CSE 422 (Artificial Intelligence)

Lab Project Report

Group: 4

1) Nafisa Khan Youkee (ID: 21301499)
2) Sazzad Hossen Himel (ID: 21301066)
3) Azwad Aziz (ID: 21301027)

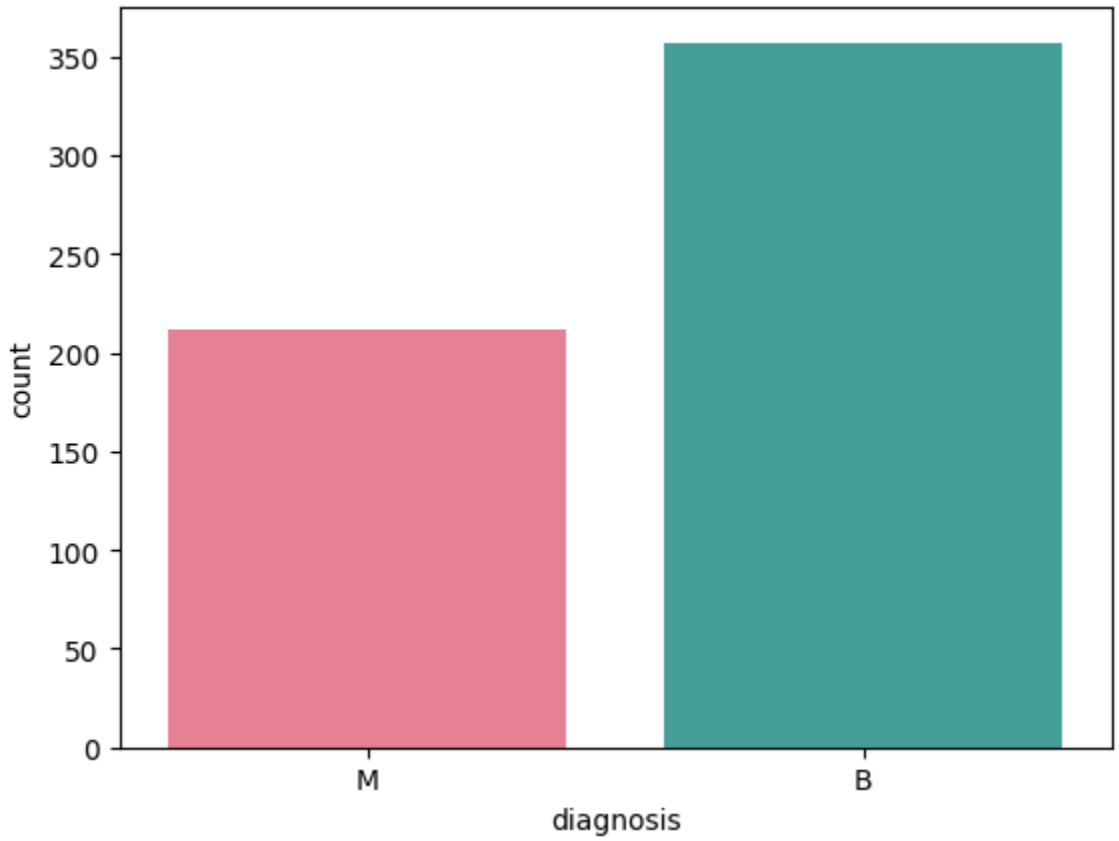
Table of Contents:

1. Introduction

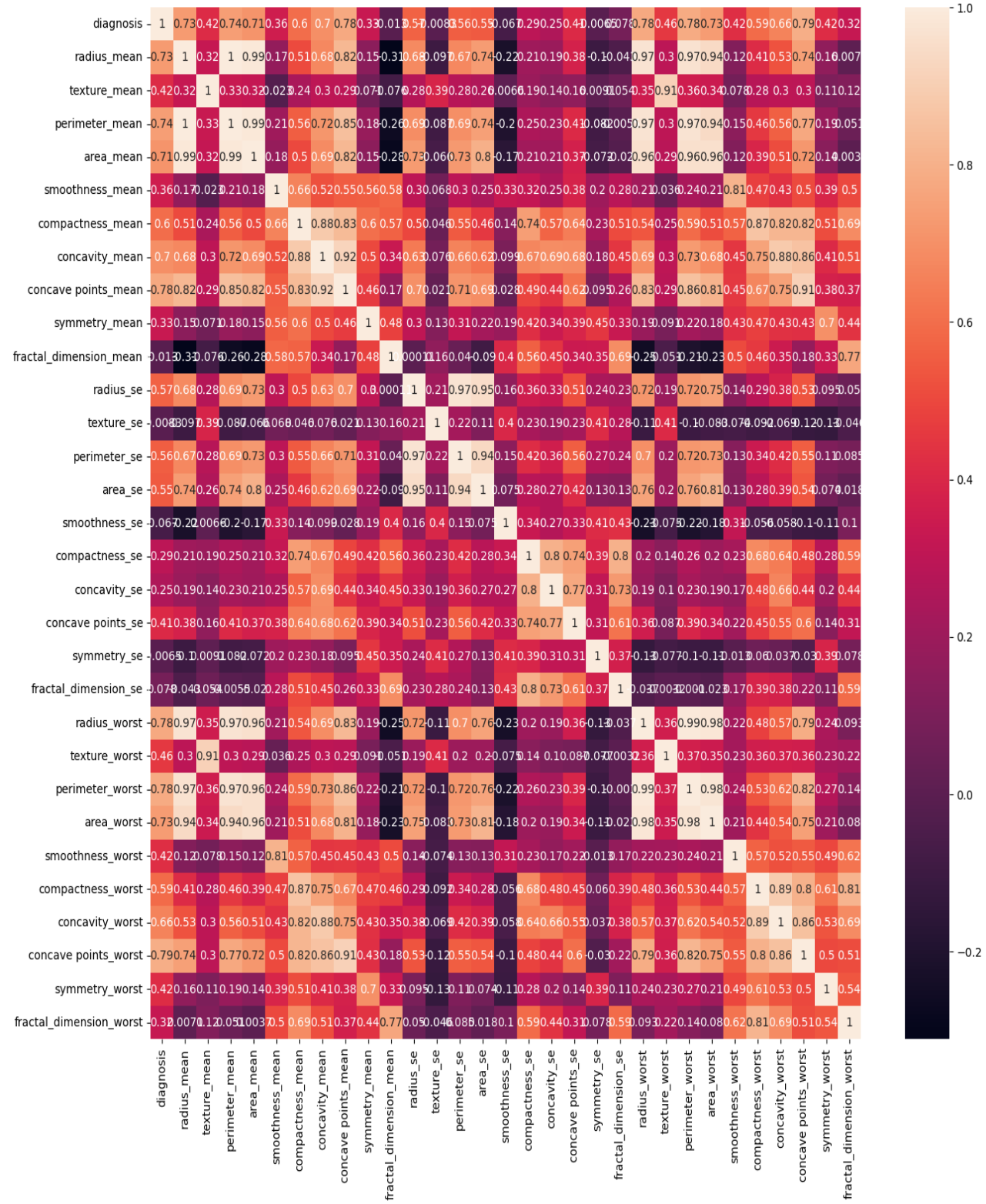
Breast Cancer is a common illness in the public health domain which continues to pose a significant health threat among women worldwide. The death rate of breast cancer patients is increasing every year. However, if it is diagnosed at a very early stage then the possibility of cure increases significantly. Machine Learning models are used effectively to classify cancer tumors by observing its input features like: radius, perimetre, area and so on. In this project, we have implemented Logistic Regression, Decision Tree and Random Forest Classifier models to classify between malignant and benign tumors. Firstly we preprocessed our selected dataset as required. After which, aforementioned models were implemented on the preprocessed dataset. Lastly, after evaluating the performance of these models, we observed the highest accuracy from Random Forest Classifier whereas Decision Tree generated the least accurate results.

2. Dataset Description:

The dataset used for this machine learning project is Breast Cancer Wisconsin (Diagnostic) Data Set from Kaggle. It contains 30 features and all of them are quantitative. These were computed from a digitized image of a breast mass and the target is to classify the tumor as Malignant or Benign which is a classification problem. It contains a total of 569 data points (rows) among which 357 are Benign and 212 are Malignant. The distribution of classes is shown in the bar chart which shows the imbalance. Also the heatmap from the correlation of a feature is given in the next page.



Heat map


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
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

3. Data Preprocessing

A number of techniques were applied for preprocessing to prepare the dataset perfectly before it is to be used to train and model and for testing:

Each preprocessing technique that has been applied is described as follows:

A) Deleting null values and unnecessary columns:

The "Unnamed:32" is a column in the dataset that entirely consists of Null values, hence it was dropped. Also, the column "ID" for the patients is not necessary for training the model and it was also removed.

B) Encoding the output label:

The output label ("diagnosis") contains two classes : Malignant and Benign and they were encoded as 1 and 0 respectively using mapping.

4. Feature Scaling

The input features are all quantitative and there is high skewness. So, to retain the correlation and also to reduce the skewness we applied a Standard Scaler so that there is no high skewness towards any particular feature.

5. Dataset Splitting

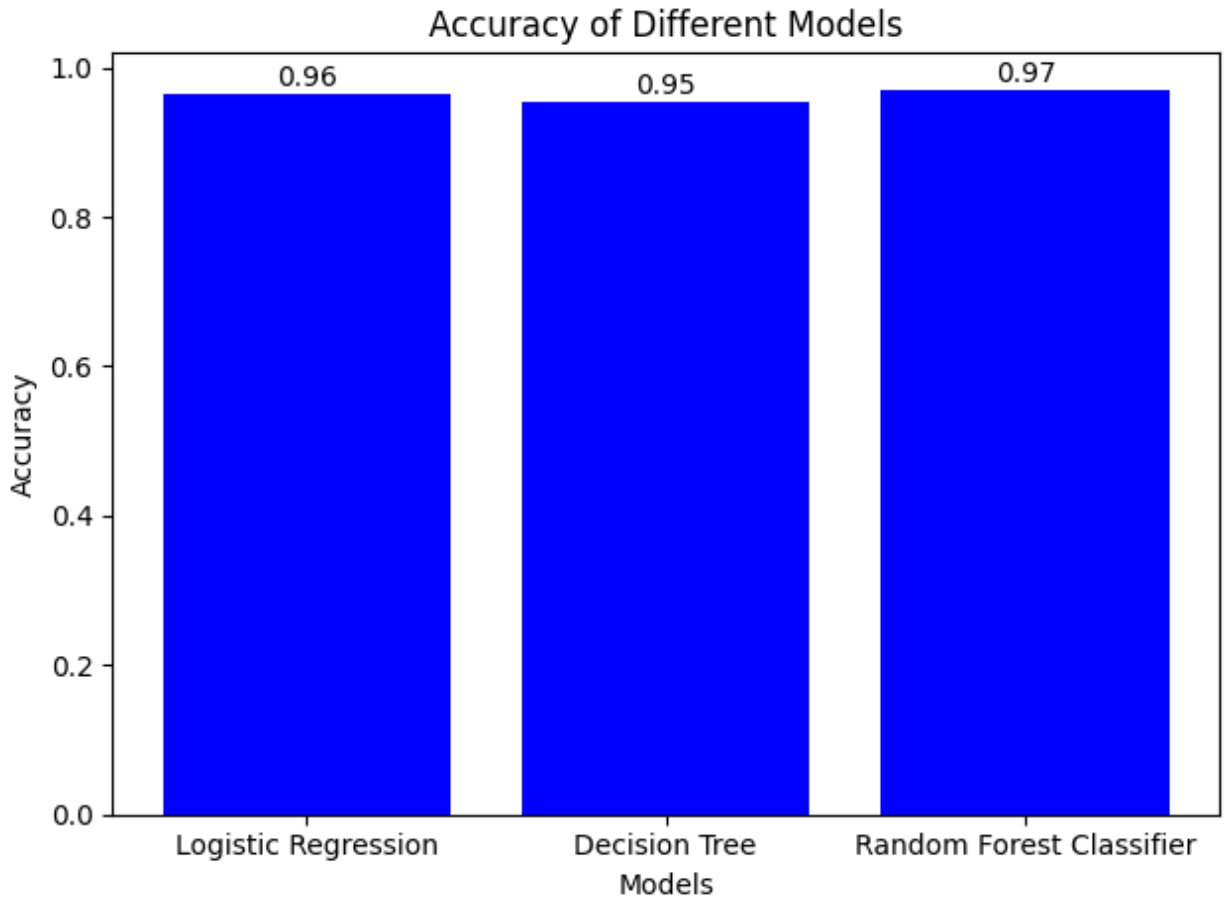
The dataset is split into two parts for training and testing. The training part contains 70% of the dataset whereas the testing part contains 30% of the dataset. Larger portion was

used for training since it was more necessary to optimize the parameters so that the model makes better predictions.

6. Model training and testing

- Logistic Regression
- Decision Tree
- Random Forest Classifier

7. Comparison Analysis:



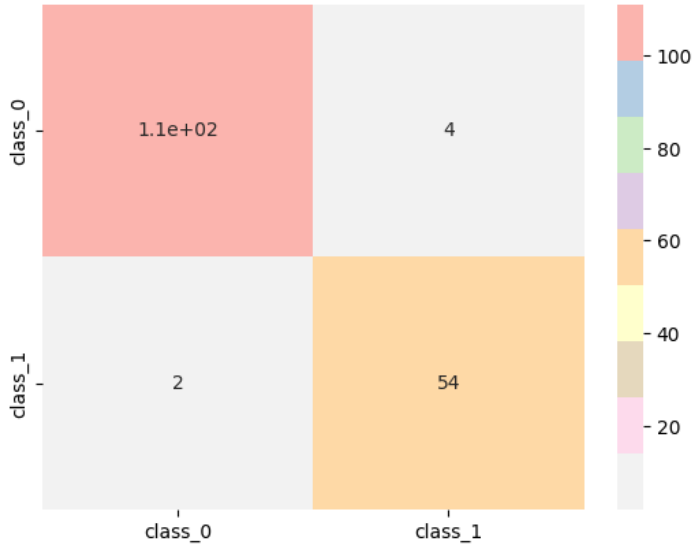
From the above bar chart, we can see that the highest accuracy is 97% which is obtained from the Random Forest classifier. On the other hand, logistic regression and decision tree shows 96% and 95% accuracy respectively.

Precision and Recall for each model:

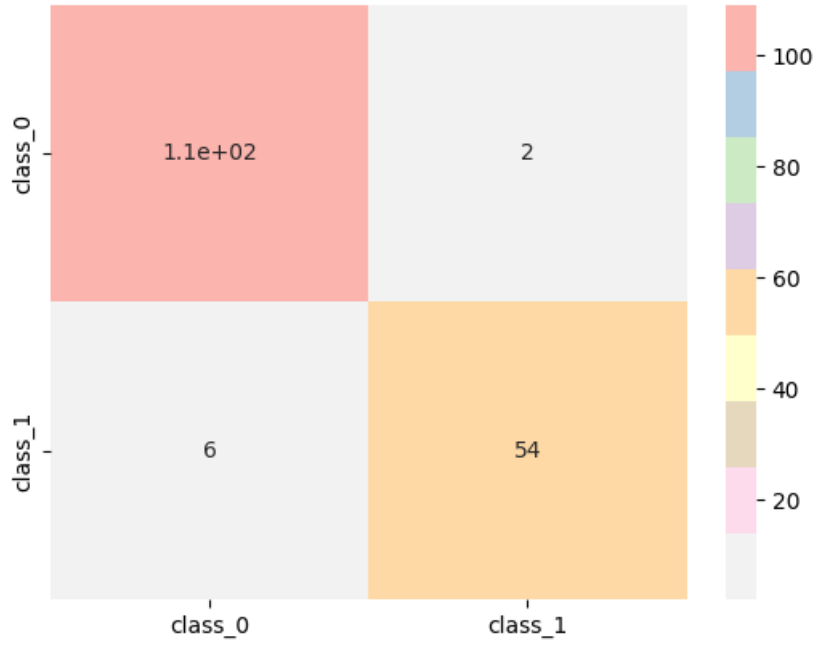
Model name	**Precision**	**Recall**
Logistic regression	97%	96%
Decision Tree	96%	95%
Random Forest Classifier	97%	97%

Confusion Matrix

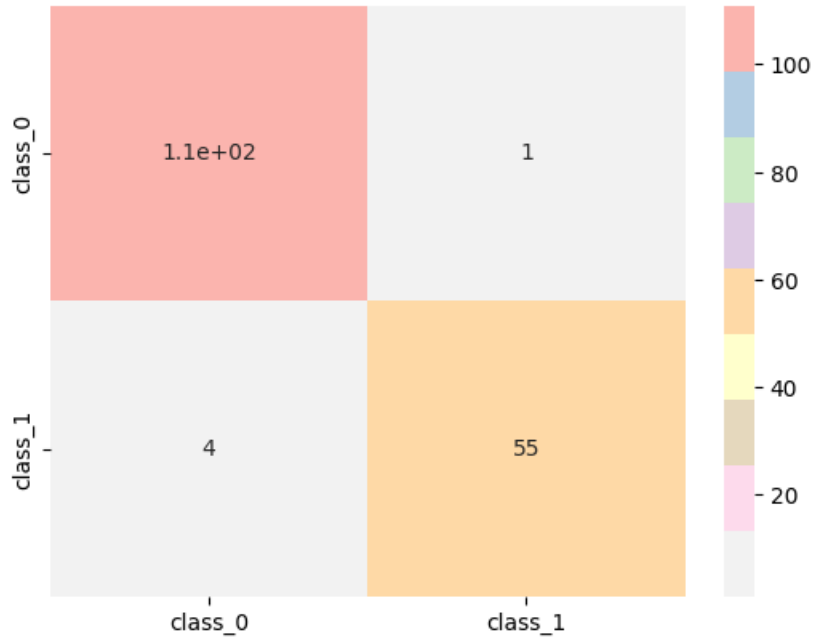
Logistic regression



Decision Tree



Random Forest Classifier



8. Conclusion

Overall, we can see that the task of breast classification can be achieved with great accuracy using the Machine Learning models. Starting from preprocessing, applying 3 different models and using various metrics for comparison, we were able to further distinguish which model performs better than others in this task. Random Forest gives the greatest accuracy since it is an ensemble model of Decision tree, hence it provides better accuracy. To sum up, we can conclude that use of machine learning models at a large scale can be very useful to delve further into the field of breast cancer classification to achieve a significant success for solving this problem.