Breast Cancer Detection & Prediction through Machine Learning

Ujwal Watgule
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
ujwal.watgule22@pccoepune.org

Dr. Pravin Game
Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
pravin.game@pccoepune.org

Abstract

Data mining play vital role in the prediction and treatment of breast cancer. To enable practitioner decision-making, huge volumes of data is processed with machine learning techniques to produce tools for prediction and accuracy. In this paper, I proposed a prediction model, which is specifically designed for detection and prediction of breast cancer using machine learning algorithms like decision tree classifier, Logistic Regression, Random Forest and SVM algorithms. Machine learning algorithms improves the accuracy and enhances the performance. This paper is an attempt to identify the type of tumour (i.e., benign or malignant), visualise, correlate and categorize different characteristic of data, split the dataset into dependent and independent dataset and eventually split the data into training and testing set thereby increasing the prediction accuracy. With the help of feature scaling and with the use of different models like logistic regression, decision tree, random forest classifier, metrics model an attempt is made to improve the accuracy of training data.

Keywords: Breast cancer, logistic regression, decision tree, random forest classifier.

2. Introduction:

Breast cancer has become the most recurrent type of health issue among women especially in middle age. Early detection of breast cancer can help women cure this disease and death rate can be reduced [2]. In the present-day scenario, breast cancer mammograms are used and they are known to be the most effective scanning technique. In this paper the detection of cancer cells is done by machine learning technique.

Breast cancer is one of the most common cancers that is responsible for high number of deaths in women every year. Despite the fact that cancer is treatable and healable in earliest stages, the huge number of patients are diagnosed with cancer very late. Data mining process and classification are an important way to predict and detect the cancer efficiently and effectively. [1]

Worldwide there are 1.67 million new breast cancer cases in women in the year 2012; this morbidity accounts for about 25% in all cancers [2]. Breast cancer can be the most treated malignancies if detected early. Mammogram and fine needle aspiration cytology are the most frequently used diagnostic techniques, however; they lack high diagnostic capability [3]. Therefore, there is a pressing need to develop better techniques for cancer diagnosis that are inexpensive, convenient and capable.

Breast cancer prediction based on dataset, compare and identify an accurate model to predict the breast cancer based on various patients' clinical records. Four data mining models are applied, i.e., support vector machine (SVM), artificial neural network (ANN), Naive Bayes classifier, AdaBoost tree. Furthermore, feature space is highly discussed its high influence on the efficiency and effectiveness of the learning process.[4]

2.1 Breast Cancer and its Symptoms

Different people have different symptoms of breast cancer. Some people do not have any signs or symptoms at all.
Some warning signs of breast cancer are—

- New lump in the breast or underarm (armpit).
- Thickening or swelling of part of the breast.
- Irritation or dimpling of breast skin.
- Redness or flaky skin in the nipple area or the breast.
- Pulling in of the nipple or pain in the nipple area.
- Nipple discharge other than breast milk, including blood.
- Any change in the size or the shape of the breast.
- Pain in any area of the breast.

3. Literature Survey

Research paper of Masud Rana Basunia; Ismot Ara Pervin; Md. Al Mahmud; Suman Saha; Mohammad Arifuzzaman proposed an ensemble method named stacking classifier which combines multiple classification techniques and efficaciously classifies the benign and malignant tumour. They applied different classification techniques over the dataset and tuned their parameters to improve accuracy [9]. This proposed Stacking classifier combined the results of those best classifiers using meta classifier and provided 97.20% accuracy for breast cancer prediction.

“A Survey on Breast Cancer Prediction Using Data Mining Techniques”, in this paper of Dona Sara Jacob; Rakhi Viswan; V Manju; L PadmaSuresh; Shine Raj, they performed a comparison of diverse classification and clustering algorithms. Varied classification algorithms and the clustering algorithm are used and outcome indicates that the classification algorithms are superior predictors than the clustering algorithms.[10]

A Data Mining Model To Predict Breast Cancer Using Improved Feature Selection Method On Real Time Data”, here researchers used Naive Bayes (NB) classifier algorithm, along with a new improved feature selection method to get Empirical verification reaffirms that hybrid feature-selection approach, using minimal set of attributes, outperforms the results obtained from solitary feature selection methods and accuracy around 98%.[11]

In the paper of K. Shilpa; T. Adilakshmi; K. Chitra, the main objective of a research paper is to find the best classification algorithm. The research paper proposed an approach that improves accuracy and enhances the performance of algorithms. The efficient Machine learning algorithms are used such as Naïve Bayes, J48, Sequential Minimal Optimization (SMO) and Instance- Based for K-Nearest neighbor (IBK) and The investigational results illustration in the proposed IBK algorithm gives the maximum accuracy of 100%. [13]

“Diagnosis Using Data Mining Algorithms for Malignant Breast Cancer Cell Detection”, in this paper used decision tree (DT) and deep learning method and found different time complexity, 96% accuracy, sensitivity.[12]

So, aim of all these researchers to predict breast cancer with higher accuracy. For that they used classification techniques over the dataset, used different algorithms like clustering algorithm, Naïve Bayes algorithm, J48, Sequential Minimal Optimization (SMO) and Instance- Based for K-Nearest neighbour (IBK) and used other technique and algorithm. They tried to get higher accuracy and they got. So concluded that performance of different data mining approaches has been evaluated and found best results with accuracy.

4. Problem Statement

In order to accomplish accuracy in classification results, purity of breast cancer dataset must be high. The noisy data is the biggest problem of data mining. Selection of appropriate purified dataset can have great impact on quality of extracted knowledge. The dataset suffers from missing data, outliers and imbalanced data. These type of impurities directly affects the accuracy of the prediction models.

Missing data: Missing data mentions that some of the data will not be available or applicable. This type of missing values severely degrades the classification algorithms. Breast cancer dataset suffers with three forms of missing data. Secondly, the unrecorded events happened in the lab. Finally, the intermediate results, which was neglected at the time of data aggregation.

Outliers: Outliers are the invariant data that may follow the general category. This type of data severely affects the prediction results. These outlier data can be detected using various approaches like including, outlier filtering, outlier correction and robust algorithms.

4.1 Methodology

The aim of supervised learning is to construct a classification model based on a given data set that contains some attributes and labelled classes. The training data set and testing data set are two necessary components that are implemented in supervised learning. The training data set is used to build the prediction model, using different algorithms. Test data is often randomly extracted from the entire database and used to validate the model.

The aim of this study is breast cancer detection system which can auto detect and predict the type of breast cancer which is malignant or benignant. In data pre-processing stage we remove the noise. Data analyzation is used to predict the cancer as benign or malignant. We can insert new data of patient and get new analyzation of data records. This system is cost effective.

A Dataset is a collection of data. Load Dataset into the program. Data Pre-processing is an important step in machine learning process. Data pre-processing is a technique that is used to convert the raw data into a clean dataset.

Data is cleaned through process such as handling the missing value, noisy data or resolving the inconsistencies in the data. Data transformation is the process of converting data from one format or structure into another format or structure. It primarily involves mapping to see how source data element will be changed for the destination.

Machine learning use so called features (i.e., variables or attributes) to generate predictive models. Use of suitable combination of features is essential for obtaining high precision and accuracy.

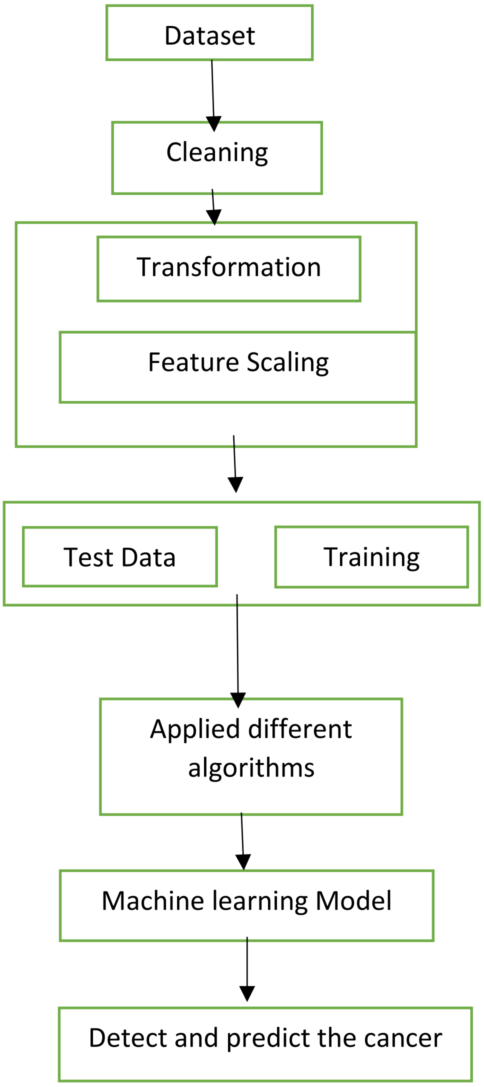


Figure 1: Research Method Overview

Here I have used random forest classifier, logistic regression, decision tree classifier, SVC for measuring accuracy.

4.1.1 Decision Tree Algorithm

Decision Tree algorithm belongs to the supervised learning algorithms. decision tree algorithm can be used for regression and classification problems. The general motive of using Decision Tree is to create a training model which can use to predict class or value of target variables by learning decision rules inferred from prior data (training data).

4.1.2 Logistic Regression

Logistic regression is a fundamental classification technique. It belongs to the group of linear classifiers and is somewhat similar to polynomial and linear regression. Logistic regression is fast and relatively uncomplicated, and it’s convenient for you to interpret the results. Although it’s essentially a method for binary classification, it can also be applied to multiclass problems.

4.1.3 Support vector machines (SVMs)

It is a set of supervised learning methods used for classification, regression and outliers’ detection.

The advantages of support vector machines are:

- Effective in high dimensional spaces.
- Still effective in cases where number of dimensions is greater than the number of samples.
- Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.
- Versatile: different Kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels.

4.1.4 Random Forest Classifier

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems.

4.1.5 Confusion Matrix

It includes a table that usually applied to represent fulfilment of a classification model, on a collection of test data for which the actual rates are known. Performance of such a method is regularly estimated utilizing data in the model.

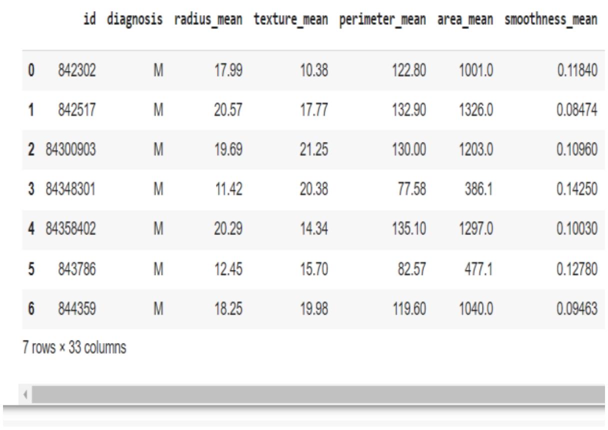
		Predicted	
		Positive	Negative
Real	Positive	True Positive (TP)	True Negative (TN)
	Negative	False Positive (FP)	False Negative (FN)

Table 1: Confusion Matrix

5. Proposed Work : Result and Discussion

Here, the purpose of this paper was to predict and detect breast cancer at early stage with high accuracy using data mining classification algorithms.

In this paper I used python language. Load breast cancer dataset from Kaggle website and read data and count the rows and columns.

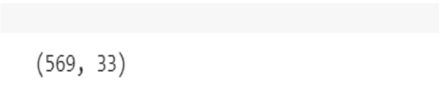


	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840
1	842517	M	20.57	17.77	132.90	1326.0	0.08474
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960
3	84348301	M	11.42	20.38	77.58	386.1	0.14250
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030
5	843786	M	12.45	15.70	82.57	477.1	0.12780
6	844359	M	18.25	19.98	119.60	1040.0	0.09463

7 rows × 33 columns

Figure 2: Read column and Rows

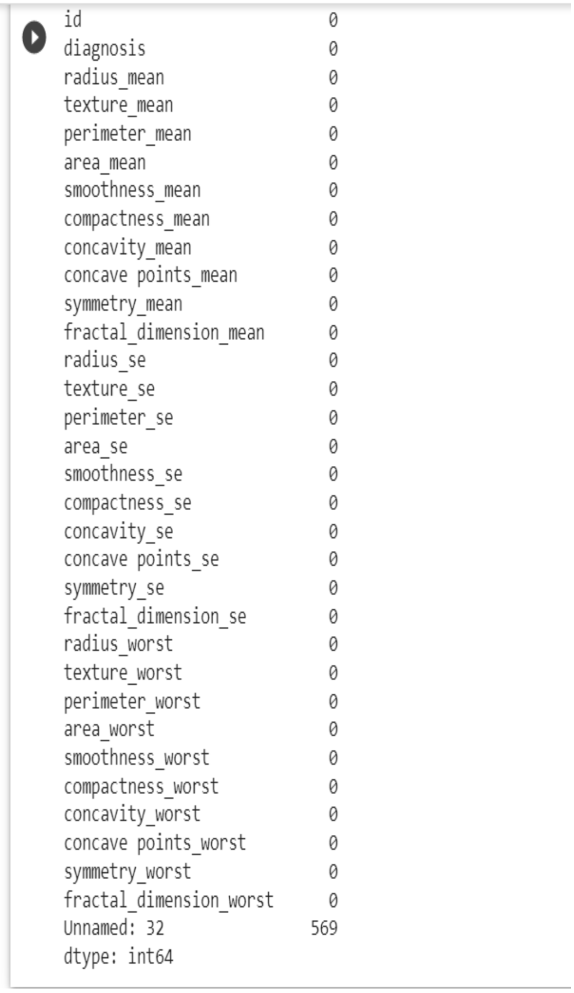
Got 569 rows and 33 columns.



```
(569, 33)
```

Next is to count the number of empty (Nan, NAN, na) values in each column.

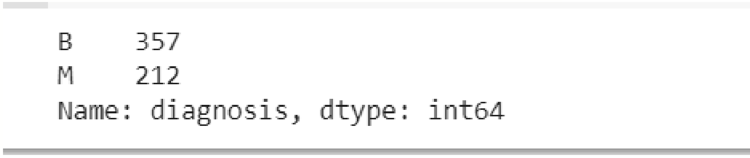
Drop the column with all missing values and get the new count of the number of rows and columns are (569, 32).



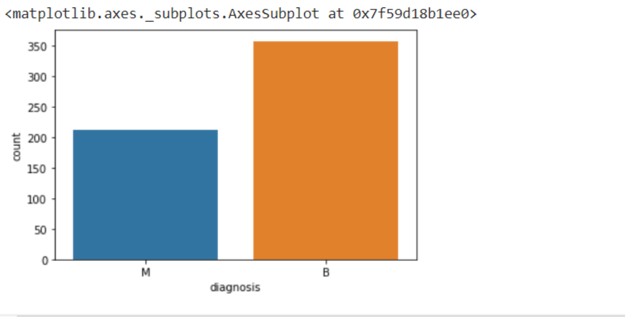
```
id                          0
diagnosis                   0
radius_mean                 0
texture_mean                0
perimeter_mean              0
area_mean                   0
smoothness_mean             0
compactness_mean            0
concavity_mean              0
concave points_mean       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

Figure 3: Measuring Parameters

Get count of the number of Malignant(M) or Benign(B) cells and visualize the count in graphical form.



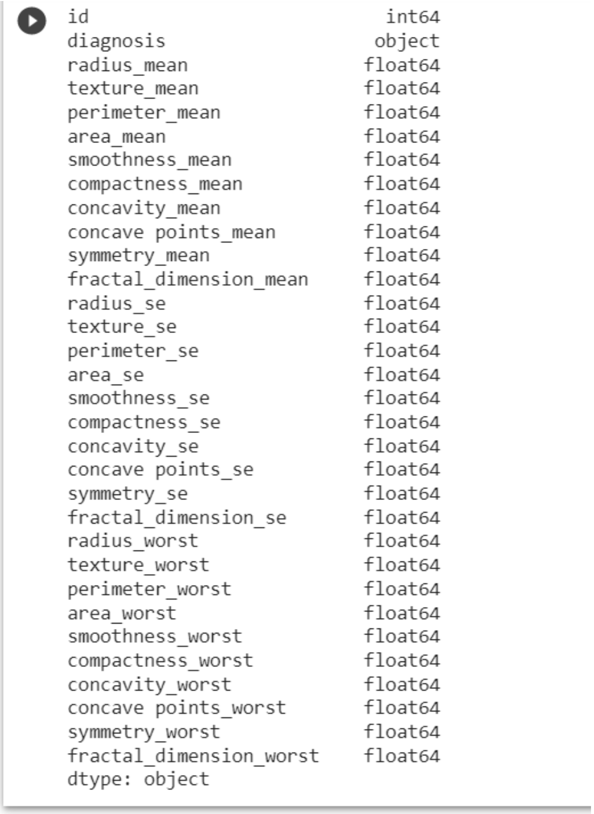
```
B    357
M    212
Name: diagnosis, dtype: int64
```



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f59d18b1ee0>
```

Figure 4: malignant & benignant tumour cases in graphical form

The data types to see which columns need to be encoded.



```
id                         int64
diagnosis                 object
radius_mean              float64
texture_mean             float64
perimeter_mean           float64
area_mean                float64
smoothness_mean          float64
compactness_mean         float64
concavity_mean           float64
concave points_mean      float64
symmetry_mean            float64
fractal_dimension_mean   float64
radius_se                float64
texture_se               float64
perimeter_se             float64
area_se                  float64
smoothness_se            float64
compactness_se           float64
concavity_se             float64
concave points_se        float64
symmetry_se              float64
fractal_dimension_se     float64
radius_worst             float64
texture_worst            float64
perimeter_worst          float64
area_worst               float64
smoothness_worst         float64
compactness_worst        float64
concavity_worst          float64
concave points_worst     float64
symmetry_worst           float64
fractal_dimension_worst  float64
dtype: object
```

Figure 5: Data types

Encoded the categorical data values and created pair plot.

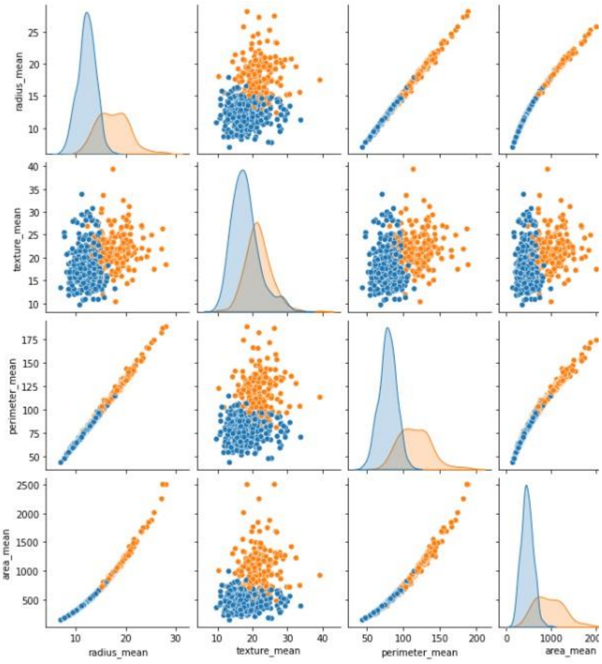
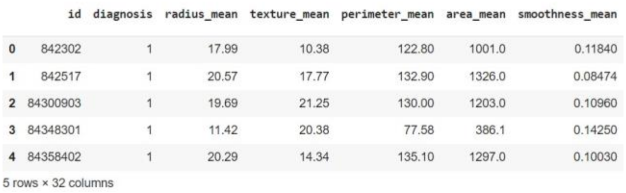


Figure 6 : Pair Plot

Print the first 5 rows of the new data.

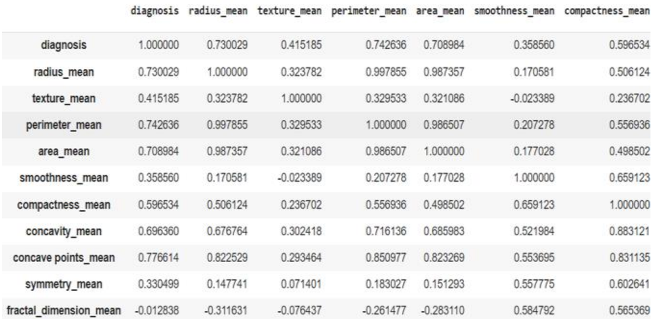


	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	842302	1	17.99	10.38	122.80	1001.0	0.11840
1	842517	1	20.57	17.77	132.90	1326.0	0.08474
2	84300903	1	19.69	21.25	130.00	1203.0	0.10960
3	84348301	1	11.42	20.38	77.58	386.1	0.14250
4	84358402	1	20.29	14.34	135.10	1297.0	0.10030

5 rows × 32 columns

Figure 7: New Data

Get the correlation column.



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
diagnosis	1.000000	0.730029	0.415185	0.742636	0.708984	0.358560	0.596534
radius_mean	0.730029	1.000000	0.323782	0.997855	0.987357	0.170581	0.506124
texture_mean	0.415185	0.323782	1.000000	0.329533	0.321086	-0.023389	0.236702
perimeter_mean	0.742636	0.997855	0.329533	1.000000	0.986507	0.207278	0.556936
area_mean	0.708984	0.987357	0.321086	0.986507	1.000000	0.177028	0.498502
smoothness_mean	0.358560	0.170581	-0.023389	0.207278	0.177028	1.000000	0.659123
compactness_mean	0.596534	0.506124	0.236702	0.556936	0.498502	0.659123	1.000000
concavity_mean	0.696360	0.676764	0.302418	0.716136	0.685983	0.521984	0.883121
concave points_mean	0.776614	0.822529	0.293464	0.850977	0.823269	0.553695	0.831135
symmetry_mean	0.330499	0.147741	0.071401	0.183027	0.151293	0.557775	0.602641
fractal_dimension_mean	-0.012838	-0.311631	-0.076437	-0.261477	-0.283110	0.584792	0.565369

Figure 8: Correlation column

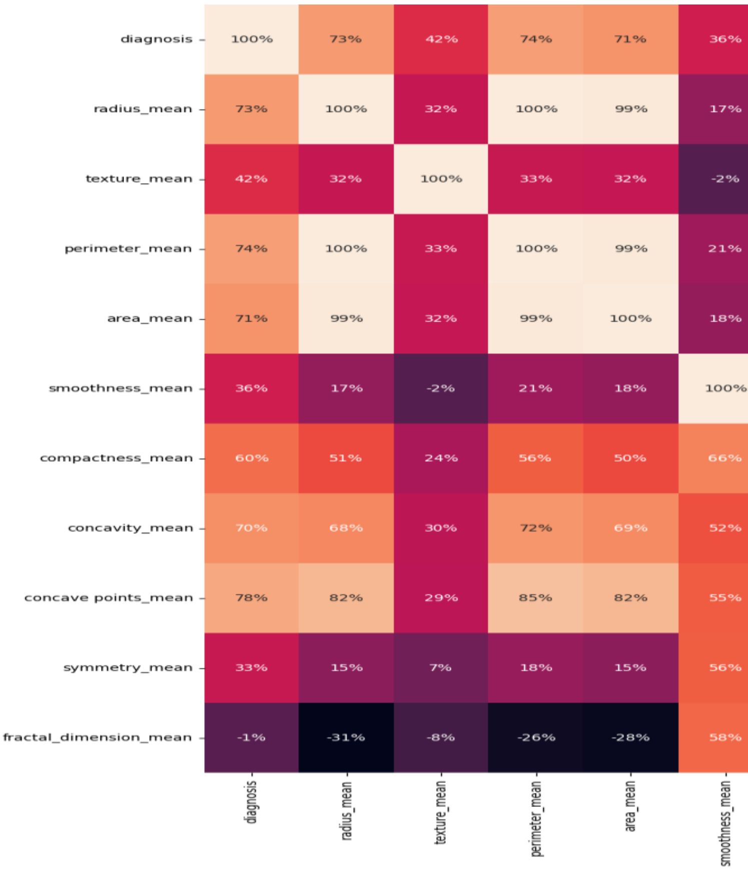
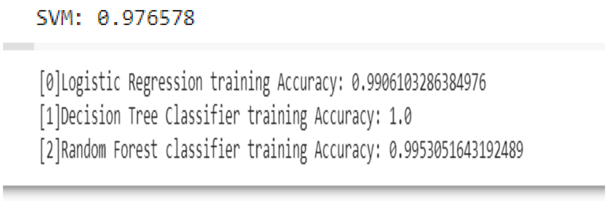


Figure 9: Visualize Correlation

After that split the data into testing data and training data and do the feature scaling with arrays.

Then created function for model selection using different algorithm as logistic regression, decision tree, random forest classifier and measure `accuracy to get following accuracy result.

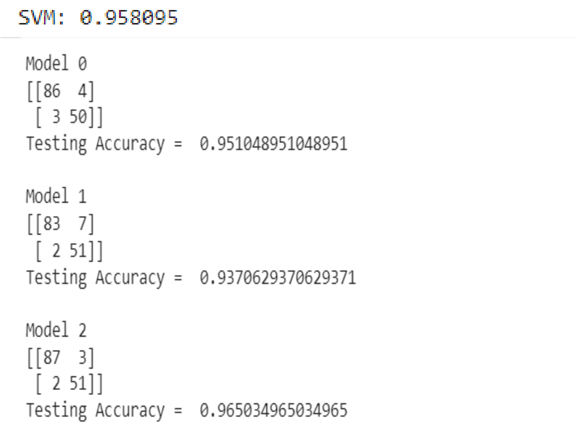


```
SVM: 0.976578

[0]Logistic Regression training Accuracy: 0.9906103286384976
[1]Decision Tree Classifier training Accuracy: 1.0
[2]Random Forest classifier training Accuracy: 0.9953051643192489
```

Figure 10: Training Data Algorithm Results

Test data with Confusion matrix and algorithm accuracy result



```
SVM: 0.958095

Model 0
[[86  4]
 [ 3 50]]
Testing Accuracy =  0.951048951048951

Model 1
[[83  7]
 [ 2 51]]
Testing Accuracy =  0.9370629370629371

Model 2
[[87  3]
 [ 2 51]]
Testing Accuracy =  0.965034965034965
```

6. Conclusion

In this paper, we have explored the application of machine learning techniques for the detection and prediction of breast cancer. The utilization of advanced algorithms, particularly Random Forest, has demonstrated promising results in accurately classifying breast cancer cases based on relevant features extracted from medical data.

The experiments conducted on the dataset showcased the effectiveness of the proposed machine learning model in distinguishing between benign and malignant tumours. The achieved accuracy and performance metrics underscore the potential of machine learning as a valuable tool in breast cancer diagnosis.

In this research, we split the data into training and testing data and applied different algorithms, confusion matrix, different model. It has been found that Random Forest performed better as compared to other algorithms with 97% accuracy.

7. Future Scope

However, there are still some points that can be studied in the future. the future scope for research in breast cancer detection and prediction using Random Forest extends beyond achieving high accuracy to encompass interpretability, integration with clinical workflows, and robust deployment in real-world healthcare settings. These directions aim to leverage the strengths of Random Forest and contribute to the development of reliable and effective tools for breast cancer diagnosis.

8. References

[1] S. N. Singh; Shivani Thakral, " Using Data Mining Tools for Breast Cancer Prediction and Analysis" , 2018 4th International Conference on Computing Communication and Automation(ICCCA), 10.1109/CCAA.2018.8777713.

[2] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods." 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (2018): 1-3.2011;61(6):409-418. doi:10.3322/caac.20134.

[3] H. L. Chen, B. Yang, J. Liu, and D. Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," Expert Syst. Appl., vol. 38, no. 7, pp. 9014–9022, 2011.

[4] V. Chaurasia and S. Pal, "Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability." 29-Jun-2017

[5] Rani, K. U., 2010, "Parallel Approach for Diagnosis of Breast Cancer using Neural Network Technique," International Journal of Computer Applications, 10(3), 1-5.

[6] Delen, D., Walker, G. and Kadam, A., 2005, "Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods," Artificial Intelligence in Medicine, 34(2), 113-127.

[7] Choi, J. P., Han, T. H. and Park, R. W, 2009, "A Hybrid Bayesian Network Model for Predicting Breast Cancer Prognosis," Journal of Korean Society of Medical Informatics, 15(1), 49-57. 8. Bellaachia, A. and Guven E., 20.

[8] Uma Ojha; Savita Goel, "A study on prediction of breast cancer recurrence using data mining techniques" , IEEE – 12-13 jaunuary2017, "2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence"

[9] Masud Rana Basunia; Ismot Ara Pervin; Md. Al Mahmud; Suman Saha; Mohammad Arifuzzaman, “ On Predicting and Analyzing Breast Cancer using Data Mining Approach”, 10.1109/TENSYMP50017.2020.9230871, 2020 IEEE Region 10 Symposium (TENSYMP).

[10] Dona Sara Jacob; Rakhi Viswan; V Manju; L PadmaSuresh; Shine Raj, “ A Survey on Breast Cancer Prediction Using Data Mining Techniques” ,IEEE- 10.1109/ICEDSS.2018.8544268, 2018 Conference on Emerging Devices and Smart Systems (ICEDSS).

[11] Amrita Sanjay; H Vinayak Nair; Sruthy Murali; K S Krishnaveni, “ A Data Mining Model To Predict Breast Cancer Using Improved Feature Selection Method On Real Time Data”, IEEE- 10.1109/ICACCI.2018.8554450, 2018 International Research on Advances in Computing, Communications and Informatics (IRACCI).

[12] K. Shilpa; T. Adilakshmi; K. Chitra, “ Applying Machine Learning Techniques To Predict Breast Cancer”, IEEE-10.1109/ICPS55917.2022.00011, 2022 International Research on Advances in Computing, Communications and Informatics (IRACCI).

[13] S. Saranya; S. Sasikala, “ Diagnosis Using Data Mining Algorithms for Malignant Breast Cancer Cell Detection”, IEEE- 10.1109/ICECA49313.2020.9297481, 2020 International Research on Advances in Computing, Communications and Informatics (IRACCI).

[14] D. Devikanniga, A. Ramu and A. Haldorai, "Efficient Diagnosis of Liver Disease using Support Vector Machine Optimized with Crows Search Algorithm", EAI Endorsed Transactions on Energy Web, pp. 164177, Jul. 2018.

[15] Shakya Subarna, "Analysis of Artificial Intelligence based Image Classification Techniques", Journal of Innovative Image Processing (JIIP), vol. 2, no. 01, pp. 44-54, 2020.

[16] H. Ayatollahi, L. Gholamhosseini and M. Salehi, "Predicting Coronary artery Disease: a Comparison between Two Data Mining Algorithms", BMC Public Health, vol. 19, no. 448, pp. 1-9, Apr. 2019.

[17] Senapati, M. R., Mohanty, A. K., Dash, S., and Dash, P. K., 2013, “Local Linear Wavelet Neural Network for Breast Cancer Recognition,” Neural Computing and Applications, 22(1), 125-131.

[18] Fallahi, A. and Jafari S., 2011, "An Expert System for Detection of Breast Cancer Using Data Pre-processing and Bayesian Network," International Journal of Advanced Science and Technology, 34, 65-70. 25.

[19] Pena-Reyes, C. A. and Sipper M., 1999, "A Fuzzy-genetic Approach to Breast Cancer Diagnosis," Artificial Intelligence in Medicine, 17(2), 131-155.

[20] DeSantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. CA Cancer J Clin. learning methods.” 2018 ElectricElectronics, Computer Science, Biomedical Engineering’s' Meeting (EBBT) (2018): 1-3.2011;61(6):409-418, 10.3322/caac.20134.

[21] Ammu P K and Preeja V. Article: Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications 61(12):39-44, January 20.