Breast Cancer Diagnosis using Python

Salma Yousaf and Shakeel Khan

September 9, 2020

Breast Cancer Diagnosis using Python

Salma Yousaf, Shakeel Khan, June 2020, Department of Computer Science, Riphah International University Lahore, Pakistan

Abstract

Breast cancer is the most common disease which is found more in women and the cause of death of them. It is spreading worldwide. In this research paper Diagnosis of breast cancer is performed in python using different classifier. Here I used the dataset of breast cancer and apply different machine learning algorithms to get different results. There are two main diagnoses categories are in breast cancer. One is called malignant diagnosis and the other one is called benign diagnosis. These two types of diagnosis are processed in python using different machine learning algorithms.

Keywords: Breast cancer, Classifiers, SVM, Naïve byes, k nearest neighbors, visualization graphs.

Introduction

Breast Cancer is common in women and every year no of people died because of breast cancer all over the world. It seems to be one of the most common type of cancer now days.

We can see that medical database is size and numbers both are growing very but mostly these types of database unable to examine and search about the valuable and hidden data and knowledge. The most advanced data mining techniques can be used to discover hidden patterns and relationships.[1]

Huge amount of time infuse in the manual diagnosing and minor amount of diagnostic system available emphasize the development of automated diagnosis for early diagnosis of the disease.[2]

In this research paper python is used for the diagnosis of breast cancer. Dataset of breast cancer have been taken. Different machine learning algorithms are applied on the dataset of breast cancer for getting different accuracies. This dataset has 569 rows and 33 columns. Ranges of entries are from 0 to 568. 357 are malignant diagnosis and 212 are benign diagnosis. Heat map graph, frequency graph and different visualization graphs shows the diagnosis value. By taking the different values for train and test data, tables have been drawn for showing the different accuracies of various classifiers.

Problem Statement

How to diagnose the malignant and benign breast cancer using different machine learning algorithms in human beings?

How to show the number of benign and malignant diagnosis in the form of bar graph?

Literature Review

1. Some patients were admitted in the Iranian Center for Breast Cancer (ICBC) program dated 1997 to 2008. The dataset which is used for this paper have 1189 records, 22 predictor variables, and one outcome variable. Various machine learning techniques and algorithms were used like Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models. The main aim of this research paper is to contrast the performance of these three well-known algorithms on our data through sensitivity, specificity, and accuracy.

2. This paper shows whether the breast cancer is benign or malignant and forecast the recurrence and non-recurrence of malignant cases after a specific period. For this they used machine learning techniques and algorithms such as Support Vector Machine, Logistic Regression, KNN and Naive Bayes. These methods are coded in MATLAB using UCI machine learning repository. They contrast the correctness of different techniques and examined the results. Their results shows that SVM is most suited for guessing inspection and KNN execute best for our comprehensive procedure.

3. This paper narrates about the application of machine learning algorithms in recognizing cancer in human. It also provides the information of neural network, its learning rules which are used here to refine the correctness of predicting breast cancer.

4. This paper narrates a new hybrid algorithm that depends on back-propagation and radial basis function-10 based neural networks for forecast. The algorithm has been developed in an open source-based environment. The algorithm was12 tested on a 13-year dataset (1995–2008). This paper compares the13 algorithm and proves its correctness and regulation with different14 platforms. Nearly80%accuracyand88%positive predictive value15 and reactivity were recorded for the algorithm. The results were16 encouraging; 40–50% of negative predictive value and specificity17 warrant further work.

5. In this scenario, correctness forecast of BC behavior assumes an important role, since it aids clinicians in their decision-making process, enabling a more personalized treatment for patients. This research work try to supply an over view of the forecast of BC recurrence using machine learning techniques. The challenge is to accurately predict recurrence events, within a binary outcome (yes/no). This challenge surround not only the choice of a good dataset (containing quality data) but also the selection of the most suitable attributes, as well as the most commanding algorithm.

6. In this paper, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. The main

objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

Methodology

Data set of breast cancer used and its shows us the different results in different form. This dataset has 569 rows and 33 columns. Ranges of entries are from 0 to 568. This data has total 569 diagnoses in which 357 diagnosis are malignant while 212 diagnoses are benign. Heat map for this dataset is shown in the figure 1.1. different data attributes means are shown like radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean etc.

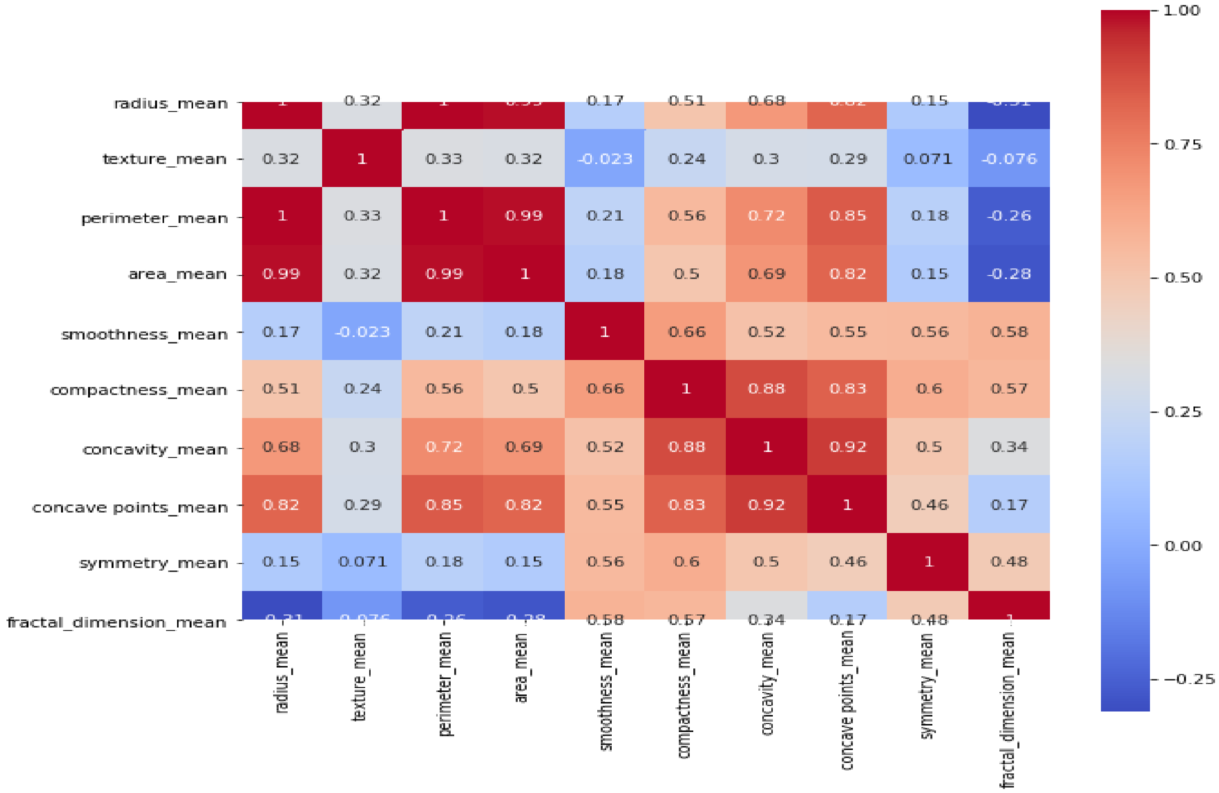


Fig 1.1 Heat map of multiple attributes

Figure 1.2 shows the visualization of the data in form of bar graph. It shows the data in easily readable and visual form. This graph shows the malignant and benign diagnosis on x-axis and number of this diagnosis on the y-axis. Letter B is for benign diagnosis while M is for malignant diagnosis.

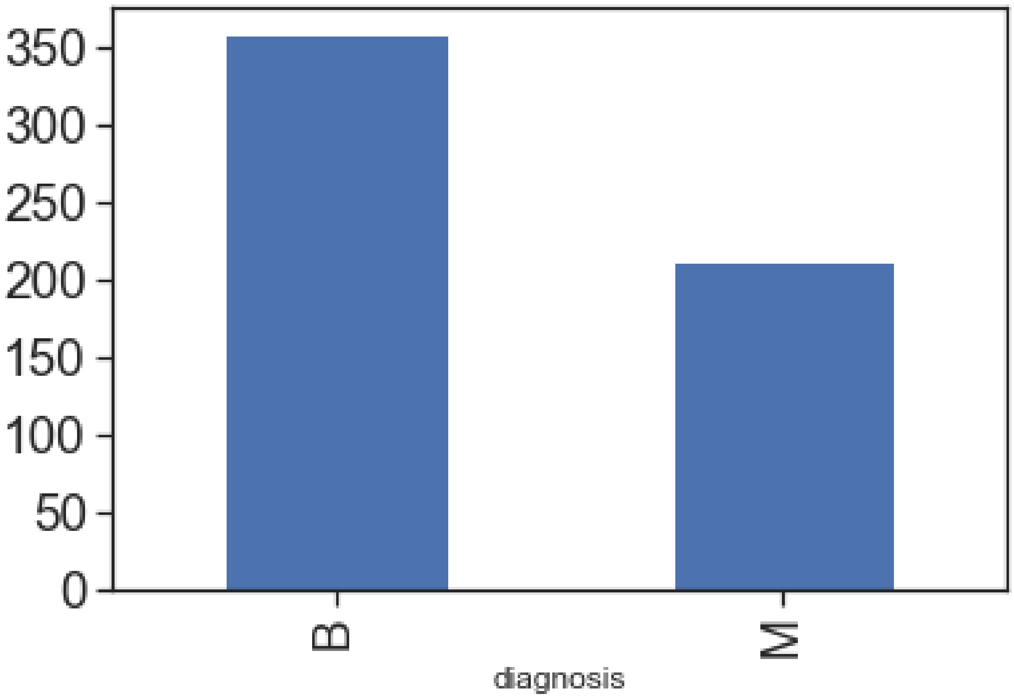


Fig 1.2 Bar graph of frequencies of benign and malignant diagnosis

Figure 1.3 shows the visualization of the attributes of a heat map in graphical form. Frequencies of malignant and benign diagnosis is shown. Two peaks are shown in the graph, one for malignant diagnosis and other for the benign diagnosis. Red color is for malignant(M) and blue color is for benign(B). Visualization graph for each attribute like radius_mean, texture_mean etc is drawn separately. In radius_mean, parameter_mean, area_mean, compactness_mean, concavity_means, concave_point means, symmetry_mean and fractal dimension_mean have blue peak high. It means they shows the highest values for benign diagnosis, while rest of the graphs shows the highest values for the malignant diagnosis. The highest frequency (nearly 70) is shown by the fractal dimension_mean for benign diagnosis and the lowest value (nearly 1.8) is of the perimeter_mean for malignant diagnosis.

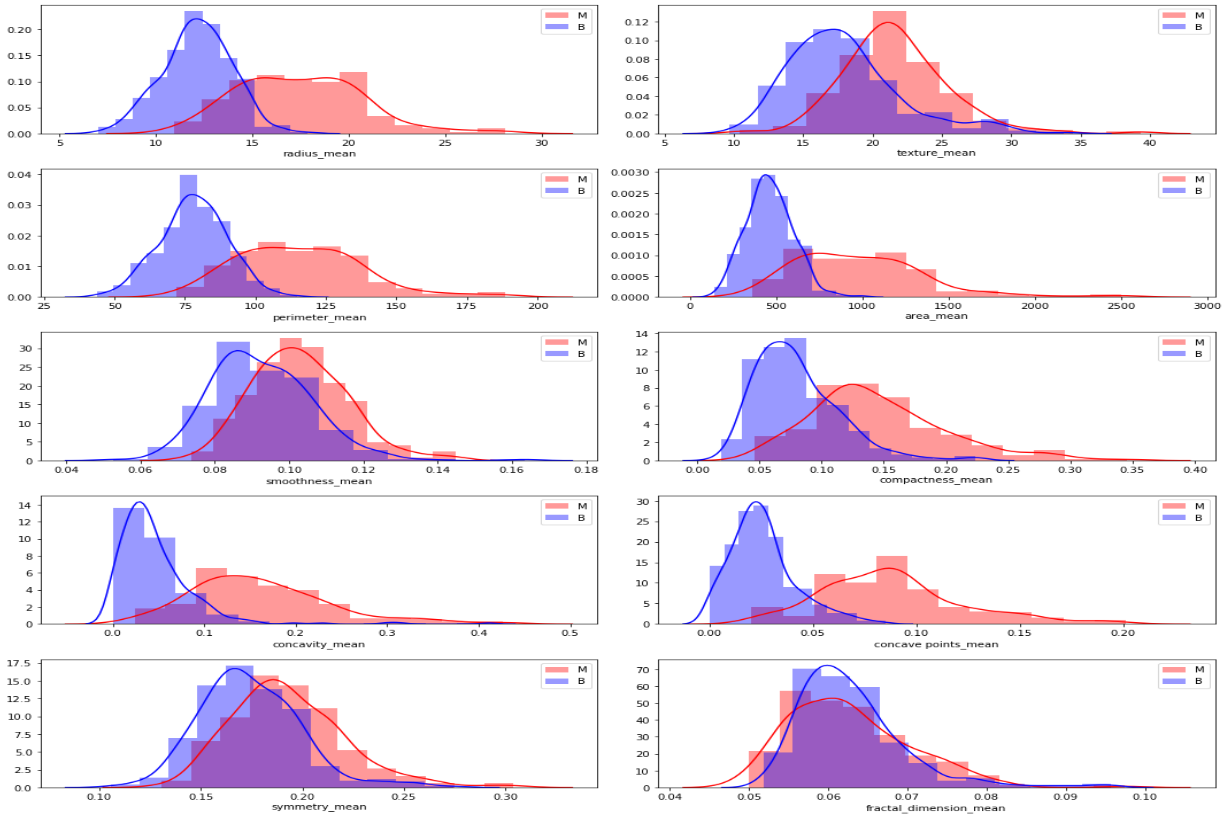


Fig 1.3 Visualization of the attributes of a heat map in graphical form

Figure 1.3 shows the graph which tells us about the diagnosis error separately for each attribute. The highest error rate is in the area_mean for malignant diagnosis which is nearly 1500. The lowest error rate is in the cocavepoint_mean and concavity_mean for benign diagnosis which is nearly zero.

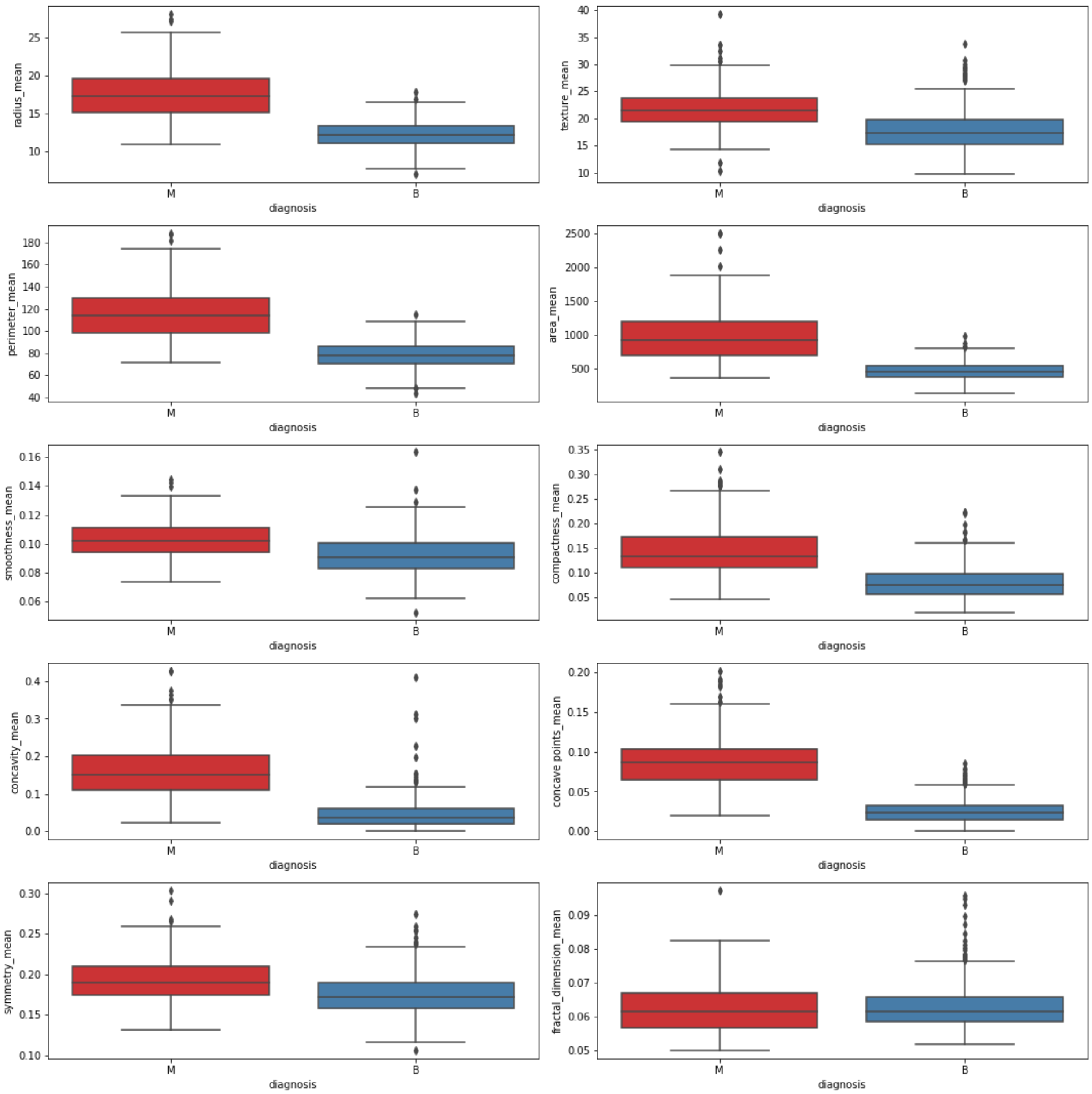


Fig 1.4Visuallization graph

SGD classifier is one of the quick approach to fitting linear classifiers and regressors under convex loss functions i.e. Logistic Regression and SVM. Here is the SGD classifier accuracy, Cross validation score and execution time is shown.

SGD Classifier Accuracy: 92.11%
Cross validation score: 79.41% (+/- 21.99%)
Execution time: 0.34122 seconds

Support Vector Classifier, Nu Support Vector Classifier and Linear Support Vector Classifiers are used and applied on the train and test data. This is the SVC classifier accuracy, cross validation score and execution time.

SVC Accuracy: 69.30%

Cross validation score: 71.70% (+/- 4.07%)
Execution time: 0.30181 seconds

NuSVC Accuracy: 69.30%
Cross validation score: 71.88% (+/- 3.97%)
Execution time: 0.22486 seconds

LinearSVC Accuracy: 81.58%
Cross validation score: 82.26% (+/- 10.88%)
Execution time: 0.24239 seconds
t KNeighborsClassifier
Accuracy: 93.86%
Cross validation score: 88.60% (+/- 6.96%)
Execution time: 0.16687 seconds

Name of Clssifier	Accuracy	Cross validation score	Execution time
SGD Classifier	92.11%	79.41% (+/- 21.99%)	0.34122 seconds
SVC Classifier	69.30%	71.70% (+/- 4.07%)	0.30181 seconds
NuSVC Classifier	69.30%	71.88% (+/- 3.97%)	0.22486 seconds
LinearSVC	81.58%	82.26% (+/- 10.88%)	0.24239 seconds
KNeighborsClassifier	93.86%	88.60% (+/- 6.96%)	0.16687 seconds
GaussianNB	94.74%	91.40% (+/- 5.03%)	0.47455 seconds
Random Forest	93.86%	93.69% (+/- 4.67%)	0.33464 seconds
Extra Trees	92.11%	94.21% (+/- 4.82%)	0.094941 seconds
Dedicion Tree	92.11%	92.45% (+/- 4.33%)	0.036977 seconds

Table 1.1 Accuracies, cross validation score and execution time of classifiers

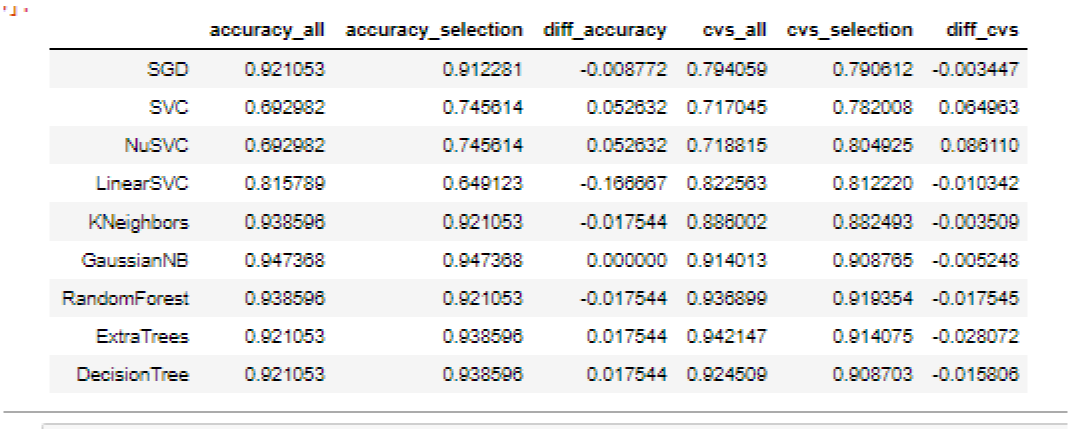
Now if change the values of train data and test data then we get the different values for each classifier. Different accuracy, cross validation score and execution time is generated. Now the below table shows the difference of values in case of each classifier.

Name of Clssifier	Accuracy	Cross validation score	Execution time
SGD Classifier	91.23%	79.06% (+/- 18.19%)	0.084925 seconds
SVC Classifier	74.56%	78.20% (+/- 8.67%)	0.1649 seconds
NuSVC Classifier	74.56%	80.49% (+/- 5.95%)	0.17763 seconds
LinearSVC	64.91%	81.22% (+/- 8.10%)	0.17589 seconds
KNeighborsClassifier	92.11%	88.25% (+/- 6.91%)	0.082948 seconds
GaussianNB	94.74%	90.88% (+/- 5.83%)	0.041973 seconds
Random Forest	92.11%	91.94% (+/- 5.46%)	0.13792 seconds

Extra Trees	93.86%	91.41% (+/- 6.47%)	0.09794 seconds
Dedicion Tree	93.86%	90.87% (+/- 4.02%)	0.033976 seconds

Table 1.2 Accuracies, cross validation score and execution time of classifiers with changed values

Now the table 1.3 shows the difference of the above two tables. It shows the values of accuracy_all, accuracy_selection, diff_accuracy, cvs_all, cvs_selection and diff_cvs for SGD, SVC, NuSVC, LinearSVC, KNeighbors, GussianNB, Random Forest, Extra Trees, Decision Tree.



	accuracy_all	accuracy_selection	diff_accuracy	cvs_all	cvs_selection	diff_cvs
SGD	0.921053	0.912281	-0.008772	0.794059	0.790612	-0.003447
SVC	0.692982	0.745614	0.052632	0.717045	0.782008	0.064963
NuSVC	0.692982	0.745614	0.052632	0.718815	0.804925	0.086110
LinearSVC	0.815789	0.649123	-0.166667	0.822563	0.812220	-0.010342
KNeighbors	0.938596	0.921053	-0.017544	0.886002	0.882493	-0.003509
GaussianNB	0.947368	0.947368	0.000000	0.914013	0.908765	-0.005248
RandomForest	0.938596	0.921053	-0.017544	0.936899	0.919354	-0.017545
ExtraTrees	0.921053	0.938596	0.017544	0.942147	0.914075	-0.028072
DecisionTree	0.921053	0.938596	0.017544	0.924509	0.908703	-0.015806

Table 1.3 Difference of the above two tables

So final accuracy for the test_size= 0.33 and random_state= 42 is given below

(0.8923884514435696, 0.9095744680851063)

Results

Visualization graphs show the diagnosis of malignant and benign. In bar graph benign diagnosis have highest value than malignant. While in other visualization graphs fluctuation occurs. Final accuracy accuracy for the test_size= 0.33 and random_state= 42 is given below (0.8923884514435696, 0.9095744680851063)

We also see the different results when applied the different classifier. By taking the different test data and train data values various classifiers show the different accuracies, cross validation score and execution time.

Conclusion/Future work

In future more machine algorithms can be applied to get the more accurate values. Other classifiers can also be used which shows the more accuracy, cross validation score and minimum execution time. Other types of graphs can also be drawn for more and easy readability. Scatter plot, line graph can also be drawn. More machine learning algorithms are applied on the dataset of breast cancer for more accuracy.

References

[1] Ahmad, L. G., Eshlaghy, A. T., Poorebrahimi, A., Ebrahimi, M., & Razavi, A. R. (2013). Using three machine learning techniques for predicting breast cancer recurrence. *J Health Med Inform*, *4*(124), 3.

[2] Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning techniques. *IJRET: International Journal of Research in Engineering and Technology eISSN*, 2319-1163.

[3] Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems*, *4*(3), 105.

[4] Baskaran, V., Guergachi, A., Bali, R. K., & Naguib, R. N. (2011). Predicting breast screening attendance using machine learning techniques. *IEEE Transactions on Information Technology in Biomedicine*, *15*(2), 251-259.

[5] Abreu, P. H., Santos, M. S., Abreu, M. H., Andrade, B., & Silva, D. C. (2016). Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, *49*(3), 1-40.

[6] Asri, H., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Using machine learning algorithms for breast cancer risk prediction and diagnosis. *Procedia Computer Science*, *83*, 1064-1069.