B.M.S College of Engineering

P.O. Box No.: 1908 Bull Temple Road,

Bangalore-560 019

DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING

Course – Technical Seminar

AY 2019-20

Report on Seminar Project

BREAST CANCER PREDICTION USING MACHINE LEARNING

Submitted to Dr. Sheela S.V Ma'am

Submitted by

PRATEEK JAIN 1BM18IS149

B.M.S College of Engineering

P.O. Box No.: 1908 Bull Temple Road
Bangalore-560 019

ABSTRACT

Breast cancer represents one of the diseases that make a large number of deaths every year. The second major cause of women's death is breast cancer (after lung cancer). A number of machine learning algorithms have been used to develop a prediction model . Among them Logistic Regression , Decision Tree , KNN , SVM are the most commonly used techniques . However , there have been very few studies about the performance of SVM based on kernel functions used in the breast cancer prediction . Therefore, the aim of this study is to fully assess the prediction performance of these algorithms in breast cancer prediction . The experimental results show that the RBF kernel of SVM and KNN outperforms all the other classifiers . .

1 .1 INTRODUCTION

Breast Cancer is a very common disease in women all over the world . It's the main cause of women's death all over the world[1] .This cancer develops in the breast tissue . They are several risk factors for breast cancer including female sex , obesity , lack of physical exercise , drinking alcohol , hormone replacement therapy during menopause , ionizing radiation , early age at first menstruation , having children late or not at all , and older age .

We can measure the seriousness of this disease by this report given by Siegal et al.that Breast cancer contributes around 12% of all new cancer cases and 25 % of all cancers in women[2] and that's why breast cancer prediction becomes an important research problem both in the medical as well as in healthcare communities .

Many machine learning algorithms and statistical techniques have been employed to develop a large variety of breast cancer prediction models. Some of the most commonly used techniques are Logistic Regression , Decision Trees , K nearest Neighbours (KNN) , Support vector machine (SVM) , etc . In this study the performance of these algorithms have been analysed and compared and then concluded which algorithm is best suited for this kind of analysis . The main focus of this study is the SVM algorithm since accuracy of an SVM model depends largely on kernel functions used . These Kernel functions include Linear , RBF (Radial Basis Function) , Poly and Sigmoid functions .

The evaluation metrics used for checking the accuracy of a model are f1-score , confusion Matrix and accuracy score .

2 LITERATURE SURVEY

All the algorithms used in this study have been discussed below : -

2.1 LOGISTIC REGRESSION

Logistic Regression is one of the most fundamental classification algorithms used in ML . Logistic Regression can be used for both binary classification as well as multi-class classification . In fact , logistic regression predicts the probability of different samples and then these samples are mapped to a discrete class based on that probability .

It uses Logistic / Sigmoid Function at its core , that's why it has been named Logistic regression [3].
The curve for this function looks as shown below[4] :

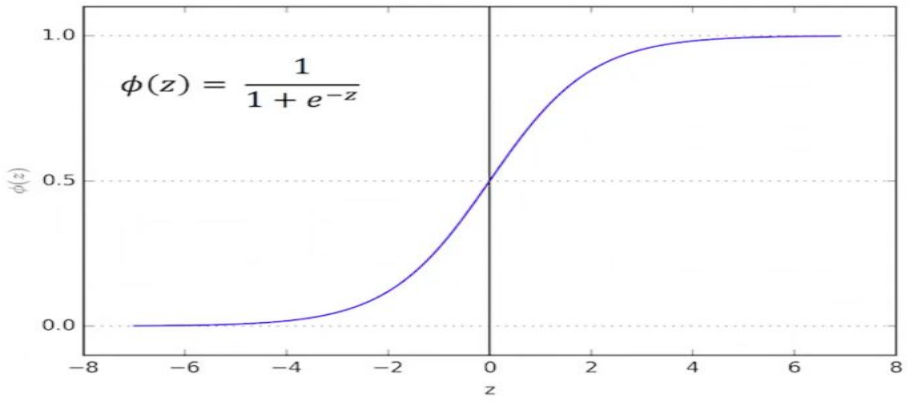
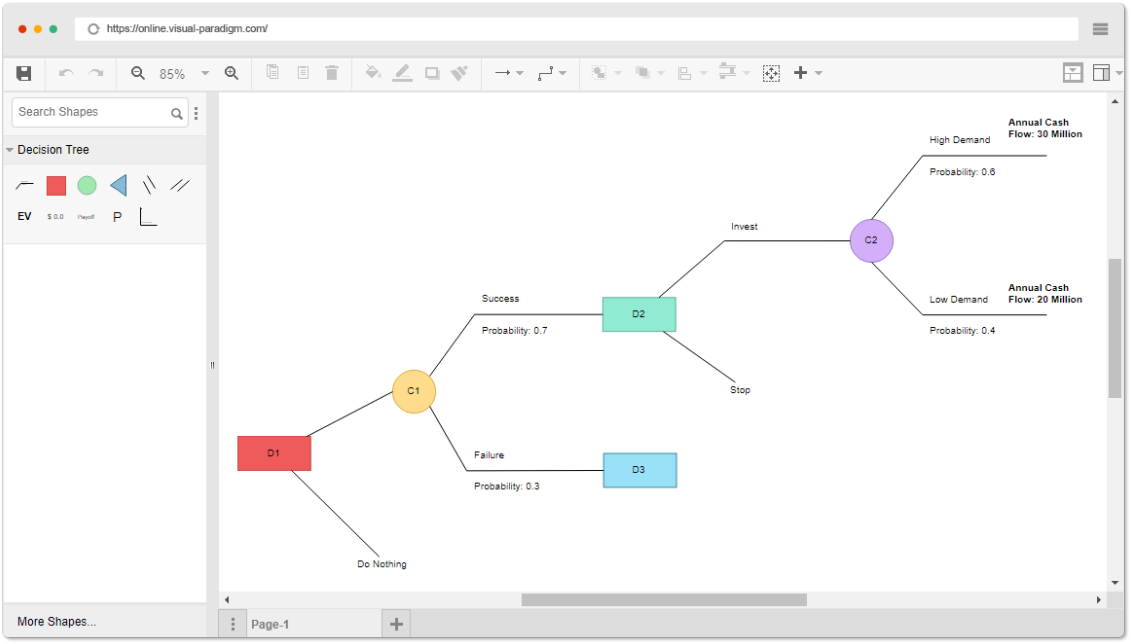


figure 2.1. Logistic curve and Logistic function

Lundin et al. used logistic regression to develop the prediction models using a large dataset (more than 200,000 cases)[5] .

2.2 DECISION TREE

The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree .
Rokach et al. explains that decision trees are built by splitting the training set into distinct nodes where one node contains all of or most of one category of the data .
It uses recursive partitioning to classify the data[6] .
Below is an example showing how decision trees can be implemented for predicting the drug prescription based on the features .



The algorithm for decision tree has been discussed below :

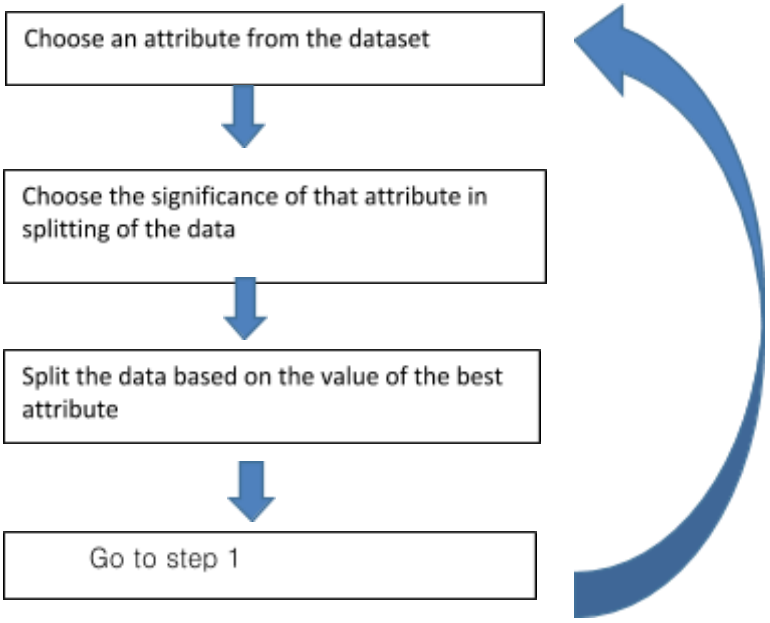


figure 2.3 Decision Tree algorithm

2.3 K-Nearest Neighbors(KNN)

The K-Nearest neighbor algorithm is a classification algorithm that takes a bunch of labelled points and uses them to learn how to label other points[7] .
This algorithm classifies cases based on their similarity to other cases . Data points that are near each other are called neighbors .
So , k in k nearest neighbours stand for the number of nearest neighbors to examine .
Below example shows how labelled points are grouped into different labels .

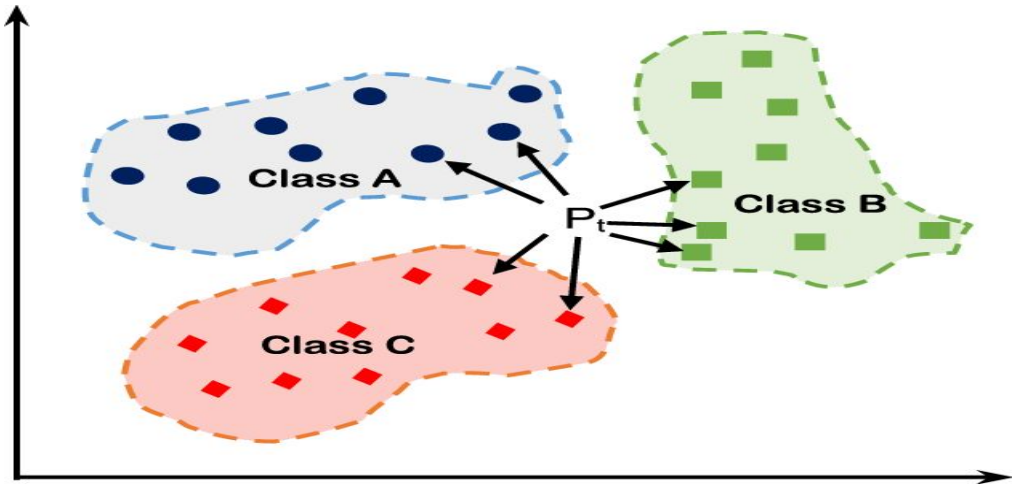


figure 2.4. Grouping of labelled points in KNN

The algorithm for KNN :

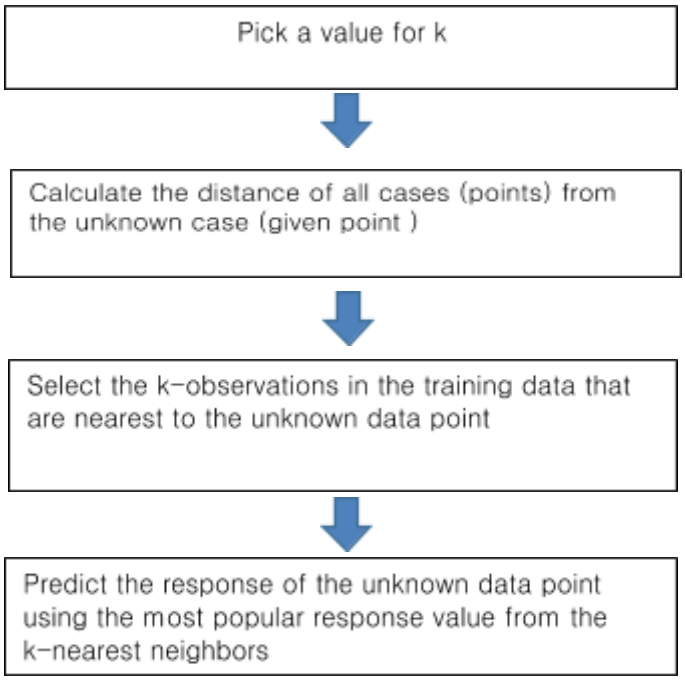


figure 2.5 . KNN algorithm

As pointed out by Bhatia and Vandana et al. K value plays a major role in determining how accurate our model is. So , what value of K to select becomes a big issue here[8]. For very low values of K , it is considered that noise has been captured in the data or one of the points that was an anomaly in the data has been chosen. On the other side of the spectrum , a higher value of K causes the model to become overly generalized .

2.4 SUPPORT VECTOR MACHINE (SVM)

Support vector machine is a supervised machine learning technique that uses classification algorithms for two-group classification problems by finding a separator. In SVM , we plot each data item as a point in n dimensional space (where n is the number of features) with the value of each feature being the value of a particular coordinate . Then a hyperplane is found which differentiates the two classes . Anything that falls on one side of a hyperplane will be considered in one class while anything that falls on the other side will be considered in the other class . The diagram below shows a hyperplane used for separating two classes .

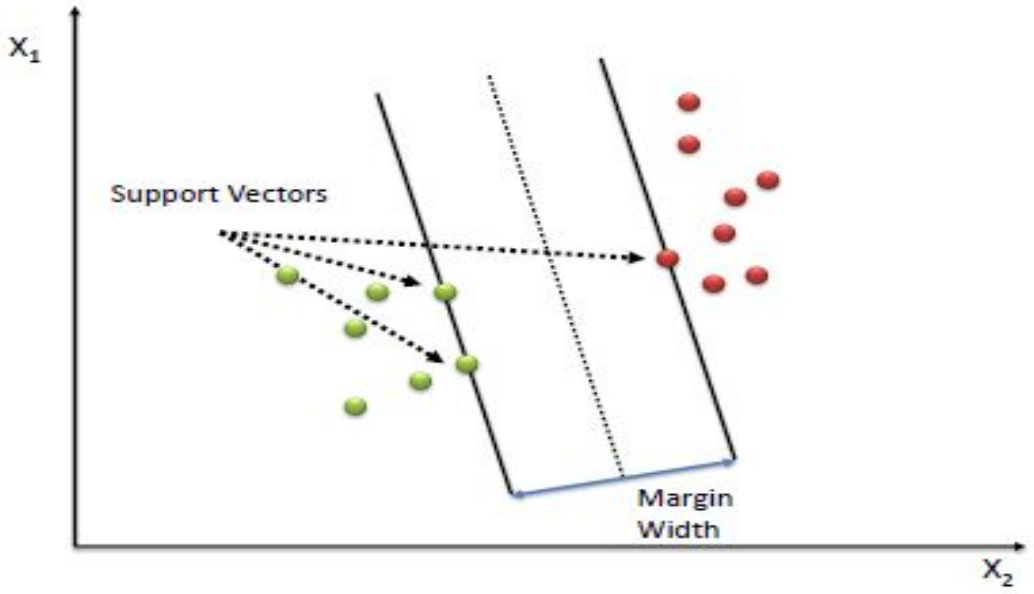


figure 2.. A hyperplane separating two classes

There can be many possible hyperplanes . But the one whose distance to the nearest points of each tag is the largest are considered the best hyperplane .
For finding the best hyperplane , support vectors are used .

Most of the time finding the hyperplane becomes quite a tough job since not every time all data points can be separated by linearly separable lines .Therefore the concept of kernel functions comes into the picture[9] . Hofmann et al. has put Kernelling basically means mapping data into a higher-dimensional space and the mathematical functions that are used for this purpose are called kernel functions[10] . There are many kernel functions out there . Linear , RBF (Radial basis functions) , Poly , and Sigmoid are some of the most used kernel functions[11] .

2.2 RELATED WORK

Many machine learning models have been created using the above algorithms . There has a number of research works been done which deal with the performance of different algorithms being used in breast cancer prediction .

One such work has been carried out by Asri , Mousannif et al. where they have compared the performance of four classifiers : Support vector machine (SVM) , C4.5 , Naive Bayes (NB) and K Nearest Neighbours (KNN) on the Wisconsin Breast cancer dataset . Not only this they have evaluated the efficiency and effectiveness of those algorithms in terms of accuracy , sensitivity , specificity and precision as well . SVM proved to be the most effective, outperforming the rest of them by reaching the highest accuracy of 97.13 while C4.5, Naive Bayes and KNN had accuracies that varied between 95.12 % and 95.28 % . In addition , SVM achieved the best performance both in terms of precision as well as in terms of low error rate[12] .

Another work has been done by Vikas et al. where they have used Decision tree , RBF kernel and Logistic Regression algorithms on the dataset obtained from University Medical Centre , Institute of Oncology , Ljubljana . They conducted this experiment using libraries obtained from the Weka machine learning environment . They concluded that simple Logistic Regression performed better than all with an accuracy of 74.47 %[13] .

Ahmad et al. used three very popular machine learning techniques , Decision Tree (C4.5) , Support vector machine (SVM) and Artificial Neural Networks (ANN) , and they compared the performances these algorithms through sensitivity , specificity and accuracy on the dataset obtained from Iranian centre for Breast cancer (ICBC) . The predictions given by the SVM model were 95.7 % accurate which is more as compared to the predictions obtained by other algorithms[14] .

Delen et al. used Artificial Neural Networks (ANN) , Decision Trees along with Logistic regression to develop the prediction models on a large dataset which contained more than 200,000 instances . They also made use of 10-fold cross validation in order to measure the unbiased estimate of three models obtained and carried out their performance comparison . Their results concluded Decision Tree to be the best predictor which gave an accuracy of 93.6 % followed by ANN and Logistic classifier which predicted the cases with the accuracies 91.2% and 89.2% respectively[15] .

Another remarkable work has been done by Huang et al. on analyzing the accuracies of SVM classifiers based on the different kernel functions used . Analyzing the prediction performances of SVM and SVM ensembles over small and large datasets is also their primary focus . They used 10-fold cross validation for better estimates of a model , and feature selection as a preprocessing step . They concluded that for a small scale dataset , linear kernel based SVM ensembles based on the bagging method and RBF kernel based SVM ensembles with the boosting method are the better choices where feature selection should be performed in the data pre-processing stage. But , for a large scale dataset, RBF kernel based SVM ensembles based on the boosting method perform better than the other classifiers[16] .

3. SYSTEM REQUIREMENT

All the work has been carried out on an i5-8th generation computing machine .
Software used - Jupyter notebook
Programming language used - Python
All the classifier models have been built and their comparisons have been carried out using Sci-Kit learn , Numpy , Pandas , Seaborn and Matplotlib
APIs of Python .

4 . EXPERIMENT METHODOLOGY AND DESIGN

We conducted multiple experiments to compare the performance of the different classifiers and then analysed their performance .

4.1 Datasets Used

The dataset used in this experiment is The Wisconsin Breast Cancer (original) dataset from the UCI Machine Learning Repository . It contains 569 instances
and has 33 features . Out of which 357 samples are benign and 212 samples are malignant . 2/3rd of the data has been used for training the model and 1/3rd has been used for testing the model .

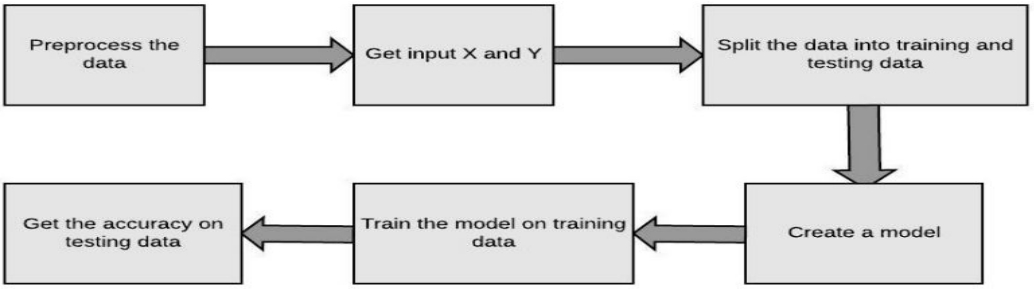


figure 4.2 Flow chart

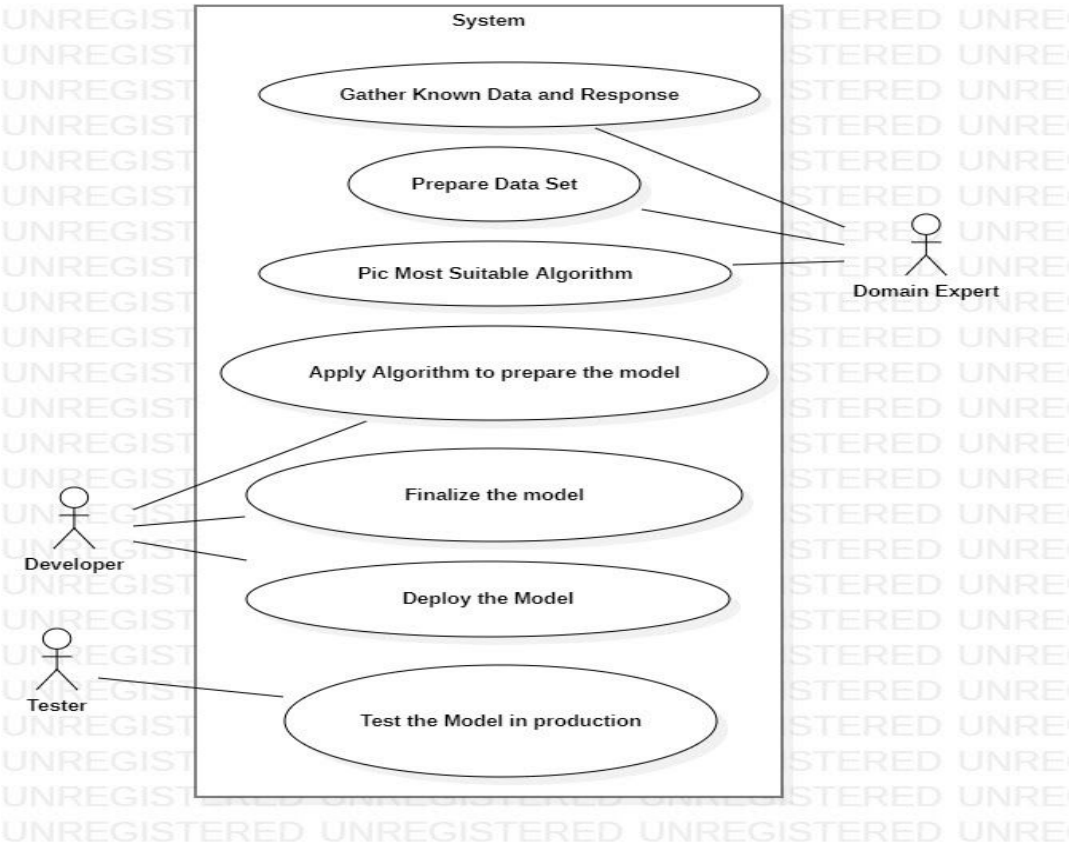


figure 4.2.2 Use case Diagram

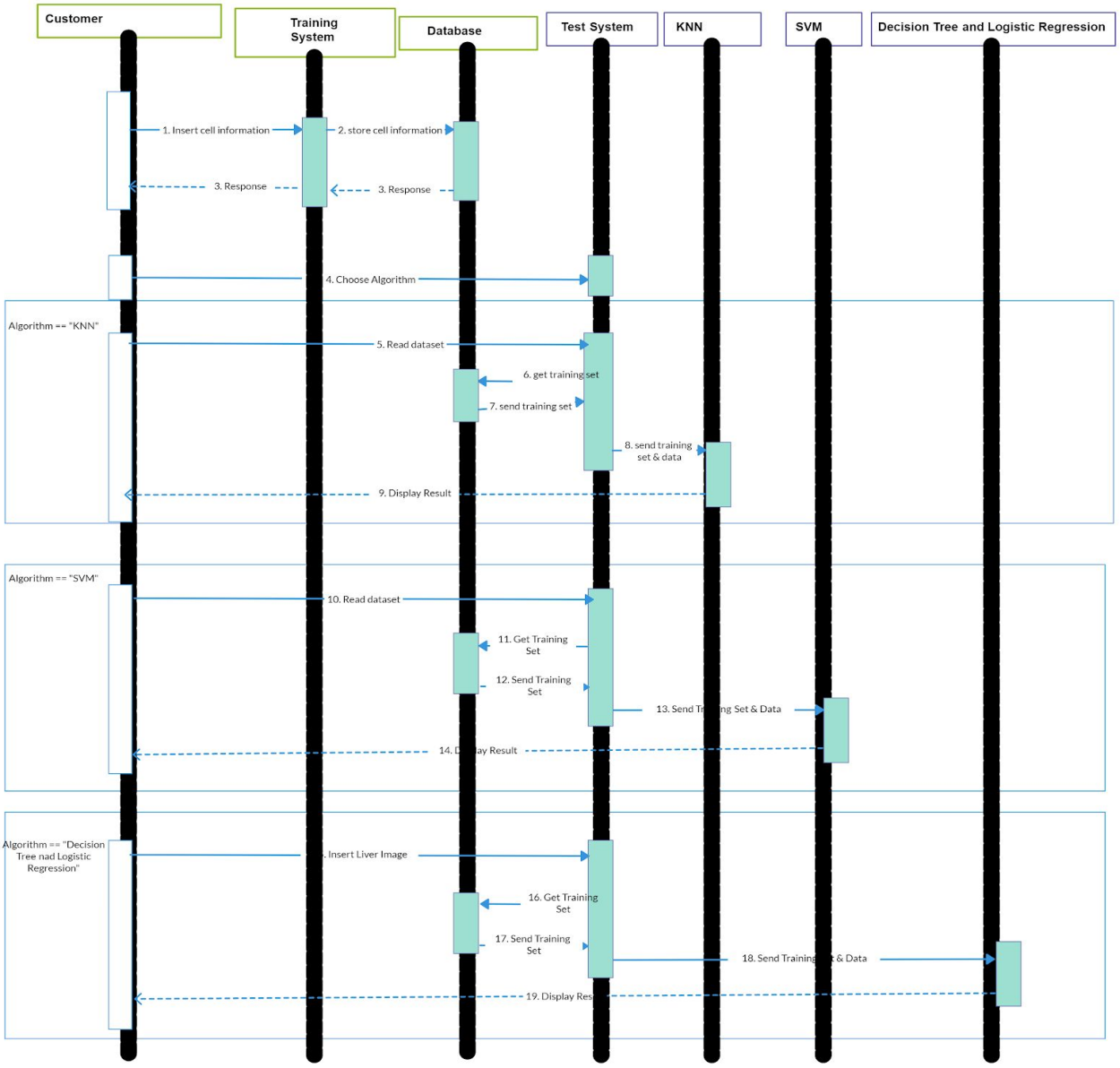


figure 4.2.3 Sequence diagram

4.2 Experimental Procedure

The data that we had can not be directly used to train the models . Therefore, we must pass it through the preprocessing step . In the preprocessing step , we scaled our input data in such a way that it has a mean of 0 and step deviation of 1 . This is required because features having more variance can dominate other features having low variance . So , to make sure this does not happen we passed it to the preprocessing step .

First part that we did in this experiment is to compare the performance of Logistic Regression , Decision Tree and K nearest Neighbors . The first step was to clean the

data to get rid of any missing values . In the next step we preprocessed the data . In the preprocessing step the input data has been scaled in such a way that it has a mean of zero and a step deviation of one. This is to make sure one feature doesn't dominate over other features . In the following step the input data is split into training and testing data then individual models based on these techniques are created . First we trained each of these models on training data then we checked their accuracy score , confusion Matrix , f1-score .
Second part of the first analysis is based on comparing the performance of Support vector machines (SVM) based on the kernel functions used .So , here also individual models based on Linear , RBF , Poly and Sigmoid kernel functions are created . After splitting the dataset into training and testing data each of these models were trained on the training data and then we calculated their accuracy score on the testing data . The first analysis has been done without making use of cross-validation .

5 . Experimental Results and DISCUSSION

In the first part of the first analysis highest accuracy is shown by KNN for k = 12 . It's accuracy stands at 99.47% . Logistic regression gives an accuracy of 96.27% . For Decision Tree , accuracy is 94.14% for depth = 5 . This comparison has been shown in fig 5.1 .

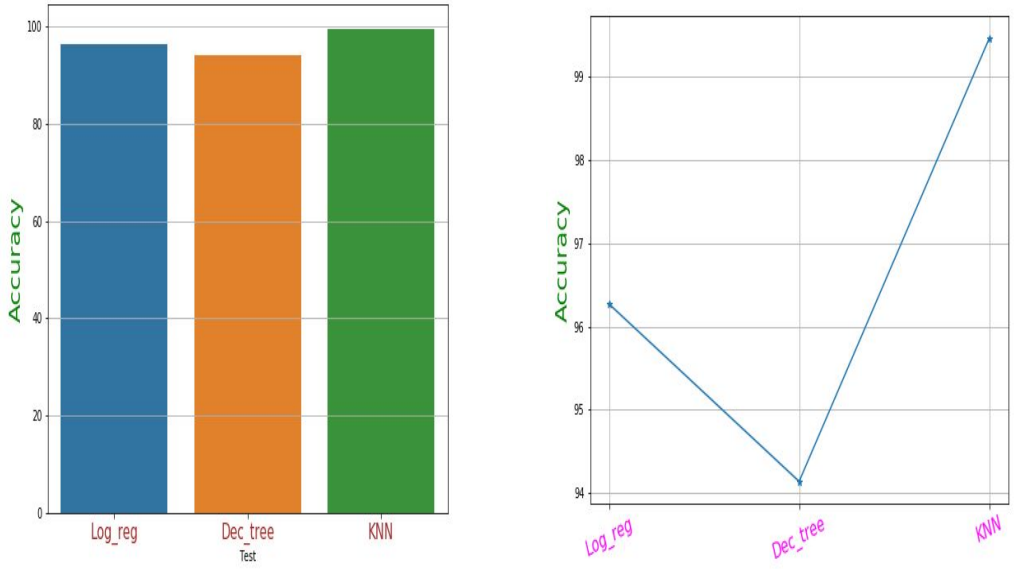


Fig 5.1 (a)Bar graph comparison , (b) comparison by plot of Log_reg , DT and Knn

In the second part , we compared different kernels of SVM . RBF kernel gives the highest accuracy of 99.47% . This is followed by the Sigmoid kernel which has done the prediction with an accuracy of 97.87% . Then linear and Poly kernels have predicted 96.28% and 95.21% of the test data correctly respectively .
Figure 5.2 gives a pictorial representation of this comparison .

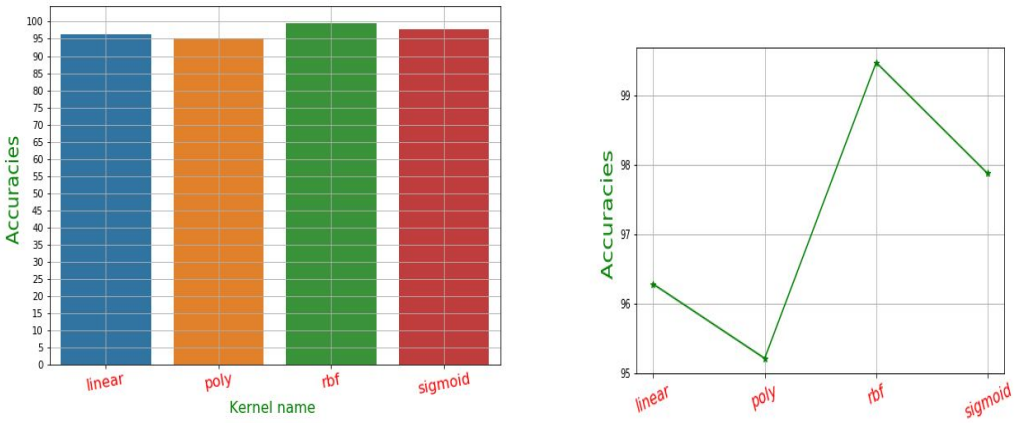


Fig 5.2. (a) Bar graph comparison , (b) comparison by plot of different kernels

6 . CONCLUSION

The experiment that we conducted on the Wisconsin Breast Cancer dataset gives the result that feature selection has improved the accuracy of all the models .
Highest accuracy is given by the RBF kernel of the SVM classifier along with that the KNN also performs extremely well . In conclusion , these two classifiers have proven their efficiency in Breast Cancer prediction and diagnosis and achieves the best performance in terms of accuracy .

7. FUTURE ENHANCEMENTS

We can further add cross validation and feature selection steps to further enhance the accuracy of these models . Along with that we can make use of classifier ensembles to further extend this project .

8. REFERENCES

[1] US Cancer Statistics Working Group. "United States cancer statistics: 1999–2012 incidence and mortality web-based report." Atlanta (GA): department of health and human services, centers for disease control and prevention, and national cancer institute (2015).
[2] Siegal, Rebecca, K. D. Miller, and Ahmddin Jemal. "Cancer statistics, 2012." Ca cancer J clin 64.1 (2014): 9-29.
[3-5] Tolles, Juliana, and William J. Meurer. "Logistic regression: relating patient characteristics to outcomes." Jama 316.5 (2016): 533-534. doi:10.1001/jama.2016.7653.ISSN 0098-7484. OCLC 6823603312. PMID 27483067
[6] Rokach, Lior, and Oded Z. Maimon. Data mining with decision trees: theory and applications. Vol. 69. World scientific, 2008.
[7] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." The American Statistician 46.3 (1992): 175-185.

[8] Bhatia, Nitin. "Survey of nearest neighbor techniques." arXiv preprint arXiv:1007.0085 (2010).
[9-11] Hofmann, Thomas, Bernhard Schölkopf, and Alexander J. Smola. "Kernel methods in machine learning." The annals of statistics (2008): 1171-1220.
[12] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.
[13] Chaurasia, Vikas, and Saurabh Pal. "Data mining techniques: to predict and resolve breast cancer survivability." International Journal of Computer Science and Mobile Computing IJCSMC 3.1 (2014): 10-22.
[14] Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., Razavi, A. R., & Ahmad, L. G. (2013). Using three machine learning techniques for predicting breast cancer recurrence. J Health Med Inform, 4(2), 124.
[15] Delen, Dursun, Glenn Walker, and Amit Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods." Artificial intelligence in medicine 34.2 (2005): 113-127.
[16] Huang, Min-Wei, Chih-Wen Chen, Wei-Chao Lin, Shih-Wen Ke, and Chih-Fong Tsai. "SVM and SVM ensembles in breast cancer prediction." *PloS one* 12, no. 1 (2017).