**PREDICTING BREAST CANCER WITH MACHINE LEARNING**

BY

NYIAM KIELLY KEVIN

VUG|CSC|18|2659

A PROJECT SUBMITTED TO

THE DEPARTMENT OF COMPUTER AND INFORMATION TECHNOLOGY

FACULTY OF NATURAL AND APPLIED SCIENCES, VERITAS

UNIVERSITY, ABUJA.

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE BACHELOR OF SCIENCES DEGREE (BSc) IN COMPUTER SCIENCE.

25th July 2022

**CERTIFICATION**

This is to certify that this project entitled “Predicting Breast Cancer using Machine Learning” was carried out by “Nyiam Kielly Kevin” with Matriculation number “VUG|CSC|18|2659” in the faculty of Natural and Applied Sciences, Veritas University, Abuja for award of Bachelor of Science in Computer Science.

Mr. JohnWendy Nwaukwa \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Project Supervisor) Sign and Date

Dr. Joe Essien \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Head of Department) Sign and Date

Prof. Simon Koma Okwute \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Dean of Faculty) Sign and Date

NAME \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(External Examiner) Sign and Date

**DEDICATION**

To the All-Powerful GOD for granting me the privilege to complete this programme.

**ACKNOWLEDGMENT**

I am sincerely grateful to the lord for seeing me through, from the start to the end of this research work. Even when I had no strength and the zeal to continue, the Lord was continually my strength.

I sincerely appreciate my hardworking, wonderful lecturer and friend, Mr. Felix Uloko, for his consistent words of encouragement and support to me throughout the period of writing the research. I am forever grateful to him. I want to thank the Head of Department, Dr. Joe Essien, for his words of advice that has helped to make me whom I am today. I specially appreciate my lecturers; Mr. Rapheal Enihe, Engr. Callistus C. Chimezie, Dr. Okochi Ajonye, Mr. Johnwendy Nwaukwa and Mr. Victor Omopariola, your impacts in my life have been massive. May the Lord reward you all richly.

A huge gratitude to my loved ones; Mrs Ann-Marie Terfa Tarhembah, Mr. Terfa Tarhembah, Mr. Peter Sebastian Gwimi, Mrs. Abigail Sebastian Gwimi, Miss Eney Benita Nyiam, Mr. Franklin Sunny Nyiam, Mrs. Atedze Matilda Nguhemen, your continuous support and help were sufficient encouragement for me to achieve this dream.

To my course mates, Orobosa Igho-Osagie, Abiola-Ajayi Damilola, Yakubu Micheal Mamman, Godwin Fidel Emmanuel, Francis Oche Divine, Kinse Annastasia Shako, your love to me during this programme was priceless.

To my friends, Napoli Daniel, Kingdavid Kelechi, Kenneth Olajide, Ekanabasi Isok Clement, Luke Ocheme, Daniel Tarhembe, Ebuka Igbokidi, Neska Sebastian Gwimi, I am grateful for all your supports.

**ABSTRACT**

Breast Cancer is one of the most common cancers among women worldwide, representing the majority of new cancer cases and cancer related deaths according to global statistics, making it a significant public health issue in today’s society. The early diagnosis of Breast Cancer (BC) can improve the prognosis and the chance of survival significantly, as it can promote timely clinical treatments to patients. This project aims to predict the type of Breast Cancer (Malignant or Benign classes) using K-Nearest Neighbor (KNN), and Logistic Regression (LR) two Machine Learning Algorithms by taking cell parameters as input. This project also aims to achieve maximum accuracy in prediction of Breast Cancer using BC datasets. This study is conducted on Wisconsin Breast Cancer Datasets (WBCD) obtained by the University of Wisconsin Hospital from UCL repository.

**TABLE OF CONTENTS**

**TITLE PAGE i**

**CERTIFICATION ii**

**DEDICATION iii**

**ACKNOWLEDGMENT iv**

**ABSTRACT v**

**LIST OF FIGURES vi**

**LIST OF TABLES vii**

**LIST OF APPENDICES viii**

**CHAPTER 1: INTRODUCTION 10**

* 1. Problem Statement 11
  2. Research Questions 11
  3. Research Aims and Objectives 12
  4. Research Motivation 12
  5. Significance of the research 14
  6. Operational Definition of Terms 14
  7. Organization of the Project 15
  8. Chapter Conclusion 15

**CHAPTER 2: LITERATURE REVIEW 16**

2.1 Related Works 17

2.2 Machine learning 20

**CHAPTER 3: PROPOSED MODELS 22**

* 1. Dataset 23
  2. Data Visualization 24
     1. Histogram 24
     2. Heatmap 25
  3. Data Preprocessing 27
     1. Categorical Variable Conversion 27
     2. Feature Selection 27
     3. Train-test split 27
  4. Algorithms 29
     1. K-Nearest Neighbor (KNN) 29
     2. Logistic Regression (LR) 29

**CHAPTER 4: RESULTS ANALYSIS 31**

4.1 Performance Metrics 31

4.1.1 Confusion Matrix 31

4.1.2 Accuracy 34

4.1.3 Precision 34

4.1.4 Recall 34

4.1.5 F1 Score 35

4.2 Model Performance 35

4.2.1 Logistic Regression (LR) 36

4.2.2 K-Nearest Neighbor (KNN) 37

**CHAPTER 5: CONCLUSION 39**

**REFERENCES 40**

**APPENDIX 45**

**LIST OF FIGURES**

3.1 Workflow of the Proposed System 11

3.2 Class Distribution 14

3.4 Correlation between all variables 16

4.1 Logistic Regression Confusion matrix 26

4.2 K-Nearest Neighbor Confusion matrix 27

**LIST OF TABLES**

4.1 Confusion Matrix 21

4.2 Scores of the two Models 24

**CHAPTER 1**

**INTRODUCTION**

Breast cancer is the most common cancer among women, but it can occur in both genders. 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 has 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 (American Cancer Society, 2003). Numerical features, such as radius, texture, perimeter, and area, can be measured from microscopic images. Data, later, 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 (American Cancer Society, 2003). In all, early prediction remains 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 two classification models; K-Nearest Neighbors (KNN), Logistic Regression (LR) have been run on the Wisconsin Breast Cancer (original) datasets. The results obtained are then measured using various performance metrics to compare among the algorithms to find out the best suited model for cancer prediction (American Cancer Society, 2003).

* 1. **Problem Statement**

With rising rate of breast cancer amongst both genders, the means/method in use to predict the possibility of a person developing cancer is lacking. There is a need to assist the medical field to aid in reducing the cancer rise.

* 1. **Research Questions**
* Can Machine learning be a good addition to the medical field?
* Does early breast cancer prediction reduce the rate of fatalities?
* Prediction using machine learning will mean patients no longer need to undergo certain tests relating to finding out about their cancer status, will that potentially be a good resolve?
* Does Knowing the class of cancer that a patient has; help reduce the disease spread in the part of the patient affected?

**1.3 Research Aims and Objectives**

* The aim of this project is to predict the type of Breast Cancer (Malignant or Benign classes) with K-Nearest Neighbor (KNN), and Logistic Regression (LR), using two Machine Learning Algorithms by taking cell parameters as input.
* This project also aims to achieve maximum accuracy in prediction of Breast Cancer using BC datasets.
* The objective of this paper is to describe a predictive model built using Machine Learning algorithms for prediction of breast cancer to improve the prognosis and chances of survival through timely clinical treatment to patients. Most influential machine learning algorithms (Wolberg, Street & Mangasarian, 1992) have been used for comparison in terms of accuracy, sensitivity, specificity, and precision.
* To attain the correct diagnosis of breast cancer and classification of patients into malignant and benign groups is the center of a lot of research.

**1.4** **Research 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 the cancer is benign or malignant additionally needs an excellent deal of effort from 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 with in 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 (Machine Learning Mastery, 2016). This cancer may be a quite common sort of cancer among girls and therefore the second highest reason behind cancer death. Within the United States, regarding one in eight girls over their time 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 tumor. The tumor can be either benign or malignant. The correct designation in determinant whether the tumor 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 an essential role is that the diagnosis of breast cancer. Therefore, machine learning techniques will facilitate doctors to create a correct identification for breast cancer and make the proper classification of been 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. I aim in this paper to compare two classification machine learning algorithms significantly to predict a benign from malignant cancer in breast cancer dataset. I aim to investigate different machine learning techniques and I will use two algorithms and apply on cancer dataset.

I will focus on machine learning algorithms: K-nearest neighbor, and Logistic regression algorithm. I will primarily study these two algorithms and analyze their result.

**1.5 Significance of the Research**

The significance of this research is for Machine learning algorithm to play a vital role in helping the medical field agents (Doctors, Nurses, Clinicians) in the primary screening of cancer, alongside the diagnosis and depiction of malignant or benign lesions, selection of treatment method and determining cancer occurrence.

**1.6 Operational Definition of Terms.**

* **MRI (Magnetic Resonance Imaging):** it is used in radiology to form pictures of the anatomy and physiological processes of the body.
* **KDD (Knowledge Discovery in Database):** is defined as a method of finding, transforming, and refining meaningful data and patterns from a raw database to be utilized in different domains or applications.
* **Malignant:** in relation to cancer; a malignant disease or growth and is likely to be harmful.
* **Benign:** a benign growth is not likely to be harmful to the patient.

**1.7 Organization of the Project**

The next chapter of this paper discusses about the similar works done before in the same field by different researchers.

The proposed model is described in the third chapter and its subsections explain the system implementation including the visualization of the dataset, train-test split and the short background of the algorithms used for this model.

The chapter four section consists of the performance metrics and the experimental results and ends with the results of the two models. Lastly, the final chapter of this paper includes few concluding summary and future prospects.

* 1. **Chapter Conclusion**

The breast comprises of identical tissues present both in men and women, therefore meaning breast is also present in men. But the occurrence in men is about 100 times less than it is in women, but for men with breast cancer the statistical rate of survival is the same as women. This chapterfocuses on breast cancer detection modalities which use breast detection dataset for analysis and follow-up detection, for better odds of survival and reduction of therapeutic sessions and treatments.

**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 (American Cancer Society, 2003). 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 supervised learning in machine learning, aims to classify unknown things supported learning existing patterns and classes from the information set and after predicting future things. The training set, that is employed to build the classifying structure, and therefore look at set, that tends to assess the classifier, are ordinarily mentioned in classification tasks (UK Breast Cancer Research, 2017).

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.

**2.1 Related Works**

There are numerous modern techniques have been evolved with the evolution of technology for the prediction of breast cancer. The work related to this field is outlined shortly as follows:

Some of the studies (Zhou, Jiang., et al, 2003, pg. 37), (Lundin., et al, 1999, pg. 281), and (Delen, Walker & Kadam, 2005, pg. 113) displayed work associated to prediction and diagnosis of diseases using machine learning techniques like decision tree for detection of cancer. KNN algorithm is well known for simplicity and versatility in implementation which makes it one of the most frequently used classification algorithms in machine learning according to Jin (Marsilin & Jiji, 2012).

Liu Lei proposed a model that uses machine learning for cancer detection. In this research, Logistic Regression algorithm of Sklearn machine learning library has been used to classify the data sets of breast cancer. Two features of maximum texture and minimum perimeter was selected, and the classification accuracy stood at 96.5% (Liu, 2018, pg. 157).

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 (Floyd et al, 1994, pg. 2944), (Fogel et al, 1995, pg.49), (Fogel et al, 1998, pg. 485), (Setiono, 1996, pg. 31), (Wilding et al, 1994, pg. 145), (Wu et al, 1993, pg. 81). ANNs have numerous successes with great progress in Breast Cancer classification and diagnosis in the very early stages (UK Breast Cancer Survival, 2017), (Floyd et al, 1994, pg. 2944), (Fogel et al, 1995, pg.49), (Fogel et al, 1998, pg. 485), (Setiono, 1996, pg. 31), (Wilding et al, 1994, pg. 145), (Wu et al, 1993, pg. 81). 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 (Abbass, 2002, pg. 265) – (Khan & Wei et al, 2001, pg. 673). The technique, however, has some limitations such as no guarantee to global optima, a lot of tuning parameters, and long training time.

Single Hidden Layer Neural Networks (SFLN) was proposed by Huang and Babri (Huang, Zhu & Siew, 2006, pg. 489) to tackle the mentioned problems with tree steps learning process that called extreme learning machine (ELM). Standard (Utomo, Kardiana & Yuliwulandari, 2014, pg. 10) and best parameterized (Utomo, Pratiwi, et al, 2014, pg. 50) ELM model was 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.

Zemouri, Omri, Devalland, Arnould, Morello, Zerhouni and Fnaiech (Zemouri, Omri, et al, 2018, pg. 159) proposed a model that uses a Breast Cancer Computer Aided Diagnosis (BC-CAD) based on joint variable selection and a Constructive Deep Neural Network” ConstDeepNet”. Wisconsin Breast Cancer Dataset (WBCD) and real data from the north hospital of Belfort (France) were used to predict the recurrence score of the Oncotype DX. They applied a method to lower the number of inputs for training a deep learning neural network. Accordingly, performance of the use of the Deep Learning architecture alone was exceeded using joint variable algorithm with ConstDeepNet.

Bellaachia and Guven (Bellaachia & Guven, 2006, pg. 10) investigated the use of Naïve Bayes, the back-propagated neural network and the C4.5 decision tree algorithms on SEER dataset which contained 16 attributes and 482,052 records. The dataset is ideal due to large amount of patient and a moderate number of attributes. From their experiment, C4.5 algorithm outperformed the rest with an accuracy of 86.7%.

Afzan Adam et al. (Afzan Adam, Computerized breast cancer diagnosis, pdf). have developed a computerized breast cancer diagnosis by combining genetic algorithm and Back propagation neural network which was developed as faster classifier model to reduce the diagnose time as well as increasing the accuracy in classifying mass in breast to either benign or malignant. In these two different cleaning processes was carried out on the dataset. In Set A, it only eliminated records with missing values, while set B was trained with normal statistical cleaning process to identify any noisy or missing values. At last Set A gave 100% of highest accuracy percentage and set B gave 83.36% of accuracy. Hence the author has concluded that medical data are best kept in its original value as it gives high accuracy percentage as compared to altered data.

Tuba kiyan (Tuba & Tubay, 2004) has discussed that statistical neural networks can be used to perform breast cancer diagnosis effectively. The scholar has compared statistical neural network with Multi-Layer Perceptron on WBCD database. Radial basis function (RBF), General Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) were used for classification and their overall performance were 96.18% for Radial Basis Function (RBF), 97% PNN, 98.8% for GRNN and 95.74% for MLP. Hence it is proved that these statistical neural network structures can be applied to diagnose breast cancer.

With due respect to all related work referred above, this paper compares the performance of the algorithms; K-Nearest Neighbors (KNN), Logistic Regression (LR) 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.

**2.2 Machine Learning**

Machine learning (ML) is one of the fields of Artificial Intelligence (AI) where statistical techniques are used to provide the computer systems with the capability to “learn” and improve by itself progressively without being explicitly programmed. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data and datasets (Kohavi, 1998, pg. 127) based on different teaching mechanisms. The term ‘machine learning’ was initially coined by Arthur Samuel in 1959 (Samuel, 1959, pg. 210). Three important categories of machine learning can be described as follows:

1. **Supervised learning:** In this form of learning, the machine uses data that is labeled and some of the data is already tagged with the correct answers to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when you have new input data (x) you can predict the output variables (Y) for that data.

2. **Unsupervised learning:** The machine is trained using data that is neither classified nor labeled. This allows the algorithm to perform calculations on its own accord without guidance. The task of the machine is to group unsorted data according to patterns and differences to find the hidden structure by itself.

3. **Reinforcement learning:** The machine or the agent learn how to behave in an environment by performing actions and seeing the results based on the action allowing it to dictate the ideal action in a specific circumstance.

In the modern times, the vast amount of data available is not feasible for human being to keep up with and analyze them. Machine learning, which is a subset of computer science and an important branch of artificial intelligence, primarily focuses on the development and building of algorithms to over this problem. The very recent advancement in this field has opened vast and almost limitless applications in fields ranging from financial industries, data security to medical fields. But there is still much space to make progress by means of using Machine Learning in social media services, disease prediction and identification, virtual assistants, search engine refining, fraud detection, manufacturing, etc. It is only going to improve and integrate in our daily lives making it easier and more convenient in the future.

**CHAPTER 3**

**PROPOSED MODEL**

With my aim being to predict whether the tumor is Benign (non-cancerous) or Malignant (cancerous), I 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 my dataset, I split the train-test ration to 70:30 in order to train and test two algorithms: K-Nearest Neighbors (KNN), and Logistic regression (LR).

**Get Dataset**

**Analyze Data**

**Feature Selection**

**Train Model**

**Performance Analysis**

**Improve Model**

Fig.3.1 Workflow

**3.1 Dataset**

The dataset chosen for this research is the Wisconsin Breast Cancer (Diagnostic) Data Set (WBCD). The dataset is publicly available on the reputed Machine Learning Repository that is UCI-Repository. WBCD was made by Dr. William H. Wolfberg, doctor at the University of Wisconsin Hospital at Madison, Wisconsin, USA. Dr. Wolfberg used Xcyt to analyze fluids samples taken from patients with solid breast masses (Wolberg et al, 1992,). Xcyt is an easy-to-use graphical computer program which is equipped to perform the investigation of cytological features based on digital scans. The dataset comprises of 569 samples and 32 attributes of visually measured atomic features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. FNA is a thin needle which is injected into the region of abnormal-appearing body fluid or tissues and collects a sample to make a diagnosis or predicting disease such as cancer. Among the 569 samples, the class distribution are 212 cancerous tumors (malignant) and other 357 non-cancerous tumors (benign). Ten features are computed from each one of the cells in the sample which are as follows:

1. Radius (mean of distances from center to points on the perimeter)
2. Texture (standard deviation of gray-scale values)
3. Perimeter
4. Area
5. Smoothness (local variation in radius lengths)
6. Compactness (perimeter² / area—1.0)
7. Concavity (severity of concave portions of the contour)
8. Concave points (number of concave portions of the contour)
9. Symmetry
10. Fractal dimension (“coastline approximation”—1)

The mean, standard error, and “worst” or largest (mean of the three largest values) of These features were computed for each image, resulting in 30 features. For instance, field 3 Is Mean Radius, field 13 is Radius SE and field 23 is Worst Radius. All feature values are Recoded with four significant digits. There are no missing values in the dataset. There is Presence of both numerical and categorical features in the dataset. ‘Diagnosis’ and ‘Unnamed 32’ are the only Column with categorical feature, I am going to predict diagnosis, which says if the cancer is M = malignant or B = benign. The rest of the features are numerical.

**3.2 Data Visualization**

**3.2.1 Histogram**

A histogram is a graphical representation of data or information using bars of different heights where each bar groups numbers into ranges. Higher bars show that more data falls in that range. The shape and spread of continuous sample data can be shown using a histogram. Figure 3.2 will show the class distribution of diagnosed malignant (M) and benign (B) tumors. There are 212 malignant tumors which is approximately 38% and other 357 benign tumors making up the rest of the 62% of the predictive class.

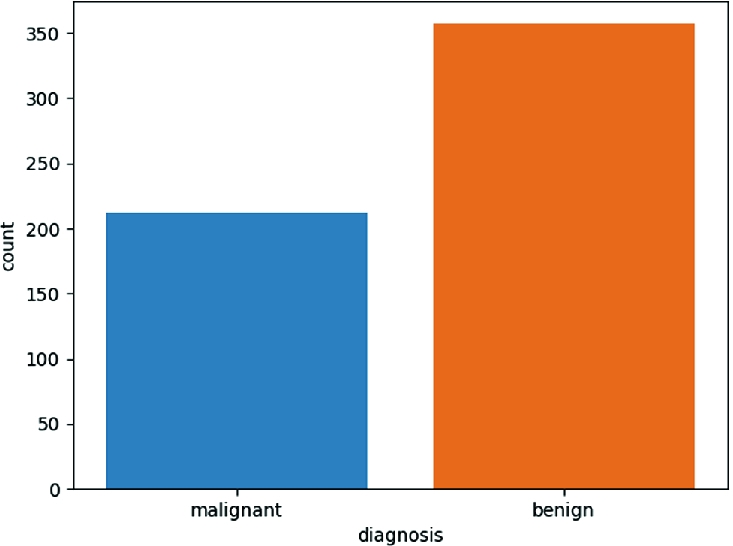


Fig. 3.2 Class Distribution

**3.2.2 Heatmap**

A heatmap is a two-dimensional representation with the help of colors for visualization of Both simple and complex information. Heatmap is an extremely useful way to see which Intersections of the values have higher concentration of the data compared to the others. Figure 3.3 will represent a correlation matrix using a heatmap. It is used to show the correlation Among all 30 features in these datasets.

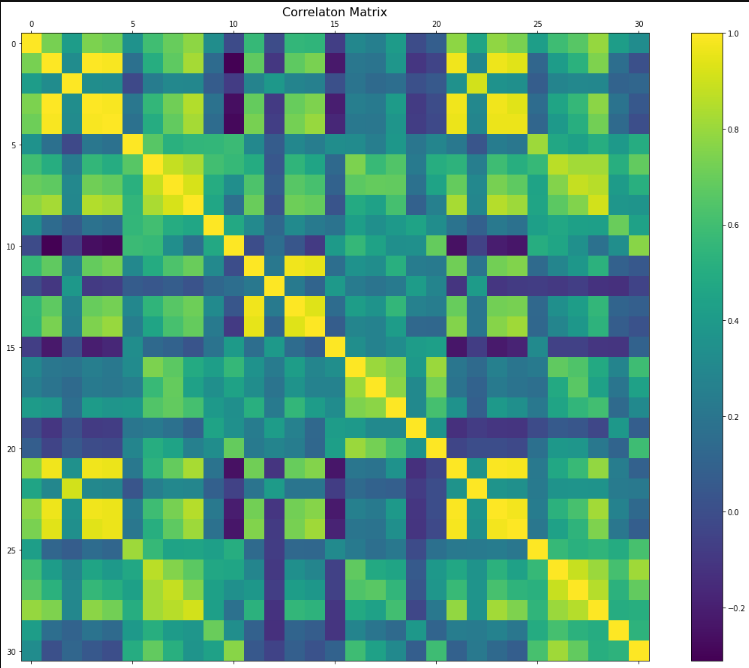


Fig. 3.3 Correlation between all variables

**3.3 Data Preprocessing**

**3.3.1 Categorical Variable Conversion**

The dataset included both numerical and categorical features. Among which the ‘Diagnosis’ column had categorical feature, which says if the cancer is M = malignant or B = benign. The rest of the features are numerical. Most of the algorithms produce better result with numerical variable. In python, library “sklearn” requires features in numerical arrays and categorical variables cannot be fitted into a regression equation in their raw form. Hence, label Encoder was used to transform non-numerical to numerical labels.

**3.3.2 Feature Selection**

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. 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.

**3.3.2 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, to make predictions on the test data. Training dataset is a part of the actual dataset that is 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 (Adi, Understanding Machine Learning, 2017).

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 project, the dataset has to be split 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, I have slotted 70% of the dataset to training. Out of the 70% dataset for training, 63 percent is kept 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.

**3.4 Algorithms**

The model deals with binary classification of labeled data and the algorithms were chosen Based on that fact. The machine learning algorithms chosen for this problem are K-Nearest Neighbors (KNN), Logistic Regression (LR), The results of the algorithms were compared to determine the best classifier for the problem.

**3.4.1 K-Nearest Neighbors**

K-Nearest Neighbors algorithm is a very simple, yet functional non-parametric method used for classification and regression (Altman, 1992, pg. 175). The model that is created are results from the training sample, associated with a distance function and the choice function of the class. Before a new element is classified, it is compared to other elements using a similarity measure. The k-nearest neighbors of the elements are then considered, and the class that is most common among the neighbors, is assigned to the element to be classified. Using the distance, the neighbors are weighted. KNN can demand a lot of memory or space to store all the data but only performs a calculation when a prediction is needed.

**3.4.2 Logistic Regression**

Logistic regression is another supervised learning technique borrowed by machine learning from the field of statistics. In Logistic Regression the output or target variable is a categorical variables, unlike Linear Regression, and is thus a binary classification algorithm that categorizes a data point to one of the classes of the data (Hosmer et al, 2013, vol 398). The general equation of Logistic Regression is:



Where, p(X) is the dependent variable,

X is the independent variable,

β0 is the intercept and

β1 is the slope co-efficient.

Input values (X) are combined linearly using coefficient values to determine an output (Y). It significantly differs from linear regression because the output value being modeled is a binary value (0 or 1) rather than a numeric value. The value of the score lies between positive and negative infinity but needs to be between [0, 1]. Therefore, the logistic function, also known as the sigmoid function, is used to make the conversion. It’s a curve that is shaped like an S, which takes any real number and maps it into a value between 0 and 1. It is widely used for binary classification problems and works better when attributes are reduced that are correlated to each other which explains the results in the future section of result analysis. It’s a fast model to learn and effective on binary classification problems.

**CHAPTER 4**

**RESULT ANALYSIS**

After the successful implementation of machine learning model, it is very important to measure that how effective the model is and how the model performs on the dataset that I have chosen for our research. I noticed that algorithms performance is different. In this project, various execution parameter to figure out which machine learning algorithm will be the best option for the detection of breast cancer. From WBCD dataset, I have taken 70% of the data for training purpose and rest 30% was used for testing to find out the best possible result and the more satisfying result.

**4.1 Performance Metrics**

This research deals with classification problem and therefore the chosen performance metrics primarily focus on classification. To predict breast cancer, the tagged variable 1(Malignant) means it is a positive instance and that clearly refers the patient is having a breast cancer. On the contrary, 0(Benign) means it is a negative instance and that indicates the patient having no breast cancer.

**4.1.1 Confusion Matrix**

Summarization of 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 classification 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.

|  |  |  |
| --- | --- | --- |
|  | **Predictive**  **Negative** | **Predictive**  **Positive** |
| **Actual**  **Negative** | **True Negative**  **(TN)** | **False Positive**  **(FP)** |
| **Actual**  **Positive** | **False Negative**  **(FN)** | **True Positive**  **(TP)** |

Table 4.1 Confusion Matrix

Important terms of a Confusion Matrix and their related discussion are given below for better understanding of a reader:

**True Positives (TP):**

Basically, in this situation both the anticipate class and the real class is correct (true) (1), i.e., while a classifier predicts the sufferer obtained a complication of breast cancer and the patient has a breast cancer. So, the classifier is predicting correct decision in this situation.

**False Positives (FP):**

This situation refers the classifiers anticipate class is correct(true) (1), but the actual class is wrong(false) (0), i.e., while a classifier predicts that a sufferer has a complication of breast cancer, but the sufferer has no breast cancer. So, the classifier is unable to predict correct decision for that case.

**True Negatives (TN):**

True Negatives state that the anticipate class and the real class are false (0), i.e., while a classifier predicts that a sufferer has a no complication of breast cancer, and the sufferer does not contain breast cancer. Therefore, this classifier is predicting the correct decision.

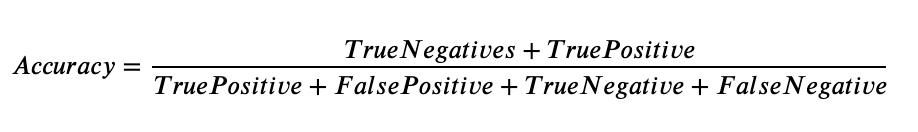
**False Negatives (FN):**

Essentially, this situation refers the classifiers anticipate class is false (0), but the real class is correct(true) (1), i.e., while the classifier predicts that the sufferer has no complication of breast cancer, but the patient has a breast cancer. So, the classifier is unable to predict correct decision similarly to False Positives.

Thus, the classifier’s accuracy is higher when more TP and TN are found inside confusion matrix. Similarly, accuracy of a classifier decreases when the amount of FP and the FN increases in a Confusion Matrix. So, the best situation would be when none of the FP and FN would be founded inside the model. If it happens, the model can give us the 100% accuracy.

**4.1.2 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:



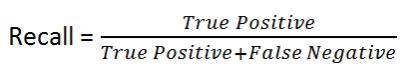
**4.1.3 Precision**

Precision is called that which generate ratio of True Positives to the summation of True Positives and False Positives. Simply high precision means that an algorithm generated mostly appropriate results than inappropriate.



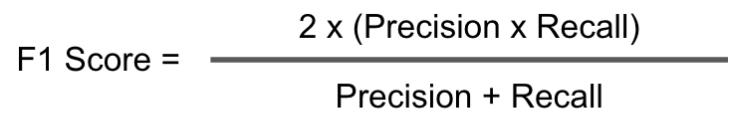
**4.1.4 Recall**

Recall is a measure that shows the proportion of patients that actually had malignant tumor was diagnosed by the algorithm as having malignant tumor. The actual positives (People having malignant tumor are TP and FN) and the people diagnosed by the model having a malignant tumor are TP. Therefore, if I want to focus more on minimizing False Negatives, I will want our Recall to be as close to 100% as possible.



**4.1.5 F1 Score**

F1 Score is the Harmonic Mean between precision and recall. Additionally, weighted average of precision and recall is known as F1 score. The span or range of F1 score is from 0 to 1. F1 score represents how accurate the classifier is and shows how durable that is at the same time.

**4.2 Model Performances**

A total set of three classification algorithms are used - Logistic Regression (LR), and K-Nearest Neighbors (KNN) Classifier have been applied on the dataset. For each experiment, the performance of the algorithms is measured using Accuracy, Precision, Recall, and F1 Score.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Accuracy** | **Precision** | **Recall** | **F1 score** |
| **Logistic Regression** | 0.947 | 0.95 | 0.904 | 0.926 |
| **K-Nearest Neighbor** | 0.929 | 0.918 | 0.888 | 0.903 |

Table 4.2 scores of Accuracies, Precision, Recall and F1 Score

In the following sub-sections, the Confusion Matrix for every algorithm is represented through figures. As shown previously in Table, the Confusion Matrix has four values - True 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, and F1 Score to evaluate the performance of each algorithm.

**4.2.1 Logistic Regression (LR)**

The figures below illustrate the Confusion Matrix of Logistic Regression (Figure 4.1). The results show that Logistic Regression scores a good accuracy of 0.947 along with figures of 0.95, 0.904 and 0.926 for precision, recall and F1 Score respectively.

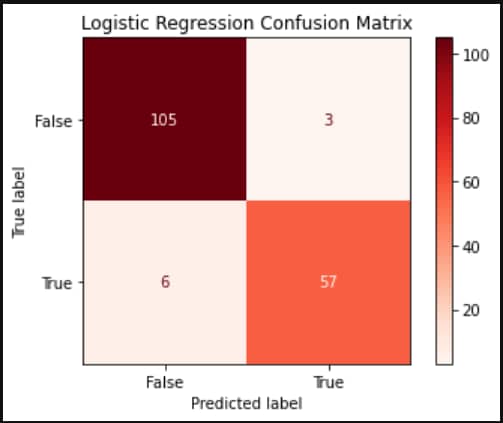
****

Fig 4.2 Logistic regression Confusion matrix

**4.2.2 K-Nearest Neighbor (KNN)**

The figures below illustrate the Confusion Matrix of the K-Nearest Neighbors on this dataset. The results show how well K-Nearest Neighbor performs with an accuracy of 0.929. Precision, Recall and F1-Score also scores good figures of 0.918, 0.888 and 0.903 respectively.

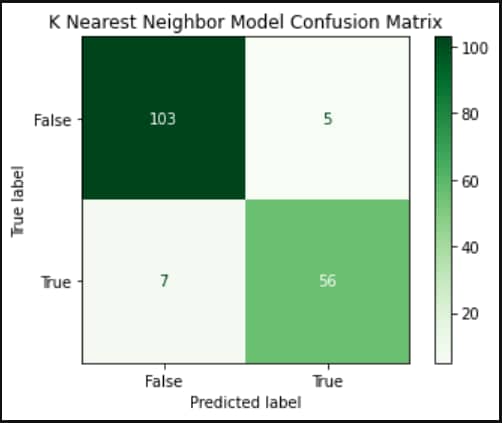


Fig 4.2 K-Nearest Neighbor Confusion matrix.

**CHAPTER 5**

**CONCLUSION**

There are different data mining techniques that can be used for the prediction of breast cancer recurrence. In this paper, I used supervised method for my research. I have used two algorithms to predict the breast cancer more accurately. I used K-Nearest Neighbor and Logistic regression algorithms. I analyzed the results of all algorithms and tried to find out the best possible one. In my work, the Logistic regression algorithm proved to be a better algorithm for the prediction task.

Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

.

**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] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3):175–185.

[4] Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). Applied logistic regression, volume 398. John Wiley & Sons.

[5] Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, 34(2):113–127.

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

[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] Kohavi, R. (1998). Glossary of terms. Special issue on applications of machine learning and the knowledge discovery process, 30(271):127–132.

[16] Liu, L. (2018). Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In 2018 International Conference on Robots & Intelligent System (ICRIS), pages 157–160. IEEE.

[17] Lundin, M., Lundin, J., Burke, H., Toikkanen, S., Pylkkänen, L., and Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. Oncology, 57(4):281–286.

[18] Marsilin, J. R. and Jiji, G. W. (2012). An efficient cbir approach for diagnosing the stages of breast cancer using knn classifier. Bonfring International Journal of Advances in Image Processing, 2(1):01–05.

[19] Bellaachia, A. and Guven, E. (2006). Predicting breast cancer survivability using data Mining techniques. Age, 58(13):10–110.

[20] Samuel, A. L. (1959). Some studies in machine learning using the game of checkers.

IBM Journal of research and development, 3(3):210–229.

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

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

[22] Afzan Adam1 Khairuddin Omar2 “Computerized Breast Cancer Diagnosis with Genetic Algorithms and Neural Network” fitt.mmu.edu.my/caiic/papers/afzaniCAIET.pdf.

[23] 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

[24] 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.

[25] Wolberg, W., Street, W. N., and Mangasarian, O. L. (1992). Breast cancer wisconsin (diagnostic) data set [uci machine learning repository].

[26] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. Knowledge and information systems, 14(1):1–37.M

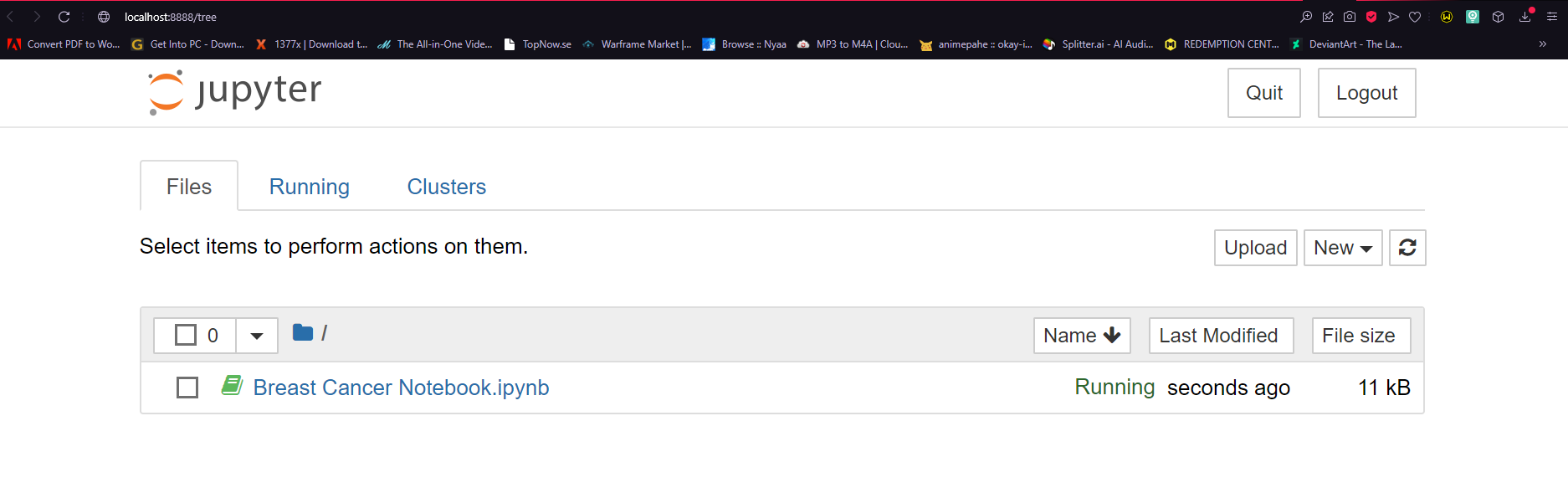
[27] Fine Needle Aspiration Biopsy of the Breast. American Cancer Society

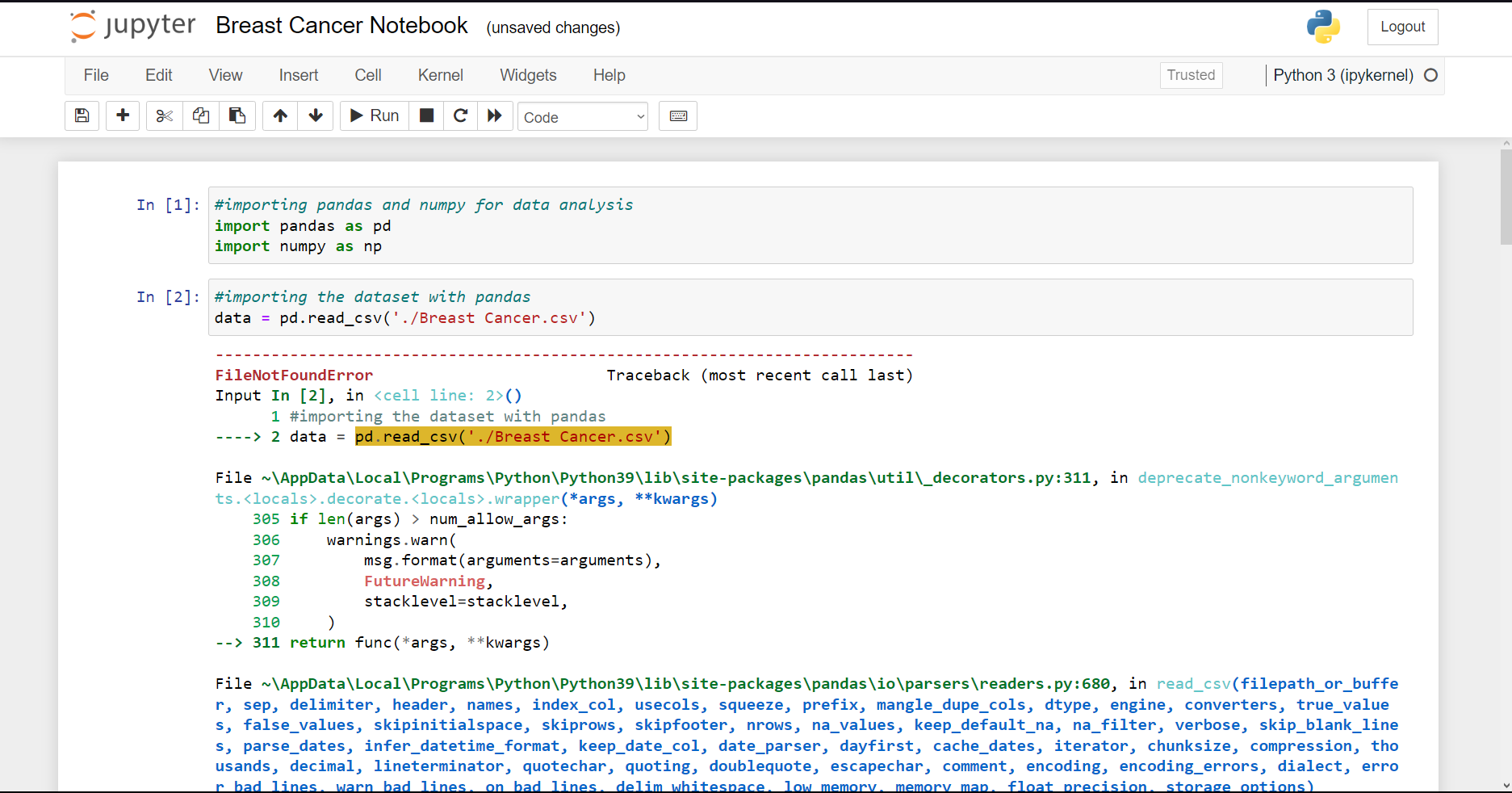
[28] Adi Bronshtein. Train/Test Split and Cross Validation in Python. Understanding Machine Learning (2017).

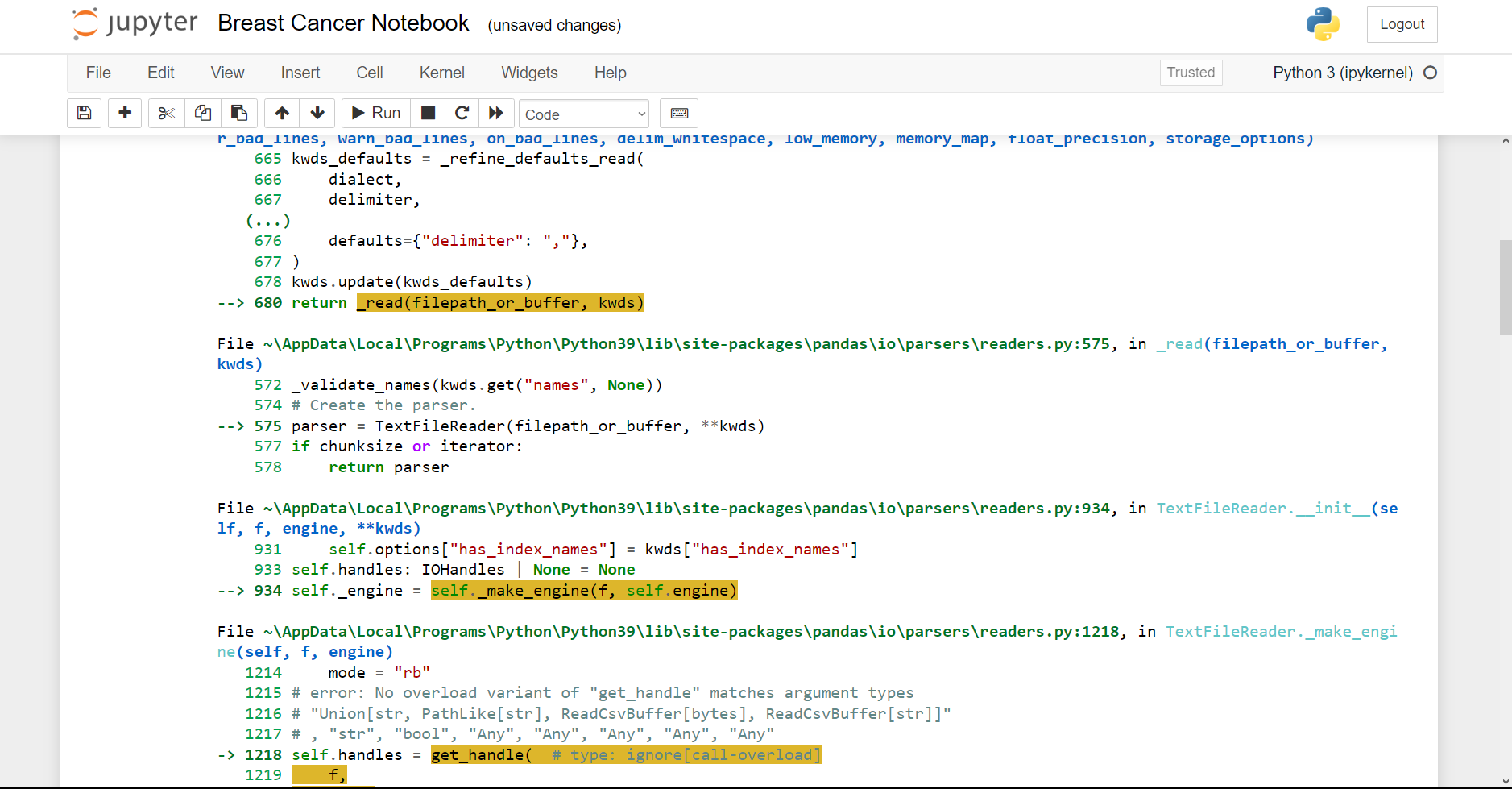
[29] Zhou, Z.-H., Jiang, Y., et al. (2003). Medical diagnosis with c4. 5 rule preceded by artificial neural network ensemble. IEEE Transactions on information Technology in Biomedicine, 7(1):37–42.

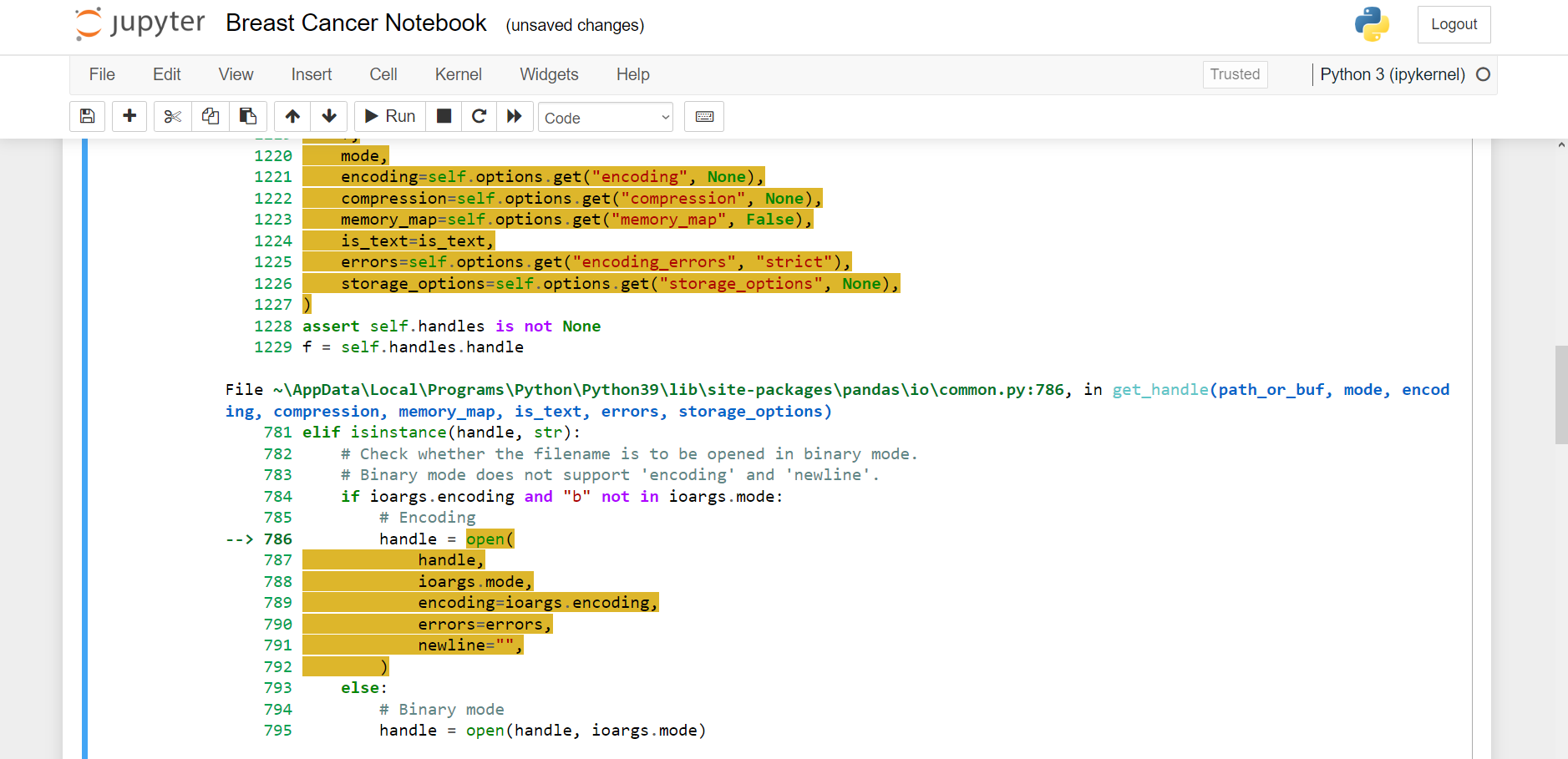
*[30] Tuba kiyan, Tulay Yildirim “Breast cancer diagnosis using statistical neural networks”, Journal of Electrical and Electronic Engineering.*

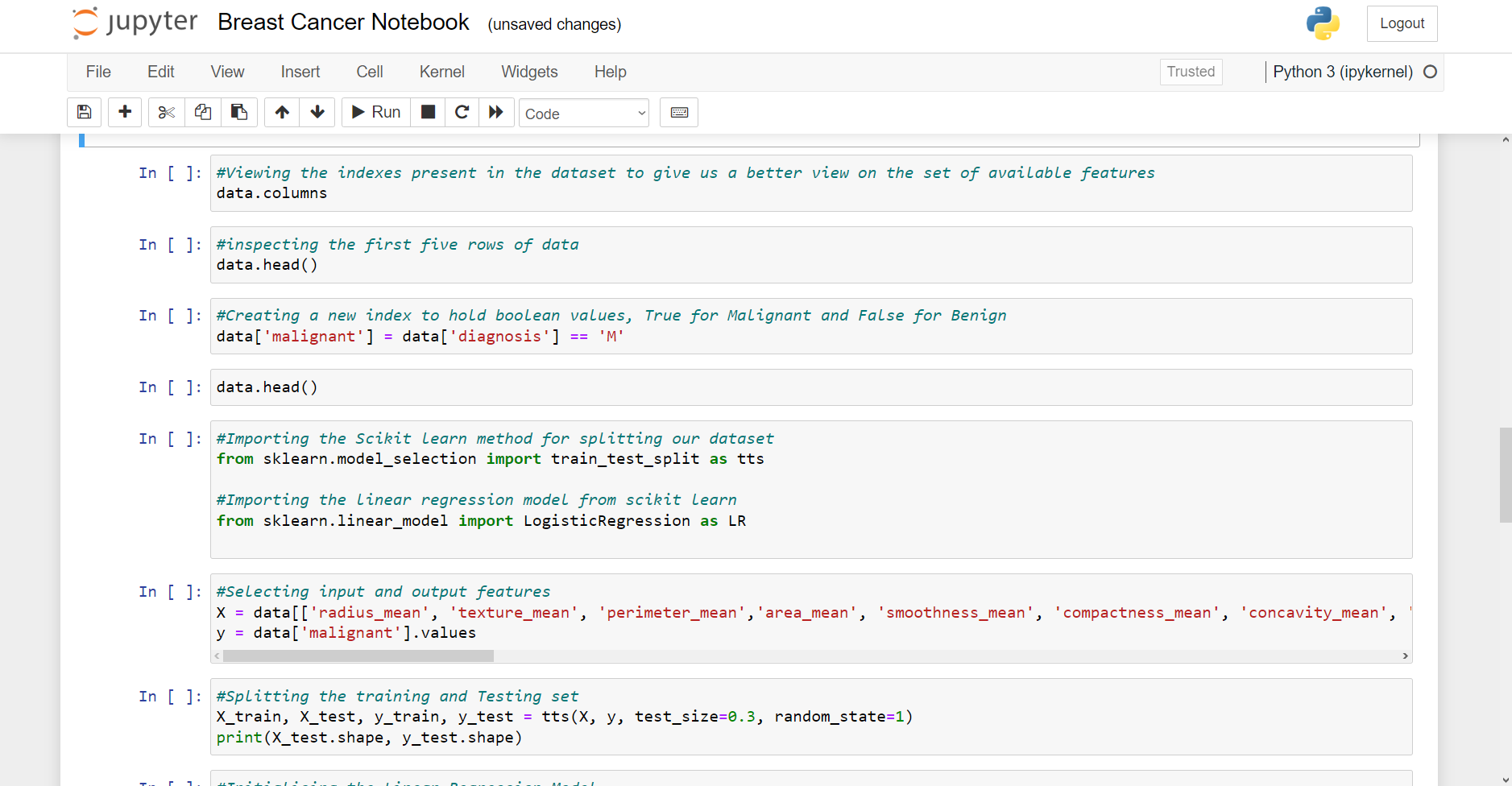
**APPENDIX 1**



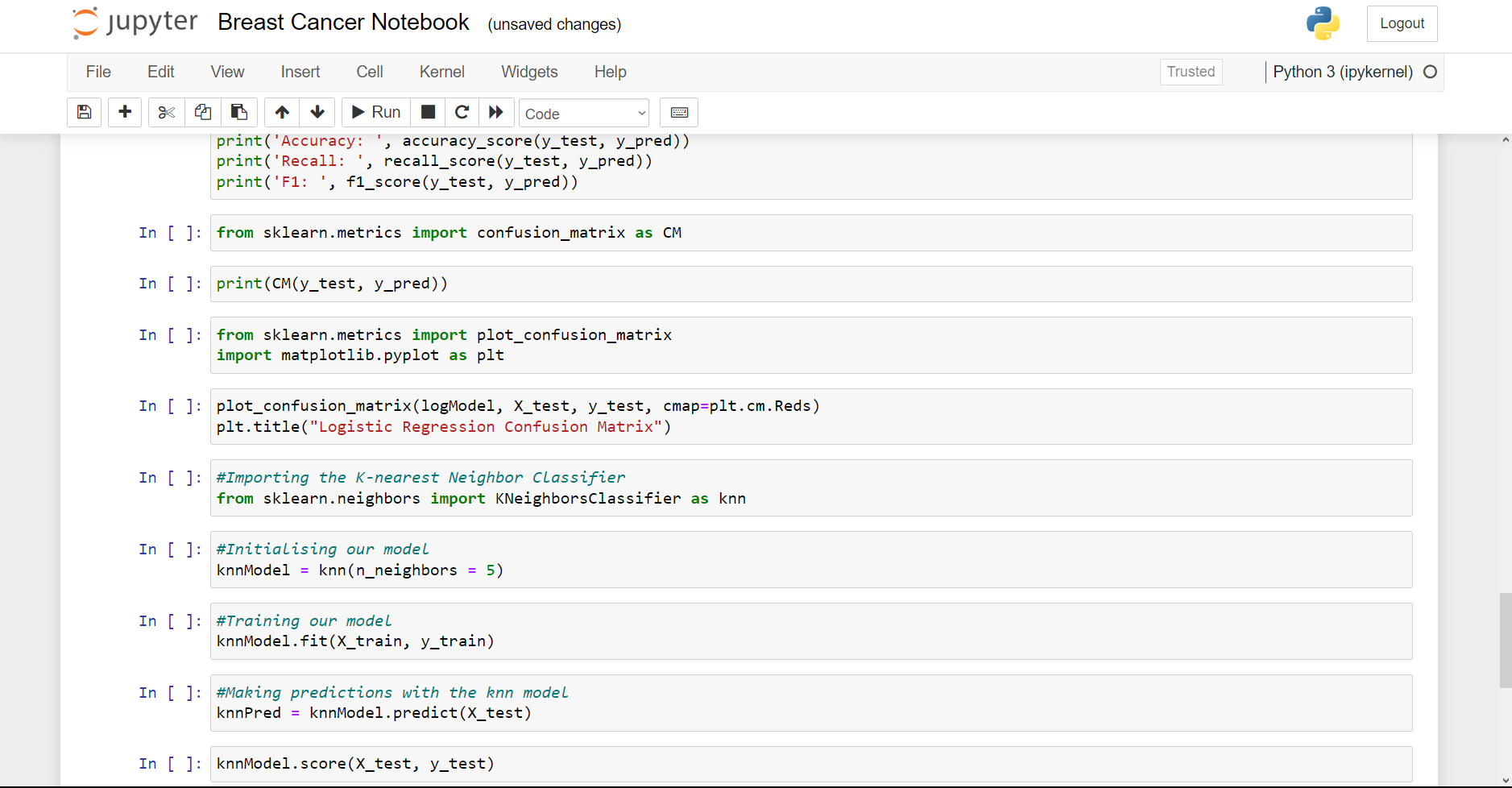


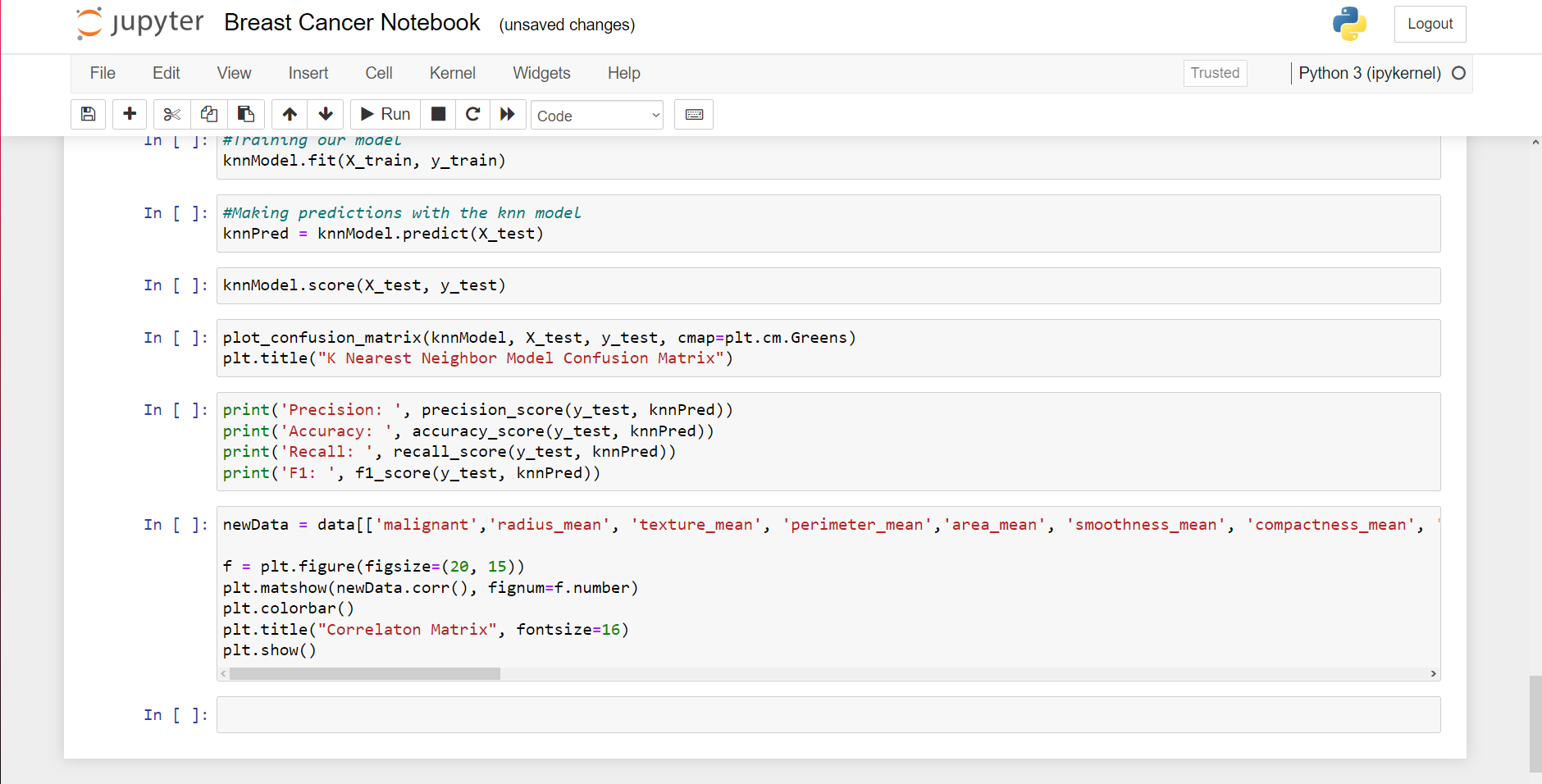












**APPENDIX 2**

#importing pandas and numpy for data analysis  
import pandas as pd  
import numpy as np  
#%%  
#importing the dataset with pandas  
data = pd.read\_csv('./Breast Cancer.csv')  
#%%  
#Viewing the indexes present in the dataset to give us a better view on the set of available features  
data.columns  
#%%  
#inspecting the first five rows of data  
data.head()  
#%%  
#Creating a new index to hold boolean values, True for Malignant and False for Benign  
data['malignant'] = data['diagnosis'] == 'M'  
#%%  
data.head()  
#%%  
#Importing the Scikit learn method for splitting our dataset  
from sklearn.model\_selection import train\_test\_split as tts  
  
#Importing the linear regression model from scikit learn  
from sklearn.linear\_model import LogisticRegression as LR  
  
#%%  
#Selecting input and output features  
X = data[['radius\_mean', 'texture\_mean', 'perimeter\_mean','area\_mean', 'smoothness\_mean', 'compactness\_mean', 'concavity\_mean', 'concave points\_mean', 'symmetry\_mean', 'fractal\_dimension\_mean', 'radius\_se', 'texture\_se', 'perimeter\_se', 'area\_se', 'smoothness\_se', 'compactness\_se', 'concavity\_se', 'concave points\_se', 'symmetry\_se','fractal\_dimension\_se', 'radius\_worst', 'texture\_worst', 'perimeter\_worst', 'area\_worst', 'smoothness\_worst', 'compactness\_worst', 'concavity\_worst', 'concave points\_worst','symmetry\_worst', 'fractal\_dimension\_worst']].values  
y = data['malignant'].values  
#%%  
#Splitting the training and Testing set  
X\_train, X\_test, y\_train, y\_test = tts(X, y, test\_size=0.3, random\_state=1)  
print(X\_test.shape, y\_test.shape)  
#%%  
#Initialising the Linear Regression Model  
logModel = LR(solver='liblinear')  
#%%  
#Training our model  
logModel.fit(X\_train, y\_train)  
#%%  
#Predicting from data  
logModel.predict([X[89]])  
#%%  
print([y[89]])  
#%%  
#Checking the performance of our model  
logModel.score(X\_test, y\_test)  
#%%  
#Further model performance metrics  
from sklearn.metrics import accuracy\_score, precision\_score, recall\_score, f1\_score  
#%%  
y\_pred = logModel.predict(X\_test)  
#%%  
print('Precision: ', precision\_score(y\_test, y\_pred))  
print('Accuracy: ', accuracy\_score(y\_test, y\_pred))  
print('Recall: ', recall\_score(y\_test, y\_pred))  
print('F1: ', f1\_score(y\_test, y\_pred))  
#%%  
from sklearn.metrics import confusion\_matrix as CM  
#%%  
print(CM(y\_test, y\_pred))  
#%%  
from sklearn.metrics import plot\_confusion\_matrix  
import matplotlib.pyplot as plt  
#%%  
plot\_confusion\_matrix(logModel, X\_test, y\_test, cmap=plt.cm.Reds)  
plt.title("Logistic Regression Confusion Matrix")  
#%%  
#Importing the K-nearest Neighbor Classifier  
from sklearn.neighbors import KNeighborsClassifier as knn  
#%%  
#Initialising our model  
knnModel = knn(n\_neighbors = 5)  
#%%  
#Training our model  
knnModel.fit(X\_train, y\_train)  
#%%  
#Making predictions with the knn model  
knnPred = knnModel.predict(X\_test)  
#%%  
knnModel.score(X\_test, y\_test)  
#%%  
plot\_confusion\_matrix(knnModel, X\_test, y\_test, cmap=plt.cm.Greens)  
plt.title("K Nearest Neighbor Model Confusion Matrix")  
#%%  
print('Precision: ', precision\_score(y\_test, knnPred))  
print('Accuracy: ', accuracy\_score(y\_test, knnPred))  
print('Recall: ', recall\_score(y\_test, knnPred))  
print('F1: ', f1\_score(y\_test, knnPred))  
#%%  
newData = data[['malignant','radius\_mean', 'texture\_mean', 'perimeter\_mean','area\_mean', 'smoothness\_mean', 'compactness\_mean', 'concavity\_mean', 'concave points\_mean', 'symmetry\_mean', 'fractal\_dimension\_mean', 'radius\_se', 'texture\_se', 'perimeter\_se', 'area\_se', 'smoothness\_se', 'compactness\_se', 'concavity\_se', 'concave points\_se', 'symmetry\_se','fractal\_dimension\_se', 'radius\_worst', 'texture\_worst', 'perimeter\_worst', 'area\_worst', 'smoothness\_worst', 'compactness\_worst', 'concavity\_worst', 'concave points\_worst','symmetry\_worst', 'fractal\_dimension\_worst']]  
  
f = plt.figure(figsize=(20, 15))  
plt.matshow(newData.corr(), fignum=f.number)  
plt.colorbar()  
plt.title("Correlaton Matrix", fontsize=16)  
plt.show()  
#%%