BEST FEATURE SELECTION AND DATA VISUALIZATION FOR BREAST CANCER PREDICTION

by

Tanjim Ahmed Hemel
14301013
Rohan Parvez
14101199

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
September 2019

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Tanjim Ahmed Hemel	Rohan Parvez
14301013	14101199

Approval

The thesis/project titled "BEST FEATURE SELECTION AND DATA VISUALIZATION FOR BREAST CANCER PREDICTION" submitted by

1. Tanjim Ahmed Hemel (14301013)
2. Rohan Parvez (14101199)

Of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 28, 2019.

Examining Committee:

Supervisor:

Dr.Md. Iftekharul Mobin
Assitant Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:

Dr.Md. Ashraful Alam
Assitant Professor
Department of Computer Science and Engineering
BRAC University

Chairperson:

Dr. Mahbub Alam Majumdar
Professor
Department of Computer Science and Engineering
Brac University

Abstract

Breast cancer (BC) is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer-related deaths according to global statistics, making it a significant public health problem in today's society. In medical diagnosis, the forecast of an infection goes about as a significant center in breaking down the therapeutic pictures. The undesirable cell development in any piece of the organ is known as tumor. The tumor might be favorable or harmful. Threatening tumor is viewed as the most risky tissue. There are different specialists learned about the forecast of bosom malignancy. This paper aims to review on various data set techniques that are specifically considered on breast cancer prediction and also to investigate which feature set is responsible for the disease and rapid growth of cancer cells as we are selecting the best features. From primarily given data set we can measure which parameter is responsible for cancer cells and which features can make the nearest perfect outcomes. It is presently possible to make precise Computer Aided Diagnosis (CAD)[7] framework so as to make the whole procedure of distinguishing a dangerous tumor more asset proficient and efficient through appropriate usage. This paper displays the relative investigation of various machine learning calculations and their outcomes in anticipating destructive tumors. For example, Decision Tree, Support Vector Machine, K-Nearest Neighbors, Linear Discriminant Analysis, Naive Bayes and Logistic Regression with and without PCA on a dataset with 30 highlights removed from a digitized picture of a Fine Needle Aspirate (FNA)[19] of a breast mass. Profound learning models like Artificial Neural System and Convolutional Neural Network are utilized and their exhibitions are looked at.

Keywords: Computer Aided Diagnosis, Convolutional Neural Network, Support Vector Machine, Breast Cancer Detection, Logistic Regression, Random Forest, K-Nearest Neighbours, Naive Bayes, PCA, FNA, Artificial Neural Network

Acknowledgement

First of all we want to express gratitude toward Almighty Allah to empower us to take a shot at this wok which has been an extraordinary learning background for us. By the grace of Allah, we could able to put our best efforts and effectively complete it on schedule. Also, we might want to pass on our appreciation to our thesis supervisor Dr. Iftekharul Mobin for his direction and bunch commitment all through the entire period of our thesis work and furthermore to compose this report. From the earliest reference point as far as possible of the work he has given every one of us sorts of assistance and roused us to push ahead to our objective. We might likewise want to recognize the help that we got from a number of assets over the Internet particularly from crafted by our individual looks into. At last, we might want to say thanks to BRAC University for giving us the chance to finish up the thesis and for allowing us to finish our Bachelor qualification.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The second major cause of woman's death is breast cancer (after lung cancer) 1. 246,660 of woman's new cases of invasive breast cancer are expected to be diagnosed in the US during 2016 and 40,450 of women's death is estimated. Breast cancer represents about 12 percent of all new cancer cases and 25 percent of all cancers in women[7]. Information and Communication Technologies (ICT) can expect potential occupations in malady care. Believe it or not, Big data has advanced the size of data just as making a motivator from it. Immense data, that transforms into a synonymous of data mining, business analytics, and business understanding, has revealed a noteworthy improvement in BI from uncovering and the decision to desire results. Data mining moves close, for instance, associated with therapeutic science focuses rise rapidly due to their prevalent in predicting results, decreasing costs of prescription, propelling patients' prosperity, improving human administration's worth and quality and in choosing consistent decision to save people's lives. There are numerous calculations for order and expectation of bosom malignant growth results. Random Forest, Naive Bayes, KNN, Logistic Regression and SVM with and without PCA on an informational index. Thirty highlights expelled from an image of a FNA of a bosom mass. Significant learning models like Artificial Neural System and CNN are used and their shows are taken a gander at. Comparative assessment, it is seen that the significant learning models outmaneuver every single distinctive classifier and achieves imperative scores over various presentation estimations, for instance, Exactness of 98.83 percent, Precision of 98.44 percent and Recall of 100 percent.

1.1 Motivation

Numerous individuals are now a days influenced from breast cancer. The main reason of this disease relies upon many circumstances and cannot be essentially decided. Furthermore, the distinguishing proof technique that decides if the disease is kind or harmful also needs an amazing arrangement of exertion from a specialists and doctors. Once many tests are concerned within the identification of breast cancer, like cluster breadth, consistency of cell size, consistency of cell form, etc. The ultimate result could also be troublesome to get even for doctors. The ailments that end various lives, demonstrative PC based applications are utilized wide. Mechanical technology is partaking in an outrageously fundamental job in operational rooms. Additionally, the gifted frameworks are presented inside the concentrated treatment

rooms. Thus, utilizing another side of artificial knowledge for bosom malignant growth assignment isn't contemptible. It's accounted for that bosom malignant growth sickness is the 2nd most occurring disease which influences young ladies, this overflowing disease globally spreading from 2003[7]. The malignant growth might be a very normal kind of disease among young ladies and along these lines the second most elevated purpose for disease passing. Inside the United State, with respect to one of every eight young ladies over their time frame incorporates a danger of creating bosom disease. With the uncontrolled division of one cell inside the bosom prompts starting to the bosom malignant growth and it brings the outcome in a noticeable pile, called a tumor. It is kindhearted or dangerous. Along these lines, the need for exact grouping inside the facility might be a clarification for decent worry for pros and specialists. This significance of machine learning has been actuated for the last twenty five years, once scientists began to comprehend the nature of taking bound choices to treat explicit infections. The work of machine learning and data processing as tools in diagnosing becomes terribly effective and one among the pivotal maladies in medications any place the order undertaking assumes an extremely basic job is that the conclusion of bosom malignant growth. Along these lines, AI methods will encourage specialists to make a right recognizable proof for bosom disease and make the best possible grouping of being generous or dangerous tumor. There are a bit inquiry that investigation of data taken from the diseased and determinations of specialists also experts are the premier essential factors inside the recognizable proof, anyway proficient frameworks and fake insight methods like AI for characterization errands, conjointly encourage specialists and authorities in a lot. We intend to explore distinctive AI strategies and we will utilize a few calculations and apply on bosom malignant growth dataset. We will concentrate on AI calculations: Naive Bayes, K-closest neighbor, strategic relapse, fortification calculation, bolster vector machine calculation. We will essentially think about these different calculations and break down their outcome.

1.2 Objectives

The target is assessing among the SVM, K-Nearest Neighbors, Naive Bayes and Logistic Regression with and without PCA classifiers to find out which method is ideal to anticipate Breast Cancer. After that our aim will be to find out the best features for our particular models. If we can find out the best output then with various different dataset we can predict cancer cells more accurately. In broader perspective, we trust the models utilized here are helpful enough for restorative professionals to settle on right choices. Certain presentation measurements for example, Accuracy, Recall, Precision, Specificity and the F1 Score have used to enable us to take a gander at and pick the best figuring. In the wake of getting the outcomes we would have the option to pick the best includes from our dataset.

1.3 Thesis Orientation

This book is made out of an aggregate of five sections. Section 1 is the present area and presents the subject of the proposition. Fundamental influences of breast cancer on woman is described in this section. Section 2 depicts the past works

here. It shows different figuring survey which uses insightful models for bosom malignant prediction. This is like manner portrays the most recent works. The confinements of these models are moreover depicted in this segment. Section 3 talks about different algorithms and various dataset. When we get a result we used PCA This part delineates the results we get from different features from our picked dataset. Likewise it shows the system use. It talks about different methods for better result analysis. Chapter 4 also tells our fundamental settings and results. A short record of the presentation estimations utilized in our report and the outcomes got of different execution estimations of each check are addressed and considered here, both for with and without applying Principal Component Analysis. Chapter 5 tells our examination and also incorporates the restrictions of our evaluation. A short record of things to come works or steps we plan to take to improve our models or research is in like way imparted here.

Chapter 2

Literature Review

Beforehand, inquire about with respect to characterization and expectation of bosom malignant growth has been completed utilizing a few information mining procedures. Characterization and collection of data are 2 wide utilized routes in data mining. Collection or grouping ways mean to extricate data from informational collection to get groups or bunches and portray the data set. Characterization otherwise called supervised learning in AI, plans to group obscure things bolstered taking in existing examples and future plans. Preparation set which is used in fabricating the arranging figure, and in this manner the investigate set, that will in general evaluate the classifier, are commonly referenced in grouping assignments. Besides, basic advancement has been done with regards to breast malignancy survivability forecast utilizing marked, unlabeled, and pseudo-named quiet information. Prognostic investigations of bosom malignancy survivability have been supported by AI calculations, which can anticipate the survival of a specific patient dependent on verifiable patient information.

Neural systems and related methods have an immense commitment with regards to foreseeing bosom malignancy. In the course of recent decades, ANNs have used progressively through an ever growing number of experts, and become a functioning examination territory. ANNs have managed various victories with incredible advancement in Breast Cancer arrangement and determination in the beginning periods. ANN model is used for ranking: input layers, hidden layers and output layers. Broad research had been finished with back proliferation counterfeit neural system (BP-ANN) technique and its varieties in bosom malignancy determination. The framework, nevertheless, has a couple of limitations, for instance, no accreditation to overall optima, a huge amount of tuning para-meters, and long getting ready time. Single Hidden Layer Neural Networks (SFLN)[8] was proposed by Huang and Babri to handle referenced issues with TLP which is called extraordinary learning machine (ELM). Better ELM model used for bosom malignant growth early expectation. Results demonstrated that it for the most part gave better exactness, explicitness, and affectability contrasted with BP ANN. Nonetheless, most existing works center around expectation execution with constrained consideration with therapeutic expert as end client and pertinence viewpoint in genuine restorative setting

With due regard to all related work alluded over, this paper thinks about the presentation of the calculations; Decision Tree, Linear Discriminant Analysis (LDA), K-Neighbors, Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM)

utilizing Wisconsin Breast Cancer (unique) dataset. The objective is to accomplish the most effective calculation to enable us to foresee bosom malignant growth at the underlying stages. To do as such, we analyze proficiency and adequacy of those methodologies as far as specific criteria, for example, accuracy, precision,specificity„ confusion and normalized matrix, recall and f1-score.

2.1 Fundamental of Machine Mearning

Machine learning is a section inside man made reasoning which has a place with the science and building of making smart machines. Robotized information procurement centered by AI through the plan and usage of calculations where exact information is acquired by calculations. Essentially this system is trained by AI relying upon the utilization of likelihood. There are distinctive sort of ways have a place with AI. Supervised learning: In supervised learning separate themselves through the related level. The target of this learning adequately recognizes new data given to it through oversaw learning and using the past educational accumulation and learning figurings can get acquainted with the technique to perceive the data. The estimations working underneath composed learning takes the wellsprings of information that the yield is beginning at now known for the reason altogether that the checks will cause the machine to discover by holding it to separate the specific yield and the extremely recognized regard test for to any degree further messes up. The machine then appears thusly. The recognized administered learning tallies wire demand, point boosting, want and descend into sin. At that point the model is adjusted by it subsequently. With such calculations, a machine make an utilization of directed figuring out how to attempt to do the expectation of mark esteems on unlabeled data by misuse suitable examples. Directed learning[4] finds the machine in such regions any place the more drawn out term occasions are normal through the recorded data. Unsupervised learning: Unsupervised learning studies anyway that frameworks will figure out how to speak to explicit info designs in a way that mirrors the connected math structure of the combination of information designs. By stand out from regulated learning or fortification learning, there aren't any express objective yields or natural assessments identified with each info; rather the unattended student gets contact past inclinations on what parts of the structure of the information should be caught inside the yield. A particular yield isn't having by unaided learning. Finding the structures and examples in the information is pointed by the learning specialist. Semi-Supervised learning: The machine is framed fit for adapting each marked and untagged data for the training reason. This especially includes preparing the machine through a little low amount of marked information to the detriment of a curiously large amount of untagged information. This can be for the levelheaded that untagged data are conservative and direct to gather. This kind of AI is utilized in many cases with the calculations like arrangement, forecast and relapse. Further, this kind of learning is utilized inside the field any place the cost of a related naming is spending too much to make on account of a completely marked instructing strategy. The praised utilization of semi-directed learning is face recognition.
Reinforcement learning: AI calculations experience the experimentation approach to manage structure positive of the exercises that offer the least unpredictable results and it sets up applications inside the field of play, course, and man-made thinking. Is commonly used for man-made cognizance, gaming, and course. There are 3 segments

that business essentially underneath this AI sort - teacher, student, nature with that the administrator do the collaboration and moreover the exercises that the expert is proposed to endeavor to do. The entire objective of support learning is to shape the administrator pick exercises which will urge to get extended reward over the perfect proportion of time. Along these lines the course of action is obvious that the help empowers the machine to pick up capability with the most clear technique to figure with to allow best results.
Collaborative learning: Recommendations make through a strategy which is shared isolating which is a basic sort of prescribed structure. It is an independent getting the hang of Clustering: Structure in collections of data where no specific structure as of late existed is found by bundling computation is a performance learning. Through the breaking down different properties of the data the bundles, regularly occur in data is found by gathering count. Batching is routinely used for dividing colossal proportion of data into more diminutive assembling and tuning examination for each social occasion, which has a spot with exploratory assessment. Requests: Classification has a spot with regulated acknowledging which requires getting ready with data that has known imprints. Through the readiness of past records structure will make sense of how to perceive the danger.

2.2 Software and Hardware

The dataset are prepared and organized in Python 3.6, This model use lots of packages like numpy and pandas. Beside this we make use of sci-kit learn. The whole model is developed with keras and tensorflow which is a power full library and used for large scale heterogeneous systems. Anaconda python 3.6 is used to run the algorithms that are used in this thesis work. The outputs shown in result analysis is taken after the algorithms being compiled in Anaconda python 3.6.

2.3 Related Researcher Works

There are various present day systems have been developed with the advancement of innovation for the expectation of bosom malignant growth. The business related to this field is laid out in a matter of seconds as pursues. A portion of the examinations related to forecast and analysis of infections utilizing AI procedures like choice tree for identification of disease. KNN calculation is outstanding for effortlessness and flexibility in usage which makes it one of the most every now and again utilized characterization calculation in AI as indicated by Jin. Liu Lei describes a model which is the utilizations AI for malignancy recognition. Logistic Regression calculation of Sklearn library has used to characterize the informational collections for bosom disease. Two highlights of most extreme surface and least edge was chosen and the order exactness remained at 95.5 percent.
Zemouri, Omri, Devalland, Arnould, Morello, Zerhouni and Fnaiech gave a model that uses a Breast Cancer Computer Aided Diagnosis (BC-CAD)[1] based on various variables and a Constructive Deep Neural Network. Wisconsin Breast Cancer Dataset (WBCD) and genuine information from the north hospital of Belfort (France) was utilized to find out the repeat score of the Oncotype DX. This model connected a technique to bring down the quantity of contributions for preparing a

profound learning neural system.
Bellaachia and Guven (2006) looked into the use of Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms on SEER dataset[2] which contained 16 attributes and 482,052 records. The dataset is seen as immaculate as a result of colossal proportion of patient and a moderate number of characteristics. From their examination, C4.5 estimation beat the the other results. They displayed another technique for chest sickness assurance by planning a significant learning-based independent component extraction count, stacked auto-encoders with ravenous layer-wise pre-getting ready computation to isolate huge features and information, with an assistance vector machine model to perceive tests with new incorporates into liberal and destructive tumors. The proposed system for significant learning-based independent component extraction was taken a stab at the Wisconsin Diagnostic Breast Cancer enlightening gathering and it inside and out improved the show of request and gave a promising method to manage chest dangerous development finding.
Sivakami proposed breast cancer Hybrid Model which integrates DT and SVM algorithms. This model was utilized to arrange patients into two classes (Benign/Malignant). The dataset containing eleven characteristics. It contains six hundred ninety nine cases where two hundred fourty one cases have a place with the harmful class and 458 cases have a place with the generous class. Sixteen instances of the dataset have missing values. The result was compared to IBL, SMO, and NAÏVE classifications techniques using Weka software. The results show that DT-SVM[5] perform well in classifying the breast cancer data, better than any other classifier algorithms. The accuracy of the Classification model was DT – SVM 91 percent. The low error rate was 2.58 percent, correctly classified instance was 459 and incorrectly classified instance were 240
There are explores in ongoing past and on-going looks into which expects from the highlights that are most useful in anticipating threatening or considerate malignancy and to overview general patterns which may be useful for choosing specific models and another parameter determinations. The sum total of what explores have been to arrive at the most noteworthy precision conceivable in the briefest time.

Chapter 3

Proposed Model

Our target is to predict the tumor is Benign or Malignant. We came up with a model which compares with other models. We will find out the numerical results we get from the algorithms. we run the test with six algorithms: Decision Tree, K-

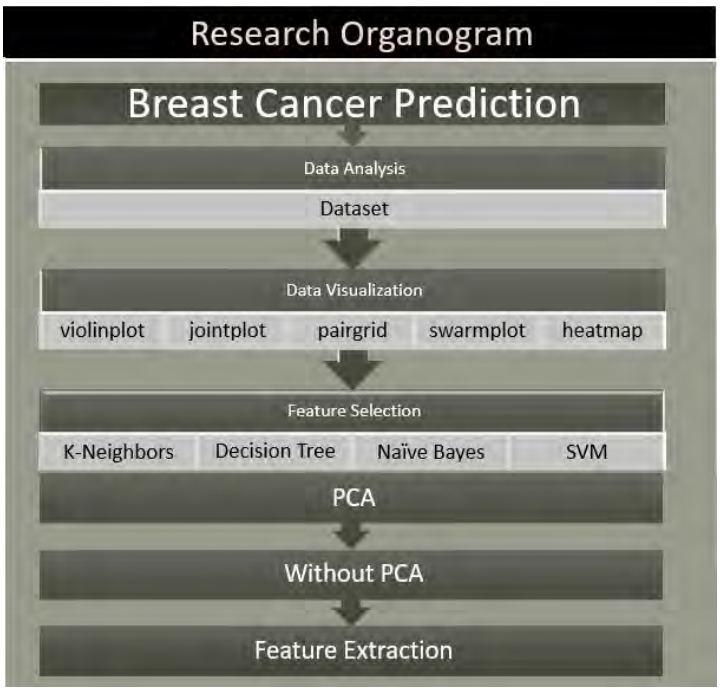


Figure 3.1: Research Organogram

Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes and Support Vector machine (SVM). Feature selection in the form of PCA have been used to decrease dimensionality of the dataset. The models are prepared by methods for preparing and testing after PCA is connected and then without PCA.

3.1 Dataset

The dataset is taken from the Wisconsin Breast Cancer (Diagnostic) Data Set(WBCD). It is free accessible in UCI-Repository. The dataset contains 569 examples and 32 traits of outwardly estimated nuclear highlights registered from a picture of a fine needle suction (FNA) of a bosom mass. FNA is a used which finds out the liquid or

tissues and gathers an example to make a determination of foreseeing illness. For example, disease. Among the 569 examples, the class conveyance are 212 harmful and other 357 non-harmful tumors. Twenty features are used for predicting better results.

3.2 Data Visualization

3.2.1 Histogram

A histogram is the data or information using bars of different statures utilizing is a graphical depiction. The size of reliable model data can be exhibited using a histogram. Figure 3.2 exhibits the class appointment of broke down unsafe (M) and liberal (B) tumors. There are 200 twelve risky tumors which are approximately thirty-eight percent and other 300 fifty-seven liberal tumors making up the rest of the sixty-two percent of the judicious class.
The features can be plotted against examination as saw in figure 3.3 from which we can see that mean estimation of cell length, edge, zone, diminutiveness, concavity, and depressed centers can be used in the portrayal of the dangerous development. Greater estimations of these parameters will all in all exhibit an association with undermining tumors. The mean estimations of surface, smoothness, equalization or fractal estimation don't show a particular tendency of one assurance over the other.

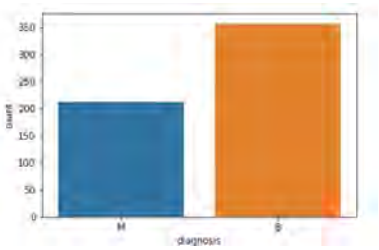


Figure 3.2: Histogram

3.2.2 Heatmap

A heatmap is the assistance of hues for perception of both basic and complex data of a two-dimensional portrayal. Heatmap is an exceptionally supportive way to deal with see which intersection purposes of the characteristics have higher gathering of the data appeared differently in relation to the others. Figure 3.4 addresses a relationship system using a heatmap. It is used to show the relationship among all of the 30 features in this dataset.

3.3 Data Processing

3.3.1 Categorical Variable Conversion

In our dataset we can get both numerical and categorical features. Among the Diagnosis segment had all out component which says if the disease is M = dangerous or B = favorable. The numbers they give are numerical. A large portion produces

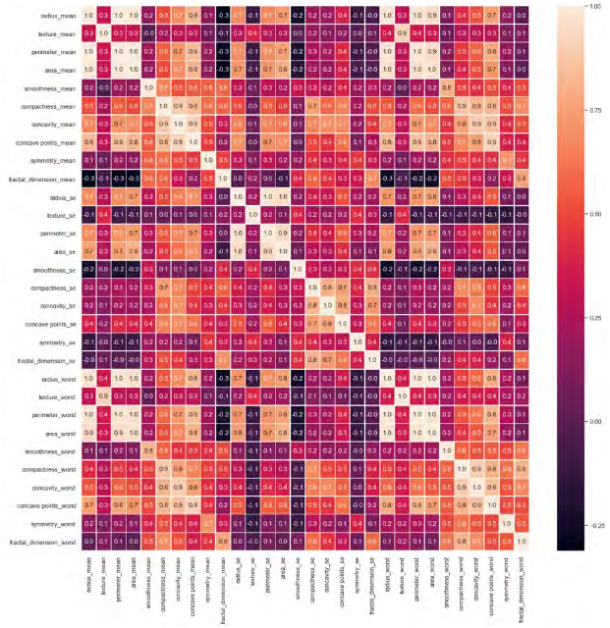


Figure 3.3: Heatmap

best outcome with numeric values. An absolute variable has such a large number of levels. This pulls down the execution level of the model. variable "postal division" would have various levels. A clear cut variable has levels which seldom happen. A large number of these levels have an insignificant possibility of having a genuine effect on model fit. For instance, a variable 'sickness' may have a few levels which would once in a while happen. There is one level that consistently happens for example for the vast majority of the perceptions in the informational collection there is just one level. Factors with such levels neglect to have a beneficial outcome on model execution because of exceptionally low variety.

3.3.2 Feature Scaling

The estimations of the characteristics we get from the dataset differs generally. It is otherwise called information standardization. This is done in light of the fact that a few calculations won't work appropriately without it and information ought to be institutionalized before applying PCA as variables with higher and lower change will be managed in a sudden manner. In this paper, StandardScaler is used to realize the organization. The standard format of a model x is resolved as:

$$z = x - u/s \tag{3.1}$$

3.3.3 Principal Component Analysis (PCA)

Head Component Analysis is a technique for measurement decrease which diminishes an enormous arrangement of factors . PCA is fundamentally a numerical strategy which changes various corresponded factors into fewer uncorrelated direct factors called head segments utilizing utilizes a symmetrical change [1]. In the wake of institutionalizing the information, PCA was associated for Random Forest, SVM, KNN, Logistic Regression, and Gaussian Naïve Bayes classifiers anyway not for the neural frameworks as both Artificial Neural Network and Convolutional Neural

Network in light of the way that the neural framework can evaluate some other non-straight planning through picking up and is free from the necessities of a non-direct model. In the wake of applying PCA, the dataset is decreased to eight head sections from past thirty properties which address each recognition.

3.3.4 Data Reshaping

Further information handling of is required as information reshaping for the contribution of CNN. The dataset was at first was (569, 30) in two dimensional structure. Utilizing NumPy, whole information goes into (569, 10, 3) in three dimensional structure.

3.3.5 Train-Test Split

Preparing information and testing information (and in some cases to three: train, approve and test) and fit our model on the train information to make forecasts on the test information. Getting ready dataset is a bit of the genuine dataset that we use to set up. This model sees and gains from the data. Test data, on the other hand, is the case of data used to give a fair examination of a keep going model fit on the planning dataset. The Test dataset gives the best use to survey the model. It is used when the model is completely arranged.
Part the dataset into preparing, approval testing sets can be resolved on two classifications. Right off the bat, it relies upon how much the all out number of tests in the information and second, on the real model the client is preparing. A few models need productive or enormous information to prepare upon, so all things considered one could advance for the bigger preparing sets. Models with not very many hyper parameters are assessed to be anything but difficult to approve and tune, so one can decrease the size of your approval set. Be that as it may, given the model has numerous hyper parameters, the client would need to have an enormous approval set also.
In this theory, we have part our dataset into 70 percent-30 percent proportion for preparing and test separately. 70 percent of the dataset to preparing. Out of the 70 percent dataset for preparing, we are keeping 63 percent for preparing. playing out the examination on one subset (the preparation set), and approving the investigation on the other subset (called the approval set or testing set). To decrease inconstancy, in numerous methodologies different types of cross-endorsement are used for using different fragments, and the endorsement results are combined (for instance landed at the midpoint of) over the rounds to give a check of the model's judicious introduction.

3.3.6 Support Vector Machine

Support Vector Machine (SVM) is an administered AI procedure that is extensively utilized in example acknowledgment and arrangement issues, particularly at the point when the dataset has definitely two classes. Reinforce vectors are data demonstrates that are closer the hyperplane and effect the position and course of the hyperplane. Using these assistance vectors, we help the edge of the classifier. Eradicating the assistance vectors will change the circumstance of the hyperplane.

These are the centers that help us collect our SVM. In determined backslide, we take the yield of the immediate limit and squash the motivating force inside the extent of [0,1] using the sigmoid limit. If the squashed worth is more unmistakable than a farthest point value(0.5) we designate it an imprint 1, else we dole out it a name 0. In SVM, we take the yield of the straight work and if that yield is more noticeable than 1, we recognize it with one class and if the yield is - 1, we perceive is with another class. Since the breaking point regards are changed to 1 and - 1 in SVM, we get this help extent of values([-1,1]) which goes about as edge.

3.4 System Implementation

In the field of Machine learning high dimentional data analysis could be a challange for researchers. With the help of computation timr, improving learning accuracy and facilitate a higher understanding for the learning model. we have a tendency to discuss many frequently used analysis measures for feature choice, and so survey supervised, unsupervised and semi-supervised feature selection strategies.Variable choice or trait choice is known as highlight choice. Programmed choice of characteristics in the information that is most significant to the prescient displaying issue. Dimensionality decrease is totally unique in relation to including choice. Every procedure solicitation to downsize the number of qualities inside the dataset, anyway a dimensionality decrease system do consequently by making a new mix of traits, any place as highlight choice systems grasp and avoid qualities present inside the information while not regularly evolving them. An exact prescient model is made to highlight choice strategies. Aiding in picking highlights will give the best or better precision while requiring less information. Distinguishing and evacuating unneeded should be possible by utilizing the element determination technique. Feature selection algorithms are 3 classes.
Filter method: Factual measure to allot assessment to each component connected by the channel highlight choice strategies. The highlights are hierarchic by the point and either chose to be whole or off. The techniques are normally uni variate and consider the component severally, or with reference to the variable amount.
Wrapper method: Wrapper ways is the determination of a gathering of choices as a hunt disadvantage. Any place totally various highlights are prepared, assessed and contrasted with various blends. A prescient model us acclimated valuate a blend of mixes and appoint a score upheld model precision.
Embedded method: Embedded strategies discover that alternatives best add to the exactness of the model while the model is being made. The principal basic sort of implanted element decision strategy is regularization techniques. Extra imperatives into the advancement of a prescient calculation are presented by Regularization techniques are additionally called punishment strategies. That inclination the model to-ward lower multifaceted nature.

3.5 Algorithms

The model works with binary information of different data. The algorithms chosen for this model are Support Vector Machine (SVM)[17], Random Forest (RF)[19], K-Nearest Neighbors (KNN), Logistic Regression (LR), Gaussian Naïve Bayes, ANN

and CNN. The results of the algorithms were compared to determine the best classifier for the problem

3.5.1 Linear Discriminant Analysis

The use of linear discriminant analysis algorithm is mainly for classifications predictive[7] modeling problems. For both preparation and application LDA is one of the simplest model. Calculating for each class depends on the statistical properties of the data consisted by LDA can be said straight forward representation of LDA. For one input variable (x) this is regularly the mean and furthermore the difference of the variable for each class. For different factors, this is frequently the indistinguishable properties determined over the variable Gaussian[22], explicitly the implies that and furthermore the co variance framework. From the information, the factual properties are determined and plug into the LDA condition to make forecasts. Some improving suppositions are made by direct discriminant examination about the information, for example, the learning from Gaussian that each factor is shaped kind of a chime bend once planned. That each quality has an indistinguishable change that estimations of each factor differ around the mean by the indistinguishable amount on the normal. Linear Discriminant Analysis makes forecasts by assessing the likelihood that a substitution set of data sources has a place with each class.

3.5.2 Logistic Regression

Logistic Regression is well known AI calculation after straight relapse[9]. From numerous perspectives, straight backslide and vital backslide are similar. In any case, the best differentiation lies in what they are used for. Direct backslide estimations are used to predict/figure regards anyway key backslide is used for gathering endeavors. There are various gathering endeavors done routinely by people. For example, gathering whether an email is a spam or not, describing whether a tumor is destructive or charitable, orchestrating whether a site is phony or not, etc. These are typical models[24] where AI figuring can make our lives a lot less difficult. An amazingly essential, fundamental and accommodating estimation for a request is the key backslide figuring. By and by, we should explore vital backslide. The general province of Logistic backslide is
log(p(X)/(1-p(X))=$\beta 0 + \beta 1X$
Where p(X) is the reliant variable, X is the free factor, 0 is the block and
1 is the slant co-effective.
Input values (x) are joined directly utilizing loads or coefficient esteems to foresee yield esteem (y). A key distinction from direct relapse is that the yield worth being demonstrated is a paired quality (0 or 1) instead of a numeric value. Where y is the anticipated yield, b0 is the inclination or capture term and b1 is the coefficient for the single info esteem (x). Every segment in your information has a related b coefficient (a steady genuine worth) that must be gained from your preparation information. The realistic portrayal[12] of the model that you would store in memory or in a document are the coefficients in the condition (the beta worth or b's).

3.5.3 Naive Bayes

The Naive Bayes Classifier framework relies upon the alleged Bayesian speculation and is particularly fit when the dimensionality of the wellsprings of information is high. Despite its straightforwardness, Naive Bayes can consistently beat progressively complex request strategies. To demonstrate the possibility of Naïve Bayes Classification[12], consider the model that appeared in the depiction above. As illustrated, the articles can be named either GREEN or RED. Our task is to arrange new cases as they arrive, i.e., decide to which class mark they have a spot, in perspective on the correct presently leaving objects. Naive Bayes classifiers can manage an emotional number of free factors whether steady or obvious. Given a ton of components, X = x1,x2,x...,xd, we have to build up the back probability for the event Cj among a ton of potential outcomes C = c1,c2,c...,cd. In a continuously conspicuous language, X is the markers and C is the plan of obvious levels present in the destitute variable. The formula for Naive Bayes hypothesis is:

$$P(\mathbf{C}|\mathbf{A}) = P(\mathbf{C})\frac{P(\mathbf{A}|\mathbf{C})}{P(\mathbf{A})} \tag{3.2}$$

Here, P (C—A) is the back probability, the probability that a hypothesis (C) is certifiable given some confirmation (A). P (C) is the prior probability, the probability of the hypothesis being legitimate. P (An) is the probability of the verification, autonomous of the theory. P (A—C) is the probability of the verification when the hypothesis is legitimate

Guileless Bayes algorithmic program is utilized for twofold and multi request assembling and may even be set up on an insignificant unmindful set that could be a titanic bit of room. In addition, it moved the issue ascending out of the scourge of spatial property somewhat. In any case, as referenced as of now, it makes the bogus supposition that the information factors are self-sufficient of each astounding[17]. This can be not the situation, in reality, enlightening lists, any place there is a couple instigated relationship between the portion variable.

3.5.4 Decision Tree

Decision tree is the the arrangement of directed learning. They can be used for both backslide and gathering issues. It uses the tree depiction to deal with the issue in which each leaf center identifies with a class imprint and properties are addressed within center of the tree. We can address any boolean limit[6] on discrete attributes. We consider the whole getting ready set as the source. Feature regards are gotten a kick out of the chance to be hard and fast. In case the characteristics are steady, by then they are discredited going before structure the model. In the records we can see they are circled recursively. We use real systems for mentioning qualities as source or within center point. As ought to be clear from that Decision Tree manages the Sum of Product structure which is generally called Disjunctive Normal Form. In the above picture, we are envisioning the usage of PC in the consistently life of the people. In Decision Tree the genuine test is to recognizing evidence of the trademark for the root center point in each level. This methodology is known as property assurance. We have two noticeable property assurance[7] measures:1. Information Gain 2.Gini Index

3.5.5 K-Neighbors

The KNN estimation expects that near things exist in closeness. Practically identical things are near each other. Find out the detachment between the request model and the present model from the data. Incorporate the detachment and the record of the manual for a masterminded aggregation. Sort the orchestrated assembling of detachments and records from most diminutive to greatest by the partitions. Pick the essential K segments from the masterminded collection. Get the signs of the picked K sections. On the occasion that backslides, return the mean of the K names. In the occasion that request, return the technique for the K names. At the point when KNN is utilized for arrangement, the yield is determined on the grounds that the class with the absolute best recurrence from the K-most comparative occasions[11]. Generally votes in favor of their group and in this manner the class with the principal votes is taken for the expectation. Class balance is determined in light of the fact that the standardized recurrence of tests that have a place with each class inside the arrangement of K most comparable examples for another information occasion. For example, during a parallel order issue (class is zero or 1):
(class=0) = count(class=0) / (count(class=0)+count(class=1))
On the occasion that using K and having an extensive number of classes it is a savvy thought to pick a K regard with an odd number to keep up a key good ways from a tie. Additionally, the opposite, use an altogether number for K when having an odd number of classes.

Chapter 4

Result Analysis

The stages that executing AI models is to find out how successful our model is. Running different algorithms on different dataset or different values gives us clear numerical values of our selected algorithms. Our dataset includes 30 percent for prediction. So as to decide and look at the exhibitions of the various calculations, a few measurements have been used for this.

4.1 Performance Metrics

Different performance metrics techniques have been used to find out the performance. The paper earnestly manages grouping issues, execution measurements identifying with arrangements are talked about here. For Breast Cancer expectation, in the event that the objective variable is 1(malignant), at that point it is a positive example, which means the patient has Breast disease. Also, in the event that the objective variable is 0 (generous)[13], at that point it is a negative case, expressing that the patient doesn't have the malignant growth.

4.1.1 Confusion Metrics

Confusion matrix summarize the classification problem. Correct and incorrect values are summarized through count values and broken classes. It is used when classification model is confused to make prediction. The types of error made by a classifier is in sighted by this.The format design picture exhibition of a calculation[15]. Each line of the measurement speaks to the occasions in a genuin class while every segment speaks to the occurrence in an anticipated class or the other way around.

4.1.2 Accuracy

Precision is a summation of every single diverse kind of forcasts made. Precision is a proper result of the objective information which are adjusted. .

$$Accuracy = TP + TN/(TP + FP + TN + FN) \tag{4.1}$$

4.1.3 Precision

Precision is the proportion of diseased who has ben examined to find out cancerous cells. .

$$Precision = TP/(TP + FP) \tag{4.2}$$

4.1.4 Recall or Sensitivity

Recall is checking the patient wheather the cells are cancerous or not. Destructive tumors are TN and Constraining FN would require Recall to be almost as normal considering the present situation.

$$Recall = TP/TP + FN \tag{4.3}$$

4.1.5 F1 Score

Normal extractness and review is known as F1 score.FN and FP are taken by the score. F1 is ordinarily extra accomodating that extractness. It is determined as pursues:

$$F1Score = 2 * PrecisionRecall/Precision + Recall \tag{4.4}$$

4.2 Model Performance

We used in total of six algorithms - Logistic Regression (LR), Support Vector Machines (SVM), Naive Bayes (NB), Decision Tree (DT), Linear Discriminant Analysis (LDA) and K Neighbors Classifier. The outputs have executed directly. Similar calculations is attached after Principal Component Analysis (PCA). Calculations are taken by using Accuracy, Precision, Recall, F1 Score and Specificity.

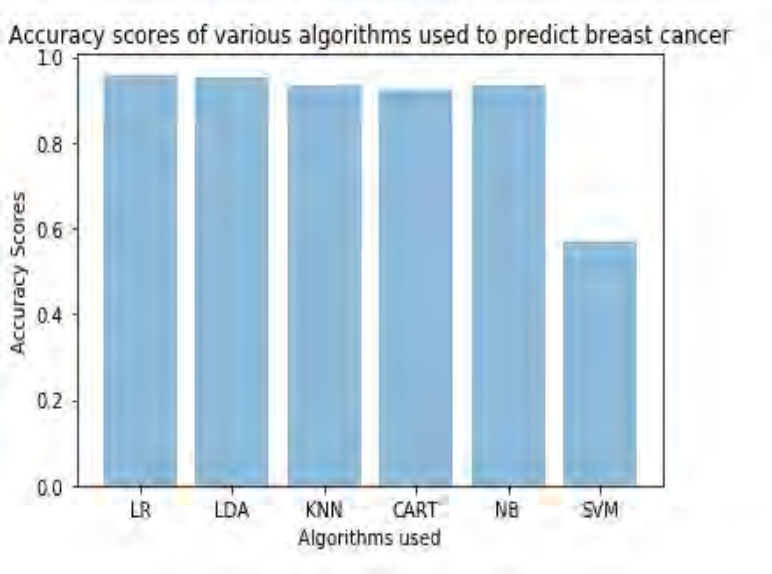


Figure 4.1: Accuracy scores without PCA.

In the following figure, The Normalized and the Confusion Matrix for every algorithms is represented through figures. As shown previously, the Confusion Matrix

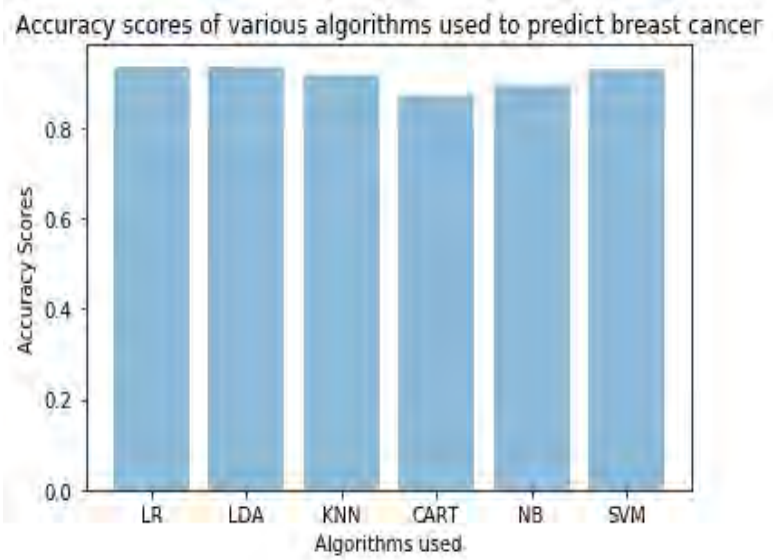


Figure 4.2: Accuracy scores with PCA.

	Accuracy	Precision	Specificity	Recall	F1 Score
Decision Tree	0.84	0.70	0.60	0.99	0.89
K-Neighbor	0.93	0.93	0.80	0.99	0.96
LDA	0.97	0.98	0.95	0.98	0.98
LR	0.92	0.92	0.78	0.97	0.94
Naive Bayes	0.96	0.964	0.91	0.99	0.98
SVM	0.91	0.9	0.75	0.989	0.95

Table 4.1: Scores of Accuracy,Precision,Recall,F1Score and Specificity without PCA

has four values-TN, False Positive, FN and TP. The blocks represent correctly predicted negative in True Negative, falsely predicted positive in False Positive, wrong prediction of negative in False Negative and correctly predicted positive in True Positive respectively, in all the figures of Confusion Matrix.

4.2.1 Linear Discriminant Analysis before and after applying PCA

The figures below illustrate The Normalized and the Confusion Matrix of Linear Discriminant Analysis (LDA)[16] without (Figure 4.3) and with applying PCA (Figure

	Accuracy	Precision	Specificity	Recall	F1 Score
Decision Tree	0.86	0.85	0.65	0.97	0.91
K-Neighbor	0.89	0.88	0.70	0.97	0.92
LDA	0.91	0.90	0.76	0.99	0.95
LR	0.88	0.84	0.65	1.0	0.91
Naive Bayes	0.91	0.90	0.75	0.99	0.94
SVM	0.90	0.87	0.70	1.0	0.93

Table 4.2: Scores of Accuracy,Precision,Recall,F1Score and Specificity with PCA

4.4). Without applying PCA. LDA has a really good accuracy of predicting breast cancer, with a score reaching 0.97. Results in other performance metrics; precision (0.985), Recall (0.977) and F1-Score (0.981) also suggest that LDA can be a reliable algorithm in predicting breast cancer. Mixed results are obtained as accuracy, precision and F1 Score records figures lower than for LDA[7] without PCA. However, application of PCA does increase the recall score (0.983 from 0.977). With Re-call being more vital in predict-ing diseases than precision, LDA is preferred to be applied with PCA.

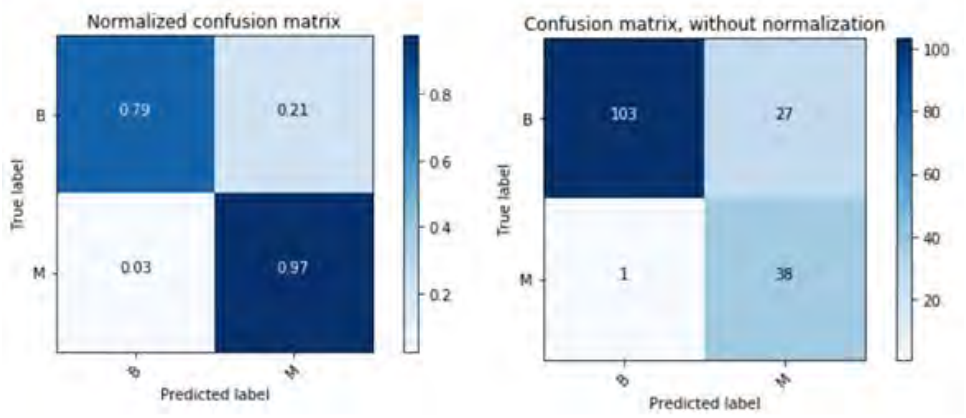


Figure 4.3: Normalized and Confusion Matrix of LDA without PCA

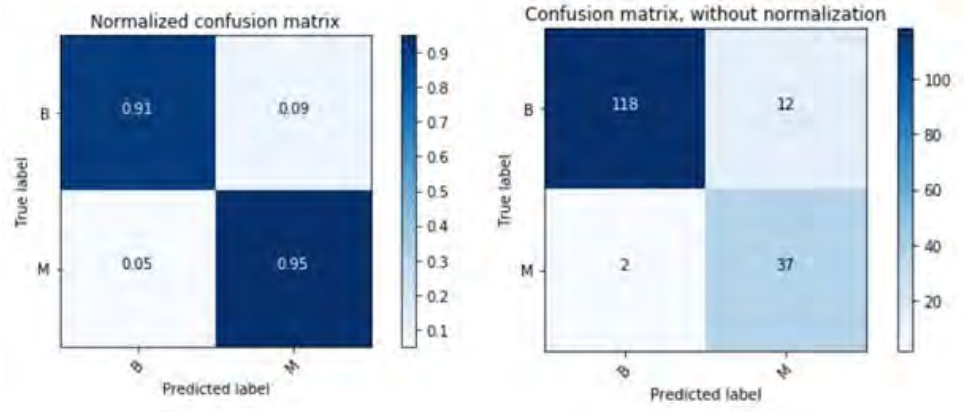


Figure 4.4: Normalized and Confusion Matrix of LDA with PCA

4.2.2 Decision Tree before and after applying PCA

The figures illustrate Normalized and the Confusion Matrix of the Decision Tree classifier both without (Figure 4.5) and with applying PCA (Figure 4.6). Numerical output data show that Decision Tree has performed moderately well for this problem with an accuracy score of 0.834 and a recall score of 0.792 and 0.99 without PCA. Introduction of PCA has a better impact on the accuracy of the decision tree as there is an increase for all three the performance metrics. However there is a decrease in the Re-call once PCA is applied. Since re-call is more important that precision in disease prediction, we can conclude that Decision tree performs better without PCA.

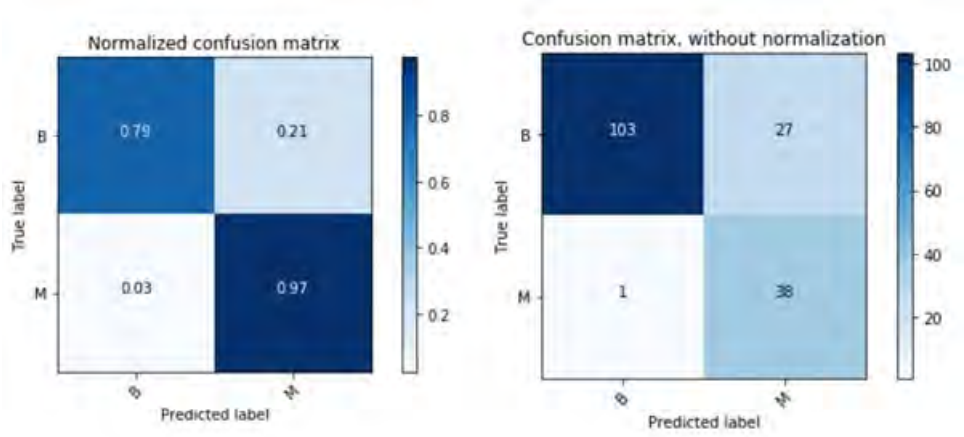


Figure 4.5: Decision Tree without PCA

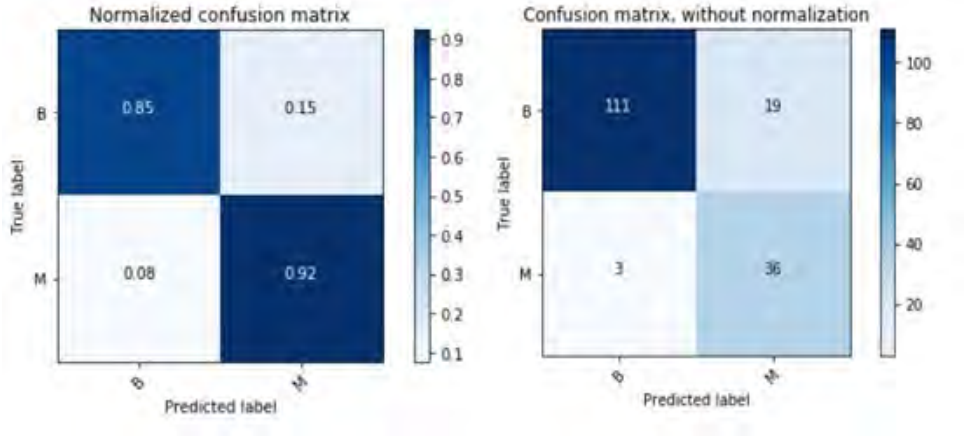


Figure 4.6: Decision Tree with PCA

4.2.3 Logistic Regression before and after applying PCA

The figures below illustrate the Normalized and the Confusion Matrix of Logistic Regression[23] both for without (Figure 4.7) and with applying PCA (Figure 4.8). The results show that Logistic Regression without PCA records a good accuracy of 0.92 along with figures of 0.923, 0.976 and 0.949 for precision. PCA on the dataset shows mixed results as the accuracy of the algorithm decreases. However, Recall scores a perfect 1.00 after PCA is applied and hence logistic regression can be applied with PCA for breast cancer prediction.

4.2.4 Naïve Bayes before and after applying PCA

The figures below illustrates the Normalized and the Confusion Matrix of Naïve Bayes[20] for both without(Figure 4.9) and with applying PCA (Figure 4.10). Naïve Bayes records a good score of 0.964 in accuracy while also having scores of 0.969 in precision, 0.984 in recall and 0.977 in F1Score. Introduction of PCA results in a decrease in the numerical count of accuracy, while the value of recall decreases by 0.001 after Naïve Bayes is implemented with. Hence it will be ideal to use this algorithm without applying PCA.

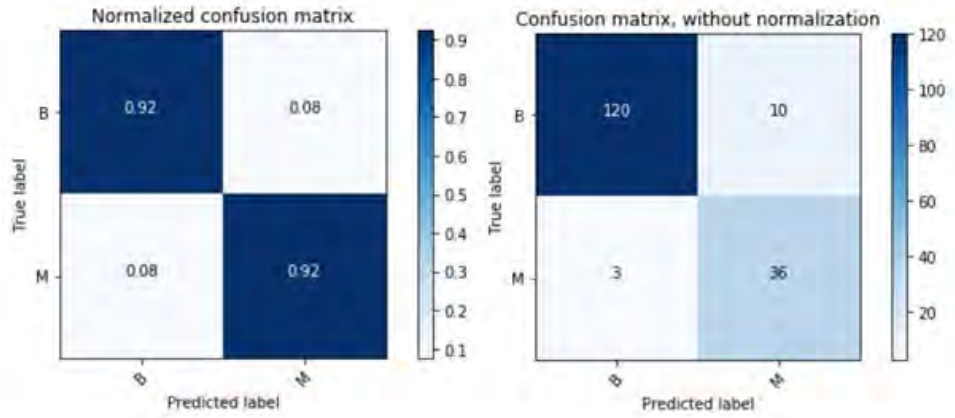


Figure 4.7: Logistic Regression without PCA

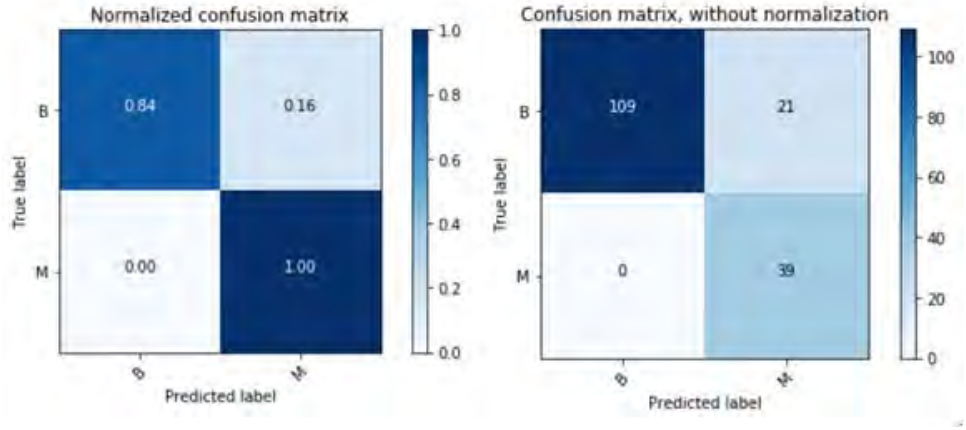
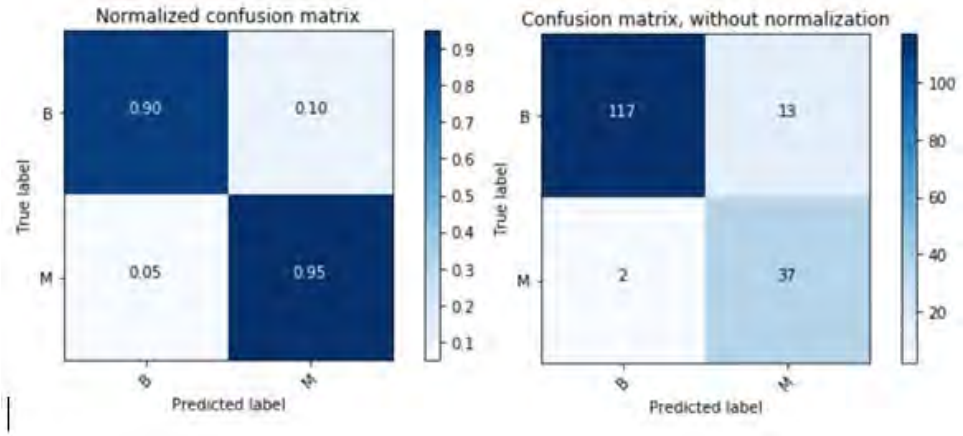


Figure 4.8: Logistic Regression with PCA



4.2.5 Support Vector Analysis (SVM) before and after applying PCA

The figures below illustrate the Normalized and the Confusion Matrix[14] of Support Vector Ma-chine (SVM) for both without (Figure 4.11) and with applying PCA (Figure 4.12). Numerical count obtained for SVM were quite satisfy-ing as SVM projected an accuracy of 0.917. Scores of 0.9 for precision, 0.991 for recall and 0.944

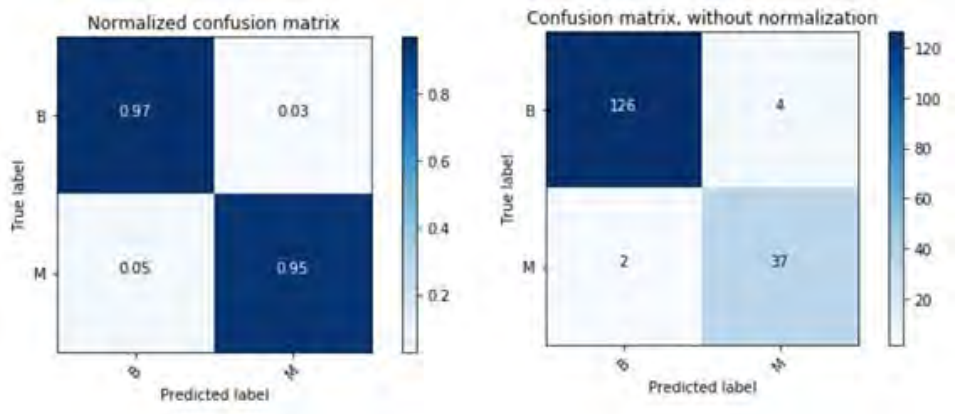


Figure 4.9: Naive Bayes without PCA

for F1 Score is also recorded for this problem. Introduction of PCA has seen a decline in case of accuracy (0.9 17 to 0.899), precision (0.9 to 0.869) and F1 Score (0.944 to 0.93). Recall, however scores an extract 1.000 after PCA is applied and hence the low score of precision can be over-looked while SVM is applied with PCA, since recall is more important in predicting disease.

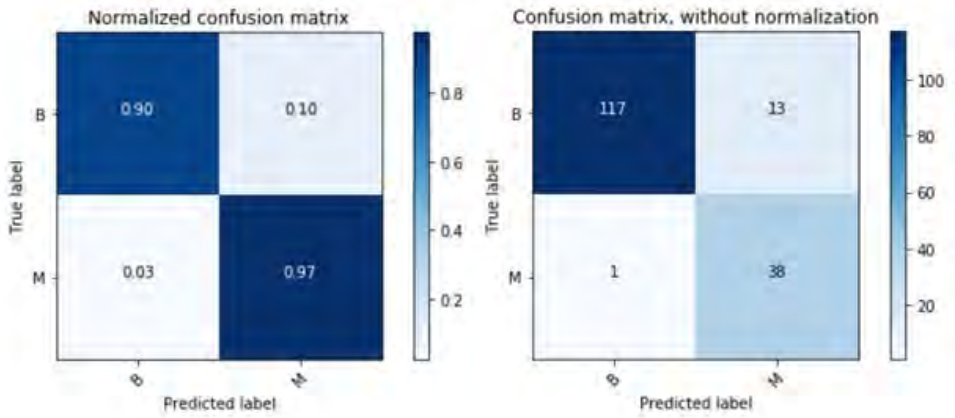


Figure 4.10: SVM without PCA

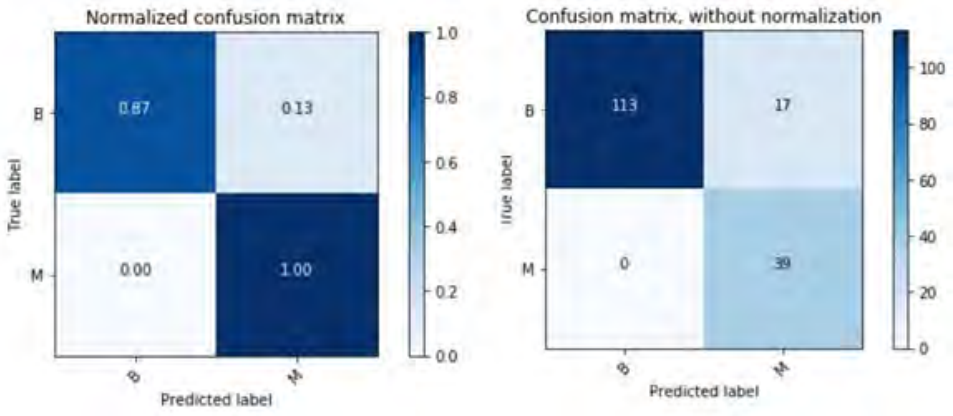


Figure 4.11: SVM with PCA

Chapter 5

Conclusion

5.1 Summary

As far as accuracy, Linear Discriminant Analysis (LDA) and Naïve Bayes, have scored high figures of 0.97 and 0.96 individually, without applying PCA. K-Neighbors (0.93) and Logistic regression (0.92) are not a long ways behind either. SVM scores 0.91 in exactness. Use of PCA decreases the exactness of the considerable number of calculations aside from Decision tree. In any case, the exactness figures are as yet higher than that of Decision tree's LDA, once more performs best after PCA is connected, despite the fact that there is a fall in air conditioning accuracy (0.917). Considering the other presentation lattice into record, a great deal can be resolved with respect to the exhibition of the calculations. Choice tree, K-Neighbors and Naïve Bayes per-shapes better without the presentation of PCA, while LDA,
Calculated Decision Tree and SVM gives better output when PCA is connected. An ideal 1.0 with regards to review, which is indispensable as far as ailment expectation, after PCA is connected, despite the fact that there are decreases in the estimations of all other presentation measurements of both the mentioned calculations for SVM and Logistic Regression. Recalling in PCA decreases the total time. We predict that the dataset we used, Logistic Regression and Support Vector Analysis with PCA performs better concerning Breast Cancer Prediction.
In light of various number of highlights, include determination procedure the outcomes have been finished up in Table 4.1 and 4.2. Various outcomes have demonstrated that equivalent precision can be accomplished even with lesser number of highlights for expectation of bosom disease in lesser computational time. Our work mainly focused in the advancement of predictive models to achieve good accuracy in predicting valid disease outcomes using supervised machine learning methods. The analysis of the results signify that the integration of multidimensional data along with different classification, feature selection and dimensionality reduction techniques can provide auspicious tools for inference in this domain. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables.

5.2 Limitation Of This Model

While we were effective at achieving results with exact correctnesses, there were sure boundary which develop while doing this theory. While we achieved exactness with

more than 90 percent without PCA. With bigger dataset it could not be possible. Getting a huge dataset will be helpful for this model for more specifications. Besides if we get huge dataset, all our algorithms will not predict better as it is proposed for short dataset. There are different complex models which we can work with huge dataset. Despite the fact we got better outcomes with our model. Training properly with our dataset can get us with better result.

5.3 Future Plan

In spite of accomplishing exact outcomes and correctnesses with the six calculations we have utilized, we wish to affirm the outcomes we acquired are not one-sided on account of the size of our dataset. Searching out an even bigger dataset and perform similar analysis and see if the results are the identical. Furthermore, since our dataset is kind of obsolete (collected within the 90s), more criteria for prediction and improved technology must have been available to attain more accurate numerical data. It would in like way put our assessment under an enhancing glass, in the event that we can see the correct parameters from our present and future datasets to convey ROC turns. In like manner, other than the models we have attempted we would conjointly wish to try, different tallies, for example, Adaboost so as to think about outcomes and proceed with our main goal for the best model for the figure. Applying other portion confirmation on the beginning at now utilized models is in like way under the thought, for example, the Recursive Feature Elimination.

References

[1] Asri, H., Mousannif, H., Al Moatassime, H., and Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, 83:1064–1069.

[2] Bellizzi, K. M. and Blank, T. O. (2006). Predicting posttraumatic growth in breast cancer survivors. *Health Psychology*, 25(1):47.

[3] Burke, H. B., Rosen, D. B., and Goodman, P. H. (1995). Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In *Advances in neural information processing systems*, pages 1063–1067.

[4] Gulli, A. and Pal, S. (2017). *Deep Learning with Keras*. Packt Publishing Ltd.

[5] Gupta, A. and Kaushik, B. N. (2018). Feature selection from biological database for breast cancer prediction and detection using machine learning classifier. *Journal of Artificial Intelligence*, 11:55–64.

[6] Huynh, H. T., Won, Y., and Kim, J.-j. (2008). An improvement of extreme learning machine for compact single-hidden-layer feedforward neural networks. *International journal of neural systems*, 18(05):433–441.

[7] Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA: a cancer journal for clinicians*, 61(2):69–90.

[8] Jiang, Y., Nishikawa, R. M., Schmidt, R. A., Metz, C. E., Giger, M. L., and Doi, K. (1999). Improving breast cancer diagnosis with computer-aided diagnosis. *Academic radiology*, 6(1):22–33.

[9] Kharya, S. (2012). Using data mining techniques for diagnosis and prognosis of cancer disease. *arXiv preprint arXiv:1205.1923*.

[10] Kline, T. S., Joshi, L. P., and Neal, H. S. (1979). Fine-needle aspiration of the breast: Diagnoses and pitfalls. a review of 3545 cases. *Cancer*, 44(4):1458–1464.

[11] Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM.

[12] Mandal, S. K. (2017). Performance analysis of data mining algorithms for breast cancer cell detection using naïve bayes, logistic regression and decision tree. *International Journal Of Engineering And Computer Science*, 6(2):20388–20391.

[13] McKinney, W. (2015). pandas: a python data analysis library. *see http://pandas. pydata. org*.

[14] Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.

[15] Rossbach, C. J., Yu, Y., Currey, J., Martin, J.-P., and Fetterly, D. (2013). Dandelion: a compiler and runtime for heterogeneous systems. In *Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles*, pages 49–68. ACM.

[16] Saika, K. and Sobue, T. (2013). Cancer statistics in the world. *Gan to kagaku ryoho. Cancer & chemotherapy*, 40(13):2475–2480.

[17] Salama, G. I., Abdelhalim, M., and Zeid, M. A.-e. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *Breast Cancer (WDBC)*, 32(569):2.

[18] Sewak, M., Vaidya, P., Chan, C., and Zhong-Hui Duan (2007). Svm approach to breast cancer classification. In *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, pages 32–37.

[19] Shah, C. and Jivani, A. G. (2013). Comparison of data mining classification algorithms for breast cancer prediction. In *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pages 1–4. IEEE.

[20] Thangavel, S. K., Bkaratki, P. D., and Sankar, A. (2017). Student placement analyzer: A recommendation system using machine learning. In *2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 1–5. IEEE.

[21] Watkins, A., Timmis, J., and Boggess, L. (2004). Artificial immune recognition system (airs): An immune-inspired supervised learning algorithm. *Genetic Programming and Evolvable Machines*, 5(3):291–317.

[22] Zemouri, R., Omri, N., Devalland, C., Arnould, L., Morello, B., Zerhouni, N., and Fnaiech, F. (2018). Breast cancer diagnosis based on joint variable selection and constructive deep neural network. In *2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME)*, pages 159–164. IEEE.

[23] Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

[24] Zurada, J. M. (1992). *Introduction to artificial neural systems*, volume 8. West publishing company St. Paul.