Journal of Computer Science

Original Research Paper

Visualization of Data Mining Techniques for the Prediction of Breast Cancer with High Accuracy Rates

[1]Vasudev Sharma, [1]Raj Kumar Rajasekaran and [2]Shreya Badhrinarayanan

[1]Data Analytics, Vellore Institute of Technology, Vellore, India
[2]Brighton and Sussex Medical School, East Sussex BN1 9PX, GB, United Kingdom

Article history
Received: 27-08-2018
Revised: 20-12-2018
Accepted: 19-01-2019

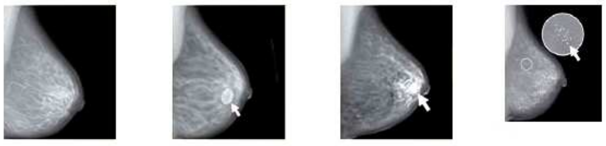
Corresponding Author:
Vasudev Sharma
Data Analytics, Vellore Institute of Technology, Vellore, India
Email: vasudevsharma74@yahoo.com

Abstract: Breast cancer is one of the leading causes of death in women worldwide. Around one in 30 women are affected by breast cancer. Mammography has helped in detecting breast cancer in the early stages which have reduced mortality. The diagnosis of breast cancer is dependent on a variety of parameters. In this paper, we aim to create the best model for predicting breast cancer through preprocessing, feature extraction, data visualization and prediction using breast cancer data. Various visualization techniques like violin plot, grid plot, swarm plot and heat plot were utilized for proper feature extraction which has improved the accuracy of our results. For the purpose of prediction, we have used algorithms like the random forest, decision tree with single and multiple predictors, along with the commonly used statistical model, logistic regression model. We have also relied on 5-fold cross-validation methods to measure the unbiasedness of the prediction models for performance reasons. An analysis of the models was carried out and the best model was selected based on its accuracy. The results showcased that the random forest model provided an accuracy rate of 94.724% with decent 5-fold cross-validation, followed by the decision tree model which had an accuracy rate of 100% with poor 5-fold cross-validation. This was followed by the logistic regression model which had an accuracy rate of 88.442% with a low 5-fold cross-validation score.

Keywords: Mammography, Data Visualization, Violin Plot, Swarm Plot, Random Forest, Logistic Regression, Decision Tree, 5-Fold Cross Validation

Introduction

Breast cancer is a type of cancer in women which emerges from the cells of the breast. There are various types of breast cancer. Treatment of breast cancer depends on many factors like its stage and concomitant surgeries and comorbidities of the individual. It is better to have knowledge of the type of cancer in the preliminary stages because early diagnosis leads to better outcomes. Diagnosis is dependent on the location of the mass and small scale calcification bunches which are vital in early identification of breast cancer. Microcalcification is little mineral stores inside the breast tissue which are appeared as little white-hued spots which may be caused by cancer. The differential diagnoses for masses in the breast can range from benign pimples (liquid-filled sacs) to malignant non-destructive strong tumors as illustrated from Fig. 1. The trouble in growth discovery is the variations from the norm from typical breast tissues are difficult to peruse due to their unpretentious appearance and vague margins. Automated devices can be used by radiologists in early identification of breast disease.



Normal mammogram

Benign cyst (not cancer)

Cancer

Calcium in your diet does not cause calcium deposits (calcification) in the breast

Fig. 1: The differences between benign and malignant tumors in breast cancer

Science Publications

Tumors can either be benign or malignant. A benign tumor survives on the local regions and cannot be spread by other means. A malignant tumor exploits nearby tissues, usually by entering the blood vessels and intruding nearby cells. Due to the high mortality seen in breast cancer, detection in the early stages with yield promising results. However, a promising tool to diagnosis breast cancer is yet to elucidate in scientific literature as it is still considered a challenging problem to solve. This paper introduces an easy and efficient approach that will aid in breast cancer prediction.

Review of Literature

Wolberg *et al.* (1994) in his paper used two main data mining algorithms which are neural networks and decision trees. The statistical model included was the logistic regression model. 10-fold cross-validation methods were also used to predict the accuracy of diagnosis of breast cancer. The dataset used for that research purpose was SEER Cancer Incidence Public-Use Database for the years 1973-2000. The files were made available on the SEER website. SEER is a part of Surveillance Research Program (SRP) at National Cancer Institute (NCI) which is responsible for gathering data from the datasets to the institutions and laboratories around the world for conducting analytical research. The SEER database is majorly used for analytical research purposes.

The results indicated that the decision tree model (C5) is the best predictor among all the 3 models with an accuracy of 93.6% on the sample, artificial neural networks came to be second with an accuracy of 91.2% and finally, logistic regression models scored an accuracy of 89.2%. A present a comparative study of all the above models along with 10- foldcross-validation provided with a glimpse into the relative predicting the ability of different data mining methods.

Chaurasia and Pal (2017a) in his paper utilized five famous data mining methods (Naïve Bayes, RBF Network, Simple Logistic, J48, and Decision Tree) to build up the expectation models utilizing a vast dataset (270 Heart sickness and 683 breast tumor cases). They additionally utilized 10-crease cross-approval strategies to quantify the fair gauge of the five prediction models for execution correlation purposes. The outcomes (in light of normal exactness of Heart and Breast Cancer informational index) showed the Naïve Bayes is the best indicator with 87.01% accuracy on the holdout test RBF Network turned out to be the second with 86.9% accuracy, Simple Logistic turned out to be third with 85.65% exactness, J48 turned out fourth with 84.85% accuracy and the Decision table models turned out to be the most exceedingly awful of the five with 83.34% accuracy.

Williams *et al.* (2015) in his paper investigated two unique data mining techniques utilized for the prediction of breast disease and their performance was compared all together assess the best classifier. Test results demonstrate the J48 choice trees is a superior model for the forecast of bosom disease dangers for the estimations of exactness, review, accuracy and mistake rates recorded for the two models which gave an accuracy of ~94% compared to navies Bayes of ~83%. Henceforth, a proficient and compelling classifier for breast cancer risk was recognized while numbers of the attribute used by the classifier can be expanded by expanding the sample size of the training set and consequently the improvement of a more exact model.

Rajesh and Anand (2012) in his paper has endeavored to characterize SEER breast malignancy information into the gatherings of "Carcinoma in situ" and "Malignant potential" utilizing C4.5 algorithm. They have utilized a training set of an arbitrary example of 500 records and afterward connected the arrangement govern set got to the full breast cancer dataset. They got a precision of ~94% in the training stage and accuracy of ~93% in the testing stage. They have compared the execution of C4.5 calculation and other arrangement methods. Future improvement of this work incorporates improvisation of the C4.5 algorithm to enhance the classification rate to accomplish more prominent accuracy.

Chaurasia and Pal (2017b) in his paper exhibits a diagnosis framework for detecting breast cancer based on RepTree, RBF Network and Simple Logistic. In test arrange, 10-crease cross approval strategy was applied to the University Medical Center, Institute of Oncology and Ljubljana, Yugoslavia database to assess the proposed framework exhibitions. The accuracy of the proposed system is 74.5%. This examination exhibited that the Simple Logistic can be utilized for decreasing the measurement of highlight space and proposed Rep Tree and RBF Network model can be utilized to get quick automatic diagnostic frameworks for other infections.

Comparative Analysis

The data constituted of 32 features which played a role in the prediction of the type ofcancer. Some of these features were irrelevant as they did not provide us withsufficient knowledge in determining tumor in the breast. Data analysis and data visualization through violin plot, swarm plot, heat plot and correlation matrix provide an insight of redundant and irrelevant features. Various visualization plots are provided to select only the important features of the dataset. This gravely improved the accuracy of models (random forest, decision tree and, the logistic regression model) to a great extent. As seen from Fig. 2, a review of Wolberg's *et al.* (1994) research illustrated that there was a maximum accuracy of only 93.6% with 10 cross fold-validation, Chaurasia and Pal (2017a) research had an accuracy of 87.01, Rajesh and Anand (2012) had an accuracy of 92.2%, Williams *et al.* (2015) had an accuracy of 94.2% and Chaurasia and Pal (2017b) obtained accuracy of merely 74.5% whereas our Random forest model resulted in an accuracy of 94.724% due to selection of important and useful features.

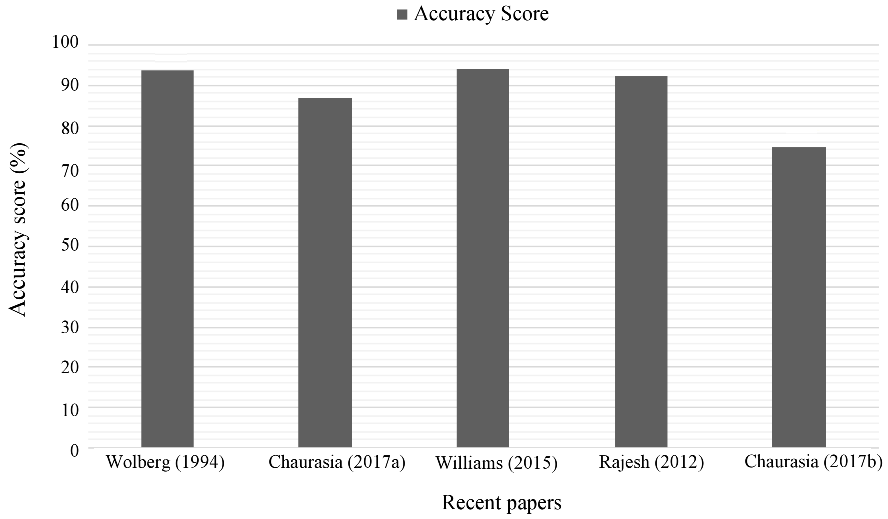


Fig. 2: The accuracies of previous models

S.No	**Attribute**	**Description**
1	ID	Describes the ID no of individual nucleus
2	Diagnosis	B=Benign, M=Malignant
3	Radius	Mean of Distances from center on the perimeter
4	Texture	Standard deviation of gray scale values
5	Perimeter	Perimeter of nucleus cell
6	Area	Area of nucleus instance
7	Smoothness	Local variations in radius lengths
8	Compactness	Measure of density of nucleus
9	Concavity	Severity of concave portions of contour
10	Concave points	Number of concave portions of contour
11	Symmetry	Measure symmetry of nucleus
12	Fractional dimension	Measure of coastline approximation

Fig. 3: Description of attributes of thedataset

The second factor was picking up which features to use in the model while performing classification. Sometimes selecting one feature resulted in the downfall of the accuracy of the decision tree model. Decision tree model when used with predictor's such as radius mean, perimeter mean, area mean, compactness mean, concave points mean provide an accuracy of 100% which is clearly case of overfitting whereas when decision tree was classified withthe only radius mean predictor resulted in an accuracy of 97.236% but with poor 5-fold cross validation model.

Dataset

The data set used in this article was taken from UCI Machine Learning (Dua and KarraTaniskidou, 2017). This dataset has 32 features (attributes) as shown in Fig. 3 and has data objects around 569 data objects where each data object is indicating information about thenucleus.

The dataset consists of following attributes:

['id','diagnosis','radius_mean','texture_mean','perimeter_mean','area_mean','smoothness_mean','compactness_mean','concavity_mean','concavepoints_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','concavepoints_worst','symmetry_worst','fractal_dimension_worst']

Attribute Information: (1) ID number (2) Diagnosis (M = malignant, B = benign) 3-32) Ten real-valued features are computed for each cell nucleus: These attributes are radius which is mean of distances from center to points on the perimeter, texture which is standard deviation of gray-scale values, perimeter,area, smoothness which is local variation in radius lengths, compactness which is perimeter^2/area - 1.0,concavity which is severity of concave portions of the contour, concave points which signifies number of concave portionsofthecontour,symmetry,fractaldimensionThemean,standarderrorand"worst" or largest (mean of the three largest values) of these features were computed for each image,resultingin30features.Forinstance,field3isMeanRadius,field13isRadiusSE and field 23 is Worst Radius. All feature values are recoded with four significant digits. Missing attribute values: None the class distribution contains 357 benign tumors and 212 malignanttumors.

Methodology

The methodology is depicted in Fig. 4. The overall flow proceeds as follows. The data available from the UCI machine repository is taken.This process is termed as the selection of the data. The data selected possess various features (attributes) around 32. All these attributes are not necessary to find out the necessary information from the data. Some of these may be redundant while others can be missing. There can also be noise in the data.

Therefore, the data must be cleaned to process it further. They can be cleaned only after we explore or analyze the data. Theattributes are viewed and data is checked what must be taken care off. Then data preprocessing steps place to remove noise or outliers. As the data contains 32 features, there may be a high probability that only a small fraction of data features might be necessary for mining of the data. Therefore a representation must be shown to the human visual system to show what all features areredundant or not redundant, which of the features can be dropped off for further processing. These stepsare taken in the Data Visualization process where data is visualized and useful information can be taken out of it. Now the data is visualized with the Violin plot, Swarm Plot, Join plot and Heat map. Through this, one can only know which of the features are really important for further processing. Here the necessary features are extracted obtained from the process of Data Visualization. This process is called feature extraction where important features are extracted. Now, the further process that is the classification of data that there are two types of breast cancer tumors benign and Malignant is carried out. The various models are applied to the training and testing data and check which of models is best for getting better accuracy in terms of the model and also cross-validation accuracy.

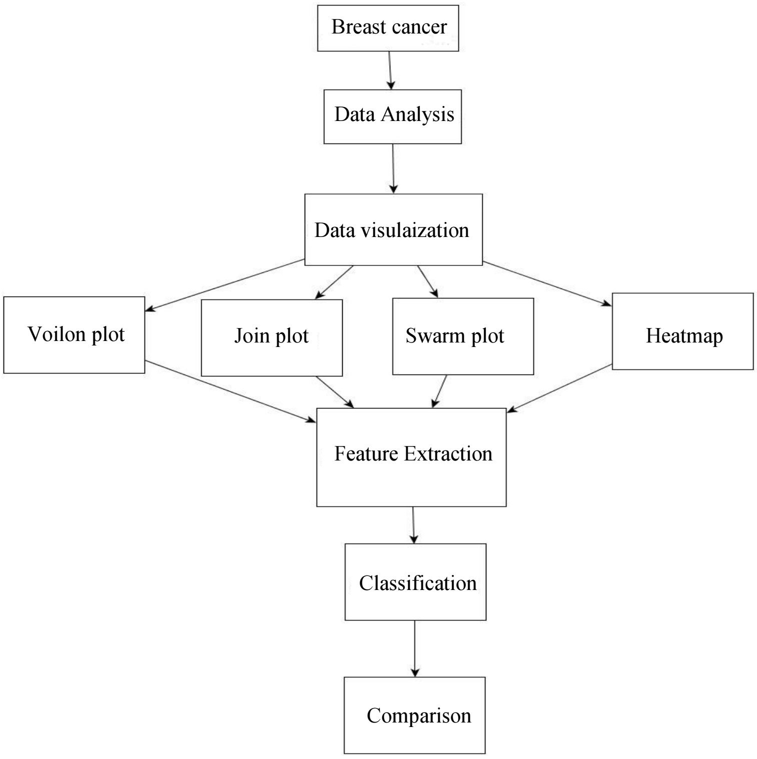
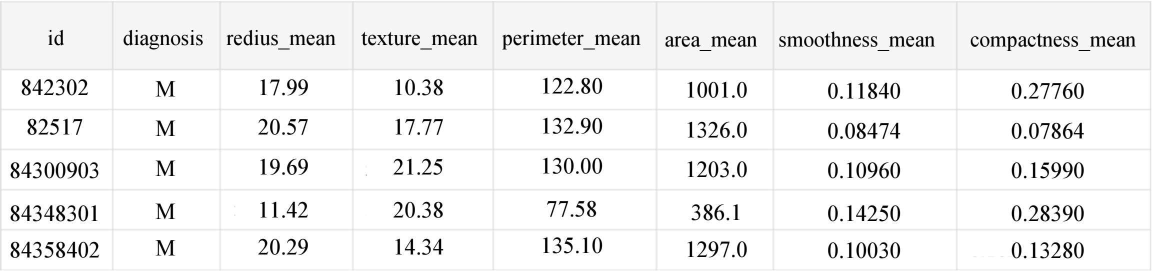


Fig. 4: Overall process flow of breast cancer detection



id	diagnosis	redius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
82517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

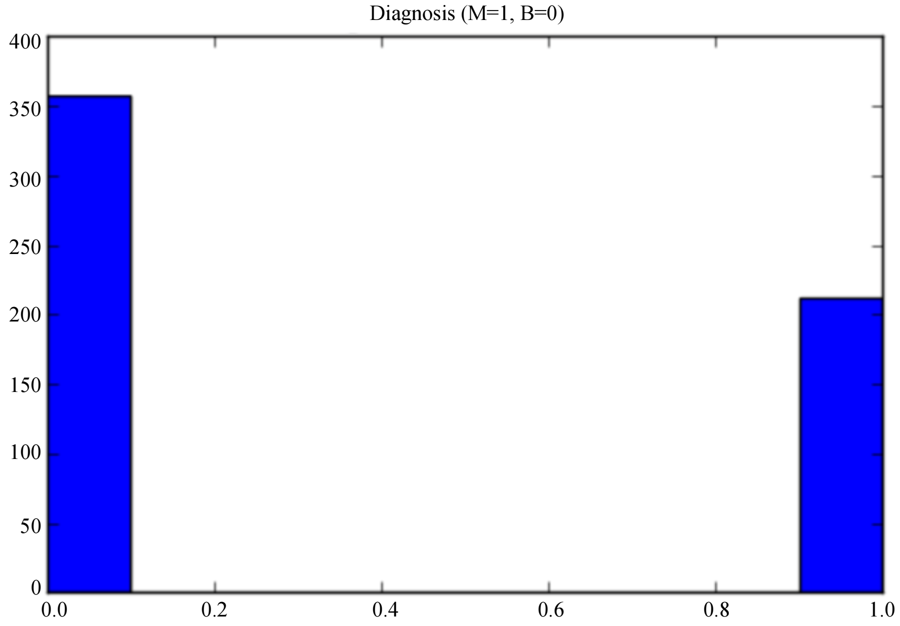


Fig. 5: Count of Malignant and Benign breast cancer tumors

Data Analysis

Before making anything like feature selection, feature extraction and classification, firstly a basic data analysis is carried out to look at the pattern of data. Let's look at features of data.

While doing data analyses, two important observations were made. Firstly, the mean values of cell radius, perimeter, area, compactness, concavity and concave points can be used in the classification of cancer. Larger values of these parameters tend to show a correlation with malignant tumors. Secondly, the mean values of texture, smoothness, symmetry or factual dimension do not show a particular preference of one diagnosis over the other. In any of the histograms, there are no noticeable large outliers that explain further cleanup. From Fig. 5, one can easily infer that the class distribution that is 357 benign tumors and 212 malignant tumors.

Data Visualization

Violin Plot

The blue region on the left part of the vertical line indicates malignant tumor and the right part indicates a benign tumor. Let's interpret the plot as illustrated in Fig. 7. For example, in texture mean feature, the median of the *Malignant* and *Benign* looks like separated so it can be good for classification. However, in thefractal dimension mean feature, median of the *Malignant* and *Benign* does not look like separated so it does not give good information for classification.

Let's interpret one more thing about the plot as shown in Fig. 8, variable of concavity worst and concave point worst looks like similar but how to decide whether they are correlated with each other or not. (Not always true but, basically if the features are correlated with either of it can bedropped).

Join Plot

From Fig. 6, one can easily conclude that concavity worst and point worst predicted from violin plot turn out to be correlated. Hence, one feature is dropped instead of taking both the features. In order to compare two features deeper, let's use joint plot as shown in Fig. 9. Look at this in joint plot below, it is really correlated. Pearson value is correlation value and 1 is the highest. Therefore, 0.86 looks enough to say that they are correlated.

Swarm Plot

Up to this point, some analyses and discoveries on the present data has been made already. In swarm plot as shown in Fig. 10 and 11, a similar analysis to violin plot (Fig. 7 and 8) is carried illustrating the first 10 and last ten features to analyze the features with respect to data.

In this plot (Fig. 10 and. 11), the variance can be seen more clearly. In these two plots which feature looks like more clear in terms of classification. Here from the area worst feature in the last swarm plot looks malignant and benign are separated not totally but mostly.

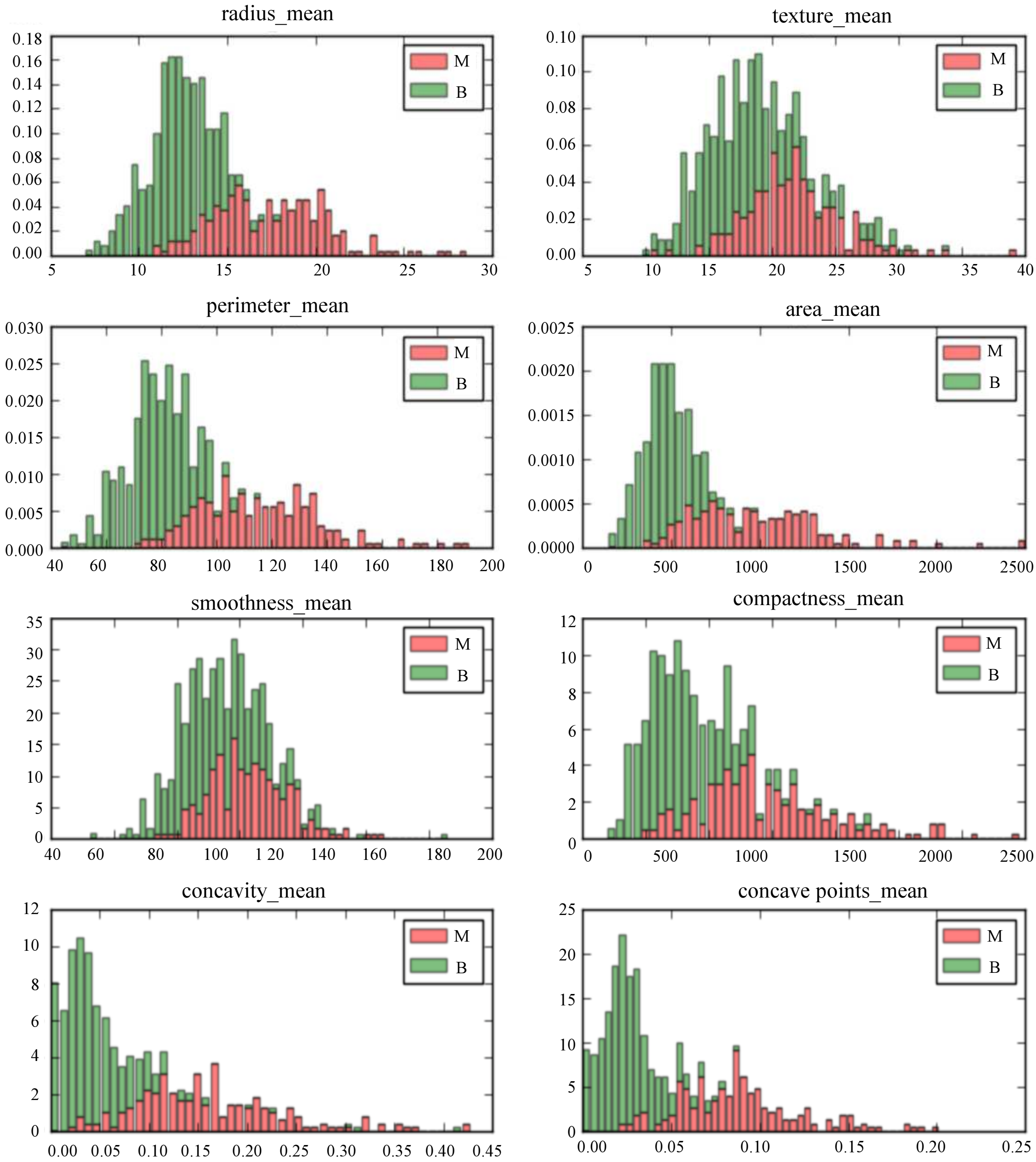


Fig. 6: Nucleus features Vs diagnosis

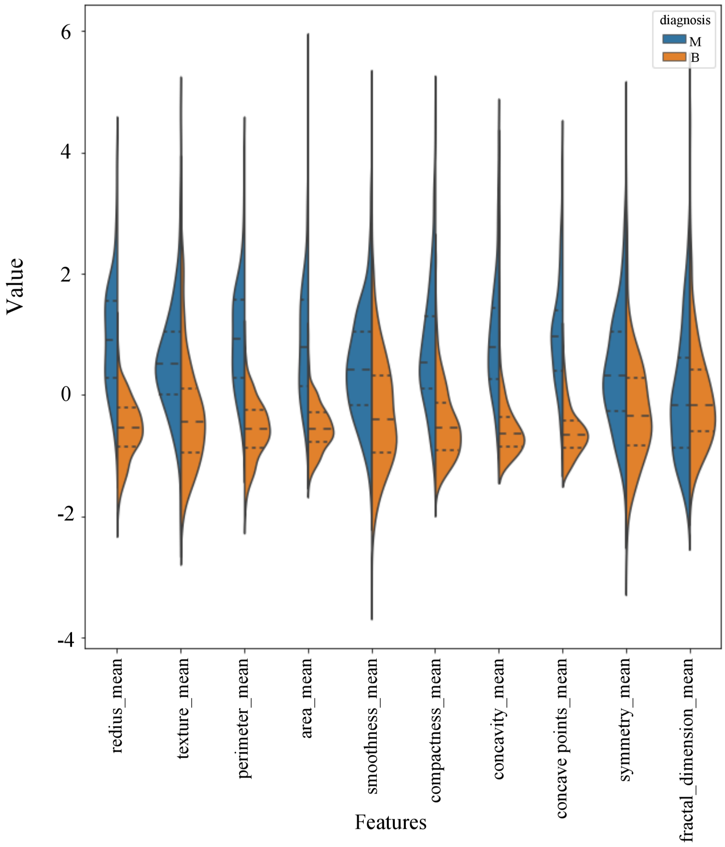


Fig. 7: Violin Plot of first 10 features

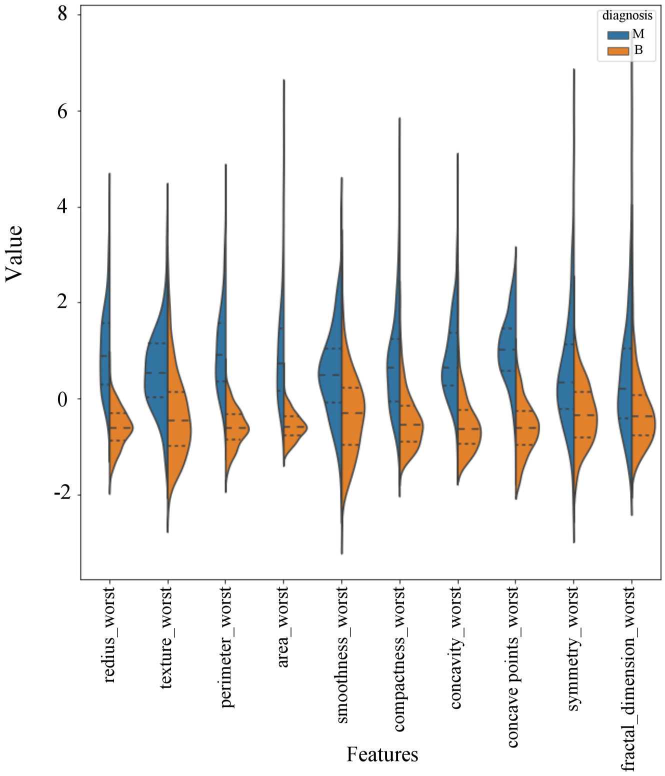


Fig. 8: Violin plot of last 10 features

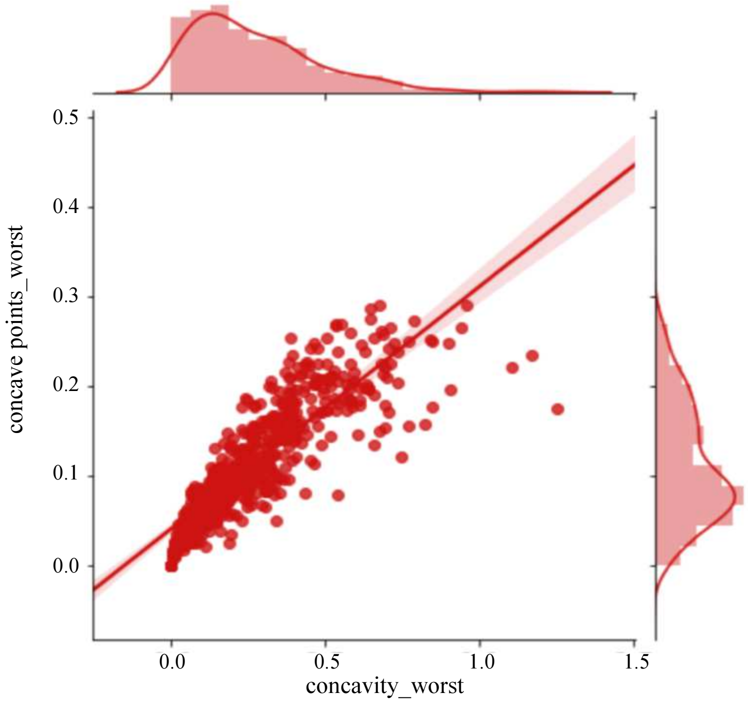


Fig. 9: Correlation graph between concavity worst and concave point worst using the join plot

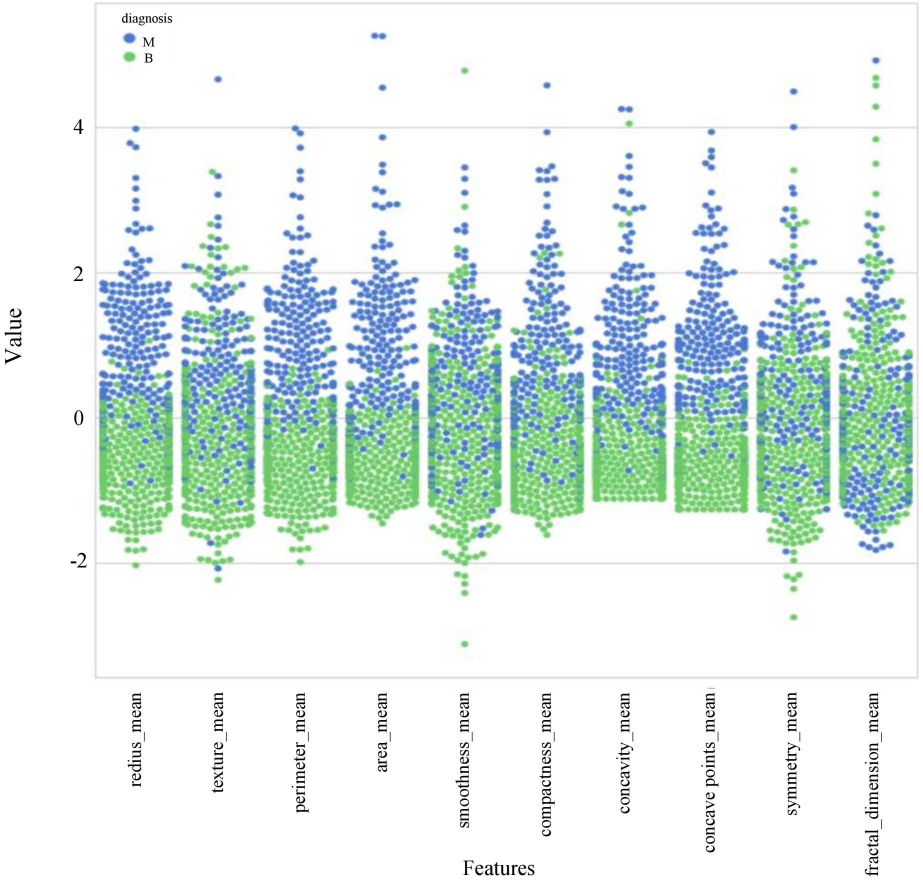


Fig. 10: Swarm Plot of first 10 feature

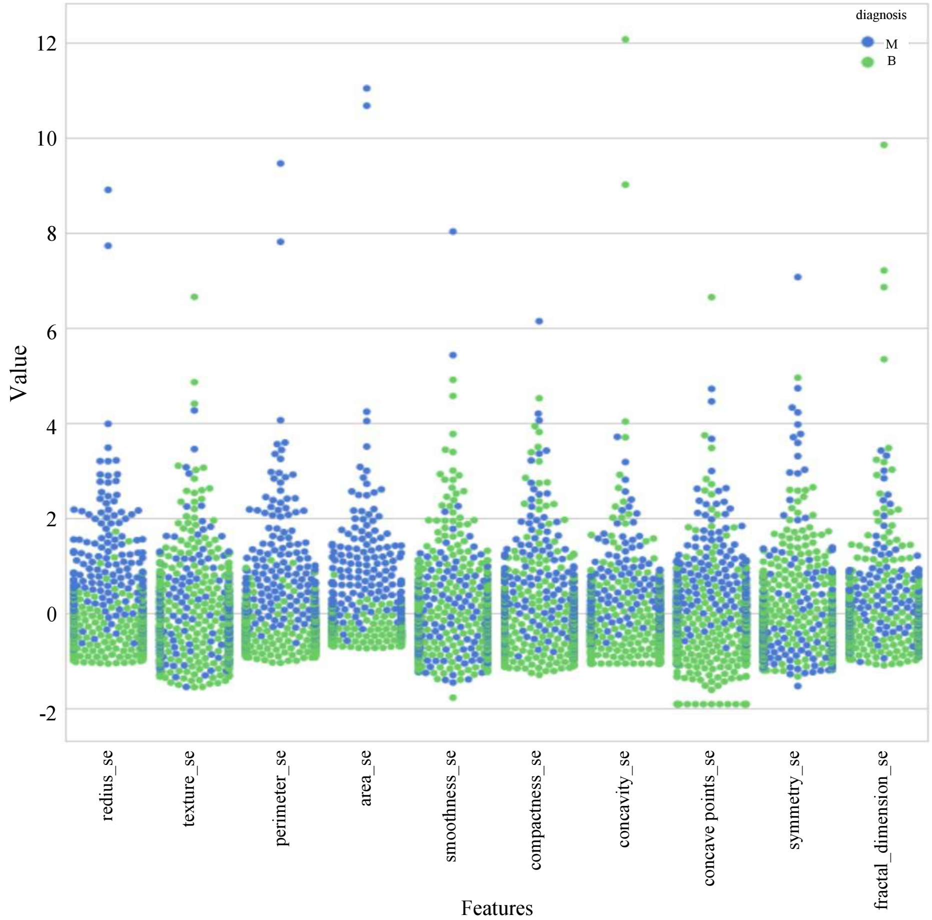


Fig. 11: Swarm Plot of last 10 features

However, smoothness se in swarm plot 2 looks like malignant and benign are mixed so it is hard to classify while using this feature.

Heat MAP

Heat MAP, illustrated in Fig. 12, shows all the correlation between features. Through which all the features which are not relevant are eradicated from this.

Classification

Logistic Regression Model

Logistic regression is widely used for classification of discrete data. In this case, we will use it for binary (1,0) classification.

Logistic regression (Hosmer *et al*., 2000; Han *et al*., 2011) is used to define relation among the variable which is to be predicted $x=(x_1, x_2, x_3, \ldots\ldots, x_p)$ and the response variable. The probability that is conditional probability that a patient has Benign Cancer is written as $P(Y=1|x)=\pi(x)$.

Now:

$$\pi(x) = \frac{e^{\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}}{1 + e^{\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p\right)}} \qquad (1.1.1)$$

where, $0 \leq \pi(x) \leq 1$.

The logit transformation is given by:

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \qquad (1.1.2)$$

where, by the method of maximum likelihood estimation, parameters $\beta = \beta_0, \beta_1, \beta_2, \ldots\ldots \beta_p$ can be found. It finds the value of parameters that maximize likelihood function:

$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \qquad (1.1.3)$$

The logarithm likelihood estimators which is L(ß) is now differentiated with respect to every parameter and put to zero to evaluate its value. There are few other methods such as Newton to find out the same.

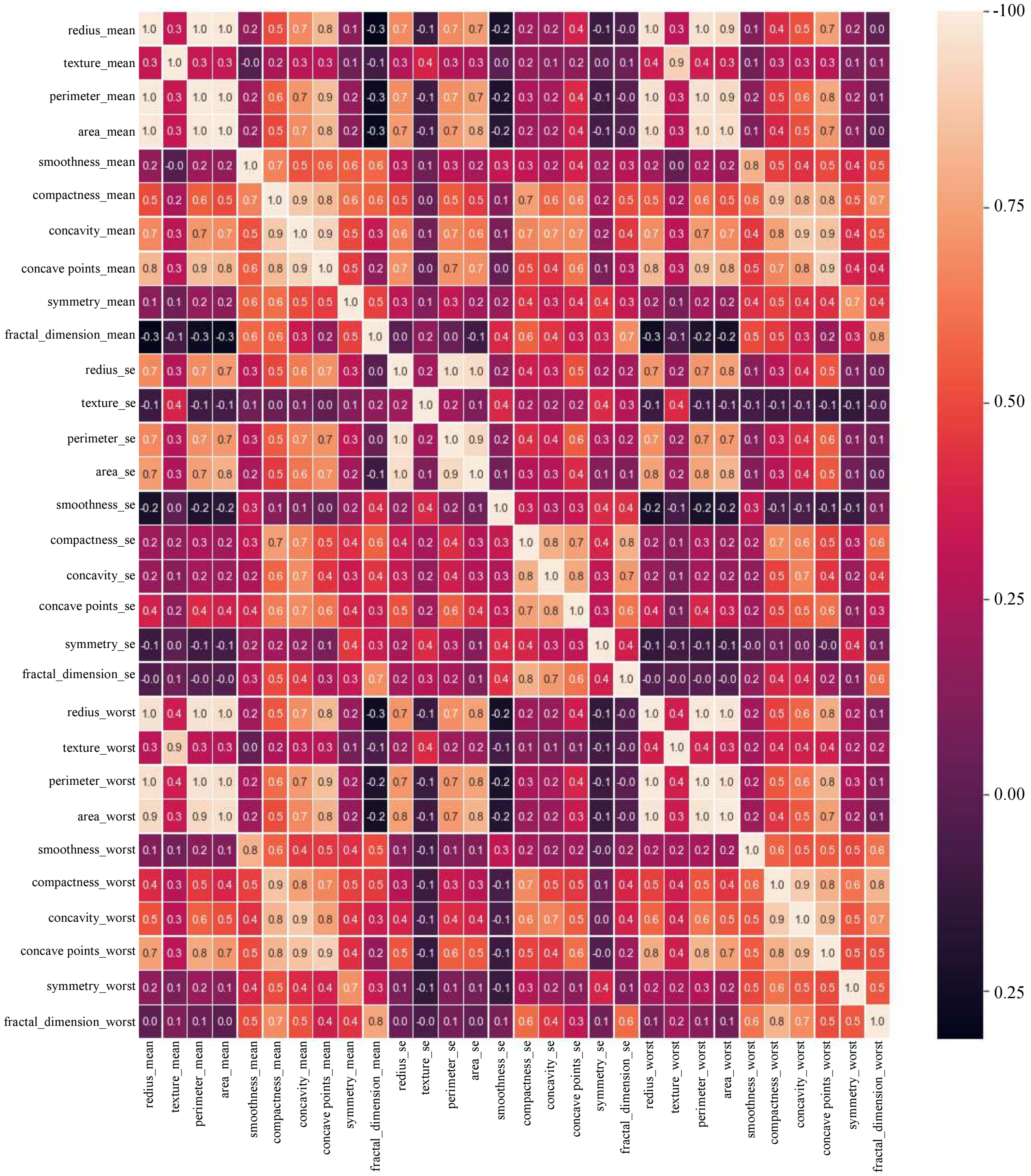


Fig. 12: Heat MAP of breast cancer data

The model is interpreted widely by Odds Ratio (OR), which associates a unit change in x_j represented by $e^{(\beta j)}$:

$$L(\beta) = \ln l(\beta)$$
$$= \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i)\left[1 - \pi(x_i)\right] \right\} \tag{1.1.4}$$

Highly correlated feature leads to unstable parameter estimation hence it is necessary to find unique features and non–redundant features. Correlation is decreased in ordinary least square regression by Ridge regression and is then applied to logistic regression to find it's estimator. β_r that is Ridge estimator is defined by Mangasarian (1990):

$$\beta_r = \left(x'vx + kI\right)^{-1} x'vx\beta_{mle} \qquad (1.1.5)$$

β_{male} => Maximum likelihood estimator
V => Diagonal matrix of maximum likelihood estimator
I => Identity matrix
k => Ridgeconstant.

Based on the observations in the histogram plots, one can reasonably hypothesize that the cancer diagnosis depends on the mean cell radius, mean perimeter, mean area, mean compactness, mean concavity and mean concave points. A logistic regression analysis using those features is performed and the result obtained is as follows.

The accuracy of the predictions are good but not great. The cross-validation scores are reasonable.

Decision Tree Model

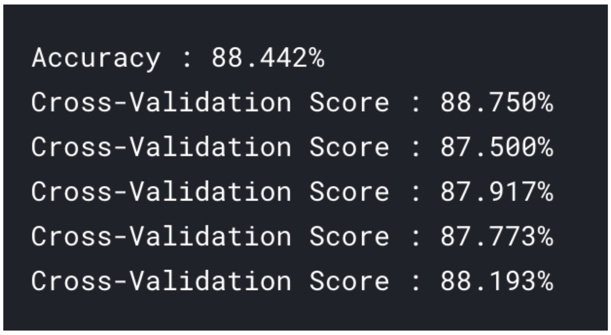
A decision tree (Chaurasia and Pal, 2017a; Quinlan, 1986; Han *et al*., 2011) classifier that is J48 was used here, which is a simplistic learning supervised technique. This Algorithm employs the concept of ID3 to build a decisiontree by a top-down approach which is a greedy search through training data where each and every attribute is tested at each node while building a tree. It uses Information gain which is a measure of the decrease in entropy of attribute after the split of the dataset. Higher the information gain higher the chances of selecting that particular attribute among others. This algorithm is suitable for both categorical and continuous variables of the dataset. A threshold value is fixed such that all values below it are only taken into consideration while building the model. The first and foremost step is to calculate information gain for each and everyattribute:

$$E(T,X) = \sum_{c \in X} P(c)E(c) \qquad (1.2.1)$$

Suppose the dataset consists of T cases, J48 algorithm considers an initial tree and then using the principle of divide and conquer the tree grows givenbelow:

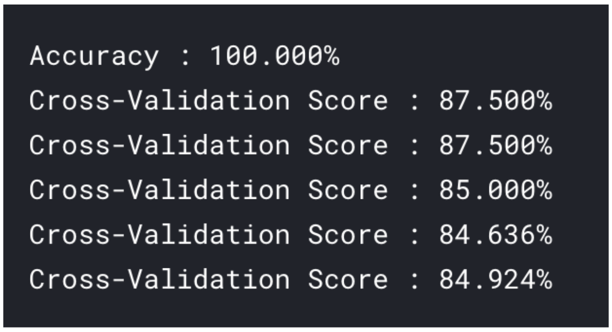
- If all cases in T are in the same class or T is very small the tree is a leaf labeled with most occurring class in S
- Else based on a single attribute with at least 2 outcomes a test case is chosen.This test is the root of the tree with one branch for each outcome of the test. Test T is portioned into corresponding subsets $T_1,T_2,T_3,\ldots,T_n$ for a dataset which contains n cases based on each case outcome and the same procedure is applied recursively to each subset

Here it is clearly the case of over-fitting the model probably due to a large number of predictors. Let use a single predictor, the obvious one is the radius of the cell.



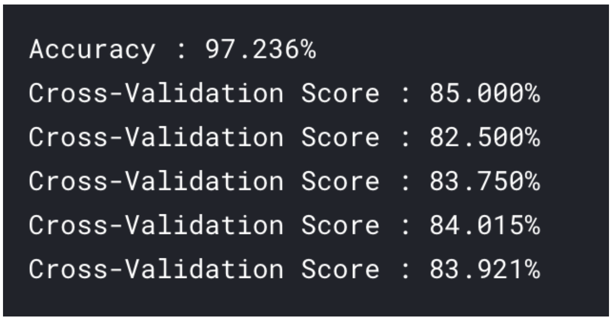
```
Accuracy : 88.442%
Cross-Validation Score : 88.750%
Cross-Validation Score : 87.500%
Cross-Validation Score : 87.917%
Cross-Validation Score : 87.773%
Cross-Validation Score : 88.193%
```

Fig. 13: Accuracy model of the logistic regression model



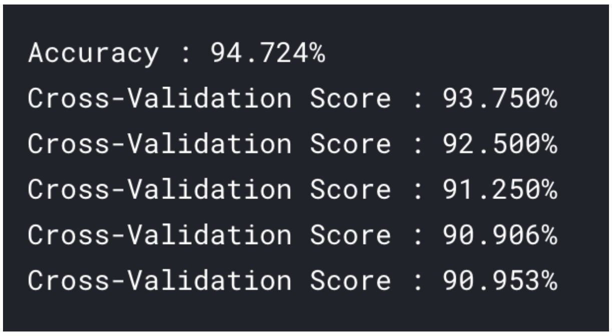
```
Accuracy : 100.000%
Cross-Validation Score : 87.500%
Cross-Validation Score : 87.500%
Cross-Validation Score : 85.000%
Cross-Validation Score : 84.636%
Cross-Validation Score : 84.924%
```

Fig. 14: Accuracy model of the decision tree model



```
Accuracy : 97.236%
Cross-Validation Score : 85.000%
Cross-Validation Score : 82.500%
Cross-Validation Score : 83.750%
Cross-Validation Score : 84.015%
Cross-Validation Score : 83.921%
```

Fig. 15: Accuracy model of decision tree model using single



```
Accuracy : 94.724%
Cross-Validation Score : 93.750%
Cross-Validation Score : 92.500%
Cross-Validation Score : 91.250%
Cross-Validation Score : 90.906%
Cross-Validation Score : 90.953%
```

Fig. 16: Accuracy model of random forest model

The accuracy of the prediction is much better here using a single predictor gives a 97% depicted in Fig. 15 prediction accuracy for this model but the cross-validation score is not that great.

Random Forest

Random forest (Han *et al*., 2011), an ensemble of decision tress, classifier creates a set of decision trees from a randomly selected subset of the training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

Using all the features improves the prediction accuracy and the cross-validation score is great as shown in Fig. 16. An advantage with Random Forest is that it returns a feature importance matrix which can be used to select features.

Result Analysis

Here different models were used, trained and tested. The accuracy was predicted for all these models and k (5) fold cross-validation was performed on the training data to improve and correctly predict the accuracy. Out of all the results, random forest came out to be the best model in terms of classification with an accuracy score of 94.724%. Figure 17 depicts that random forest turns out to be the most accurate model in terms of accuracy and 5-fold cross validation score with an overall accuracy of 94.724. Although the decision tree model was more accurate-100% (Fig. 14) in contrast to logistic regression model-88.42% (Fig. 13), 5-fold cross-validation accuracy turns out to be opposite. Decision tree with single predictor having an accuracy of 97.236% (Fig. 15), has the least 5-fold cross-validation accuracy as depicted in Table 1 fall behind Logistic regression model in terms of accuracy.

Table 1: Comparative analysis of different models

Model	Accuracy	Cross Validation (1)	Cross Validation (2)	Cross Validation (3)	Cross Validation (4)	Cross Validation (5)
Logistic Regression	88.42	88.750	88.500	87.917	87.773	88.193
Decision tree (overfitting)	100	87.500	87.500	85.000	84.636	84.924
Decision tree using radius Predictor	97.236	85.000	82.500	83.750	84.015	83.921
Random Forest	94.724	93.750	92.500	91.250	90.906	90.953

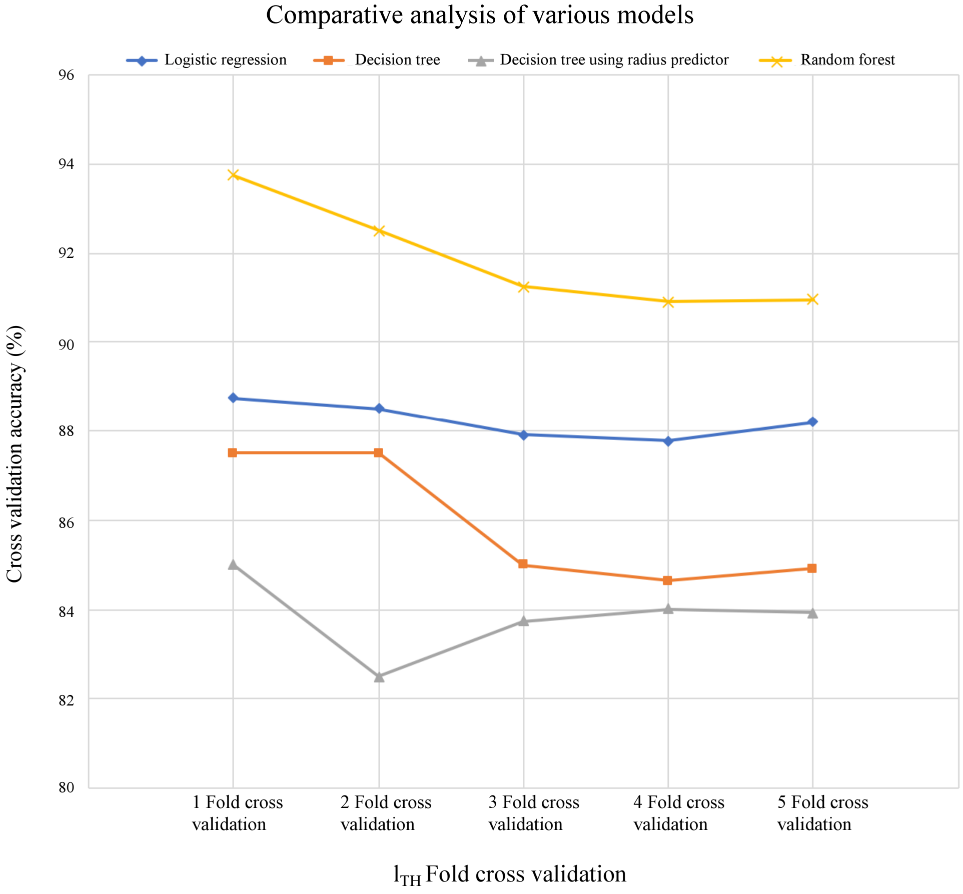


Fig. 17: Measure of cross-validation accuracy between various models

Conclusion

Breast cancer detection with data visualization can be utilized to eliminate some features and to find out which of the features were important like join plot swarm lot, heat map and grid plot. An analysis of all the models was done considering accuracy and cross-validation as parameters (*k*=5).

The best model to be used for diagnosing breast cancer as found in the analyses is the Random forest Model with top 5 predictors 'Concave points_mean','area_mean','perimeter_mean'. It gives a prediction accuracy of ~95% and a cross-validation score of ~93% for the test data as shown in Fig. 16. Other models such as the decision tree model also gave a reasonable amount of good accuracy around ~100% (Fig. 14). However, this model lagged in the cross-validation score very much behind other models and had the case of overfitting. Therefore, one can easily conclude that Random forest classifier is the best among all these models in terms of accuracy for this data set in effectively predicting breast cancer with careful feature selection through data visualization.

Acknowledgment

The authors would like to thank the reviewers for their valuable comments and suggestions that contributed to the improvement of this work.

Author's Contributions

All authors equally contributed in this work.

Ethics

This article is original and contains unpublished material. The authors confirm that is no conflict of interest involved.

References

Chaurasia, V. and S. Pal, 2017a. Performance analysis of data mining algorithms for diagnosis and prediction of heart and breast cancer disease. Rev. Res., 3: 1-13.

Chaurasia, V. and S. Pal, 2017b. Data mining techniques: To predict and resolve breast cancer survivability.

Dua, D. and E. Karra Taniskidou, 2017. UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA.

Han, J., J. Pei, and M. Kamber, 2011. Data mining: Concepts and techniques. Elsevier.

Hosmer, D., S. Lemeshow and R.X. Sturdivant, 2000. Applied Logistic Regression. 1st Edn., A Wiley-Interscience Publication, New York.

Mangasarian, O.L. and W.H. Wolberg, 1990. Cancer diagnosis via linear programming. SIAM News, 23: 1-18.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn., 1: 81-106. DOI: 10.1023/A:1022643204877

Rajesh, K. and S. Anand, 2012. Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. Int. J. Adv. Res. Comput. Commun. Eng., 1: 72-77.

Williams, K., P.A. Idowu, J.A. Balogun and A.I. Oluwaranti, 2015. Breast cancer risk prediction using data mining classification techniques. Trans. Netw. Commun. DOI: 10.14738/tnc.32.662

Wolberg, W.H., W.N. Street and O.L. Mangasarian, 1994. Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett., 77: 163-71. DOI: 10.1016/0304-3835(94)90099-x