BREAST CANCER PREDICTION USING MACHINE LEARNING

BY

RAIHAN JAMI KHAN
ID: 201-15-13620

This Report Presented in Partial Completion of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Co-supervised By

Ms. Johora Akter Polin
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project titled "**BREAST CANCER PREDICTION USING MACHINE LEARNING**", submitted by Raihan Jami Khan ID No:201-15-13620 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 19-01-2023.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Chairman

Narayan Ranjan Chakraborty
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Raja Tariqul Hasan Tusher
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Md. Safaet Hossain
Associate Professor & Head
Department of Computer Science and Engineering
City University

External Examiner

DECLARATION

I hereby declare that, this project has been done by us under the supervision of **Dr. Md Zahid Hasan, Associate Professor & Program Director MIS, Department of CSE at Daffodil International University**. I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Md Zahid Hasan
Associate Professor & Program Director MIS
Department of CSE
Daffodil International University

Co-Supervised by:

Ms. Johora Akter Polin
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Jami Khan

Raihan Jami Khan
ID: 201-15-13620
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

I really grateful and wish my profound my indebtedness to **Dr. Md Zahid Hasan**, **Associate Professor & Program Director MIS**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of my supervisor in the field of "***Machine Learning***" to carry out this project. His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Dr. Touhid Bhuiyan**, Professor, and Head**,** Department of CSE, for his kind help to finish my project and also to other faculty member and the staff of CSE department of Daffodil International University.

I would like to thank all of my classmates at Daffodil International University who participated in this discussion while also attending class.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

ABSTRACT

Nowadays, Breast cancer is one of the foremost causalities, and it's the second most familiar reason for death for women. Circulation of distal organ tumors is the primary cause of death from breast cancer. Breast cancer has now become a common health issue, and its expansion and harmony have increased recently. Sometimes breast cancer spreads without a family history. Also, heightened chances of breast cancer retaining aging, genes, dense breast tissue, obesity, and radiation exposure. Sometimes women don't even know they have breast cancer. There are two distinct kinds of malignant and benign tumors, and physicians should use a reliable diagnostic strategy to differentiate between them. The main ambition of this paper is to utilize the most outcomes in developing a classification and related strategies. Earlier detection of breast cancer will assist in the survival of patients with breast cancer. Machine learning helps build planning models to predict planning models that can be utilized to predict consequences for individual patients. Data mining and machine learning help in the early detection of breast cancer. The goal of this study is to review the role of machine learning methods in the prediction and diagnosis of breast cancer. Most of these methods focus on predicting breast cancer by using machine learning.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

1.1 Introduction

The second primary cause of women's death is carcinoma (after heart disease). In the past few years, More than 12 million women have died from carcinoma. Carcinoma is a type of cancer that begins in the breast. Carcinoma can spread when cancer cells penetrate the blood or lymphatic system. Breast cancer tissues are usually not detected on X-ray and MRI. It is diagnosed when a tumor forms from this tissue. A report in 2019 revealed that Some 12,764 women are seen with breast cancer in Bangladesh annually, and 6,844 of them die of the disease. This number is getting bigger day by day. There are many considerable algorithms for classifying breast cancer prediction. The present paper compares the performance of 7 classifications. Those are Logistic Regression, KNN, Decision Tree Classifier, Random Forest, Support Vector Machine, Naive Bayes, Artificial neural networks (ANNs).These classifications can determine the stage of the disease. Patients must undergo carcinoma surgery, chemotherapy, radiotherapy, and endocrine to stop cancer from spreading. The classification aims to identify and classify malignant and benign patients and to parametrize my classification methods to gain high accuracy. I am analyzing multiple datasets and additional added on how machine learning algorithms identify stages. I want to scale the error rates again with the highest accuracy. I use the data lore platform to train and test datasets, which a Machine Learning platform used in JUPYTER to evaluate the info and analyze data in terms of effectiveness and efficiency.

1.2 Overview of Breast Cancer

Breast tissue that is fibrous or contains animal tissue, including breast cells, can develop breast cancer. Breast cancer is a malignant disease with rapidly expanding, gradually worsening tumors that often result in death. Despite being more prevalent in females, carcinoma rarely happens in guys. A tumor can either be benign (not harmful to health) or malignant (potentially dangerous). The risk of carcinoma might be increased by elements including age and a case history of the condition.

1.3 Types and stages of Breast Cancer

Options for treating carcinoma are supported by cancer kind and stage. The categorization of carcinoma depends on the condition of the cell surface receptors.

Tumors generally fall into two categories:

- **Benign:** This particular tumor kind seldom results in death and is not harmful to a person's health.This form of tumor has restricted growth and only affects one area of the body.

- **Malignant:** This more lethal tumor kind, known as carcinoma, is more harmful. When cells within the breast tissue proliferate improperly, cancerous growth results. the majority of carcinoma types are:

    1. **Ductal carcinoma in place (DCIS): i**s a treatable form of the earliest kind of carcinoma.

    2. **Invasive Ductal Carcinoma (IDC):**The most prevalent kind of cancer starts in the milk duct.

    3. **Invasive Lobular Carcinoma (ILC):**originate in a very breast lobule. Human papillomavirus has the ability to quickly spread to the body's lymph nodes and other regions.

The primary stages of carcinoma

Breast cancer stage can describe the size of the tumor and indicate if the disease has spread. The majority of carcinoma phases include:

Stage I: The cancer has not progressed to the lymph nodes or other tissues; it is contained to a limited location.

Stage II: Despite expanding, the cancer has not yet spread.

Stage III: The tumor could have expanded and migrated to nearby lymph nodes or other tissues.

Stage IV: Ymy body's organs or other regions have been infected by cancer. Additionally known as metastatic or advanced cancer, this stage.

1.4 Treatment of carcinoma

Patients are occasionally given one form of treatment or a mix of treatments based on the age, type, and stage of the woman's cancer. The majority of cancer therapies are:

1. **Surgery:** For cancer, there are primarily two surgical approaches. Breast conserving surgery, often known as a lumpectomy, is the most common type of operation. The goal of surgery is to remove a cancerous part of the breast along with some surrounding healthy tissue. The second type of surgery might be a mastectomy, which involves removing the entire breast.
2. **Radio Therapy:** It harnesses nonparticulate radiation to destroy cancer cells.
3. **Chemo Therapy:** It will utilize cytotoxic medications to eradicate cancer cells throughout the body, including those in the breast.
4. **Hormone Therapy:** It is frequently used following surgery to help lower the risk of cancer recurrence or treat cancer that has spread to other body areas. typically taken for at least five years.
5. **Biological Therapy:** It includes modern medications that work differently from chemotherapy. It lessens the likelihood of breast cancer developing.

1.5 Prevalence and survival rates of carcinoma

As the most prevalent disease diagnosed globally, female breast cancer has now overtaken lung cancer. In 2020, it is anticipated that 2,261,419 new cases of cancer in women. In the United States, there will likely be 530 men and 43,250 women who pass away from breast cancer this year. Female breast cancer is the sixth most common cause of mortality in the world. Around the world, 684,996 women are anticipated to pass away from breast cancer in 2020. After lung cancer, breast cancer is the second most prevalent reason for cancer-related deaths among women in the US. However, because to advances in early identification and treatment, the number of women who have died from breast cancer has fallen by 42% between 1989 and 2019. As a consequence, throughout that time, more than 431,800 breast cancer deaths were avoided. Black women had a 41% higher mortality risk from the illness despite receiving fewer breast cancer diagnoses than White women. I have now discussed the working procedure starting with the first step. a succinct explanation on how to setup an environment (Anaconda).

1.6 Motivation

At this time, cancer is affecting a lot of individuals. The disease's causation is dependent on external factors and is unknown. Additionally, this serves as a screening technique to identify if the malignancy is benign or aggressive. It brings a lot of work to accomplish this. Many test questions, including the following, are connected to cancer detection: B. The clamp's thickness, the consistency of cell size, and the consistency of cell shape. Even motivators find the final product difficult to achieve, and in recent years, the use of computer science and machine learning as universal diagnostic tools has grown. Applications for computer diagnostics frequently exploit diseases that have resulted in numerous fatalities. In the operating room, robots play a crucial role.Other artificial intelligences are also frequently employed to identify cancer, and the critical care unit has an effective method. Cancer, which first arrived on the planet in 2002, is now known to be the second most prevalent kind of cancer among females. In the United States, one in eight females at this time is at risk of developing cancer. Unchecked breast cell division might result in cancer and eventually noticeable fillers (called tumors). Typically, the tumor is benign. The necessity for precise categorization in the clinic can be a severe issue for doctors and other healthcare workers, regardless of whether the condition is benign or malignant since the proper labeling of the determinants will result in survival. The significance of artificial intelligence has increased during the past 25 years. The most significant disease screening assignment in the medical sector is the use of computers and machine learning as diagnostic tools, which has become fatal as scientists grasp the necessity of making strong conclusions about how to treat specific diseases. A role has a significant impact. The definition of cancer is one of the most significant roles of illness. Doctors can detect, diagnose, and categorize cancer as benign or malignant using machine learning technologies. Although selecting physicians and specialists and evaluating patient data might be difficult, cognitive systems and computational techniques (like machine learning for categorization) will ultimately be helpful to medical practitioners. I am investigating a wide range of methods to learn categorization at the time of writing. I investigated several machine learning approaches utilizing various algorithms and cancer data sets.

1.7 Objectives

This paper's major objective is to test and assess four classifiers in order to choose the best one for predicting cancer at an early stage. In general, I anticipate that the paradigm presented here is clear enough for physicians to comprehend legally. Most of the time, certain performance metrics, including accuracy, can assist us in comparing and selecting the best algorithm. Since this is a subject that is of utmost importance at the moment, I look forward to greater investigation to provide more accurate and dependable results. Here, I highlight certain goals for my work -

1. An overview of breast cancer, including its kinds, risk factors, symptoms, and management.
2. Gathering of data.
3. Using the Wisconsin breast cancer data set, demonstrate Extra Trees Classifier, Ada Boost Classifier, Light Gradient Boosting Machine, Random Forest Classifier, Linear Discriminant Analysis, Ridge Classifier, Quadratic Discriminant Analysis, Extreme Gradient Boosting, Naive Bayes, Gradient Boosting Classifier, Logistic Regression, K Neighbors Classifier, Decision Tree Classifier, SVM - Linear Kernel, Dummy Classifier which is operate in Pycaret classification are machine learning classifiers for predicting breast cancer.
4. To lessen the incidence of fatalities related to breast cancer.

5. Offer the patient suggestions.

1.8 Expected Outcome

This research provides to my present collection of information. exploratory discoveries that could result in new understandings or inspire further study. Results might be quantitative, qualitative, or a combination of both. I should make an effort to make the results as quantitative as possible.

My expected outcomes are -

1. Development of medical site.
2. I will easily predict the disease.
3. Usage of Machine learning the error rate is very low.
4. my physician can take a step at proper time for a Patient.
5. My model willeasily predict Breast Cancer Patient.
6. Give high accuracy by analyzing all the data.

1.9 Challenges

Nowadays, machine learning is becoming more and more popular. Machine learning is a vast area of artificial intelligence, as I all know. I encounter several hurdles every day when I go to work in my industry, and I'm not the only one. The issues i have are outlined below:

1. Since i needed a lot of data for my research and cancer patients are notoriously difficult to locate, it was challenging for me to gather all of their information.
2. With a typical setup device or computer, processing this system was rather challenging.
3. Sometimes, the specific training procedure required a very lengthy time to complete.
4. Finally, i can say that processing procedure is very difficult and very challenging because the technique requires high configured machine.

1.10 Fundamental of Machine Learning

A sub-field of artificial intelligence known as "machine learning" uses a number of statistical, probabilistic, and optimization approaches to help computers "learn" from previous experiences and identify difficult-to-find patterns in vast, noisy, or complicated data sets. Medical applications, particularly those that rely on intricate proteomic and genomic data, are especially well-suited for this capacity. Machine learning is therefore widely utilized in the identification and diagnosis of cancer. Machine learning has more recently been used for cancer prognosis and prediction. A lot of published research also seem to lack the necessary level of testing or validation. Machine learning techniques may be used to significantly (15–25%) increase the accuracy of predicting cancer susceptibility, recurrence, and death, according to the better conducted and verified research. On a more fundamental level, it is also clear that machine learning is assisting in bettering my fundamental comprehension of the onset and course of cancer. Classification, prediction, and regression algorithms frequently employ this kind of machine learning. This kind of learning is likewise covered in areas where the fully labeled coaching technique can generate any labeled value. Face recognition software for digital cameras uses semi-supervised learning. Advanced learning: In accordance with this classification of machine learning, machine learning algorithms build positive actions via trial and error to simplify outcomes and discover uses in artificial intelligence, navigation, and gaming. Artificial intelligence, video games, and navigation all make extensive use of it. The three key factors that influence this machine learning are the agent, the novice, and the environment. In addition, the agent engages in trials and actions. Reinforcement learning's primary goal is to assist the agent in selecting the course of action that will yield the most benefit at the desired moment. The strategy is so obvious: use machine learning to learn simple concepts through reinforcement to achieve the optimal outcomes. Collaborative learning: A method called collaborative filtering is used to create proposals. The most common kind of recommendation system is this one. Users may compare settings between themselves and other users to discover the best goods depending on a variety of parameters. related agonists of the filter domain. The data acquisition structure when the clustering algorithm is unable to detect a certain pre-existing structure is conventional learning. This is the observable learning clustering. Make sure the clustering algorithm can identify the numerous cluster input properties that are naturally present in the data.Large data sets are frequently partitioned into smaller groups using clustering, which entails performing exploratory analysis to tailor the analysis for each group.

Classification: The process of classifying involves supervised learning, which calls for learning from data.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

Although certain tumors may now be treated more readily and individuals may get fully unwell if they receive the right care, cancer is not an entirely immunological illness. Cancer treatments are regularly being developed. Treatment gets simpler and more specialized as time goes on.

The outcomes have greatly improved over time. Radiation therapy, chemotherapy, hormone therapy, and cancer surgery are the primary cancer treatment modalities. There are several immunotherapy techniques employed today, and there appear to be effective medications (or targeted medications) available. There are several anticancer medications. They are generally combined when utilized. One of the cancers that is easiest to treat is breast cancer. There are a ton of amazing therapy options available today that can tackle the intricate cell mix that makes up every cancer.

2.2 Literature Review

Here are many technical reviews and evaluations. articles on how to diagnose breast cancer using data mining techniques. Numerous innovative new technologies have been developed for the diagnosis of breast cancer as a result of the advancement of medical research. Here is a quick summary of the study that is relevant to this area:

S. P. T. Vikas Chourasia.[4] They employed a 10-fold cross-validation approach and a stratified sample strategy to split the dataset into 10 mutually exclusive divisions in order to compare the three algorithms' unbiased prediction accuracy. For each of the three prediction models, they then carried out this procedure once again. This gave them a less biased way to compare the three models' prediction performance. According to the results, Native Bayes provided them with the most accuracy (97.36%), followed by RBF Network (96.77%), which came in second, and J48 (Decision Tree), which was able to provide the third-best accuracy (93.41%). They evaluated the sensitivity and specificity of Native Bayes, RBF Network, and J48 (Decision Tree). Md. MIlon Islam.[5] This research provided a comparative study of machine learning methods for predicting breast cancer prognosis. Of the nine learning algorithms, five have been used. These include logistic regression, K-nearest neighbors, support vector machines, random forests, and artificial neural networks (ANNs) (LR). Each of the five machine learning approaches' fundamental characteristics and operations was demonstrated. In contrast to the lowest accuracy gained from RFs and LR, which is 95.7%, the maximum accuracy attained by ANNs is 98.57%. Medical diagnostic techniques may be time- and resource-intensive, and costly. This system demonstrates the application of machine learning techniques as diagnostic tools for breast cancer, and if the ANN-developed model is better reliable than any of the other strategies described above, and is very good for new physicians or therapists, and can change the prediction to cause breast cancer the scope of cancer. T. B PadmaprIya.[6] Since carcinoma is one of the primary causes of death in women, this study takes clinical facts pertaining to it into consideration. The CART set of rules was chosen for this study's investigation of carcinoma datasets because it provides more accurate clinical information units than the other frequently used decision tree algorithms and CART. Three distinct cancer datasets have been the subject of several trials to see whether the same characteristic choosing method may result in higher accuracy for different datasets within the same domain. For the study of datasets, they specifically employed three well-liked data processing techniques: J48, ADTree, and Cart. The J48 classifiers are 98.1% accurate, Three has the greatest accuracy value of 98.5%, which is found in CART, with a score of 97.7%. It can be seen from the findings that the CART algorithm does a good job of categorizing mammography pictures. ADTree Algorithm performs poorly in this job, with J48 acting as an intermediary. In the future, they'll employ different classification algorithms in an effort to forecast how well they'll do when analyzing similar mammography pictures. G. SIngh.[7] In this study, the classification parameters have been examined across four machine-learning techniques. These include k-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Naive Bayes, all of which are accessible on the Wisconsin breast cancer dataset at the

UCI Machine Learning Repository. This comparison analysis's goal was to identify the most precise machine learning algorithm that may be used to aid in the detection of breast cancer. The best accuracy for the given dataset is achieved by k Nearest Neighbor, which is consistent with the prediction findings. K closest neighbor achieves the necessary performance in their study article in terms of accuracy, which was 0.99(99%) correspondingly. And in terms of accuracy, Logistic Regression achieves the second performance with 0.97 (97%) accordingly. In terms of accuracy, Support Vector Machine and Naive Bayes performed at 0.96 (96%) and 0.97 (95%) respectively. Therefore, this demonstrates that k Nearest Neighbor consistently outperforms SVM, LR, and Naive Bayes for the prediction of breast cancer.[8] This work established the machine learning-based automated diagnostic method for breast cancer. These algorithms operate in three stages. In the first phase, a world-class clustering method was used to separate the data into multiple groups. The dataset is decreased, which drastically cuts down on the computation time. The next phase utilized the Outlier Detection Algorithm to identify outliers in the breast cancer dataset. The J48 method is used to determine if the malignancy is benign or malignant in the prepossessed dataset as the final and third phases. The execution of this system was evaluated using the Wisconsin Diagnosis Breast Cancer and Wisconsin Breast Cancer datasets.These tests were conducted using Waikato Environment for Knowledge Analysis version 3.7.13. The suggested two metrics may be superior than current research for the same data set, according to experimental findings. For WBCD data sets and 99.6% for WDBC data sets, the highest accuracy was 99.9%. This research will aid in the early detection of breast cancer and aid in the treatment of patients. [9] In order to create a hybrid classification method, this work combines the genetic algorithm and the k-nearest neighbor algorithm. It could aid physicians treating early-stage breast cancer in their forecasting. Medical professionals employed data mining techniques to aid in the early diagnosis of breast cancer. Optimizing methods is the Genetic Algorithm's primary goal. While using the k-nearest neighbor technique for classification, it was utilized to choose the higher quality for KNN as well as optimize the value of K. This machine learning dataset was extracted from the Wisconsin Breast Cancer database and is available in the UCI repository. Researchers examined 699 occurrences and 10 traits at WBCD, while just 569 instances and 31 features were examined at WDBC. The suggested algorithm produced superior results for the researchers when compared to the outcomes of other classification algorithms used in this investigation. The proposed method's accuracy was 99%. [10] A hybrid breast cancer model that combines the Decision Tree (DT) and Support Vector Machine (SVM) algorithms is proposed in this paper. This approach was used to divide patients into two groups: benign patients and malignant patients. This is a machine-learning dataset from the UCI repository that was collected from Wisconsin Breast Cancer. It has 699 occurrences and 11 characteristics, 458 of the cases fall under the benign category, and 241 falls under the malignant category. There are 16 occurrences of missing values in this dataset. Using WEKA software, the findings were distinguished using the IBL, SMO, and NAVE classification techniques. The test results demonstrate that DT + SVM outperforms every other method in terms of data classification. The DT-SVM classification model has a 91% accuracy rate. The correctly identified case was number 459, the wrongly classified example was number 240, and the lowest error rate was 2.58%. [11] In this work, data mining is used to predict the occurrence of breast cancer. The goal of this analysis was to create a breast cancer prognostic model that used neural network classification techniques. For possible enhancement, a sorted strategy was employed. Lazy IBK, Tree Random Forest, Lazy K Star Classifier, and NNG rules were subjected to the classification techniques. Data from the WBCD dataset are gathered for this investigation. The WEKA software investigated the observation. There are 286 occurrences in the dataset, 201 of which were benign and 85 of which were malignant. A total of ten variables, including age, tumor size, and class, were used to define these instances. The Tree Random Classifier's classification accuracy was 98% in the end. For 100% accuracy, the researchers advised using Ensemble classified analysis. [12] Breast Cancer Classification Using Machine Learning was given in this study. A collection of tools called "machine learning" was developed and is employed for prediction, evaluation of algorithms, and categorization. Data gathering, model selection, model training, and model testing are the four phases that makeup machine learning. For this categorization, researchers employed two machine learning models. These are the K-nearest neighbor and the Naive Bayesian Classifier. Their goal in employing two classifiers in a data set for this classification is to create efficient machine-learning methods for cancer classification. They brought the provided data set from UCI to the University of California, Irvine as a data set. 35% of the tumors in this dataset are malignant, while 65% of the tumors are benign. The naive Bayesian Classifier provided an accuracy of 96.19% after training, whereas K-nearest Neighbor provided an accuracy of 97.51%. It is accurate enough. Researchers add that if there is an excessive volume of data, accuracy and training data may vary. [13] The researcher attempted to create a brand-new prototype for a clinical issue involving the identification and treatment of breast cancer patients. Ten significant clinical characteristics were chosen, including age, tumor size, and node size. Data from web mining was utilized to uncover hidden patterns in datasets on breast cancer. Data were gathered from the repository of UCI datasets. The WEKA Open-Source environment was used to analyze the experiment. Patients' diagnoses and prognoses for breast cancer were determined using 37 classification criteria. Patients in good or bad health might attend the diagnostic class. The experimental findings showed that 76% of the 37 classifieds were in good health, and 24% were unwell; 13 out of 37 were correctly diagnosed, with 13 more accurate results. Bayes-Net, SMO, Logistics, Multilayer-Perceptron, J48, SGD, Simple-Logistics, AdabostM1, Attribute Selected, and Filtered Classifier were among the thirteen classifiers and Regression, Multi-Class Classification, and LMT. Researchers advise repeating the prototype construction process to forecast the best categorization utilizing the Tanagra and Kamala data mining tools. In the future, academics will use tanager and Orange data mining methods since they wish to create generic prototypes for many domains like e-commerce, power, or many other sectors. [14] The approach under consideration seeks to identify the lowest subset of characteristics that can provide extremely precise categorization. Data preparation, which comes before classification, comprises data cleansing, dimension reduction, and transformation. The WDBC breast cancer dataset employed in the study was taken from the UCI machine learning repository. The dataset consists of 569 instances and 32 characteristics, with 30% of the examples used for testing and 70% used for training. Three classification techniques, including Naive Bayes (NB), Logistic Regression (LR), and Decision Tree (DT), were applied to the test data to determine if the cell was benign or cancerous. The findings demonstrate that the logistic regression classifier offered the most accurate categorization. Comprising a smaller number of characteristics (four), thus this algorithm's complexity is lower than that of the other two classifications. [15] The diagnosis and prognosis of breast cancer were given using machine learning (ML) techniques in this research. The Wisconsin Breast Cancer Database was also employed by the researchers as they concentrated on studies employing artificial neural networks (ANNs), support vector machines (SVM), decision trees (TT), and K-Nearest Neighbor (K-NN) approaches. Using machine learning approaches, categorization and prediction accuracy have increased with great competence. He used the dataset the researchers had supplied, which he had obtained from WBCD. A table including this information, along with references, algorithms, sample methods, and accuracy, was displayed. Creating more sophisticated algorithms is still required, even though researchers think several algorithms employing Wisconsin have achieved exceptionally high accuracy of the Breast Cancer Database (WBCD). Going forward, Researchers intend to thoroughly examine the Frederick Ataxia (FRDA) dataset using machine learning techniques to create an effective FRDA healthcare system.

CHAPTER 3

METHODOLOGY

3.1 **Introduction**

Carcinoma is the most common cancer in women in industrialized and developing nations. It's also the second most important factor in women dying from cancer. According to a WHO analysis from 2013, "it is projected that over 508 000 women globally died in 2011 as a result of breast cancer." The majority of women (97%) might live for at least five years with early diagnosis. In the diagnosis and prognosis of cancer, data processing, and machine learning are frequently used.Additionally, data processing and machine learning help medical researchers identify links between factors and prepare them to forecast illness outcomes using previous records. Machine learning may be used to improve cancer detection and diagnosis and minimize over treatment. Additionally, it could support correct higher cognitive function. As a result, this study aims to examine how machine learning and data processing techniques are used in the identification and diagnosis of cancer.

3.2 Methodology

I am using seven fundamental machine learning methods for my classification problem. Those are:

1. LogisticRegression
2. KNN
3. Decision Tree Classifier
4. Random Forest
5. Support Vector Machine
6. Naive Bayes
7. Artificial neural networks (ANNs)

Logistic Regression: Logistic regression is the most well-known machine learning technique after linear regression. Although they are highly distinct, logistic regression and linear regression have many similarities. To forecast the value, a linear regression approach is employed. Accepting input data and goal variables may transform logistics rules into supervised rules for training models. In contrast to the regression of the protection average, the output or target variable in the logic rule might be a categorical variable. As a result, this is a rule of binary categorization that places knowledge objects in one of the information categories.

The known logistic regression equation is

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

K Neighbors Classifier: It is the non-parametric slow method known as the K-nearest neighbor. The nearest neighbors are chosen based on the Euclidean distance between the equation's x and y vectors.

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

KNN outcomes vary for various K values. A high K value will overlap, whereas a low K value will result in more calculations.

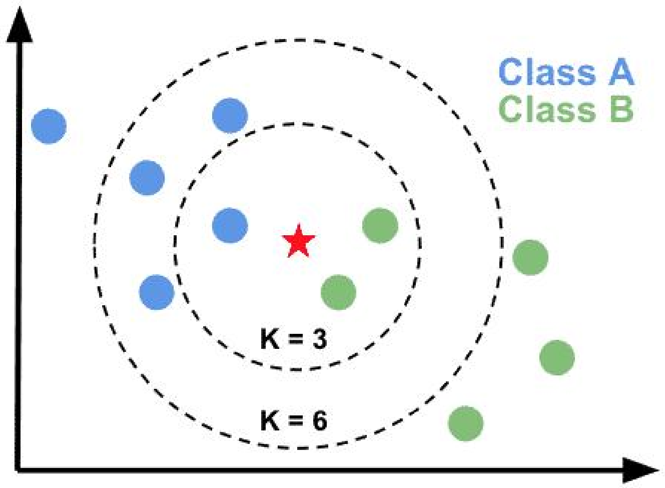


Figure 3.1: K Neighbor

Decision Tree Classifier: Introduction Decision Trees are a sort of supervised machine learning in which the training data is continually segmented based on a particular parameter, with you describing the input and the associated output. Decision nodes and leaves are the two components that may be used to explain the tree.

Random Forest Classifier: Supervised machine learning algorithms like random forest are frequently employed in classification and regression issues. On various samples, it constructs decision trees and uses their average for classification and the majority case for regression.

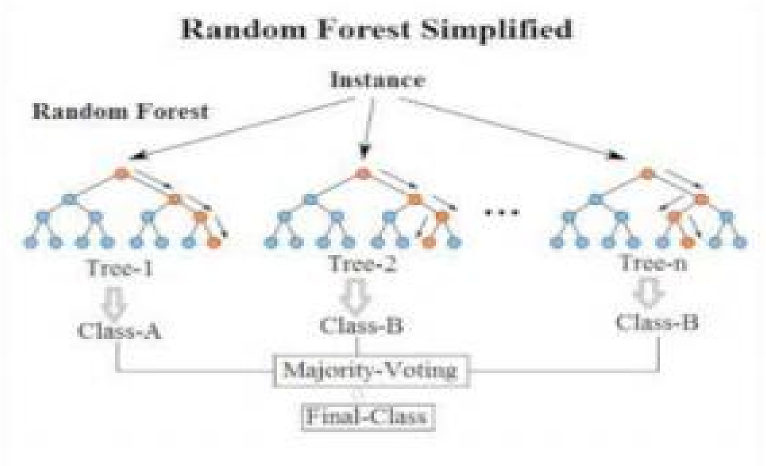


Figure 3.2: Random Forest

Support Vector Machine

The Support Vector Machine (SVM) technique is a supervised machine learning technique used for regression and classification. Even if I also refer to regression issues, classification is the best fit. The SVM classifier seeks a hyperplane in an N-dimensional space that is classify the data points.

Naive Bayes: The naive Bayes algorithm is a supervised technique to solve prediction classification issues based on the Bayes theorem. It is mostly employed in text categorization with a large training set.

Artificial neural networks (ANNs)

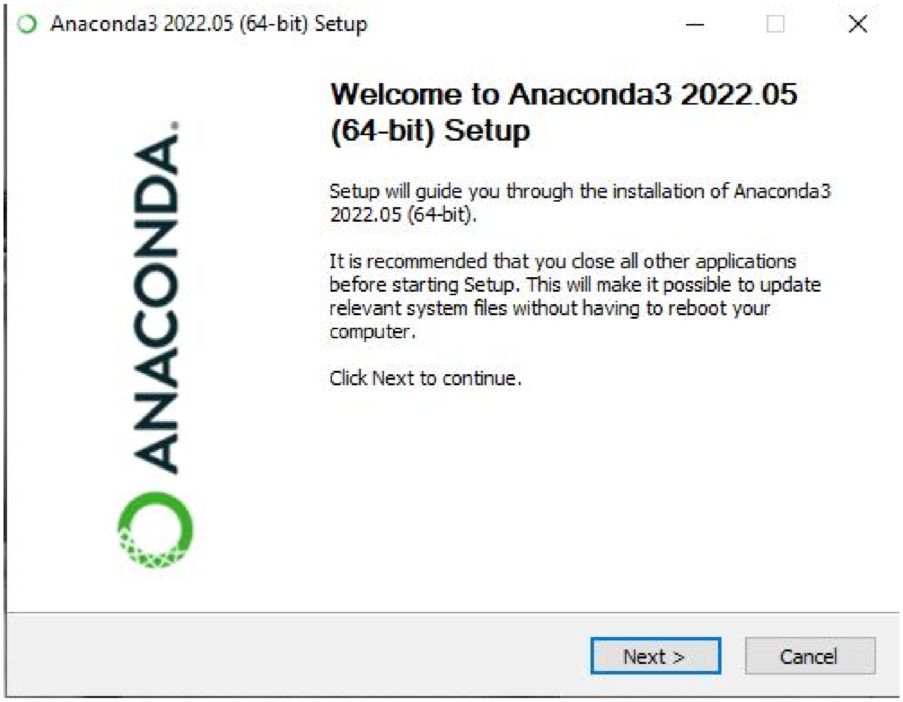
Artificial Neural Networks (ANN) are brain-inspired algorithms that foresee problems and model complex patterns. The Artificial Neural Network (ANN) is a deep learning technique inspired by the biological neural networks in the human brain.

3.3 Visual environment setup and installation

Installing Anaconda: Python and R are both available for free in Anaconda. It's a specific type of open-source platform that is free. Additionally, it makes deployment as well as package and library management easier. It is applied to machine learning.Data scientists, IT professionals, and business leaders are a few examples.

Why I use Anaconda?

It is free and open source to start. It also contains more than 1500 Python packages. Because AI and machine learning have more tools, I can acquire more data from more sources. Testing, development, and training are all done using the same equipment.



Anaconda3 2022.05 (64-bit) Setup

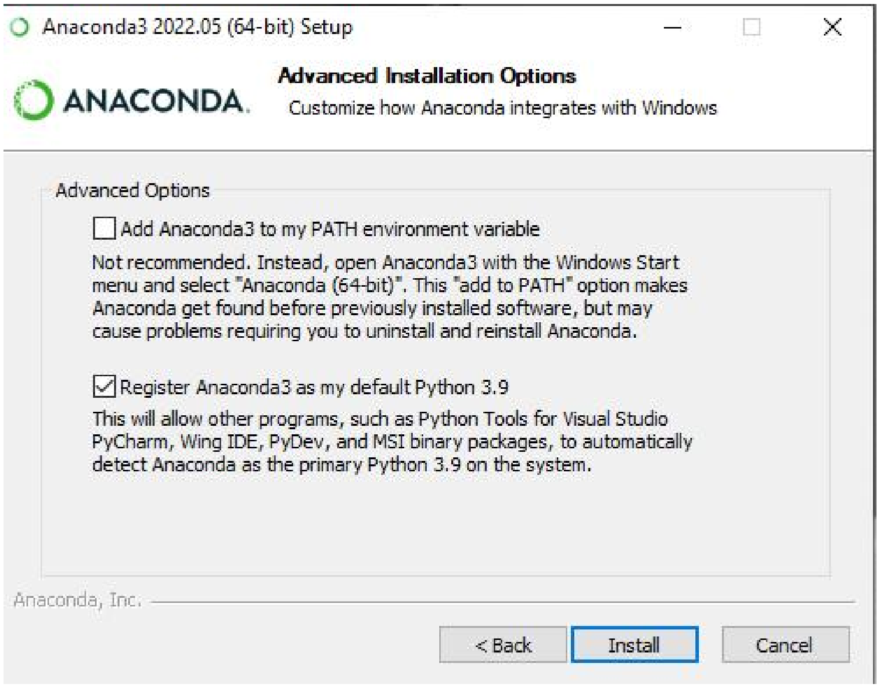
Welcome to Anaconda3 2022.05 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2022.05 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Figure 3.3: Installing Anaconda (1st step)



Advanced Installation Options

Customize how Anaconda integrates with Windows

Advanced Options

□ Add Anaconda3 to my PATH environment variable

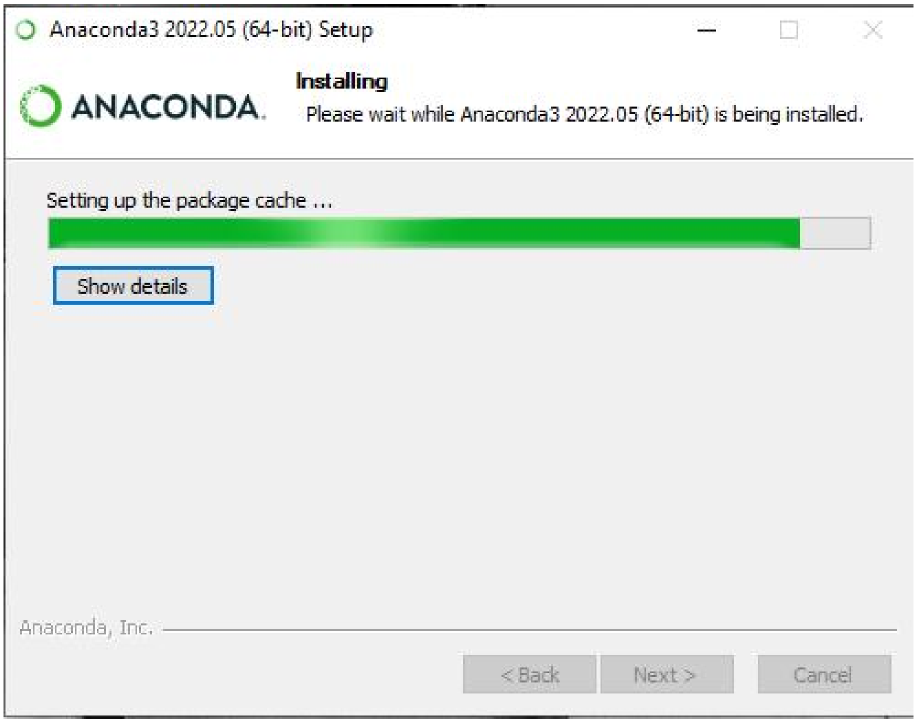
Not recommended. Instead, open Anaconda3 with the Windows Start menu and select "Anaconda (64-bit)". This "add to PATH" option makes Anaconda get found before previously installed software, but may cause problems requiring you to uninstall and reinstall Anaconda.

☑ Register Anaconda3 as my default Python 3.9

This will allow other programs, such as Python Tools for Visual Studio PyCharm, Wing IDE, PyDev, and MSI binary packages, to automatically detect Anaconda as the primary Python 3.9 on the system.

Figure 3.4: Select Anaconda environment (2nd step)

Here I can select or needed environment.



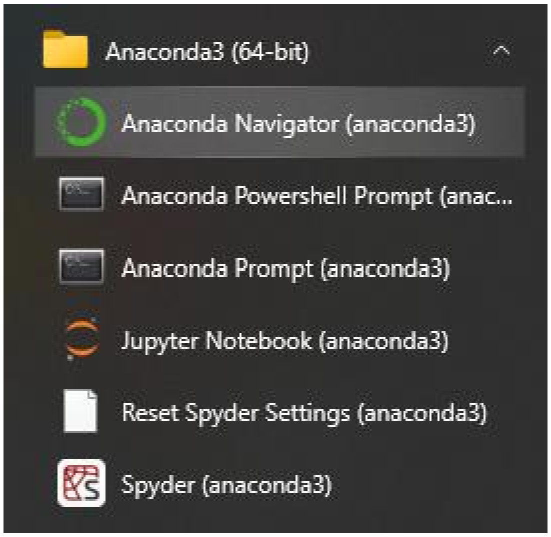
Installing

Please wait while Anaconda3 2022.05 (64-bit) is being installed.

Setting up the package cache ...

Figure 3.5: Installing Anaconda (Final step)

Datalore was launched once Anaconda had been installed. I know that the Anaconda is too sociable for us to use here. Since it has given us many functionalities found in machine learning software. Datalore is one of them.



Anaconda3 (64-bit)

Anaconda Navigator (anaconda3)

Anaconda Powershell Prompt (anac...

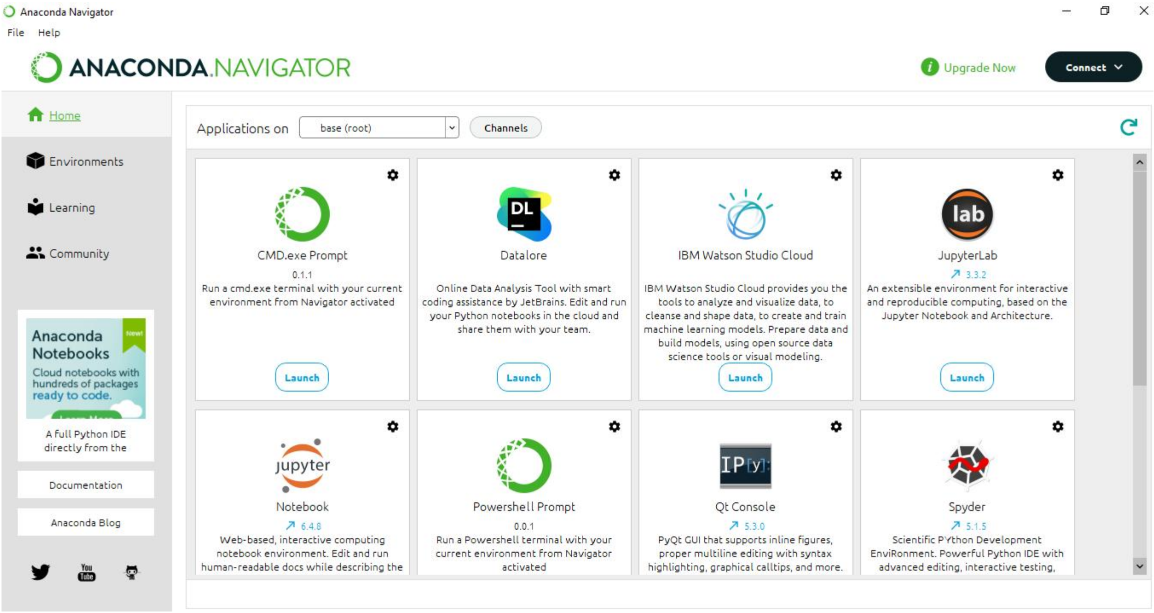
Anaconda Prompt (anaconda3)

Jupyter Notebook (anaconda3)

Reset Spyder Settings (anaconda3)

Spyder (anaconda3)

Figure 3.6: Platform of Anaconda



Anaconda Notebooks New!

Cloud notebooks with hundreds of packages ready to code.

A full Python IDE directly from the

CMD.exe Prompt 0.1.1
Run a cmd.exe terminal with your current environment from Navigator activated
Launch

Datalore
Online Data Analysis Tool with smart coding assistance by JetBrains. Edit and run your Python notebooks in the cloud and share them with your team.
Launch

IBM Watson Studio Cloud
IBM Watson Studio Cloud provides you the tools to analyze and visualize data, to cleanse and shape data, to create and train machine learning models. Prepare data and build models, using open source data science tools or visual modeling.
Launch

JupyterLab 3.3.2
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch

Notebook 6.4.8
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the

Powershell Prompt 0.0.1
Run a Powershell terminal with your current environment from Navigator activated

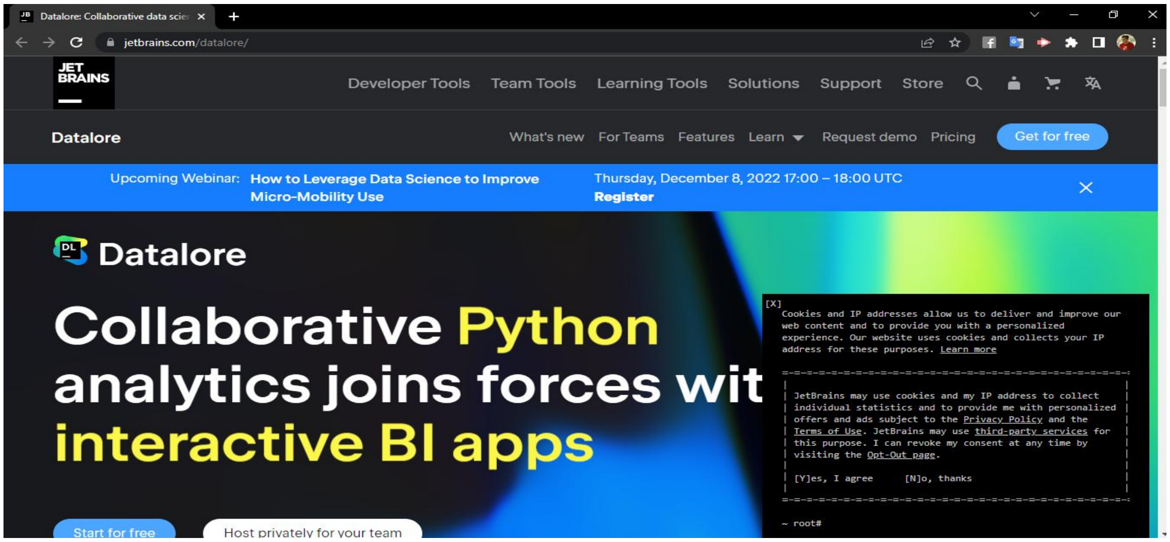
Qt Console 5.3.0
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.

Spyder 5.1.5
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing,

Figure 3.7: Interface of Anaconda

Opening Datalore:

Then open Datalore



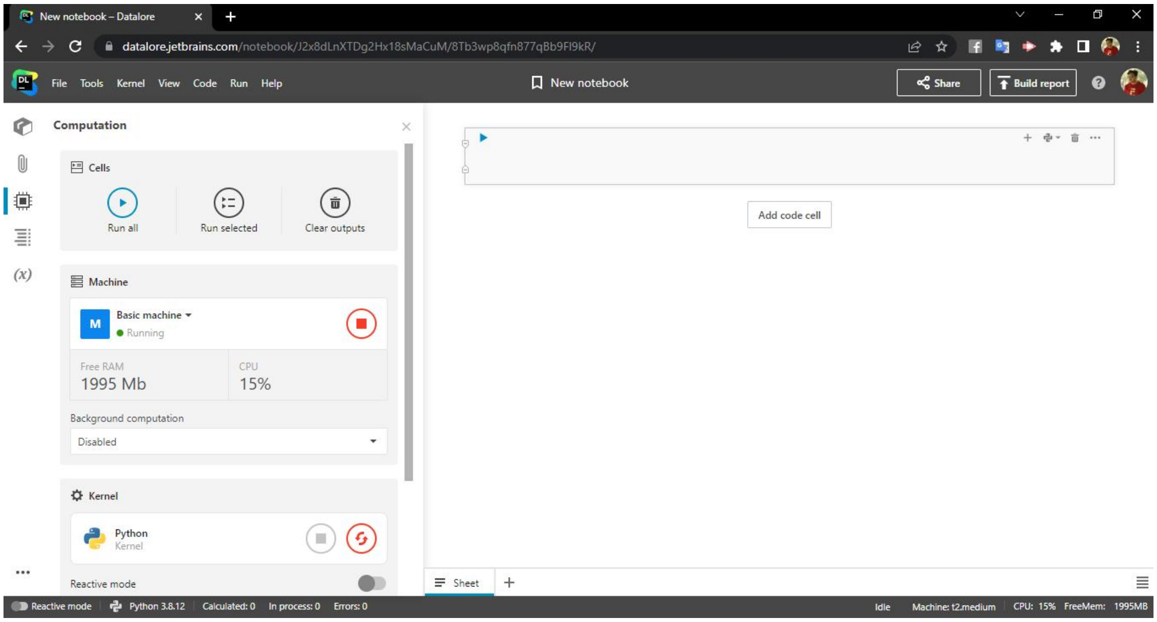
Upcoming Webinar: **How to Leverage Data Science to Improve Micro-Mobility Use** Thursday, December 8, 2022 17:00 – 18:00 UTC **Register**

Datalore

Collaborative Python analytics joins forces wit interactive BI apps

```
[X]
   Cookies and IP addresses allow us to deliver and improve our
   web content and to provide you with a personalized
   experience. Our website uses cookies and collects your IP
   address for these purposes. Learn more

   =-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-:
    |                                                          |
    | JetBrains may use cookies and my IP address to collect   |
    | individual statistics and to provide me with personalized |
    | offers and ads subject to the Privacy Policy and the     |
    | Terms of Use. JetBrains may use third-party services for |
    | this purpose. I can revoke my consent at any time by     |
    | visiting the Opt-Out page.                               |
    |                                                          |
    | [Y]es, I agree     [N]o, thanks                          |
    |                                                          |
   =-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-:

   ~ root#
```

Figure 3.8: Opening Datalore



Basic machine
Running

Free RAM 1995 Mb

CPU 15%

Background computation: Disabled

Python Kernel

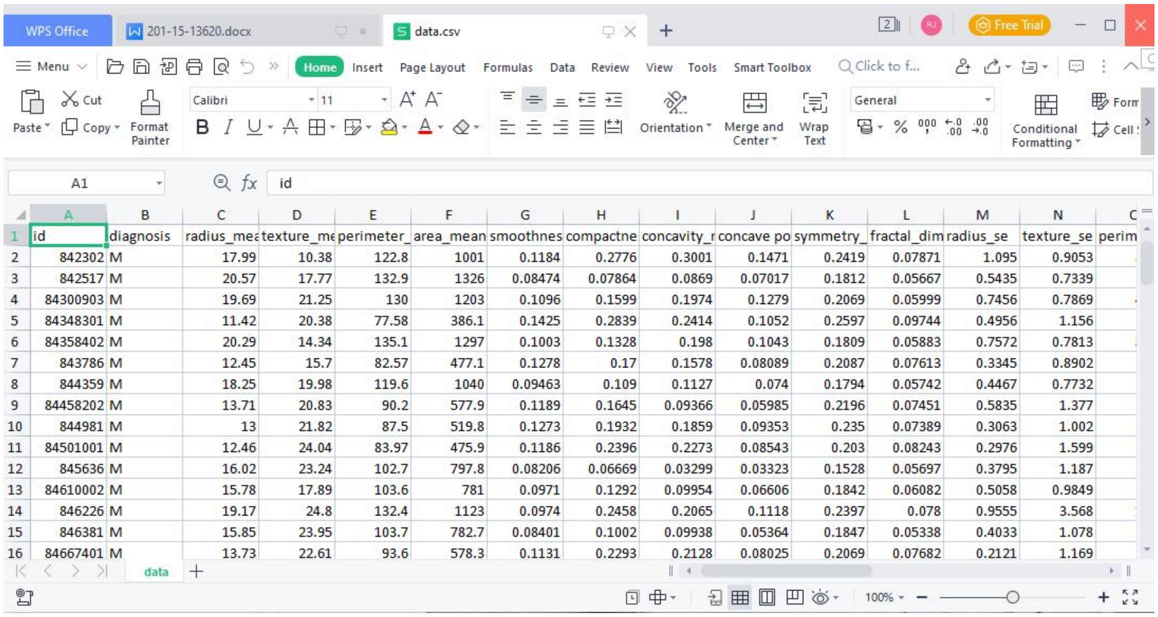
Reactive mode

Reactive mode | Python 3.8.12 | Calculated: 0 | In process: 0 | Errors: 0 | Idle | Machine: t2.medium | CPU: 15% | FreeMem: 1995MB

Figure 3.9: Interface of DataLore

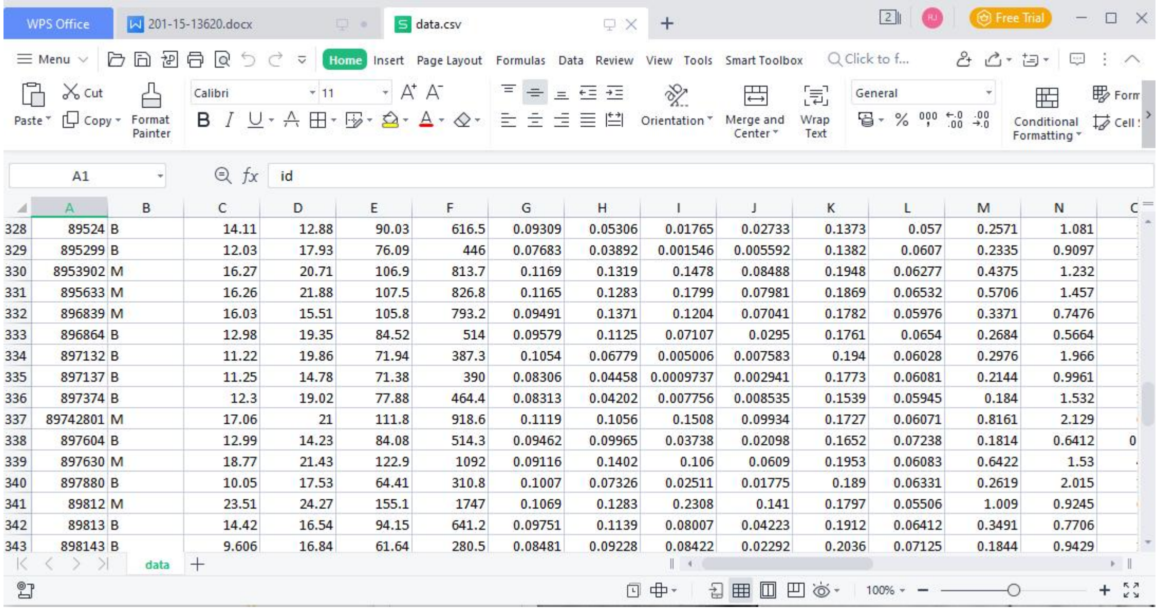
After launching Datalore, I built my python file and got to work on my workflow.

Sample of Data:



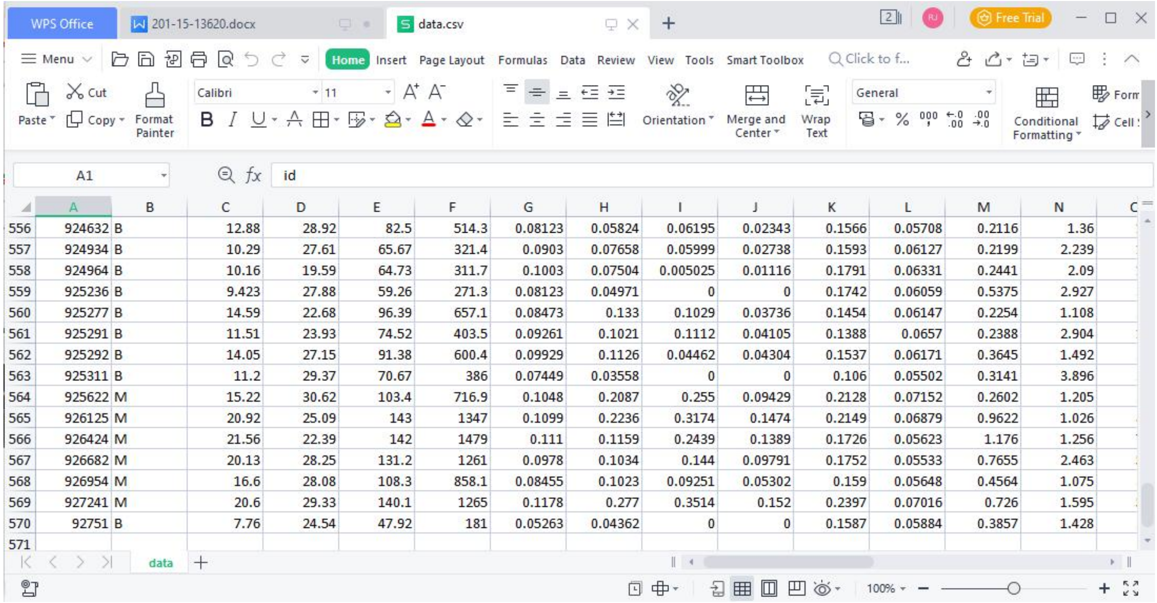
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	C
1	id	diagnosis	radius_mea	texture_me	perimeter_	area_mean	smoothnes	compactne	concavity_r	concave po	symmetry_	fractal_dim	radius_se	texture_se	perim
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	
15	846381	M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338	0.4033	1.078	
16	84667401	M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682	0.2121	1.169	

Figure 3.10: Sample of Data.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	C
328	89524	B	14.11	12.88	90.03	616.5	0.09309	0.05306	0.01765	0.02733	0.1373	0.057	0.2571	1.081	
329	895299	B	12.03	17.93	76.09	446	0.07683	0.03892	0.001546	0.005592	0.1382	0.0607	0.2335	0.9097	
330	8953902	M	16.27	20.71	106.9	813.7	0.1169	0.1319	0.1478	0.08488	0.1948	0.06277	0.4375	1.232	
331	895633	M	16.26	21.88	107.5	826.8	0.1165	0.1283	0.1799	0.07981	0.1869	0.06532	0.5706	1.457	
332	896839	M	16.03	15.51	105.8	793.2	0.09491	0.1371	0.1204	0.07041	0.1782	0.05976	0.3371	0.7476	
333	896864	B	12.98	19.35	84.52	514	0.09579	0.1125	0.07107	0.0295	0.1761	0.0654	0.2684	0.5664	
334	897132	B	11.22	19.86	71.94	387.3	0.1054	0.06779	0.005006	0.007583	0.194	0.06028	0.2976	1.966	
335	897137	B	11.25	14.78	71.38	390	0.08306	0.04458	0.0009737	0.002941	0.1773	0.06081	0.2144	0.9961	
336	897374	B	12.3	19.02	77.88	464.4	0.08313	0.04202	0.007756	0.008535	0.1539	0.05945	0.184	1.532	
337	89742801	M	17.06	21	111.8	918.6	0.1119	0.1056	0.1508	0.09934	0.1727	0.06071	0.8161	2.129	
338	897604	B	12.99	14.23	84.08	514.3	0.09462	0.09965	0.03738	0.02098	0.1652	0.07238	0.1814	0.6412	0
339	897630	M	18.77	21.43	122.9	1092	0.09116	0.1402	0.106	0.0609	0.1953	0.06083	0.6422	1.53	
340	897880	B	10.05	17.53	64.41	310.8	0.1007	0.07326	0.02511	0.01775	0.189	0.06331	0.2619	2.015	
341	89812	M	23.51	24.27	155.1	1747	0.1069	0.1283	0.2308	0.141	0.1797	0.05506	1.009	0.9245	
342	89813	B	14.42	16.54	94.15	641.2	0.09751	0.1139	0.08007	0.04223	0.1912	0.06412	0.3491	0.7706	
343	898143	B	9.606	16.84	61.64	280.5	0.08481	0.09228	0.08422	0.02292	0.2036	0.07125	0.1844	0.9429	

Figure 3.11: Sample of Data 2

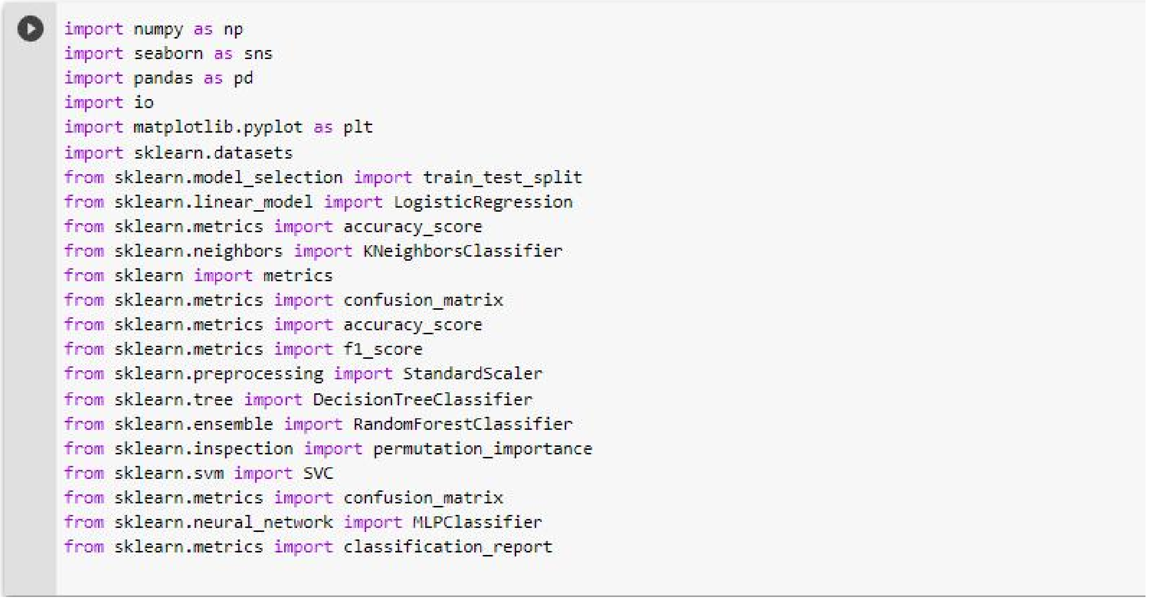


A1 | fx | id

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
556	924632	B	12.88	28.92	82.5	514.3	0.08123	0.05824	0.06195	0.02343	0.1566	0.05708	0.2116	1.36
557	924934	B	10.29	27.61	65.67	321.4	0.0903	0.07658	0.05999	0.02738	0.1593	0.06127	0.2199	2.239
558	924964	B	10.16	19.59	64.73	311.7	0.1003	0.07504	0.005025	0.01116	0.1791	0.06331	0.2441	2.09
559	925236	B	9.423	27.88	59.26	271.3	0.08123	0.04971	0	0	0.1742	0.06059	0.5375	2.927
560	925277	B	14.59	22.68	96.39	657.1	0.08473	0.133	0.1029	0.03736	0.1454	0.06147	0.2254	1.108
561	925291	B	11.51	23.93	74.52	403.5	0.09261	0.1021	0.1112	0.04105	0.1388	0.0657	0.2388	2.904
562	925292	B	14.05	27.15	91.38	600.4	0.09929	0.1126	0.04462	0.04304	0.1537	0.06171	0.3645	1.492
563	925311	B	11.2	29.37	70.67	386	0.07449	0.03558	0	0	0.106	0.05502	0.3141	3.896
564	925622	M	15.22	30.62	103.4	716.9	0.1048	0.2087	0.255	0.09429	0.2128	0.07152	0.2602	1.205
565	926125	M	20.92	25.09	143	1347	0.1099	0.2236	0.3174	0.1474	0.2149	0.06879	0.9622	1.026
566	926424	M	21.56	22.39	142	1479	0.111	0.1159	0.2439	0.1389	0.1726	0.05623	1.176	1.256
567	926682	M	20.13	28.25	131.2	1261	0.0978	0.1034	0.144	0.09791	0.1752	0.05533	0.7655	2.463
568	926954	M	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	0.05302	0.159	0.05648	0.4564	1.075
569	927241	M	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	0.152	0.2397	0.07016	0.726	1.595
570	92751	B	7.76	24.54	47.92	181	0.05263	0.04362	0	0	0.1587	0.05884	0.3857	1.428
571														

Figure 3.12: Sample of Data 3

3.4 Importing Necessary Libraries



```
import numpy as np
import seaborn as sns
import pandas as pd
import io
import matplotlib.pyplot as plt
import sklearn.datasets
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import f1_score
from sklearn.preprocessing import StandardScaler
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.inspection import permutation_importance
from sklearn.svm import SVC
from sklearn.metrics import confusion_matrix
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report
```

Figure 3.13: Importing Liberties.

3.4.1 Pandas

Pandas is a Python open-source library mostly utilized for machine learning and data science jobs. It is built upon NumPy, a different library that supports multi-dimensional arrays.

3.4.2 NumPy

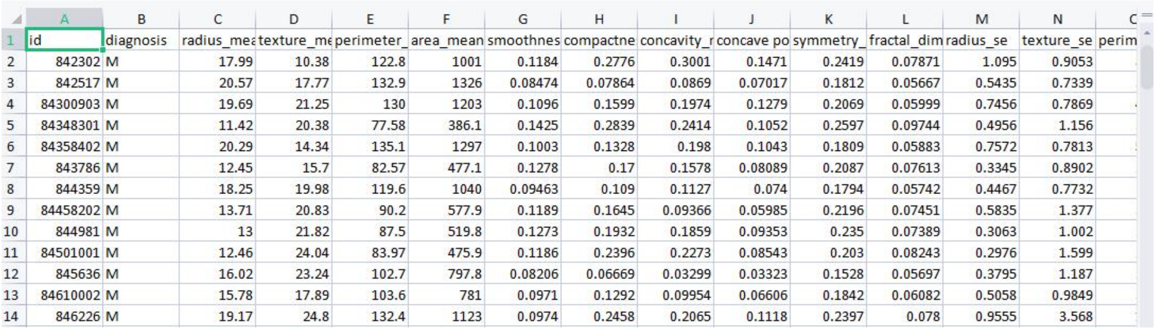
The Python package NumPy is used to manipulate arrays. Additionally, it contains matrices, Fourier transform, and functions for working in linear algebra. In the year 2005, Travis Oliphant developed NumPy. You can use it for free because it is an open-source project.

3.4.3 Matplotlib

A Python charting library is known as Matplotlib. A comprehensive library for animated, static, and interactive visualization in Python is another name for it.

3.4.4 CSV

To structure the tabular data, CSV is employed. Data from an excel file that this library has used as input.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	C
1	id	diagnosis	radius_mea	texture_me	perimeter_	area_mean	smoothnes	compactne	concavity_r	concave po	symmetry_	fractal_dim	radius_se	texture_se	perim
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	
11	84501001	M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243	0.2976	1.599	
12	845636	M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697	0.3795	1.187	
13	84610002	M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082	0.5058	0.9849	
14	846226	M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078	0.9555	3.568	

Figure 3.14: CSV Data.

3.4.5 Seaborn

Seaborn's data visualization library is constructed in Python on top of Matplotlib. It allows users to create complex and beautiful visualizations.

3.4.6 Sklearn (Scikit-learn)

With the Python programming language, this library is utilized as a machine learning library. It can provide features for myimage processing in addition to classification, the K-means clustering approach, and DBSCAN. This library is designed to work with Python's NumPy and SciPy scientific and numerical libraries.

3.5 Dataset Preparing & Analyzing

3.5.1 Dataset

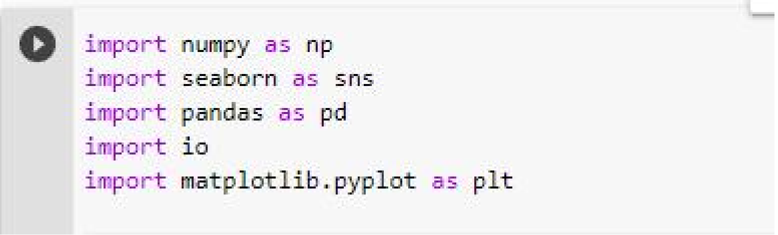
The Wisconsin Breast Cancer (Diagnostic) Data Set includes perhaps the Breast Cancer Dataset (BCD) that I used [16]. There are 32 characteristics, the first of which is the ID I wish to delete (rather than the feature I want to give the category). There are 31 requirements in my categorization section to evaluate whether a tumor is benign. The last characteristic has a binary value of either good or bad (0 for benign, 1 for malign).569 clinical cases are included in this dataset. My data set is restricted to 569 samples because the original BCD only had 32 unidentified values and no missing observation data. Without including target properties, mydataset's attributes are listed below-

Table 3.1: Dataset Attributes

1.radius (mean)	16.compactness (standard error)
2.texture (mean)	17.concavity (standard error)

3.perimeter (mean)	18.concave points (standard error)
4.area (mean)	19.symmetry (standard error)
5.smoothness (mean)	20.fractal dimension (standard error)
6.compactness (mean)	21.radius (worst)
7.concavity (mean)	22.texture (worst)
8.concave points (mean)	23.perimeter (worst)
9.symmetry (mean)	24.area (worst)
10.fractal dimension (mean)	25.smoothness (worst)
11.radius (standard error)	26.compactness (worst)
12.texture (standard error)	27.concavity (worst)
13.perimeter (standard error)	28.concave points (worst)
14.area (standard error)	29.symmetry (worst)
15.smoothness (standard error)	30.fractal dimension (worst)

The first step in using Python is to import the necessary packages. Additionally, I have already imported all of the required libraries for gathering and analyzing data.

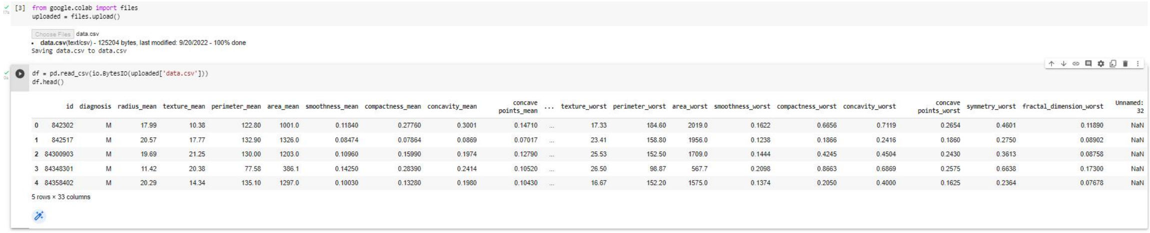


```
import numpy as np
import seaborn as sns
import pandas as pd
import io
import matplotlib.pyplot as plt
```

Figure 3.15: Initially Import Libraries.

3.5.2 Load Data

I first need to use pandas to import data.



```
[3] from google.colab import files
    uploaded = files.upload()
```

data.csv
- **data.csv**(text/csv) - 125204 bytes, last modified: 9/20/2022 - 100% done

```
Saving data.csv to data.csv
```

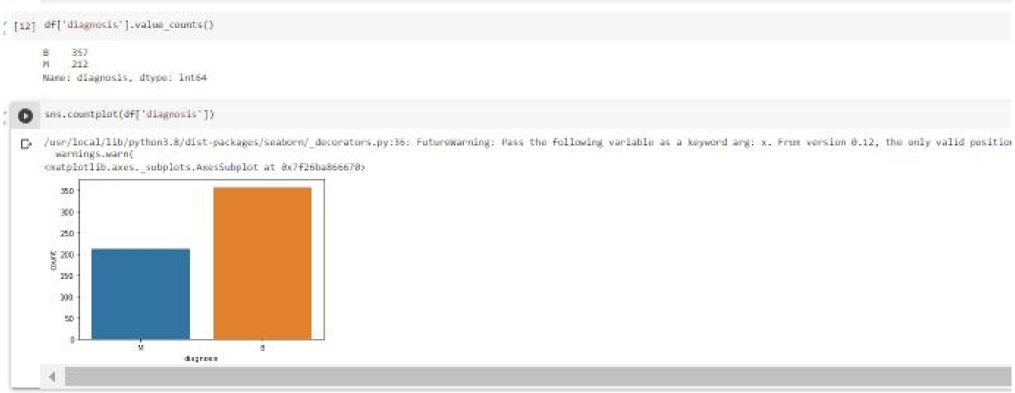
```
df = pd.read_csv(io.BytesIO(uploaded['data.csv']))
df.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	texture_worst	perimeter_worst	area_worst	smoothness_worst	compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst	Unnamed: 32
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	17.33	184.60	2019.0	0.1622	0.6656	0.7119	0.2654	0.4601	0.11890	NaN
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	23.41	158.80	1956.0	0.1238	0.1866	0.2416	0.1860	0.2750	0.08902	NaN
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	25.53	152.50	1709.0	0.1444	0.4245	0.4504	0.2430	0.3613	0.08758	NaN
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	26.50	98.87	567.7	0.2098	0.8663	0.6869	0.2575	0.6638	0.17300	NaN
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	16.67	152.20	1575.0	0.1374	0.2050	0.4000	0.1625	0.2364	0.07678	NaN

5 rows × 33 columns

Figure 3.16: Exploring my data.

After importing the data, I gather information about my dataset's number of features. And I'm looking for the missing value here. What datatype does each feature have? By analyzing this dataset, I got 569 rows and 31 columns here. There isn't a value missing. In each feature, float 64. Now that I've looked at my dataset, I see two different patient categories, one of which is malignant and the other benign.



```
[12] df['diagnosis'].value_counts()
```

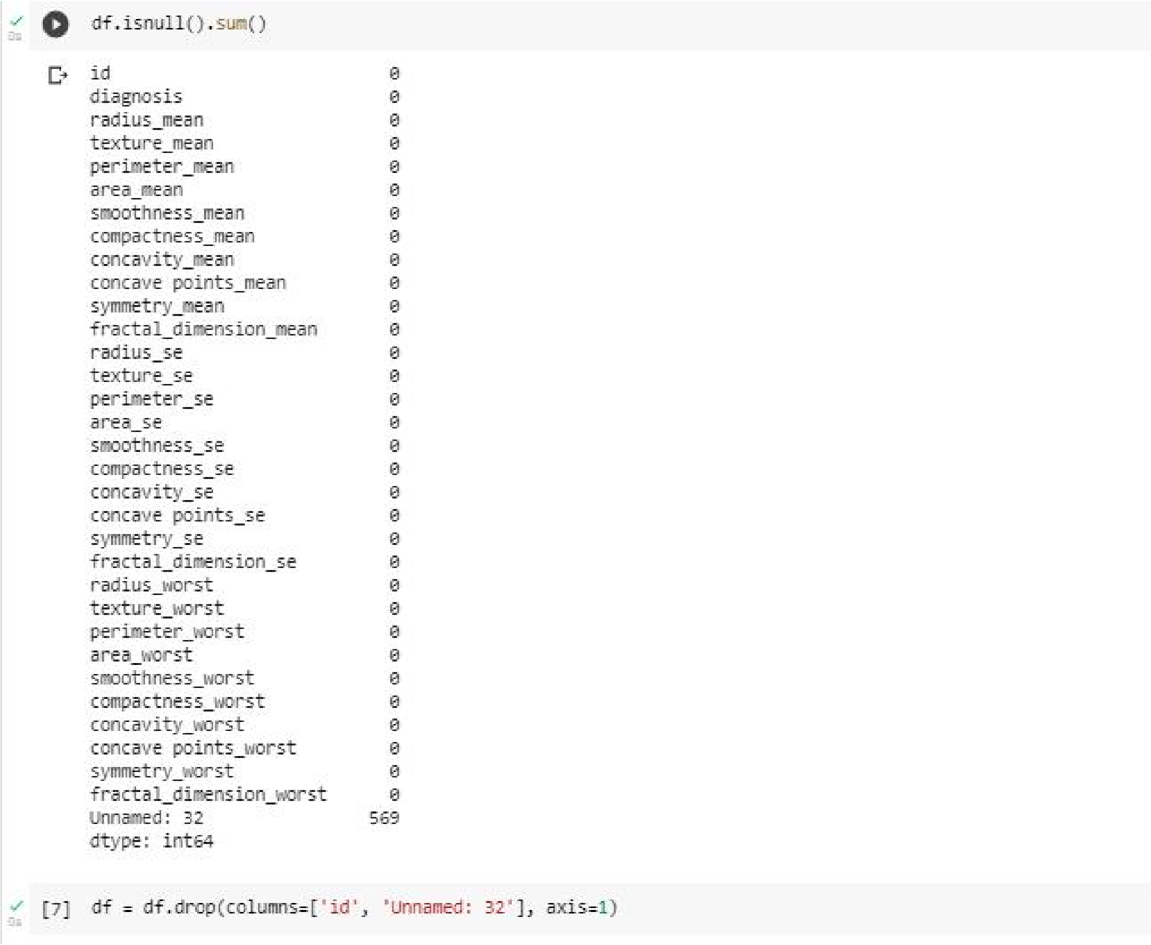
```
B    357
M    212
Name: diagnosis, dtype: int64
```

```
sns.countplot(df['diagnosis'])
```

```
/usr/local/lib/python3.8/dist-packages/seaborn/_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positio
  warnings.warn(
<matplotlib.axes._subplots.AxesSubplot at 0x7f26ba866670>
```

Figure 3.17: Target Column.

According to Figure 3.5.2.2, 357 tumors are benign and 212 are malignant.
But here I found an unnamed column, so now I have to drop this column unmanned.



```
df.isnull().sum()
```

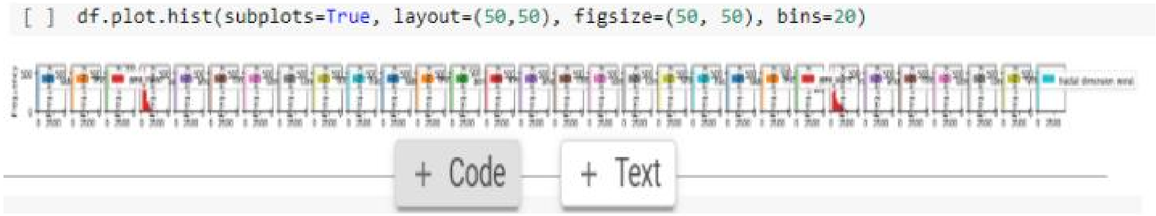
```
id                           0
diagnosis                    0
radius_mean                  0
texture_mean                 0
perimeter_mean               0
area_mean                    0
smoothness_mean              0
compactness_mean             0
concavity_mean               0
concave points_mean          0
symmetry_mean                0
fractal_dimension_mean       0
radius_se                    0
texture_se                   0
perimeter_se                 0
area_se                      0
smoothness_se                0
compactness_se               0
concavity_se                 0
concave points_se            0
symmetry_se                  0
fractal_dimension_se         0
radius_worst                 0
texture_worst                0
perimeter_worst              0
area_worst                   0
smoothness_worst             0
compactness_worst            0
concavity_worst              0
concave points_worst         0
symmetry_worst               0
fractal_dimension_worst      0
Unnamed: 32                569
dtype: int64
```

```
[7] df = df.drop(columns=['id', 'Unnamed: 32'], axis=1)
```

Figure 3.18: drop unnamed column.

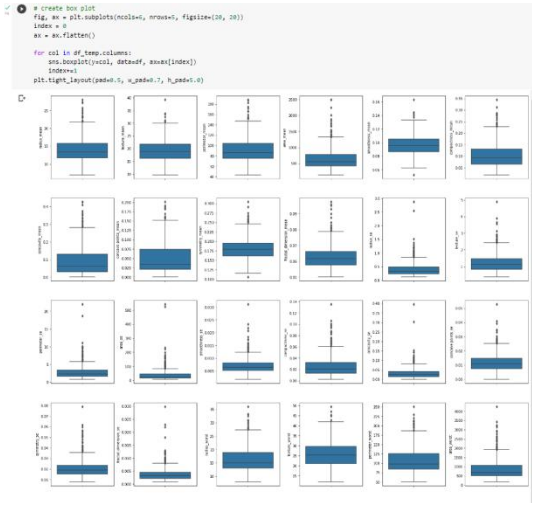
3.5.3 Creating plot

In this context, I will discuss how to create a plot with feature and value of this dataset. There are many plot types such as scatter plot, line plot, dot plot, dist plot, box plot, hist plot, etc. Some of them are -



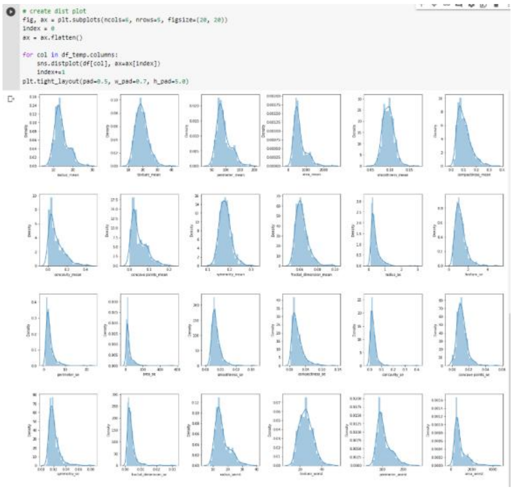
```
[ ]  df.plot.hist(subplots=True, layout=(50,50), figsize=(50, 50), bins=20)
```

Figure 3.19: hist plot



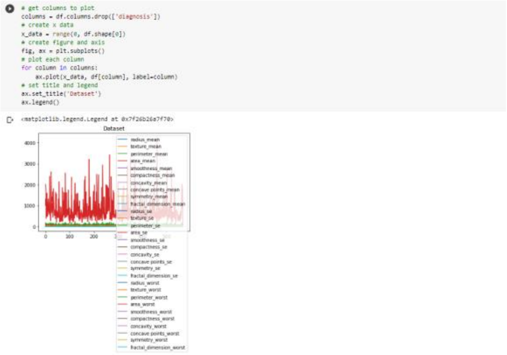
```
# create box plot
fig, ax = plt.subplots(ncols=6, nrows=5, figsize=(20, 20))
index = 0
ax = ax.flatten()

for col in df_temp.columns:
    sns.boxplot(y=col, data=df, ax=ax[index])
    index+=1
plt.tight_layout(pad=0.5, w_pad=0.7, h_pad=5.0)
```

Figure 3.20: box plot



```
# create dist plot
fig, ax = plt.subplots(ncols=6, nrows=5, figsize=(20, 20))
index = 0
ax = ax.flatten()

for col in df_temp.columns:
    sns.distplot(df[col], ax=ax[index])
    index+=1
plt.tight_layout(pad=0.5, w_pad=0.7, h_pad=5.0)
```

Figure 3.21: dist plot



```
# get columns to plot
columns = df.columns.drop(['diagnosis'])
# create x data
x_data = range(0, df.shape[0])
# create figure and axis
fig, ax = plt.subplots()
# plot each column
for column in columns:
    ax.plot(x_data, df[column], label=column)
# set title and legend
ax.set_title('Dataset')
ax.legend()
```

```
<matplotlib.legend.Legend at 0x7f26b26a7f70>
```

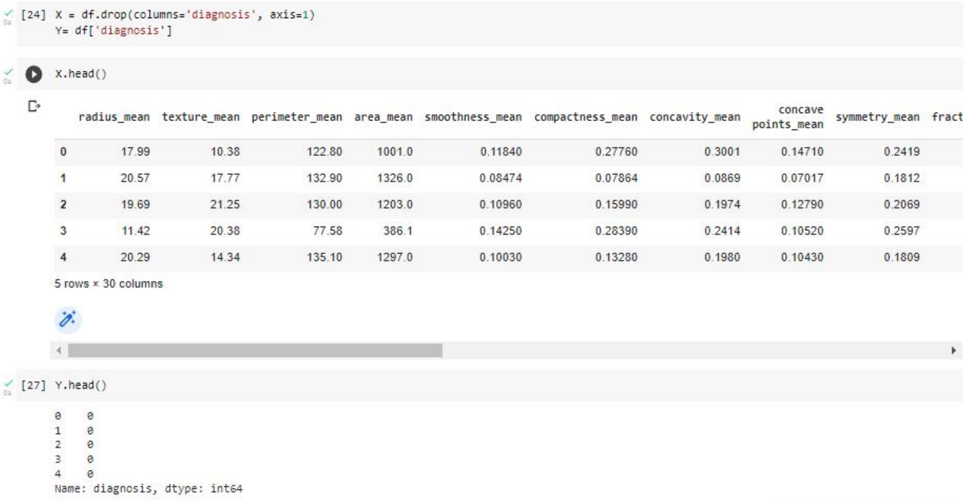
Figure 3.22: legend plot

Co-relation Matrix

The correlation matrix is essentially a table indicating the correlation. This measurement works best when the variables are in a linear relationship. A scatter chart can display the consistency of the data. The variables are shown in rows and columns of the correlation matrix.

3.6 Train Methodology

I'll talk about the training methods in this aspect. The data I input for learning will be understood by the computer during training. So here is a description of how to train, how the trained data will be used, and how it will be processed. I must first classify the variables into dependent and independent categories to train the model.



```
[24] X = df.drop(columns='diagnosis', axis=1)
     Y= df['diagnosis']
```

```
X.head()
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fract
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	

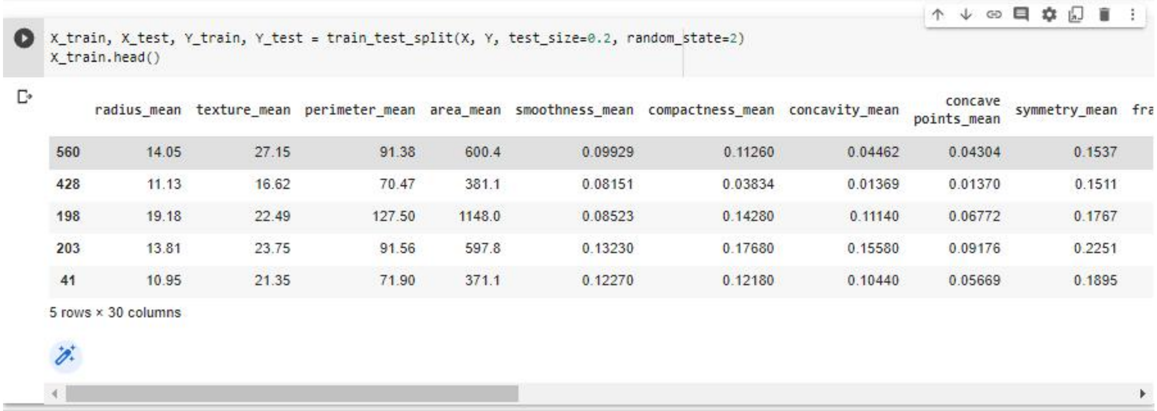
5 rows × 30 columns

```
[27] Y.head()
```

```
0    0
1    0
2    0
3    0
4    0
Name: diagnosis, dtype: int64
```

Figure 3.23: Independent and Depended variable.

My dataset has to be split into training and testing halves. Additionally, a training_test_split library is needed for the divide.



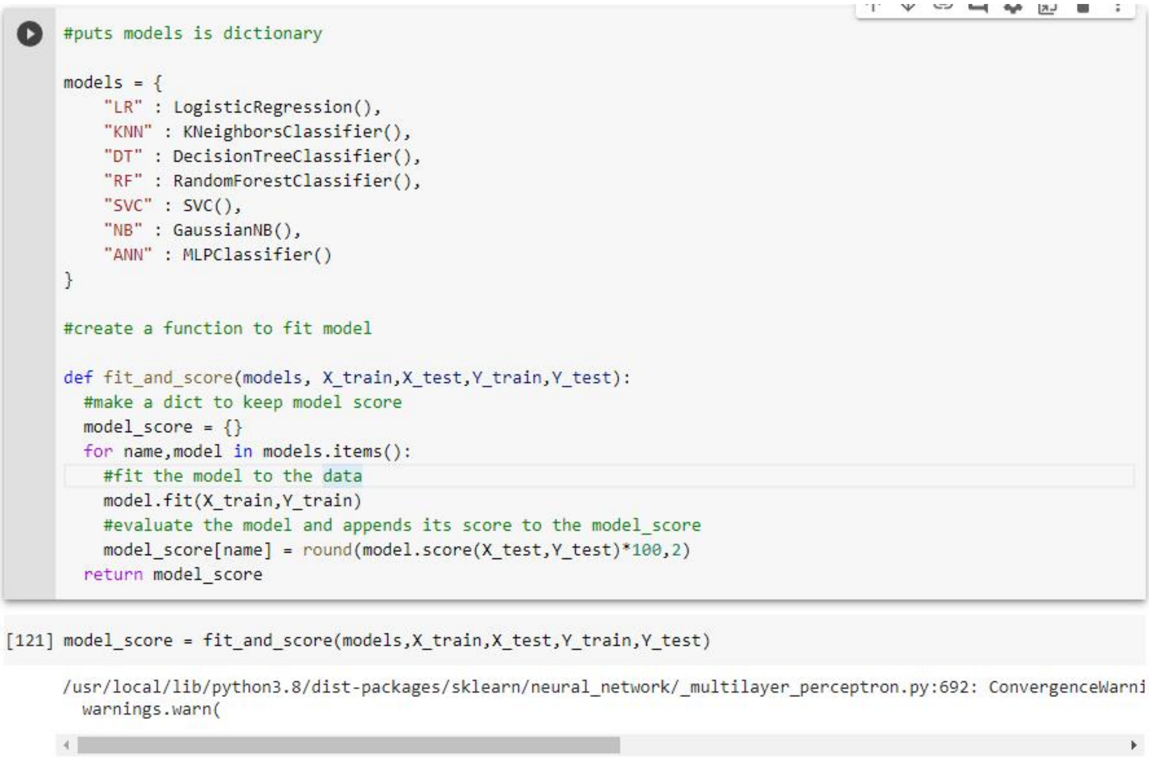
```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
X_train.head()
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	symmetry_mean	fra
560	14.05	27.15	91.38	600.4	0.09929	0.11260	0.04462	0.04304	0.1537	
428	11.13	16.62	70.47	381.1	0.08151	0.03834	0.01369	0.01370	0.1511	
198	19.18	22.49	127.50	1148.0	0.08523	0.14280	0.11140	0.06772	0.1767	
203	13.81	23.75	91.56	597.8	0.13230	0.17680	0.15580	0.09176	0.2251	
41	10.95	21.35	71.90	371.1	0.12270	0.12180	0.10440	0.05669	0.1895	

5 rows × 30 columns

Figure 3.24: Training and test data

Here, 80% of the data from the complete set of training data and 20% of the data for testing. I am now starting my work period by training datasets. The initial step in training is to choose a model. and the requirement to produce model objects.



```
#puts models is dictionary

models = {
    "LR" : LogisticRegression(),
    "KNN" : KNeighborsClassifier(),
    "DT" : DecisionTreeClassifier(),
    "RF" : RandomForestClassifier(),
    "SVC" : SVC(),
    "NB" : GaussianNB(),
    "ANN" : MLPClassifier()
}

#create a function to fit model

def fit_and_score(models, X_train,X_test,Y_train,Y_test):
  #make a dict to keep model score
  model_score = {}
  for name,model in models.items():
    #fit the model to the data
    model.fit(X_train,Y_train)
    #evaluate the model and appends its score to the model_score
    model_score[name] = round(model.score(X_test,Y_test)*100,2)
  return model_score
```

```
[121] model_score = fit_and_score(models,X_train,X_test,Y_train,Y_test)

      /usr/local/lib/python3.8/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:692: ConvergenceWarni
        warnings.warn(
```

Figure 3.25: Training my model

I create an object for training the model. And I train (fit) My every model separately.

CHAPTER 4

EXPERIMENTAL RESULTS & ANALYSIS

4.1 Introduction

I'll discuss about my model's accuracy level and performance evaluation in this part.

4.2 Results

I've already trained my whole datasets using a total of seven models. I can then compare the accuracy of the various models. Below is a list of their accuracy results-

Table 4.1: Accuracy Outcome.

Model name	Accuracy
Logistic Regression	97.37%
KNN	97.37%
Decision Tree	93.86%
Random Forest	94.74%
SVC	96.49%
Naive Bayes	93.86%
ANN	95.61%

Table 4.2.1 demonstrates that using the same dataset, the seven models (Logistic Regression, KNN, Decision Tree, Random Forest, SVC, Naive Bayes & ANN) provide me with seven different levels of accuracy. Now, in figure 4.2.1, the accuracy comparison is displayed.

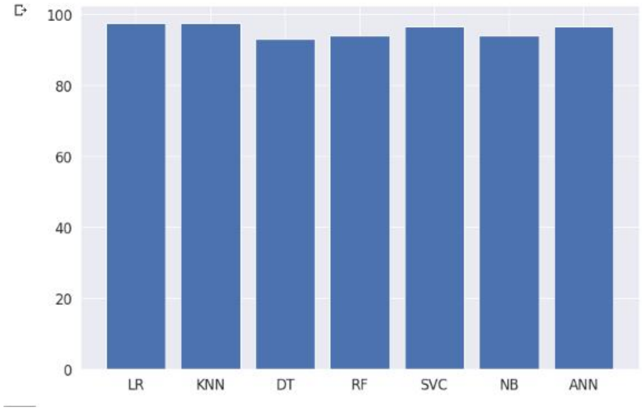
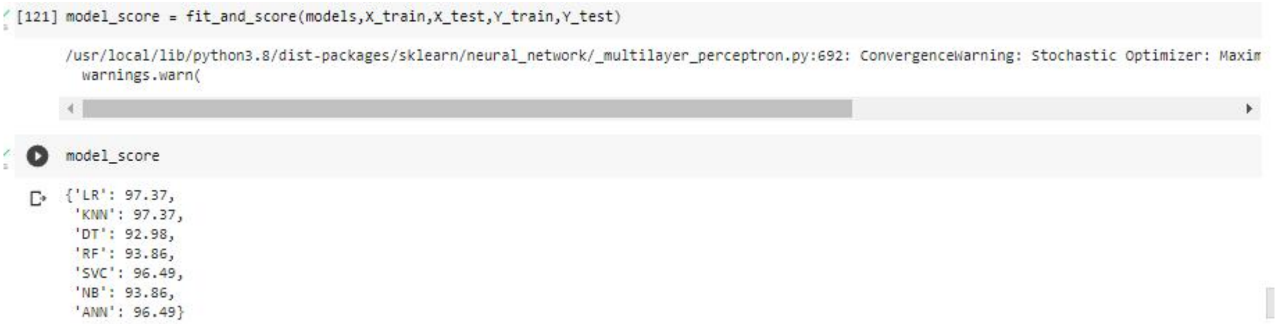


Figure 4.1: Accuracy Comparison.



```
[121] model_score = fit_and_score(models,X_train,X_test,Y_train,Y_test)

      /usr/local/lib/python3.8/dist-packages/sklearn/neural_network/_multilayer_perceptron.py:692: ConvergenceWarning: Stochastic Optimizer: Maxim
        warnings.warn(
```

```
model_score
```

```
{'LR': 97.37,
 'KNN': 97.37,
 'DT': 92.98,
 'RF': 93.86,
 'SVC': 96.49,
 'NB': 93.86,
 'ANN': 96.49}
```

Figure 4.2: Model Accuracy.

4.3 Model Optimization

Now that I have a baseline model, I am aware that early predictions from the model are only sometimes good or accurate. Therefore, the next step I'm doing to enhance the model is:

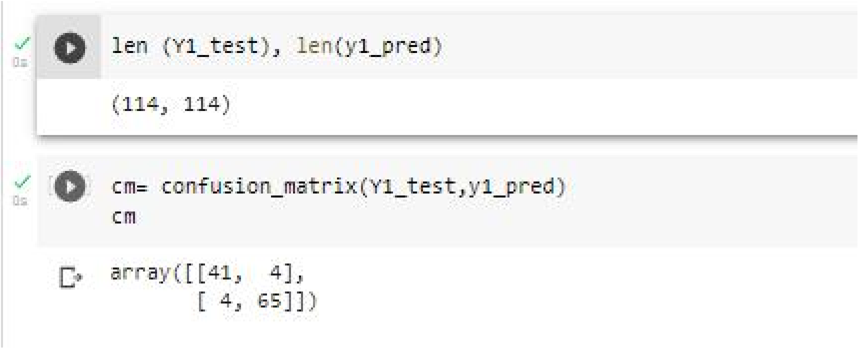
1. Confusion Matrix
2. Classification report

4.3.1 Confusion Matrix

An illusion's matrix condenses the classification problem's prediction outcomes.

Predictions that were accurate and inaccurate are tallied and split by each square. The matrix of confusion may be unlocked with this. The confusion matrix needs to be clarified when it displays your classification models instead of making predictions. This not only provides you with information into the mistakes made by your categorization, but more crucially, it identifies the many kinds of mistakes that might happen.

Let see my test data length,



```
len (Y1_test), len(y1_pred)
```

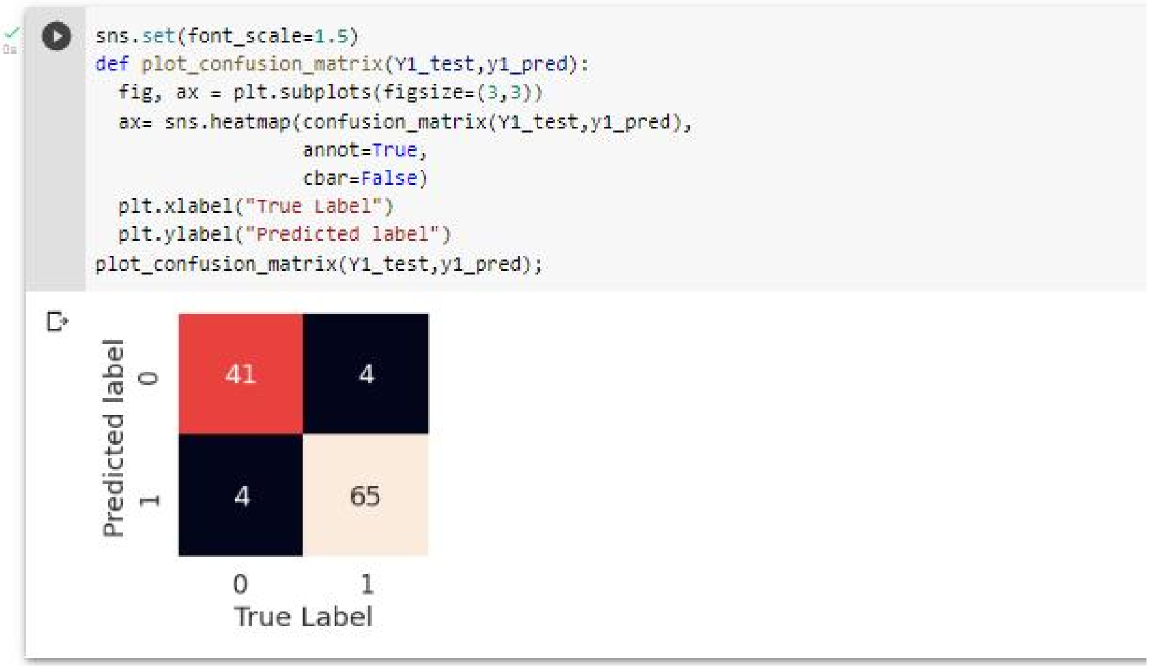
```
(114, 114)
```

```
cm= confusion_matrix(Y1_test,y1_pred)
cm
```

```
array([[41,  4],
       [ 4, 65]])
```

Figure 4.3: Test Data Length and Confusion Matrix

There are 114 test data visible here. The outcome will be determined by a confusion matrix using these data.



```
sns.set(font_scale=1.5)
def plot_confusion_matrix(Y1_test,y1_pred):
  fig, ax = plt.subplots(figsize=(3,3))
  ax= sns.heatmap(confusion_matrix(Y1_test,y1_pred),
                  annot=True,
                  cbar=False)
  plt.xlabel("True Label")
  plt.ylabel("Predicted label")
plot_confusion_matrix(Y1_test,y1_pred);
```

Figure 4.4: Confusion Matrix by using Heatmap

The whole test data is visible via this confusion matrix. And from the above data, 106 correct predictions and 8 erroneous predictions were made. The true negative prediction number is 4, and the false positive prediction number is 4.

In this instance, "positive" and "negative" are just names for the projected classes. There are four ways to determine if the forecasts were accurate or not:

1. True Negative(TN) A case that was negative and was expected to be negative is a true negative.

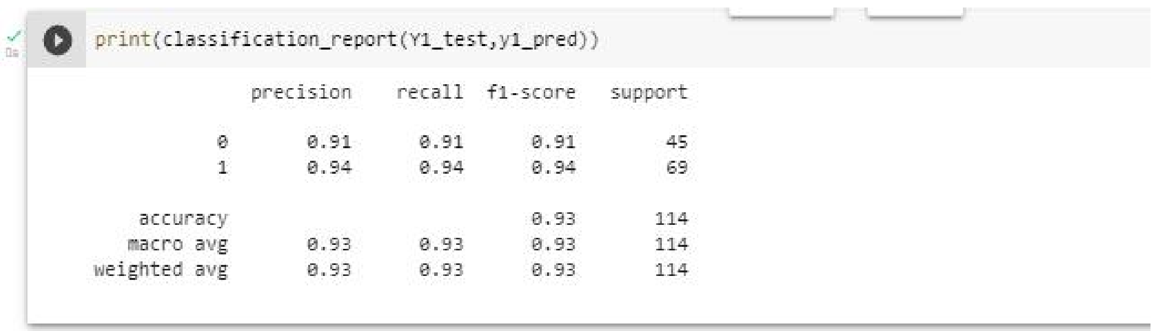
2. True Positive (TP) refers to a case that was both positive and projected to be positive.

3. False Negative (FN) when a case was positive, although the result was projected to be negative

4. False Positive (FP) when a case was negative but was anticipated to be positive

4.3.2 Classification Report

The accuracy of predictions made by a classification algorithm is evaluated using a classification report. How many of the forecasts came true, and how many were wrong? More specifically, the metrics of a classification report random forest are predicted using True Positives, False Positives, True Negatives, and False Negatives, as illustrated below-



```
print(classification_report(Y1_test,y1_pred))
```

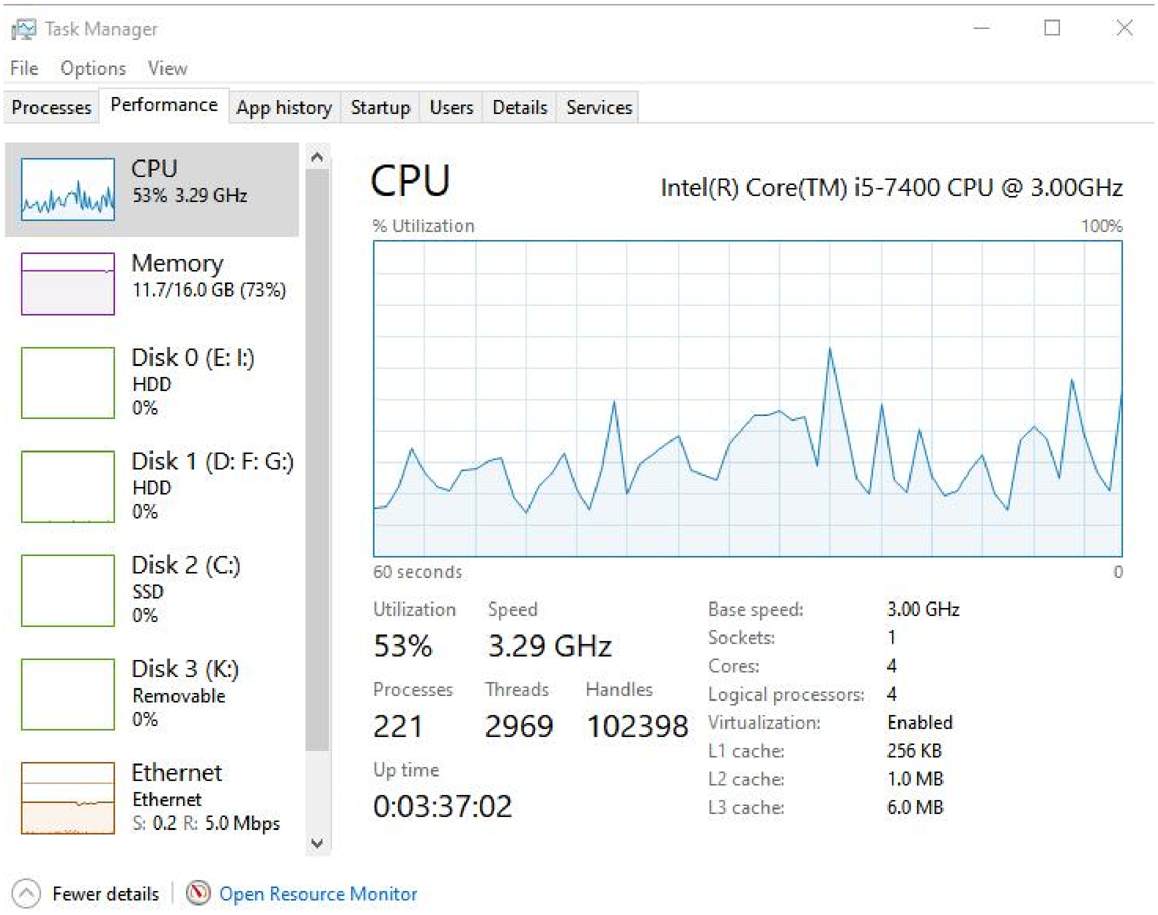
```
              precision    recall  f1-score   support

           0       0.91      0.91      0.91        45
           1       0.94      0.94      0.94        69

    accuracy                           0.93       114
   macro avg       0.93      0.93      0.93       114
weighted avg       0.93      0.93      0.93       114
```

Figure 4.5: Classification report for KNN.

4.4 Requirements Accessories

These conditions must be satisfied for image processing to operate once the approach has been described and my model has been fully trained.

- Operating System (Windows 10)
- Hard Disk (minimum 1000)
- Ram (Minimum 12 GB)
- GPU(Recommended) – I have used 4GB (AMD RX550)



CPU
53% 3.29 GHz

Memory
11.7/16.0 GB (73%)

Disk 0 (E: I:)
HDD
0%

Disk 1 (D: F: G:)
HDD
0%

Disk 2 (C:)
SSD
0%

Disk 3 (K:)
Removable
0%

Ethernet
Ethernet
S: 0.2 R: 5.0 Mbps

CPU Intel(R) Core(TM) i5-7400 CPU @ 3.00GHz

Utilization 53%
Speed 3.29 GHz
Processes 221
Threads 2969
Handles 102398
Up time 0:03:37:02

Base speed: 3.00 GHz
Sockets: 1
Cores: 4
Logical processors: 4
Virtualization: Enabled
L1 cache: 256 KB
L2 cache: 1.0 MB
L3 cache: 6.0 MB

Figure 4.6: CPU Performance at the time of working

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary

Seven algorithms are used in this paper. Those are Logistic Regression, k Nearest Neighbor, Decision Tree Classifier, Random Forest, Support Vector Machine, Naive Bayes, Artificial neural networks (ANNs). I used the Wisconsin breast cancer dataset here as the dataset. k Nearest Neighbor and logistic regression both had excellent accuracy scores of 97.37%. Support Vector Machine 96.49% performs best among these seven models in terms of accuracy, followed by Artificial neural networks 95.61%, Random Forest 94.74%, Naive Bayes 93.86%, and Decision Tree with an accuracy of 93.86%. Nearest Neighbor and Logistic Regression both provide the same result. Therefore, it is clear that among these seven models, logistic regression and k Nearest Neighbor both perform most well. Then, as I am aware that cross validation is intended to identify overfitting, I apply 5 K-fold cross validation for simply Logistic Regression. Thus, I may conclude that mymodel is not overfitting. I may draw the conclusion that, given the dataset in question, Logistic Regression outperforms other methods for predicting breast cancer.

5.2 Future Work

A more precise and comprehensive breast cancer diagnostic model will be developed using data gathered from various parts of the globe. Future studies will also concentrate on data collecting during the most recent time period to uncover new possible predictors that may be used in decision-making. This study may be built upon and enhanced to attain the automated detection of breast cancer. Although the data collection is not huge, it is a well-known standard data set. By keeping a sizable data collection and introducing other functions, such as the stage identification of carcinoma, moving forward, I will enhance my work in the near future.In the future, anyone can work with datasets since I'm making a CSV test dataset and filtered dataset that will allow for further development of that report. I used logistic regression, but if someone wanted to apply it with a more potent methodology, they might get a more accurate result.

5.3 Conclusion

In the actual world, automated breast cancer prediction is a severe medical issue. The key to treating breast cancer is predicting it early. In order to model the actual prediction of breast cancer for local and conventional treatments as well as other methods, this article explains how to use machine learning algorithms like logistic regression, k Nearest Neighbor, decision tree classifier, random forest, support vector machine, naive Bayes, and artificial neural networks (ANNs).The Wisconsin's breast cancer dataset, which is from the UCI Machine Learning Repository, served as the basis for my examination of the classification criteria using seven machine learning techniques, including logistic regression, k-nearest neighbor, decision tree classifier, random forest, support vector machine, naive Bayes, and artificial neural networks (ANNs). I examined every one of the algorithms' outputs. I also attempted to get the most accurate information possible from them. With a classification accuracy of 97.37%, the findings indicated that Logistic Regression and k Nearest Neighbor performed best. The Support Vector Machine came in second with a classification accuracy of 96.49%, while artificial neural networks came in last with a classification accuracy of 95.61%. Finally, I looked for a reliable classifier that might identify cancer in patients and maybe save their lives. Although everything is within the power of My Almighty, I shall make every effort to use early cancer prediction to address the cancer issue.

Reference:

[1]Cruz, J. A., & Wishart, D. S. (2006). Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics, 2, 117693510600200. Available at: https://sci-hub.ru/10.1177/117693510600200030

[2]Amrane, M., Oukid, S., Gagaoua, I., & Ensari, T. (2018). Breast cancer classification using machine learning. 2018 Electric Electronics, Computer Science, Biomedical Engineerings'Meeting(EBBT). Available at: https://sci-hub.ru/10.1109/EBBT.2018.8391453

[3]Sathya, S., Joshi, S., & Padmavathi, S. (2017). Classification of breast cancer dataset by different classification algorithms. 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS). Available at: https://sci-hub.ru/10.1109/ICACCS.2017.8014573

[4]S. P. T. Vikas Chaurasia, "Prediction of benign and malignant breast cancer using data mining techniques," *Journal of Algorithms & Computational Technology,* vol. 12, no. 2, pp. 119-126, 2018.

[5]M. R. H. I. M. H. H. N. K. Md. Milon Islam, "Breast Cancer Prediction: A Comparative Study Using Machine," *Springer Nature Singapore Pte Ltd 2020,* vol. 1, no. 5, p. 290, 18 August 2020.

[6]T. B.Padmapriya, "Classification Algorithm Based Analysis of Breast Cancer Data," *International Journal of Data Mining Techniques and Applications,* vol. 5, no. 1, pp. 43-49, June 2016.

[7]G. Singh, "Breast Cancer Prediction Using Machine Learning," *International Journal of Scientific Research in Computer Science,* vol. 6, no. 4, pp. 278-284, 30 July 2020.

[8]D. M. I. D. R Delshi Howsalya Devi, "Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer," *International Journal of Advanced Engineering Technology,* vol. 7, no. 2, pp. 93-98, 2016.

[9]K. S. A. J. S. M. M. F. A. Baraa M. Abed, "A hybrid classification algorithm approach for breast cancer diagnosis," IEEE Industrial Electronics and Applications Conference (IEACon), Kota Kinabalu, Malaysia, 2016.

[10]K.Sivakami, "Mining big data: Breast cancer prediction using DT-SVM Hybrid model," *International Journal of Scientific Engineering and Applied Science (IJSEAS),* vol. 1, no. 5, pp. 418-429, 2015.

[11]R. Chandrasekar, "Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis," *IOSR Journal of Computer Engineering (IOSR-JCE),* vol. 15, no. 5, pp. 39-44, 2013.

[12]S. O. G. a. T. E. M. Amrane, "Breast cancer classification using machine learning," *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT),* vol. VII, no. II, pp. 1-4, 2018.

[13]M. R. J. P. Miss Jahanvi Joshi, "DIAGNOSIS AND PROGNOSIS BREAST CANCER USING CLASSIFICATION RULESng," *International Journal of Engineering Research and General Science.,* vol. 2, no. 6, pp. 315-23, 2014.

[14]M. SK, "Performance analysis of data mining algorithms for breast cancer cell detection using Naïve Bayes, logistic regression and decision tree," *International Journal of Engineering and Computer Science,* vol. 6, no. 2, pp. 20388-91, 2017.

[15]Z. W. C. P. a. X. L. Wenbin Yue, "Learning with Applications in Breast Cancer Diagnosis and Prognosis," *MDPI,* vol. V, no. VIII, pp. 1-17, 2018.

[16]D. W. H. Wolberg, "Breast Cancer Wisconsin (Original) Data Set," UC Irvine, [Online].Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)?fbclid=IwAR0p0UkTfRhdFFpxGLY1b30atI7cWNRutiO93w57CoOEc1XkjVMyfRe7ZDI. [Accessed 25 August 2018].

BREAST CANCER PREDICTION

ORIGINALITY REPORT

28%	23%	12%	19%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	4%
3	www.iosrjournals.org Internet Source	3%
4	Submitted to University of Ruhuna Matara Student Paper	1%
5	Submitted to Coventry University Student Paper	1%
6	Submitted to Purdue University Student Paper	1%
7	link.springer.com Internet Source	1%
8	Submitted to University of Stirling Student Paper	1%
9	Submitted to University of Wales Institute, Cardiff Student Paper	1%