### Table of Content

Declaration… (ii)

Certificate… (iii)

Acknowledgement… (iv)

Abstract… (v)

Chapter 1. Purpose of the Study. 10

* 1. [Introduction 10](#_TOC_250010)
  2. [Some Risk Factors of Breast Cancer 11](#_TOC_250009)
     1. [Age 11](#_TOC_250008)
     2. [Personal History of Breast Cancer 11](#_TOC_250007)
     3. Family History of Breast Cancer 11
     4. Genetic Factors…………………………………………………………….11
     5. Childbearing and Menstrual History………………………………………11

Chapter 2. Background of the Study… 12

**Chapter 3.** Future Works… 14

Chapter 4. Literature 16

Chapter 5. Research Gap… 18

**Chapter 6**. Aims and Objectives of the Project 19

Chapter 7. Scope of the Project… 20

Chapter 8. Methodology… 21

* 1. Flowchart… 21
  2. [Development Tool… 22](#_TOC_250006)
  3. [Implementation Setup… 23](#_TOC_250005)

[Chapter 9. Result and Discussion 29](#_bookmark0)

Chapter 10. Conclusion… 30

[References 31](#_TOC_250000)

# List of Figures

**Figure 1.1.** Early Detection 10

**Figure 8.1.1.** Flowchart 21

**Figure 8.3.1.** Importing Python Libraries. 23

**Figure 8.3.2.** Loading the Dataset 23

**Figure 8.3.3.** Counting Rows and Columns and Empty Value in Dataset 24

**Figure 8.3.4.** Dropping All Missing Values and Counting Rows and Columns 24

**Figure 8.3.5.** Visualization of The Count of M and B**.** 25

**Figure 8.3.6.** Visualization Using Pair Plot 25

**Figure 8.3.7.** Visualization of Correlation. 26

**Figure 8.3.8**. Splitting of Dataset and Creation of Models. 27

**Figure 8.3.9.** Checking Accuracy of Models on Training and testing set 28

**Figure 8.3.10.** Output 28

## Chapter-1

### Purpose of the Study

#### Introduction

According to the world health organization (WHO), Breast cancer is the most frequent cancer among women, impacting 2.1 million women each year, and also causes the greatest number of cancer-related deaths among women. In 2020, it is estimated that 685,000 women died globally from breast cancer-that is approximately 15% of all cancer deaths among women. While breast cancer rates are higher among women in more developed regions, rates are increasing in nearly every region globally.



#### Figure 1.1: Early Detection of Breast Cancer.

In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Early diagnosis strategies focus on providing timely access to cancer treatment by reducing barriers to care and/or improving access to effective diagnosis services. The goal is to increase the proportion of breast cancers identified at an early stage, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Since early detection of cancer is key to effective treatment of breast cancer we use various machine learning algorithms to predict if a tumor is benign or malignant, based on the features provided by the data.

**1.2 Some Risk Factors of Breast Cancer**

**1.2.1 Age**

The chance of getting breast cancer increases as women age.Nearly 80 percent of breast cancers are found in women over the age of 50.

**1.2.2 Personal history of breast cancer**

A woman who has had breast cancer in one breast is at an increased risk of developing cancer in her other breast.

**1.2.3 Family history of breast cancer**

A woman has a higher risk of breast cancer if her mother, sister or daughter had breast cancer, especially at a young age (before 40). Having other relatives with breast cancer may also raise the risk.

**1.2.4** **Genetic Factors**

Women with certain genetic mutations, including changes to the BRCA1 and BRCA2 genes, are at higher risk of developing breast cancer during their lifetime. Other gene changes may raise breast cancer risk as well.

**1.2.5 Childbearing and Menstrual History**

The older a woman is when she has her first child, the greater her risk of breast cancer. Also at higher risk are:

**1.2.5.1** Women who menstruate for the first time at an early age (before 12)

**1.2.5.2** Women who go through menopause late (after age 55)

**1.2.5.3** Women who’ve never had children

The use of computer-based learning models has become a predominant area of cancer research. In recent years, several researchers have focused on building systems, both hybrid and fully automatic systems, that could facilitate the diagnosis, prognosis, and prediction of breast cancer outcomes taking a leap using Statistics and Artificial Intelligence. The development of these systems requires different techniques, where the most common are machine learning (ML) algorithms. Several scientific studies have published algorithms and nomograms predicting the pathologic stage of patients with clinically localized cancer or Gleason score upgrading . Specifically, ML allows the integration or combination of different layers of data, such as those from medical images, laboratory results, clinical outcomes, biomarkers, and biological features for better prognostication and stratification of patients toward personalized medicine.

## Chapter-2

### Background of the Study

The following machine learning classification techniques are used:

*A. Decision Tree Classifier*

In 1980, J. Ross Quinlan developed ID3 (Iterative Dichotomise) which is a decision tree algorithm. The decision tree classifier is an example of supervised machine learning A decision tree works on possible solutions to a decision based on certain conditions. It classifies conditions at every node to find a solution.

Algorithm

* 1. Starts at root node
  2. Root value compared with record real dataset attribute
  3. Jump to next node based on comparison
  4. Compare attribute value with sub-node value and jump accordingly
  5. Repeat till left node of tree is reached.

*B. Support Vector Machine*

Support Vector Machines (SVMs) are machine learning algorithms which are used to deal with both classification and regression problems. SVM linear classifier is built around the margin maximization principle. SVM algorithm is used to create a decision boundary, called hyperplane that segregates n-dimensional space into different classes so that new data points can be added easily. This linear classifier plans to broaden the space between the decision hyperplane and the closest data points by finding the most suitable hyperplane.

Algorithm

1. Correctly classify the training dataset on the basis of lines/boundaries
2. Selects the one having maximum distance from the closest data point out of the lines/boundaries

*C. K-Nearest Neighbour*

K-Nearest Neighbour is a Machine Learning algorithm which belongs to the Supervised Learning technique. K-NN algorithm can solve both classification and regression problems but is mainly used for classification problems. It is a non-parametric classification method, which doesn’t make assumptions based on underlying data. It is also known as a lazy learner algorithm because the training set is not learned immediately, instead, it stores the dataset and then at classification time, it performs an action on it. In the K-NN algorithm, similarity is assumed between a new case and available data, and the new case is placed in the category with the most similarity to the available cases. All the available data is stored and the classification of new data point based is done based on similarity. As a result, when new data appears, the K-NN algorithm can be easily used to classify it into an appropriate category. K-NN stores datasets during the training phase, and when it gets new datasets, it categorizes them into a category that is similar to the new data.

Algorithm

1. Load training and testing data from dataset
2. Choose the value of nearest data point (K)
3. For each data point, calculate Euclidean (distance between testing and rows of training data) distance and sort them ascendingly
4. Choose top K row from sorted array and allot a class to test point based on the most frequent class

*D*.*Decision Tree:*

DTs apply a top-down approach to data so that given a knowledge set, they struggle to group and label observations that are similar between them, and appearance for the simplest rules that split the observations that are not the same between them until they reach a certain degree of similarity. They use a layered splitting process, where at each layer they struggle to separate the info into two or more groups, in order that data that fall under an equivalent group is most like every other (homogeneity), and groups are as die-rent as possible from one another.

## Chapter-3

### Future Work

The analysis of the results signifies that the integration of multidimensional data along with different

classification, feature selection and dimensionality reduction techniques can provide auspicious tool for inf

erence in this domain.

Further research in this field should be carried out for the better performance of the classification techniques

so that it can predict on more variables. We are intending how to parametrize our classification techniques

hence to achieve high accuracy. We are looking into many datasets and how further Machine Learning algorithms can be used to characterize Breast Cancer. We want to reduce the error rates with maximum accuracy.

Machine learning has been widely implemented in many areas such as speech recognition, object detection, predict the protein structure, and it is used to detect and diagnose human cancers especially lung, liver, brain, and breast cancers (C. Deng et al., 2020). Its ability to learn from a tremendous amount of data makes it a powerful tool to support decision making in recent decades. For the time being, many methods of machine learning have been used to help a physician in breast cancer detection and diagnosis. The capability of these methods not only helps for revealing the early stage of breast cancer but also the prediction of cancer occurrence and the possibility of the recurrence of cancer after treatment can be provided (Kajala & Jain, 2020). Moreover, during a specific time, the possibility of the death or survival rate can be addressed. Besides, using machine learning can significantly reduce the mistakes taken by humans and enhance the robustness and reliability of outcomes to help to build stable systems (Kajala & Jain, 2020). However, with all of these abilities of machine learning techniques, it still far away from behaving smartly like what human beings do.

## Chapter-4

### Literature Survey

* + According to Shubham Sharma in their research paper **(DOI: doi.org/10.1109/CTEMS.2018.8769187)**

The most frequently occurring cancer among Indian women is breast cancer. There is a chance of fifty percent for fatality in a case as one of two women diagnosed with breast cancer die in the cases of Indian women [1]. This paper aims to present comparison of the largely popular machine learning algorithms and techniques commonly used for breast cancer prediction, namely Random Forest, kNN (k-Nearest-Neighbor) and Naïve Bayes. The Wisconsin Diagnosis Breast Cancer data set was used as a training set to compare the performance of the various machine learning techniques in terms of key parameters such as accuracy, and precision. The results obtained are very competitive and can be used for detection and treatment.

* According Siham A. Mohammed their research paper

Breast cancer is the second leading cause of death among women worldwide [[1](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR1)]. In 2019, 268,600 new cases of invasive breast cancer were expected to be diagnosed in women in the U.S., along with 62,930 new cases of non-invasive breast cancer [[2](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR2)]. Early detection is the best way to increase the chance of treatment and survivability. Data mining has become a popular tool for knowledge discovery which shows good results in marketing, social science, finance and medicine [[19](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR19), [20](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR20)]. Recently, multiple classifiers algorithms are applied on medical datasets to perform predictive analysis about patients and their medical diagnosis [[6](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR6), [9](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR9), [10](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR10), [21](https://link.springer.com/chapter/10.1007/978-981-15-7205-0_10#ref-CR21)]. For example, using machine learning techniques to assess tumor behavior for breast cancer patients. One problem is that there is a class imbalance in the training data, since the probability of not having this disease is higher than the one of having it. This paper introduces a comparison between three different classifiers: J48, NB, and SMO with respect to accuracy in detection of breast cancer. Our aim is to prepare the dataset by proposing a suitable method that can manage the imbalanced dataset and the missing values, to enhance the classifier’s performance. All tasks were conducted using Weka 3.8.3.

Breast cancer is considered to be one of the significant causes of death in women. Early detection of breast cancer plays an essential role to save women’s life. Breast cancer detection can be done with the help of modern machine learning algorithms. In this paper, we focus on how to deal with imbalanced data that have missing values using resampling techniques to enhance the classification accuracy of detecting breast cancer. In our work, three classifiers algorithms J48, NB, and SMO applied on two different breast cancer datasets. Results show that using the resample filter in the preprocessing phase enhances the classifier’s performance. In the future, the same experiments will apply to different classifiers and different datasets.

* According to Wenbin Yue in their research paper **(** [**https://doi.org/10.3390/designs2020013**](https://doi.org/10.3390/designs2020013)**)**

In this paper, we have provided explanations of different ML approaches and their applications in BC diagnosis and prognosis used to analyse the data in the benchmark database WBCD. ML techniques have shown their remarkable ability to improve classification and prediction accuracy. Various methods have been shown with references, algorithms, sampling strategies and classification accuracies, providing a clear and intuitive catalogue of information. Although lots of algorithms have achieved very high accuracy in WBCD, the development of improved algorithms is still necessary. Classification accuracy is a very important assessment criteria but it is not the only one. Different algorithms consider different aspects, and have different mechanisms. Although for several decades ANNs have dominated BC diagnosis and prognosis, it is clear that more recently alternative ML methods have been applied to intelligent healthcare systems to provide a variety of options to physicians.

* According to Sharmin, Selina, et al. IEEE Access (2023) in their research Paper

This research presents an enhanced model for breast cancer detection. It proposes a breast cancer detection model that leverages the power of a pre-trained DL model, ResNet50V2, for efficient feature extraction. Then, it incorporated several ML algorithms, including Decision Tree (DT), Random Forest (RF), Extra Tree (ET), Ada Boosting (AbB), Histogram Gradient Boosting Classifier (HGBC), Gradient Boosting Classifier (GBC), Extreme Boosting Classifier (XGB), and LGB into our framework. By combining the strengths of both DL and ML, it present a hybrid dependable breast cancer prediction model which allows us to capture complex patterns and relationships in breast cancer data while benefiting from traditional machine learning algorithms’ interpretability and generalization abilities.

## Chapter-5

### Research Gaps

## Genetics (knowledge of genetic changes, their effects, and interactions)

## Initiation of breast cancer (how developmental signaling pathways cause ductal elongation and branching at the cellular level and influence stem cell dynamics, and how their disruption initiates tumor formation);

## Progression of breast cancer (deciphering the intracellular and extracellular regulators of early progression, tumor growth, angiogenesis and metastasis);

## therapies and targets (understanding who develops advanced disease);

## disease markers (incorporating intelligent trial design into all studies to ensure new treatments are tested in patient groups stratified using biomarkers);

## prevention (strategies to prevent oestrogen-receptor negative tumors and the long-term effects of chemoprevention for oestrogen-receptor positive tumors);

## psychosocial aspects of cancer (the use of appropriate psychosocial interventions, and the personal impact of all stages of the disease among patients from a range of ethnic and demographic backgrounds).

## Chapter-6

### Aims and Objectives

The objective of this project is to **increase the proportion of breast cancers identified at an early stage**, allowing for more effective treatment to be used and reducing the risks of death from breast cancer. Here are some potential objectives for such a project:

• To create an ML Model to classify malignant and benign tumor by training it using a dataset of both types of tumor cells. Malignant tumors are the ones which are cancerous and benign are not.

• To complete this ML project, we are using the supervised machine learning classifier algorithm. To build the best model, we must train and test the dataset with multiple Machine Learning algorithms then we can find the best ML model.

• The project has lot of scope in the field of medicine and healthcare, helping doctors in the detection of breast cancer patients with better accuracy.

The motivation behind predicting breast cancer is that cancer in women always has a huge incidence rate and mortality rate. Breast cancer alone is estimated to account for 25% of all new cancer diagnoses worldwide and 15% of cancer deaths in women worldwide, according to the latest cancer statistics. Every 1 in 8 Women in USA develop breast cancer in her lifetime. In case of any sign or symptom, people usually visit a doctor immediately, who may refer you to an oncologist for help. An oncologist can diagnose breast cancer by: Examining the patient's medical history thoroughly, examining both breasts, and even checking for swelling or hardening of any lymph nodes in the armpits. Here in this project, we have used the Wisconsin Breast Cancer Dataset (WBCD) and with the dataset we have used machine learning algorithms to predict whether a patient has breast cancer. Machine learning algorithms (MLs) are indicated as one Option of human vision and experience to make final decisions with high accuracy.

## Chapter-7

### Scope of the Study

The breast cancer classification project opens several avenues for scope and development:

Feature Engineering: Further investigation into feature engineering techniques can be explored to enhance the performance of the logistic regression model. Feature selection, transformation, and extraction methods may be employed to identify the most informative and relevant features for breast cancer classification.

Model Optimization: Although logistic regression yielded satisfactory results, further optimization of the model can be pursued. Techniques such as regularization, hyperparameter tuning, and ensemble methods may be employed to improve the model's performance and generalization.

Integration of Advanced Techniques: Integration of logistic regression with other advanced machine learning techniques, such as support vector machines, random forests, or deep learning models, may be explored to leverage their complementary strengths and enhance classification accuracy.

Validation on Diverse Datasets: The performance and generalizability of the logistic regression model should be validated on larger and more diverse datasets. Evaluating the model's performance across different populations and healthcare settings will enhance its reliability and practicality.

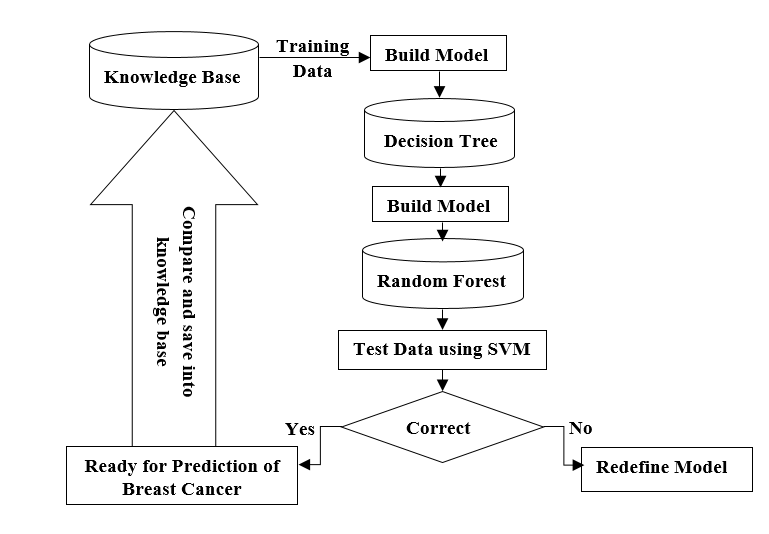
Clinical Integration: Collaboration with healthcare professionals and experts in the field of breast cancer diagnosis and treatment is essential for integrating the logistic regression model into clinical practice. The model's implementation in real-world scenarios and its impact on clinical decision-making can be studied and validated.

Development of Decision Support Systems: The logistic regression model can be integrated into decision support systems that aid healthcare professionals in making accurate and timely breast cancer diagnoses. These systems can provide valuable insights and recommendations, potentially leading to improved patient care and outcomes.

## Chapter-8

### Methodology

**8.1 Flowchart**



#### Figure 8.1.1: Flowchart

#### Development Tool

Different steps that we are going to follow in order to accomplish are goal are: -

* Data acquisition
* Data pre-processing
* Data analysis
* Data Visualization
* Model Designing
* Training and Testing of Model
* Result

Each phase will have some tasks and deliverables.

In order to achieve all this, we are going to use various tools and libraries, and these are –

#### Language:

**PYTHON** - Python is a High-level-programming language. It’s design philosophy emphasises code readability with use of significant indentation. We are going to use python extensively for our project.

#### Tools:

**GOOGLE COLLABORATORY** – Google Collaboratory web based interactive computing platform. It combine live code, narrative text and visualizations. It support various languages that are popular in data science like python, Julia, Scala, R etc.

**MICROSOFT EXCEL** - Microsoft Excel is spreadsheet tool that features calculations or computation capabilities, graphing tools, pivot tables etc.

#### Libraries:

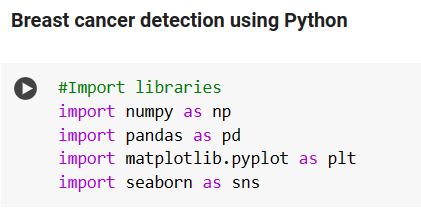
**NUMPY** - It is a python library used for scientific computation in python and work with arrays. It also has functions for working with linear algebra and matrices.

**PANDAS** - It is written for python from data manipulation and analysis, it offers data structures and operations for manipulating numerical tables and time series.

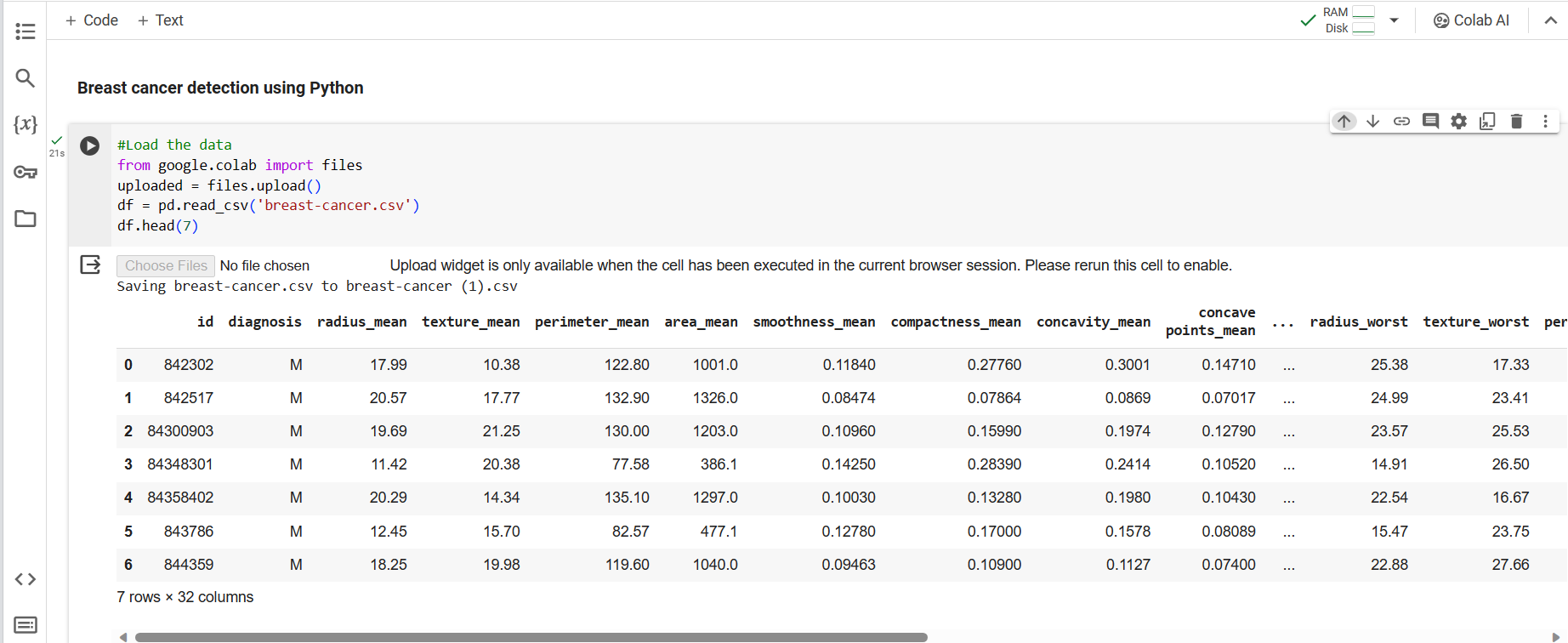
**MATPLOTLIB** - It is plotting library for the python and its numerical extension Numpy. It provides an object-oriented API for embedding plots and visualize data.

**SEABRON-** Seaborn is a library for making statistical graphics in python. It builds on top of matplotlib and integrates closely on with pandas data structures. Seaborn helps us to explore your data.

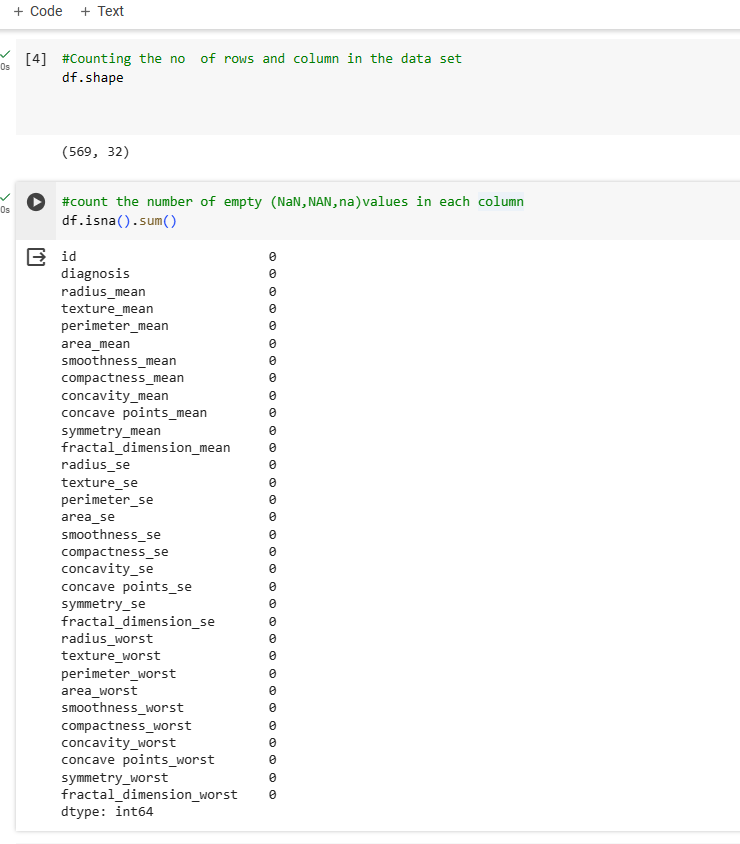
**8.3 IMPLEMENTATION STEPS:**



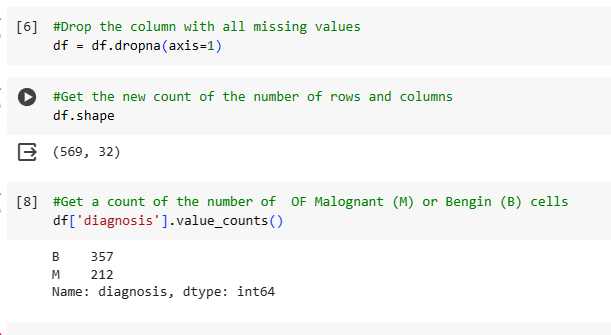
**Figure 8.3.1: Importing python libraries.**



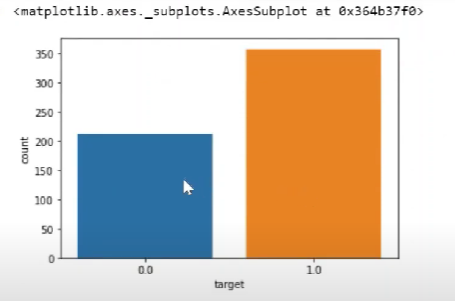
**Figure 8.3.2: Loading the Dataset.**

****

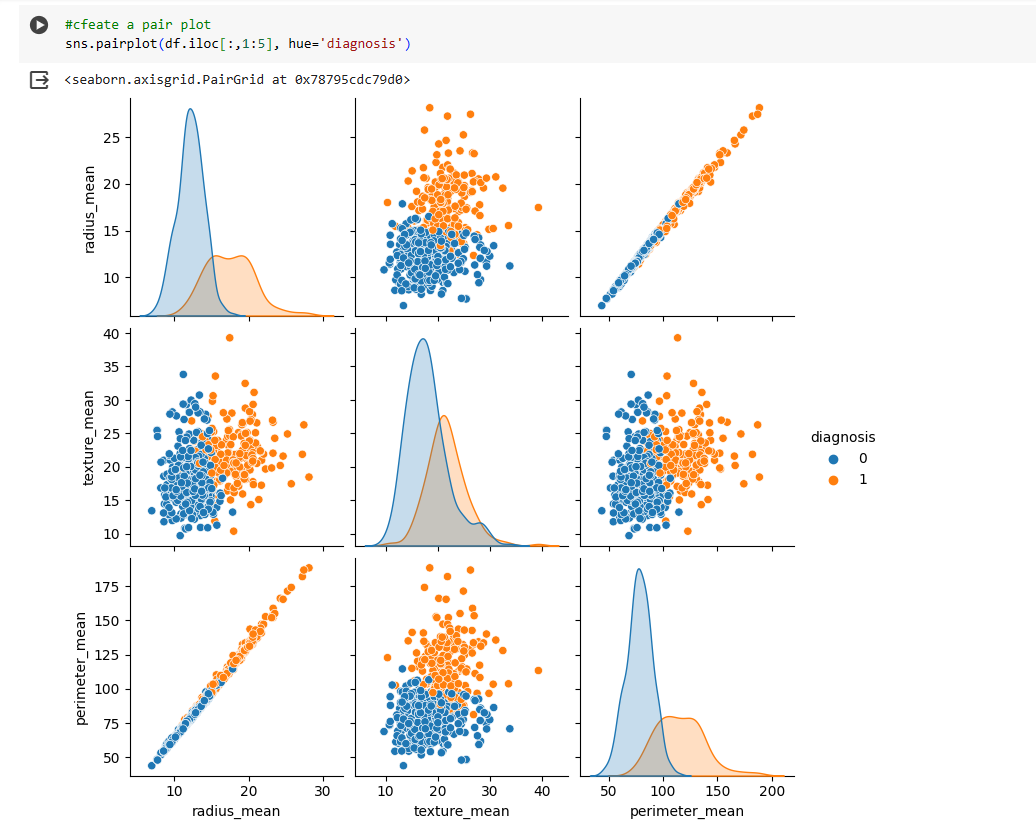
**Figure 8.3.3: Counting rows and columns and empty value .**

****

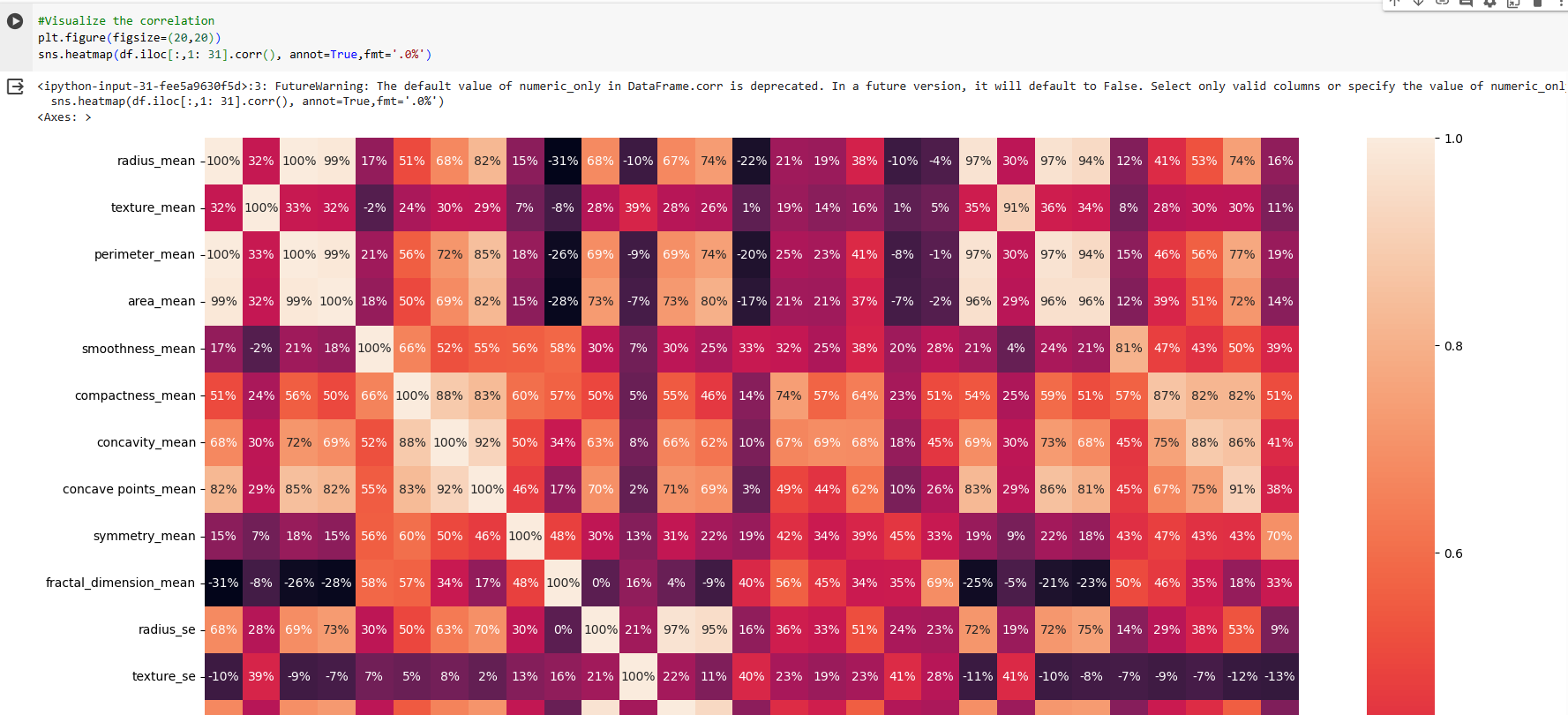
**Figure 8.3.4: Dropping all missing values and counting rows and columns .**

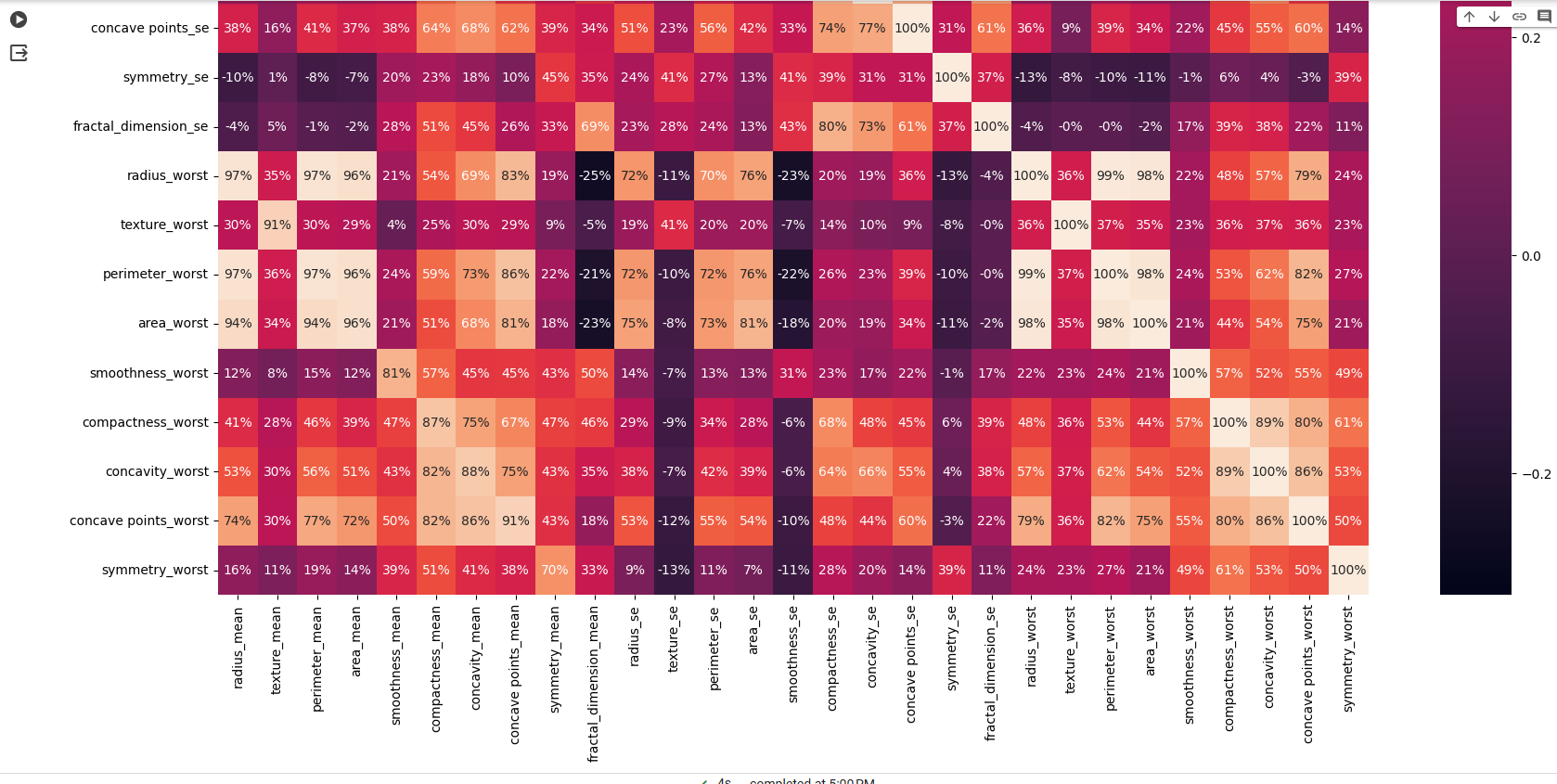
****

**Figure 8.3.5: Visualization of the count of M and B.**

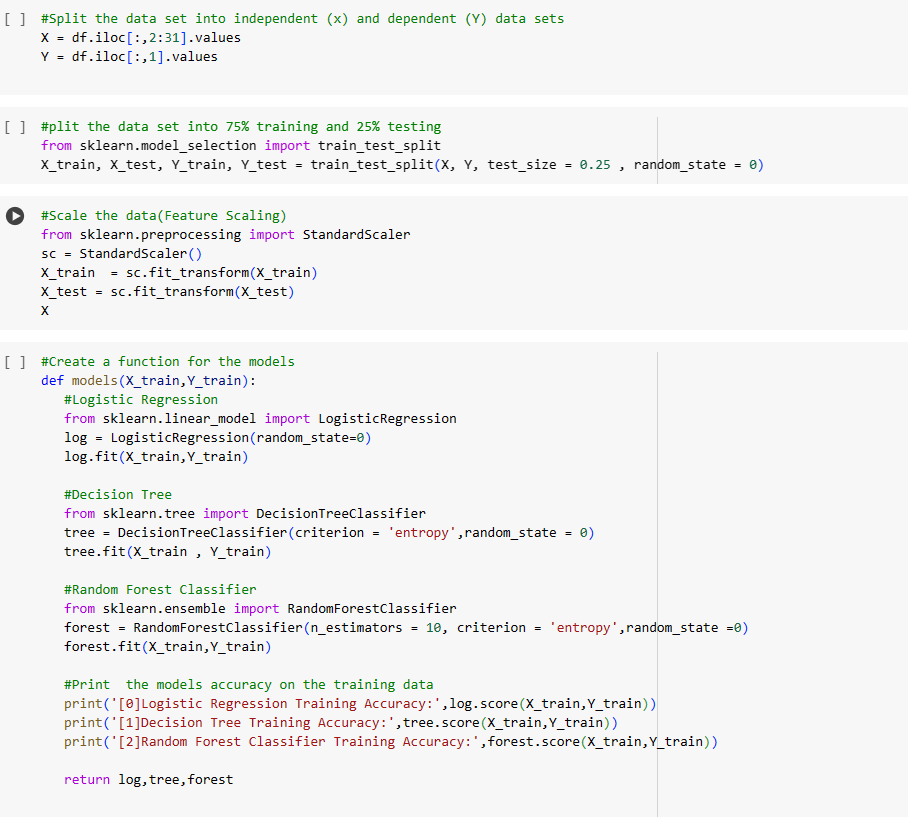
****

**Figure 8.3.6: Visualization using Pair Plot.**

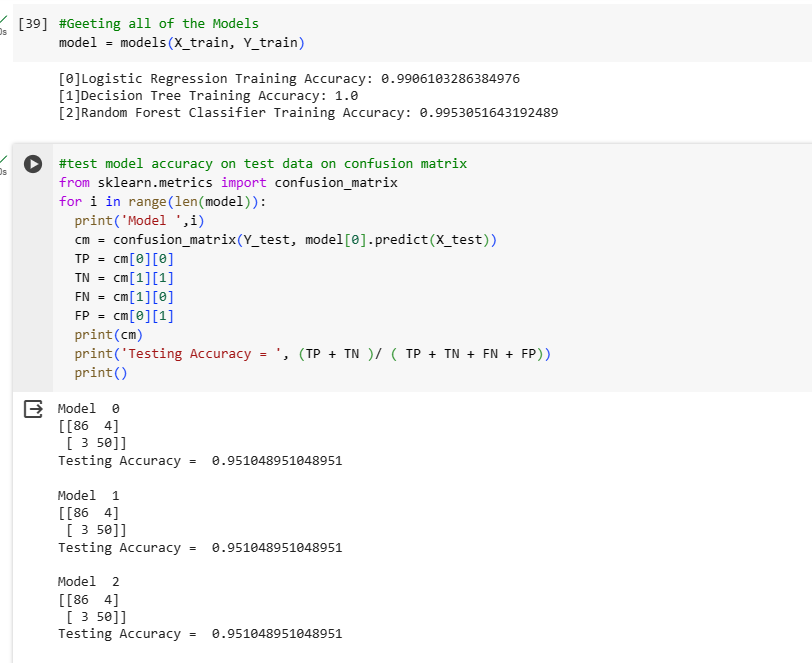
****

****

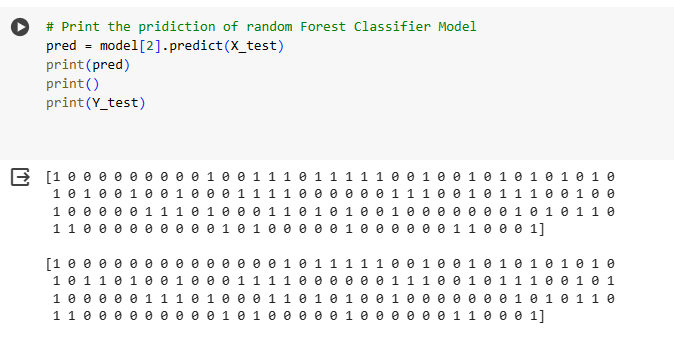
**Figure 8.3.7: Visualization of correlation.**

****

**Figure 8.3.8: Splitting of dataset And Creation of models.**

****

**Figure 8.3.9: Checking Accuracy of Models on Training and testing set**

****

**Figure 8.3.10: Output**

## Chapter-9

### Result and Discussion

Breast cancer is the leading cause of death among middle aged and older women. The present study demonstrates the potential of machine learning (ML) approaches for detecting, analyzing, and classifying breast cancer. Using ML, we were able to evaluate different features of a digitized image of a fine needle aspirate (FNA) of a breast mass made available to researchers by Wolberg . The FNA of a breast mass describes the characteristics of the cell nuclei present in the image. FNA is a type of biopsy procedure where a very thin needle is inserted into an area of abnormal tissue or cells with a guide of computerized tomography (CT) scan or ultrasound monitors. The collected sample is then transferred to a pathologist to study it under a microscope and examine whether cells in the biopsy are normal or abnormal. The results generated based on different feature values indicated that among the 569 patients diagnosed with breast cancer, 63% were benign and 37% were malignant. We found that the mean value of each feature for benign tumor (non-cancerous) is lower when compared to each corresponding feature for malignant tumor (cancerous), suggesting that malignant tumor spread to the other parts of the body . Based on these features, we were able to differentiate between benign and malignant tumors . Cancer cells have the ability to spread to other parts of the body through the blood and lymphatic systems.

Medical researchers and physicians usually identify geometrical features and textural features by viewing biopsy images. Multiple classifiers algorithms are applied on medical datasets to perform predictive analysis about patients and their medical diagnosis. For example, one analysis using a combination of mammograms and ML approaches has led to an accurate diagnosis of breast cancer. Analyses using histopathological images and automatic grading systems have been applied to successfully determine the Gleason grade of breast cancer, and prostate cancer. In addition, several previously published methods have shown the potential of ML methods for automatic breast cancer and prostate cancer detection and grading on digital histopathology images.

## Chapter-10

### Conclusion

Breast cancer is one of the leading causes of mortality among women worldwide and it is important to develop novel approaches to screen, diagnose, and treat breast cancer. This project presents a novel computer-aided diagnosis system for the prediction, diagnosis, and classification of breast cancer using ML.

In particular, we discussed the concepts of ML and outlined its application in the classification of breast cancer. The features including radius, texture, perimeter, area, compactness, concavity, and concave points of the cell image allow us to differentiate between benign and malignant breast cancer.

Other features including smoothness, symmetry, and fractual dimension of the cell image do not indicate a particular preference of one diagnosis over the other. Some benign tumors may progress to malignant tumors. We believe that ML will soon become much more commonplace in many clinical and hospital settings. Our results based on the ML can be translated into tools for future clinical treatment decision-making.

### References

[1] Loizidou, Kosmia, Rafaella Elia, and Costas Pitris. "Computer aided breast cancer detection and classification in mammography: A comprehensive review." Computers in Biology and Medicine (2023): 106554.

[2] Sharmin, Selina, et al. "A hybrid dependable deep feature extraction and ensemble-based machine learning approach for breast cancer detection." IEEE Access (2023).

[3] Z. Jiang, and W. Xu, “Classification of benign and malignant breast cancer based on DWI texture features,” ICBCI 2017 Proceedings of the International Conference on Bioinformatics and Computational Intelligence 2017.

[4] Nover, Adam B., et al. "Modern breast cancer detection: a technological review." Journal of Biomedical Imaging 2009 (2009): 1-14.

[5] Joy, Janet E., Edward E. Penhoet, and Diana B. Petitti, eds. Saving women's lives: strategies for improving breast cancer detection and diagnosis. Washington, DC: national academies press, 2005.

[6] Fear, Elise C., Paul M. Meaney, and Maria A. Stuchly. "Microwaves for breast cancer detection?." IEEE potentials 22.1 (2003): 12-18.

[7] Asri H, Mousannif H, Al Moatassime H, Noel T. Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. In: *Procedia Computer Science*. Vol 83. ; 2016. doi: 10.1016/j.procs.2016.04.224