**A**

**Project Report on**

**BREAST CANCER DATA ANALYSIS**

**Submitted to the partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Anish Thakur (RA1511003030317)**

**Shobhit Aggarwal (RA1511003030375)**

**Shahvez Malik (RA1511003030327)**

**Priyam Kumar (RA1511003030348)**

**Supervised by:**

**Ms. Shivangi Tyagi**

**Assistant Professor**

****

**SRM Institute of Science and Technology**

**(Deemed to be University u/s of UGC Act, 1956)**

**NCR Campus, Modinagar**

**2019**

|  |  |  |
| --- | --- | --- |
|  | **TABLE OF CONTENTS** |  |
| **TOPICS** |  | **PAGE** |
|  |  |  |
| Bonafie certificate |  | 2 |
| Acknowledgement  Declaration |  | 3  4 |
| Abstract  List of figures   1. 1. Introduction    1. Background    2. Puropose of the project    3. Problem statement 2. . Literature review 3. Methodology and Requirments |  | 5  8 |

4. Algorithm Used

4.1 KNN Alogorithm

4.2 Principal Component Analysis

4.3 K Means

5. Hardware and Software Requirments

5.1 Hardware Requirments

5.2 Software Requirments

5.3 Library Used

6. Diagram

6.1 Flowchart

7. Conclusion

8. Refrences

**LIST OF FIGURES**

Fig 2.1 Heat Map of Data Attributes

Fig 3.2 Value Counts of Malign and benign data

Fig4.2 Cummulative graphs of eigon values

Fig 4.3.1 Selecting K value Plot

Fig 4.3.2 3D plot of K means Clustering

**1. INTRODUCTION**

**1.1 Background**

Breast cancer is a type of cancer which occurs mostly among women throughout worldwide, the majority of breast cancer cases representation and cancer-related deaths increasing drastically which makes it the most significant health issue in women.

The ratheripe diagnosis of Breast Cancer can improve the pessimistic prediction and chance of survival significantly, as it can increase timely clinical treatment to patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, the correct diagnosis of Breast Cancer and classification of patients into malignant or benign groups is the subject which needs more research.Machine learning is widely recognized as the methodology of choice in Breast Cancer pattern classification and classification modelling.

Data mining and classification methods both are effective ways to classify the data. Especially in medical field, where these methods are used very much in diagnosis and analysis to make decisions.

When healthy cells in the breast of female change and grow out of control which is the evidence of the formation of breast cancer, they form a excessive layer of cells or large mass that is known as tumor. A tumor could be benign or malignant. A benign tumor is not so dangerous because the tumor can grow but will not spread. Malignant tumor is cancerous , meaning it can grow and spread to other parts of the body. Breast cancer spreads when the cancer starts to grow in other parts of the body or when breast cancer cells comes in contact with other part of body through the lymph or blood vessels. This is known as metastasis.Our study covers locally advanced breast cancer and early-stages of cancer, which includes stages I, II, and III. The stage of breast cancer helps us to describes in which part of body the cancer is located, how much the cancer has grown, and if or where it has spread.

Although mostly the breast cancer spreads to nearby lymph nodes, it can also starts to spread further through the body to areas such as the lungs, liver, bones and brain. This is known as metastatic or stage IV breast cancer.

If breast cancer tumor develop after the initial treatment, it can be cure again , meaning in the breast and/or regional lymph nodes. The regional lymph nodes are those nearby the breast, such as the lymph nodes under the arm. It can also recur elsewhere in the body, called a distant recurrence or metastatic recurrence.

**Objective :**

The main objectives of our project is to develop a flexible diagnostic model which can assist in :

* a better understanding of breast cancer survivability in the absence of data,
* providing better insights into factors associated with patient survivability, and
* Establishing cohorts of patients that share similar properties.

**1.2 Purpose of the Project**

The “Breast Cancer Data Analysis” is a predicting model which uses machine learning techniques and a diagnostic model which focuses on the prediction of cancer .

The Breast Cancer Data analysis model works on the basis of symptoms of various attributes present in patient report. It can take the attributes like tumor length, height, perimeter and radius etc. we can implement the various algorithm in our data frame in which knn is the best classification algorithm among other algorithm because it gives good recall score in comparison to others. After classification algorithm we implement feature extraction technique(Principle Component Analysis) to remove redundancy (attributes having common values). After applying feature extraction technique we conclude that some attributes can also be applicable to predict cancer . But, by removing redundancy recall value decreases . So, we don’t remove redundancy from dataframe. That’s why we use ensemble technique .

Main purpose of the model is to solve few problems which helps in daily life :

* To raise the awareness of breast cancer among women .
* To predict whether the patient have enough time for treatment .
* Prediction of the stage of breast cancer.
* Improve the accuracy of model whether it predicts the cancer is benign or malign using different techniques.

**1.3 Problem Statement :**

In today’s world health is the biggest issue as people are more effecting due to the food they eat and the environment in which they live.

Doctors are not able to judge the cancer before the time period as doctors are not aware of the symptoms the patient carry which leads them to the cancer. Doctors are able to predict cancer of mostly last stage because of less information availability. So, machine learning model helps the doctor to predict at the early stage of the cancer by taking all the necessary points into consideration. Machine Learning models are fast and flexible to predict the cancer in the patient. As model are not replacing the doctors but give the more flexibility to analyze the data given of the patient.

As predicting cancer is classification problem so we applied various classification models to predict. As all the models are not predicting well on the dataset where KNN comes out of picture and perform well on the dataset as KNN is the distance measure model so features of benign and malign cancer are less distance. Hence KNN performing best in all of other models. Using PCA which help to remove the redundancy given to the standalone model and it will improve the accuracy by removing the redundancy.

In the end use the k fold technique to work on the full data and provide the measure of Spread or variance of the data to the model and then it will perform best on the test data or unseen data.

**2. LITERATURE REVIEW**

Literature review is the most essential part of any project construction. We perform literature review over many articles and research papers before starting of our project construction to get help in not just building and development of the project but all to know how and from we can start over the prior researches that were already been done.

Konstantina Kourou, Themis P. Exarchos (2015) write a research paper “Machine learning applications in cancer prognosis and prediction” variety of these techniques, including Support Vector Machines (SVMs) and Decision Trees (DTs) have been widely applied in cancer research for the development of predictive models, resulting in effective and accurate decision making. Even though it is evident that the use of ML methods can improve our understanding of cancer progression, an appropriate level of validation is needed in order for these methods to be considered in the everyday clinical practice.

Hiba Asri, Hajar Mousannif (2016) “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis” In this paper, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB)

and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. 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, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy

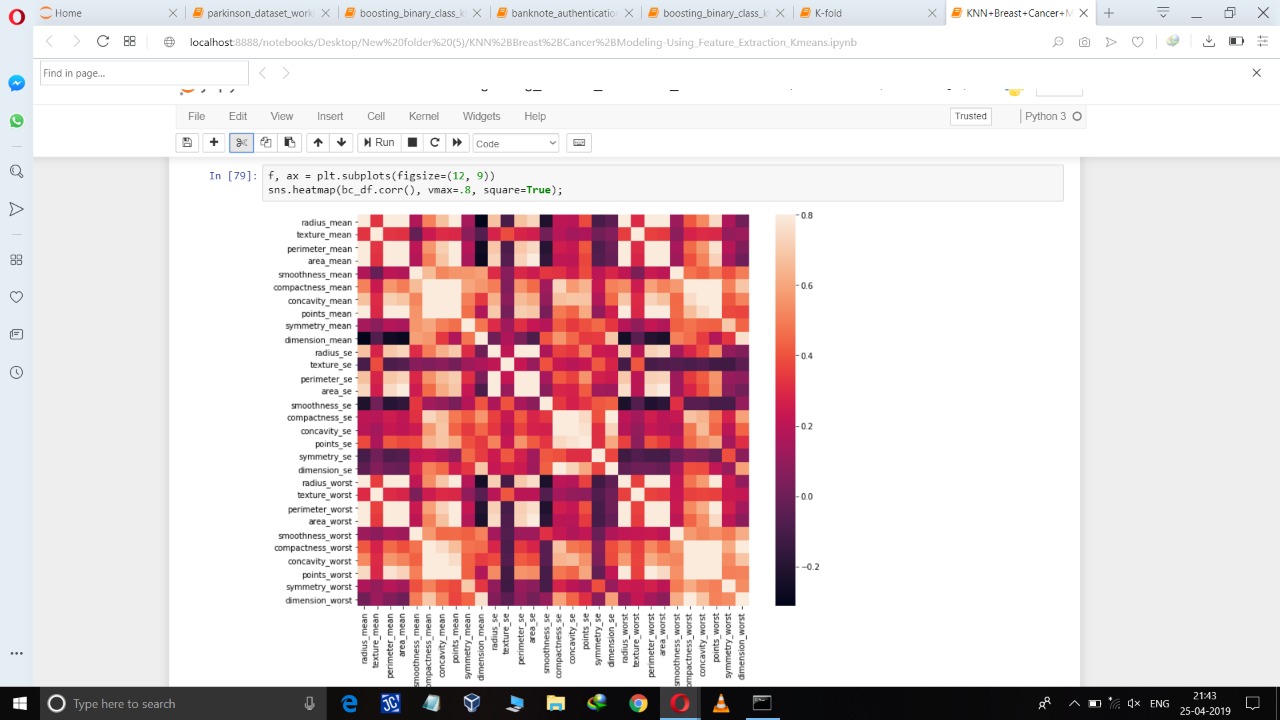
(97.13%) with lowest error rate.

Ahmed Abd El-Hafeez Ibrahim published a paper “Performance analysis of various Open Source Tools on four Breast Cancer Dataset using Ensemble Classifiers Technique”.Analysis the performance of different classification algorithms shows that using Ensemble Classifiers Techniques improved the accuracy .

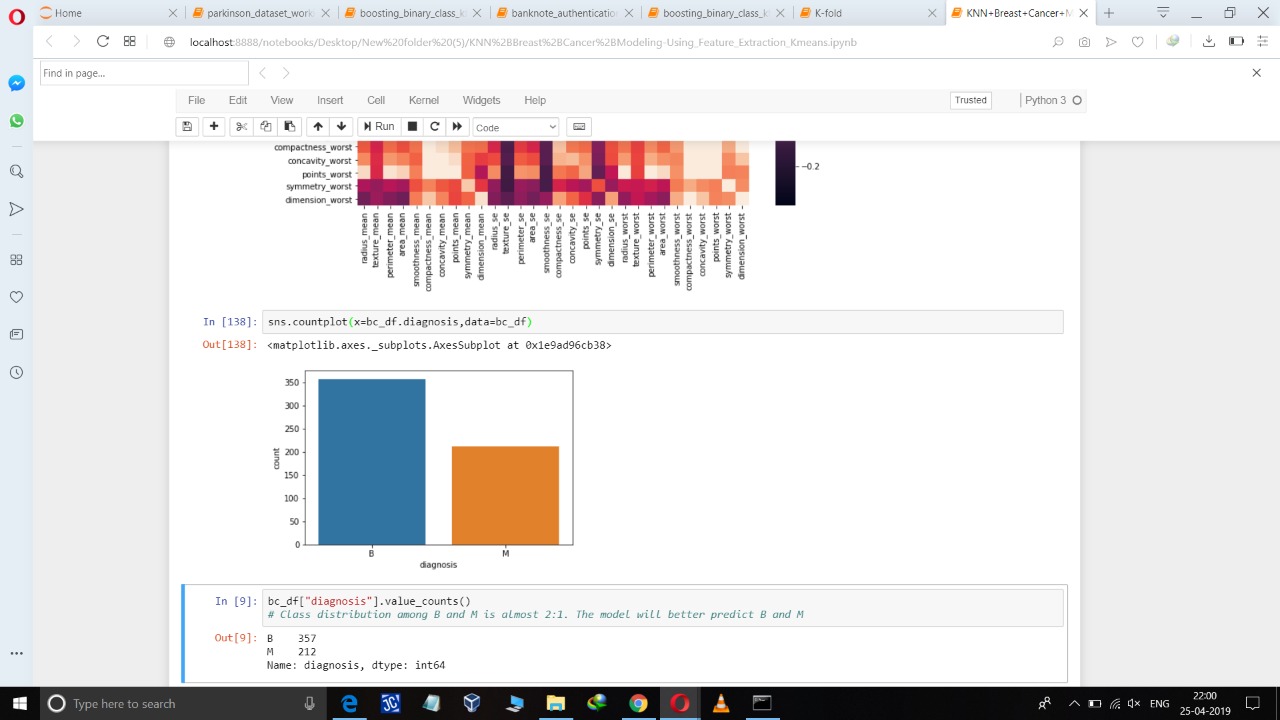
As different models use different approaches for the prediction of breast cancer . Each approach give its own accuracy rate for the prediction of breast cancer. Mostly models use SVM (Support Vector Machine) for the prediction as a classification algorithm to find accuracy . In our model we use a different approach to predict the cancer .Using KNN algorithm on the data collected of breast cancer patient reports we apply ensemble technique . As a result our model can predict the cancer with a accuracy rate of 98 % which is more than any of the approach used be another researchers.

**3. METHODOLOGY AND REQUIRMENTS**

This section gives insights the dependency of target variables on independent variables using machine learning techniques to predict the type of cancer .The dependent variable is “diagnosis” whereas independent variables i.e. area\_mean, perimeter\_mean, radius\_mean, texture\_mean etc. are assumed to be predictors . For the analysis on these variables are done in two different ways.



**Fig 2.1 Heat Map of Data Attributes**

Firstly, classification algorithms are used to check the accuracy score of every classification algorithm and KNN is used to evaluate the value of target variable that is diagnosis 

**Fig 3.2 Value Counts of Malign and benign data**

While evaluating the data model various types of errors are occurred like over fitting, introduced from having too large of a training set and bias occur due to too small of a test set. So we applied PCA on the data to remove the redundancy of the model and to decrease the complexity of the model by removing the extra features from the model and inputting the spread of the data at each dimension.

While evaluating the data is divided into two parts, first is the training set and second is the test set. e

Firstly, k-fold cross validation or k-fold CV randomly divides the set of data into the number of K splits as value of K will be decided on the basis of the shape of the data , then using k-1 blocks datasets are trained and to test the performance of different algorithm, the remaining blocks are. Repeat the whole process k times and average value is calculated.

Second method is percentage split method in which we use train test split to split the data into percentage. Mainly 70% of the dataset is used for training and 30% of the dataset is used for testing.

We can evaluate the effectiveness of model by classification matrix, used to calculate the accuracy,precision,recall,f1-score.

Recall = TP/TP+FN

Precision=TP/TP+FP

Accuracy=TN+TP/TN+TP+FN+FP

F1-score=2\*precision\*recall/ precision+ recall

**\**

|  |  |  |
| --- | --- | --- |
| **n= 569** | **Predicted :**  **B** | **Predicted :**  **M** |
| **Actual : B** | **107** | **2** |
| **Actual : M** | **7** | **55** |

**Table 3.1 Confusion Matrix**

**4. ALGORITHMS**

This section gives insights the dependency of target variables on independent variables using machine learning techniques to predict the type of cancer .The dependent variable is “

diagnosis” whereas independent variables i.e. area\_mean, perimeter\_mean, radius\_mean, texture\_mean etc. are assumed to be predictors . For the analysis on these variables are done in two different ways.

Firstly, classification algorithms are used to check the accuracy score of every classification algorithm and KNN is used to evaluate the value of target variable that is diagnosis

While evaluating the data model various types of errors are occurred like over fitting, introduced from having too large of a training set and bias occur due to too small of a test set.

Analysis of cancer can be done in two different ways:

1. By doing the laboratory test
2. Test done by the doctors which is time consuming and predict after the particular stage has been crossed and patients are more prone to death.

**4.1 KNN**

KNN is non-parametric, it means that it does not make any assumptions on the underlying data distribution.

In other words, the model structure is determined from the data. KNN probably should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution data. KNN is a simple yet powerful classification algorithm.

* KNN is also known as a lazy algorithm since does not use the training data points to do any generalization.
* Lack of generalization means that KNN keeps all the training data. To be more exact, all (or most) the training data is needed during the testing phase
* KNN does not try to learn a function from the training data, so it is called Nonparametric model, It memorize the pattern from the dataset.
* KNN requires no training for making predictions, which is typically one of the most difficult parts of a machine learning algorithm.

The intuition behind the KNN algorithm is one of the simplest of all the supervised machine learning algorithms:

1. It simply calculates the distance of a new data point to all other training data points. The distance can be of any type e.g., Euclidean or Manhattan etc.

2. It then selects the K-nearest data points, where K can be any integer.

3. Finally it assigns the data point to the class to which the majority of the K data points belong.

4. This algorithm segregates unlabeled data points into well-defined groups.

KNN uses two distance matrix named as euclidian and manhattan. We use Euclidian distance instead of manhattan distance because it gives better result then manhattan distance

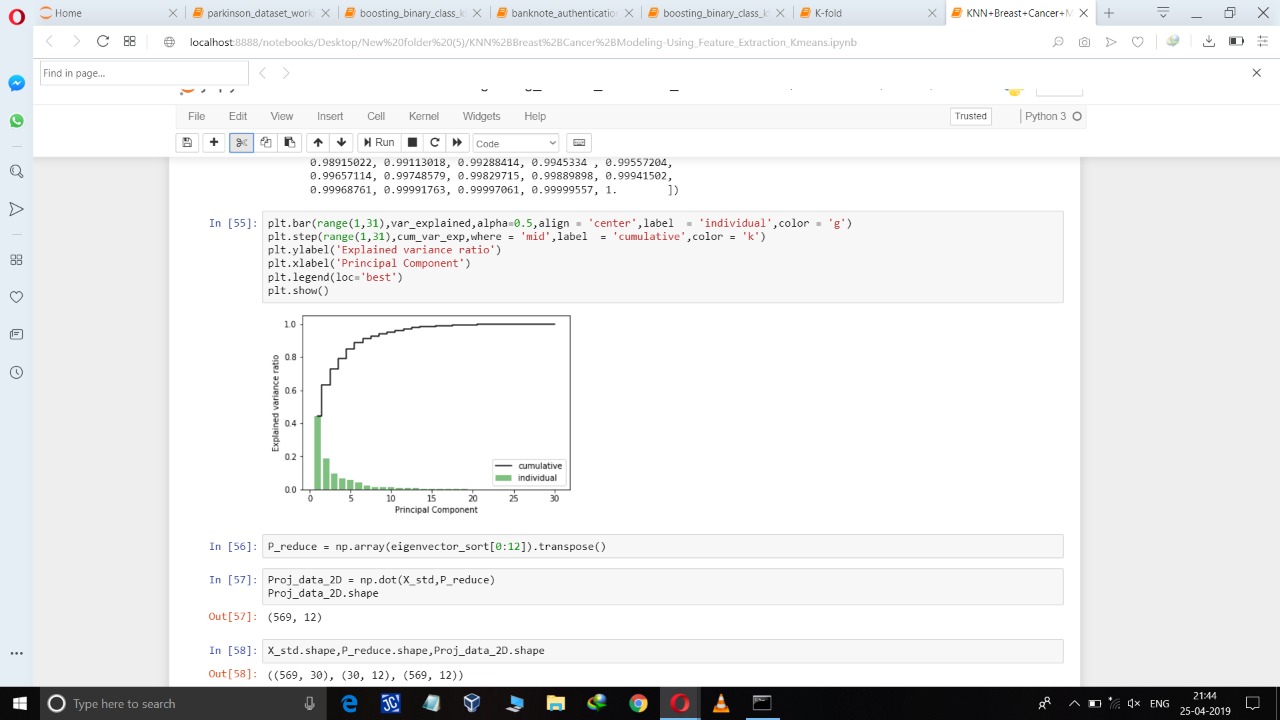
KNN uses two voting methods one is weighed and unweighed.

**4.2 PCA**

In model we pass the center value to the data .Model does not know covariance present in the data and PCA is the way to tell the covariance or the spread of the data.PCA is done on the independent variable then only it will tab the information otherwise model is not able to tab the information related to the covariance of the data.

X1,X2 are the independent be passed in the model .Instead of passing X1 and X2 we passed the relation between X1 and X2 to Y. So multicolinearity is taken care by PCA process which helps to reduce redundancy effect in the model. Redundancy is the process present in the data in the form of noise in the dataset.

As noise decreases information tends to increase.



**Fig4.2 Cummulative graphs of eigon values**

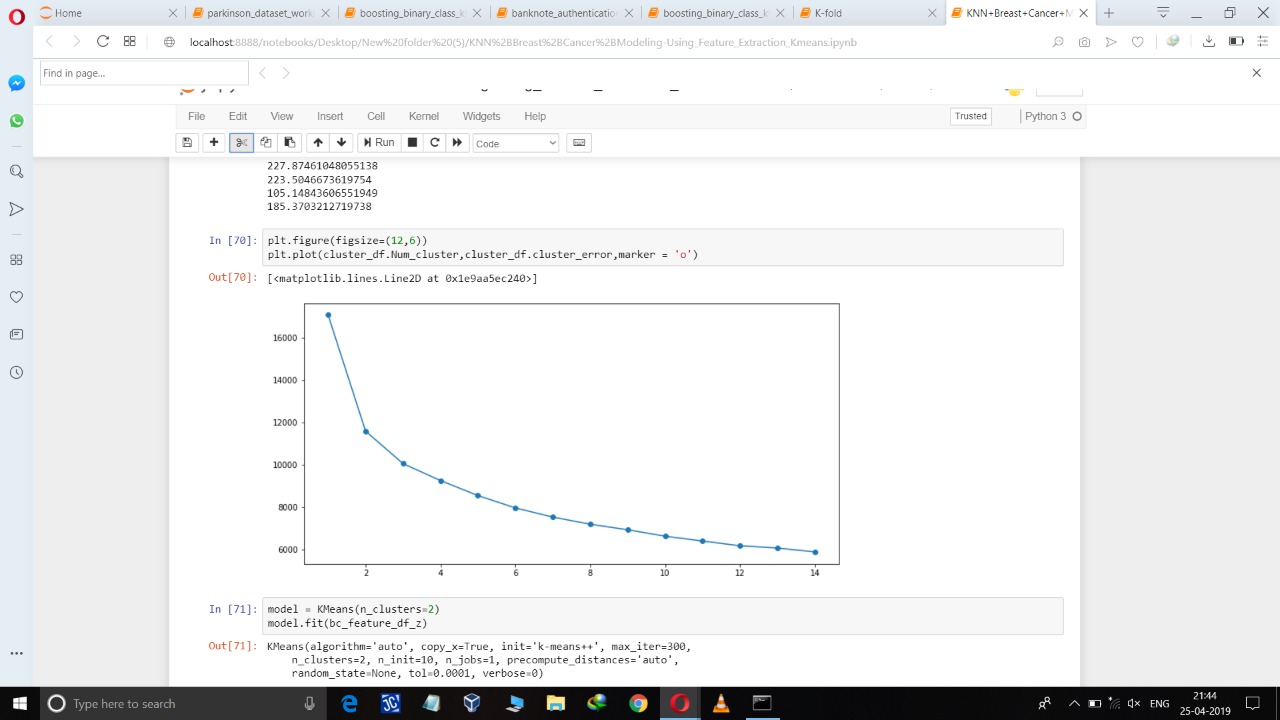
Methods to imply PCA on dataset.

1. Firstly,we standardize the data (using z-score)
2. Secondly, convert the data into covariance matrix.
3. Using linear algebra we convert the covariance matrix into eigen vectors and eigen values.
4. Then, we sort the eigen vectors according to eigen values.
5. Then, creating the cumulative plot of the eigen values.
6. Then, we reduce the number of features and apply dot product on standardize data.

**4.3 K Means**

K means is an unsupervised technique which helps to group the data on the basis distance apart from one another.

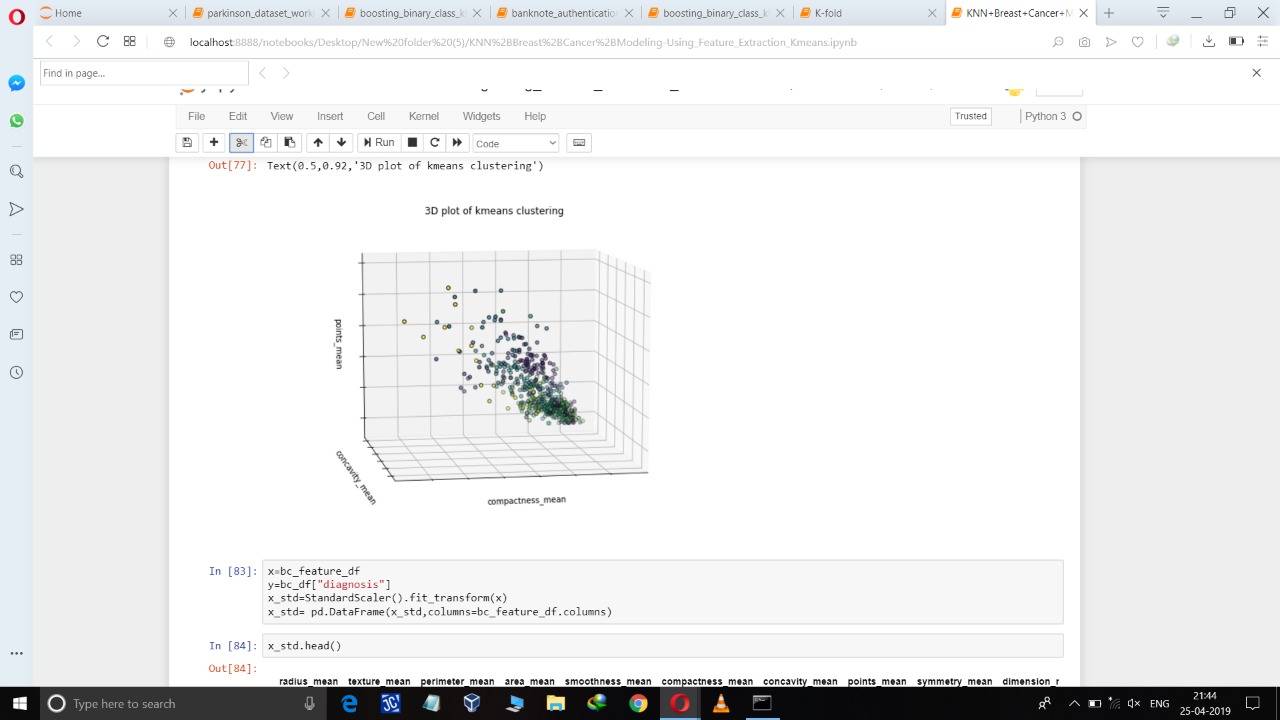
K means cluster our data as much as possible to the equal density(equal variance group)



**Fig 4.3.1 Selecting K value Plot**

The working of algorithm :

1. First we input the k value according to the elbow plot.
2. We take mean of every data point and the mean which is closest to the cluster mean then the data point will attach to that cluster .
3. We repeat the process for a given k value we have our clusters.



**Fig 4.3.2 3D plot of K means Clustering**

**4.4 Ensemble techniques**

By combining several models we can improve the machine learning results and this technique is known as ensemble technique .This approach gives better predictive performance compared to single model.

Ensemble techniques are used to decrease the variance as well as bias error and ensemble learning helps to stack more then one machine learning model.

There are three types of ensemble learning

1 Bagging

2 Boosting

3 Stacking

**4.4.1 BAGGING :**

Bagging is a method in which resampling of data with replacement .It is a technique which works only on decreasing the variance error .It works on the principle of Bootstrap sampling.

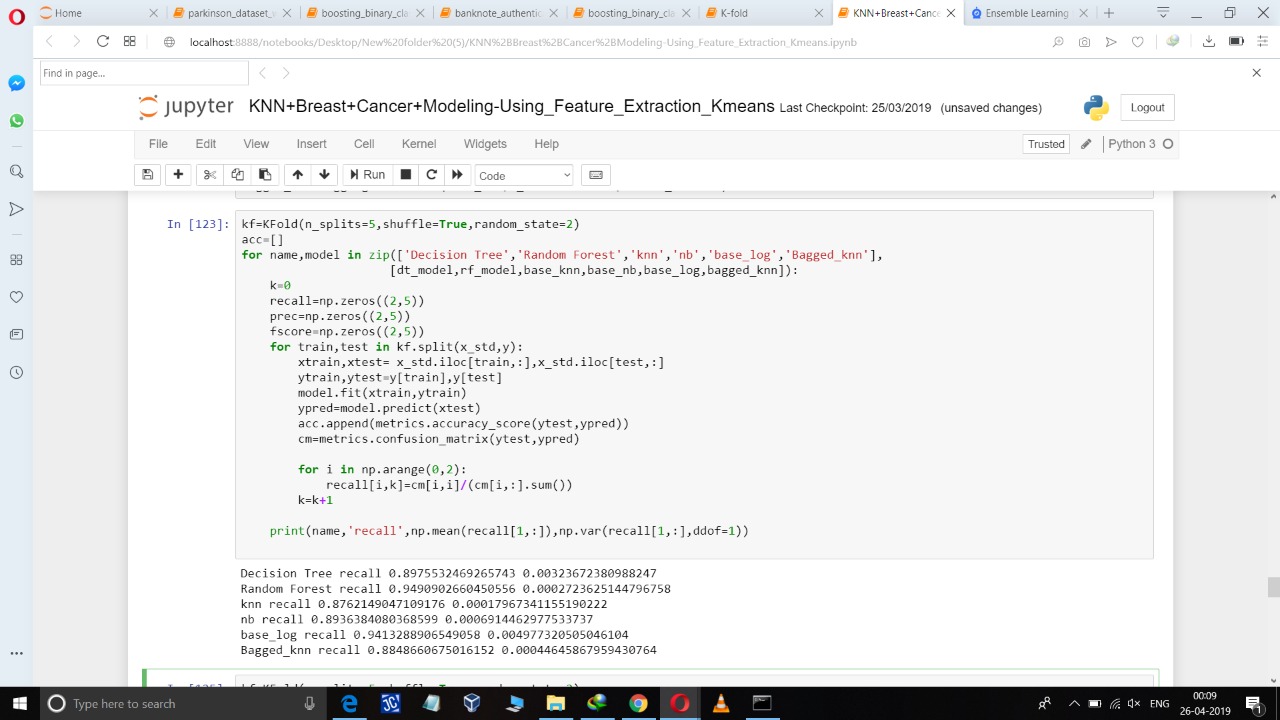
Bootstrap sampling is a technique in which chances of same data point occurs in a bag comes fourth times if we take five bags .So there are so many duplicates or triplicate in every bag .Because records are picked with replacement .So there are certain records which are unpicked known as out of bag records(OOB) and typically we use OOB records as testing .By doing this we are providing generalization to the model.

**4.4.2 BOOSTING :**

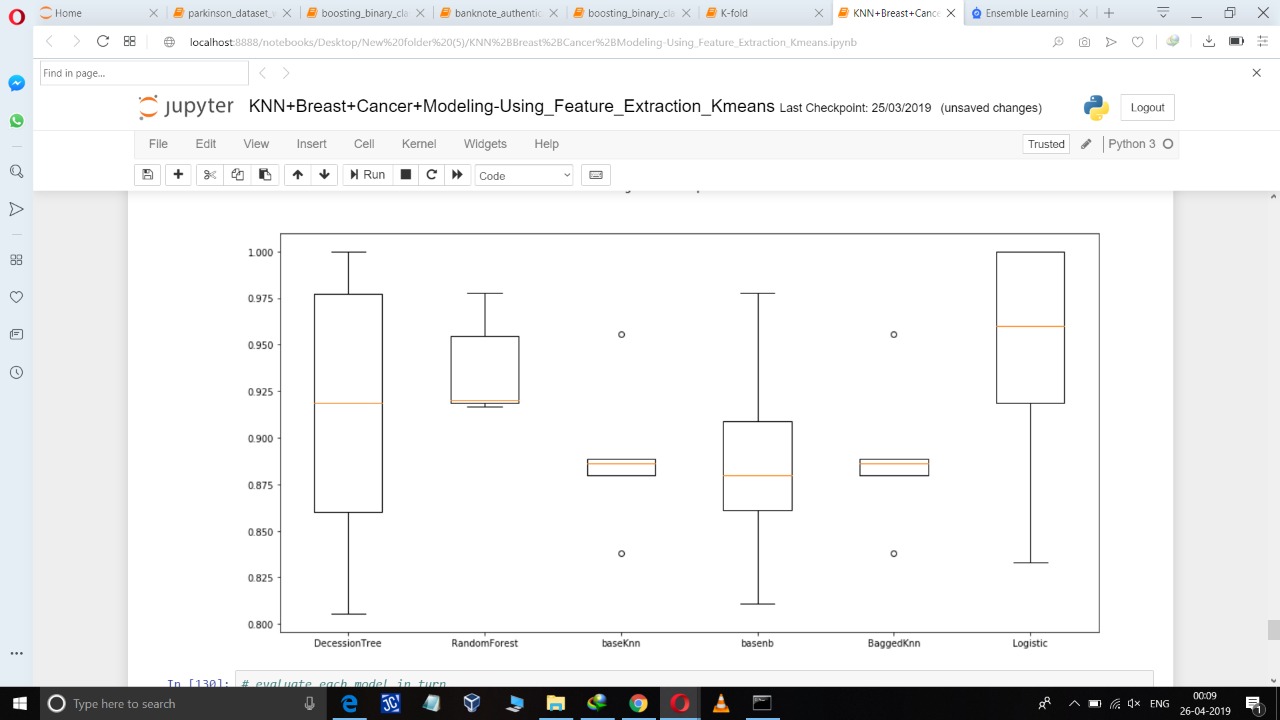
Boosting is same like bagging but in bagging all bags works parallely. But in boosting one bag has to complete the process first then next bag starts doing this process .The first bag has to complete this task which picks all the records with replacement .In first bag all records have equal probability but for second bag the probability is not equal as we changed the probability of some data points .The probability of certains data points gets changed because bag one misclassify the certain data points so it increases its probability due to which bag two will solve the data point with higher probability and last bag is the highest complex bag because highly misclassified data points which contain the high probability value makes the model bag more complex so it is a sequential approach.

**4.4.3 STACKING :**

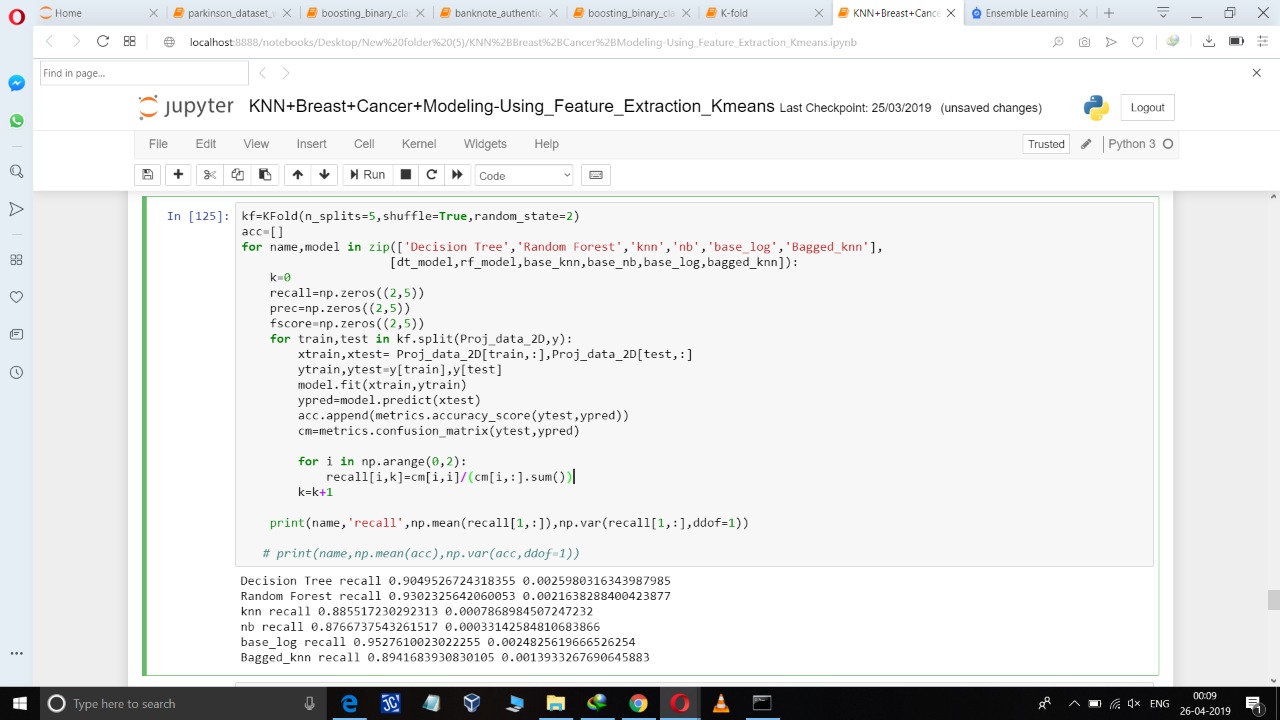
Stacking is the simplest approach in which there are two or three models which performs better on the dataset .will be group together and weighed voting is used to evaluate the model.



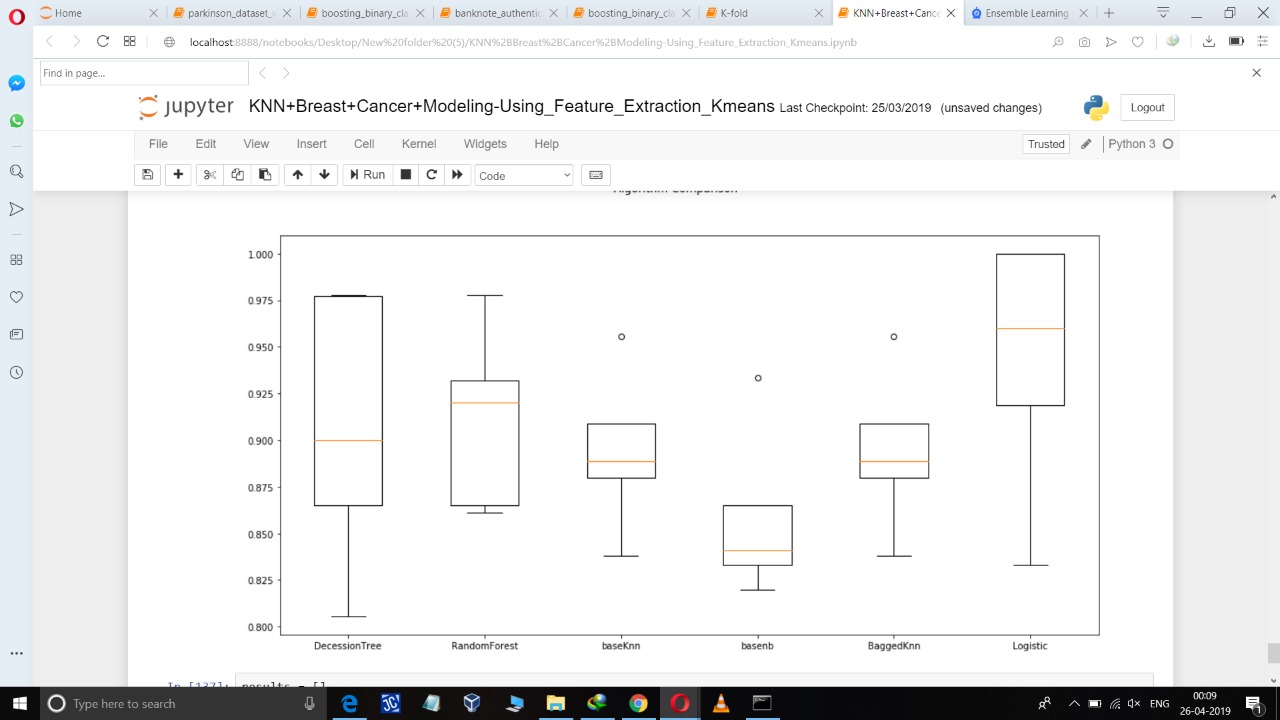
**Fig 4.4.1 : Bagged KNN model With variance and bias error using standardize data**



**Fig 4.4.2 : Graphical Repersentation Std Data**



**Fig 4.4.3 : Bagged KNN model With variance and bias error using PCA data**



**Fig 4.4.4 : Graphical Repersentation local Data**

**3.3 Hardware and Software Requirements**

**3.3.1 Hardware Requirements**

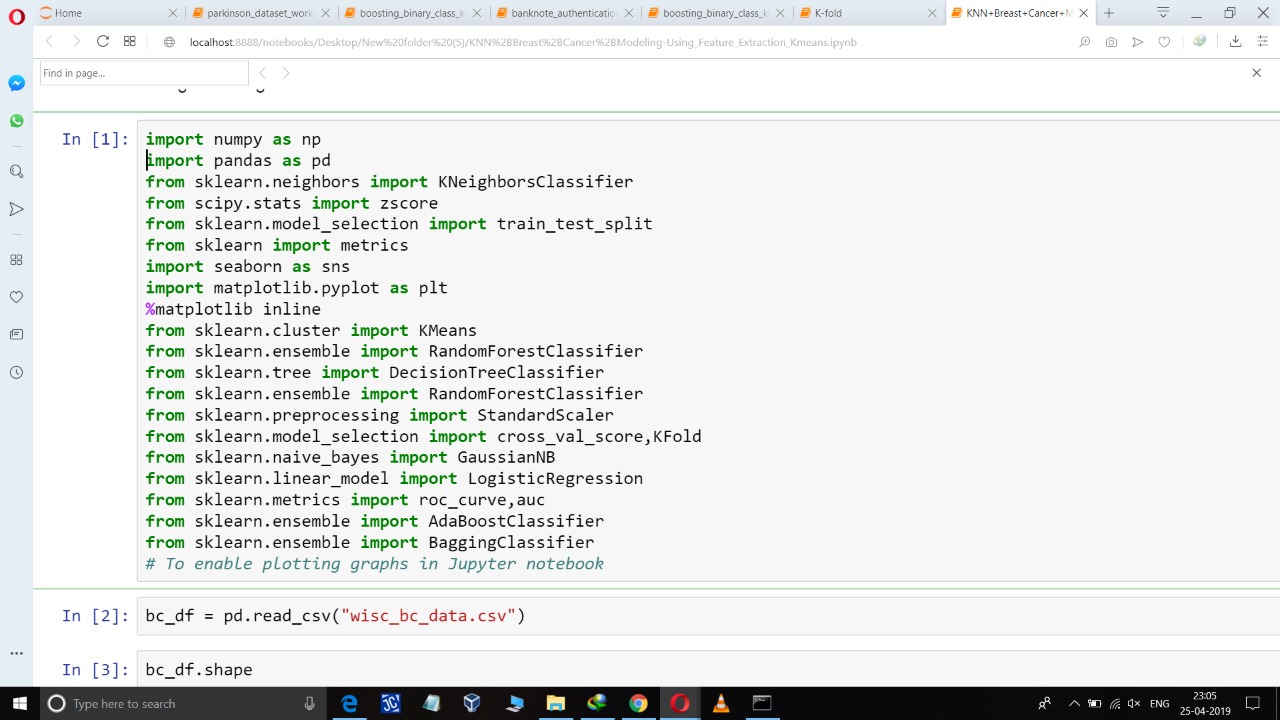
* Personal computer /laptop
* Internet connection 4G LTE
* Atleast 4GB RAM
* I5 processor octacore

* 20 GB storage minimum

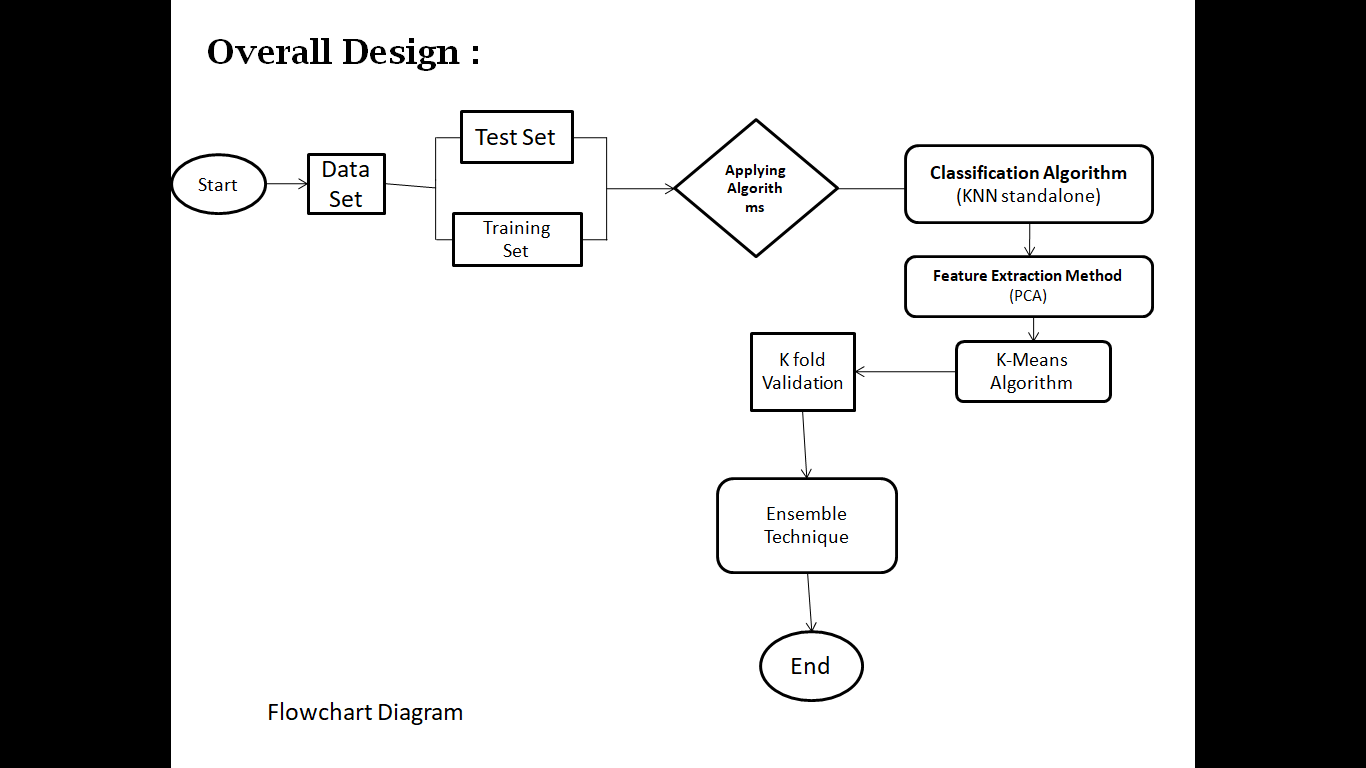
**3.3.2 Software Requirements**

* Python 3.0
* Jupyter Notebook
* Anaconda
* Web Browser

**3.3.3 Library Used**



**FlOWCHART :**

****

**CONCLUSION :**

This paper deals with KNN algorithm to classify cancer tumors as either benign or malignant. We applied feature selection on the dataset to remove duplicate and irrelevant features. we applied symmetrical uncertainty attribute evaluation in WEKA for feature selection. Our proposed approach is evaluated and compared using Wisconsin breast cancer dataset. The experimental result showed that accuracy, precision, recall, and F-measure are increased by our proposed method when compared with different models. In future, we will work on feature selection techniques to improve the accuracy of the model.