

A Term Paper on
“BREAST CANCER DETECTION USING MACHINE LEARNING AND DEEP
LEARNING”

Submitted in partial fulfillment of the requirement for the award of the degree

BACHELOR OF SCIENCE IN PHYSICS, MATHS, COMPUTER SCIENCE

SUBMITTED BY

DIJIN DOMINIC (20PMC31)

UNDER THE GUIDANCE OF

DR. JAYATHI BHABRA



ST. JOSEPH'S COLLEGE(AUTONOMOUS)

BENGALURU CITY UNIVERSITY

DECLARATION BY THE CANDIDATE SUBMITTING THE REPORT



I, Dijin Dominic (20PMC31), an Undergraduate student in the Department Physical Science declare that the work embodied in this term paper report is a result of our own Bonafede work carried out with our personal effort and submitted by us under the supervision of Dr. Jayati Bhadra at St. Joseph's College (Autonomous), Bengaluru. I declare that I have faithfully acknowledged and given credit and referred to the researchers wherever their works have been cited in the body of the thesis.

Station: Bengaluru

Date: 23-10-2022

DIJIN DOMINIC

ACKNOWLEDGEMENT

I thank Rev. Dr. Fr. Victor Lobo, SJ, Principal of St. Joseph's College (Autonomous), who constantly motivates and believes that every Josephite has the potential to reach world standards.

I especially thank Dr. Jayati Bhadra, Head of Department, Department of Big Data Analytics, for her ever-encouraging attitude and support in every possible way and my hearty sense of gratitude for her constant motivation, valuable suggestions, and excellent guidance during the research. I take this opportunity to thank all those who have helped me in bringing this project successfully. Above all, I submit my thoughts and actions to the almighty, who always guides me in the path of righteousness and whose blessings have enabled me to complete this work successfully.

Place: Bengaluru

Date: 23-10-2022

DIJIN DOMINIC

LIST OF FIGURES

<i>Fig. 1: Parts of a Breast</i>	<i>8</i>
<i>Fig. 2: Output of the describe method()</i>	<i>20</i>
<i>Fig. 3: Plot of Malignant and Benign Cases v/s count</i>	<i>21</i>
<i>Fig. 4: Data of first 5 patients after dropping two columns</i>	<i>21</i>
<i>Fig. 5: Total number of null values in the dataset</i>	<i>22</i>
<i>Fig. 6: Heatmap showing the correlation between the attributes.....</i>	<i>24</i>
<i>Fig. 7: Scatter Plot of Mean Values</i>	<i>25</i>
<i>Fig. 8: Upper Triangle of Seaborn heatmap</i>	<i>27</i>
<i>Fig. 9: Total columns after dropping reductant attributes.....</i>	<i>28</i>
<i>Fig. 10: Heatmap displaying correlation with fewer reductant attributes</i>	<i>29</i>

LIST OF ABBREVIATIONS

1. ML : Machine Learning
2. LR: Linear Regression
3. DT : Decision Tree
4. RF: Random Forest.
5. KNN: K Nearest Neighbour
6. SVM: Support Vector Machine
7. NB: Naive Bayes

ABSTRACT

Breast cancer is a common factor today. Despite this, not all general hospitals are capable of diagnosing breast cancer by mammography. The longer you wait to be diagnosed with breast cancer, the more likely it is to spread. A computer-assisted breast cancer diagnosis has therefore been developed to reduce the time required to diagnose breast cancer and reduce mortality. This paper summarizes breast cancer diagnostic investigations using six machine-learning algorithms and methods used to improve cancer prediction accuracy.

TABLE OF CONTENTS

1. INTRODUCTION

1. What is Cancer?.....	8
2. Breast Cancer.....	9
3. Types of Breast Cancer.....	10
4. Risks for Breast Cancer.....	11
5. Breast Cancer in the World and India.....	12

2. REVIEW OF LITERATURE.....13

3. METHODOLOGY USED.14

1. Machine Learning Algorithms.....	14
2. Exploring Machine Learning Algorithms.....	15
3. Dataset Description.....	19
4. Exploring about Attributes.....	19
5. Into the Data Analysis Part.....	21
6. Building Models.....	31

4. Result Analysis.....33

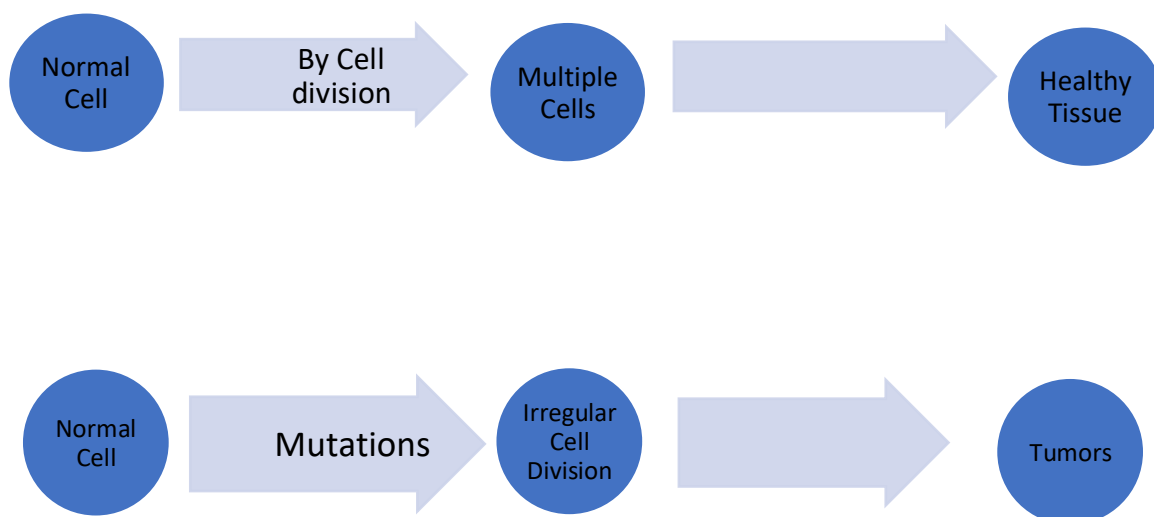
5. Conclusion.....34

6. References.....35

INTRODUCTION

What Is Cancer?

Our body is made up of cells. The division of these cells gives rise to tissues, organs, and organ systems. Cancer is a condition when a few of the body's cells grow out of control and spread to other regions of the body. Normally, Human cells (excluding germ cells) grow and multiply through a process called Cell division (Mitosis). Generally, a new cell replaces an old one when the cells die or are damaged. Sometimes, this structured approach fails, causing damaged cells to develop when they shouldn't. This leads to Tumors or lumps of tissue. These tumors are of two types Benign and Malignant. Benign tumors are less problematic compared to Malignant tumors. Benign tumors stay in their location whereas Malignant tumors spread to nearby tissues or even can travel to distant places in our body causing the formation of different tumors.



Cancer is a genetic disorder. It is caused due to the mutations in genes that control cell division and growth. Mutations can be caused by carcinogens like tobacco smoke, UV rays, and even by some viruses like Human Papilloma Virus, etc... Cancer is hereditary if the mutated gene is present in the sperm or egg of the parent. Inheriting a cancer gene doesn't mean that the person will get cancer. It will just increase the risk of getting cancer.

Breast Cancer

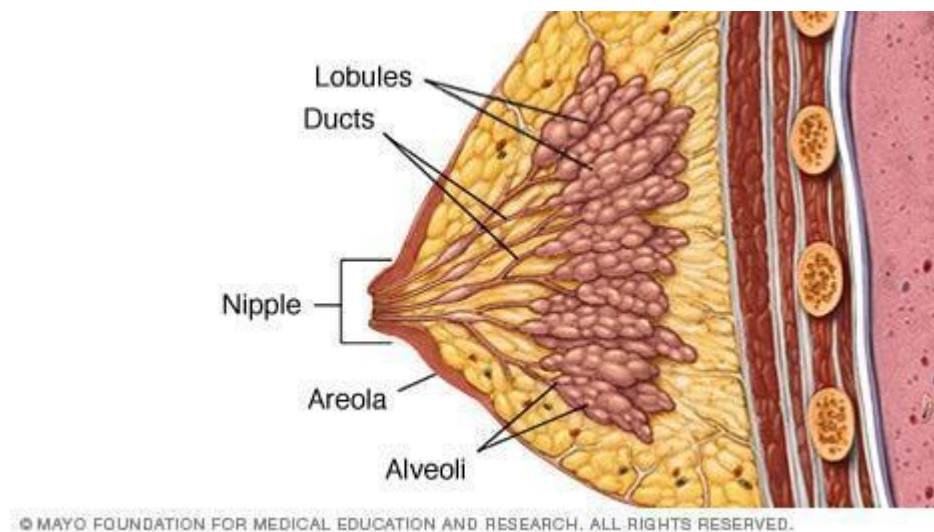
According to the U.S. Cancer Statistics Working Group, Breast cancer is the second major cause of women's death after lung cancer. Early diagnosis is very important to keep the survival rate greater than the mortality rate. Treatment for breast cancer can be effective, especially if the disease is discovered early (1). There are various Machine Learning algorithms that can predict Breast Cancer. Finding the most suitable and appropriate algorithm for the prediction of breast cancer is one of the important tasks.

Breast Cancer is not an infectious disease. Breast Cancer can be caused due to various reasons which are discussed in the later part of this paper. Aging, being overweight, drinking alcohol, radiation exposure history, family history of breast cancer, family history of radiation exposure, reproductive history (such as the age of first pregnancy and age of first menstrual period), smoking, and postmenopausal hormone therapy are a few of the risk factors.

Breast cancer is a disease in which cells in the breast grow out of control. There are different kinds of breast cancer. The kind of breast cancer depends on which cells in the breast turn into cancer. Breast cancer can begin or be malignant in different parts of the breast (2). The structure of the female breast is complex — including fat, glandular and connective tissue, as well as lobes, lobules, ducts, lymph nodes, blood vessels and ligaments (3). The ducts are

tubes that carry milk to the nipple. The connective tissue (which consists of fibrous and fatty tissue) surrounds and holds everything together.

Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized or the tumor is said to be Malignant.



Fig(1): Parts of a Breast

[Types of Breast Cancer](#)

According to the American Cancer Society, A breast cancer type is determined by the specific cells in the breast that become cancer. Breast Cancer can be divided into these ways,

I. Ductal or lobular carcinoma.

Most breast cancers are carcinomas, which are tumors that start in the epithelial cells that line organs and tissues throughout the body. When carcinomas form in the breast, they are usually a more specific type called adenocarcinoma, which starts in cells in the milk ducts or the glands in the breast that make milk (Lobules).

II. In situ and invasive breast cancers.

The type of breast cancer refers to whether cancer has spread or not. In situ breast cancer (ductal carcinoma in situ or DCIS) is a pre-cancer that starts in a milk duct and has not grown into the rest of the breast tissue. The term invasive breast cancer is used to describe any type of breast cancer that has spread into the surrounding breast tissue.

III. Special types of invasive breast cancers.

Triple Negative and Invasive Breast Cancer are special invasive breast cancers that are less common but can be more serious than other types of breast cancer.

IV. Less common types of breast cancer.

Paget Disease of Breast, Angiosarcoma, Phyllode tumors are types of breast cancers that start to grow in other types of cells in the breast. These cancers are much less common in the population.

[Risks for Breast Cancer](#)

According to the Centers for Disease Control and Prevention, the risk for breast cancer is due to a combination of factors. The main factors that influence your risk include being a woman and getting older. Most breast cancers are found in women who are 50 years old or older. Some women will get Breast Cancer even without any risk factors. Women older than 50 years of age are advised to visit doctors at regular intervals or take Mammograms or other tests under the advice of a medical practitioner.

Risk factors that cannot be altered	Risk factors that can be altered
Getting Older	Not being physically active.
Genetic Mutations	Being overweight or having obesity after menopause
Reproductive history (Periods and Menopause)	Taking hormones
Having Dense Breast	Drinking alcohol
Personal history of breast cancer or certain non-cancerous breast diseases	Never having a full-term pregnancy
Family history of breast or ovarian cancer.	Having the first pregnancy after age 30, not breastfeeding
Previous treatment using radiation therapy	Eating unhealthy food and unnecessary medications

Breast Cancer in the World and in India.

With an estimated 1.67 million new cancer cases identified in 2012, breast cancer is the most prevalent cancer in women worldwide, accounting for almost a quarter (25%) of all cancer cases. Women in less developed regions experience slightly more cases than those in more developed countries. Although India's age-adjusted incidence rate of breast cancer is lower than the United Kingdom's, the death is comparable to the United Kingdom.

According to both international and Indian studies, the incidence of morbidity and death linked to cancer have significantly increased in the Indian subcontinent. Previously, cervical

cancer was the most frequent type of cancer among Indian women, but now it's breast cancer. Breast cancer has ranked number one cancer among Indian females with age adjusted rate as high as 25.8 per 100,000 women and mortality 12.7 per 100,000 women (4).

REVIEW OF LITERATURE

Breast Cancer is one of the most dangerous cancers among women. In all countries, many people including men are affected by Breast cancer. The rate of occurrence of breast cancer is less in Men but still, men are affected by it. . It is the second largest disease that is responsible for women's death. Traditional cancer detection methodologies are based on the “Gold Standard” method, which consists of three tests: clinical examination, radiographic examination, and pathological examination (5). There are many research papers available on the internet which did the analysis on the Wisconsin Breast Cancer dataset. The dataset was publicly made available in the 1990s. In many countries, late diagnosis of Breast Cancer is increasing the mortality rate of women. Hence early detection of Breast cancer is so crucial. Predicting diseases using machines will be a landmark in the medical sector. In the case of Breast cancer, as early detection plays a major role in curing the disease, machine learning algorithms help a lot in predicting the probability of the disease.

Many researchers have published papers that provide a comparative analysis of machine learning algorithms such as Support Vector Machine, K Nearest Neighbour, Naive Bayes, Decision Tree, K-means and Artificial Neural Networks. This implies that most of the research papers used a combination of Linear, non-linear algorithms.

Non-linear algorithms include Random Forest, Naive Bayes, Support Vector Machine and K Nearest Neighbor etc.. . In all research papers the results show that SVM or Support Vector Machine has the highest accuracy rate. In the research paper by M. K. Keles we can see that the accuracy rate of Random forest Algorithm is 92.2%. (6).

S. K. Maliha et al. (7) used Naive Bayes, K Nearest Neighbours and J48 algorithm for the prediction of breast cancer and other 8 cancers. They found that NB and KNN have a higher accuracy rate than the J48 decision tree algorithm.

A. Bharat et al. (8) used SVM, DT, NB, and KNN algorithms for predicting breast cancer. The result showed that SVM has an accuracy of 97.13% with less execution time. It is also found that SVM has higher execution time than KNN algorithm.

METHODOLOGY USED

Machine Learning Algorithms.

The utilization of data science and machine learning approaches in medical fields proves to be prolific as such approaches may be considered of great assistance in the decision-making process of medical practitioners [9]. Machine Learning algorithms are the efficient way to predict something based on the trained information from datasets. There are different types of learning algorithms. The best and most efficient algorithm is chosen based on the task and the available dataset. In our dataset, we are classifying whether the cancer is Benign(B) or Malignant(M). In this paper, our main focus is to find the best suitable algorithms for the prediction of Breast Cancer by comparing their accuracy rates. Machine Learning algorithms are of different types. Mainly it is classified into Linear, Non-linear, and ensemble algorithms. Linear algorithms include Linear Regression, Logistic Regression, Linear Discriminant Analysis, nonlinear includes Classification and Regression Tree, Naive

Bayes, K-Nearest Neighbor, Support Vector Machine and ensemble algorithms include Decision Tree, Random Forest, Boosting and AdaBoost.

[Exploring Machine Learning Algorithms.](#)

Machine Learning Algorithms are primarily divided into three.

1. Supervised Learning Algorithms

Algorithms that are trained and tested under the supervision/ guidance from the external are referred to as Supervised Learning Algorithms. A great example of Supervised Learning is the classification of whether the mail is spam or not.

2. Unsupervised Learning Algorithms

3. Reinforcement Algorithms

Supervised Learning Algorithm includes:

A. Regression

Regression algorithms are used if there is a relationship that exists between input and output variables.

- a. Linear
- b. Non-Linear
- c. Polynomial

B. Decision Tree

C. Random Forest

D. Classification

- E. Classification algorithms are used when the output can be categorized. For example,
If the outputs are Yes/No, Male/Female, True/False, etc...

- a. *KNN*
- b. *Tress*
- c. *Logistic Regression*
- d. *Naive-Bayes*
- e. *SVM (Support Vector Machines)*

Unsupervised Learning Algorithm includes:

A. Clustering

- a. *SVD*
- b. *PCA*
- c. *K-means*

B. Association Analysis

- a. *Apriori*
- b. *FP-Growth*

C. Hidden Markov Model.

SUPERVISED MACHINE LEARNING ALGORITHM	NON-SUPERVISED MACHINE LEARNING ALGORITHM	REINFORCEMENT MACHINE LEARNING ALGORITHM
A technique of accomplishing a task by giving the training, input and output labelled data patterns to the system.	A self-learning technique in which the system has to discover the patterns based on the provided unlabelled dataset.	A learning technique in which the algorithm does trial and error method to complete a task

Email Spam Detection with training dataset of already defined labelled emailed is an example of a training dataset.	Clustering similar documents based on textual content is an example of non-supervised learning.	Provides an approach to AI.
Used for Prediction.	Used for analysis.	Interactive and adaptive in nature.

In this research paper, we will be discussing 6 algorithms for Breast Cancer Prediction. They are Logistic Regression, Decision Tree Model, Random Forest Algorithm, K Nearest Neighbour(K NN), Support Vector Machine and finally Naive Bayes.

1) Logistic Regression:

This is a supervised learning algorithm with many dependent variables. The response of this algorithm is in binary format. Logistic regression can provide continuous results for specific data. This algorithm consists of a statistical model with binary variables.

2) Decision Tree Model:

This algorithm comes under the Classification and Regression model algorithm. Here the given dataset is divided into various subsets and thereby predicted with a higher level of accuracy.

3) Random Forest Algorithm:

This algorithm comes under the classification of Supervised learning.

RF algorithm is used to solve problems related to classification and regression. It is considered the building block of machine learning that is used to predict new data on the basis of the old available data.

4) K Nearest Neighbour(K NN)

As the word implies this algorithm will look for patterns in the neighbouring data.

This is an excellent approach for predicting breast cancer. Each class is given equal weightage to see the pattern. K Nearest Neighbour extracts similarly labelled data from large data sets and classifies large datasets based on feature similarity.

5) Support Vector Machine

It is also called the SVM algorithm. Similar to RF algorithms, this also comes under the category of Supervised Learning which is used for solving problems which are based on classification and regression. It consists of theoretical and numerical functions for solving the regression problem. It offers the highest accuracy rate when predicting large datasets. This is a powerful machine learning technology based on 3D and 2D modelling.

6) Naive Bayes Algorithm

It is also called the NB algorithm. Calculation of probability is done using this method. Provides the highest accuracy in computing probabilities for noisy data used as input. This is the analogy classifier used to compare the training data set and training tuples. (10)

Data Set Description

The Breast Cancer Wisconsin (Diagnostic) Data Set is publicly available and obtained from the UCI Machine Learning Repository [11]. The dataset is created by Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, W. Nick Street, Computer Sciences Dept., the University of Wisconsin and Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. This dataset can be also found on UCI Machine Learning Repository (11). The dataset contains twenty attributes. They are Radius mean, Texture Mean, Perimeter mean, Area mean, Smoothness Mean, Compactness Mean, Concavity mean, Concave points mean, Symmetry SE, Fractal Dimension SE, Radius Worst, Texture Worst, Perimeter worst, Area worst, Smoothness Worst, Compactness Worst, Concavity Worst, Concave point worst, Symmetry worst and Fractal Dimension Worst.

Exploring about Attributes

1. Radius: Each cell has a radius. Radius defines the distances from the centre to the perimeter of the breast cell mass.
2. Texture: Defines the mean of the standard deviation of grey-scale values.
3. Perimeter and Area: Denotes the perimeter and area of the respective cell nuclei.
4. Smoothness: Mean of local variation in the radius length.
5. Compactness: It refers to the cell nucleus's compactness. It is given by the formula:
$$\text{Compactness} = (\text{Perimeter}^2)/\text{Area}-1$$
6. Concavity: It refers to the concave portions in the contour plot.
7. Concave points: The number of concave portions of the contour plot of the nucleus.

8. Symmetry: Finding the longest line across the center of the nucleus from the boundary point is the first step in determining symmetry. The relative lengths of the lines parallel to the longest line parallel to the boundary in both directions are then measured. The nucleus through which the longest line passes should receive special consideration. Symmetry refers to the cell nucleus's shape.
9. Fractal Dimension is approximated using the formula

Coastline Approximation – 1

The attributes Radius and Area refers to the size of the nucleus whereas the shape is denoted by the attributes Smoothness, Concavity, Compactness, Concave points, Symmetry and fractal dimension. Here the attribute Perimeter denotes both the shape and size of the nucleus. For each of the attributes, the respective mean, Standard error, and Worst values are calculated. Worst values refer to the mean of the large values of the given attribute which is obtained from the Ultrasound fine Needle aspirate images. The mean and Standard Error is given by,

$$Mean_x = 1 \div N \times \sum_{i=0}^N x_i$$

where x is the attribute, N is the total set of observations of the attribute x and x_i is the i^{th} set of observations.

Similarly, Standard Deviation is given by the equation:

$$Standard\ Error_x = sd \div \sqrt{N}$$

Where N is the number of observations for the attribute x,

sd is the standard deviation

Therefore, including the ID number and the diagnosis, we have a total of 33 attributes. ID number is the unique number given to each person. Diagnosis classifies the patients into two categories. The patient with Benign cancer and the person with Malignant cancer. All the attribute values are taken with four significant figures.

[Into the Data Analysis Part](#)

The dataset is a CSV file of 123kb named as data. The first step is to import the required library. Hence we are importing pandas as pd, NumPy as np, and seaborn as sns. Pandas is an open-source python package that is used for data analysis and tasks of machine learning. NumPy is also an open-source library that is used for the approach of Linear algebra and other few mathematical operations. Seaborn is a data visualization library used for plotting data. Seaborn is used because it is more comfortable in handling data frame.

Pandas has two main data structures

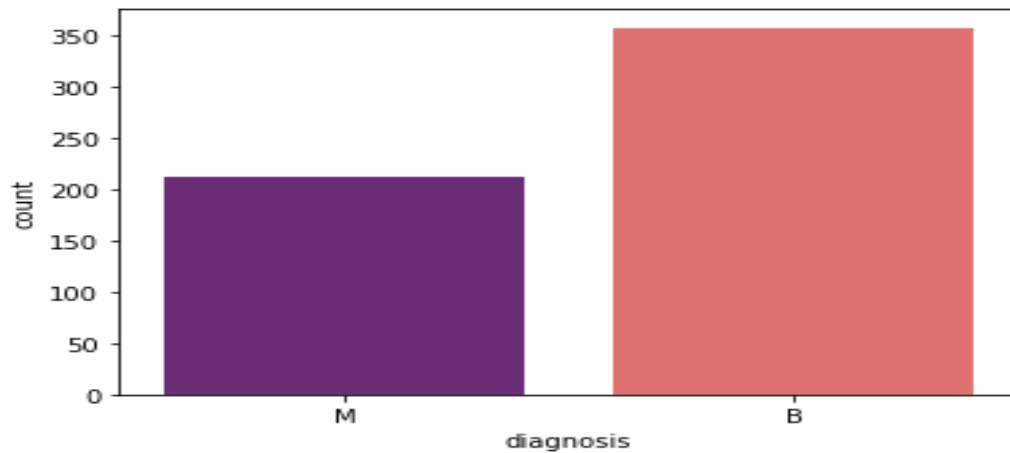
1. Series: One could think of a series as a one-dimensional array. Any data type, such as texts, integers, floats, and Python objects, can be stored in the Series.
2. Data Frames: A table and the data frame data structure are comparable. That is, it is two-dimensional. A simple data frame will consist of three components Rows, Columns, and data.

In this paper, as the obtained data is in the form of a table, we can use data frames to do the analysis. The CSV file data is loaded using the python code. To know more about the data, we use the shape keyword which returns the number of rows and columns and the output is "(569, 33)". It implies that we have 569 rows and 33 columns in our dataset.

	count	mean	std	min	25%	50%	75%	max
id	569.0	3.037183e+07	1.250206e+08	8670.000000	869218.000000	906024.000000	8.813129e+06	9.113205e+08
radius_mean	569.0	1.412729e+01	3.524049e+00	6.981000	11.700000	13.370000	1.578000e+01	2.811000e+01
texture_mean	569.0	1.928965e+01	4.301036e+00	9.710000	16.170000	18.840000	2.180000e+01	3.928000e+01
perimeter_mean	569.0	9.196903e+01	2.429898e+01	43.790000	75.170000	86.240000	1.041000e+02	1.885000e+02
area_mean	569.0	6.548891e+02	3.519141e+02	143.500000	420.300000	551.100000	7.827000e+02	2.501000e+03
smoothness_mean	569.0	9.636028e-02	1.406413e-02	0.052630	0.086370	0.095870	1.053000e-01	1.634000e-01
compactness_mean	569.0	1.043410e-01	5.281276e-02	0.019380	0.064920	0.092630	1.304000e-01	3.454000e-01
concavity_mean	569.0	8.879932e-02	7.971981e-02	0.000000	0.029560	0.061540	1.307000e-01	4.268000e-01
concave points_mean	569.0	4.891915e-02	3.880284e-02	0.000000	0.020310	0.033500	7.400000e-02	2.012000e-01
symmetry_mean	569.0	1.811619e-01	2.741428e-02	0.106000	0.161900	0.179200	1.957000e-01	3.040000e-01
fractal_dimension_mean	569.0	6.279761e-02	7.060363e-03	0.049960	0.057700	0.061540	6.612000e-02	9.744000e-02
radius_se	569.0	4.051721e-01	2.773127e-01	0.111500	0.232400	0.324200	4.789000e-01	2.873000e+00
texture_se	569.0	1.216853e+00	5.516484e-01	0.360200	0.833900	1.108000	1.474000e+00	4.885000e+00
perimeter_se	569.0	2.866059e+00	2.021855e+00	0.757000	1.606000	2.287000	3.357000e+00	2.198000e+01
area_se	569.0	4.033708e+01	4.549101e+01	6.802000	17.850000	24.530000	4.519000e+01	5.422000e+02
smoothness_se	569.0	7.040979e-03	3.002518e-03	0.001713	0.005169	0.006380	8.146000e-03	3.113000e-02
compactness_se	569.0	2.547814e-02	1.790818e-02	0.002252	0.013080	0.020450	3.245000e-02	1.354000e-01
concavity_se	569.0	3.189372e-02	3.018606e-02	0.000000	0.015090	0.025890	4.205000e-02	3.960000e-01
concave points_se	569.0	1.179614e-02	6.170285e-03	0.000000	0.007638	0.010930	1.471000e-02	5.279000e-02
symmetry_se	569.0	2.054230e-02	8.266372e-03	0.007882	0.015160	0.018730	2.348000e-02	7.895000e-02
fractal_dimension_se	569.0	3.794904e-03	2.646071e-03	0.000895	0.002248	0.003187	4.558000e-03	2.984000e-02
radius_worst	569.0	1.626919e+01	4.833242e+00	7.930000	13.010000	14.970000	1.879000e+01	3.604000e+01
texture_worst	569.0	2.567722e+01	6.146258e+00	12.020000	21.080000	25.410000	2.972000e+01	4.954000e+01
perimeter_worst	569.0	1.072612e+02	3.360254e+01	50.410000	84.110000	97.660000	1.254000e+02	2.512000e+02
area_worst	569.0	8.805831e+02	5.693570e+02	185.200000	515.300000	686.500000	1.084000e+03	4.254000e+03
smoothness_worst	569.0	1.323686e-01	2.283243e-02	0.071170	0.116600	0.131300	1.460000e-01	2.226000e-01
compactness_worst	569.0	2.542650e-01	1.573365e-01	0.027290	0.147200	0.211900	3.391000e-01	1.058000e+00
concavity_worst	569.0	2.721885e-01	2.086243e-01	0.000000	0.114500	0.226700	3.829000e-01	1.252000e+00
concave points_worst	569.0	1.146062e-01	6.573234e-02	0.000000	0.064930	0.099930	1.614000e-01	2.910000e-01
symmetry_worst	569.0	2.900756e-01	6.186747e-02	0.156500	0.250400	0.282200	3.179000e-01	6.638000e-01
fractal_dimension_worst	569.0	8.394582e-02	1.806127e-02	0.055040	0.071460	0.080040	9.208000e-02	2.075000e-01
Unnamed: 32	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig(II): Output of the describe method()

To obtain the description of the dataset we use the describe() method. It returns the description of the dataset such as count, mean, etc... The output after transposing is given in the figure(II).As we know, our task is to predict whether the cancer is Benign or Malignant. Hence let's find out the number of Benign and Malignant cases using a plot. For that we are using seaborn data visualization library. *The plotted graph is:*



Fig(III):Plot of Malignant and Benign Cases v/s count

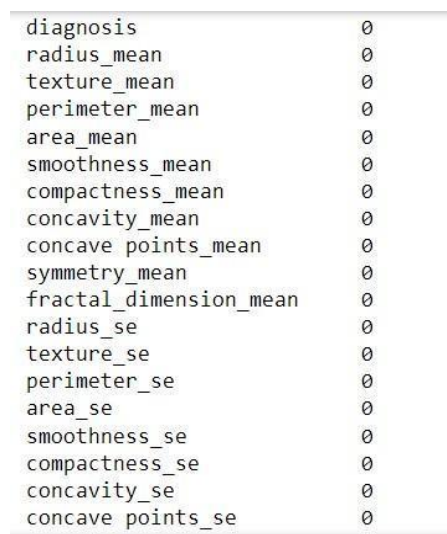
In the fig(III), it is observed that Benign cases are more than Malignant cases. To get the specific count we can use the “value_counts()” method. This gives the output that there are 357 Benign cases and 212 Malignant cases. As it is already mentioned there are 33 attributes out of which we can remove the attributes “id” and: Unnamed 32” by dropping them. We can drop those columns. *After dropping the dataset of first 5 patients(because of using head() method the dataset with first 8 attributes is shown in fig(IV)*

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869
2	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974
3	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414
4	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980

5 rows × 9 columns

Fig(IV): Data of first 5 patients after dropping two columns

To perform a machine learning task or to train an algorithm we have to convert the attribute 'diagnosis' values M and B to 1 and 0 respectively because it will not understand M and B. Hence if the value of the Diagnosis attribute is 1 the cancer is Malignant and if it's 0 then it's Benign. Null values in a dataset affects proper prediction. To confirm whether we have got any null values in a dataset we use the method 'isnull()' with the count of null values given by sum() method.



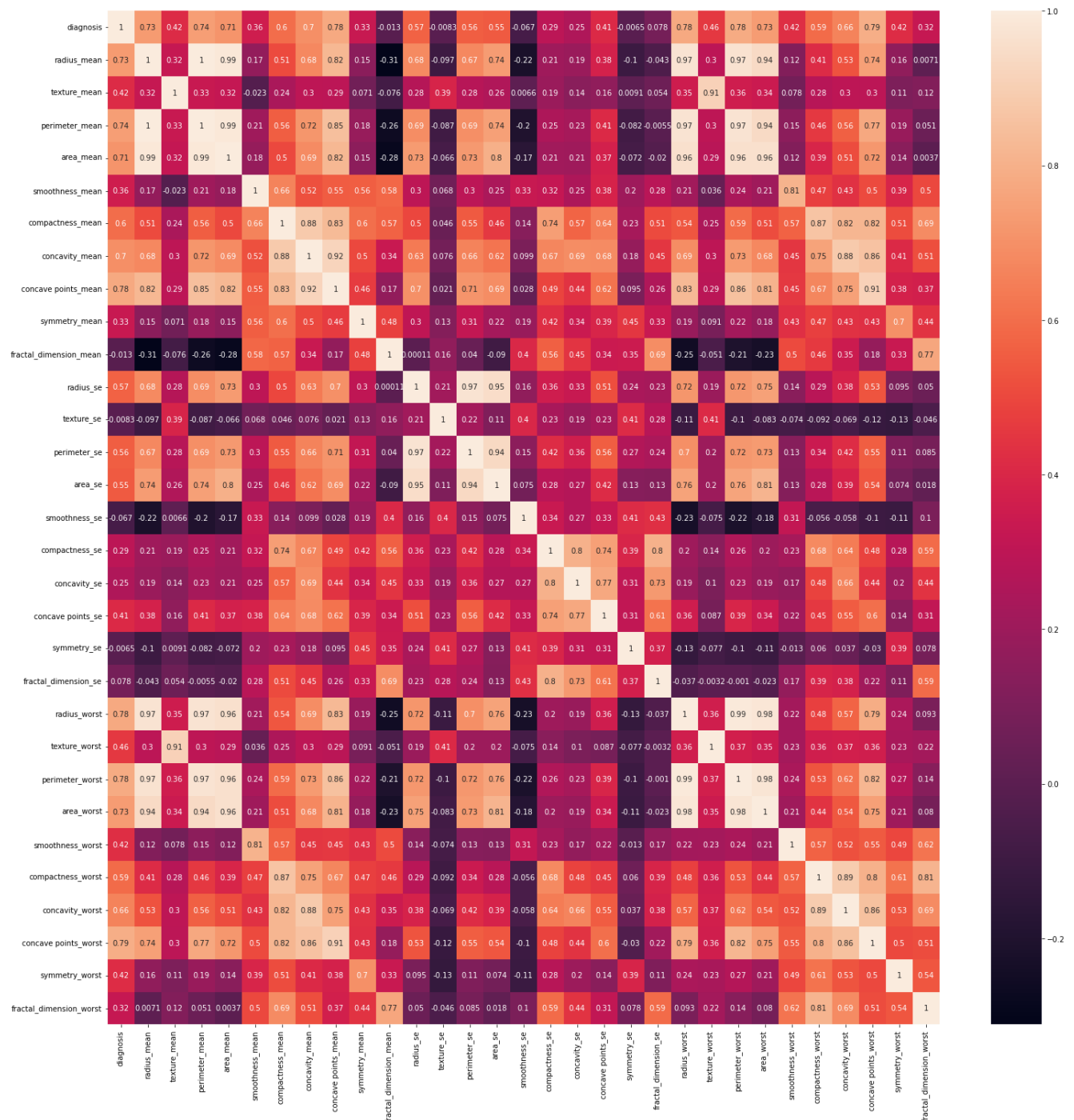
diagnosis	0
radius_mean	0
texture_mean	0
perimeter_mean	0
area_mean	0
smoothness_mean	0
compactness_mean	0
concavity_mean	0
concave points_mean	0
symmetry_mean	0
fractal_dimension_mean	0
radius_se	0
texture_se	0
perimeter_se	0
area_se	0
smoothness_se	0
compactness_se	0
concavity_se	0
concave points_se	0

Fig(V) : Total number of null values in the dataset

The output shows that there are zero null values in all attributes. That is the dataset has no null values. A part of the output is fig(V).

Now we have to check for, we can move to the correlation part. A group of attributes or another attribute whose information content is partially contained by another attribute is called a Correlated Attribute. If the correlation is high it implies that the attributes are more interdependent or it says that reductant attributes exist. The reductant attribute is the attribute which is derived from other attributes which led to high correlation. Hence, we

have to clean the data in such a way that there are no redundant attributes. For that let's check the correlation between the attributes by plotting a heatmap using Seaborn. For that we first use `figure()` method to set the plot size. Then use the '`sns.heatmap`' functionality to obtain the colour coded heatmap. This will give us a heatmap. It is a 2D graphical representation of data with the values in the matrix represented as colour. The output for the correlation between attributes in our dataset obtained in fig (VI).

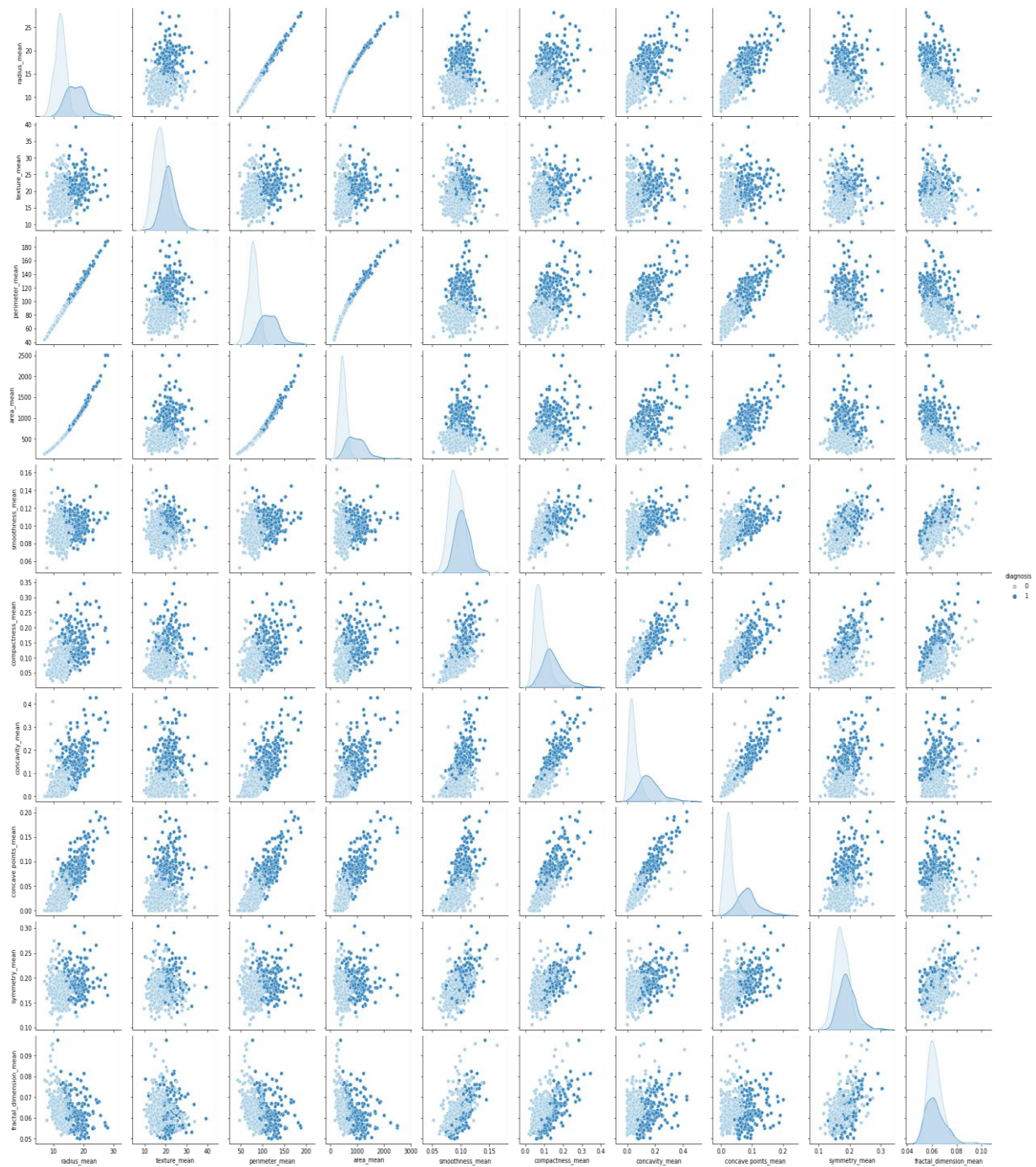


Fig(VI): Heatmap showing the correlation between the attributes

In this heatmap, we can see that a high correlation exists between many attributes. These attributes should not be considered while predicting the types of Cancer. For better clarity let's do the scatter plot of the 'mean' values. Scatter plot is used to plot data points on the horizontal and vertical axes to show how much one variable is influenced by another.

Therefore we can find the existence of correlation between the attributes. Scatter plot of mean values is plotted by creating a data frame of mean values and using 'sns.pairplot' functionality.

This will result in the following output as in fig (VII)

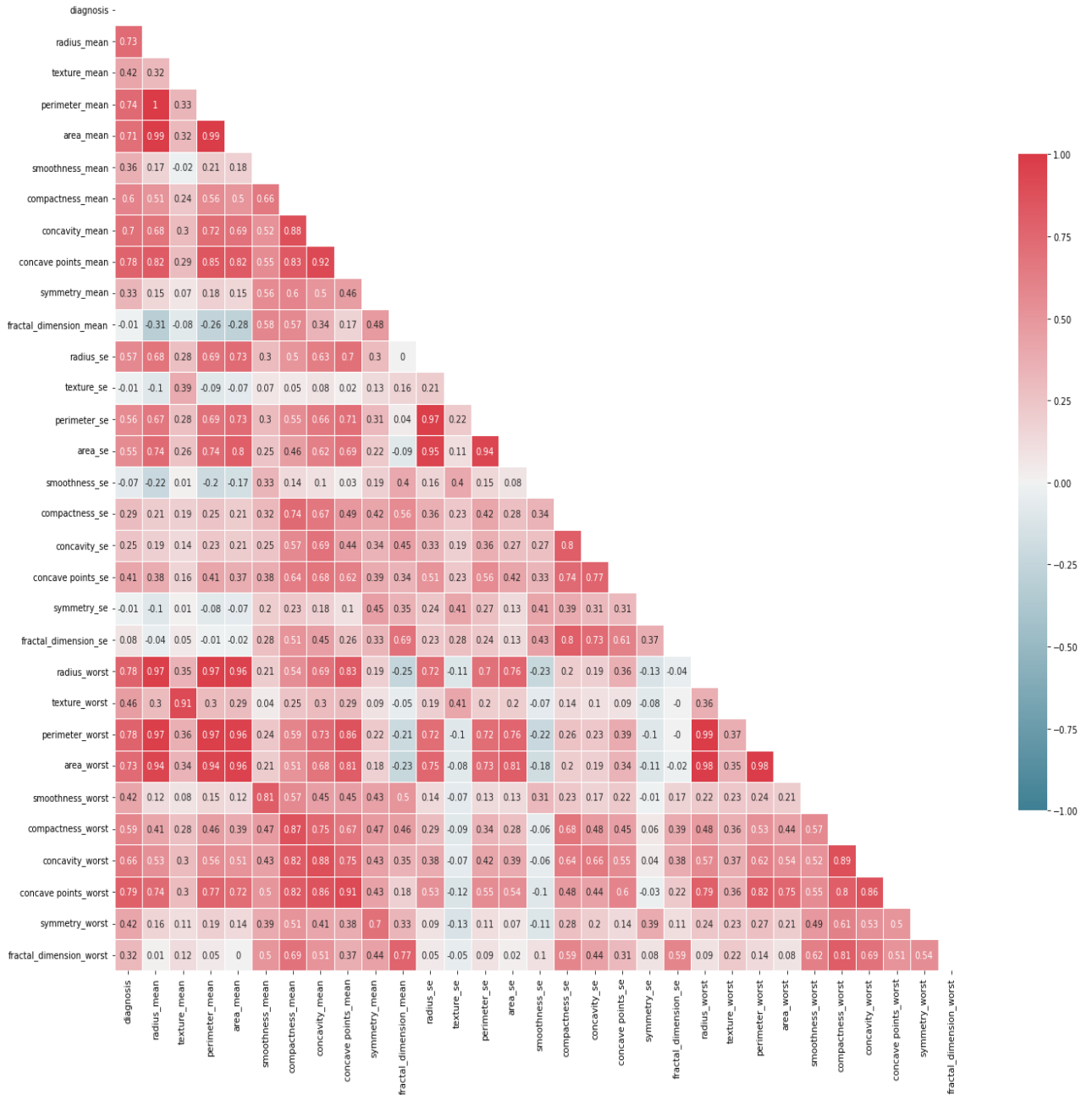


Fig(VII): Scatter Plot of Mean Values

If we closely observe these scatter plots we can see that some plots are highly linear whereas some are not. High multicollinearity exists between radius, area and perimeter mean. Multicollinearity is a statistical concept in which several independent variables are correlated in the model. It is decided based on the value of the Correlation Coefficient. The heatmap plotted above displays the value of the correlation coefficient in different colors. Its value ranges from -1 to 1. A correlation coefficient is a statistical measure of the strength of a linear relationship between two variables. Its value ranges from -1 to 1. A correlation coefficient of -1 indicates a perfect negative or inverse correlation. In this correlation, the values of one series increase and the values of the other series decrease, and vice versa. A factor of 1 indicates a perfect positive or direct relationship. A correlation coefficient of 0 means that there is no linear relationship [12]. The attributes concavity, concave_points and compactness are also showing multicollinearity but less compared to the attributes mentioned above.

To know more about the attribute's relation, let's look into the upper triangle. The Upper triangle is drawn using the correlation matrix and using the data visualization library Seaborn. The Upper triangle is plotted by tight layout () method as shown in fig(VIII).

From this visualization of the correlation matrix, we can verify that multicollinearity exists between some variables. Example, here we can see that the radius mean has a correlation of .99 with the area mean and 1 with the perimeter mean. It is because all these three attributes reflect the physical characteristics of the cell(nuclei).



Fig(VIII): Upper Triangle of Seaborn heatmap

Therefore for the prediction purpose, we have to pick one of the three attributes to reduce reductant attributes in our dataset. Similarly, we can notice that a high correlation exists between radius mean and radius_worst column, texture mean and texture worst column, perimeter mean and area_worst or perimeter worst etc... Therefore, to reduce the multicollinearity we can drop the 'mean' and 'worst' columns. Also, it is noticeable that

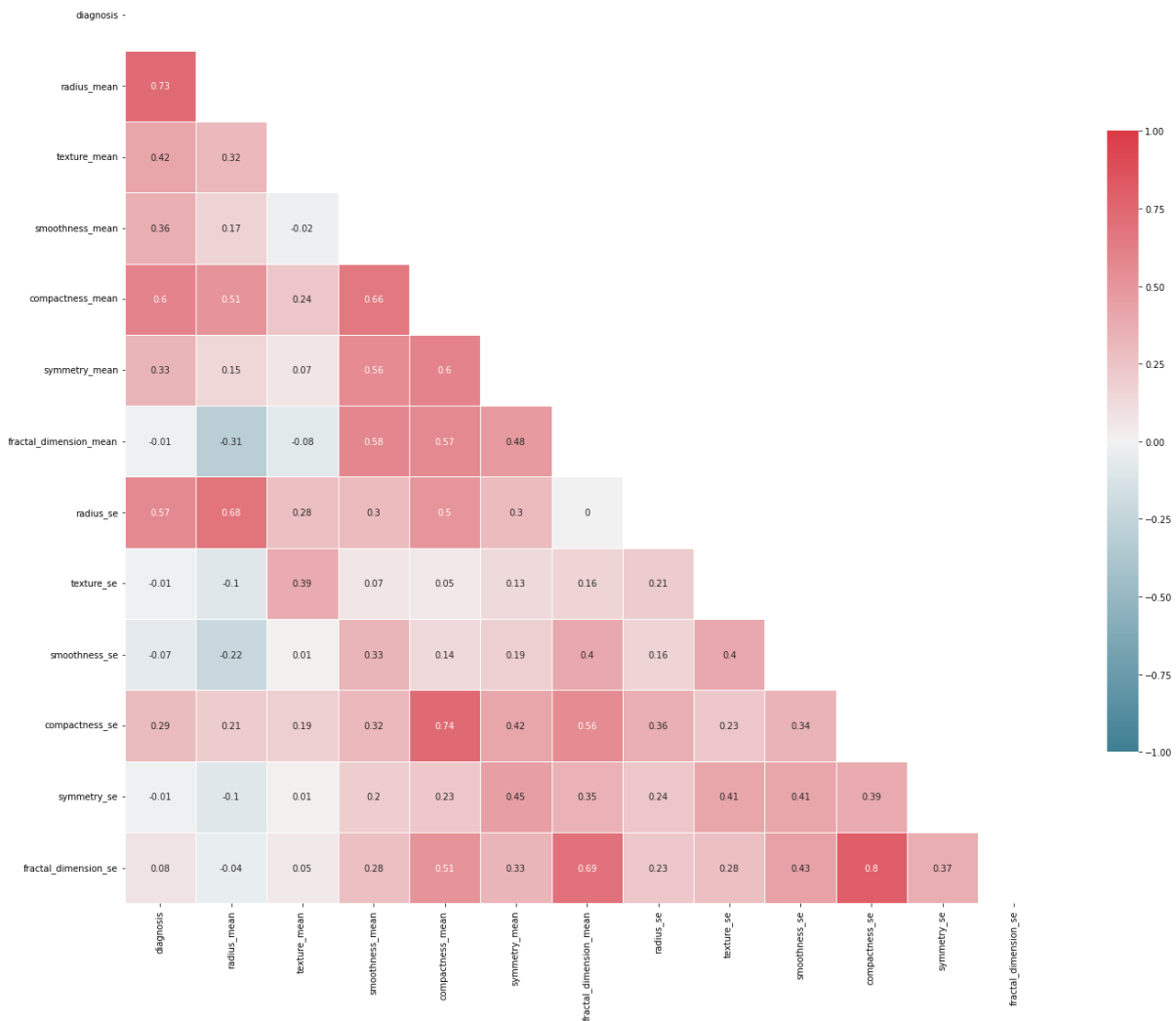
there is multicollinearity between compactness, concavity, and concave points. Hence, we can drop the attributes of concavity and concave points and use compactness for our prediction.

So, let's drop all the attributes which show multicollinearity. Creating data frames of the respective attributes and dropping them using drop() method will give the following output as in Fig(IX).

```
Index(['diagnosis', 'radius_mean', 'texture_mean', 'smoothness_mean',  
      'compactness_mean', 'concavity_mean', 'concave points_mean',  
      'symmetry_mean', 'fractal_dimension_mean', 'radius_se', 'texture_se',  
      'smoothness_se', 'compactness_se', 'concavity_se', 'concave points_se',  
      'symmetry_se', 'fractal_dimension_se'],  
      dtype='object')
```

Fig (IX): Total columns after dropping reductant attributes.

Now we have obtained a dataset with attributes which have less correlation. Let's verify the correlation using heatmap. The correlation between attributes can be verified by creating a heatmap. The heatmap will result in a correlation which is less than 1 or 0.9.



Fig(X): Heatmap displaying correlation with fewer reductant attributes

From the heatmap given in figure(X), we can understand that the correlation between the selected attributes is very less. Thus we can move on to building models for the Prediction of Breast Cancer.

Building Models

For building models, we have to separate the training and testing data. Here, first let's drop the diagnosis attribute and store the dataset in 'X' and let y be the data frame consisting of the diagnosis attribute data. Here, X has the attributes except 'diagnosis' whereas y has the 'diagnosis' attribute.

We are using the scikit library to apply the algorithms. It is an efficient and powerful library for machine learning tasks. For dividing our dataset into training and testing dataset, we are importing 'train_test_split' from scikit-learn library. We are splitting the dataset into four subsets X_train, X_test, y_train, and y_test. The next step is Feature Scaling. It is done by using the StandardScaler algorithm. It standardizes the values to a standard form. In the code for standardizing, we use the transform () method. *Now the datasets are ready for the implementation of the algorithm.*

Implementation of Logistic regression on the dataset gives a prediction accuracy of 95.96%.

This algorithm identified 110 True positive values, 5 False positive values, 2 False negative values and 54 True negative values. The accuracy is calculated using the formula:

$$Accuracy = (TP + TN) \div (TP + TN + FN + FP)$$

where TP is the accurately identified cases.

TN is the incorrectly identified case.

FN is the incorrectly rejected cases.

FP is the accurately identified cases.

Similarly, for the decision tree model, there were 105 true positive cases, 10 false positive cases, 8 false negative cases and 48 true negative cases. It will result in an accuracy of 89.47% which is less than Logistic Regression Algorithm.

For the Random forest algorithm, the number of true positive, false positive, false negative and true negative cases are 105, 10, 7, 49 respectively. The accuracy is obtained as 90.05% which is greater than the accuracy of the decision tree algorithm. For K Nearest Neighbour, Support Vector machine and Naive Bayes Algorithms we get 90.95%, 96.49% and 92.98%

respectively. SVM finds a total of 112 true positive cases, 3 false positive cases, 3 false negative and 53 true negative cases. Hence by applying all the 6 algorithms Support Vector machine Algorithm gives the highest prediction accuracy of 96.49 % with a Standard deviation of .0376.

Result Analysis

In this research paper, I have used six machine learning algorithms namely Logistic Regression, Decision Tree Model, Random Forest Algorithm, K Nearest Neighbour, Support Vector Machine and Naive Bayes Algorithm. The following table shows the accuracy of prediction. All works regarding this research paper were done on a laptop with AMD Ryzen 5 5500U processor with Radeon Graphics 2.10 GHz, 16 GB RAM and NVIDIA GeForce GTX 1650 GPU. All the coding was done in a Jupyter notebook.

ALGORITHMS	ACCURACY (IN%)
Logistic Regression	95.96
Decision Tree Model	89.47
Random Forest Algorithm	90.05
K Nearest Neighbour	90.05
Support Vector Machine	96.49
Naive Bayes Algorithm	92.008

All algorithms showed greater accuracy above 85%. The Support Vector Machine algorithm shows larger accuracy compared to other algorithms. Logistic Regression is the second in predicting breast cancer with higher accuracy. The least prediction accuracy was found in the Decision tree algorithm. The high prediction accuracy of the SVM algorithm is found in many research papers of Breast Cancer Prediction.

Conclusion

In this study, we have developed a comparison between 6 ML algorithms. The prediction was performed using Wisconsin's Breast Cancer Diagnostic Database. The aim of this paper was to classify patients into two cancer types Benign and Malignant. Initially we made a correlation between all the attributes in the dataset. Those attributes which had correlation higher than .9 were neglected. After that we implemented all the 6 algorithms and found the accuracy rate. The result showed that Support Vector Machine Algorithm has higher prediction accuracy than all other 5 selected algorithms.

References.

1. WHO. "Breast cancer." *World Health Organization (WHO)*, 26 March 2021, <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>. Accessed 23 October 2022.
2. CDC. "What Is Breast Cancer? - What Is Breast Cancer?" *CDC*, https://www.cdc.gov/cancer/breast/basic_info/what-is-breast-cancer.htm. Accessed 23 October 2022.
3. Mayo clinic. "Cancer - Symptoms and causes." *Mayo Clinic*, <https://www.mayoclinic.org/diseases-conditions/cancer/symptoms-causes/syc-20370588>. Accessed 23 October 2022.
4. Epidemiology of breast cancer in Indian women. "Home." *YouTube*, <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUK Ewj47P-Q2fb6AhXH9zgGHfeBB3AQFnoECA0QAQ&url=https%3A%2F%2Fdoi.org%2F10.1111%2Fajco.12661&usg=AOvVaw0VQmEuwwfGdfUWjZPu7uhW>. Accessed 23 October 2022.
5. N. Fatima et al.: Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis
6. M. K. Keles, "Breast cancer prediction and detection using data mining classification algorithms: A comparative study," *Tehnički Vjesnik*, vol. 26, no. 1, pp. 149–155, 2019.
7. S. K. Maliha, R. R. Ema, S. K. Ghosh, H. Ahmed, M. R. J. Mollick, and T. Islam, "Cancer disease prediction using naive Bayes, K-nearest neighbor and J48 algorithm," in *Proc. 10th Int. Conf. Comput., Commun. Netw. Technol. (ICCCNT)*, Jul. 2019, pp. 1–7.

8. A. Bharat, N. Pooja, and R. A. Reddy, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," in Proc. 3rd Int. Conf. Circuits, Control, Commun. Comput. (IC), Oct. 2018, pp. 1–4.
9. Learning Algorithms on the Wisconsin Diagnostic Dataset by Abien Fred M. Agarap
Department of Computer Science Adamson University Manila, Philippines.
<https://doi.org/10.1145/3184066.3184080>
10. N. Fatima et al.: Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis
11. UCI Machine Learning Repository:
<https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29>
12. <https://www.investopedia.com/terms/c/correlationcoefficient.asp>