BACHELOR OF SCIENCE IN
COMPUTER SCIENCE AND ENGINEERING

Inspiring Excellence

Prediction

Machine Learning as an Indicator for Breast Cancer Prediction

Authors
Tahsin Mohammed Shadman
Fahim Shahriar Akash
Mayaz Ahmed

Supervisor
Dr.Md.Ashraful Alam
Assistant Professor
Department of CSE

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

Department of Computer Science and Engineering
BRAC University, Dhaka - 1212, Bangladesh

December 2018

Declaration

It is hereby declared that this thesis /project report or any part of it has not been submitted elsewhere for the award of any Degree or Diploma.

Authors:

Tahsin Mohamed Shadman
Student ID: 14101060

Fahim Shahriar Akash
Student ID: 14101146

Mayaz Ahmed
Student ID: 14101143

Supervisor:

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

December 2018

The thesis titled Machine Learning as an Indicator for Breast Cancer Prediction
Submitted by:
Name: Tahsin Mohammed Shadman Student ID: 14101060
Name: Fahim Shahriar Akash Student ID: 14101146
Name: Mayaz Ahmed Student ID: 14101143
of Academic Year 2018 has been found as satisfactory and accepted as partial fulfillment of the requirement for the Degree of Computer Science)

1.	Dr.Md.Ashraful Alam Assistant Professor Address	Chairman
3.	Md. Abdul Mottalib Professor and Chairperson Address	Member
4.	Name of Internal Member Designation Address	Member
5.	Name of Internal Member Designation Address	Member
6.	Name of External Member Designation Address	Member

Acknowledgements

Foremost, we would like to express our sincere gratitude to our advisor Assistant Prof. Dr.Md. Ashraful Alam for the continuous support of my research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped us in all the time of research and writing of this thesis. We could not have imagined having a better advisor and mentor for our undergraduate research and thesis.

A special thanks to our families. Words cannot express how grateful we are to our parents for all of the sacrifices that they have made on our behalf. Your prayers for us were what sustained us thus far. We would also like to thank all of our friends who motivated us to strive towards our goal. Lastly, we offer our regards and blessings to all of those who supported us in any respect during the partial completion of our thesis.

Tahsin Mohammed Shadman

Abstract

Affecting roughly around 10 percent of the women across the globe in some stage of their lives, Breast Cancer has stood out to be one of the most feared and frequently occurring cancers at present among women [1]. While the cure for this cancer is now available in almost all first world and some of the third world nations, the main dilemma takes place when the cancer cannot be correctly identified at the very initial stages. Machine Learning, in this field has proved to play a vital role in predicting diseases such as cancers alike. Classification and data mining methods so far have been reliant and an effective way to classify data. Especially in medical field, these methods have been used to predict and to make decisions. In this paper, we have successfully used six classification techniques in the form of Decision Tree, K-Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) on the Wisconsin Breast Cancer (original) datasets, both before and after applying Principal Component Analysis. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, recall, specificity and F1 Score. Experimental results have shown that Logistic Regression (recall score =1.000) and Support Vector Analysis (recall score =1.000) with PCA performs better when it comes to Breast Cancer Prediction for this dataset.

Keywords: Classification; Decision tree; Machine learning; Support vector machine; Principal Component Analysis, Recall, 10-Fold cross-validation

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Breast cancer (BC) is the most common cancer in women, affecting about 10 percent of all women at some stages of their life. In modern times, the rate keeps increasing and data show that the survival rate is 88 percent after five years from diagnosis and 80 percent after 10 years from diagnosis. Early prediction of breast cancer so far have made heaps of improvement, death rate of breast cancer by 39 percent, starting from 1989.. Due to varying nature of breast cancers symptoms, patients are often subjected to a barrage of tests, including but not limited to mammography, ultrasound and biopsy, to check their likelihoods of being diagnosed with breast cancer. Biopsy, is the most indicative among these procedures, which involves extraction of sample cells or tissues for examination. The sample of cells is obtained from a breast fine needle aspiration (FNA) procedure and then sent to a pathology laboratory to be examined under a microscope [27]. Numerical features, such as radius, texture, perimeter and area, can be measured from microscopic images. Data, later on, obtained from FNA are analyzed in combination with various imaging data to predict probability of the patient having malignant breast cancer tumor. An automated system here would be hugely beneficial in this scenario. It will likely expedite the process and enhance the accuracy of the doctor's predictions. In addition, if supported by abundance dataset and the automated system consistently performs well, it will potentially eliminate the needs for patients to go through various other tests, such as mammography, ultrasound, and MRI, which subject patients to significant amount of pain and radiation. In all, early prediction remains is one of the vital aspects in the follow-up process. Data mining methods or classification can help to reduce the number of false positive and false negative decisions. Consequently, new ways like data discovery in databases (KDD) has become a preferred tool for medical researcher. In this paper, using six classification models; Decision Tree, K-Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes and Support Vector Machine (SVM) have been run on the Wisconsin Breast Cancer (original)

datasets, both before and after applying Principal Component Analysis. The results obtained are than measured using various performance metrics to compare among the algorithms in order to find out the best suited model for cancer prediction.

1.1 Motivation

Many people are affected from breast cancer at the present time. Causing of this disease depends on man factors and cannot be simply determined. In addition, the identification method that determines whether or not the cancer is benign or malignant additionally needs an excellent deal of effort from a doctors and physicians. Once many tests are concerned within the identification of breast cancer, like clump thickness, uniformity of cell size, uniformity of cell form, etc., the ultimate result could also be troublesome to get, even for doctors. This has given an increase within the previous few years to the utilization of machine learning and computing generally as diagnostic tools. The diseases that take numerous lives, diagnostic computer-based applications are used wide. Robotics are taking part in an awfully necessary role in operational rooms. Also, the skilled systems are conferred within the intensive treatment rooms. In turn, using another side of Artificial intelligence for breast cancer designation isn't unworthy. It's reported that breast cancer illness is that the second commonest cancer that affects girls, and was the rife cancer within the world by the year of 2002[21]. This cancer may be a quite common sort of cancer among girls and therefore the second highest reason behind cancer death. Within the United State, regarding one in eight girls over their time period includes a risk of developing breast cancer. With the uncontrolled division of one cell inside the breast leads to beginning to the breast cancer which results in a visible mass, called a tumour. The tumour can be either benign or malignant. The correct designation in determinant whether or not the tumour is benign or malignant may result in saving lives. Therefore, the necessity for precise classification within the clinic may be an explanation for nice concern for specialists and doctors. This importance of Artificial intelligence has been actuated for the last twenty five years, once scientists began to understand the quality of taking bound selections to treat specific diseases. The employment of machine learning and data processing as tools in diagnosing becomes terribly effective and one amongst the crucial diseases in medicines wherever the classification task plays a really essential role is that the diagnosis of breast cancer. Therefore, machine learning techniques will facilitate doctors to create an correct identification for breast cancer and make the proper classification of being benign or malignant tumor. There is little question that analysis of information taken from the patient and selections of doctors and specialists are the foremost necessary factors within the identification, however knowledgeable systems and artificial

intelligence techniques like machine learning for classification tasks, conjointly facilitate doctors and specialists in a great deal.

We aim in this paper from to compare different classification learning algorithms significantly to predict a benign from malignant cancer in breast cancer dataset. We aim to investigate different machine learning techniques and we will use several algorithms and apply on breast cancer dataset. We will focus on machine learning algorithms: Naïve bayes, K-nearest neighbor, logistic regression, reinforcement algorithm, support vector machine algorithm. We will primarily study these various algorithms and analyze their result.

1.2 Objectives

The aim of this thesis is to compare and evaluate among the six classifiers to see which classifier is best suited to predict Breast Cancer at the very initial stage. In broader perspective, we hope the models used here are useful enough for medical practitioners to make right decisions. Certain performance metrics such as Accuracy, Recall, Precision, Specificity and the F1 Score have been used to assist us compare and choose the best algorithm.

1.3 Thesis Orientation

This dissertation book is composed of a total of seven chapters. Chapter 1 is the current chapter and introduces the topic of the thesis. The basic impacts of breast cancer on women is briefly highlighted here along with the steps of early prediction of the cancer in the form of machine learning. Chapter2 describes the previous contributions in this eld. It describes different algorithms regarding used as predictive models for breast cancer prediction. It also describes the most recent works in this eld. The limitations of this eld are also described in this chapter.

Chapter 3 states the proposed model of our research; it affirms the dataset we used in our research, the predictive models we selected and how we generated results for both before and after applying Principal Component Analysis. Chapter 4 describes the dataset we selected which Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, United States of America had created. Details of the dataset along with the 10 real valued features of each cell is given here. Chapter 5 presents the system implementation. It talks about sub sections such as Train Test split; the ratio in which the dataset used in our research was split into training and testing models, feature selection and gives a brief account of Principal Component Analysis. Chapter 6 rst describes our experimental settings and results. A brief account of the performance metrics used in our research and the results

obtained of various performance metrics of each algorithm are illustrated and compared here, both for with and without applying Principal Component Analysis. Chapter 7 summarizes our research and also highlights the limitations of our research. A brief account of the future works, or steps we intend to take to improve our models or research is also stated here.

1.4 Fundamentals of Machine Learning

Machine learning is a part within artificial intelligence which belongs to the science and engineering of making intelligent machines. Automated knowledge acquisition focused by machine learning through the design and implementation of algorithms where empirical data is required by algorithms. Basically the techniques for learning of a machine is taught by machine learning depending on the use of probability. There are different kind of ways belong to machine learning.

Supervised learning: In supervised learning starting with the datasets which contains training examples, which can identify themselves through the associated level those have. It does it by running data through a learning algorithm. The goal of supervised learning is, correctly identify the new data given to it through the supervised learning and using the previous data set and learning algorithms can learn the technique to identify the data. The algorithms operating below supervised learning takes the inputs that the output is already known for the reason in order that the algorithms will create the machine to find out by holding it compare the particular output with the already known output to test for to any extent further errors. The machine is then modeled consequently. The famous supervised learning algorithms include classification, gradient boosting, prediction and regression. Then the model is modified by it consequently. With such algorithms, a machine create a use of supervised learning to try and do the prediction of label values on unlabeled information by exploitation appropriate patterns. Supervised learning finds the appliance in such areas wherever the longer term events are expected through the historical information.

Unsupervised learning: Unsupervised learning studies however systems will learn to represent specific input patterns in a manner that reflects the applied math structure of the assortment of input patterns. By contrast with supervised learning or reinforcement learning, there aren't any express target outputs or environmental evaluations related to every input; rather the unattended learner brings in contact previous biases on what aspects of the structure of the input ought to be captured within the output. A specific output is not having by unsupervised learning. Finding the structures and patterns in the data is aimed by the learning agent.

Semi-Supervised learning: Under this machine learning sort, the machine is formed

capable of learning each labelled and untagged information for the coaching purpose. This particularly involves training the machine through a tiny low quantity of labelled data at the expense of an oversized quantity of untagged data. This can be for the rationale that untagged information are economical and straightforward to assemble. This sort of machine learning is employed oftentimes with the algorithms like classification, prediction and regression. Further, this sort of learning is employed within the field wherever the price of an associated labeling is splurging to create thanks to a totally labelled coaching method. The celebrated application of semi-supervised learning is face recognition through a digital camera.

Reinforcement learning: Under this machine learning sort, the machine learning algorithms run through the trial and error approach to form positive of the actions that offer the simplest results and it founds applications within the field of play, navigation, and artificial intelligence. Is usually used for artificial intelligence, gaming, and navigation. There are 3 elements that employment primarily below this machine learning sort - the agent, learner, the atmosphere with that the agent do the interaction and also the actions that the agent is meant to try and do. The entire objective of reinforcement learning is to form the agent choose actions which will facilitate to get maximized reward over the desired amount of time. Therefore the plan is evident that the reinforcement helps the machine learn the simplest policy to figure with to allow best results.

Collaborative learning: Recommendations generate through a technique which is known as collaborative filtering which is a primary type of recommender system. Among the large number of choices and based on comparison of preferences between users it helps the users to find item of relevance. Collaborative filtering is domain agnostic. It is an unsupervised learning

Clustering: Structure in collections of data where no specific structure previously existed is discovered by clustering algorithm, is a unsupervised learning. Through the examining different properties of the input data the clusters, naturally occur in data is discovered by clustering algorithm. Clustering is often used for dividing large amount of data into smaller group and tuning analysis for each group, which belongs to exploratory analysis.

Classifications: Classification belongs to supervised learning which requires training with data that has known labels. Application involving classification like by train using a set of spam and non-spam massages System will eventually learn to detect unwanted email. Through the training of previous records system will learn to identify the risk. Overall, the branches of machine learning can be identified from the following picture:

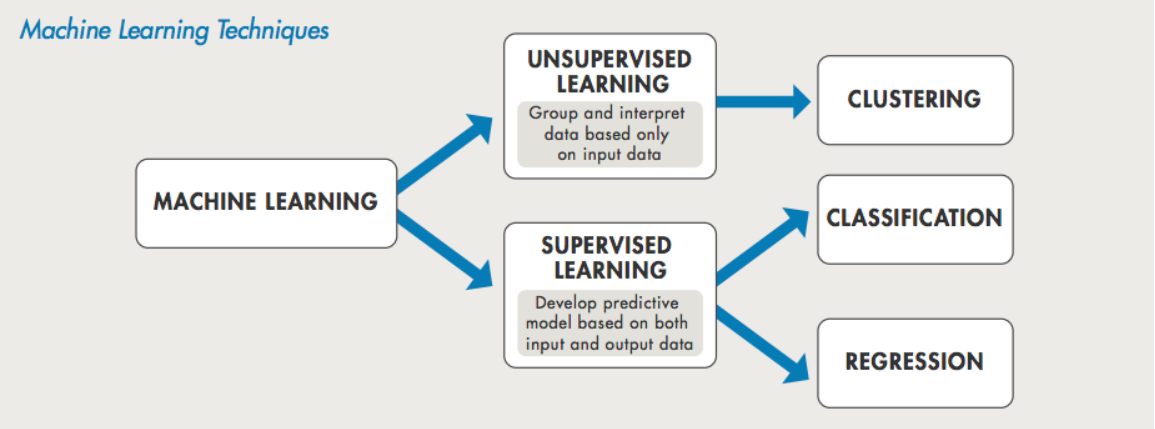


Fig. 1.1 Machine Learning Techniques

Chapter 2

Literature Review

Previously, research regarding classification and prediction of breast cancer has been carried out using several data mining techniques. Classification and agglomeration are 2 wide used ways in information mining [1]. Agglomeration or clustering ways aim to extract information from data set to get teams or clusters and describe the information set. Classification also known as super-vised learning in machine learning, aims to classify unknown things supported learning existing patterns and classes from the information set and after predict future things. The training set, that is employed to build the classifying structure, and therefore the take a look at set, that tends to assess the classifier, are ordinarily mentioned in classification tasks [2].

Furthermore, essential progress has been carried out when it comes to breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. Prognostic studies of breast cancer survivability have been aided by machine learning algorithms, which can predict the survival of a particular patient based on historical patient data.

Neural networks and related techniques have a vast contribution when it comes to predicting breast cancer. Over the past few decades, Artificial Neural Networks have been employed increasingly by more and more researchers, and become an active research area [7-12]. ANNs have afforded numerous successes with great progress in Breast Cancer classification and diagnosis in the very early stages [2,7-12]. A typical ANN model is made up of a hierarchy of layers: input, hidden and output layers. Extensive research had been done with backpropagation artificial neural network (BP-ANN) method and its variations in breast cancer diagnosis [13]–[14]. The tech-nique, however, has some limitations such as no guarantee to global optima, a lot of tuning para-meters, and long training time. Single Hidden Layer Neural Networks (SFLN) was proposed by Huang and Babri [15] to tackle the mentioned problems with tree steps learning process that called extreme learning machine (ELM). Standard [16] and best parameterized [17] ELM model were proposed for breast

cancer early prediction. Results showed that it generally gave better accuracy, specificity, and sensitivity compared to BP ANN. However, most existing works focus on prediction performance with limited attention with medical professional as end user and applicability aspect in real medical setting

With due respect to all related work referred above, this paper compares the performance of the algorithms; Decision Tree, K-Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) using Wisconsin Breast Cancer (original) datasets in both diagnosis and analysis to make decisions. The goal is to achieve the most efficient algorithm to help us predict breast cancer at the very initial stages. To do so, we compare efficiency and effectiveness of those approaches in terms of certain criteria such as accuracy, precision, specificity,, confusion and normalized matrix, recall and f1-score.

Chapter 3

Proposed Model

With our aim being to predict whether the tumor is Benign (non-cancerous) or Malignant (cancerous), we have outlined a simple model to come with the most accurate predictions. The first objective was to attain a dataset of numerical values of various instances. Upon finalizing our dataset, we split the train-test ration to 70:30 in order to train and test six algorithms: Decision Tree, K-Neighbors, Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes and Support Vector machine (SVM). Feature selection in the form of Principal Component Analysis is used to reduce dimensionality of the dataset. The models are trained again by means of training and testing after PCA is applied and finally compared the results with that of the previous results we reached without PCA.

The workflow below outlines a basic review of the entire thesis:

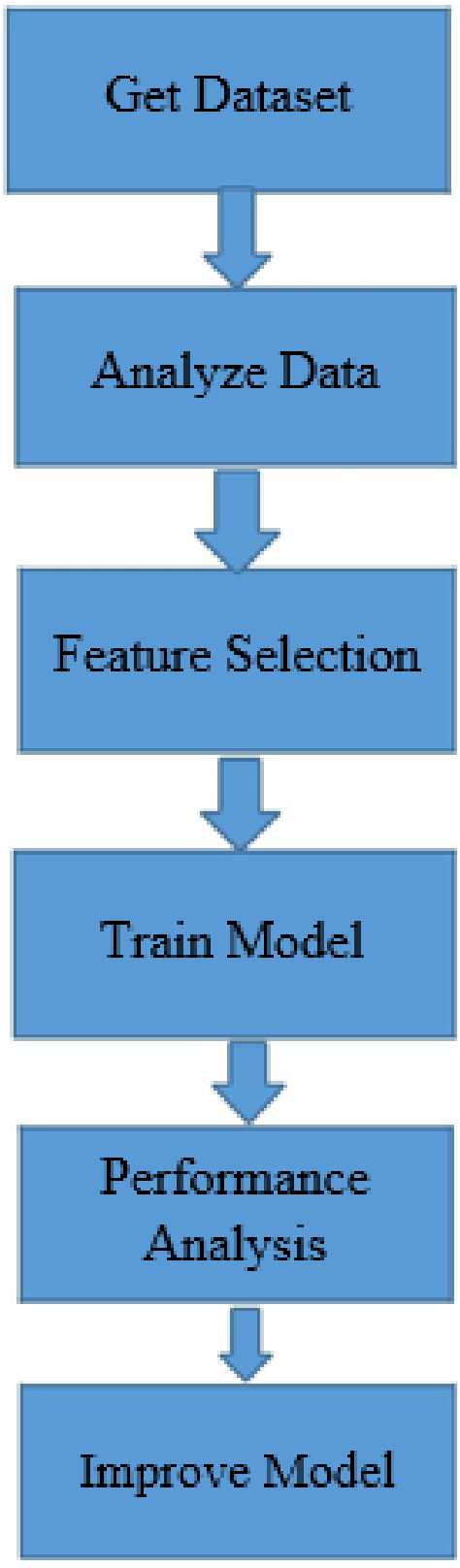


Fig. 3.1 Workflow

Chapter 4

Dataset

The dataset used for this paper is publicly available and was created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin, United States of America. To create the dataset, Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector. The dataset contains 357 cases of benign breast cancer and 212 cases of malignant breast cancer. The dataset contains 32 columns, with the first column being the ID number, the second column being the diagnosis result (benign or malignant), and followed by the mean, standard deviation and the mean of the worst measurements of ten features. There were no missing values in the dataset [18].

Attribute Information:

1. ID number 2) Diagnosis (M = malignant, B = benign) 3–32)

The ten real-valued features computed for each cell nucleus together with description are listed in the following table:

Radius	Average of distances from center to points the perimter
Texture	Standard deviation of gray-scale values
Perimeter	The total distance between the snake points constitutes the nuclear perimeter
Area	Number of pixel on the interior of the snake and adding one-half of the pixel in the Perimeter
Smoothness	Local variation in radius length, quantified by measuring the difference between the length
Compactness	Perimeter ^2 / area
Concavity	Severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	The length difference between lines perpendicular to the majority axis to the cell boundary in both directions
Fractional Dimension	Coastline approximation. A higher value corresponding to a less regular contour and thus to a higher probability of malignancy

Table 4.1 Ten real-valued features computed for each cell nucleus

Chapter 5

System Implementation

5.1 Feature Selection

Within the fields of machine learning high dimensional data analysis could be a challenge for re-searchers and engineers. Solving drawback by removing immaterial and redundant data through an efficient way provided by feature selection, which might cut back the computation time, improve learning accuracy, and facilitate a higher understanding for the learning model or data. During this study, 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, that are wide applied in machine learning issues, like classification and clustering. Variable selection or attribute selection is known as feature selection. Automatic selection of attributes in the data that are most relevant to the predictive modeling problem. Dimensionality reduction is completely different from feature selection. Each strategies request to scale back the quantity of attributes within the dataset, however a dimensionality reduction methodology do thus by making new combination of attributes, wherever as feature selection strategies embrace and exclude attributes present within the data while not ever-changing them. An accurate predictive model is created by feature selection methods. Helping in choosing features will provide best or better accuracy whilst requiring less data. Identifying and removing unneeded can be done by using the feature selection method. There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods.

Filter method:

Statistical measure to assign evaluation to each feature applied by the filter feature selection methods. The features are hierarchic by the score and either selected to be unbroken or off from the dataset. The methods are typically univariate and take into account the feature severally, or with reference to the variable quantity.

Wrapper method:

Wrapper ways think about the selection of a group of options as a search drawback, wherever completely different features are ready, evaluated and compared to different mixtures. A predictive model us accustomed valuate a mixture of combinations and assign a score supported model accuracy.

The search method is also organized like a best-first search, it should random like a random hill-climbing formula, or it should use heuristics, like forward and backward passes to feature and take away options.

Embedded method:

Embedded strategies learn that options best contribute to the accuracy of the model whereas the model is being created. the foremost common kind of embedded feature choice methods are regularization methods. Additional constraints into the optimization of a predictive algorithm is introduced by Regularization methods are also called penalization methods. That bias the model to-ward lower complexity.

5.2 Principal Component Analysis

The main plan of principal component analysis (PCA) is to cut back the dimensionality of a data set consisting of the many variables related with one another, either heavily or gently, whereas holding the variation present within the data set, up to the utmost extent[22] The identical is finished by remodeling the variables to a replacement set of variables, that are referred to as the principal elements (or merely, the PCs) and are orthogonal, ordered specified the retention of variation present within the original variables decreases as we tend to move down within the order. So, during this method, the first principal element retains most variation that was gift within the original elements. The principal elements are the Manfred Eigen vectors of a co variance matrix, and therefore they're orthogonal. Importantly, the dataset on that PCA technique is to be used should be scaled. The results are sensitive to the relative scaling. As a layman, it's a technique of summarizing information. Imagine some wine bottles on a board. every wine is delineate by its attributes like color, strength, age, etc. however redundancy can arise as a result of several of them can live connected properties. Thus what PCA can neutralize this case is summarize every wine within the stock with less characteristics. Intuitively, Principal part Analysis will provide the user with a lower-dimensional image, a projection or "shadow" of this object once viewed from its most in-formative viewpoint.

5.3 Train-Test Split

Data, in machine learning, in most scenarios are split into training data and testing data (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. Training dataset is a part of the actual dataset that we use to train the model. The model sees and learns from this data. Test data, on the other hand, is the sample of data used to provide an unbiased analysis of a final model fit on the training dataset. The Test dataset provides the ideal standard used to evaluate the model. It is used once the model is completely trained [25].

Splitting the dataset into training, validation testing sets can be determined on two categories. Firstly, it depends on how much the total number of samples in the data and second, on the actual model the user is training. Some models need efficient or large data to train upon, so in that case one could optimize for the larger training sets. Models with very few hyper parameters are estimated to be easy to validate and tune, so one can possibly reduce the size of your validation set. However, given the model has many hyper parameters, the user would want to have a large validation set as well.

In this thesis, we have split our dataset into 70%-30% ratio for training and test respectively (the first 400 instances for training while the next 169 instances for testing the model). Keeping in mind that training the model, making the machine learn, is vital, we have slotted 70% of the dataset to training. Out of the 70% dataset for training, we are keeping 63 percent for training and 7 percent for cross validation test. A round of cross-validation comprises separating a section of data into complementary subsets, performing the analysis on one subset (the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce variability, in most methods multiple rounds of cross-validation are performed using different partitions, and the validation results are combined (e.g. averaged) over the rounds to give an estimate of the model's predictive performance.

5.4 Algorithms

5.4.1 Logistic Regression

After linear regression, logistical regression is the most famous machine learning algorithm. Linear regression and logistic regression are similar in many ways. But what they are used for is the biggest distinction. Algorithms for linear regression are used to predict values, but logistic regression is used for classification tasks. Logistic rule may be a supervised rule that trains the model by taking input variables and a target variable. In logistical rule the output or target variable may be a categorical variable, in contrast to regression to-

ward the mean, and is therefore a binary classification rule that categorizes a knowledge purpose to one of the categories of information. The general equation of logistic regression is:

$$\text{logit}\ (p) = b0 + b1X1 + b2X2 + bk\ Xk$$

Where p is the probability of presence of the characteristic of interest.

Logistic regression measures the link between the variable quantity, the output, and therefore the freelance variables, the input. By estimating chances exploitation its underlying supply perform. It uses 1.2 penalty for regularization. Supply regression formula conjointly uses an equation with freelance predictors to predict a worth. The expected worth are often anyplace between negative eternity to positive eternity. The resultant chances are then born-again to binary values zero or one by the supply perform, conjointly referred to as the sigmoid function. The Sigmoid perform takes any real-valued variety and maps it into a worth between the vary 0-1 excluding the bounds themselves. Afterwards, a threshold classifier transforms the result to a binary worth. One in every of the first assumption of supply regression is that the input options ought to be freelance of every alternative. One variable ought to have very little or no co-linearity with the opposite variable. Hence, PCA is dead on the info beforehand, to convert the related variables to a collection of unrelated variables. For the creation of a model on breast cancer, Logistic regression was used. The system have developed consists within the estimation of unknown dependencies in a very system from a given knowledge set to make a helpful and general model to analyze new incoming knowledge.

5.4.2 Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithmic rule which might be used for each classification or regression challenges. However, it's principally utilized in classification issues. In this algorithmic rule, we plot each data item as a point in n-dimensional space where n is number of features one has with the value of each feature being the value of a particular coordinate [23]. Then, we perform classification by finding the hyper-plane that differentiates the two classes well shown in the figure below:

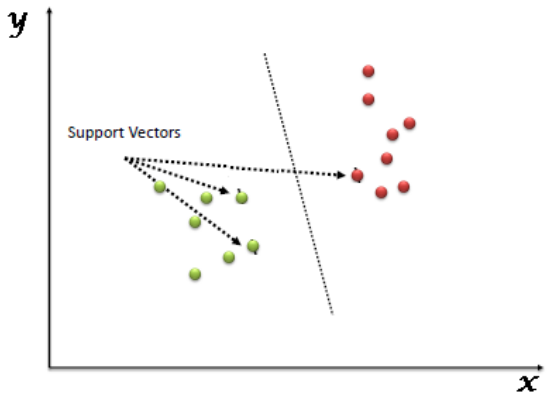


Fig. 5.1 Support Vector Machine

Often researchers tend to plot every knowledge item as some extent in n-dimensional area with the worth of every feature being the worth of a selected coordinate. Then, to perform classification by finding the hyper-plane that differentiate the 2 categories fine. It is a non-probabilistic binary linear classifier, how-ever are often manipulated during a manner that it will perform non-linear and probabilistic classification also, creating it versatile algorithmic program. AN SVM model could be an illustration of the instances as points in area mapped, so they will be categorized and divided by a transparent gap. New instances are then mapped into the identical area and foreseen that within which class it would be supported which aspect of the gap they fall in. the most advantages of SVM is that the indisputable fact that it's effective in high dimensional areas [19-20]. To boot, it is conjointly memory efficient since it uses a set of coaching points within the call operate.

Pseudo code for SVM:
initialize yi = YI for i I
REPEAT
compute SVM solution w, b for data set with imputed labels
compute outputs fi = (w, xi) + b for all xi in positive bags
set yi = sgn(fi) for every i I, YI = 1
FOR (every positive bag BI)
IF (PiI (1 + yi)/2 == 0)
compute i = arg maxiI fi
set yi = 1
END
END

WHILE (imputed labels have changed)
OUTPUT (w, b)

5.4.3 Naive Bayes

Naive Bayes classifiers are a group of classification algorithms supported Bayes' Theorem. it's not one algorithmic rule however a family of algorithms wherever all of them share a typical principle, each try of options being classified is freelance of every different. Bayes theorem uses the contingent probability that successively uses previous information to calculate the probability that a future event can happen. In Naive Bayes classifier, it's assumed that the input variables are freelance of every alternative which all options can separately contribute to the chance of target variable. So, the existence of one feature variable doesn't have an effect on the opposite feature variables. This can be why it's known as naive. However, in real knowledge sets, the feature variables are addicted to one another therefore this can be one among the drawbacks of Naive Bayes classifier. Naive Bayes classifier though, works fine for giant knowledge sets and generally per-form higher than the difficult classifiers. The formula for Naive Bayes theorem is:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Here, P (C|A) is the posterior probability, the probability that a hypothesis (C) is true given some evidence (A). P (C) is the prior probability, the probability of the hypothesis being true. P (A) is the probability of the evidence, irrespective of the hypothesis. P (A|C) is the probability of the evidence when hypothesis is true

Naive Bayes algorithmic program is employed for binary and multi category classification and might even be trained on small low information set that could be a huge advantage. It's addition-ally terribly quicker and climbable. Moreover, it migrated the matter arising from the curse of spatial property to some extent. However, as mentioned before, it makes the false assumption that the input variables are freelance of every different. This can be not the case in reality information sets, wherever there is several advanced relationships between the feature variable.

Steps to calculate Prediction:

Step1: Converting the data set into a frequency table

Step2: Creating likelihood table by finding the probabilities

Step3: Using Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

5.4.4 Decision Tree

Decision tree may be a supervised learning rule that's used for classification and regression. It works by splitting the info into 2 or additional subsets supported the values of input variables. A value operate or cacophonous criterion is employed to see the most effective split among all the split points. The info is split recursively into teams till the leaves contain just one sample. During this model, associate degree optimized version of the CART rule is employed to implement the choice tree classifier. Call trees are straightforward to interpret and perceive, compared to differ-ent classification algorithms. Moreover, call trees need very little preprocessing as outliers don't have an effect on the performance. Moreover, they're not supported the Euclidian distance. Hence, feature scaling isn't needed. Also, feature scaling may lead to wrong assumptions being tacit since the values would be modified. Call trees will handle each categorical and numerical variables as input therefore it's acceptable for this model, since the info set contains each variable varieties. During this model the link between the feature variable and target variable is complicated and high non-linear. Therefore a call tree contains a larger likelihood of outperforming lin-ear models like provision regression. While call Tree have many benefits, they even have some disadvantages. One is that, call Trees will cause over fitting by creating a tree that's too complicated and thus doesn't predict well on new information. Finally, since call Trees are greedy algo-rithms, the optimum tree isn't essentially came back.

Pseudocode for Decision Tree: 1. Checking for the base cases. 2.For each attribute a, finding the normalized information gain ratio from splitting on a. 3.Taking $a_{bestbetheattributewiththehighestnor}$

5.4.5 Linear Discriminant Analysis

The use of linear discriminant analysis algorithm is mainly for classifications predictive 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 often the mean and also the variance of the variable for every category. For multiple variables, this is often the identical properties calculated over the variable Gaussian, specifically the means that and also the co variance matrix. From the data the statistical properties are calculated and

plug into the LDA equation to make predictions. Some simplifying assumptions are made by linear discriminant analysis about the data such as the knowledge from Gaussian that every variable is formed sort of a bell curve once premeditated. That each attribute has the identical variance that values of every variable vary round the mean by the identical quantity on the average. LDA makes predictions by es-timating the probability that a replacement set of inputs belongs to every category. [21]That gets the very best probability is that the output class and a prediction is formed.

5.4.6 K-Neighbors

The k-nearest neighbor's algorithmic program is one of the simplest machine learning algorithms. It has merely supported the concept that objects that are 'near' every alternative can additionally have similar characteristics. So if it can recognize the characteristic options of one of the objects, it will be additionally predicted for its nearest neighbor." k-NN is associate improvisation over the nearest neighbor technique. It is based mostly on the plan that any new instance will be classified by the majority vote of its 'k' neighbors, - wherever k is a positive number, sometimes a little variety.

kNN is one amongst the foremost easy and simple data processing techniques. It is known as Memory-Based Classification as the coaching examples have to be in the memory at run-time. Once handling continuous attributes the distinction between the attributes is calculated Euclidean distance. a serious drawback once dealing with the Euclidean distance formula is that the big values frequency swamps the smaller ones.

When KNN is employed for classification, the output is calculated because the category with the very best frequency from the K-most similar instances. Every instance in essence votes for their class and therefore the class with the foremost votes is taken for the prediction.

Class probabilities is calculated because the normalized frequency of samples that belong to every class within the set of K most similar instances for a new data instance. For instance, during a binary classification problem (class is zero or 1):

$$(\text{class}=0) = \text{count}(\text{class}=0) / (\text{count}(\text{class}=0)+\text{count}(\text{class}=1))$$

If using K and having an even number of classes (e.g. 2) it is a good idea to choose a K value with an odd number to avoid a tie. And the inverse, use an even number for K when having

an odd number of classes.

Pseudocode of K-Neighbors [24]:

1. Load the training and test data
2. Choose the value of K
3. For each point in test data:

- find the Euclidean distance to all training data points
- store the Euclidean distances in a list and sort it
- choose the first k points
- assign a class to the test point based on the majority of classes present in the chosen points

4. End

Chapter 6

Result Analysis

The next step after applying implementing machine learning models is to seek out how effective is that the model, i.e. how the models performed on the datasets. This is carried out by running the models on the test dataset which was set earlier. The test dataset comprised of 30% of the dataset for Breast Cancer prediction.10-fold cross-validation was also done for Breast cancer pre-diction. In order to determine and compare the performances of the different algorithms, several metrics have been used.

6.1 Performance metrics

Several performance metrics have been used to figure out the performance of the Machine Learning algorithms in this our thesis. As the paper sincerely deals with classification problems, performance metrics relating to classifications are discussed here. For Breast Cancer prediction, if the target variable is 1(malignant), then it is a positive instance, meaning the patient has Breast cancer. And if the target variable is 0 (benign), then it is a negative instance, stating that the patient does not have the cancer.

6.1.1 Confusion Matrix

Summarization the performance of a classification algorithm is based on a technique which is known as confusion matrix. It is arguably the easiest way to regulate the performance of a classi-fication model by comparing how many positive instances are correctly/incorrectly classified and how many negative instances are correctly/incorrectly classified. In a confusion matrix, as shown here, the rows represent the actual labels while the columns represent the

predicted labels.

	Predicted Negative	**Predicted Positive**
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 6.1 Confusion Matrix

True Positives (TP):

These are the occurrences where both the predictive and actual class is true (1), i.e., when the patient has complications (breast cancer in this case) and is also classified by the model to have complications.

True Negatives (TN):

True negatives are the occurrences where both the predicted class and actual class is False (0), i.e., when a patient does not have complications and is also classified by the model as not having complications.

False Negative (FN):

These are occurrences where the predicted class is False (0) but actual class is True (1), i.e., case of a patient being classified by the model as not having complications even though in reality, they do.

False Positive (FP):

False positives are the occurrences where the predicted class is True (1) while the actual class is False (0), i.e., when a patient is classified by the model as having complications even

though in reality, they do not.

Normalized matrix

Normalized Confusion Matrix represents results in a more efficient way. The results are similar to that of the confusion matrix. The values are distributed within the range of 0-1. An even distribution of data makes prediction easier.

Accuracy

Evaluation of classification models is done by one of the metrics called accuracy. Accuracy is the fraction of prediction. It determines the number of correct predictions over the total number of predictions made by the model. The formula of accuracy is:

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

Recall:

It is a measure of the proportion of patients that were predicted to have the complications among those patients that actually have the complications. Recall can be calculated as follows:

$$Recall = \frac{TP}{TP + TN}$$

Precision

It is described as a measure of proportion of patients that actually have complications among those classified to have complications by the model. The formula for Precision is as follows:

$$Precision = \frac{TP}{TP + FP}$$

Specificity:

Classifier's performance to spot negative results is related by Specificity. It is exactly the nega-tive of Recall. It is a measure of the number of patients who are classified as not having compli-cations among those who actually did not have the complications. Specificity is

calculated as fol-lows:

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

F1 Score:

Weighted average of precision and recall is known as F1 score. Therefore false positives and false negatives are taken by this score into the consideration. Intuitively it is not as simple to grasp as accuracy, however F1 is typically additional helpful than accuracy. It is calculated as follows:

$$F1Score = \frac{Precision * Recall}{Precision + Recall} * 2$$

6.2 Model Performances:

A total of set of six classification algorithms are used - Logistic Regression (LR), Support Vector Machines (SVM), Naïve Bayes (NB), Decision Tree (DT), Linear Discriminant Analysis (LDA) and K Neighbors Classifier have been applied on the dataset. For Table, the algorithms have been implemented directly, while in Table, the same algorithms have been applied after Principal Component Analysis (PCA). For each experiment, the performance of the algorithms are measured using Accuracy, Precision, Recall, F1 Score and Specificity.

The table below demonstrates the results of different metrics for the algorithms to predict Breast Cancer without Principal Component Analysis:

	Accuracy	**Precision**	**Specificity**	**Recall**	**F1 Score**
Decision Tree	0.834	0.732	0.585	0.99	0.88
K-Neighbor	0.935	0.931	0.804	0.984	0.957
LDA	0.97	0.985	0.947	0.977	0.981
Logistic Regression	0.923	0.923	0.783	0.976	0.949
Naive Bayes	0.964	0.969	0.902	0.984	0.977
SVC	0.917	0.9	0.745	0.991	0.944

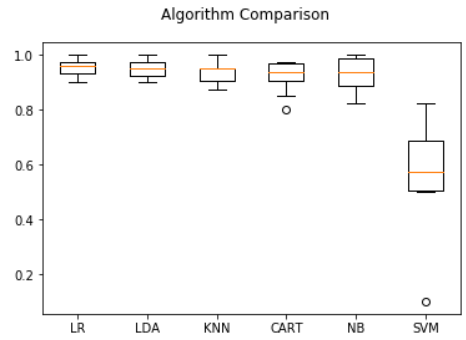
Table 6.2 Scores of Accuracy, Precision, Recall, F1 Score and Specificity without PCA

The table below demonstrates the results of different metrics for the algorithms to predict Breast Cancer with Principal Component Analysis:

	Accuracy	**Precision**	**Specificity**	**Recall**	**F1 Score**
Decision Tree	0.869	0.854	0.655	0,974	0.91
K-Neighbours	0.887	0.877	0.692	0.974	0.923
LDA	0.917	0.908	0.755	0.983	0.944
Logistic Regression	0.876	0.838	0.65	1	0.912
Naïve Bayes	0.911	0.9	0.74	0.983	0.94
SVM	0.899	0.869	0.696	1.000	0.93

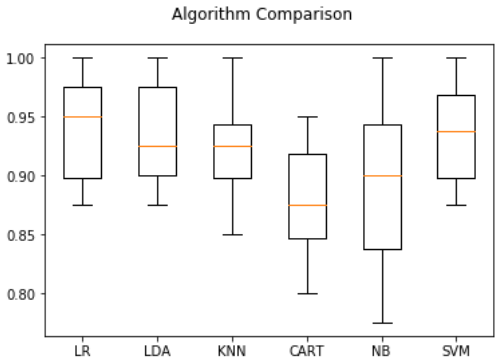
Table 6.3 Scores of Accuracy, Precision, Recall, F1 Score and Specificity with Principal Component Analysis

Further comparisons of all the six algorithms on the training data, both with and without PCA, is projected in the following bar charts:




```
[0.9574358974358974, 0.9525, 0.9347435897435897, 0.9223717948717949, 0.9323076923076924,
0.5667307692307693]
```

Fig. 6.1 Algorithm Comparison Without PCA




```
[0.9372435897435898, 0.9347435897435897, 0.9197435897435898, 0.8746153846153846,
0.8948076923076922, 0.9322435897435899]
```

Fig. 6.2 Algorithm Comparison With PCA

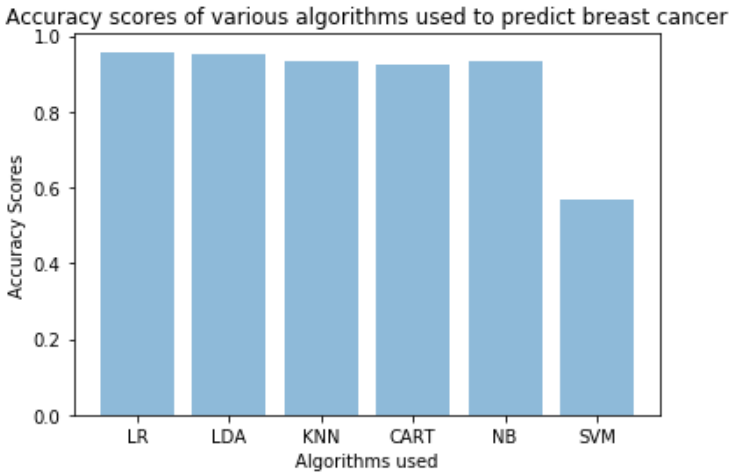


Fig. 6.3 - 5 Accuracy scores of the six algorithms on training data without Principal Component Analysis

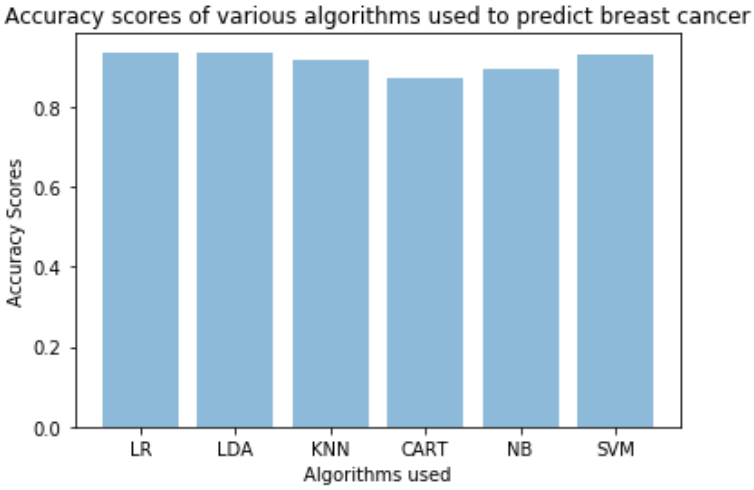


Fig. 6.4 - 5 Accuracy scores of the six algorithms on training data with Principal Component Analysis

In the following sub-sections, The Normalized and the Confusion Matrix for every algorithms is represented through figures. As shown previously in Table, the Confusion Matrix has four values-Ture Negative, False Positive, False Negative and True Positive. 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. These values are later

vital in obtaining the Accuracy, Precision, Recall, Specificity and F1 Score to evaluate the performance of each algorithm

Decision Tree before and after applying PCA

The figures below illustrate The Normalized and the Confusion Matrix of the Decision Tree classifier both without (Figure 6.5) and with applying PCA (Figure 6.6). The results show that Decision Tree has performed moderately well for this problem with an accuracy score of 0.834 and a precision and recall score of 0.792 and 0.99 respectively without PCA. Introduction of PCA has a positive impact on the accuracy, precision and F1 Score 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.

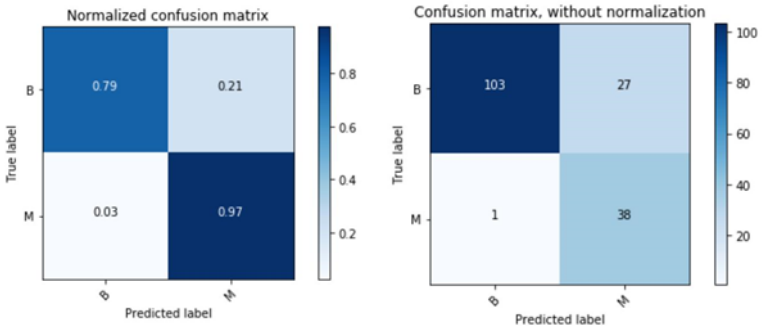


Fig. 6.5 - Normalized and Confusion Matrix of Decision Tree without PCA)

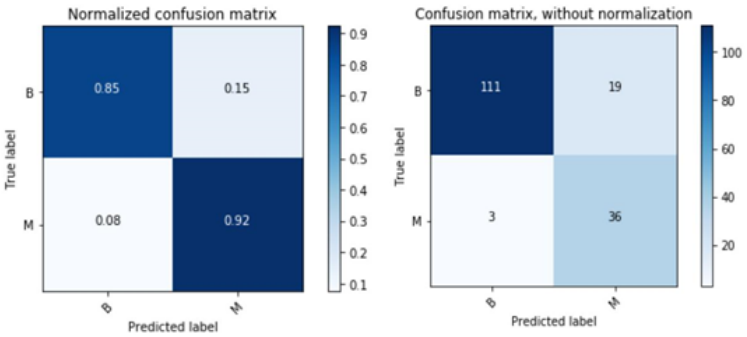


Fig. 6.6 -Normalized and Confusion Matrix of Decision Tree with PCA

K-Neighbor before and after applying PCA

The figures below illustrate The Normalized and the Confusion Matrix of the K-Neighbors both without (Fig 6.7) and with applying PCA (Figure 6.8) on this dataset. The results show how well K-Neighbor performs with an accuracy of 0.935. Precision, Recall and F1-Score also scores good figures of 0.931, 0.984 and 0.957 respectively without applying PCA. Introduction of PCA has been slightly disappointing as results in all four fronts of accuracy (0.935 to 0.887), precision (0.931 to 0.877), re-call(0.984 to 0.974) and F1 Score(0.957 to 0.923) decreases, once PCA is applied. K-neighbor has therefore is preferred to use on datasets without introducing PCA.

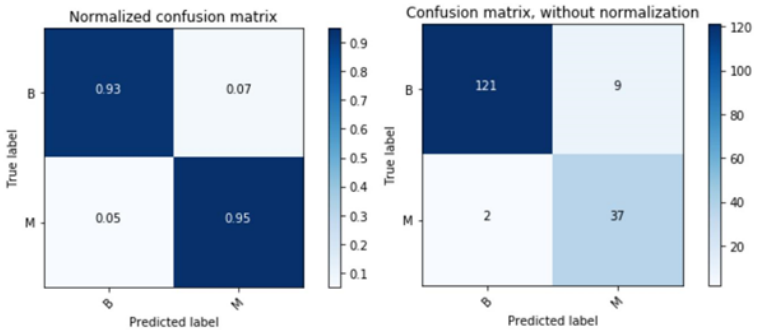


Fig. 6.7 -Normalized and Confusion Matrix of K-Neighbors without PCA

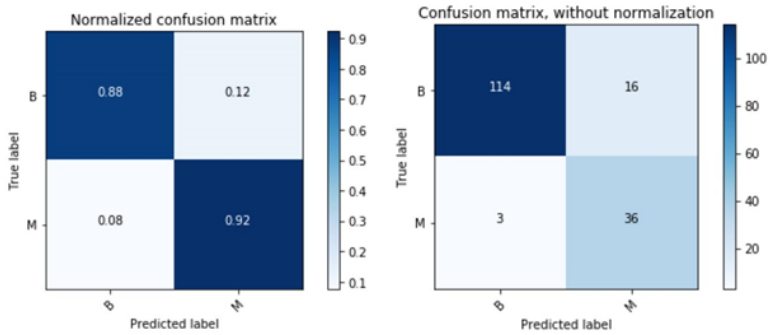


Fig. 6.8 -Normalized and Confusion Matrix of K-Neighbors with PCA

Linear Discriminant Analysis before and after applying PCA

The figures below illustrate The Normalized and the Confusion Matrix of Linear Discriminant Analysis (LDA) without (Figure 6.9) and with applying PCA (Figure 6.10). Without applying

PCA. LDA has a really good accuracy of predicting breast cancer, with a score reaching 0.970. 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 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.

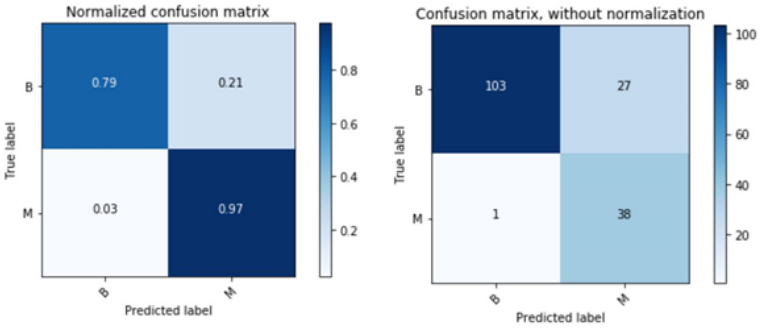


Fig. 6.9 -Normalized and Confusion Matrix of LDA without PCA

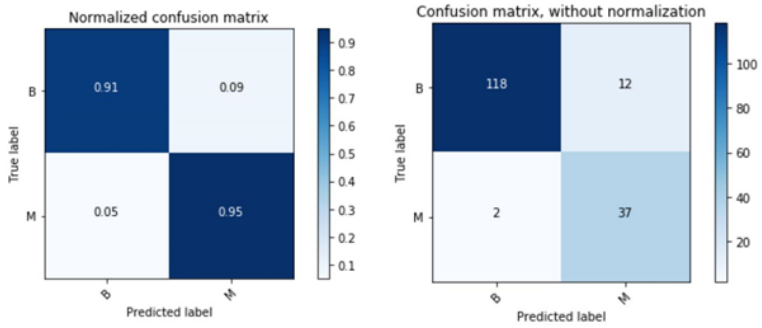


Fig. 6.10 -Normalized and Confusion Matrix of LDA with PCA

Logistic Regression before and after applying PCA

The figures below illustrate the Normalized and the Confusion Matrix of Logistic Regression both for without (Figure 6.11) and with applying PCA (Figure 6.12). The results show that Logistic Regression without PCA records a good accuracy of 0.923 along with figures of 0.923, 0.976 and 0.949 for precision, recall and F1 Score respectively. Introduction of PCA

on the dataset shows mixed results as the accuracy, precision and F1 Score 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.

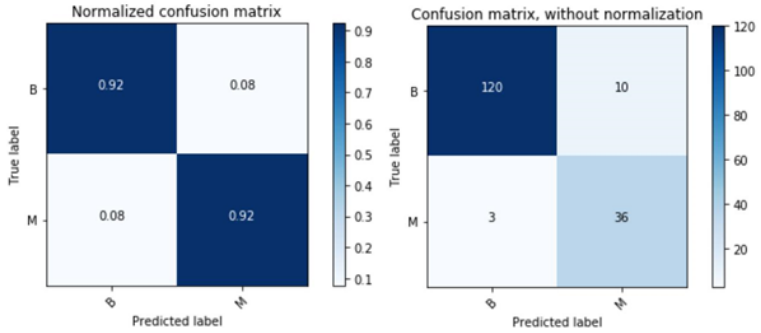


Fig. 6.11 -Normalized and Confusion Matrix of Logistic Regression without PCA

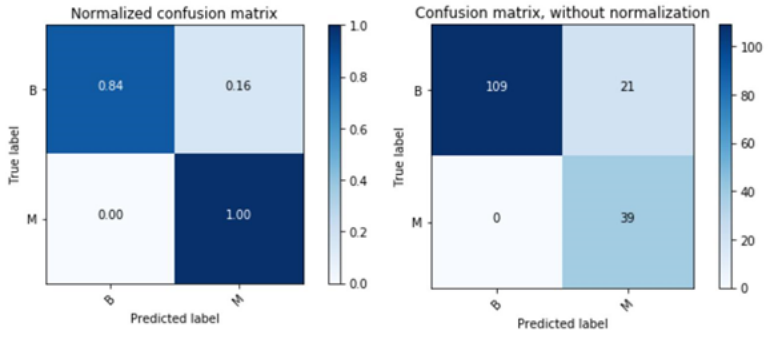


Fig. 6.12 -Normalized and Confusion Matrix of Logistic Regression with PCA

Naïve Bayes before and after applying PCA

The figures below illustrates the Normalized and the Confusion Matrix of Naïve Bayes for both without(Figure 6.13) and with applying PCA (Figure 6.14). 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 values of accuracy, precision and F1 score, 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.

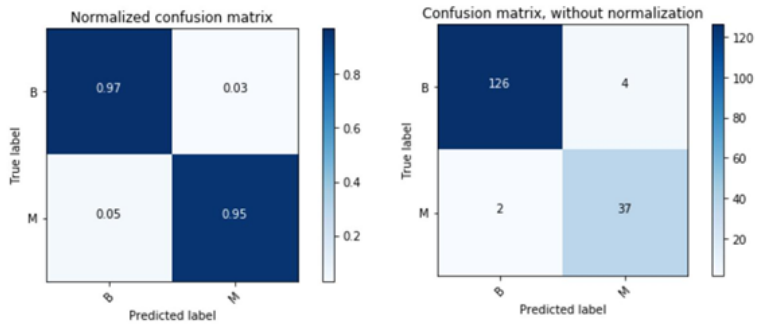


Fig. 6.13 -Normalized and Confusion Matrix of Naive Bayes without PCA

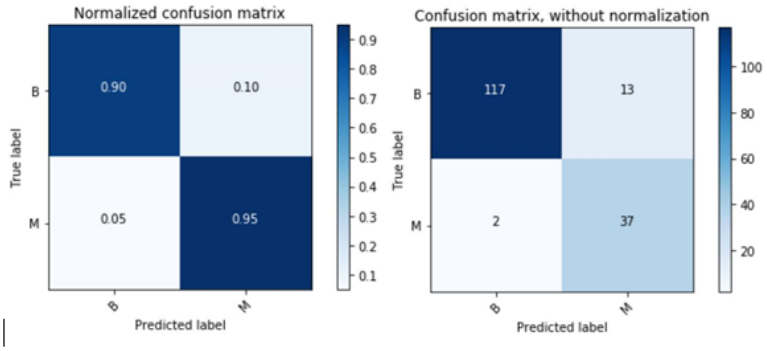


Fig. 6.14 -Normalized and Confusion Matrix of Naive Bayes with PCA

Support Vector Analysis (SVM) before and after applying PCA

The figures below illustrate the Normalized and the Confusion Matrix of Support Vector Ma-chine (SVM) for both without (Figure 6.15) and with applying PCA (Figure 6.16) on the dataset. Results 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 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.

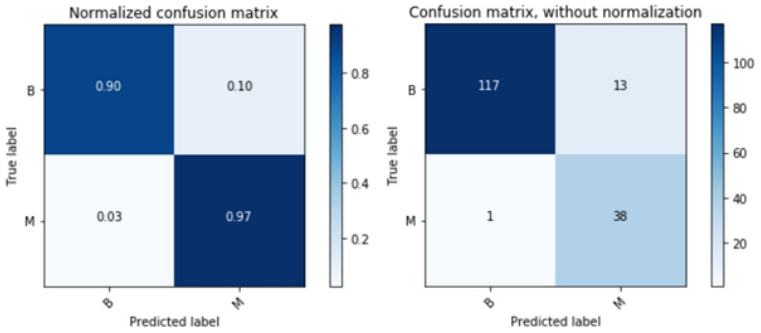


Fig. 6.15 -Normalized and Confusion Matrix of SVM without PCA

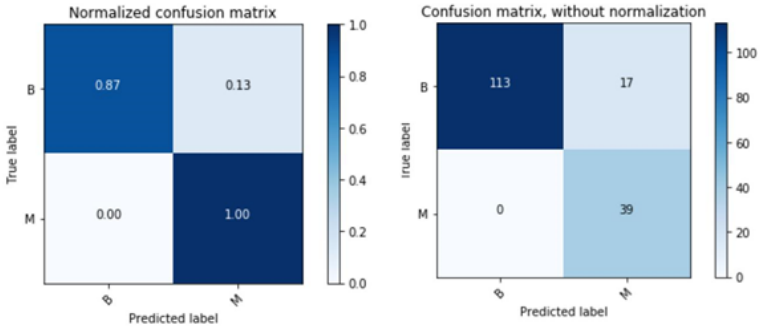


Fig. 6.16 -Normalized and Confusion Matrix of SVM with PCA

Chapter 7

Conclusion

7.1 Summary

In terms of accuracy, Linear Discriminant Analysis (LDA) and Naïve Bayes, have scored high figures of 0.9704 and 0.964 respectively, without applying PCA. K-Neighbors (0.9349) and Logistic regression (0.923) are not far behind either. SVM scores 0.917 in accuracy. Decision Tree performs the worst among all six resulting 0.834. Application of PCA declines the accuracy of all the algorithms except Decision tree. However, the accuracy figures are still higher than that of Decision tree's LDA, again, performs best after PCA is applied, even though there is a fall in ac-curacy (0.917). Considering the other performance matrix into account, a lot can be determined regarding the performance of the algorithms. Decision tree, K-Neighbors and Naïve Bayes per-forms better without the introduction of PCA, while LDA, Logistic Regression and SVM per-forms better after PCA is applied to the dataset. SVM and Logistic Regression scores a perfect 1.000 when it comes to recall, which is vital in terms of disease prediction, after PCA is applied, even though there are declines in the values of all other performance metrics of both the men-tioned algorithms. Keeping in mind that PCA reduces the run time exponential to huge extends in datasets (both small and large alike) and keeping the recall score into consideration, we can conclude that Logistic Regression and Support Vector Analysis with PCA performs better when it comes to Breast Cancer Prediction for this dataset used.

7.2 Limitations

While we were successful at attaining results with precise accuracies, there were certain hin-drance which build up while carrying out this thesis. The initial issue was the lack of a signifi-cantly large dataset. While we did achieve accuracy with over 90% without PCA for all algo-rithms except decision tree, it cannot be denied that the algorithms could have been tested better with a large dataset. Availability of a large dataset could also test the runtime of algorithms run with PCA, since it is very difficult to trace out exactly how fast the algorithms run after PCA is applied on current dataset. Furthermore there is a lack of complex models used in this thesis. Even though we obtained better results with the models we used, use of more complex models can capture complex interactions among features.

7.3 Future Works

Despite attaining accurate results and accuracies with the six algorithms we have used, we wish to confirm the results we obtained are not biased thanks to the scale of our dataset. We would like to search 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 also put our analysis to the test, if we can identify the right parameters from our current and future datasets in order to generate ROC curves. Additionally, besides the models we have tried, we would conjointly wish to attempt other algorithms such as Adaboost in order to compare results and continue our search for the best model for prediction. The idea of applying other feature selection on the currently used models is also under consideration, such as the Recursive Feature Elimination and the Correlation Heat Map.

References

[1] Breast cancer facts and figures 2003-2004 (2003). American Cancer Society.

[2] Stages | Mesothelioma | Cancer Research UK Breast cancer survival statistics September 26, 2017

[3] Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.

[4] deepsense.ai What is reinforcement learning? The complete guide July 05, 2018

[5] Hacker Noon Absolute Fundamentals of Machine Learning – Hacker Noon January 15, 2018

[6] Furundzic, D.; Djordjevic, M.; Bekic, A.J. Neural networks approach to early breast cancer detection. J. Syst. Archit. 1998, 44, 617–633. [CrossRef]

[7] Floyd, C.E.; Lo, J.Y.; Yun, A.J.; Sullivan, D.C.; Kornguth, P.J. Prediction of breast cancer malignancy using an artificial neural network. Cancer 1994, 74, 2944–2948. [CrossRef]

[8] Fogel, D.B.; Wasson, E.C.; Boughton, E.M. Evolving neural networks for detecting breast cancer. Cancer Lett. (1995), 96, 49–53. [CrossRef]

[9] Fogel, D.B.; Wasson, E.C.; Boughton, E.M.; Porto, V.W.; Angeline, P.J. Linear and neural models for classifying breast masses. IEEE Trans. Med. Imaging (1998), 17, 485–488. [CrossRef] [PubMed]

[10] Setiono, R. Extracting rules from pruned neural networks for breast cancer diagnosis. Ar-tif. Intell. Med. (1996), 8, 37–51. [CrossRef]

[11] Wilding, P.; Morgan, M.A.; Grygotis, A.E.; Shoffner, M.A.; Rosato, E.F. Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. Cancer Lett. (1994), 77, 145–153. [CrossRef]

[12] Wu, Y.; Giger, M.L.; Doi, K.; Vyborny, C.J.; Schmidt, R.A.; Metz, C.E. Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology (1993), 187, 81–87. [CrossRef] [PubMed]

[13] H. A. Abbass, “An evolutionary artificial neural networks approach for breast cancer diagnosis.” Artif.Intell. Med., vol. 25, no. 3, pp. 265–81, Jul. 2002.

[14] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, “Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.,” Nat. Med., vol. 7, no. 6, pp. 673–9, Jun. 2001.

[15] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, (Dec. 2006.) vol. 70, no. 1–3, pp. 489–501.

[16] C. P. Utomo, A. Kardiana, and R. Yuliwulandari, “Breast Cancer Diagnosis using Artifi-cial Neural Networks with Extreme Learning Techniques,” Int. J. Adv. Res. Artif. Intell, vol. 3, no. 7, pp. 10–14, 2014

[17] C. P. Utomo, P. S. Pratiwi, A. Kardiana, I. Budi, and H. Suhartanto, “Best-Parameterized Sigmoid ELM for Benign and Malignant Breast Cancer Detection,” pp. 50–55, 2014

[18] William H Wolberg, W Nick Street, and Olvi L Mangasarian. (1992). Breast cancer Wisconsin (diagnostic) data set. UCI Machine Learning Repository.

[19] Cristianini N, Shawe-taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, (2000) London: Cambridge University Press.

[20] Joachims T. Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Learning. (1998) MIT Press, Cambridge, MA, 169-184.

[21] Machine Learning Mastery Discriminant Analysis for Machine Learning September 22, (2016)

[22] DUNTEMAN, G. H. Principal component analysis. quantitative applications in the social sciences series (vol. 69), 1989.

[23] Puneet Yadav, Rajat Varshney, Vishan Kumar Gupta. Diagnosis of Breast Cancer using Decision Tree Models and SVM (2016)

[24] Rohith Gandhi. Nearest Neighbor. Understanding Machine Learning (2018)

[25] Adi Bronshtein. Train/Test Split and Cross Validation in Python. Understanding Machine Learning (2017).

[26] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. ”Regularization paths for generalized linear models via coordinate descent.” Journal of statistical software 33.1 (2010): 1

[27] Fine Needle Aspiration Biopsy of the Breast. American Cancer Society