Group 19

Breast Cancer Prediction
by

Mst. Sumiya Siddika (Exam Roll: 192292)
Md. Shakil Ahmed (Exam Roll: 192330)
Amit Azim Amit (Exam Roll: 192347)

A Project report submitted to the Institute of Information Technology in partial fulfillment of the requirements for the degree of Bachelor of Science (B.Sc.) in Information and Communication Technology

Supervisor: Md. Mahmudur Rahman

Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342

DECLARATION

This project report is submitted to the Institute of Information Technology, Jahangirnagar University, Savar, Dhaka in partial fulfillment of the requirements for having the B.Sc. (Hons.) degree in ICT. This is also needed to certify that the project work is under the 4th Year 1st Semester course of the IIT "ICT-4102: Artificial Intelligences Lab". So, we are here to declare that this project report has not been submitted elsewhere for the requirement of any kind of degree, diploma, or publication.

Signature of the Candidate
Mst. Sumiya Siddika

Signature of the Candidate
Md. Shakil Ahmed

Signature of the Candidate
Amit Azim Amit

ACCEPTANCE

This project report is submitted to the Institute of Information Technology, Jahangirnagar University, Savar, Dhaka in partial fulfillment of the requirements for having the B.Sc. (Hons.) degree in Information and Communication Technology.

Md. Mahmudur Rahman

Lecturer

Institute of Information Technology,

Jahangirnagar University, Savar, Dhaka - 1342, Bangladesh.

ACKNOWLEDGEMENT

First and foremost, we are grateful to God for providing us with the means to effectively complete this work. We would like to take this opportunity to thank everyone who has contributed to this project, helped us along the way, and shared their experiences and insightful opinions in us in order to get the necessary information for our project. We are appreciative of our parents' unwavering support. Above all, we are appreciative of our honorable supervisor, who made the time to mentor us and give us all the resources and information we needed, which was a crucial necessity.

Finally, we convey our regards to our honorable teacher **Md. Mahmudur Rahman** sir for giving us the opportunity to learn the subject particularly practically.

ABSTRACT

The project aims to develop a breast cancer risk prediction model using advanced data analytics techniques. Accurately estimating a woman's risk of breast cancer using their medical history, genetic markers, lifestyle choices, and demographic data is the main goal. The identification of patients who are more at risk could be greatly aided by this predictive model, allowing for early intervention and individualized care plans. The proposed system is developed by using Machine Learning(ML) Algorithm like Support Vector Machine(SVM)[15], K-Nearest Neighbors (KNN)[14] and Logistic Regression algorithm[13].

Contents

1 Introduction

1.1 Motivation

Breast cancer is one of the most common cancers affecting women. The primary causes of breast cancer are still unknown, despite the identification of several risk factors. Breast cancer is the second most common type of tumor and the leading cause of death for women. The World Health Organisation (WHO) reports that 2.3 million women worldwide had a breast cancer diagnosis in 2020, resulting in 6,85,000 deaths worldwide, or 19.9 deaths per 100,000 women annually.

A machine learning method called support vector machine (SVM), K-Nearest Neighbors(KNN), and Logistic Regression can be applied to classification and regression tasks. SVM can be used to pinpoint patients who have a higher chance of getting the disease or who are most likely to experience a recurrence of the illness in the setting of breast cancer prediction. One kind of cancer that starts in the breast is called cancer of the breast. The second most significant cause of cancer death for women is the most prevalent cancer among females. The creation of an accurate and dependable approach for utilizing SVMs to predict breast cancer falls under the problem domain of breast cancer prediction using SVM. This is a difficult task because breast cancer is a complex disease with numerous risk factors.

1.2 Objective

The purpose of the experiment is to develop a prediction model for the likelihood of breast cancer, which may be used to determine which women are most susceptible to the illness. The program hopes to accomplish this aim and give medical professionals a tool that improves their capacity to recognize those who are more likely to develop breast cancer, prompting the creation of more focused and efficient preventive methods for early detection and tactics.

The main objectives are:

1. Gather information on a large number of women, such as their reproductive histories, lifestyle variables, and personal and family histories of breast cancer.
2. Determine which variables have the strongest correlations with the risk of breast cancer by using machine learning methods.
3. Use a different dataset of women with breast cancer diagnoses to validate the model.

1.3 Literature Review

With the advancement of technology, a plethora of contemporary methods for breast cancer prognosis have emerged. The following is a brief summary of what has been done in this area of study.

Azar and colleagues [3] introduced a new method for the identification. of the breast kind. The methodology employed three categorization techniques known as probabilistic radial basis function (RBF), multi-layer perceptrons (MLP), and neural networks (PNN). The method that learned the dataset's characteristics for breast cancer and testing procedures was also used. A method called GA-MOO NN has been suggested by Ahmad et al. [6] for the identification of breast cancer. The dataset is split into three sections by the method: training (50%), testing (25%), and validation (25%).

By adopting a hybrid neurogenetic framework that combines train feedback and genetic algorithms to categorize the features of a breast cancer dataset, the researchers in [8] demonstrated an inventive method for predicting the likelihood of breast cancer. Overfitting results from the framework being trained in a method that leaves one out. The structure achieves a total precision of 97%.

1.4 Our Contribution

In this project, we used the SVM algorithm for predicting breast cancer and then we used grid search to increase the performance of the model for improving accuracy.

2 Methodology

2.1 Overview

A detailed description of the proposed machine learning methods used to create predictive algorithms to foretell the detection of breast cancer. Create a prediction model for the diagnosis of breast tumors. In this endeavor, the database included 569 independent data, 35% of which were from breast cancer patients with 31 attributes in which records were obtained from Kaggle. The Support Vector Machine(SVM)[15], K-Nearest Neighbors(KNN)[14], and Logistic Regression[13] Machine Learning Algorithm were used in this study. The five portions of the analysis are stored as Jupiter notebooks within this repository. First of all, we identify the problem and data sources that we used for our project. Secondly Exploratory data analysis. Thirdly Pre-processing the data, predictive model using SVM and finally comparison between different classifier algorithms.

2.2 Required Equipment

- Seaborn[6]
- Matplotlib[7]
- Pandas[8]
- Numpy[9]
- Scikit Learn[10]

2.3 Algorithms

2.3.1 Support Vector Machines

Support Vector Machine (SVM) is a technique for recognizing outliers and regression analysis and linear or nonlinear classification. SVMs are useful for many different kinds of applications, including face recognition, anomaly detection, handwriting recognition, text classification, picture classification, spam detection, and gene expression analysis[15].

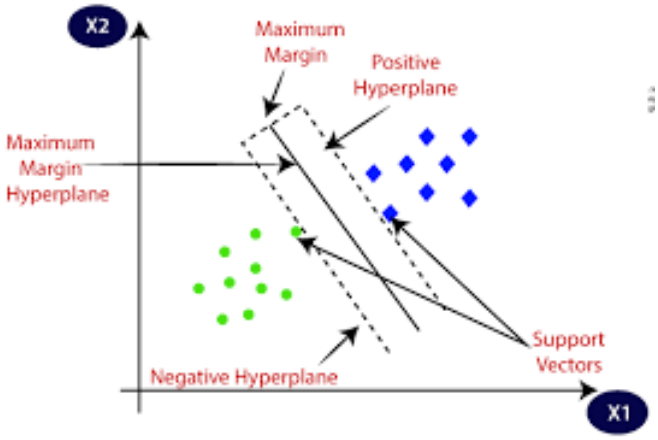

Fig 2.3.1 - Support Vector Machine

Working procedure:

1. Load the dataset from sklearn.datasets
2. Keep goal variables and input characteristics apart.
3. Use the RBF kernel to construct and train the SVM classifiers.
4. Draw the input feature scatter plot.
5. Plot the limit of the choice.

2.3.2 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a popular and simple machine-learning method for classification and regression issues. It's a non-parametric instance-based learning technique that bases its predictions only on the data, without making any assumptions about the data's distribution. KNN may be applied to both supervised and unsupervised learning tasks while being most commonly associated with supervised categorization[14].

Algorithm steps:

1. Decide which neighbor's K number to choose.
2. Compute the K number of neighbors' Euclidean distance.
3. Choose the K nearest neighbors using the calculated distance from Euclid.
4. Determine how many data points there are in each category among these k neighbors.
5. Put the additional data points in the category where the neighbor count is at its highest.
6. The model is prepared.

2.3.3 Logistic Regression

Logistic regression is an approach used in statistics and machine learning for binary and multi-class classification. It is a popular and easily understood method that simulates the likelihood of a binary result or the likelihood of falling into a certain class in a situation with many classes. Logistic regression is not used for regression; rather, it is used for classification.[13]

Steps:

1. Pre-processing of the data.
2. Matching the Training Set with Logistic Regression.
3. Forecasting the exam outcome.
4. Test result accuracy (Confusion Matrix Creation).
5. Displaying the test set output visually.

2.4 System Model

2.4.1 Identifying the Problem and Data Sources

As the second most prevalent cause of cancer-related mortality among women, breast cancer accounts for almost one in three cancer diagnoses among women in the United States. It is also the most common cancer among women overall. A tumor, which is the usual term for Breast cancer is caused by aberrant cell growth in the breast tissues. benign tumors, or non-cancerous ones, pre-malignant tumors, or cancerous tumors, can all occur from tumors, hence a tumor does not necessarily indicate cancer[11].

```
data.head()
```

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	concave points_mean	...	rac
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	...	
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	...	
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	...	
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	...	
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	...	

5 rows × 32 columns

Fig 2.4.1 - Datasets

Figure 2.4.1 displays the number of columns and rows in the datasets as well as five rows of elements. Character data can be found in the diagnosis columns of the dataset. Use label encoding to transform the character data into integer data.

```
data.columns
```

```
Index(['diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst'],
      dtype='object')
```

Fig 2.4.2 - Features of dataset

Expanding on these insights, Figure 2.4.2 represents the columns or feature names in our dataset.

```
data.describe()
```

	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	poin
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	56
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	

8 rows × 30 columns

Fig 2.4.3- Describe dataset

Figure 2.4.3 describes the dataset including the count, mean, max, and standard deviation values.

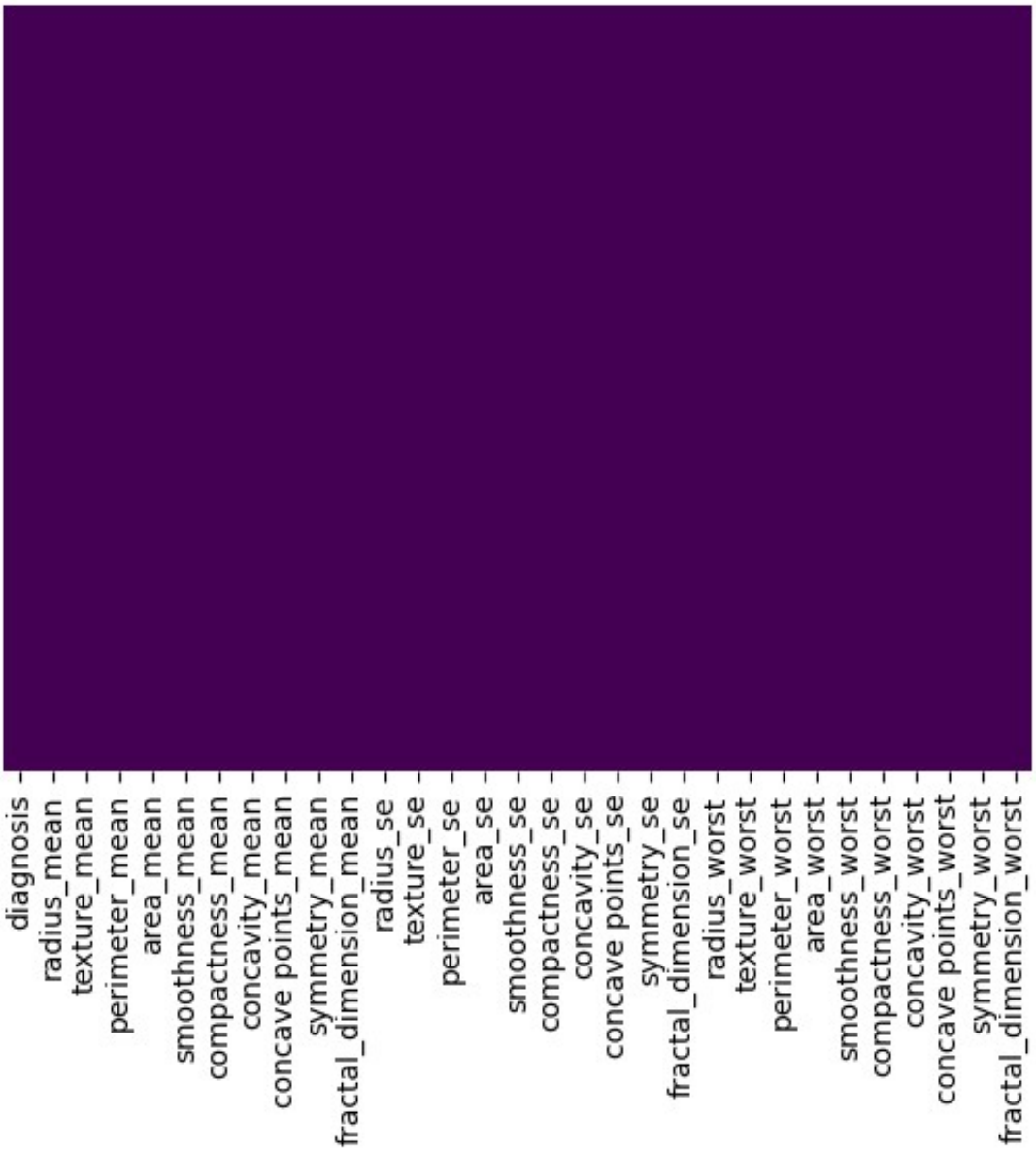

Fig 2.4.4 -Heatmap

Figure 2.4.4 represents There are 32 attributes in total and 569 observations in our dataset. We do not need to handle missing data because the isnull() command did not reveal any missing values in our dataset.

2.4.2 Exploratory Data Analysis

Exploratory data analysis (EDA)[3] is a crucial stage that needs to be completed before modeling. This is due to the fact that the ability to comprehend the nature of the data without drawing any assumptions is crucial for a data scientist. Understanding the distribution of values, the structure of the data, the presence of radical ideals, and the relationships among the data can all be greatly enhanced by the outcomes of data exploration. There are two model approaches to examining the data Descriptive statistics and Visualization.

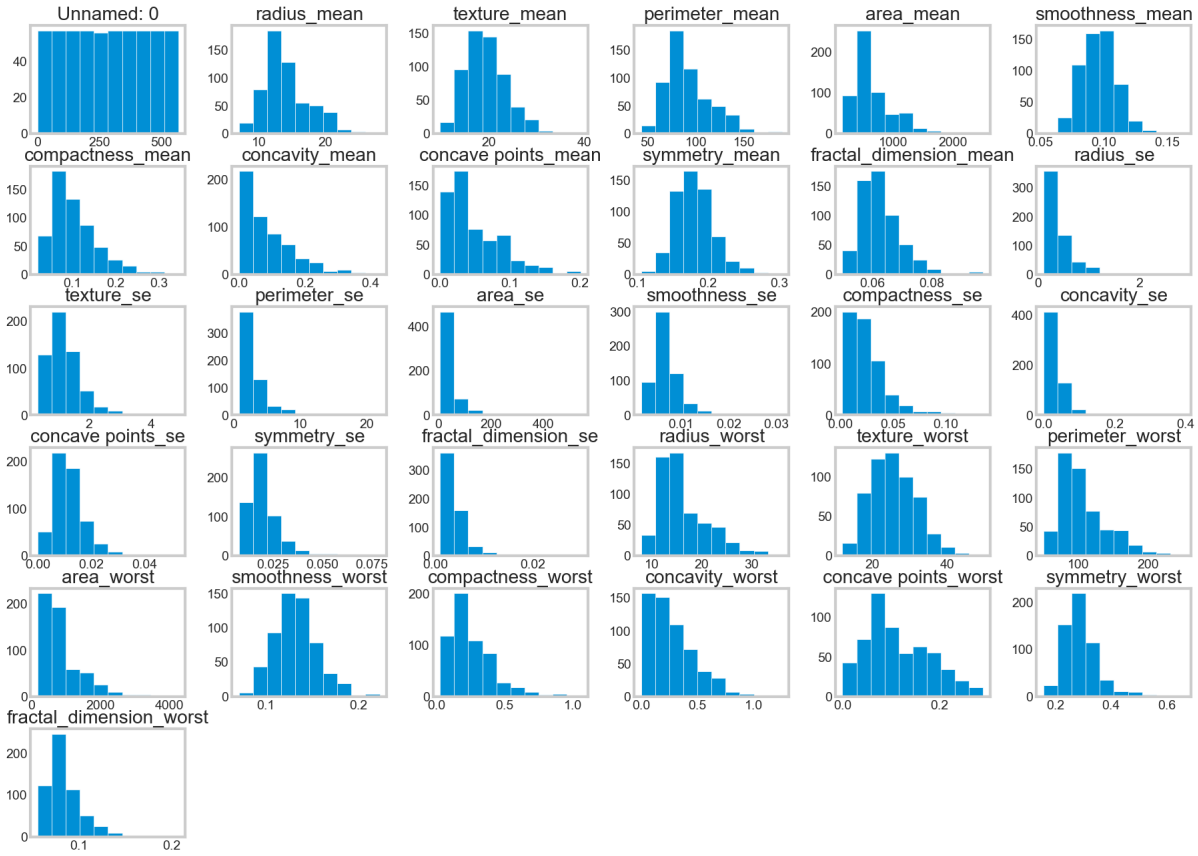

Fig 2.4.5-Histogram plot

Numerical variables are frequently visualized using histograms. Upon the contents of the parameter are divided through an infinite number of periods (bins), a histogram resembles a bar graph. Histograms count how many findings there are for each bin after grouping the data into bins. It can also help you see possible outliers. As shown in the above figure 2.4.5

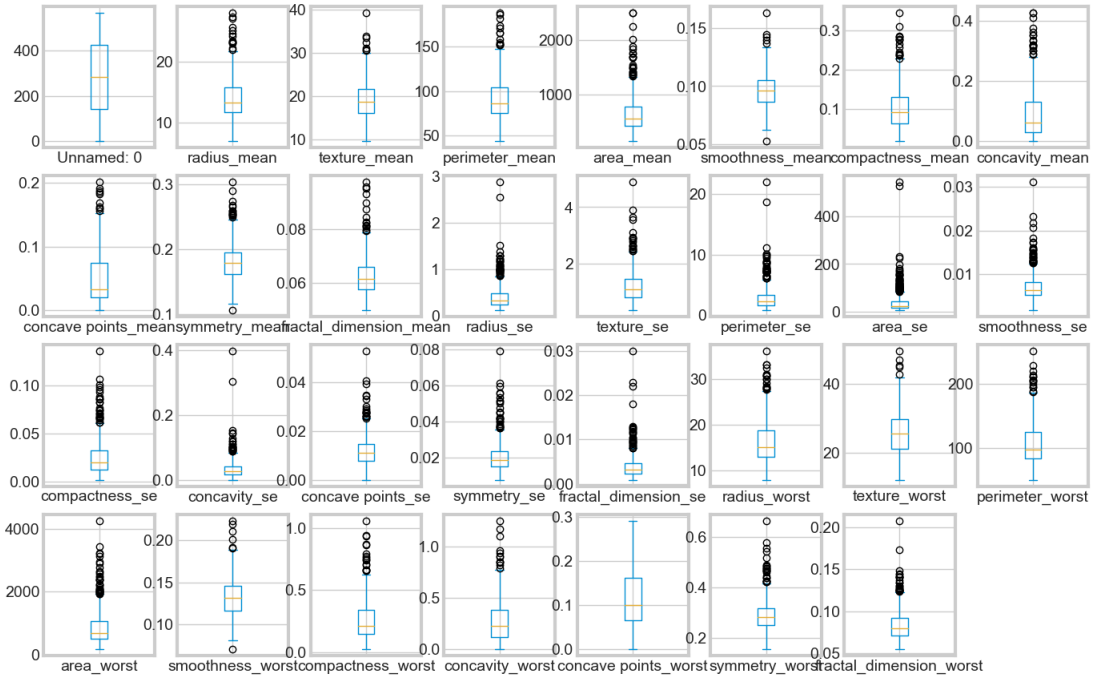

Fig 2.4.6-Box plot

Figure 2.4.6 shows the box plot that is made to show the properties of the minimum, the maximum values, the third quartile, the median, and the first quartile of a set of data. Here, the data that will be displayed is indicated by the x-axis, and the y-axis, on the other hand, shows the frequency distribution[6].

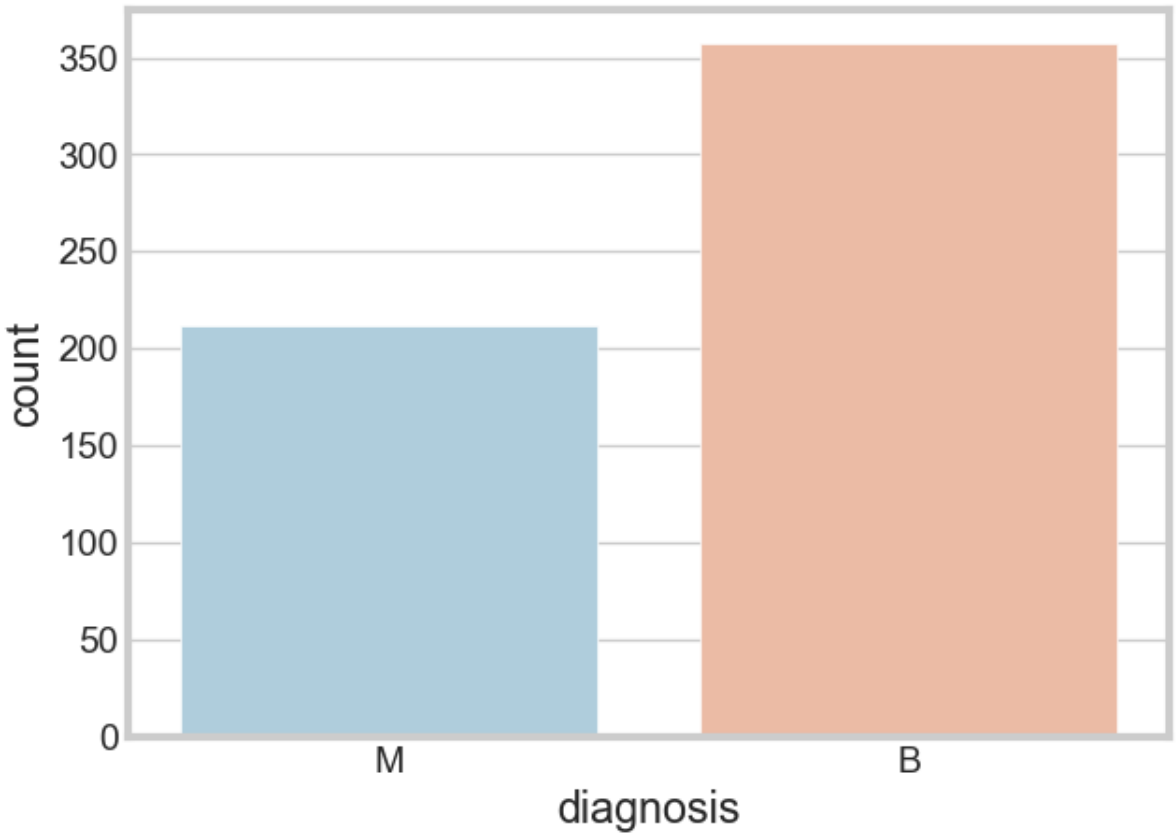

Fig 2.4.7-Frequency cancer diagnosis

Figure 2.4.7 expresses the number of Malignant=1 (shows that there are cancer cells present) and the number of Benign=0 (indicates the absence of cancer cells). 212 findings show the presence of cancer cells, while 357 show that there are no cancer cells present[5].

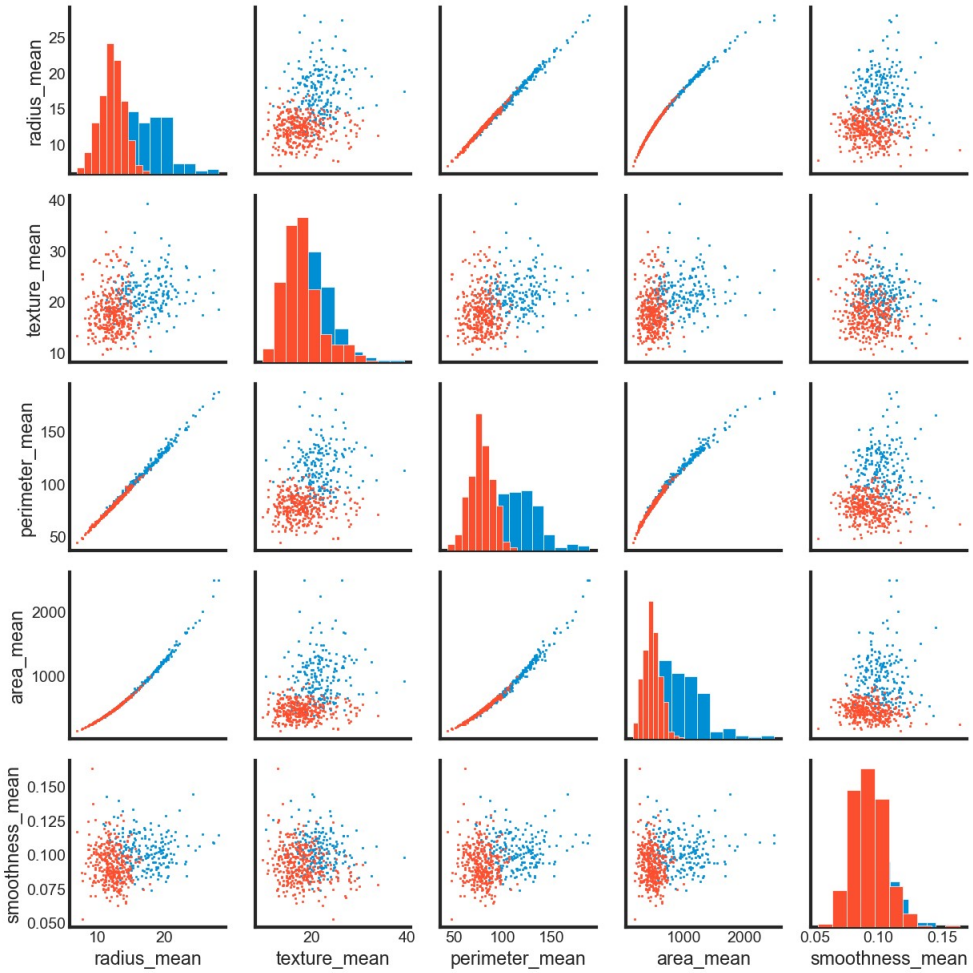

Fig 2.4.8- Histograms

As can be observed from the strong positive relationship in the mean variables between 1-0.75, the mean area of a tissue nucleus shows a strong positive correlation with the mean values of radius and parameter.

2.4.3 Data Preprocessing

Label encoding: In order for machine learning models that can only accept numerical data to fit categorical columns, a technique known as label encoding is used to transform them into numerical ones. In a machine-learning project, it is a crucial pre-processing step.

```
#Assign predictors to a variable of ndarray (matrix) type
array = data.values
X = array[:,1:31]
y = array[:,0]
```

```
#transform the class labels from their original string representation (M and B) into integers
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)

#Call the transform method of LabelEncorder on two dummy variables
#le.transform (['M', 'B'])
```

Fig 2.4.9-Label Encoding

Figure 2.4.10, The class labels are converted from their original string representation (M and B) into integers before being assigned to a NumPy array called X along with the 30 features. Calling the transform method of LabelEncorder on two dummy variables illustrates how the malignant tumors have been designated as class 1 (i.e., presence of cancer cells) and the benign tumors are represented as class 0 (i.e., no cancer cells detected), respectively, after encoding the class labels (diagnosis) in an array y[2].

Principal component analysis(PCA) is a widely used technique for analysing large datasets with many dimensions or features per observation, improving data interpretability while retaining as much information as possible, and facilitating multifarious information visualization [7].

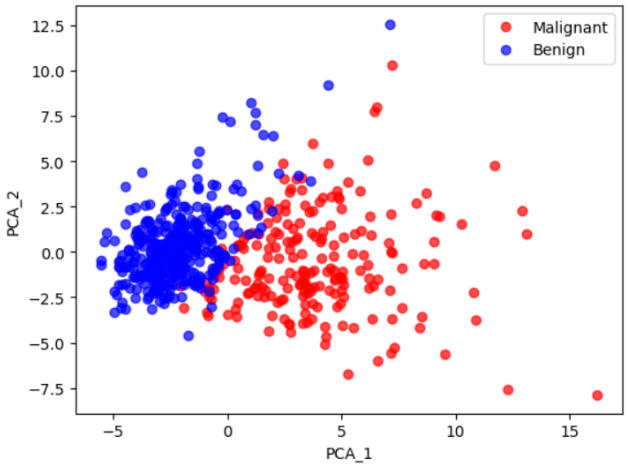

Fig 2.4.10- Feature decomposition using PCA

Using any of the dimensionality-reducing techniques makes sense in order to try to use as many features as possible and maintain as much information as possible when working with only two dimensions, as the combined plot shows that many feature pairs divide the data nicely and to a degree that is comparable.

2.4.4 Predictive Model using SVM

The learning algorithm known as support vector machines (SVMs) [15]will be employed to construct the predictive model. SVMs, one of the most widely used classification algorithms, have a sophisticated method for fitting a linear model to nonlinear data using a linear algorithm.

Kernel that uses SVMs, the crucial variables are

- Regularization parameter C.
- The choice of the kernel,(linear, radial basis function(RBF) or polynomial).
- Kernel-specific parameters.

The model's level of difficulty is determined by both gamma and C, where larger values in either parameter lead to a more complex model. As a result, good values for both parameters are typically highly correlated, meaning that C and gamma should be changed simultaneously[15].

Train Test Split

In our project, our primary goal is to predict whether breast cancer cells are present in women or not. Here the target variable is the diagnosis of the women and other columns are the features of our predictive algorithm. To assess the effectiveness of our machine learning model, we randomly split the dataset into two subsets: a test set as well as a training set. The algorithm is trained using the training set, which contains 60% of the data, and the model's performance on unobserved data is assessed using the test set, which comprises 40% of the data. The training set has 341 rows and 30 columns the test set has 227 rows and diagnosis columns.

Feature Selection

In our dataset, comprising 31 features where 30 features are variable and one is a target variable, a strategic decision was made to enhance the model's efficiency. Subsequently, we embarked on a meticulous analysis to identify the optimal number of features for our model. For each set, we trained the model and evaluated its accuracy. Through this iterative process, we found that the accuracy of the model increased with a greater amount of characteristics.

Classification

Classification is a supervisory technique that categorizes the data into the desired number of classes. The goal of this work is to find out the women have breast cancer or not. The area In a Curve (AUC) is a frequently used measure of performance in statistical modeling and machine learning that indicates how accurate a model can be in binary categorization tasks[2].

3 Result and Discussion

3.1 Dataset Analysis

The Breast Cancer datasets are accessible through the University of California, Irvine's artificial intelligence repository. The collection contains 569 samples of cancerous and benign tumor cells. The dataset's first two columns include, respectively, the samples' unique ID codes, together with the corresponding diagnostic (M = malignant, B = harmless). Thirty real-value qualities are included in Columns 3-32 and may be utilized to build a model to identify the malignant or benign nature of a tumor. These elements were calculated using digital images of the cell nuclei[4].

Correlation heatmap: A machine learning method for visualizing the correlation between variables in a dataset heatmap. This figure illustrates The correlation coefficients using correlation matrices between every pair of variables in the dataset. A dataset's patterns and connections between variables can be found using correlation heatmaps, which are particularly useful for identifying highly correlated variables that may lead to multicollinearity and affect the performance of machine learning models[9].

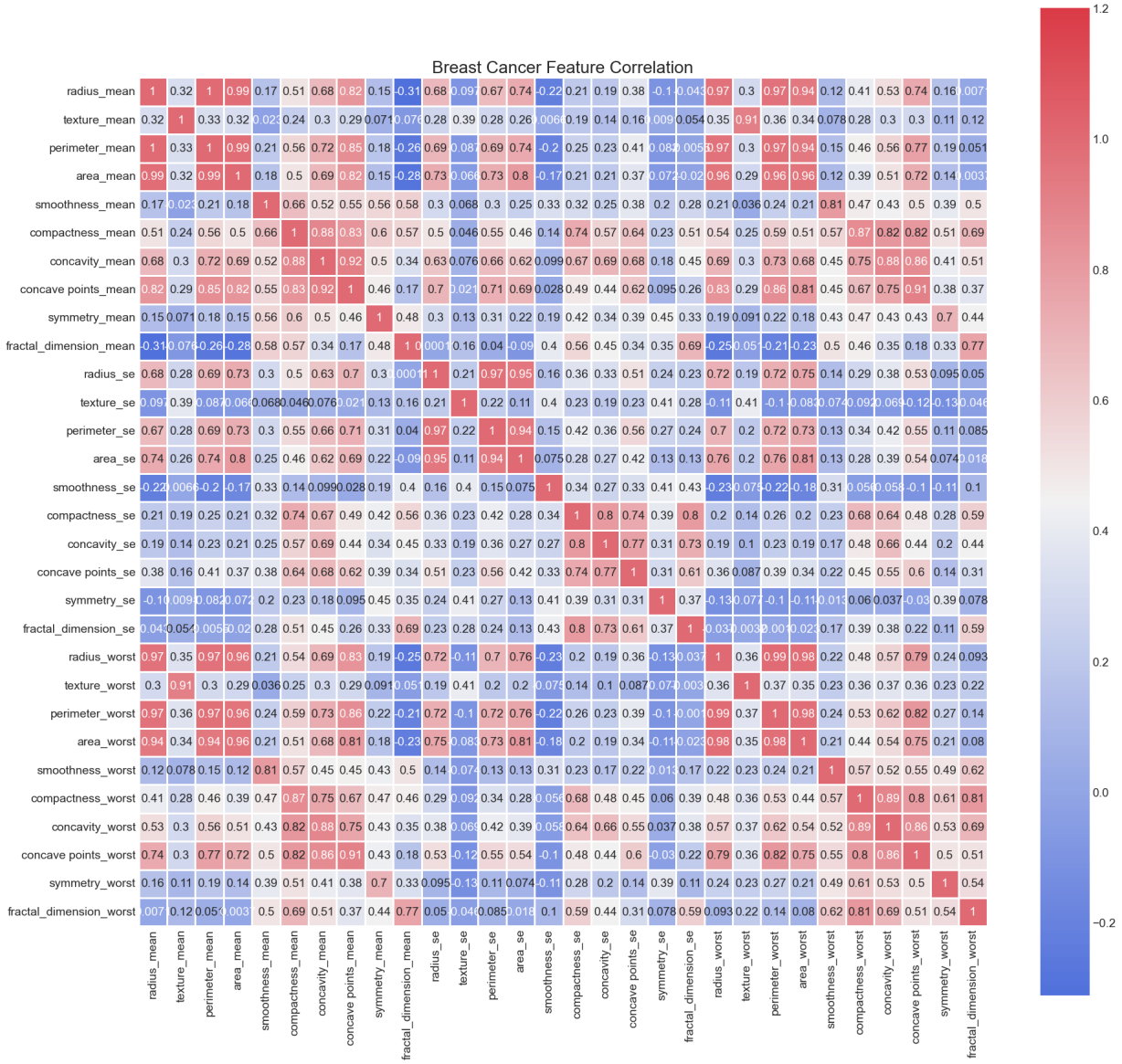

Fig 3.1-Correlation Heatmap

Figure 3.1 shows the correlation heatmap for our dataset. A higher value represents a strong correlation. We can see the correlation between a number of variables related to student performance.

Pairplot: In machine learning, a visualization technique called a data pair plot is used to show the pairwise correlations between variables in a dataset. Every variable in the dataset is compared to every other variable using this type of scatterplot matrix. The representations of each variable in a pair plot are a scatterplot of the combined distribution of all the variables and a diagonal plot of the distribution of each variable[6].

Our dataset, which has 31 predictor variables, is shown in Figure 3.2. Plotting the pair plot for our dataset would produce an excessively large 31 by 31 matrix.

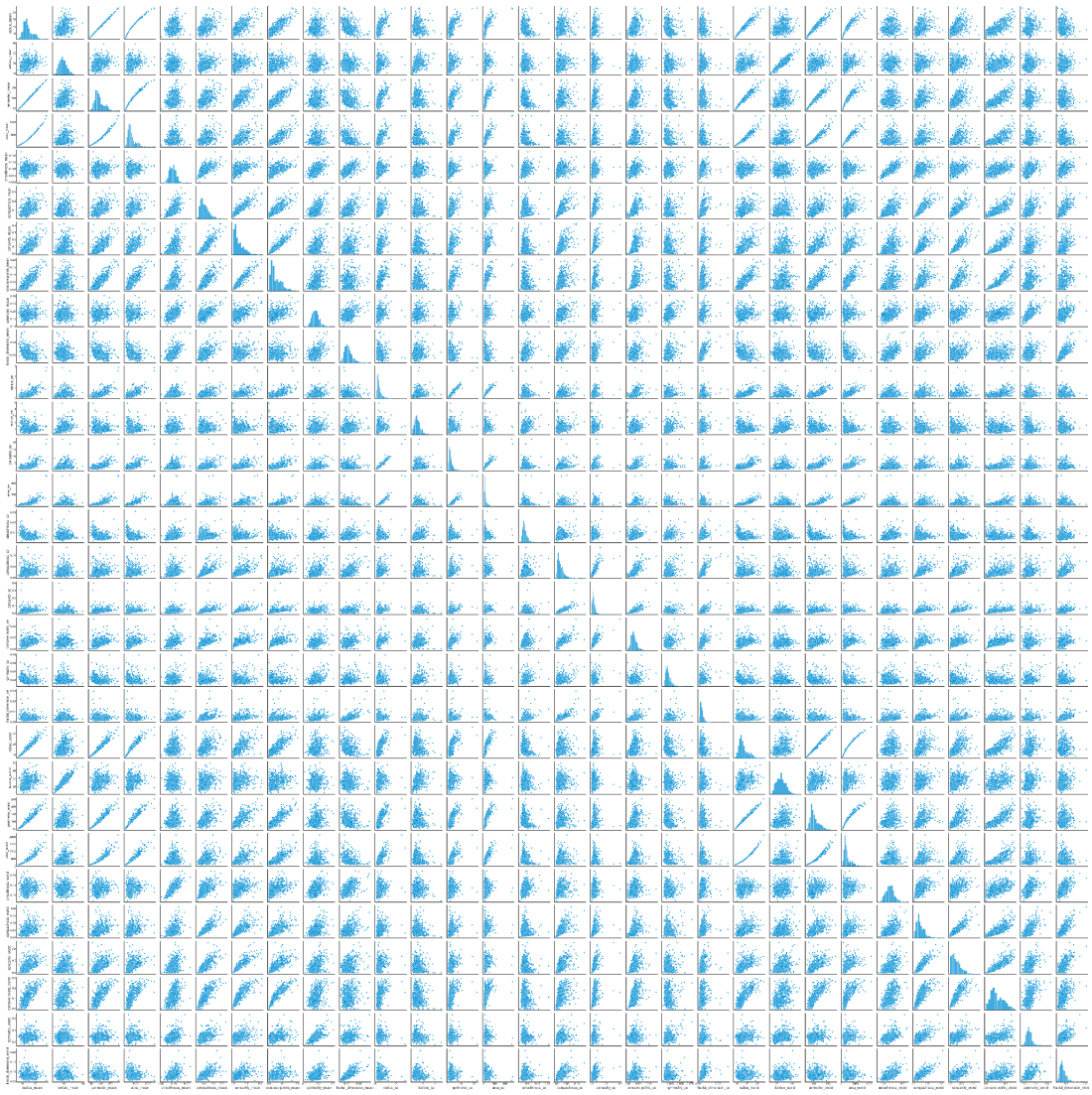

Fig 3.2-Pair plot

3.2 Result

```
[1]
The Women have Malignant (Cancerous) - Present in Breast
```

3.2.1 Accuracy

A very effective and practical metric for evaluating machine learning prediction accuracy in both contexts is accuracy. One of the most commonly used metrics in research is that, in order to focus on algorithmic approach improvements, clean and balanced datasets are usually present. AI accuracy can be defined as the ratio of correctly classified data to all other forecasts combined, or the percentage of correctly predicted data to all other forecasts produced by a trained machine learning model. True positives (TP) and true negatives (TN) make accurate predictions. The whole set of positive (P) and negative (N) instances make up each forecast[10].

The formula is given by:

$$y = \frac{Number\ of\ correct\ Prediction}{Total\ number\ of\ Prediction}$$

```
The classifier accuracy score is 0.97
```

3.2.2 Confusion Matrix

An effective tool for comprehending a classification model's performance is a confusion matrix. It can be used to direct model improvements and helps to identify instances in which the model is erroneous. Furthermore, a confusion matrix can be used to assess how well various classification models perform and select the most appropriate model for a particular issue[16].

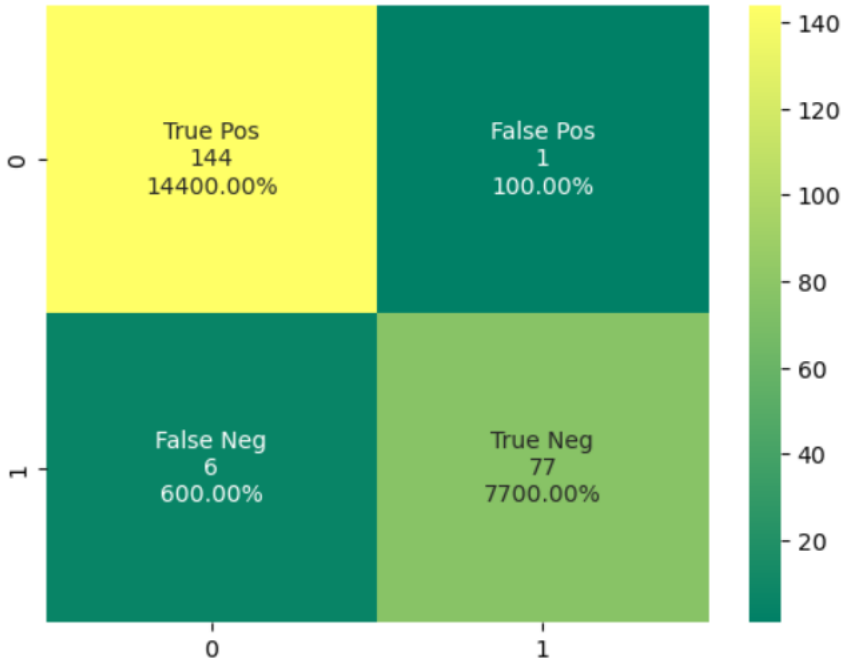

Fig 3.3- Confusion Matrix for the performance of the algorithm

3.2.3 Precision, Recall, and F1 score

Common evaluation metrics used in machine learning to assess a classification model's performance are precision, recall, and F1 score. Typically, a model with an F1 score, recall, and precision that are high indicates that the model is good. By dissecting the entire dataset, SVM had the best F1 score, recall, and precision[16].

	precision	recall	f1-score	support
0	0.96	0.99	0.98	145
1	0.99	0.93	0.96	83
accuracy			0.97	228
macro avg	0.97	0.96	0.97	228
weighted avg	0.97	0.97	0.97	228

Fig 3.4- Using SVM

Fig 3.4- Using the support vector Machine algorithm the accuracy is 97%.

	precision	recall	f1-score	support
0	0.93	0.96	0.94	105
1	0.94	0.88	0.91	66
accuracy			0.93	171
macro avg	0.93	0.92	0.93	171
weighted avg	0.93	0.93	0.93	171

Fig 3.5- Using KNN

When using K Nearest Neighbors of our datasets the accuracy of that model is 93% less than the SVM shown in Figure 3.5.

	precision	recall	f1-score	support
0	0.92	0.96	0.94	105
1	0.93	0.86	0.90	66
accuracy			0.92	171
macro avg	0.93	0.91	0.92	171
weighted avg	0.92	0.92	0.92	171

Fig 3.6- Using Logistic Regression

Logistic regression has 92% accuracy which is less than the SVM and KNN shown in Figure 3.6

3.3 Discussion

We have created a model based on SVM that has a 97% accuracy rate in predicting breast cancer. 40% of the information and 60% of the data were used for training and 60% had been utilized for evaluation. Using Gridsearch will greatly improve the model's performance[14].

4 Future Work and Conclusion

4.1 Future work

Our future objective will be:

- Develop a more accurate risk prediction model.
- Develop a real-time risk prediction model.
- Develop models that can predict cancer treatment.

4.2 Conclusion

Breast cancer is a serious disease, if we early detect the cancer and start the treatment then it can significantly improve the survival rate. Machine learning models help to identify women's breast cancer and play a significant role. Here we use SVM[15], KNN[14], and Logistic Regression[13] models. The model was trained on a dataset of the set contained 569 specimens of tumor cells, both harmless and malignant. The model developed in this project has an accuracy of over 97%, and it also has a high specificity and sensitivity.

5 References

[1] Van't Veer, L. J., H. Dai, et al. "Gene expression profiling predicts clinical outcome of breast cancer." Nature 415(6871): 530-536, 2002.

[2] Ahmad Taher Azar, Shaimaa Ahmed El-Said, "Probabilistic neural network for breast cancer classification," Neural Computing and Applications, Springer, vol. 23, pp.1737-1751, 2013.

[3] Weigelt, B., Z. Hu, et al. "Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer." Cancer research 65(20): 9155, 2005.

[4] Na KY, Kim KS, LeeJE, Kim HJ, Yang JH, Ahn SH,et al. The 70-gene prognostic signature for Korean breast cancer patients. J Breast Cancer 2011;14:33-8

[5] Dataset, Available: `https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic` [Accessed date: Octobor23,2023]

[6] seaborn,"Seaborn", Available:`https://seaborn.pydata.org/index.html`[Accessed October 22,2023,11:00pm]

[7] matplotlib, "Matplotlib", Available: `https://matplotlib.org/stable/`[Accessed: October 22, 2023,11:00pm]

[8] pandas," Pandas", Available: `https://pandas.pydata.org/`[accessed: October 23,2023,12:00pm]

[9] numpy, "Numpy",Available: `https://numpy.org/doc/stable/` [Accessed: October 21,2023,8:00am]

[10] scikit-learn, "sikit-learn", Available: `https://scikit-learn.org/stable` [Accessed/: September 23,2023,8:0pm]

[11] github. "Github", Available: `https://github.com/topics/breast-cancer-prediction` [Accessed: September 10,2023,23:00pm]

[12] github, "Github" Available: `anindya-saha.github.io` [Accessed October 22,2023, 11:00pm]

[13] LogisticRegression,Available:`https://www.geeksforgeeks.org/understanding-logistic-regression/` [Accessed: September 25,2023,6:00 pm]

[14] KNN, Available:`https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4` [Acessed: October 20,2023,4:00pm].

[15] SVM,Available:`https://medium.com/swlh/the-support-vector-machine-basic-concept-a5106bd3cc5f` [Accessed: September 23,2023,11:00pm].

[16] Available: `https://medium.com/swlh/confusion-matrix-and-classification-report-88105288d48f` [Accessed: October,23,2023,6:00pm]