Artificial Intelligence J Component Report

submitted by

PRABHVEER SINGH KHURANA(19BCE0280)

KAARTIKEYA PANJWANI(19BCE2124)

UTKARSH VERMA - 19BCE0078

On

Breast Cancer Detection using supervised learning classifier Algorithm

in partial fulfilment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

I, Kaartikeya Panjwani, hereby declare that the thesis entitled “Breast Cancer Detection using supervised learning classifier Algorithm" submitted by me, for the award of the degree of Bachelor of Technology in Computer Science and Engineering to VIT, Vellore is a record of bonafide work carried out by me under the supervision of Dr. Shalini L.

I further declare that the work reported in this thesis has not been submitted

and will not be submitted, either in part or in full, for the award of any other degree or

diploma in this institute or any other institute or university.

Place: Vellore

Date: 1st December, 2021

Signature of the Candidate

Introduction:

In the present world, approximately 7 million people die of cancer,10% of them being breast cancer patients. Breast cancer is a dominant cancer in women and is increasing in developing countries. Majority of these cases are diagnosed in later stages, resulting in minimal chance of survival. Breast cancer is one of the leading reasons for deaths among all cancers for women. Correct diagnosis and treatment is important to reduce these numbers. Traditional diagnosis of breast cancer is highly dependent on the experience of doctors and visual inspections. Thus, the traditional diagnosis are subject to human errors. Computer diagnosis/ machine learning methods can help physicians improve the accuracy of their diagnosis. Tumours are classified as benign or malignant. Benign tumours are not cancerous or life threatening. Malignant tumours are life threatening. The paper focuses on analysing various ML algorithms such as Random Forest, Naive Bayes,etc for breast cancer detection and calculate their accuracy. For implementing the Ml algorithms, the dataset is partitioned into testing and training phase. The most accurate algorithm can be used to classify the tumour as benign or malignant.

Abstract:

Breast cancer is a dangerous disease for women.

Worldwide near about 12% of women affected by breast cancer and the number is still increasing.

If it does not identify in the early-stage then the result will be the death of the patient. It is a common cancer in women worldwide. The doctors do not identify each and every breast cancer patient. That’s the reason Artificial Intelligence Engineer comes into the picture because they have knowledge of maths and computational power.

Various factors are driving interest in the application of artificial intelligence (AI) for breast cancer (BC) detection, but it is unclear whether the evidence warrants large-scale use in population-based screening.

Two types of Tumours:

1. Benign: Non- cancerous, slow growth, non spreadable, no recurrence after treatment.

2. Malignant: Cancerous, lethal, fast growing mass that spreads rapidly, can spread to other organs.

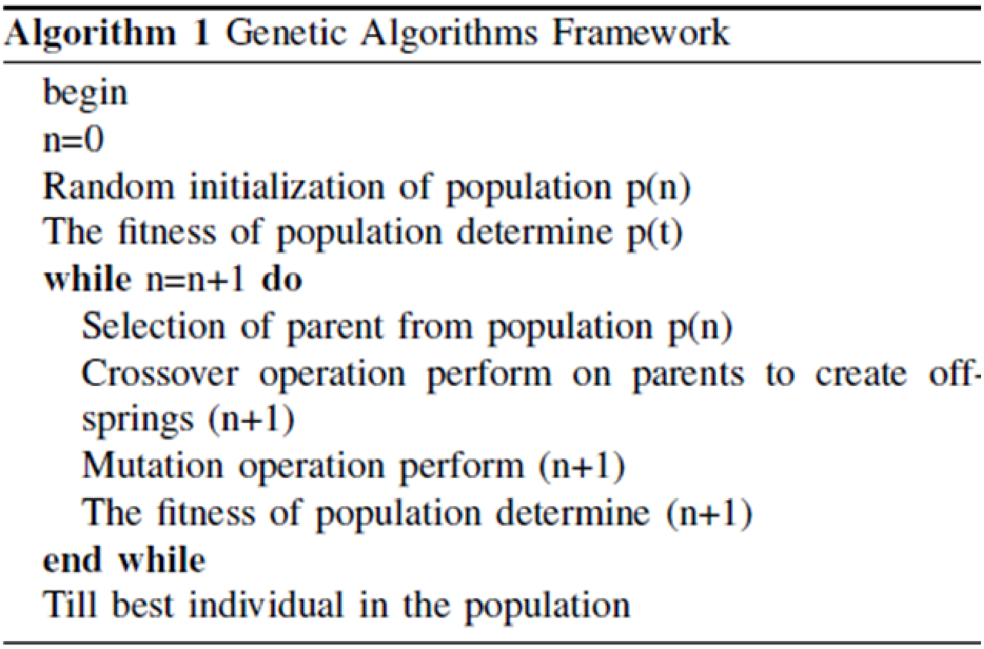
Related work/Literature Survey:

[1]Sunny, Jean & Rane, Nikita & Kanade, Rucha & Devi, Sulochana. (2020). Breast Cancer Classification and Prediction using Machine Learning. International Journal of Engineering Research and. V9. 10.17577/IJERTV9IS020280.

The paper presents a comparison of six popular machine learning algorithms: Naive Bayes, Support Vector Machine, Random Forest, Artificial Neural Network, Nearest Neighbour(KNN), Decision Tree on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. Other techniques such as blood analysis and ultrasonography are also used. Ultrasound characterisation helped doctors observe that breast cancer occurs when breast cells start growing abnormally. Genetic algorithm based weighted average method is used for prediction in various ML models. The proposed system presents a comparison between the algorithms mentioned above. The dataset used is divided into testing set and training set. The dataset has 32 attributes and total instances are 569. The ultra sound image is digitalised and becomes a feature of the dataset. Now the model will detect if the tumour is benign or malignant.

[2] P. Chauhan and A. Swami, "Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2018, pp. 1-8, doi: 10.1109/ICCCNT.2018.8493927.

Breast cancer prediction is a classification problem that can be solved using various ML algorithms. The main area of research is how to improve the accuracy of these algorithms/models. In this paper, the authors use an ensemble algorithm that combines predictions of various algorithms. The classical weighted average method had some limitations. This paper introduces a new GA weighted average method that overcomes the above limitations. In the classical weighted average method, the prediction of various models is combined giving them different weights manually. Now, the limitation is that the weights are given to the predictions, manually. An evolutionary algorithm is used that optimizes weights assigned which overcomes the limitation of the classic weighted average method. The Wisconsin dataset is used for testing the model. For the final ensemble method top three models are used(accuracy wise). Adaboost, SVM and random forest were the models with highest accuracy so, they are used.



Algorithm 1 Genetic Algorithms Framework

```
begin
n=0
Random initialization of population p(n)
The fitness of population determine p(t)
while n=n+1 do
    Selection of parent from population p(n)
    Crossover operation perform on parents to create offsprings (n+1)
    Mutation operation perform (n+1)
    The fitness of population determine (n+1)
end while
Till best individual in the population
```

[3] Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction, by Yixuan Li, Zixuan Chen October 18, 2018

Data mining methods provide effective way to extract useful information to from complex databases. Clustering and classification can be done on the extracted information. 5 different classification models namely Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Neural Network (NN) and Logistics Regression (LR) are used to explore the relationship between breast cancer and some medical attributes to the death probability of the patient. Two datasets are used in the research: Wisconsin Breast Cancer Dataset and Breast Cancer Coimbra Dataset (BCCD). Performance evaluation of the 5 models are done using three parameters: accuracy level, F- measure metric and AUC values. F-measure metric describes the efficiency of the model. 0 is the worst value and 1 is the best value for the metric.

The metrics used in this section include:

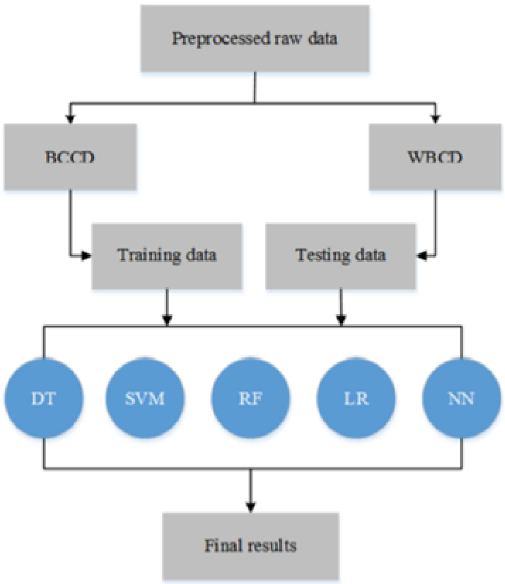
$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{FN} + \text{FP} + \text{TN} + \text{TP}}$$

$$\text{F} - \text{measure metric} = \frac{2 * \text{Precsion} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

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

$$\text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}}$$

The process of the proposed model is as follows:



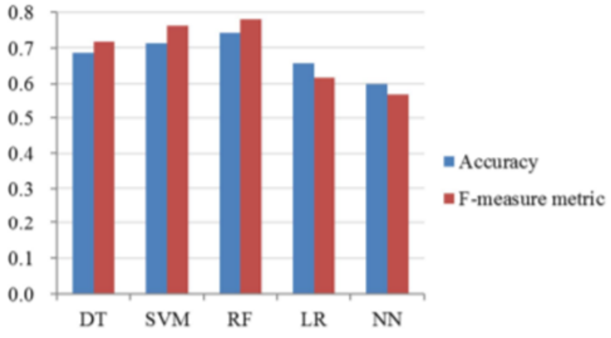


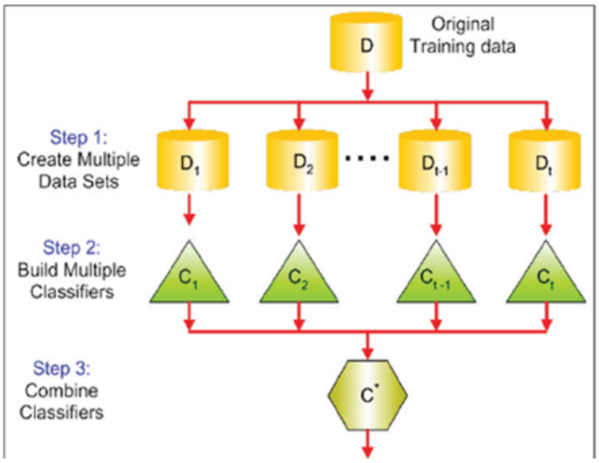
Figure 2. *The accuracy and F-measure metric of five classification models for BCCD data.*

Random forest outperforms other models and is used for clinical purposes and primary analytic purposes.

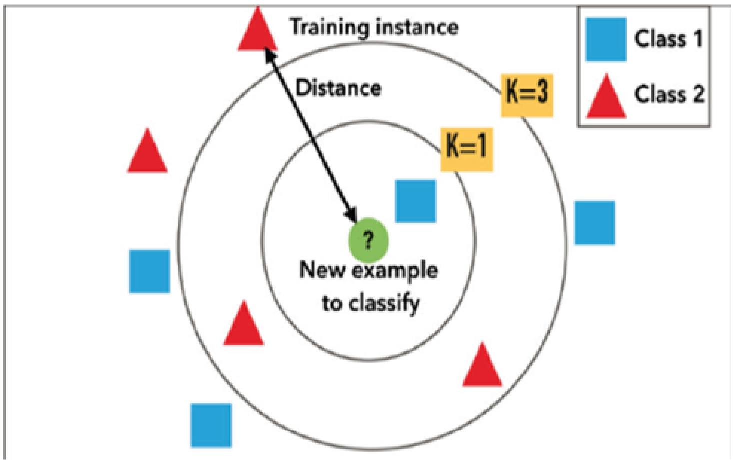
[4] Kaya Keleş, Mümine. (2019). Breast Cancer Prediction and Detection Using Data Mining Classification Algorithms: A Comparative Study. Tehnicki Vjesnik. 26. 149-155. 10.17559/TV-20180417102943.

The aim of the paper was to detect and predict breast cancer early using data mining algorithms. Experimental result was that Bagging, IBk, Random Committee, Random Forest, and SimpleCART algorithms were most successful and had approximately 90% accuracy of prediction. The research uses a dataset from the measurements of an antenna. A data mining tool known as Weka is used in the research because it is known to be used in the medicine field. The classification algorithms use one or more discrete variables, based on the attributes.

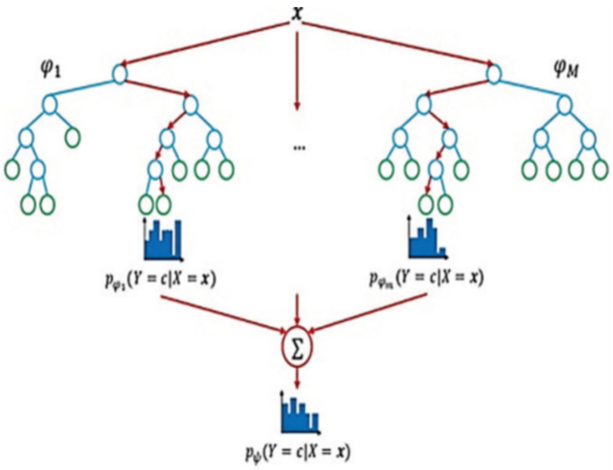
Working principle of Baging Algorithm:


D
Original Training data
Step 1: Create Multiple Data Sets
D_1
D_2
D_{t-1}
D_t
Step 2: Build Multiple Classifiers
C_1
C_2
C_{t-1}
C_t
Step 3: Combine Classifiers
C^*

Working principle of IBk algorithm:



Working principle of Random Forest:



The Weka data mining tool is used to detect breast cancer using data mining classification obtained from attributes of the datasets. The classification algorithms with highest accuracy were used.

[5] Muhammad Hammad Memon, Jian Ping Li, Wang Zhou, "Breast Cancer Detection in the IOT Health Environment Using Modified Recursive Feature Selection", University of Electronic Science and Technology of China, China, 2019

In this research, the authors have utilized the set of IOT devices combined with the help of machine learning techniques to identify and detect breast cancer. The dataset used in this research is the famous Wisconsin Breast Cancer Dataset available on the UCI repository.

In different countries with advanced technology in medical science, the 5-year survival rate of initial phase breast cancer is 75–80% and drops to 25% for breast cancer diagnosis at the initial stage.

The main aim of this research is to propose an IOT-based predictive system based on machine learning to successfully diagnosis people with breast cancer and healthy people. Machine learning predictive model SVM was used for classification of breast cancer in malignant and benign people. The recursive feature selection algorithm (REF) was adopted for the selection of features that improve the classification performance of the SVM classifier.

The authors have adopted the recursive feature selection algorithm for appropriate feature selection in this study because the classification performance of recursive feature selection algorithm FS-based method is good as compared with other methods of classification for breast cancer and healthy people. These works used other feature selection algorithms such as LASSO, MRMR, LLBFS, relief with BFO, relief, and two-stage feature selection method. The training and testing splits validation method has been used in order to select the best hyper-parameters for best model evaluation.

Performance evaluation metrics such as classification accuracy, sensitivity, specificity, F1-score, Matthews's correlation coefficient (MCC), and model execution time were used to check the performance of the proposed system. The proposed system has been tested on breast cancer dataset which is available at the UCI repository.

[6] Zaakia Salod, Yashika Singh, "Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol", University of KwaZulu-Natal, Durban, South Africa, 2019

In this research, the authors propose the use of eight machine learning algorithms:

i. Logistic Regression;
ii. Support Vector Machine;
iii. *K*-Nearest Neighbors;
iv. Decision Tree;
v. Random Forest;
vi. Adaptive Boosting;
vii. Gradient Boosting;
viii. eXtreme Gradient Boosting,

The blood test results using breast cancer Cambria Dataset (BCCD) from the University of California Irvine (UCI) online database to create models for breast cancer prediction. To ensure the models' integrity and robustness, the authors have employed:

i. Stratified *k*-fold Cross- Validation;
ii. Correlation-based Feature Selection (CFS); and
iii. parameter tuning.

The models have been validated on test sets and the validation of breast cancer Cambria Dataset for full features and reduced features. Feature reduction has an impact on algorithm performance. Seven metrics will be used for model evaluation, including accuracy.

Expected impact of the study for public health: The CFS together with highest performing model(s) can serve to identify important specific blood tests that point towards breast cancer, which may serve as an important breast cancer biomarker. Highest performing model(s) may eventually be used to create an Artificial Intelligence tool to assist clinicians in breast cancer screening and detection.

The important features from breast cancer Cambria Dataset found from the modeling process in this study, together with CFS algorithm, could potentially serve to discover cheap and effective BC biomarkers. This will be a subset of blood tests. Highest performing model(s) from this study will serve as basis for future work.

[7] Adel S. Assiri, Saima Nazir, “Breast Tumor Classification Using an Ensemble Machine Learning Method”, Queen Mary University of London, London, 2020

This research proposes a novel ensemble classification method for breast tumor classification using machine learning classification methods. The authors have evaluated the performance of the following classification algorithms: simple logistic Regression learning, SVM learning with stochastic gradient descent optimization and multilayer perceptron network, random decision tree method, random decision forest method, SVM learning with sequential minimal optimization, K-nearest neighbor classifier, and Naïve Bayes classification.

The predictions of all the three best classification algorithms are then used for ensemble classification. For ensemble classification, the authors have used unweighted voting mechanisms including majority-based voting and four minimum probabilities, maximum probabilities, product of probabilities, and the average of probabilities. The performance of the proposed approach is evaluated on the publicly available Wisconsin Breast Cancer Dataset (WBCD).

Majority-based voting is widely used in ensemble classification. It is also known as plurality voting. In the approach proposed here, after applying the three above-mentioned classification algorithms, a majority-based voting mechanism is used to improve the classification results. Each of these model classification results is computed for each test instance and the final output is predicted based on the majority results. In majority voting, the class label *y* is predicted via majority (plurality) voting of each classifier *C*:

$$y = mode\ \{\ C1(x), C2(x), .., Cn(x)\}$$

The performances of eight different classification algorithms are evaluated on the WBCD testing dataset. Multilayer perceptron network takes longer than the other algorithms. For simple logistic regression, the L2 regularization method was used with the conjugate gradient (CG) method for optimization. SGD used optimal learning rate with alpha = 0.0001 and L2 regularization method for SVM. The hidden layer size for MLP was set to 100, with ‘relu’ activation function and stochastic gradient-based optimizer for weight optimization with learning rate set to 0.002.

Random decision tree used Gini impurity to split the tree. The nodes were expanded until all leaves are pure or until all leaves contained less than two. An SMO optimization method normalized the training data and polynomial kernel for classification using SVM.

For KNN, k = 4 nearest neighbors were searched using a Euclidean distance measure, and each neighbor was weighted equally for classification purposes. For Naïve Bayes classification, numeric estimator precision values were chosen based on the analysis of the training data.

[8] Hiba Masood, “Breast Cancer Detection using Machine Learning Algorithm”, 2016

Breast cancer (BC)is 1 of the most common menaces in women. Early diagnosis of Breast Cancer and metastasis between the patients based on a precise system can upsurge survival of the patients to >86%. Breast cancer starts when malignant lumps which are cancerous begin to nurture from the breast cells. Doctors may erroneously detect benign tumor (which is noncancerous) as malignant tumor. The basis of Machine learning is that enables the Systems to study themselves repeatedly and to recover their performance through experience without any directives by programmer. The primary objective is to appraise the performance in cataloguing data with respect to efficiency and effectiveness of each algorithm in terms of cataloguing, test accuracy, precision, and recall. Here the input is a labelled set of training data given to the program to learn. The program has to find the cluster according to the branded input. Supervised learning is named because the programme imparts the learning process about what results should be found from the training data, as a guide to the algorithms. Supervised techniques administered the datasets based on its previously known output. If the output is unremitting, the regression algorithms are painstaking the best choice. Predicting the being of cancer is not the only interest of the researchers. But predicting the enduring possibility take a vital place in the health field

In this , they projected a system that can perceive breast cancer and show machine learning algorithm (ML) can recover the early detection and diagnosis of breast cancer. The support vector machine (SVM) is one of the furthermost influential machines learning (ML) algorithm that is able to model the human empathetic of organizing data. It can find the connection between data and separates them accordingly. They try to recommend the finest (accuracy) results for diagnosis and cataloguing in breast cancer

[9] Mamta Jadhav, Zeel Thakkar, Prof. Pramila M. Chawan, “Breast Cancer Prediction using Supervised Machine Learning Algorithms”, 2017

We found the breast cancer dataset of Wisconsin Breast Cancer analysis dataset and cast-off jupyter notebook as the platform for the persistence of coding. Our procedure comprises use of supervised learning algorithms and classification technique like Decision Tree, Random Forest and Logistic Regression, with Dimensionality Reduction technique.

1) Data processing

2) Categorical data

3) Feature scaling

4) Model selection

5) Supervised learning

In this, diverse types of models are revised and their precisions are figured and associated with each other, so that the best cancer prediction model can be used by doctors in real life to recognize breast cancer moderately earlier than previous methods. Above scrutinized writing study, projected that the Random Forest Classification algorithm is competently exploited and effective for detection of breast cancer as related to Decision tree and Logistic Regression algorithms.

[10] Sweta Bhise, Simran Bepari, Deepmala Kale, Dr. Shailendra Aswale, Aishwarya Singh Gaur, Shrutika Gadekar, "Breast Cancer Detection using Machine Learning Techniques", 2017

According to the Centers for Disease Control and Prevention (CDC)Trusted Source, breast cancer is the most common cancer in women. Breast cancer survival rates vary broadly supported by many factors. 2 of the most important factors are the type of cancer women have and the stage of cancer they obtain a diagnosis. Breast cancer is a cancer that progresses in breast cells. Cancer also can occur within the adipose tissue or the fibrous connective tissue within your breast. The abandoned cancer cells often conquer other healthy breast tissue and may visit the lymph nodes under the arms.

In this, they observed different machine learning techniques for breast cancer detection. They achieved a qualified analysis of CNN, KNN, SVM, Logistic regression, Naïve Bayes and Random Forest. It was detected that CNN outstrips the existing methods when it comes to accuracy, meticulousness and also size of the data set.

[10] Habib Dhahri , Eslam Al Maghayreh,Awais Mahmood,Wail Elkilani, Mohammed Faisal Nagi, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", 2018

Breast cancer is a ubiquitous cause of death, and it is the only type of cancer that is extensive among women worldwide. Many techniques have been developed for early detection and treatment of breast cancer and to reduce the number of deaths and many supported breast cancer diagnosis methods have been used to increase the diagnostic accuracy

In this, the Wisconsin Breast Cancer dataset was obtained from the UCI Machine Learning Repository. This is the same dataset used by Bennett to detect cancerous and noncancerous tumors.

This study tries to solve the problem of automatic detection of breast cancer using a machine learning algorithm. In the present, the algorithm proceeds in different stages.

In the 1st , they proved that the three algorithms can give the same performance after effective configuration. The 2nd experiment focused on that the combining features selection methods improves the accuracy performance. In the 3rd experiment, they reduced how to automatically design the machine learning supervised classifier. Going to the GP algorithm, they attempted to resolve the hyper-parameter problem, which presents a challenge for machine learning algorithms. The proposed algorithm selected the correct algorithm from among the many configurations. All experiments were performed using the Python library. The important results were derived from the proposed method by evaluating an ensemble of approaches, they encountered a significantly higher time consumption rate. In the end, the proposed model is naturally well-matched for control parameter setting of the machine learning algorithms in one side and automated breast cancer diagnosis on the other side.

Modules:

- Module 1:

Dataset creation

- Module 2:

Libraries Used->

-numpy (for numeric calculation)

-pandas (for data analysis)

-matplotlib.pyplot (for data visualisation)

-seaborn (for data visualisation)

- Module 3:

Training and testing the dataset with multiple ML algorithms such as Naïve Bayes' Classifier, Logistic Regression, K Nearest Neighbour Classifier, etc.

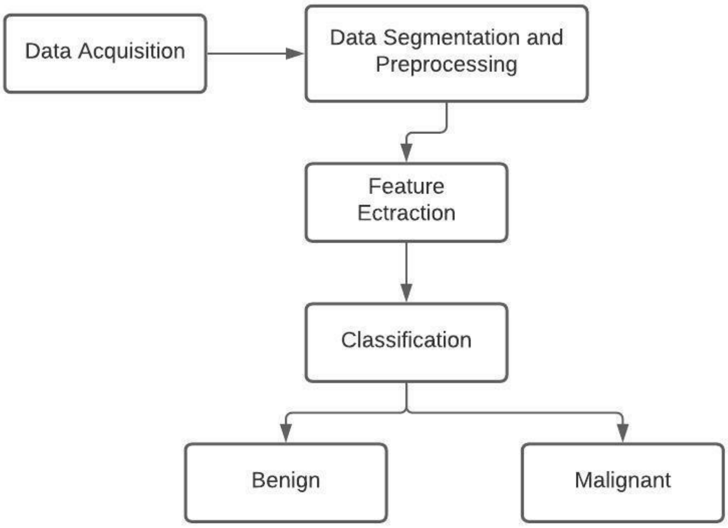
- Module 4:

Classification report of the ML model

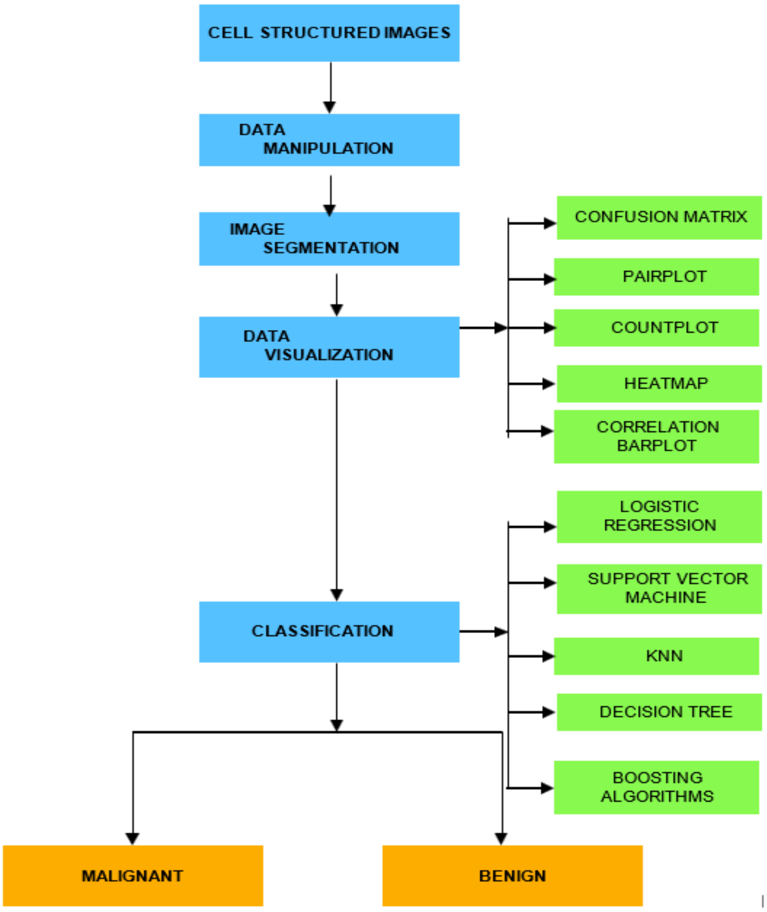
- Module 5:

Validation of the ML model

SYSTEM ARCHITECTURE:



FLOW DIAGRAM:

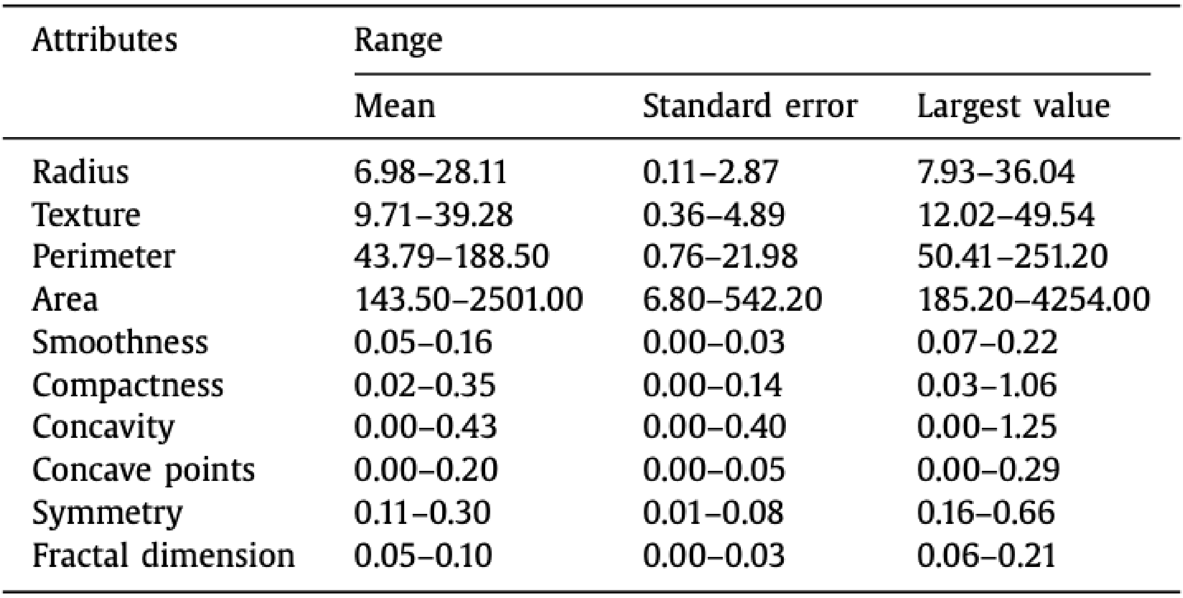


DATASET AND LIST OF ATTRIBUTES

For this assignment, the Wisconsin Breast Cancer Dataset from the University of California, Irvine (UCI) data repository(also available on Kaggle: https://www.kaggle.com/uciml/breastcancer-wisconsin-data/version/2) is used for experimentation and determination of various factors which help to determine the type of cancer(benign or malignant), intensity and probability of spreading of cancerous cells.The WBC dataset includes 699 observations (65.52% benign, 34.48% malignant). Sixteen instances that include missing values for attribute bare nuclei are removed from the dataset during data preprocessing. It contains the following attributes and parameters:-

1. Diagnosis (benign or malignant)
2. radius_mean
3. texture_mean
4. perimeter_mean
5. area_mean
6. smoothness_mean
7. compactness_mean
8. concavity_mean
9. concave points_mean
10. symmetry_mean
11. fractal_dimension_mean
12. radius_se (se -> standard error)
13. texture_se
14. perimeter_se
15. area_se
16. smoothness_se
17. compactness_se
18. concavity_se
19. concave points_se
20. symmetry_se
21. fractal_dimension_se
22. radius_worst (worst calculated variation of the attribute)
23. texture_worst
24. perimeter_worst
25. area_worst
26. smoothness_worst
27. compactness_worst
28. concavity_worst
29. concave points_worst
30. symmetry_worst

31. fractal_dimension_worst



Attributes	Range		
	Mean	Standard error	Largest value
Radius	6.98–28.11	0.11–2.87	7.93–36.04
Texture	9.71–39.28	0.36–4.89	12.02–49.54
Perimeter	43.79–188.50	0.76–21.98	50.41–251.20
Area	143.50–2501.00	6.80–542.20	185.20–4254.00
Smoothness	0.05–0.16	0.00–0.03	0.07–0.22
Compactness	0.02–0.35	0.00–0.14	0.03–1.06
Concavity	0.00–0.43	0.00–0.40	0.00–1.25
Concave points	0.00–0.20	0.00–0.05	0.00–0.29
Symmetry	0.11–0.30	0.01–0.08	0.16–0.66
Fractal dimension	0.05–0.10	0.00–0.03	0.06–0.21

Performance metric evaluation methods for assessment of model performances

A Confusion Matrix (CM) is a table showing actual *versus* predicted labels of a ML model for the various classes in a dataset. Since our dataset contains two classes with the classes ‘breast cancer tumor present’ and ‘breast cancer tumor absent’, the table forms a two-by-two dimension.

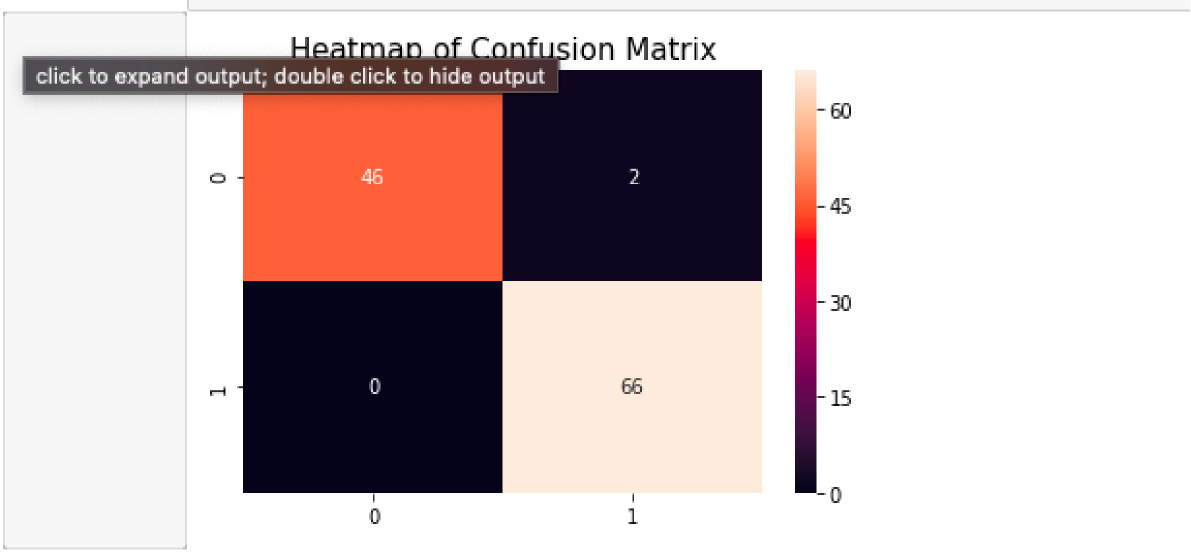
Using CM, we compute the following:

- True Positive (TP) in row one, column one: The ‘breast cancer tumor present’ class is correctly classified as having breast cancer.

- False Negative (FN) in row one, column two: The actual ‘breast cancer tumor present’ group is incorrectly classified as not having breast cancer.

- False Positive (FP) in row two, column one: The ‘breast cancer tumor absent’ group is incorrectly classified as having breast cancer.

- True Negative (TN) in row two, column two: The ‘breast cancer tumor absent’ group is correctly classified as not having breast cancer.



Furthermore, the following have been considered for valuation of models:

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

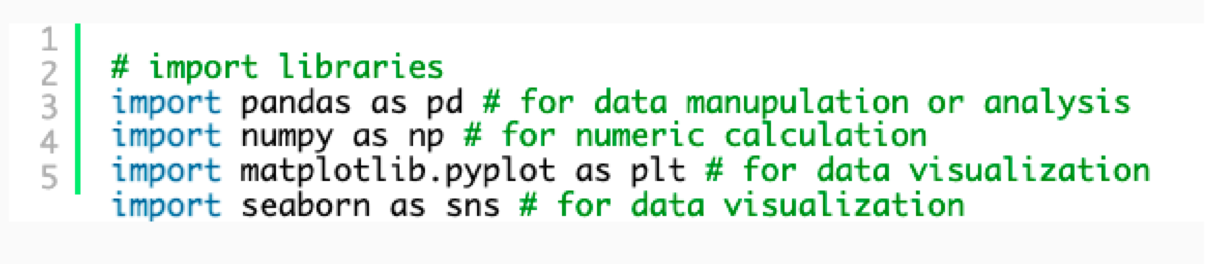
$$Precision = \frac{TP}{TP+FP}$$

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

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

All the test in this experimental analysis have conducted using Jupiter notebook through the following pylon libraries:

1. Scikit-Learn (sklearn)
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn



```
# import libraries
import pandas as pd # for data manupulation or analysis
import numpy as np # for numeric calculation
import matplotlib.pyplot as plt # for data visualization
import seaborn as sns # for data visualization
```

EXPERIMENTATION AND ANALYSIS OF MACHINE LEARNING ALGORITHMS

1. Logistic regression

$$h_\theta(x) = \frac{1}{1+exp(-\theta^T x)}$$

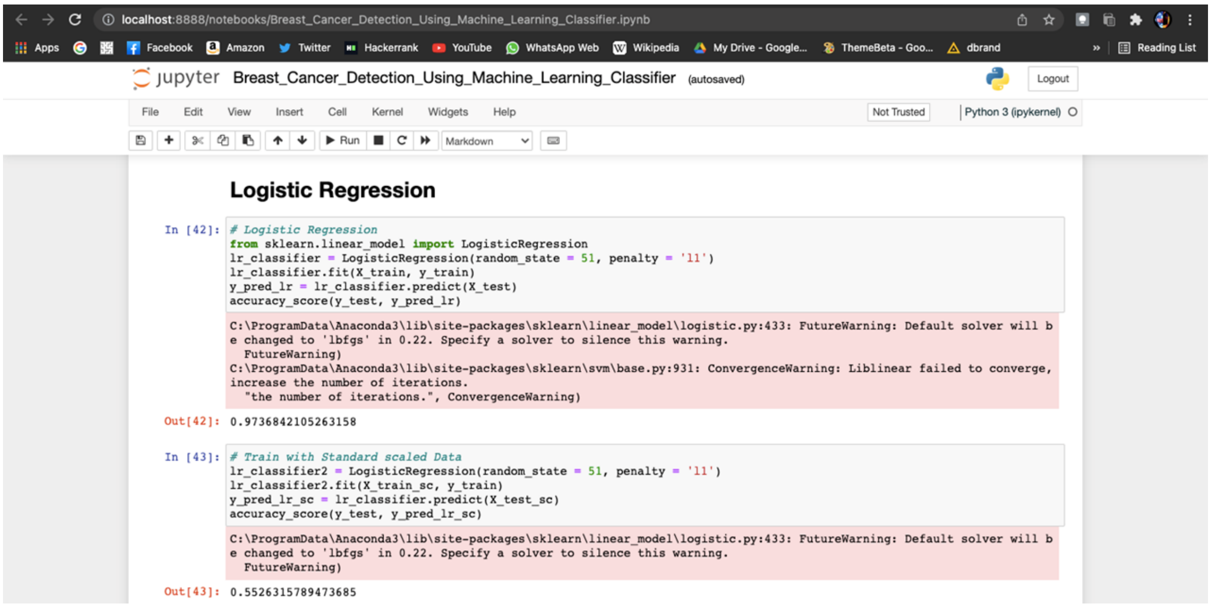
The Logistic regression algorithm is a binary classification algorithm. A prediction (symbolized by ŷ) is made by utilizing the logistical function, which is given by the equation:

The Logistic regression algorithm makes decisions which are primarily based on the probability of what is identified when the logistical function is optimally provided with a specific set of instructions and features. In general, the final output would be two or more classes. In this particular case, this function would output either of the following two classes:

i) Firstly , where breast cancer tumor is present (classification equals ‘2’); or
ii) Secondly breast cancer where the tumor is absent (classification equals ‘1’).

This algorithm is chosen because it is generally the first algorithm attempted for machine learning problems and is quite popularly known to produce good results for particularly binary classification cases.

Library used for Logistic Regression in python for experiment: LogisticRegression from sklearn



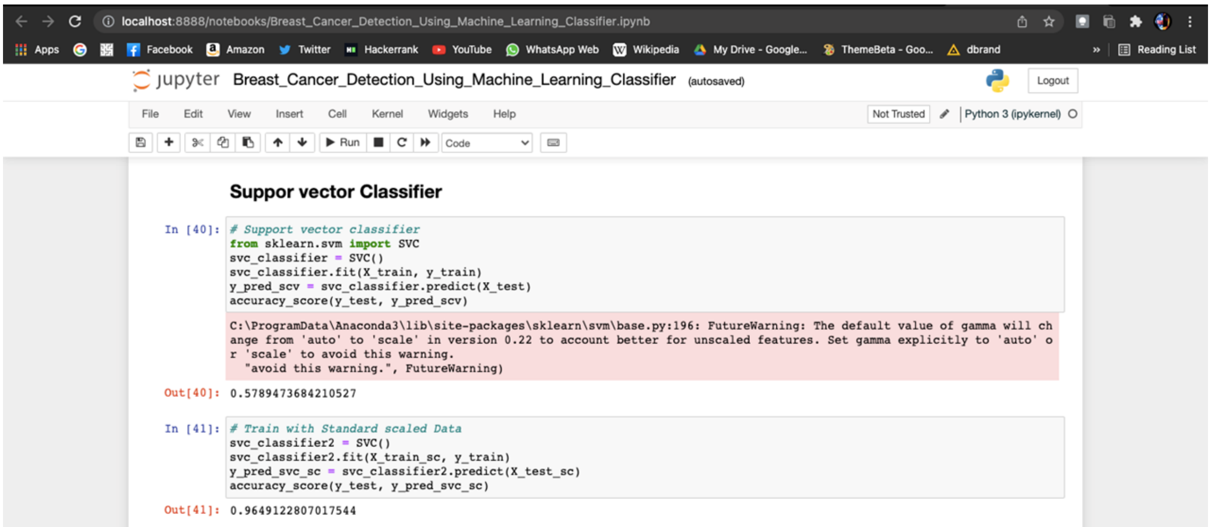
Logistic Regression

```
In [42]: # Logistic Regression
         from sklearn.linear_model import LogisticRegression
         lr_classifier = LogisticRegression(random_state = 51, penalty = 'l1')
         lr_classifier.fit(X_train, y_train)
         y_pred_lr = lr_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_lr)

         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:433: FutureWarning: Default solver will b
         e changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
           FutureWarning)
         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\svm\base.py:931: ConvergenceWarning: Liblinear failed to converge,
         increase the number of iterations.
           "the number of iterations.", ConvergenceWarning)

Out[42]: 0.9736842105263158

In [43]: # Train with Standard scaled Data
         lr_classifier2 = LogisticRegression(random_state = 51, penalty = 'l1')
         lr_classifier2.fit(X_train_sc, y_train)
         y_pred_lr_sc = lr_classifier.predict(X_test_sc)
         accuracy_score(y_test, y_pred_lr_sc)

         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:433: FutureWarning: Default solver will b
         e changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.
           FutureWarning)

Out[43]: 0.5526315789473685
```

2. Support vector machine

The Support vector machine algorithm is a classifier algorithm that takes the raw data as input and outputs the most optimal line of the same, also known as the decision boundary, in the middle which further best separates different groups of data it finally, finds on a graph. This decision boundary is then drawn right in the middle of the peripheral metadata points it finds and identifies, referred to as the support vectors of the groups. In its simplest form, the Support vector machine's process includes separation of data into two groups into a two-dimensional graphical representation. The goal of Support vector machine is to obtain the decision boundaries to classify data into the classes 'breast cancer tumor present' and 'breast cancer tumor absent' groups. This algorithm has been chosen for this assignment because it is known to perform well on medical data for classification into two groups, and it is a very popular algorithm amongst machine learning projects involving healthcare applications.

Library used for Support vector machine algorithm in python for experiment: SVC from sklearn



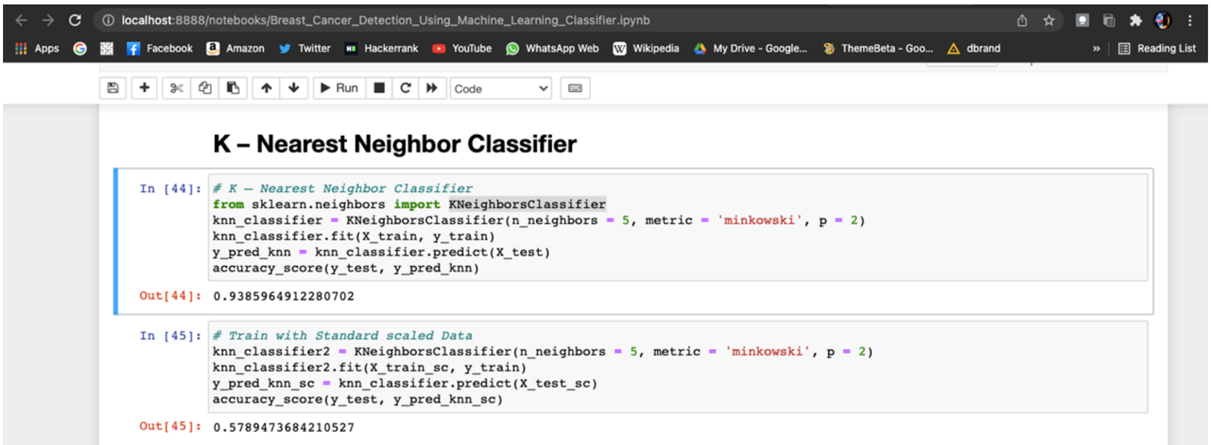
Suppor vector Classifier

```
In [40]: # Support vector classifier
         from sklearn.svm import SVC
         svc_classifier = SVC()
         svc_classifier.fit(X_train, y_train)
         y_pred_scv = svc_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_scv)

         C:\ProgramData\Anaconda3\lib\site-packages\sklearn\svm\base.py:196: FutureWarning: The default value of gamma will ch
         ange from 'auto' to 'scale' in version 0.22 to account better for unscaled features. Set gamma explicitly to 'auto' o
         r 'scale' to avoid this warning.
           "avoid this warning.", FutureWarning)

Out[40]: 0.5789473684210527

In [41]: # Train with Standard scaled Data
         svc_classifier2 = SVC()
         svc_classifier2.fit(X_train_sc, y_train)
         y_pred_svc_sc = svc_classifier2.predict(X_test_sc)
         accuracy_score(y_test, y_pred_svc_sc)

Out[41]: 0.9649122807017544
```

3. K-nearest neighbors

The *K-nearest neighbors* algorithm is a simple technique which is commonly used to classify un-labelled meta-data points, which are based on the classification of neighboring labelled data points, shown on a graphical representation. The neighboring data points are be those which are closest in distance to the current data point. The total number of the closest data points to the reference point is determined by the value of a parameter which is referred to as *k*. This entire process is essentially a voting method . The value of *k* is arbitrarily chosen, but the process of choosing the optimal value for *k* is quite important in order to ensure that a suitable number of neighbors are utilized during the process of voting so that the errors can cancel out each other and the algorithm identifies the adequate and correct patterns during the execution of this process. In its simplest and truest form, *K-nearest neighbors algorithm* works on the data in a two-dimensional graphical representation . In this particular experiment, *K-nearest neighbors* is used to classify breast cancer tumor present class and ‘breast cancer tumor absent class.

Library used for *K-nearest neighbors* algorithm in python for experiment: KNeighborsClassifier from sklearn



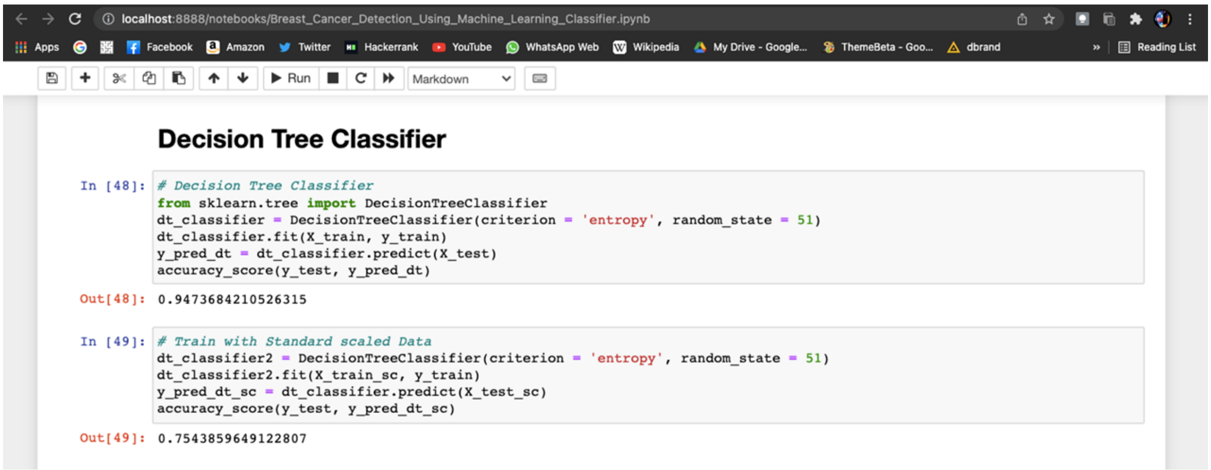
K – Nearest Neighbor Classifier

```
In [44]: # K – Nearest Neighbor Classifier
         from sklearn.neighbors import KNeighborsClassifier
         knn_classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
         knn_classifier.fit(X_train, y_train)
         y_pred_knn = knn_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_knn)

Out[44]: 0.9385964912280702

In [45]: # Train with Standard scaled Data
         knn_classifier2 = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
         knn_classifier2.fit(X_train_sc, y_train)
         y_pred_knn_sc = knn_classifier.predict(X_test_sc)
         accuracy_score(y_test, y_pred_knn_sc)

Out[45]: 0.5789473684210527
```

4. Decision tree

The Decision tree classifier algorithm is known to be highly efficient and provides easy interpretation due to its rule-based flowchart-like structure. The Decision Tree classifier algorithm belongs to the group of supervised learning algorithms. However, unlike the rest of the supervised learning algorithms, the decision tree algorithm can be utilized for solving classification and regression problems too. This algorithm starts with a question at the top, which is also termed a 'root node'. In a standard Decision tree, the answers for this underlying question have two options. The algorithm traverses the branches of the tree depending on responses to questions asked previously, until finally, the destination of the leaf node is reached, indicating whether the person under examination falls under the classification of breast cancer tumor present class or breast cancer tumor absent class.

Library used for Decision Tree algorithm in python for experiment: DecisionTreeClassifier from sklearn



Decision Tree Classifier

```
In [48]: # Decision Tree Classifier
         from sklearn.tree import DecisionTreeClassifier
         dt_classifier = DecisionTreeClassifier(criterion = 'entropy', random_state = 51)
         dt_classifier.fit(X_train, y_train)
         y_pred_dt = dt_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_dt)

Out[48]: 0.9473684210526315

In [49]: # Train with Standard scaled Data
         dt_classifier2 = DecisionTreeClassifier(criterion = 'entropy', random_state = 51)
         dt_classifier2.fit(X_train_sc, y_train)
         y_pred_dt_sc = dt_classifier.predict(X_test_sc)
         accuracy_score(y_test, y_pred_dt_sc)

Out[49]: 0.7543859649122807
```

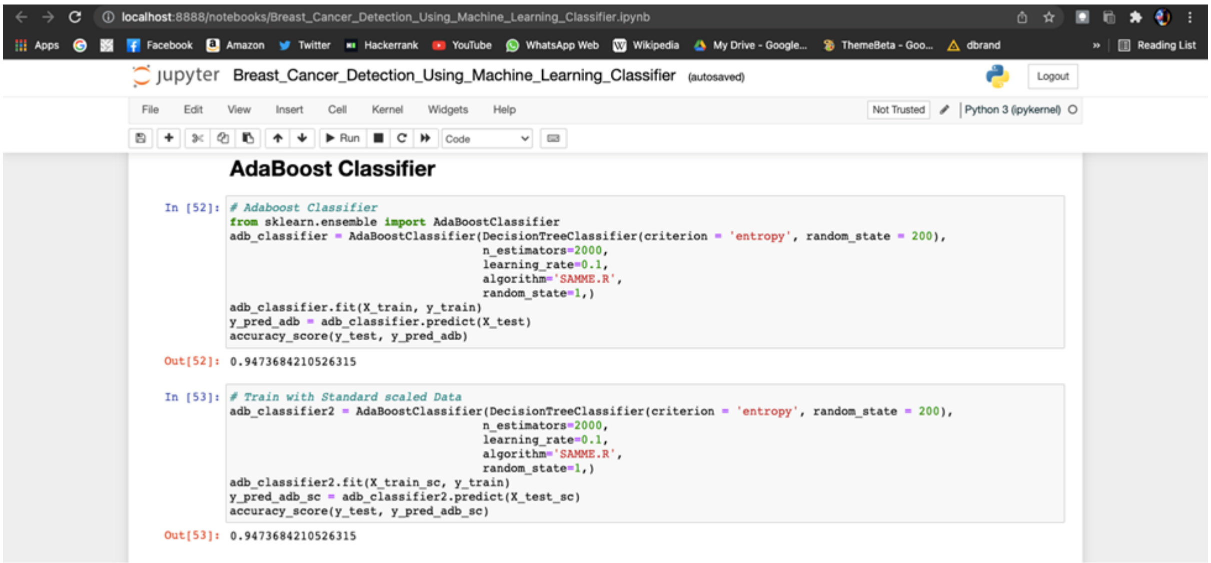
5. Boosting Algorithms

Boosting algorithms is a category of ensemble algorithms which increases the raw performance of multiple weak algorithms by adjusting the weights of observations from earlier classifications. If an observation is misclassified, boosting attempts to increase the weight of this observation and *vice versa*. For example, AdaBoost can be used to boost the performance of any machine learning algorithm. It is best used with weak learners. These are models that achieve accuracy just above random chance on a classification problem.

The AdaBoost was created by Freund and Schapire: GBM was invented by Friedman; and XGBoost was initiated by Chen and Guestrin. These algorithm is chosen because they are known to be powerful and perform well. Furthermore, XGBoost is known to be a state-of-the-art algorithm and it is also scalable.

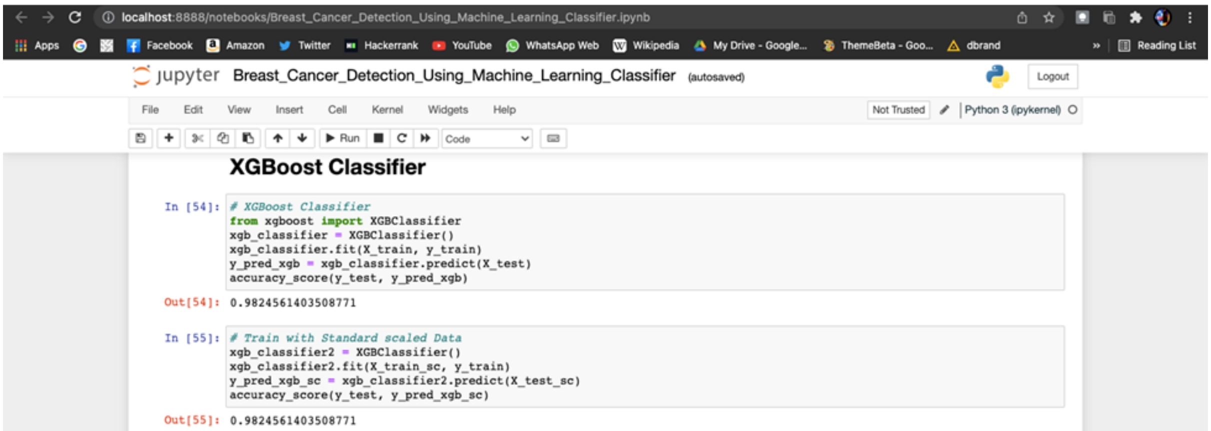
Library used for AdaBoost in python for experiment: AdaBoostClassifier from sklearn

Library used for XGBoost in python for experiment: X


localhost:8888/notebooks/Breast_Cancer_Detection_Using_Machine_Learning_Classifier.ipynb
jupyter Breast_Cancer_Detection_Using_Machine_Learning_Classifier (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

AdaBoost Classifier

```
In [52]: # Adaboost Classifier
         from sklearn.ensemble import AdaBoostClassifier
         adb_classifier = AdaBoostClassifier(DecisionTreeClassifier(criterion = 'entropy', random_state = 200),
                                             n_estimators=2000,
                                             learning_rate=0.1,
                                             algorithm='SAMME.R',
                                             random_state=1,)
         adb_classifier.fit(X_train, y_train)
         y_pred_adb = adb_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_adb)

Out[52]: 0.9473684210526315

In [53]: # Train with Standard scaled Data
         adb_classifier2 = AdaBoostClassifier(DecisionTreeClassifier(criterion = 'entropy', random_state = 200),
                                             n_estimators=2000,
                                             learning_rate=0.1,
                                             algorithm='SAMME.R',
                                             random_state=1,)
         adb_classifier2.fit(X_train_sc, y_train)
         y_pred_adb_sc = adb_classifier2.predict(X_test_sc)
         accuracy_score(y_test, y_pred_adb_sc)

Out[53]: 0.9473684210526315
```

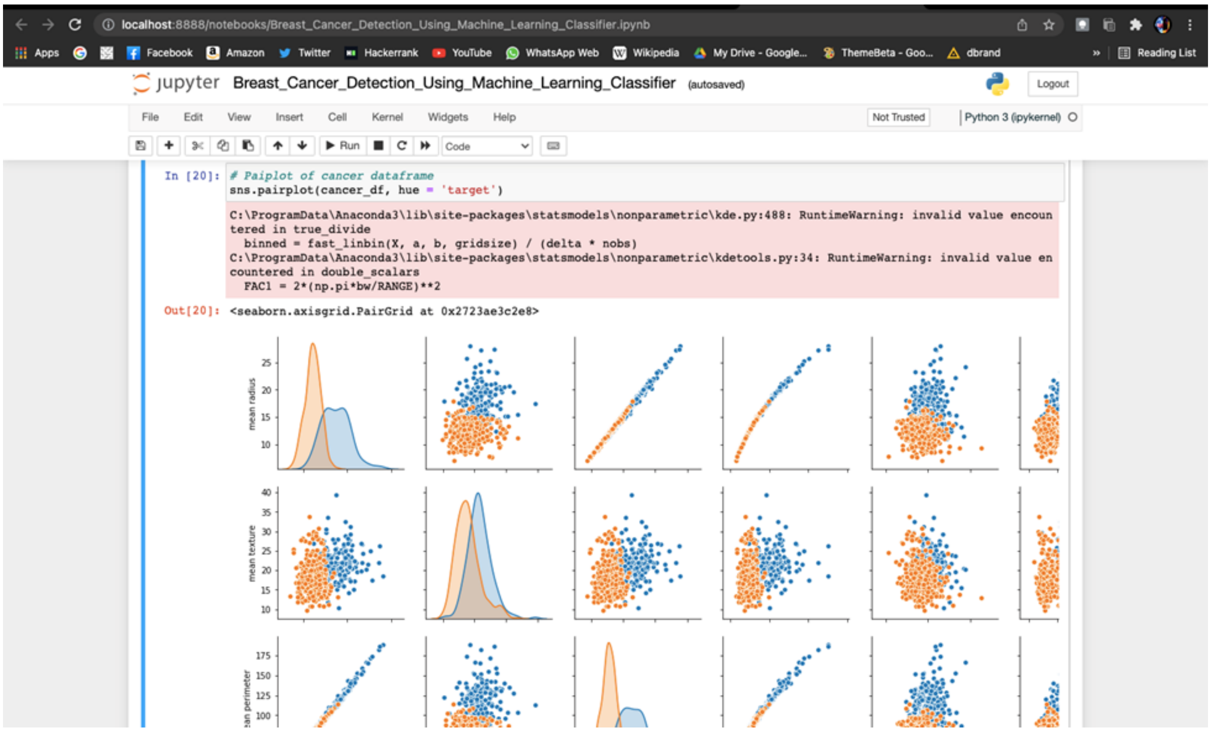

localhost:8888/notebooks/Breast_Cancer_Detection_Using_Machine_Learning_Classifier.ipynb
jupyter Breast_Cancer_Detection_Using_Machine_Learning_Classifier (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code

XGBoost Classifier

```
In [54]: # XGBoost Classifier
         from xgboost import XGBClassifier
         xgb_classifier = XGBClassifier()
         xgb_classifier.fit(X_train, y_train)
         y_pred_xgb = xgb_classifier.predict(X_test)
         accuracy_score(y_test, y_pred_xgb)

Out[54]: 0.9824561403508771

In [55]: # Train with Standard scaled Data
         xgb_classifier2 = XGBClassifier()
         xgb_classifier2.fit(X_train_sc, y_train)
         y_pred_xgb_sc = xgb_classifier2.predict(X_test_sc)
         accuracy_score(y_test, y_pred_xgb_sc)

Out[55]: 0.9824561403508771
```

GBClassifier from xgboost

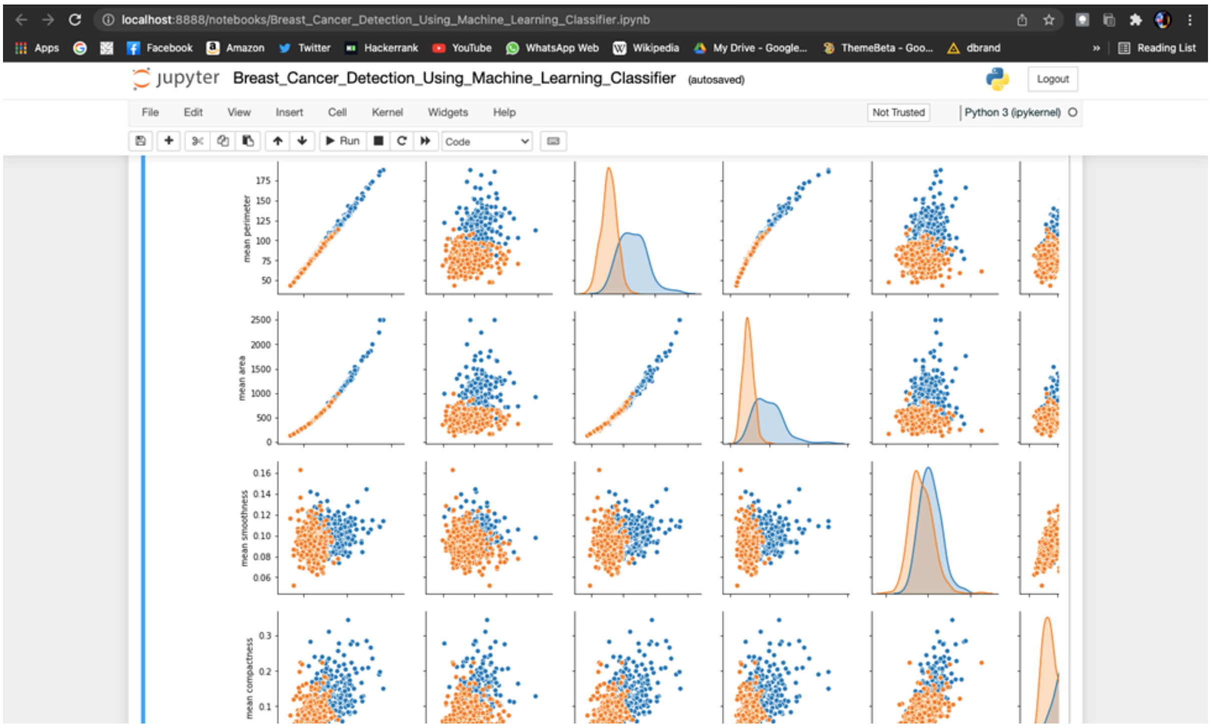
DATA VISUALIZATION



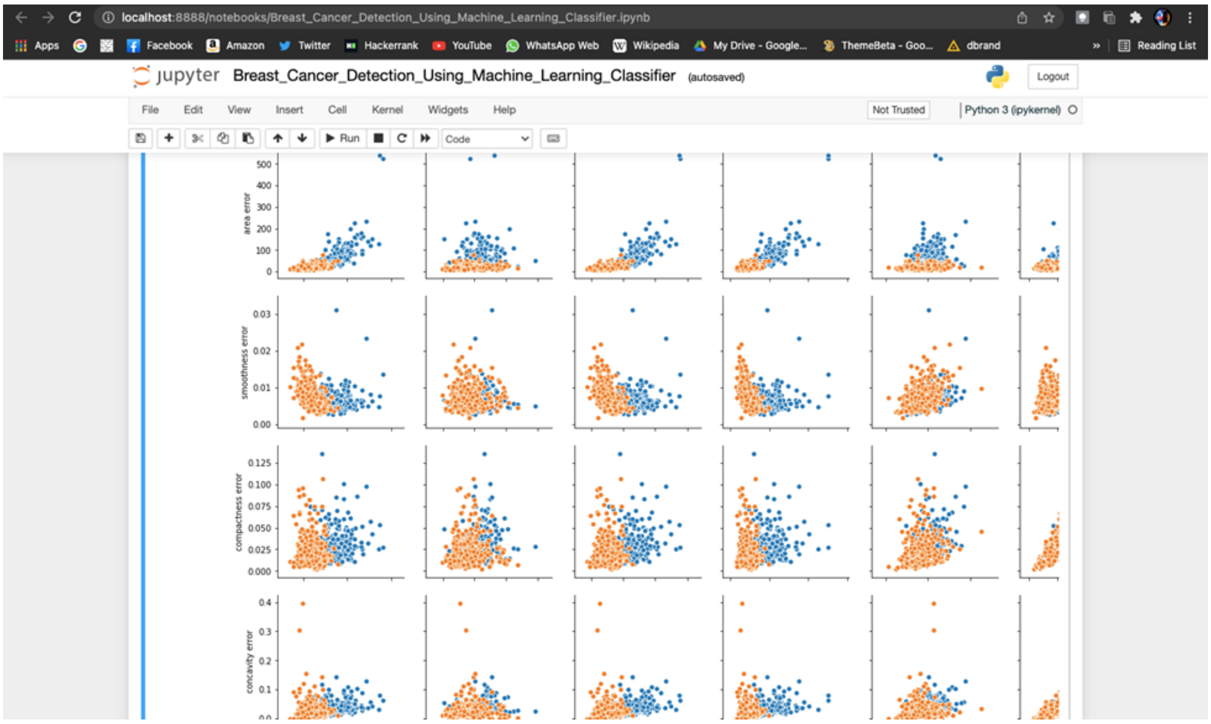
localhost:8888/notebooks/Breast_Cancer_Detection_Using_Machine_Learning_Classifier.ipynb
jupyter Breast_Cancer_Detection_Using_Machine_Learning_Classifier (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
mean radius
25
20
15
10
mean texture
40
35
30
25
20
15
10
ean perimeter
175
150
125
100

```
In [20]: # Paiplot of cancer dataframe
         sns.pairplot(cancer_df, hue = 'target')

C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\nonparametric\kde.py:488: RuntimeWarning: invalid value encoun
tered in true_divide
  binned = fast_linbin(X, a, b, gridsize) / (delta * nobs)
C:\ProgramData\Anaconda3\lib\site-packages\statsmodels\nonparametric\kdetools.py:34: RuntimeWarning: invalid value en
countered in double_scalars
  FAC1 = 2*(np.pi*bw/RANGE)**2

Out[20]: <seaborn.axisgrid.PairGrid at 0x2723ae3c2e8>
```

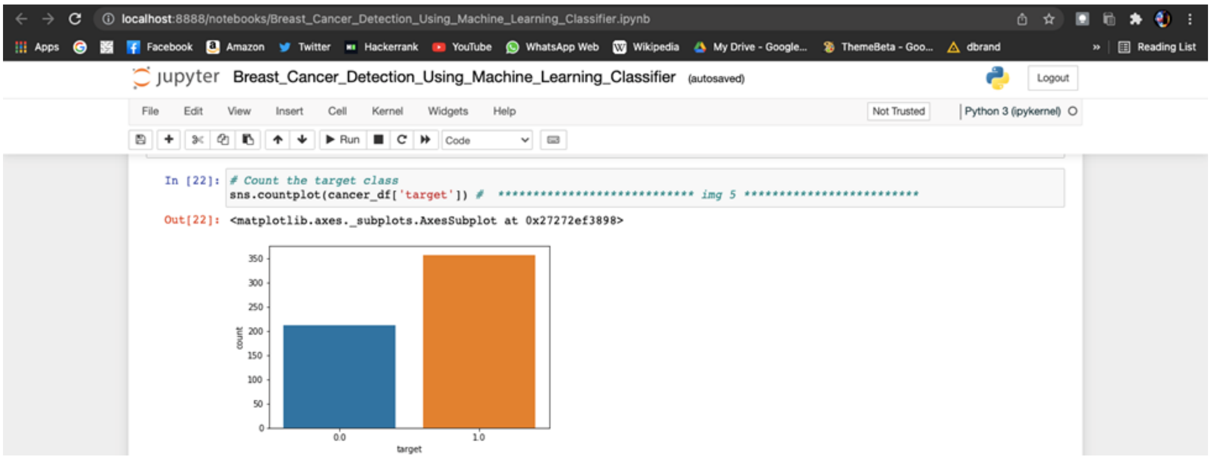
1. Pairplot (To plot multiple pairwise bivariate distributions in the dataset)



localhost:8888/notebooks/Breast_Cancer_Detection_Using_Machine_Learning_Classifier.ipynb
jupyter Breast_Cancer_Detection_Using_Machine_Learning_Classifier (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
mean perimeter
175
150
125
100
75
50
mean area
2500
2000
1500
1000
500
0
mean smoothness
0.16
0.14
0.12
0.10
0.08
0.06
mean compactness
0.3
0.2
0.1



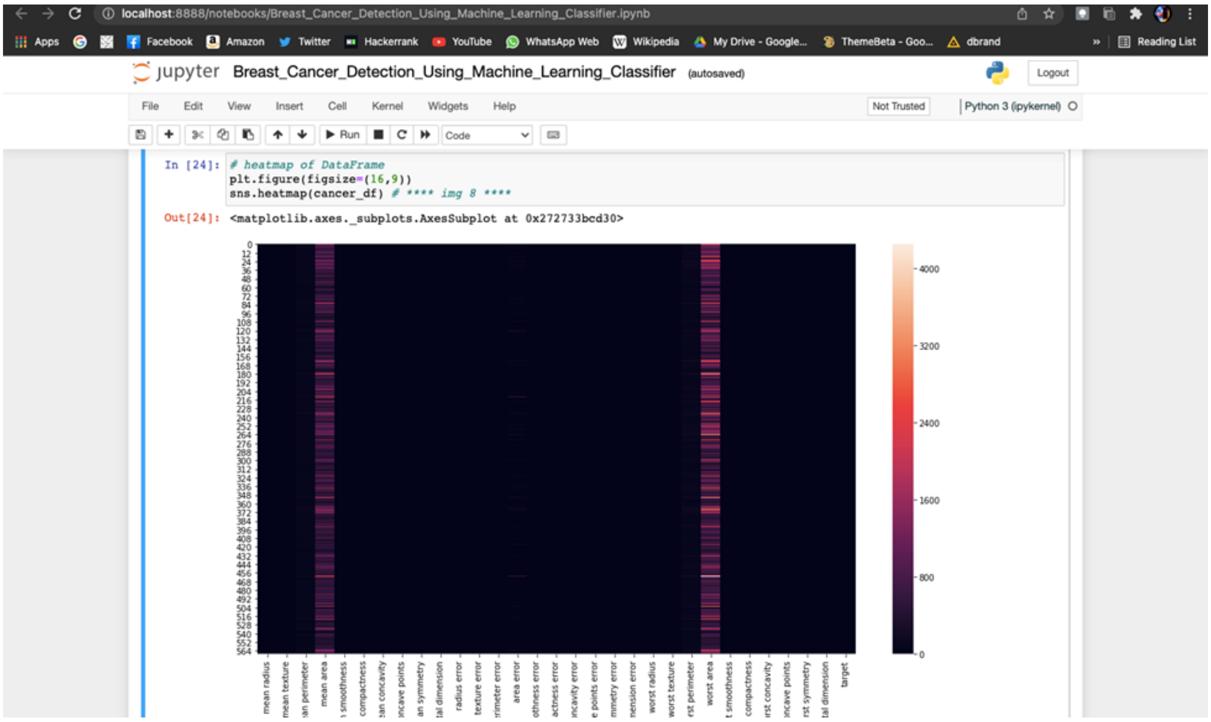
2. Countplot (to Show the counts of observations in each categorical bin using bars)



```
In [22]: # Count the target class
         sns.countplot(cancer_df['target']) #  **************************** img 5 *************************
```

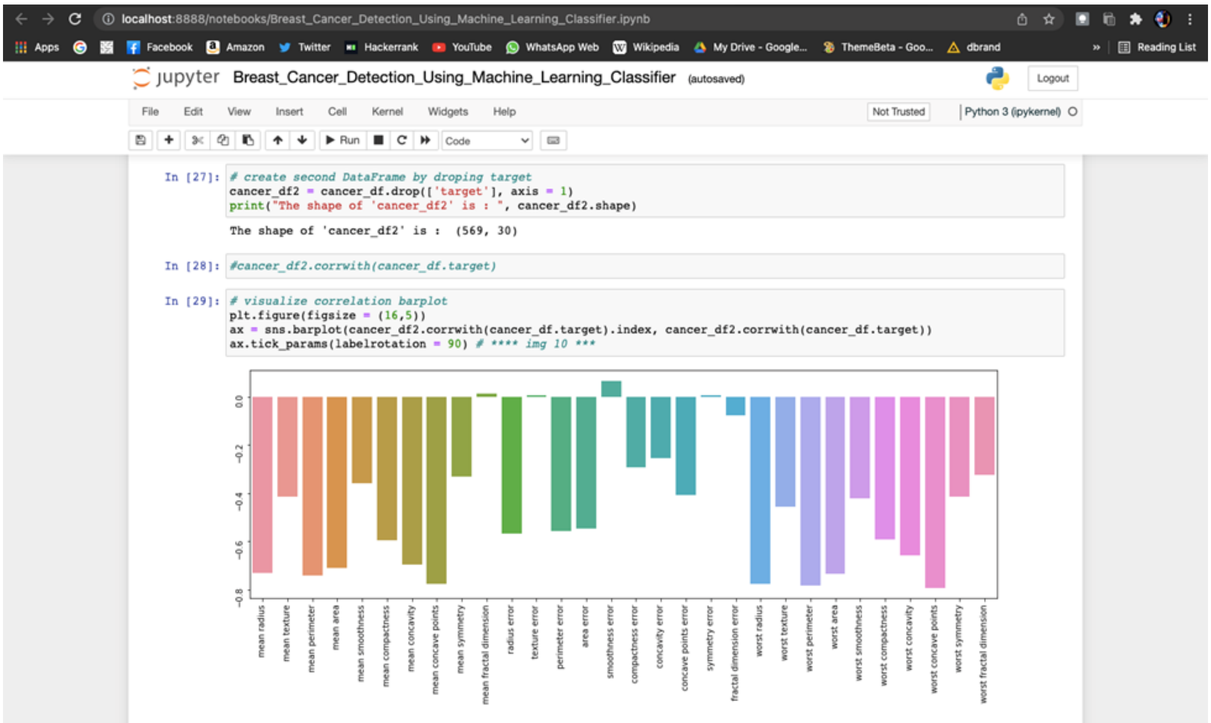
```
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0x27272ef3898>
```

3. Heatmap (to Show the care where intensity of tumor is high)



```
In [24]: # heatmap of DataFrame
         plt.figure(figsize=(16,9))
         sns.heatmap(cancer_df) # **** img 8 ****

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0x272733bcd30>
```

4. correlation barplot (to visualize pairwise correlation)



```
In [27]: # create second DataFrame by droping target
         cancer_df2 = cancer_df.drop(['target'], axis = 1)
         print("The shape of 'cancer_df2' is : ", cancer_df2.shape)

         The shape of 'cancer_df2' is :  (569, 30)

In [28]: #cancer_df2.corrwith(cancer_df.target)

In [29]: # visualize correlation barplot
         plt.figure(figsize = (16,5))
         ax = sns.barplot(cancer_df2.corrwith(cancer_df.target).index, cancer_df2.corrwith(cancer_df.target))
         ax.tick_params(labelrotation = 90) # **** img 10 ***
```

CONCLUSION

In this experiment, I have compared 5 ML techniques namely Logistic Regression, Support Vector Machines, K- nearest neighbors, XGBoost and ADAboost. All the algorithms had accuracy greater than 50%. Support Vector Machine leads the results and yields the highest accuracy of 96.49%.

FUTURE WORK

In the future, there are several aspects that can be extended based on this research. The SVM ensemble model structures can be used for other breast cancer datasets; relevant feature selection and extraction techniques can be applied to the model feature preparation process. Then, the SVM ensemble learning can also be used for other disease diagnoses, such as thyroid cancer, oral cancer, and diabetes. Also, in terms of computation time, parallel computation techniques can be helpful to accelerate the training process for the proposed ensemble algorithm model.