JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ENSEMBLE MACHINE LEARNING TECHNIQUES FOR CANCER PREDICTION

[1]Reuel Philip, [2]Y Aananditha, [3]Gnana Teja, [4]Dr. N Ch. Sriman Narayana Iyengar
[1,2,3]IV Year Students and [4]Professor & Dean, Library
Dept. of Information Technology, Sreenidhi Institute of Science and Technology(A), Hyderabad, India.

Abstract: Cancer is mostly recognized in humans, the reason for the mounting rate. The discovery of disease detection takes a lengthy time physically and little convenience of the system, an obligation to alter automatic diagnosis classification primary identification of cancer. For cancer tumour finding, castoff classification methods of machine learning and ensemble machine learning for earliest data can guess the kind of novel input data. In this paper emphasis on employment of models is complete on the dataset. The outcome of accuracy, precision, recall and False Positive Rate The efficiency of ensemble machine learning algorithms is measured and compared with other algorithms' results. From statistical analysis, the ensemble model predicts the accuracy rate is 89.3 % for decision tree and random forest models. 90.4% for logistic regression and support vector models. If you compare these two ensemble models with individual models, they provide a higher accuracy rate.

Keywords: Ensemble Machine Learning, Cancer, Prediction, Accuracy.

I. INTRODUCTION

The detection of group prognostic heritable changes might assist in important the consequence of cancer handling, permitting the social stratification of affected roles into different cohorts for discerning therapeutic protocols [1]. Cancer is one of the most dangerous diseases in the world. It affects more than a million throughout the world. India also affects more than 1 lakh people with cancer [2]. The most important things are the symptoms, diagnosis and treatment of cancer. The following diagram shows the two different cells. 1 indicates the cancer cells and 2 indicates the normal cells [3].

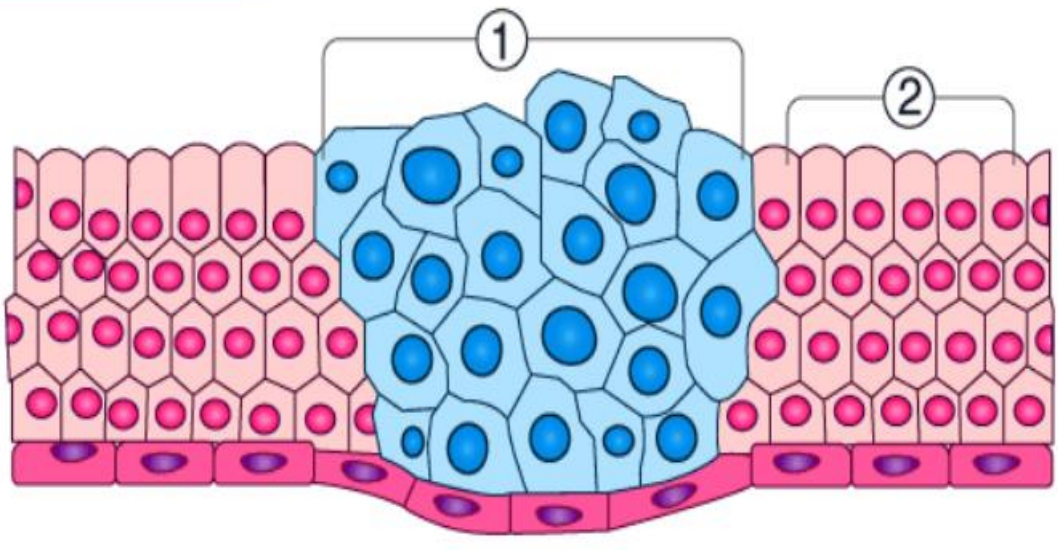


Figure 1: Cancer cells

A tumour is categorized into three kinds based on its ability to spread. **A benign Tumour is** localized at a specific place in the body [4]. **Malignant tumours** sense that they will produce rapidly and blow out to other regular tissues of the body [5]. Premalignant Tumour tumour may be benign but is observed to have the characteristics of a malignant tumour. It may not have metastasized yet, but it has the potential to turn cancerous [6].

The following paper is arranged in order of introduction in section one. Section two proposes the system and architecture. Section three states the consequences and analysis. Concludes the paper with the final section.

II. PROPOSED SYSTEM AND ARCHITECTURE

The proposed system is constructed with machine learning algorithms with ensemble learning of different models [7]. These models combine and predict the result optimally. Different phases are included in this architecture. The procedure is going in the normal way of a machine learning project. But one thing is different from methodology-integrated models for retrieving the optimal result [8].

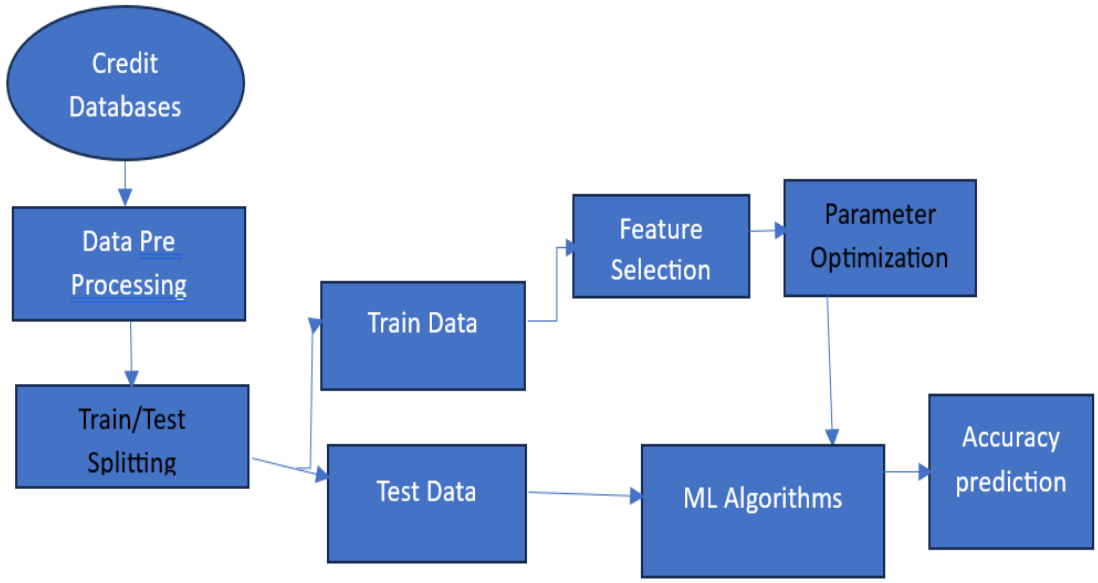


Figure 2: Diagrammatic representation of our proposed system

The architecture of the proposed system starts with input data. After input data given to the system, pre-process the data to generate accurate results to remove abnormal values [9]. Then select the suitable algorithms for given dataset [11]. The data can be divided into test and training examples. Feature selection also most important for accurate result generation. After that prepare the model and train the model and run the system, generate the results [10].

III. RESULTS AND ANALYSIS

Ensemble learning is used for generating the optimal results using different machine learning algorithms [11]. Different phased are included in this process, already explained in a proposed system with neat architecture diagram 2.

3.1 Data Collection

Data collection is also the most important phase for finding the shape of the dataset.

Table 1: Dataset

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

The following table 1 shows the dataset with 569 rows and 33 columns with 33 attributes using in our experiment.

3.2 Data Visualization

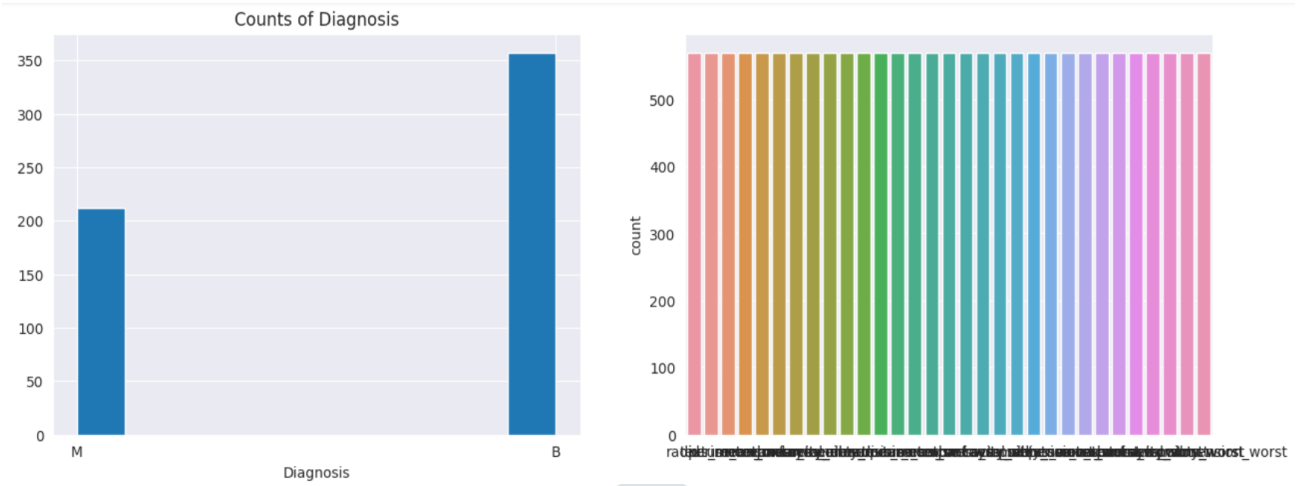


Figure 3: Visualization of data in bar chart

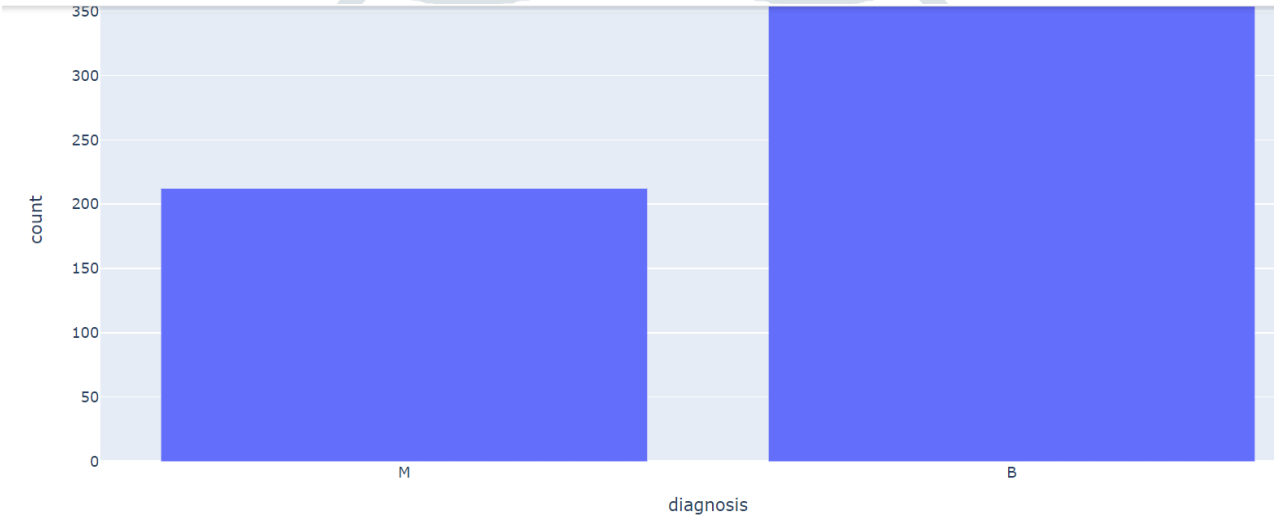
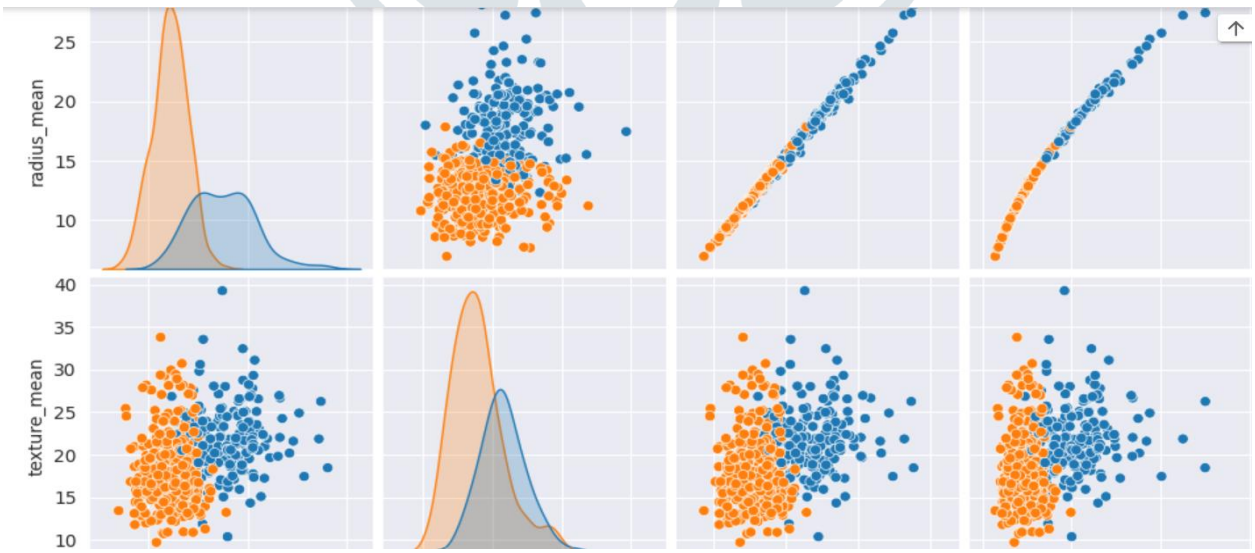


Figure 4: Histogram representation of data for diagnosis



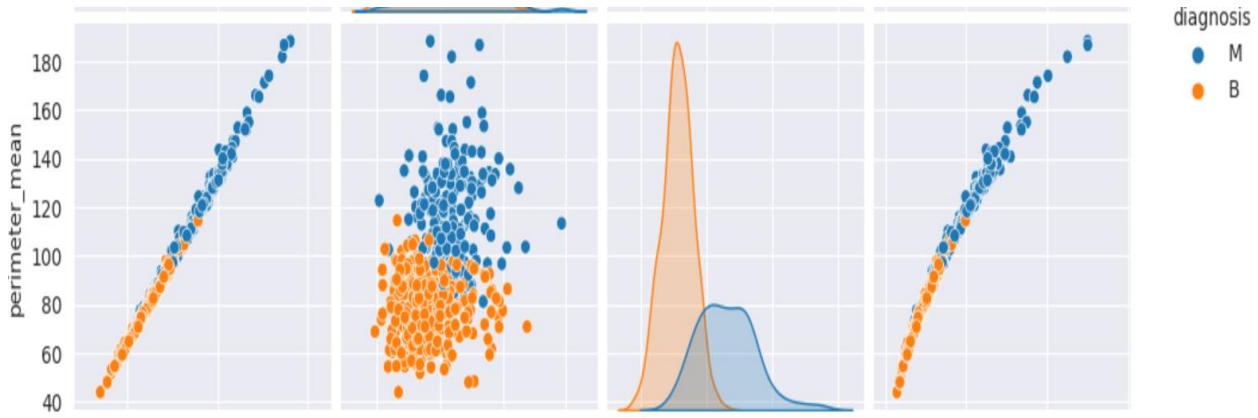


Figure 5: Visualization of attribute representation of data with multiple graphs

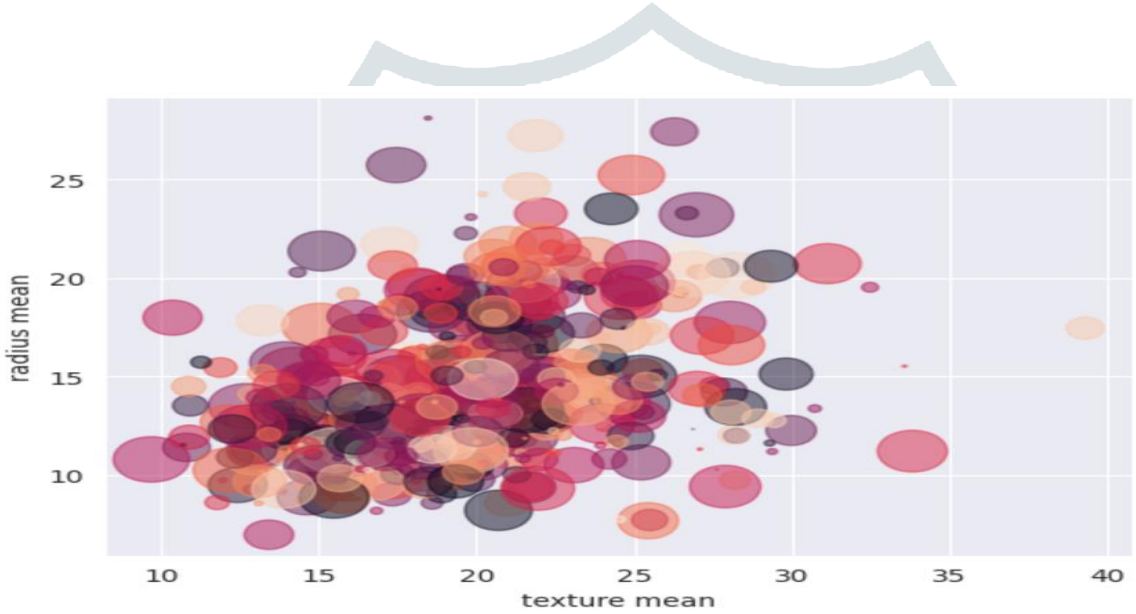


Figure 6: randomly mean representation of data cells

3.3 Data Filtering

Data filtering means preprocessing of the data for removing abnormal values. We can see the outcome classify values converted into 0 and 1. Here find the correlation between other features, mean features only.

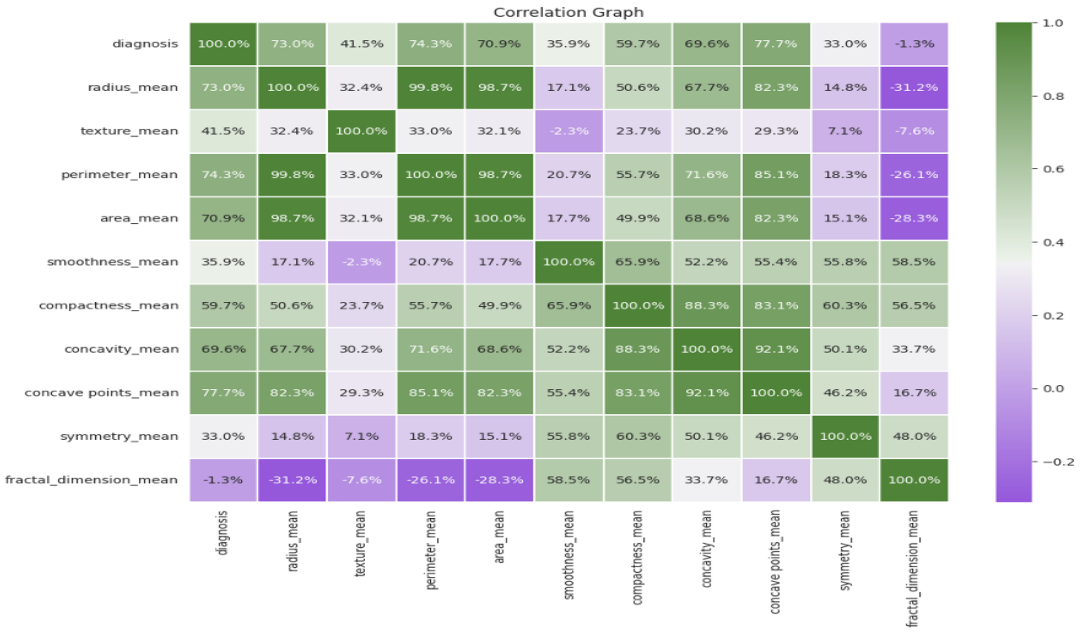


Figure 7: correlation graph of different attributes

The correlation of attributes is related to some other attributes. Few attributes are independent and redundant. Identify the redundant attributes and remove them from the dataset. The figure shows the relation of every attribute. In machine learning implementation heat maps also represent the graphical implementation of correlation of different variables. It finds near-by relations of different variables.

A heatmap is a graphical implement that shows the correlation between numerous variables as a matrix. It's shows that shows how closely connected dissimilar variables. The figure represents the correlation.

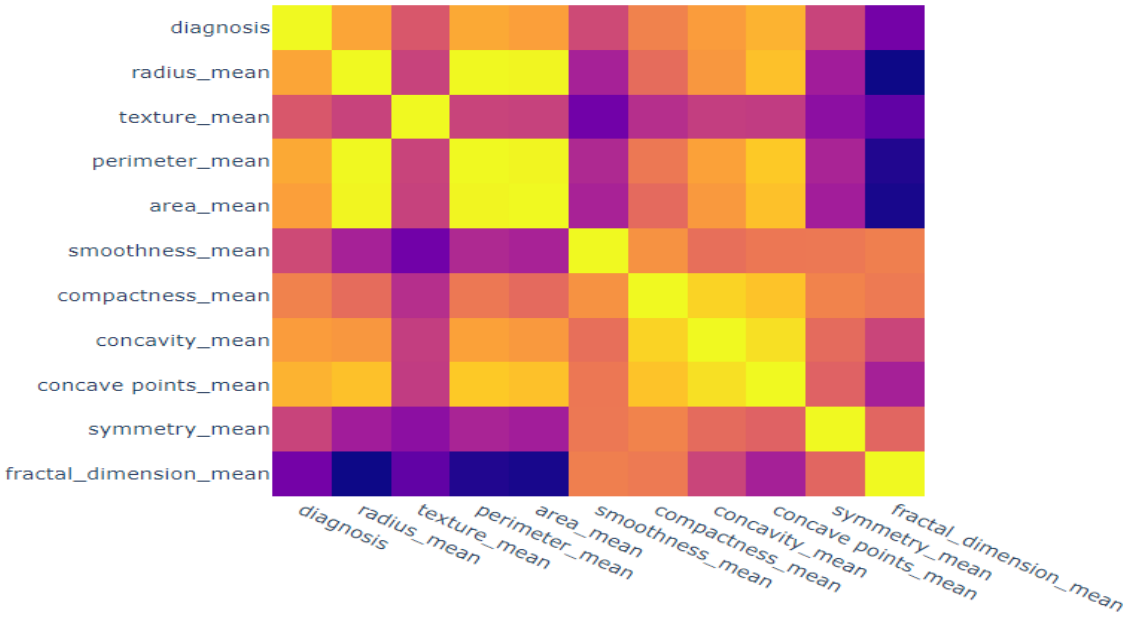


Figure 8: Heat Map representations

3.4 Implementation of Model

Model implementation concentrates on the following phases.

- Train and Test Splitting of the data.
- Scaling and model assortment.
- Import Models.
- Check the Model Accuracy.
- Errors and its Validations.

III.4.1 Selection of Features

For feature selection and then prediction of result based on the dependent and independent variables.

	radius_mean	perimeter_mean	area_mean	symmetry_mean	compactness_mean	concave points_mean
0	17.99	122.80	1001.0	0.2419	0.27760	0.14710
1	20.57	132.90	1326.0	0.1812	0.07864	0.07017
2	19.69	130.00	1203.0	0.2069	0.15990	0.12790
3	11.42	77.58	386.1	0.2597	0.28390	0.10520
4	20.29	135.10	1297.0	0.1809	0.13280	0.10430
...	...	...	...	...	...	...
564	21.56	142.00	1479.0	0.1726	0.11590	0.13890
565	20.13	131.20	1261.0	0.1752	0.10340	0.09791
566	16.60	108.30	858.1	0.1590	0.10230	0.05302
567	20.60	140.10	1265.0	0.2397	0.27700	0.15200
568	7.76	47.92	181.0	0.1587	0.04362	0.00000

The data can be split into training and test sets by 33% with 15 fixed records. Feature scaling is useful for optimal outcomes. The general score of a sample x is measured as:

$$X = (y - u) / t \quad (1)$$

3.5 Model Selecting and Pred Prediction

3.5.1 Model Construction

Now, we are ready to build our prediction model, for the I made function for model building and performing prediction and measuring its prediction and accuracy score. Now, Train the models one by one and show the classification report of particular models. Table 2 shows the precision, recall, f1 score, support and accuracy values of classification models.

Table 2: Logistic Regression and Random Forest Report

Classification Report of 'LogisticRegression '

	precision	recall	f1-score	support
0	0.90	0.96	0.93	115
1	0.92	0.84	0.88	73
accuracy			0.91	188
macro avg	0.91	0.90	0.90	188
weighted avg	0.91	0.91	0.91	188

Classification Report of 'RandomForestClassifier '

	precision	recall	f1-score	support
0	0.92	0.96	0.94	115
1	0.93	0.88	0.90	73
accuracy			0.93	188
macro avg	0.93	0.92	0.92	188
weighted avg	0.93	0.93	0.93	188

The following Table 3 shows the precision, recall, f1 score, support and accuracy values of decision tree and support vector tree algorithms.

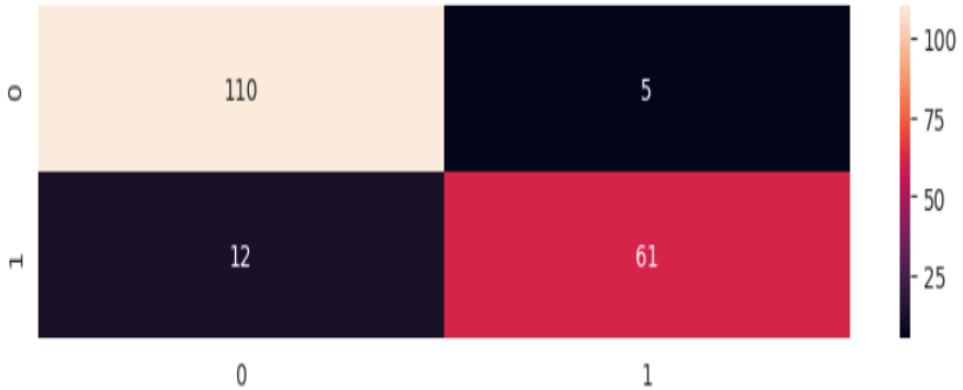
Table 3: Decision Tree and Support Vector Report

Classification Report of 'DecisionTreeClassifier '

	precision	recall	f1-score	support
0	0.90	0.96	0.93	115
1	0.92	0.84	0.88	73
accuracy			0.91	188
macro avg	0.91	0.90	0.90	188
weighted avg	0.91	0.91	0.91	188

Classification Report of 'SVC '

	precision	recall	f1-score	support
0	0.90	0.97	0.93	115
1	0.94	0.84	0.88	73
accuracy			0.91	188
macro avg	0.92	0.90	0.91	188
weighted avg	0.92	0.91	0.91	188



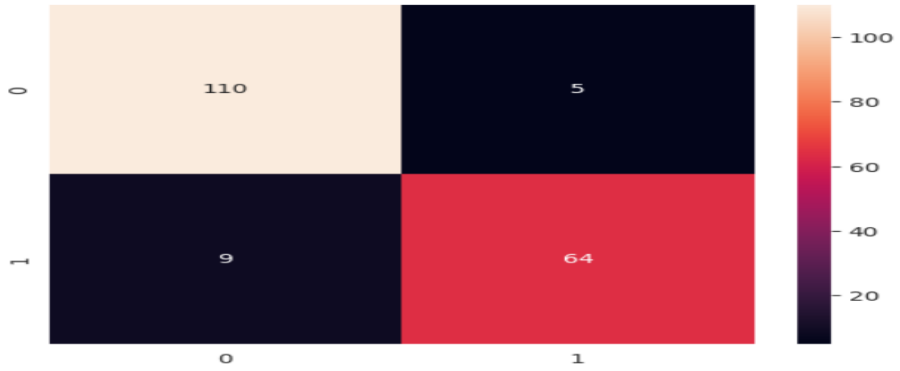


Figure 9: Confusion Metric graph values representation

The figure 9 shows that confusion metric of classification algorithms x label shows the negative positive and y label shows the True Positive. While predicting we can store model's score and prediction values to new generated data frame [12]. The following table 4 describe the comparative study of different models. Mostly all modes generate near by accuracy rate. Random Forest generate highest accuracy rate among four [13].

Table 4: accuracy of four models

	model_name	score	accuracy_score	accuracy_percentage
0	LogisticRegression	0.916010	0.909574	90.96%
1	RandomForestClassifier	0.992126	0.925532	92.55%
2	DecisionTreeClassifier	1.000000	0.909574	90.96%
3	SVC	0.923885	0.914894	91.49%

III.5 Ensemble Models

Ensemble model means

3.5.1 Decision Tree and Random Forest

```
from sklearn.ensemble import VotingClassifier

dt=DecisionTreeClassifier(max_features='sqrt', min_samples_leaf=3);
rf=RandomForestClassifier(min_samples_leaf=2, min_samples_split=5);

ensemble_model = VotingClassifier(estimators=[('decision_tree', dt), ('random_forest', rf)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.8936170212765957
```

3.5.2 Support Vector Machine and Logistic Regression

```
from sklearn.ensemble import VotingClassifier

svc=SVC(C=10, gamma=0.001);
lr=LogisticRegression(C=0.001, solver='liblinear');

ensemble_model = VotingClassifier(estimators=[('svc', svc), ('logistic_regression', lr)], voting='hard')

ensemble_model.fit(X_train, y_train)

# Make predictions on the testing data
predictions = ensemble_model.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, predictions)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.9042553191489362
```

If you observe the 3.6.1 and 3.6.2 ensemble models the accuracy rate is 89.3 % for decision tree and random forest models. 90.4% for logistic regression and support vector models. If you compare these two ensemble models with individual models provide a higher accuracy rate. (Table 4).

IV. CONCLUSION

For cancer tumour finding, castoff classification methods of machine learning and ensemble machine learning for earliest data can guess the kind of novel input data. In this paper emphasis on employment of models is complete on the dataset. The outcome of accuracy, precision, recall and False Positive Rate The efficiency of ensemble machine learning algorithms is measured and compared with other algorithms' results. From statistical analysis, the ensemble model predicts the accuracy rate is 89.3 % for decision tree and random forest models. 90.4% for logistic regression and support vector models. If you compare these two ensemble models with individual models provide a higher accuracy rate.

REFERENCES

[1] Mandeep Rana, Pooja Chandorkar, Alishiba Dsouza, “Breast cancer diagnosis and recurrence prediction using machine learning techniques”, International Journal of Research in Engineering and Technology Volume 04, Issue 04, April 2015.

[2] Vikas Chaurasia, BB Tiwari and Saurabh Pal – “Prediction of benign and malignant breast cancer using data mining techniques”, Journal of Algorithms and Computational Technology

[3] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper.

[4] Nidhi Mishra, Naresh Khuriwal.- “Breast cancer diagnosis using adaptive voting ensemble machine learning algorithm”, 2018 IEEMA Engineer Infinite Conference (eTechNxT), 2018.

[5] Logistic Regression for Machine Learning - Machine Learning Masteryhttps://machinelearningmastery.com/logistic-regression-for machine-learning/

[6] Ebrahim Edriss Ebrahim Ali1 , Wu Zhi Feng2- “Breast Cancer Classification using Support Vector Machine and Neural Network” International Journal of Science and Research(IJSR) Volume 5 Issue 3, March 2016.

[7] Ch. Shravya, Pravallika and Subhani Shaik,” Brest cancer prediction using machine learning Techniques”, International Journal of Innovative Technology and Exploring Engineering, Vol. 8, Issue 6, 2019.

[8] Shiva Keertan J and Subhani Shaik,” Machine Learning Algorithms for Oil Price Prediction”, International Journal of Innovative Technology and Exploring Engineering, Volume-8 Issue-8, 2019.

[9] Dr. R. Vijaya Kumar Reddy, Dr. Shaik Subhani, Dr. G. Rajesh Chandra, Dr. B. Srinivasa Rao,” Breast Cancer Prediction using Classification Techniques”, International Journal of Emerging Trends in Engineering Research, Vol. 8, No.9,2020.

[10] Ms. Mamatha, Srinivasa Datta and Subhani Shaik,” Fake Profile Identification using Machine Learning Algorithms”, International Journal of Engineering Research and Applications (IJERA), Vol.11, Series-2, July-2021.

[11] R. Vijaya Kumar Reddy, Subhani Shaik, B. Srinivasa Rao, “Machine learning based outlier detection for medical data” Indonesian Journal of Electrical Engineering and Computer Science, Vol. 24, No. 1, October 2021.

[12] Neeraja, Anupam, Sriram, Subhani Shaik and V. Kakulapati,” Fraud Detection of AD Clicks Using Machine Learning Techniques”, Journal of Scientific Research and Reports, Volume 29, Issue 7, Page 84-89, June-2023.

[13] Subhani Shaik and Dr. Uppu Ravibabu, "Detection and Classification of Power Quality Disturbances Using curvelet Transform and Support Vector Machines", in the 5th IEEE International Conference on Information Communication and Embedded System (ICICES-2016) at S.A Engineering college, Chennai, India on 25th -26th, February 2016.