here we can see that the maximum accurecy is around 91 %, minimum accurecy is around 87 % and the mean accurecy is around 89 % .

```
Setting levels:  control = 0, case = 1
Setting direction:  controls < cases
```




```
Call:
roc.default(response = test_data$status, predictor = order(p2),     plot = TRUE, main = "ROC curve for Line

Data: order(p2) in 195 controls (test_data$status 0) < 1012 cases (test_data$status 1).
Area under the curve: 0.817
```

Therefore area under the curve (AUC) is 0.82.Thus we can say that the SVM model using linear karnel gives a good fit to the data.

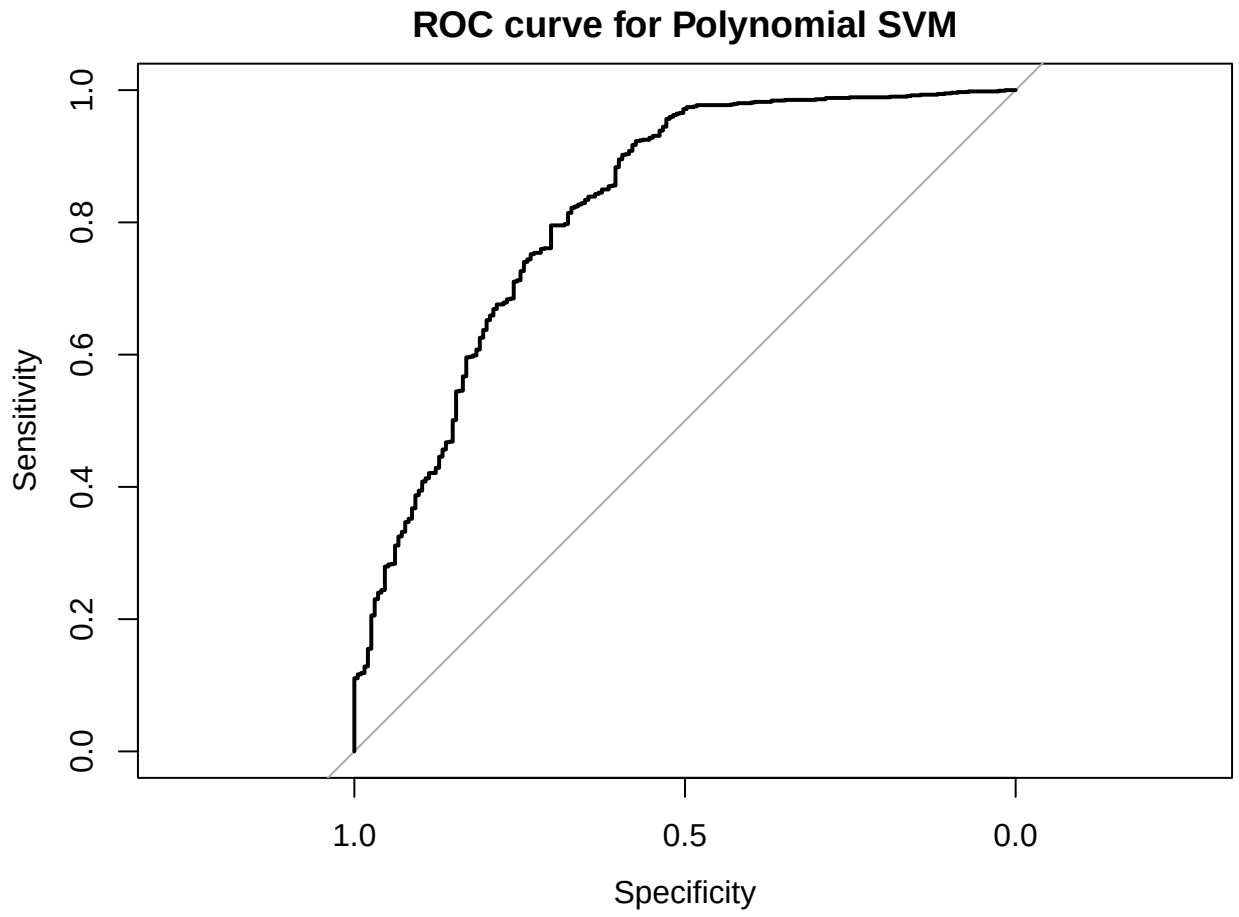
Non-linear SVM :

a). Polynomial Kernel

```
[1] 90.058
[1] 86.24689
[1] 87.94118
```

Here we can see that the maximum accurecy is around 90 %, minimum accurecy is around 86 % and the mean accurecy is around 88% .

```
Setting levels:  control = 0, case = 1
Setting direction:  controls < cases
```




```
Call:
roc.default(response = test_data$status, predictor = order(p3),     plot = TRUE, main = "ROC curve for Poly

Data: order(p3) in 195 controls (test_data$status 0) < 1012 cases (test_data$status 1).
Area under the curve: 0.8186
```

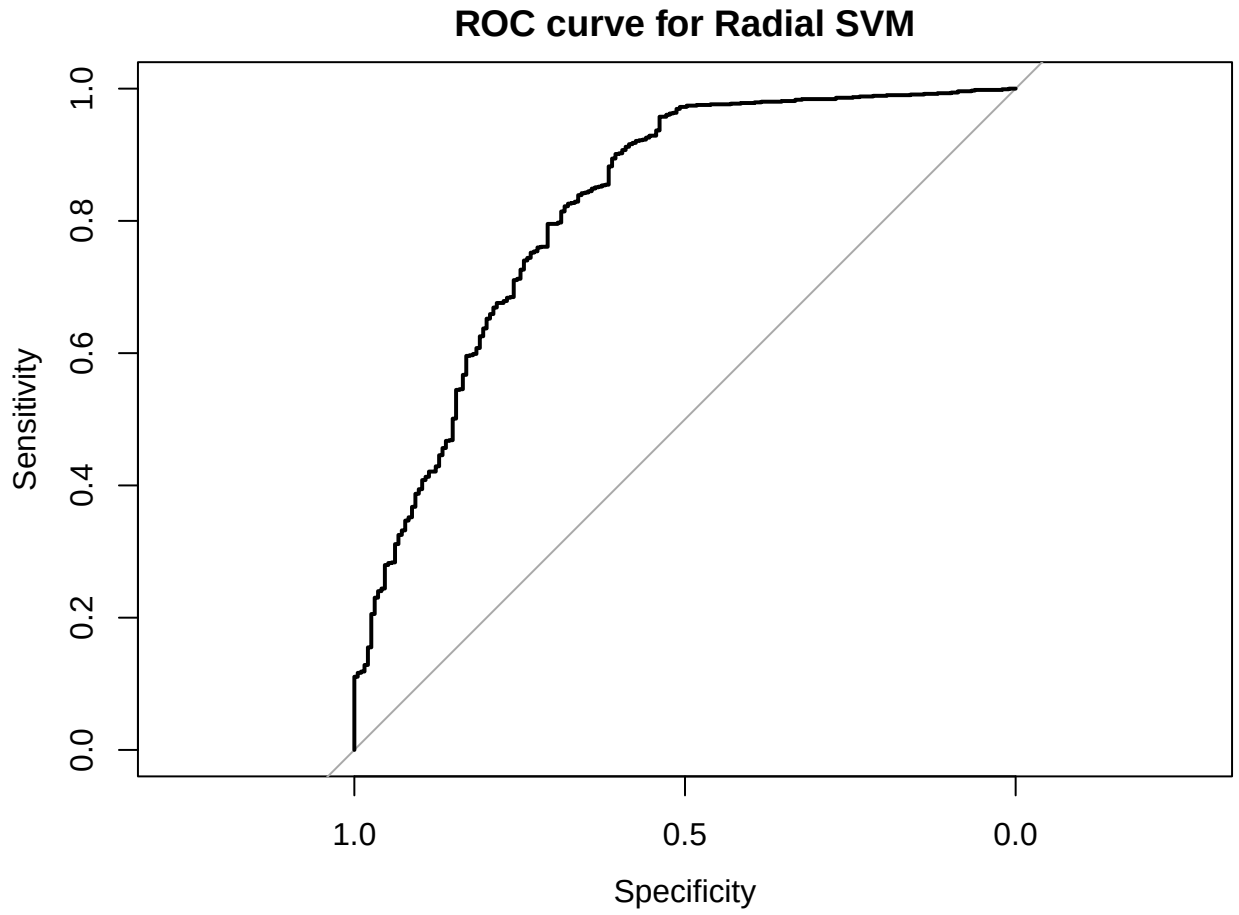
Therefore area under the curve (AUC) is 0.82. Thus we can say that the SVM model using polynomial karnel gives a good fit to the data.

b). Radial Kernel :

```
[1] 90.80365
[1] 87.15824
[1] 88.77051
```

Here we can see that the maximum accurecy is around 90.8 %, minimum accurecy is around 87 % and the mean accurecy is around 88.7% .

```
Setting levels:  control = 0, case = 1
Setting direction:  controls < cases
```




```
Call:
roc.default(response = test_data$status, predictor = order(p4),     plot = TRUE, main = "ROC curve for Radi

Data: order(p4) in 195 controls (test_data$status 0) < 1012 cases (test_data$status 1).
Area under the curve: 0.8196
```

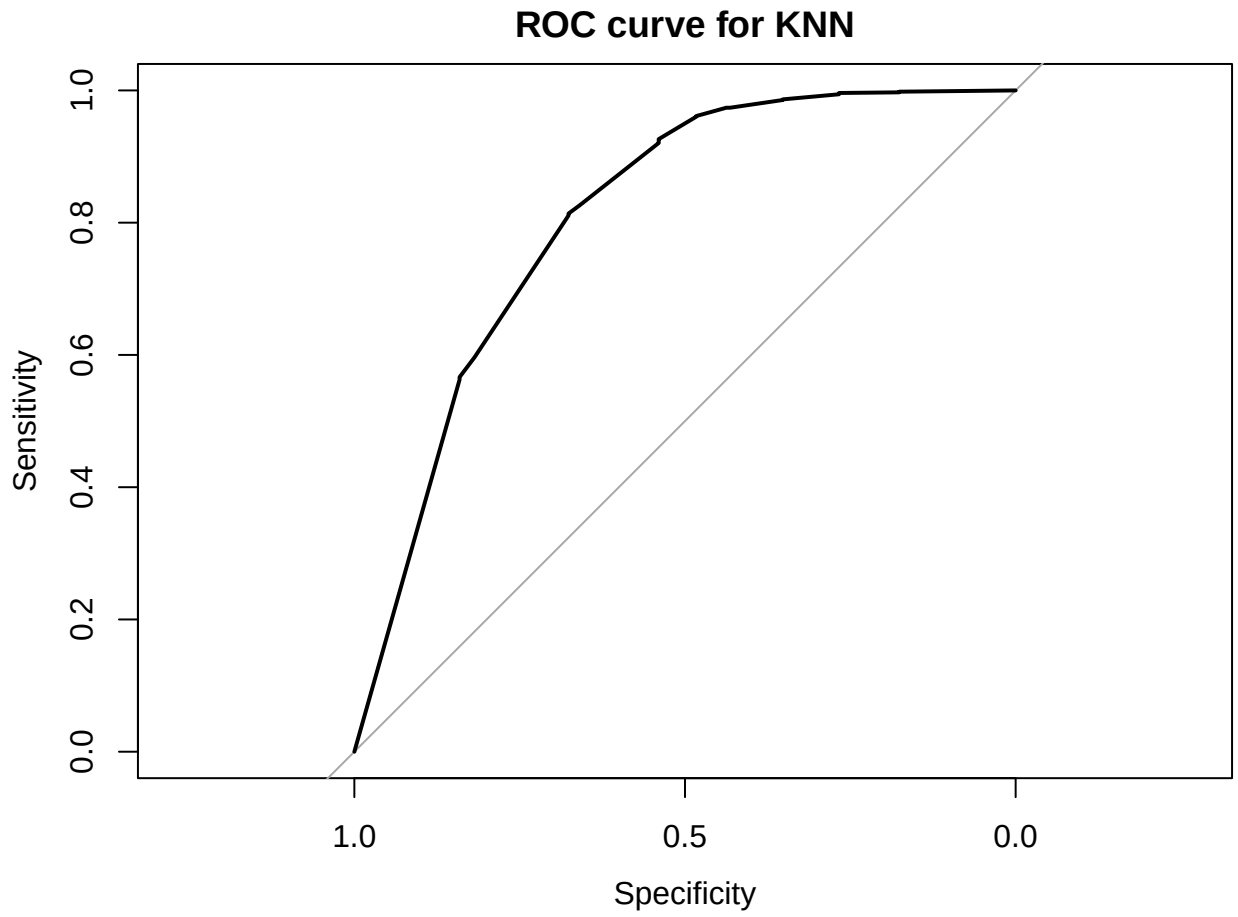
Therefore area under the curve (AUC) is 0.82. Thus we can say that the SVM model using radial karnel gives a good fit to the data.

iii. K-Nearest Neighbourhood :

```
  k
3 9
[1] 91.0522
[1] 87.98674
[1] 89.28832
```

Here we can see that the maximum accurecy is around 91 %, minimum accurecy is around 88 % and the mean accurecy is around 89 % .

```
Setting levels:  control = 0, case = 1
Setting direction:  controls < cases
```

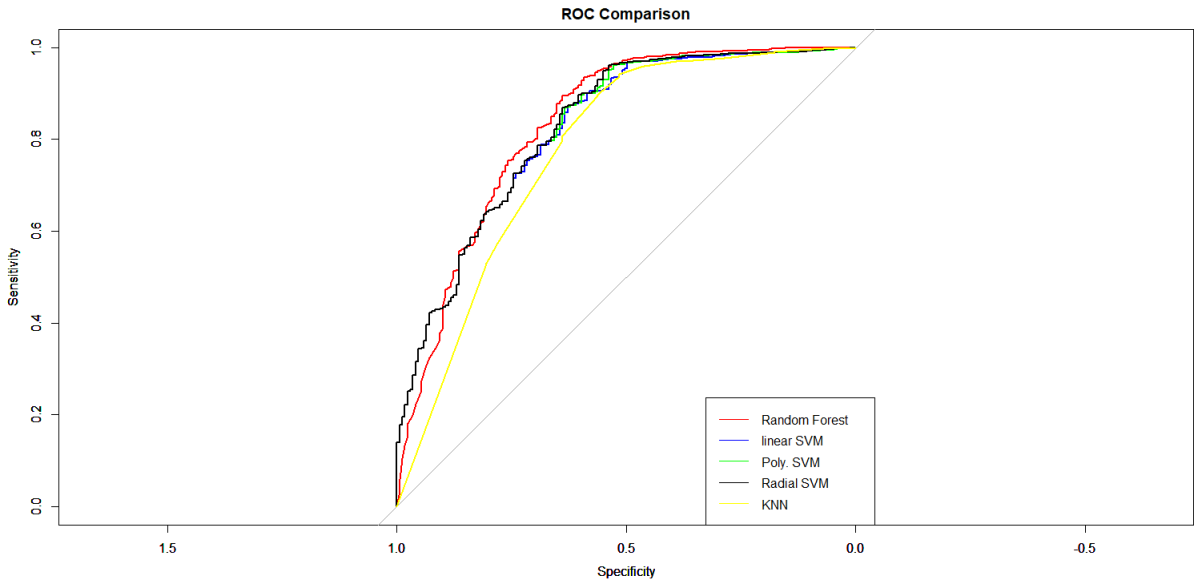



```
Call:
roc.default(response = test_data$status, predictor = p5[, 2],     plot = TRUE, main = "ROC curve for KNN")

Data: p5[, 2] in 176 controls (test_data$status 0) < 1031 cases (test_data$status 1).
Area under the curve: 0.8081
```

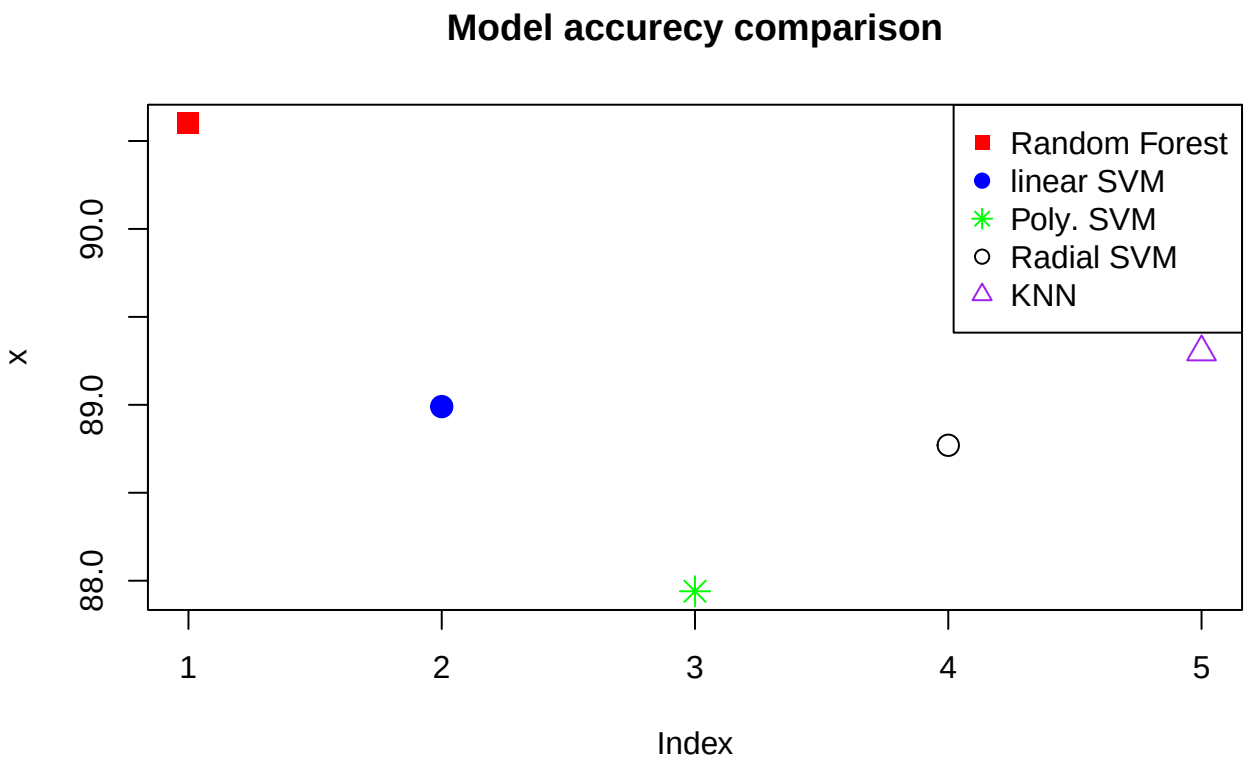
Therefore area under the curve (AUC) is 0.80.Thus we can say that the KNN model gives a good fit to the data.

Comparison :



From the above figure we can see that the area under the curve for random forest is slightly higher than the others. And AUC for the linear SVM and non-linear SVM are almost same.And KNN has the lowest AUC among them.

Now comparing the mean accurecy rate of the models we get the following figure :



From the figure it is clear that Random forest model has the maximum accurecy and the polynomial SVM has the minimum accurecy.

R Codes

```
data<- read.csv("E:\\4th sem project\\Dataset\\new data.csv")

table(data$status)
data$X<- NULL
data$Status<-NULL
 # Convert all columns to factor
data <- as.data.frame(unclass(data),stringsAsFactors = TRUE)

str(data)
```

Data Visualization :

```
barplot(prop.table(table(data$status)),col=rainbow(2),main="class distribution")
par(mfrow=c(3,2))
hist(data$Age,probability = TRUE, main = "Age distribution of patients",col="green")
plot(data$Race,main="Distribution of Race of the patients",col="light blue")
plot(data$Marital.Status,main="Distribution of Marital Status",col="light blue")
plot(data$Estrogen.Status,main="Distribution of Estrogen Status ",col="green")
plot(data$Progesterone.Status,main="Distribution of Progesterone Status",col="green")
plot(data$Tumor.Size,main="Distribution of Tumor Size ")
par(mfrow=c(2,2))
plot(data$N.Stage,main="Distribution of N Stage ",col="green")

plot(data$X6th.Stage,main="Distribution of X6th Stage ",col="light blue")

plot(data$A.Stage,main="Distribution of A Stage ",col="light blue")

plot(data$Grade,main="Distribution of Grade ",col="green")

df=data[,4:14]
library(knitr)
suppressWarnings(suppressMessages(library(ggplot2)))
suppressWarnings(suppressMessages(library(ggcorrplot)))
suppressWarnings(suppressMessages(library(tidyverse)))

model.matrix(~0+., data=df)%>%
 cor(use="pairwise.complete.obs")%>%
 ggcorrplot(show.diag=FALSE, type="lower", lab=TRUE, lab_size=2)
```

Random Forest :

```
set.seed(123)
suppressWarnings(suppressMessages(library(lattice)))
suppressWarnings(suppressMessages(library(caret)))
cross_val=trainControl(method='cv',number=5)
suppressWarnings(suppressMessages(library(randomForest)))
rfm_model_accuracy=c()

for(i in 1:100)
{
```

```
 split<-createDataPartition(data$status,p=0.7,list=FALSE)
 train_data<-data[split,]
 test_data=data[-split,]
 RFM<-randomForest(factor(status)~Age+Race+Marital.Status+T.Stage+N.Stage +                  Grade+A.Stage +

 pmmodel1= as.numeric(predict(RFM,newdata = test_data,type="response"))
 tab1= table(pmmodel1,test_data$status)
 rfm_model_accuracy[i]=sum(diag(tab1))/sum(tab1)*100
}

max(rfm_model_accuracy)
min(rfm_model_accuracy)
mean(rfm_model_accuracy)

varImpPlot(RFM,main="variable importance plot")

suppressWarnings(suppressMessages(library(pROC)))
p1 <- predict(RFM, newdata=test_data,type = "prob")
roc(test_data$status,p1[,2],plot=TRUE,main="ROC curve for Random Forest Model")
```

Linear Support Vector Machine :

```
set.seed(123)
suppressWarnings(suppressMessages(library (e1071)))
svm_accuracy=c()
for(i in 1:100)
{
 split<-createDataPartition(data$status,p=0.7,list=FALSE)
 train_data<-data[split,]
 test_data=data[-split,]
 svmfit<-svm(factor(status)~Age+Race+Marital.Status+T.Stage + N.Stage +                Grade+A.Stage +Tumo
 pmmodel2= as.numeric(predict(svmfit,newdata = test_data,type="response"))
 tab2= table(pmmodel2,test_data$status)
 svm_accuracy[i]=sum(diag(tab2))/sum(tab2)*100

}
max(svm_accuracy)
min(svm_accuracy)
mean(svm_accuracy)

suppressWarnings(suppressMessages(library(pROC)))
p2 <- predict(svmfit, newdata=test_data,type = "prob")
roc(test_data$status,order(p2),plot=TRUE,main="ROC curve for Linear SVM")
```

Non-Linear SVM :

```
set.seed(123)
svm_accurecy1=c()
for(i in 1:100)
{
 split<-createDataPartition(data$status,p=0.7,list=FALSE)
 train_data<-data[split,]
 test_data=data[-split,]

 svmfit<-svm(factor(status)~Age+Race+Marital.Status+T.Stage + N.Stage +                Grade+A.Stage +Tumo
```

```
 pmmodel21= as.numeric(predict(svmfit,newdata = test_data,type="response"))
 tab21= table(pmmodel21,test_data$status)
 svm_accurecy1[i]=sum(diag(tab21))/sum(tab21)*100
}
max(svm_accurecy1)
min(svm_accurecy1)
mean(svm_accurecy1)
```

```
suppressWarnings(suppressMessages(library(pROC)))
p3 <- predict(svmfit, newdata=test_data,type = "prob")
roc(test_data$status,order(p3),plot=TRUE,main="ROC curve for Polynomial SVM")
```

```
set.seed(123)
svm_accurecy2=c()
for(i in 1:100)
{
 split<-createDataPartition(data$status,p=0.7,list=FALSE)
 train_data<-data[split,]
 test_data=data[-split,]

 svmfit<-svm(factor(status)~Age+Race+Marital.Status+T.Stage + N.Stage +                Grade+A.Stage +Tumo
  pmmodel22= as.numeric(predict(svmfit,newdata = test_data,type="response"))
 tab22= table(pmmodel22,test_data$status)
 svm_accurecy2[i]=sum(diag(tab22))/sum(tab22)*100

}
max(svm_accurecy2)
min(svm_accurecy2)
mean(svm_accurecy2)
```

```
suppressWarnings(suppressMessages(library(pROC)))
p4 <- predict(svmfit, newdata=test_data,type = "prob")
roc(test_data$status,order(p4),plot=TRUE,main="ROC curve for Radial SVM")
```

KNN :

```
set.seed(123)
suppressWarnings(suppressMessages(library(lattice)))

library(caret)
cross_val_kn=trainControl(method='cv',number=5)

knn_model_accuracy=c()
for(i in 1:100)
{
  split<-createDataPartition(data$status,p=0.7,list=FALSE)
 train_data<-data[split,]
 test_data=data[-split,]

     knn1<- train(factor(status)~Age+Race+Marital.Status+T.Stage + N.Stage +                Grade+A.Stage

    pmmodel3= as.numeric(predict(knn1,newdata = test_data))
  tab3= table(pmmodel3,test_data$status)
```

```
  knn_model_accuracy[i]=sum(diag(tab3))/sum(tab3)*100
  }

print(knn1$bestTune)
#print(knn_model_accuracy)#..model accuracy for test data
max(knn_model_accuracy)
min(knn_model_accuracy)
mean(knn_model_accuracy)
```

```
suppressWarnings(suppressMessages(library(pROC)))
p5 <- predict(knn1, newdata=test_data,type = "prob")
roc(test_data$status,p5[,2],plot=TRUE,main="ROC curve for KNN")
```

Comparison :

```
roc(test_data$status,p1[,2],plot=TRUE,col="red",main="ROC Comparison")
par(new=TRUE)
roc(test_data$status,order(p2),plot=TRUE,col="blue")
par(new=TRUE)
roc(test_data$status,order(p3),plot=TRUE,col="green")
par(new=TRUE)
roc(test_data$status,order(p4),plot=TRUE,col="black")
par(new=TRUE)
roc(test_data$status,p5[,2],plot=TRUE,col="yellow")
legend("bottomright",legend=c("Random Forest","linear SVM","Poly. SVM","Radial SVM","KNN"),col=c("red","blu
```

```
x=c(90.6,88.99,87.94,88.77,89.3)
plot(x,main="Model accurecy comparison",pch=c(15,19,8,21,24),cex=1.5,col=c("red","blue","green","black","pu
legend("topright",legend=c("Random Forest","linear SVM","Poly. SVM","Radial SVM","KNN"),col=c("red","blue",
```

Conclusion

abv

Future Prospectus

For the lack of time we could not think of building any model which can fulfill the classification goal with more accuracy.

References

- An Introduction to Statistical Learning: with Applications in R. by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
- Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis: Mohammed Amine Naji a, Sanaa El Filalib Kawtar Aarikac, EL Habib Benlahmard, Rachida Ait Abdelouhahide, Olivier Debauchef
- Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques and their Analysis : NOREEN FATIMA , LI LIU , HONG SHA , AND HAROON AHMED
- A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications: Industrial Engineering Department, Antalya Bilim University