

‘Breast Cancer Prediction’ using classification models

Final Report

Group Members:   
Deep Singh 251122489  
Mandeep Singh 251122474  
Sanket Salunke 251102392

Table of Contents

[I. INTRODUCTION 1](#_Toc35684913)

[II. PROBLEM DEFINITION AND MOTIVATION 1](#_Toc35684914)

[*A.* *Problem Definition* 1](#_Toc35684915)

[*B.* *Motivation* 1](#_Toc35684916)

[III. RELATED WORK 1](#_Toc35684917)

[IV. METHODOLOGY 1](#_Toc35684918)

[*A.* *Understanding Business Problem* 1](#_Toc35684919)

[*B.* *Problem Formalization* 1](#_Toc35684920)

[*C.* *Requirement Analysis* 2](#_Toc35684921)

[*D.* *Exploratory Data Analysis* 2](#_Toc35684922)

[*1.* *Target Value Analysis* 3](#_Toc35684923)

[2. *Visualizations* 3](#_Toc35684924)

[*E.* *Data Preprocessing* 4](#_Toc35684925)

[V. MODELLING (IMPLEMENTATION OF MODELS) 5](#_Toc35684926)

[VI. MODEL EVALUATION 5](#_Toc35684927)

[VII. OBSERVATION 5](#_Toc35684927)

[VIII. REFERENCES 6](#_Toc35684928)

**Figures**

[Figure 1 Workflow 1](#_Toc35606147)

[Figure 2 2](#_Toc35606148)

[Figure 3 2](#_Toc35606149)

[Figure 4 Target value Analysis 3](#_Toc35606150)

[Figure 5 Univariate Analysis – Box plot 3](#_Toc35606151)

[Figure 6 Correlation Matrix - using Heatmap 3](#_Toc35606152)

[Figure 7 Univariate Analysis - Histogram 4](#_Toc35606153)

[Figure 8 Bivariate Analysis - Histogram and Scatterplot 4](#_Toc35606154)

[Figure 9 4](#_Toc35606155)

[Figure 10 5](#_Toc35606156)

[Figure 11 5](#_Toc35606157)

[Figure 12 Conversion of Categorical Data 5](#_Toc35606158)

[Figure 13 Feature Scaling (Range 0-1) 5](#_Toc35606159)

[Figure 14 5](#_Toc35606159)

[Figure 15 5](#_Toc35606159)

[Figure 16 Confusion Matrix for KNN 5](#_Toc35606159)

[Figure 17 Confusion Matrix for SVM Linear Kernel 5](#_Toc35606159)

[Figure 18 Confusion Matrix for SVM Polynomial Kernel 5](#_Toc35606159)

[Figure 19 Confusion Matrix for SVM Gaussian Kernel 5](#_Toc35606159)

**Tables**

Table 1 Dataset Description …………………………………………………………………………………………………… 2

Table 2 Project Status …….……………………………………………………………………………... ……………………. 5

Table 3 Comparison of different SVM Kernels………………………………………………………... ……………………... 5

Table 4 Comparison of different models …...…………………………………………………………... ……………………. 5

# INTRODUCTION

Breast cancer is the most common cancer among Canadian women (excluding non-melanoma skin cancers). It is the second leading cause of death from cancer in Canadian women [1]. More than 80% of deaths due to cancerous cells occurs in underdeveloped and developing countries. Cancer in an essence is a group of diseases, causing an abnormal growth of body cells which lead to development in malign cells, also termed as cancerous cells. As per some statistics of Canadian government, breast cancer accounts for 13 to 25% of overall deaths due to cancer. It is estimated that in year 2019, around 30,000 cases were diagnosed with breast cancer and 5000 in severe cases.

# PROBLEM DEFINITION AND MOTIVATION

## *Problem Definition*

Presently there are many ML models, which can be used to for predictions. However, the accuracy of these models is relatively low. Under these circumstances, the chances for false predictions become certainly high. Since, these predictions affect the life of individuals, the error introduced in predication models can be fatal. The goal of this project is to access ML models and detect the cancerous cells with a higher degree of accuracy.

## *Motivation*

The early diagnosis of Breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Accurate classification of benign tumors can assist medical practitioner in devising correct course of treatment. Machine learning (ML) algorithms are widely employed in pattern classification and forecast modelling, because of its unique advantages in critical features detection from complex datasets.

# RELATED WORK

Feature selection, in machine learning is the process of selecting a subset of the most relevant features from a number of other available feature subsets. Choosing relevant attributes is an integral part of prediction algorithms in machine learning as it plays an important role in creating a more accurate predictive model.

There are various benefits to applying the attribute selection methods such as:

1. It is more effective and faster in training the machine learning model.
2. It decreases the complexity of a model and makes it easier to interpret.
3. It improves the accuracy of an algorithm if the right subset is chosen.
4. It reduces overfitting.

Some features may have a complex interrelation between them making it difficult to select the best subset of features. Different approaches have been proposed in this project for breast cancer diagnosis [1-5]. Usually, there are three types of feature selection methods which are: filter, wrapper, and embedded methods.

# METHODOLOGY

This project will be built on Python programming language. Python platform has an extensive list of libraries and tools for building an ML project. There are number of predefined libraries, which will be employed in various stages of project, from exploratory data analysis, modelling and optimization.

Below figure will gives the brief idea of project’s workflow:

Figure 1 Workflow

Each activity is discussed as under:

## *Understanding Business Problem*

The data sample can be analyzed by ML algorithm to detect the presence of malign cell. Using different models, we will be predicting whether a person is diagnosed with cancer or not.

## *Problem Formalization*

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

In this case, based on the features like radius, texture, perimeter, area etc. given in the data set, model should predict whether it is malignant or benign where malignant means infectious and benign is not harmful. Both Supervised and Unsupervised techniques can be employed to obtain the outcomes. With python library Scikit-Learn (sklearn), different ML algorithms like Logistic regression, KNN, Decision tree, SVM can be used.

## *Requirement Analysis*

The dataset chosen to address this problem is obtained from Kaggle repository under the heading of Breast Cancer Wisconsin (Diagnostic) Data Set. The dataset is owned by University of Wisconsin, available to students and researchers under open source license. There are 32 features in data set, including

TABLE 1 DATASET DESCRIPTION

|  |  |  |
| --- | --- | --- |
|  | Feature Name | Feature Description |
|  | ID number |  |
|  | Diagnosis | M = Malign  B = Benign |

remaining features are based on real-valued features from images of cell nucleus, for example:

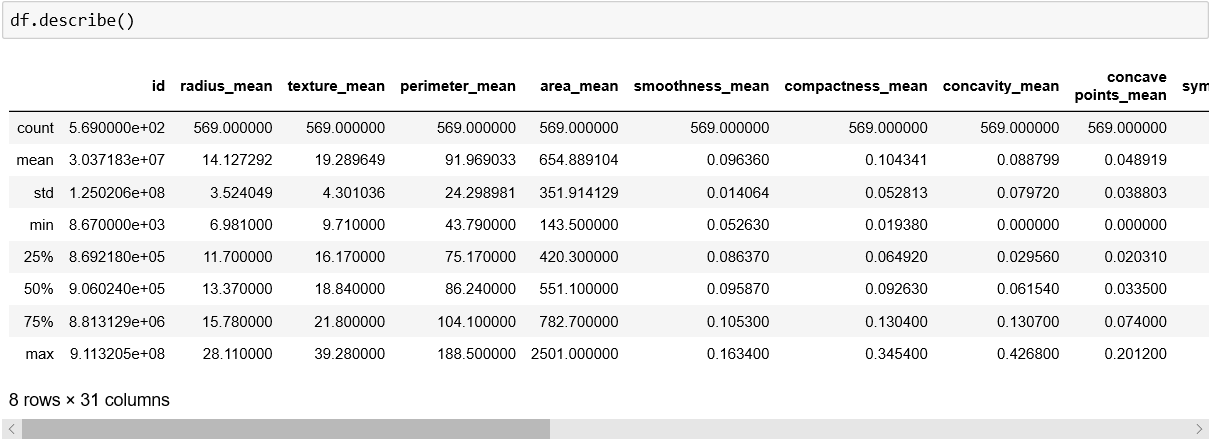
|  |  |  |
| --- | --- | --- |
|  | Radius | mean of distances from center to points on the perimeter |
|  | Texture | standard deviation of gray-scale values |
|  | Perimeter |  |
|  | Area |  |
|  | Smoothness | local variation in radius lengths |
|  | Compactness | perimeter^2 / area - 1.0 |
|  | Concavity | severity of concave portions of the contour |
|  | Concave points | number of concave portions of the contour |
|  | Symmetry |  |
|  | Fractal Dimension | "coastline approximation" - 1 |

## *Exploratory Data Analysis*

In order to summarize main characteristics of dataset EDA techniques are used. During the initial data analysis certain parameters were interrogated, viz.

1. Presence of Outliers
2. Presence of non-normal
3. Association between features
4. Pattern in data sets.

For this several analysis techniques/tools are employed, viz. Box Plot, Histogram and Scatterplot



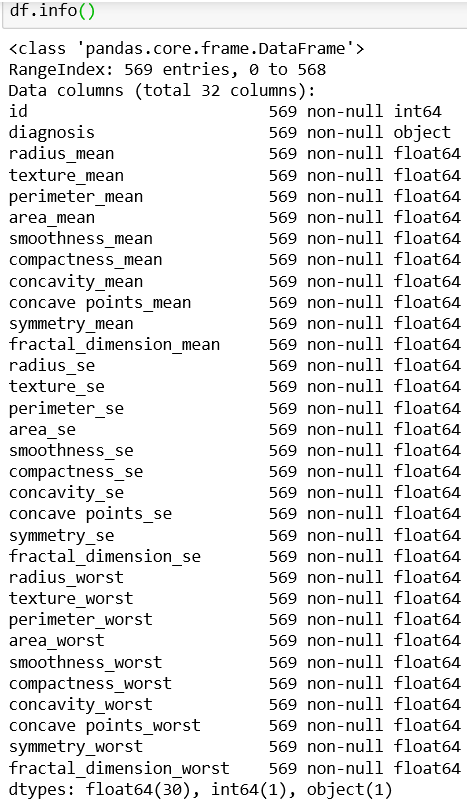


Figure 2

Inferences Figure 2:

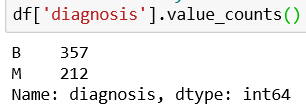
* Dataset comprises of 569 observations and 32 features
* All data are in float format except for ‘id’ which is in integer format
* No variable column has missing value

Figure 3

Inferences Figure 3:

* Mean value for all features is greater than median value (50%)
* Noticeable difference between 75% and max value for all features, this indicates presence of outliers in data set.

### *Target Value Analysis*



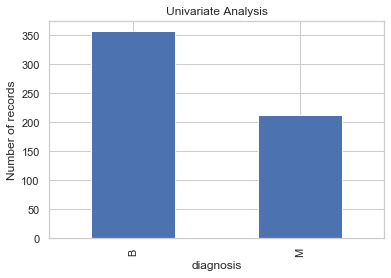


Figure 4 - Target value Analysis

Inferences Figure 4

* Target/dependent feature is categorical in nature with two possible outcomes
* With 357 Benign and 212 Malign outcomes
* Dataset has 60 to 40 ratio for outcomes, which is a balanced ratio.

### Visualizations

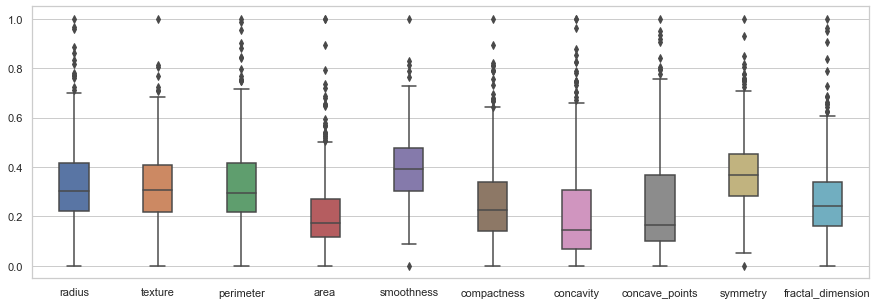
1. Correlation Matrix
2. Univariate Analysis (Single Variable)
3. Bivariate Analysis (Correlations)

Figure 5 Univariate Analysis – Box plot

1. Correlation Matrix – is done to inspect relation between varriables

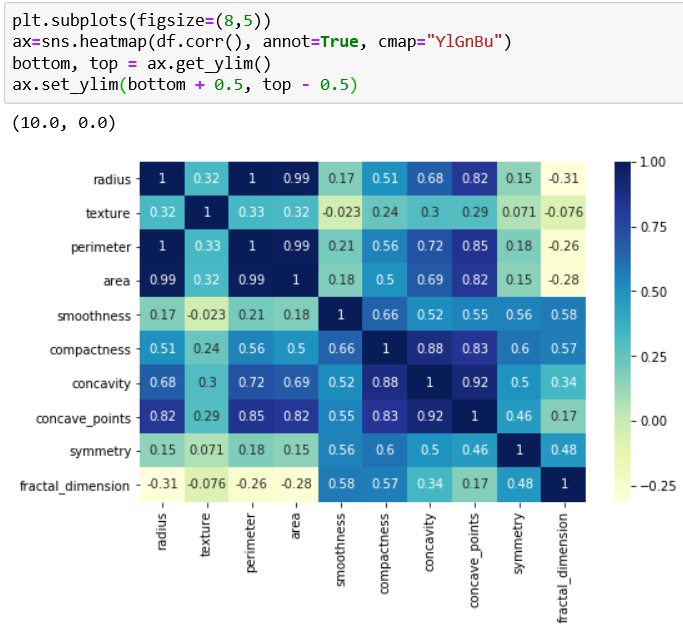


Figure 6 Correlation Matrix - using Heatmap

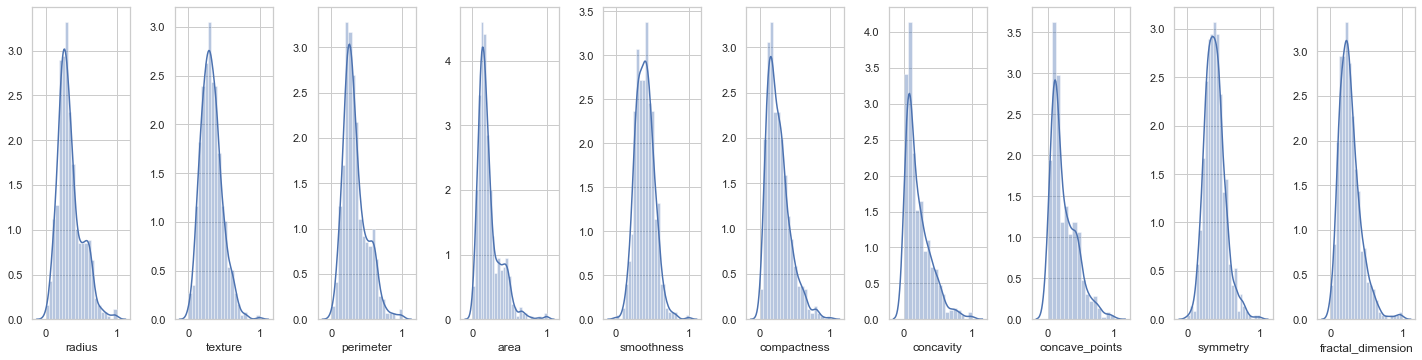
Inferences Figure 5

* Group 1: (Radius, perimeter, area, concavity and concave\_points) have strong positive correlation
* Group 2: (Texture, smoothness, symmetry, fractal\_dimension) have relatively low or negative correlation to Group 1 features

1. *Univariate analysis* – is done by performing frequency counts of different features.

Inferences Figure 6:

* Presence of outliers is in 4th quadrant, and rarely in 3rd and 1st as in case of smoothness and symmetry

Figure 7 Univariate Analysis - Histogram

Inferences Figure 7:

* Texture, smoothness and symmetry are normally distributed
* Remaining variables are skewed

1. *Bivariate analysis* is done to identify association between different features. Plot analysis between (Radius, diagnosis) and (texture, diagnosis) indicates the reason for skewness of radius and normal distribution of texture.

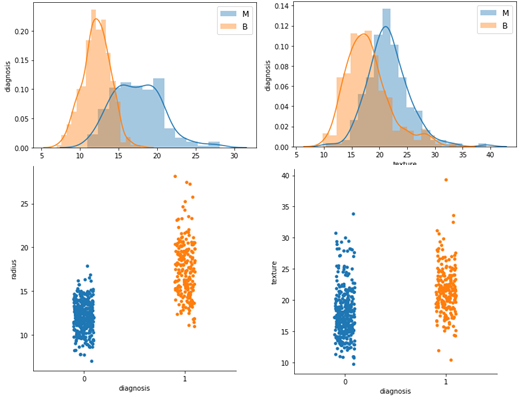


Figure 8 Bivariate Analysis - Histogram and Scatterplot

## *Data Preprocessing:*

Before starting Data preprocessing, the data quality must be inquired on parameters like:

1. Completeness (Cases of missing data): Presence of any NaN value can heavily impact the decision-making ability. There can be various methods to tackle empty or non-null values in dataset, like replacing empty values with mean, median or zero values. In our case the data consistency implies that no such operation is needed.

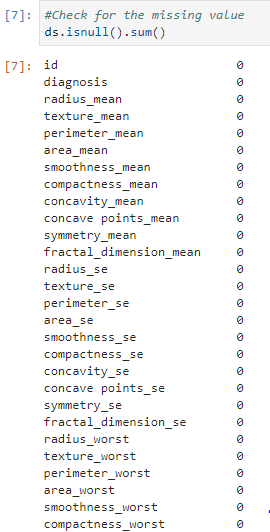


Figure 9

1. *Data Constraints***:**  i.e. on basis of Data Types (values of particular type) in a column, data range

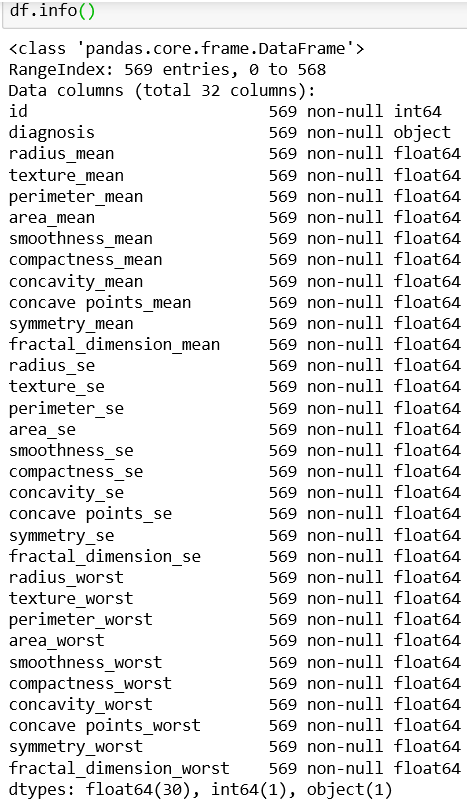
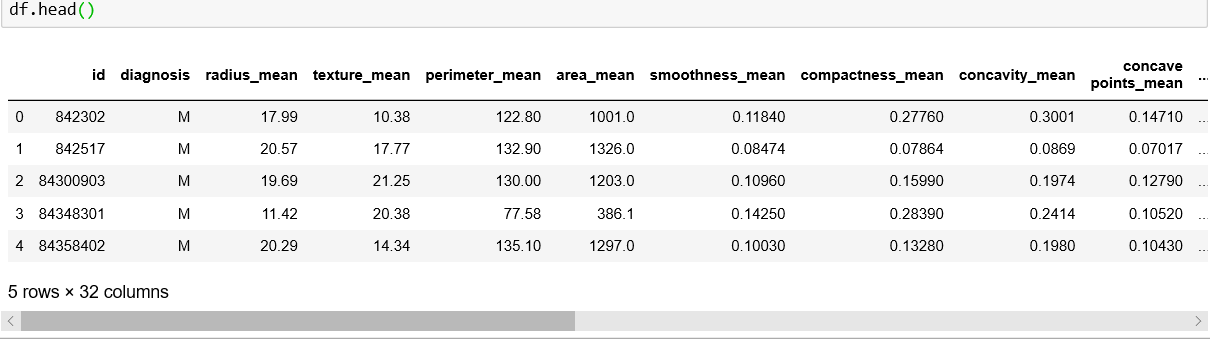
Figure 10

Figure 11

1. *Presence/Conversion of Categorical Data***:** Since most ML model work with numerical values, it is imperative to convert data to suitable forms. To ensure this the categorical feature: diagnosis is mapped to 0s and 1s for benign and malignant.

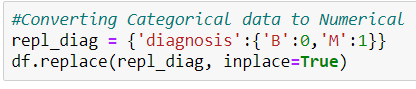


Figure 12 Conversion of Categorical Data

Data preprocessing: includes cleaning, instance selection, normalization, transformation, feature extraction and selection. Dataset doesn’t require cleaning, instance selection or transformation as there is no corrupt or missing data. However, there is a requirement for scaling/normalization of data as certain feature are in range of 0-1, while other like area and perimeter are in order of 0-1000.



Figure 13 Feature Scaling (Range 0-1)

# MODELLING (IMPLEMENTATION OF MODELS)

1. *KNN Classifier:*

KNN machine learning model is used for both regression and classification problems. It is based on feature similarity. It is widely used in industry as it requires less calculation time and easy interpreting. KNN is used in our project because in our dataset we have two cases either the cells are malignant or benign. And we are classifying it on the basis of different parameters.

To implement KNN it is very important to choose right value of k as it has very high impact on the models prediction.

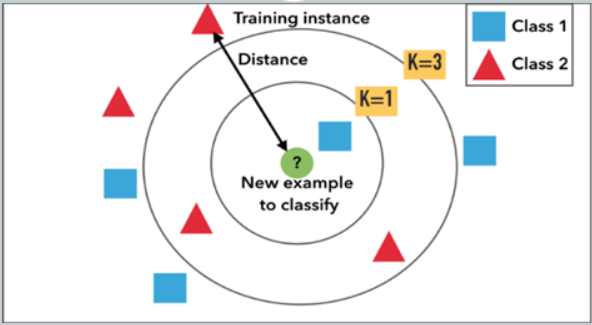


Figure14

It can be seen in the figure there are two classes and the new data point will be classified depending on the value of k.

If k=1 the only one data point near to new data will be consider and hence it will be classified as class one. But, if k=3 then three nearest data points will be considered and it will be classified as class 2.

Therefore, value of k plays very important role in the model.

Steps to Implement a KNN model:

1. Loaded the data and divided in training and testing set.
2. Changed categorical values of diagnosis to 1 for malignant and 0 for benign.
3. For getting the value of k did parameter tuning by iterate from 1 to 50 with step size of 2

* Calculated the accuracy, score and confusion matrix for each value of k and stored in an array.
* Used ‘cross\_val\_score’ method to cross validate k value with different folding values.

1. Plotted a graph of accuracy vs. Value of k shown in figure and finalized the best k value for our dataset i.e. k=13

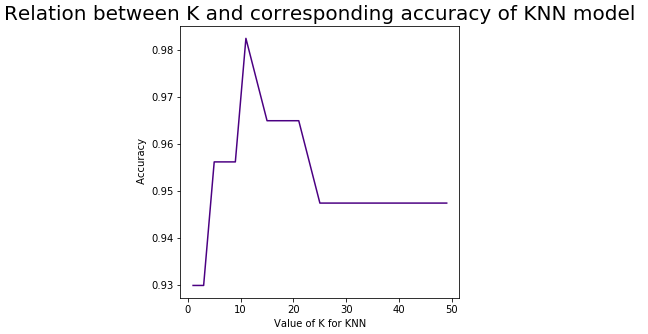


Figure 15

1. Did performance evaluation of the model by calculated confusion matrix, accuracy and precision.

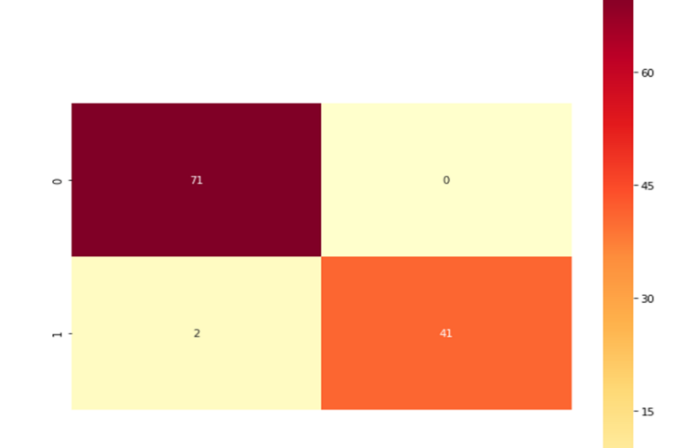


Figure 16

Confusion matrix is one of the performance evaluation matrixes of machine learning models.

For our model we got 71 True Negative(TN) values which means we predicted no and the patient doesn’t have the cancer then, we got 41 True Positive(TP) values that means we predicted yes and the patient do have the cancer. Whereas, our model calculated two False Negative(FN) values means model predicted no but patient actually have the cancer and zero False Positive(FP) values, model predicted yes but actually patient doesn’t have the cancer.

Using the values from confusion matrix we calculated accuracy and precision of the model using below formulas.

Accuracy = (True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative) **(0.9824561403508771)**

Precision = True Positive / (True Positive + False Positive) **(1.0)**

1. *Random Forest Classifier:*
2. *SVM Classifier:*

It is a type of supervised machine learning classification algorithm. It constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. We have used three different kernels to fit our model i.e. Linear, Polynomial and Gaussian. After applying these three kernels, we have compared them and identified the most accurate kernel for our dataset. Before implementation, we divided the dataset into training and testing set using ‘train\_test\_split’ and then changed the categorical values of ‘diagnosis’ feature to 1 for malignant and 0 for benignant as part of preprocessing.

After this, the model was trained using the three different kernels (Linear, Polynomial and Gaussian) and accuracy and RMSE was calculated for each of them. We have also used cross validation to calculate RMSE for cross validation using ‘cross\_val\_score’ with 5 folds to compare for each kernel.

1. *Linear Kernel:* For the Linear kernel, we have calculated an accuracy of 0.9415204678362573 and RMSE was calculated as 0.2418254167033372

The RMSE for cross validation using 5 folds was obtained as 0.3130596254931703

The following confusion matrix was obtained:

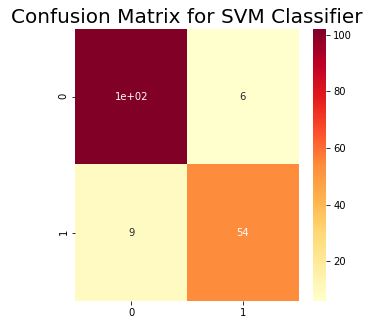


Figure 17

We have obtained 102 values for True Positive, 6 for False Positive, 9 for False Negative and the remaining 54 for True Negative.

1. *Polynomial Kernel:* For the Polynomial kernel, we have calculated an accuracy of 0.6666666666666666 and RMSE was calculated as 0.5773502691896257

The RMSE for cross validation using 5 folds was obtained as 0.6576473218982952

The following confusion matrix was obtained:

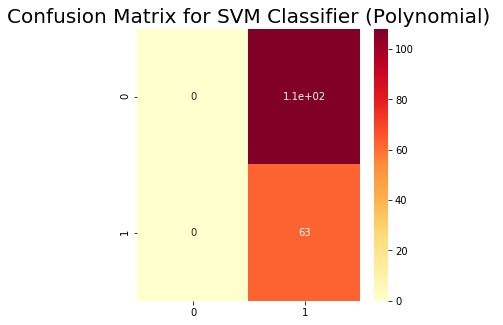


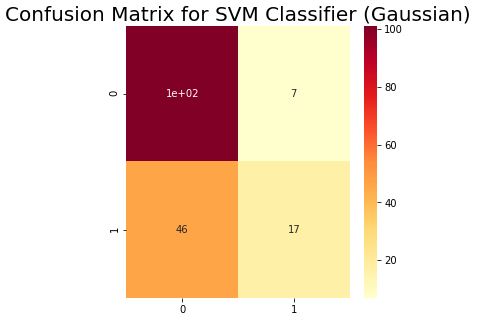
Figure 18

We have obtained 0 values for True Positive, 102 for False Positive, 0 for False Negative and the remaining 63 for True Negative.

1. *Gaussian Kernel:* For the Gaussian kernel, we have calculated an accuracy of 0.7017543859649122 and RMSE was calculated as 0.5461186812727502

The RMSE for cross validation using 5 folds was obtained as 0.5695045592776011

The following confusion matrix was obtained:



*Figure 19*

We have obtained 102 values for True Positive, 7 for False Positive, 46 for False Negative and the remaining 17 for True Negative.

*Comparing Linear, Polynomial and Gaussian Kernels:*

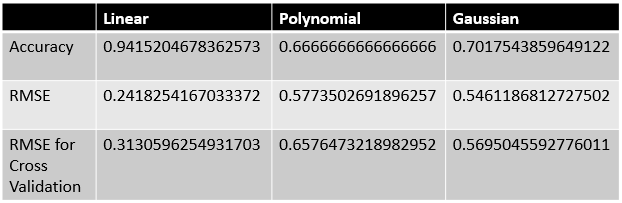


Table 3

In the above table, it can be clearly seen that the Linear kernel gave us the best results as accuracy for Linear kernel was the best among the three kernels and if we look at the RMSE and RMSE for Cross Validation, Linear kernel has the least values which makes it the best among other kernels. On the other hand, the polynomial kernel gave us the worst results having an accuracy of 0.66 and higher values of RMSE and RMSE for Cross Validation as compared to Linear and Gaussian Kernels.

# MODEL EVAULUATION

The aim of the project was to design and develop a machine learning model to predict the breast cancer by classifying cells as benign or malignant depending on the features of cells. Three different machine learning models were implemented and compared based on different evaluation matrix of ML models to find highly tuned model which can predict with less errors, higher accuracy and reliability.

The object of the project has been accomplished by implementing ML model and tuning the parameter to increase the accuracy levels of prediction.

The following table shows the comparison of different models applied to dataset:

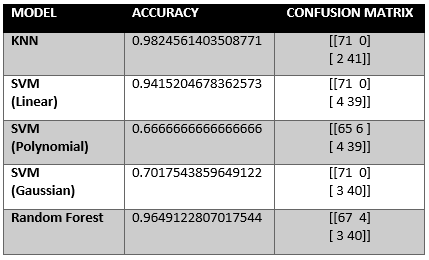


Table 4

# OBSERVATION

After comparing three algorithms it can be concluded that KNN performs better than SVM and Random forest as it has highest accuracy. As shown, parameter tuning gives better predication/output. It is important to choose the right combinations of parameter values to get higher accuracy.

# REFERENCES

1. M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240–3247, 2009.
2. Y. Sun, C. F. Babbs, and E. J. Delp, “A comparison of feature selection methods for the etection of breast cancers in mammograms: adaptive sequential floating search vs. genetic algorithm,” in *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pp. 6532–6535, Shanghai, China, September 2005.
3. B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.
4. E. Aličković and A. Subasi, “Breast cancer diagnosis using GA feature selection and Rotation Forest,” *Neural Computing and Applications*, vol. 28, no. 4, pp. 753–763, 2017.
5. M. Banaie, H. Soltanian-Zadeh, H.-R. Saligheh-Rad, and M. Gity, “Spatiotemporal features of DCE-MRI for breast cancer diagnosis,” *Computer Methods and Programs in Biomedicine*, vol. 155, pp. 153–164, 2018.
6. <https://stackoverflow.com/questions/19629331/python-how-to-find-accuracy-result-in-svm-text-classifier-algorithm-for-multil>