Final Report: Breast Cancer Prediction Analysis

1. Introduction

Background

Breast cancer is an increasingly common and dangerous disease for women that originates in the breast tissue. It is the second most prevalent type of cancer and nearly 12% of women worldwide are affected by the disease. Early detection is vital for successful treatment of the disease and positive prognosis. While advancements have been made in diagnostic techniques over the years, the high volume of data collected from them remains difficult to parse. Machine learning algorithms can help improve breast cancer diagnosis by analyzing the data for relevant trends and the most important factors.

Problem Statement

Can we classify tumors as benign or malignant with a minimum accuracy of 80% based on a set of nine features in the dataset?

References:

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8156889/
2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/

2. Wisconsin Breast Cancer Dataset (1992)

The dataset obtained from OpenML.org contains 39366 entries with the following fields:

1. **id** - Patient ID
2. **Clump_Thickness** - Indicates grouping of cancer cells in a multilayer.
3. **Cell_Size_Uniformity** - Indicates metastasis to lymph nodes.
4. **Cell_Shape_Uniformity** - Identifies cancerous cells of varying size.
5. **Marginal_Adhesion** - Quantifies loss of adhesion in cells.
6. **Single_Epi_Cell_Size** - Quantifies the size of the epithelial cells.
7. **Bare_Nuclei** - Quantifies the presence of bare nuclei in the cells.
8. **Bland_Chromatin** - Quantifies the presence of bland chromatin in the cells.
9. **Normal Nucleoli** - Quantifies the presence of normal nucleoli in the cells.
10. **Mitoses** - Quantifies the stage of Mitoses in the cells.
11. **Class** - The target variable that qualifies tumors as malignant (1) or benign (0)

3. Data Cleaning and Wrangling

The dataset is relatively polished with no missing values because the data is reported from clinical cases. Additionally, the data types for each feature are consistent with the information that each feature provides.

Invalid Data Detection and Removal

The literature for the dataset states the range of the nine features to be between 1 and 10. Therefore any values for those fields less than 1 or greater than 10 must be considered invalid. Upon further analysis, there were ~50 total records that had values less than 1 for 'Clump_Thickness', 'Cell_Size_Uniformity', 'Bare_Nuclei', or 'Normal_Nuclei'. Since these entries represent a very small and insignificant portion of the dataset, I could remove them without affecting the results.

Contrary to that, there is a significant subset of the dataset with values greater than 10 representing each attribute. Removing all such records from the dataset would remove close to 5,000 entries or about 13% of the data. Removing such a quantity of data would undoubtedly compromise the integrity and accuracy of the results. Hence, I decided to include those entries in further analysis.

4. Exploratory Data Analysis

Preliminary Look

The data is imbalanced with respect to the class labels as seen in Figure 1 since there are twice as many benign cases as malignant. The heavy skew in the class labels can lead to a bias in the trained classification models, therefore, the issue must be addressed during modeling. Additionally, the data distributions depicted in Figure 2 for each feature reveal that all attributes except 'Clump_Thickness' have a heavy data skew to the right. 'Clump_Thickness' is evenly distributed to some extent while 'Bare_Nuclei' has a second small peak at 10 which could be explored further.

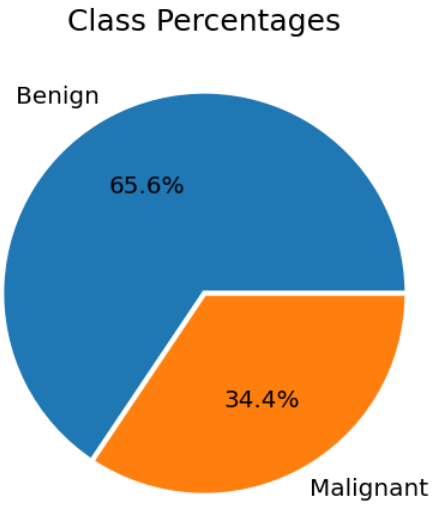

Figure 1. Pie chart depicting the percentage of records belonging to benign or malignant label

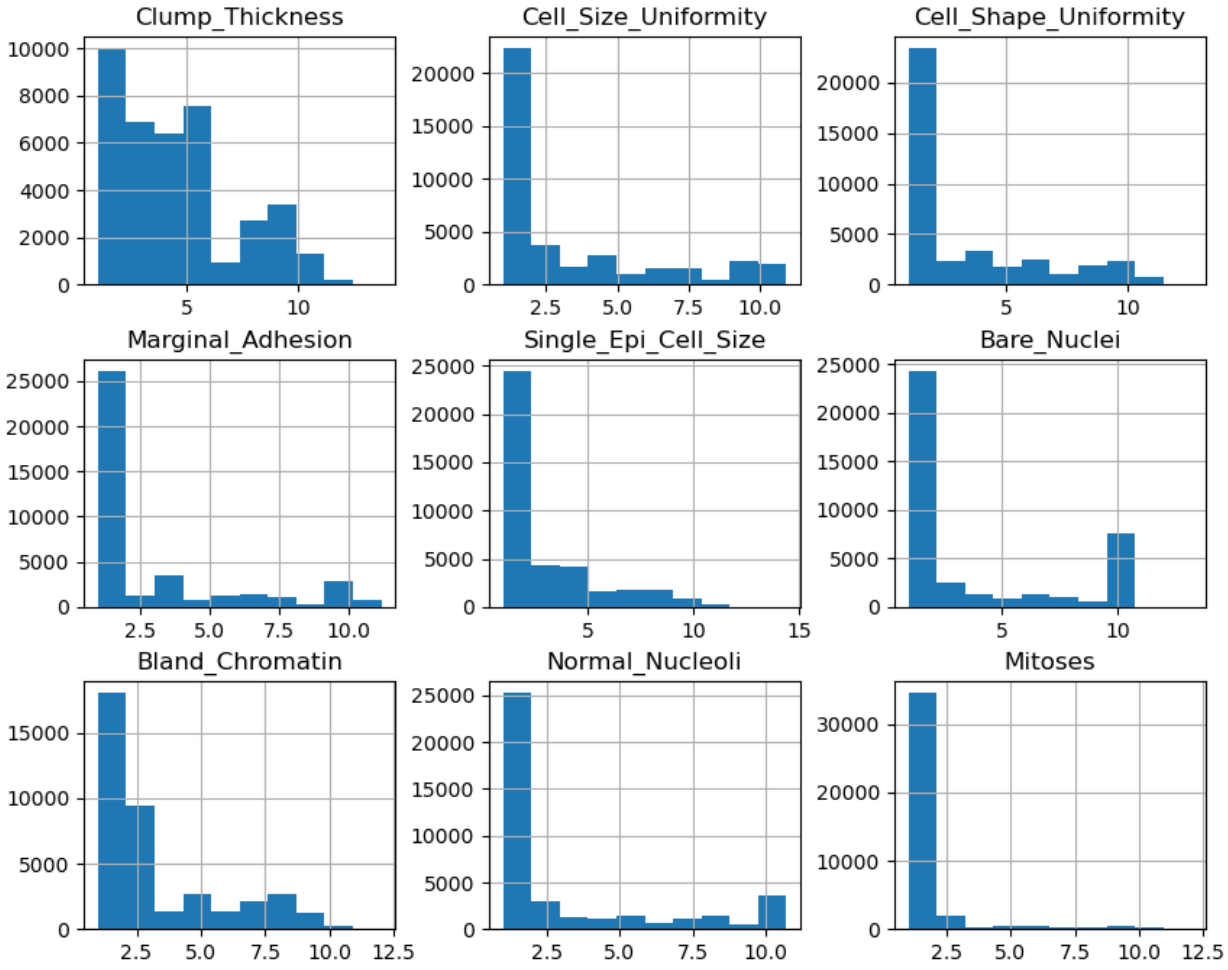

Figure 2. Data distribution for each major feature.

Features and their relationship to the Class Labels

In order to better understand the features and their relationships to the class labels, I plotted a probability distribution for each feature to analyze its skewness, which remained consistent with the results in Figure 2. I also plotted a box plot for each feature with respect to each class label to visualize the mean and the quantiles for each group. In these plots, the difference in the means and quantiles for malignant and benign labels for each feature are significant visually. Lastly, I plotted a bar chart for each feature grouped by the class label to compare the data distribution between each class.

In Figures 3-5, bar charts are shown for 'Mitoses', 'Normal_Nucleoli', and 'Bland_Chromatin'. These features were deemed to be the most important by multiple classifiers hence their plots are included in the report. One trend consistent among all three plots is the distribution of benign labels is heavily skewed to the right with very few data points in the higher value bins on the x-axis. In each of the plots, there are consistently more data points with malignant labels in the 4+ value bins for each feature. These trends point to a clear distinction in the dataset between the malignant and benign records.

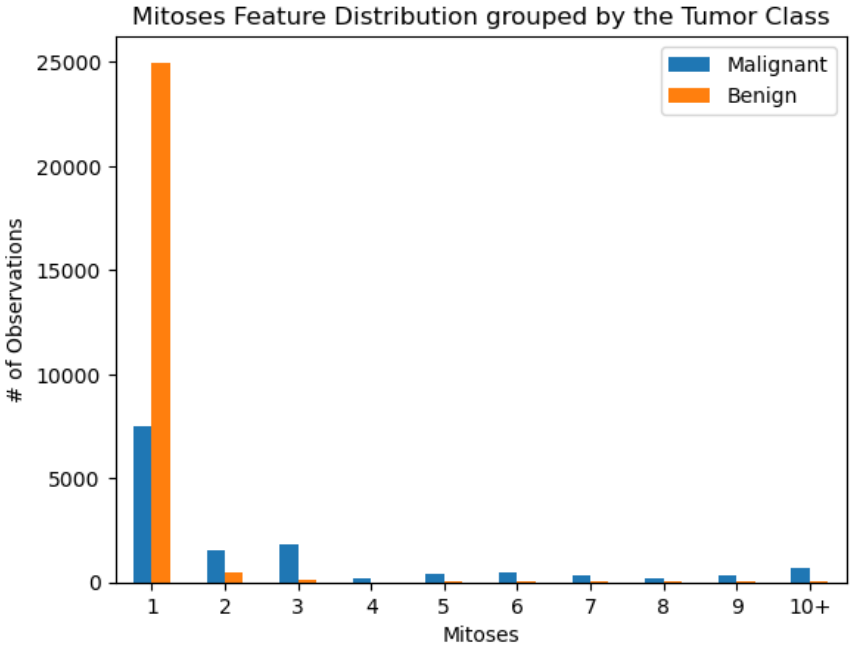

Figure 3. Bar chart depicting the data distribution of the 'Mitoses' feature grouped by the class labels

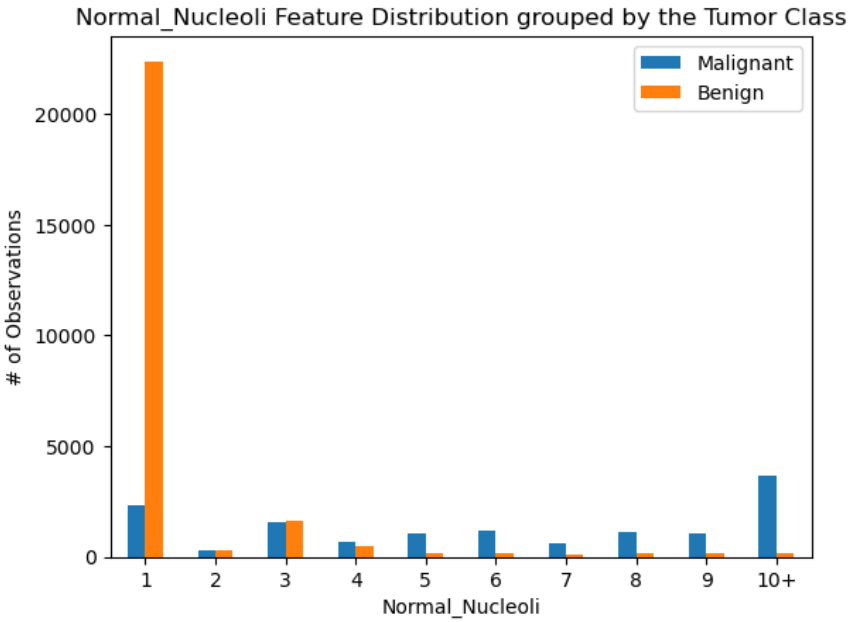

Figure 4. Bar chart depicting the data distribution of the 'Normal_Nucleoli' feature grouped by the class labels

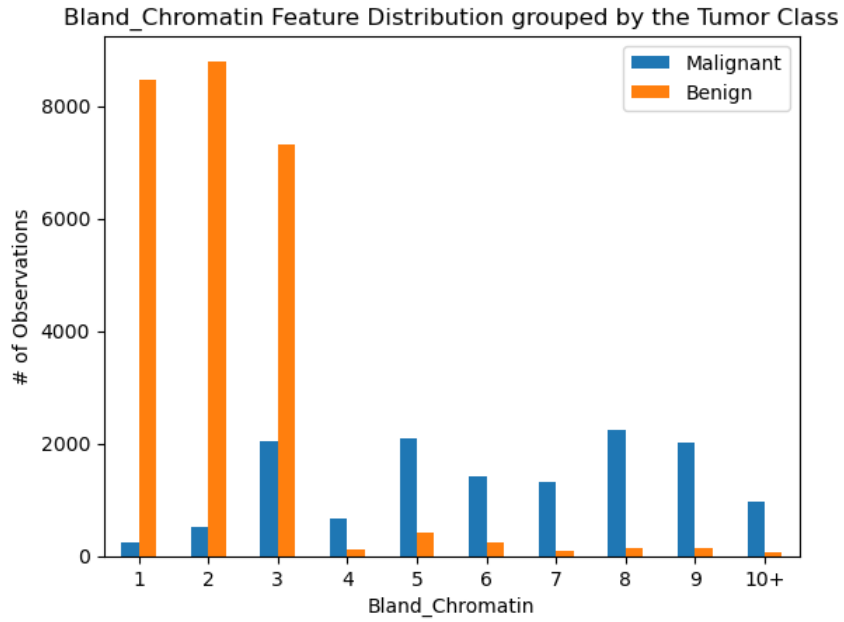

Figure 5. Bar chart depicting the data distribution of the 'Bland Chromatin' feature grouped by the class labels

Relationship between the features

In order to examine the relationship between the features, I first created a pair plot. However, it didn't yield any useful information as there were no noticeable visual patterns. Furthermore, the high volume of data masked any potentially useful trends in the scatterplots. I then created a heatmap of the correlation between the features which can be viewed in figure 6.

The heatmap shows that ‘Mitoses’ is not strongly correlated with any other feature in the dataset. Additionally, ‘Cell_Size_Uniformity’ has a strong correlation with most of the other features, and the strongest correlation with ‘Cell_Shape_Uniformity’. These observations are important to note while creating an appropriate model for the data set.

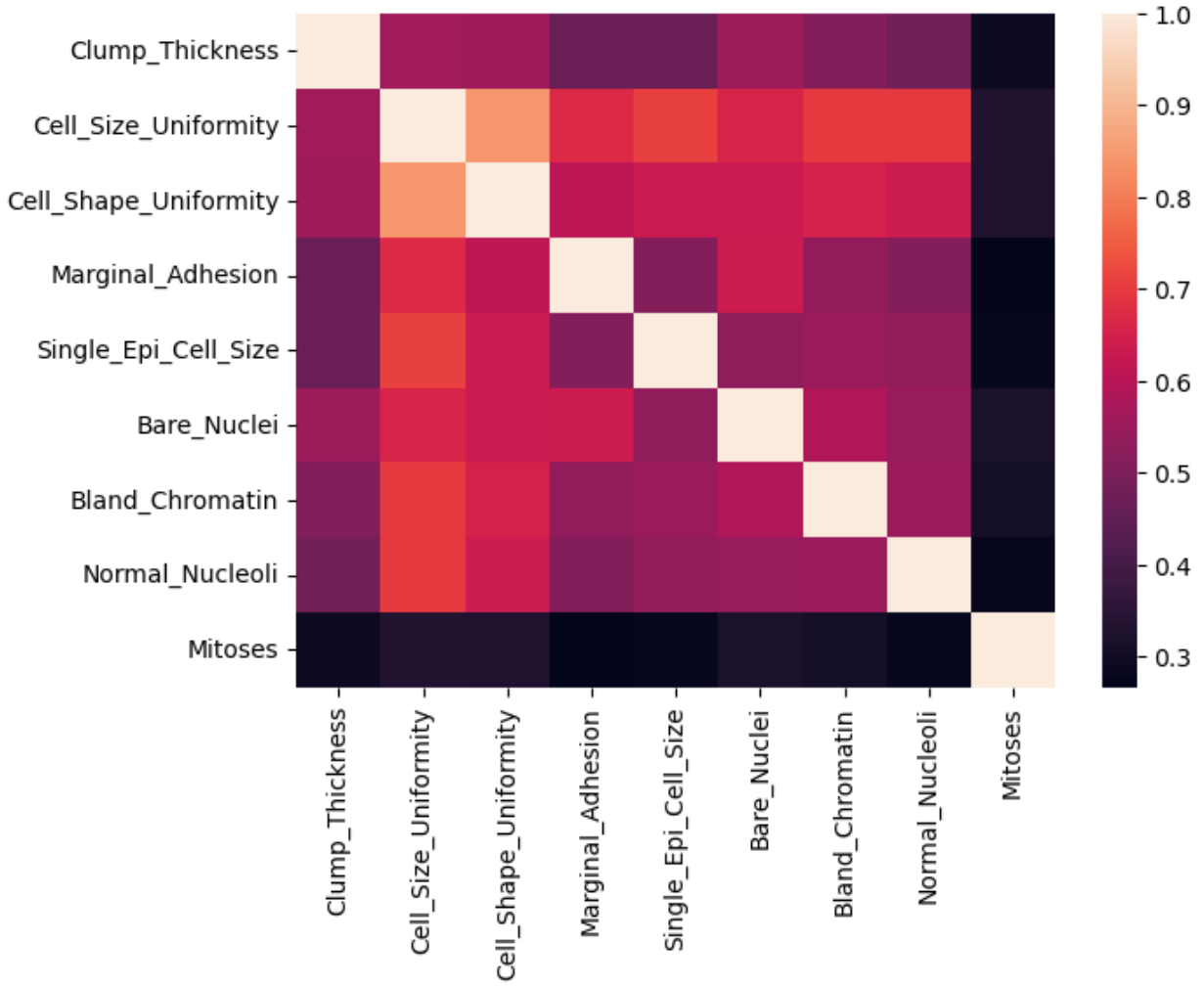

Figure 6. Heatmap depicting the correlation between each feature in the dataset.

5. Modeling

Preprocessing

To prepare the dataset for modeling, I removed the ‘id’ column and mapped class labels to integer values (Benign: 0, Malignant:1). From the original dataset, I created the train and test sets with 75-25 split and a random state of 100 to ensure reproducibility of results. I used the following classifiers for modeling:

1. Logistic Regression

2. Random Forest
3. XGBoost

Since this is a classification problem, the performance metrics used to evaluate the model are the following:

1. Precision, Recall, F1-score, support (Classification Report)
2. ROC Curve
3. Confusion Matrix
4. R^2

Logistic Regression Model

I chose logistic regression as the first model to have a baseline to compare against the other two models. Logistic regression is ideal because it is easier to implement, interpret, and efficient to train. The hyperparameters I used for this model were penalty = 'L2' and C = 10. After fitting the model on the training dataset, the R^2 of the test dataset yielded 0.979, and the confusion matrix showed a very small rate of false positives and false negatives. I also performed 5-fold cross-validation to account for the imbalance dataset as discovered during EDA. The mean CV score was 0.997 which indicates there is no overfitting on the dataset. The classification report also produced exceptionally high scores in all four metrics. Lastly, Figure 5 is the ROC curve for the logistic regression model which also confirms the exceptional performance of the model.

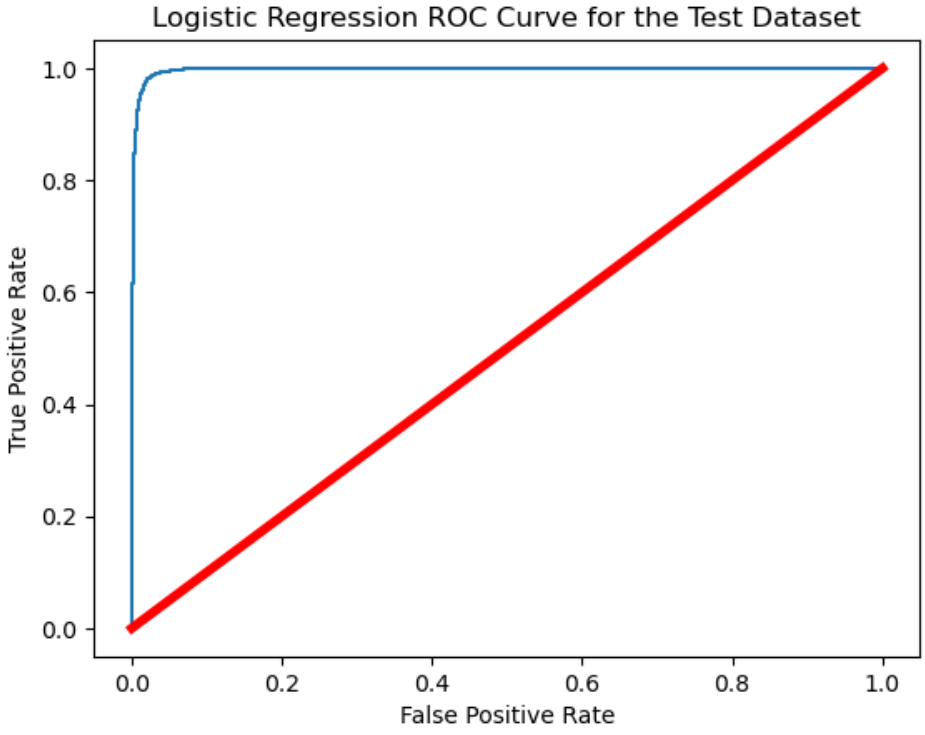

Figure 7. ROC curve for the logistic regression model

Random Forest Model

Random Forest is a good choice for classification problems due to its high accuracy, robustness, feature importance, versatility, and scalability. The hyperparameters I used for this model were enable bootstrapping (in order to combat the data imbalance), n_estimators = 100, and criterion = 'entropy'. the R^2 of the test dataset yielded 0.985 after fitting the model on the test set. The confusion matrix showed a slightly smaller rate of false positives and false negatives compared to the logistic regression model. I performed a 5-fold cross-validation similar to the previous model and discovered the mean CV score to be 0.998 which indicates there is no overfitting on the dataset. The classification report once again produced exceptionally high scores in all four metrics. In Figure 8, the ROC curve is depicted for the random forest model which looks very similar to the logistic regression model. I also evaluated the feature importance according to the RF model and plotted the data as shown in Figure 9. The figure shows that 'Mitoses', 'Normal_Nucleoli', and 'Bland_Chromatin' are significantly more important to the model in comparison to the other features.

The troubling trend of exceptionally high performance among both models demanded more attention, therefore, after discovering the most important features in the RF model, I re-evaluated the model performance after removing the most important feature. Quite surprisingly, the performance remained very similar despite losing the 'Mitoses' feature. This hints at the presence of multi-colinearity among the feature variables.

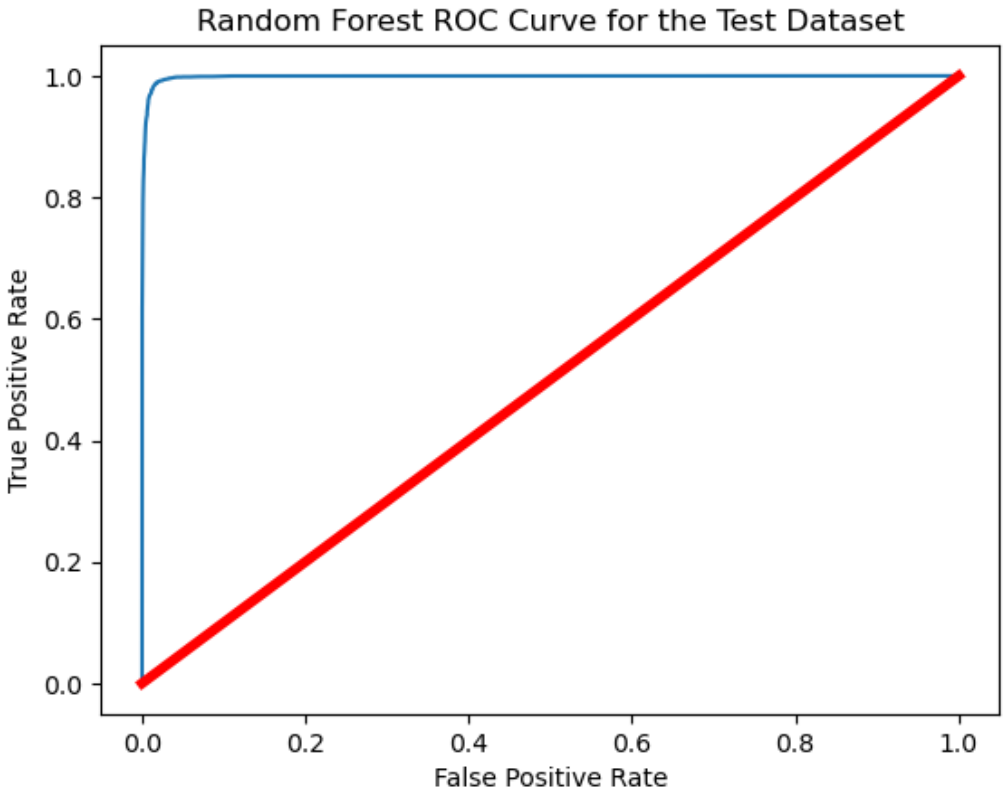

Figure 8. ROC curve for the random forest model

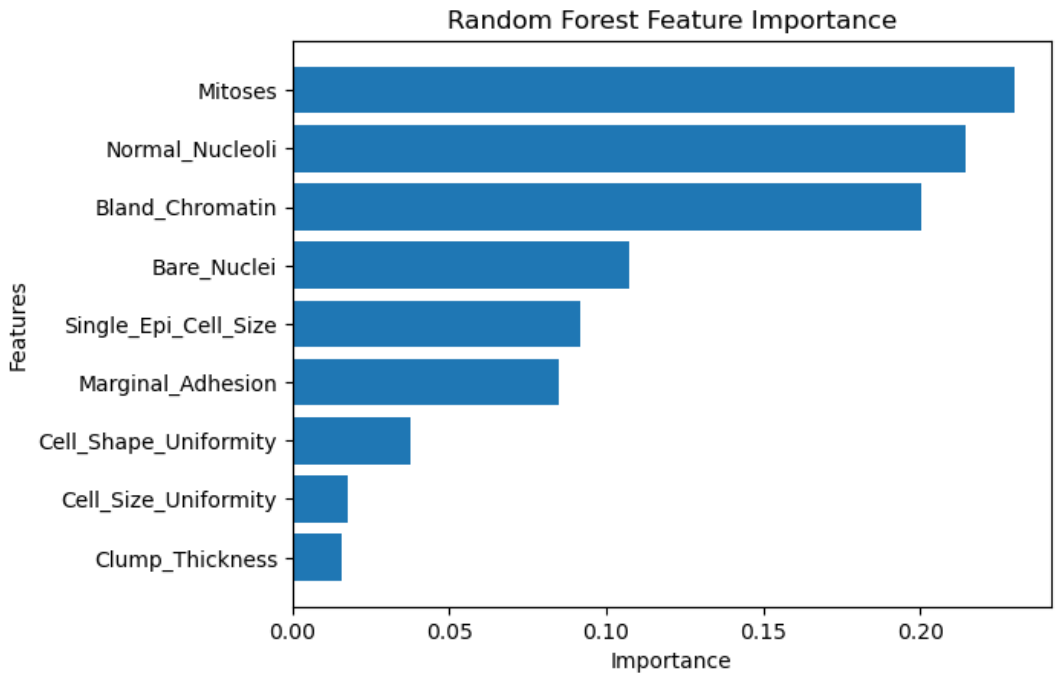

Figure 9. Feature importance plotted for random forest model

XGBoost Model

The last model I used on the dataset is XGBoost. This classifier is a good choice for classification problems due to its versatility, efficiency, and use of regularization to avoid overfitting. The R^2 of the test dataset yielded 0.985 after fitting the model on the test set, similar to the other two models. The confusion matrix showed the smallest rate of false positives and false negatives compared to the other models. I performed a 5-fold cross-validation and discovered the mean CV score to be 0.997. The classification report yet again produced exceptionally high scores in all four metrics. I also evaluated the feature importance according to the XGBoost model and plotted the data as shown in Figure 10. The ‘Mitoses’ feature for XGBoost is overwhelmingly significant for the XGBoost model in comparison to the random forest model.

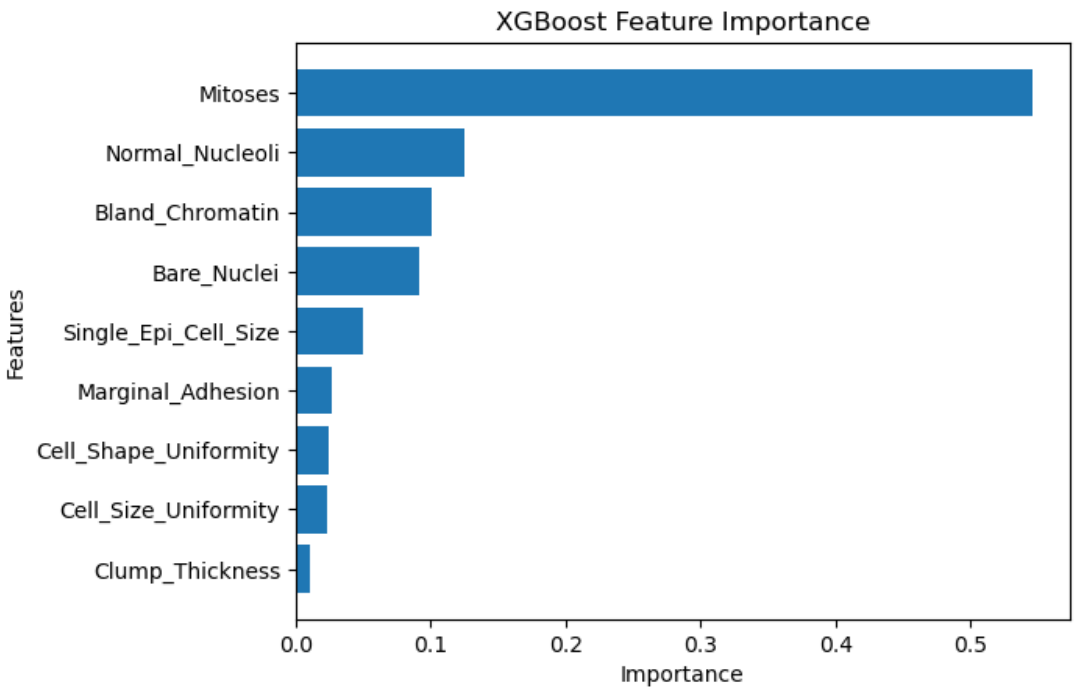

Figure 10. Feature importance plotted for XGBoost model

Model Performance Comparison

As mentioned previously, I evaluated each model in terms of model accuracy score, and 'ROC-AUC' score for both the training and test data, and plotted them. Among the three models, there isn't a significant difference in performance. Figure 11 plots the model accuracy scores for purposes of visual comparison and the difference is not visible. Similarly, the ROC-AUC scores in Figure 12 for each model are nearly identical.

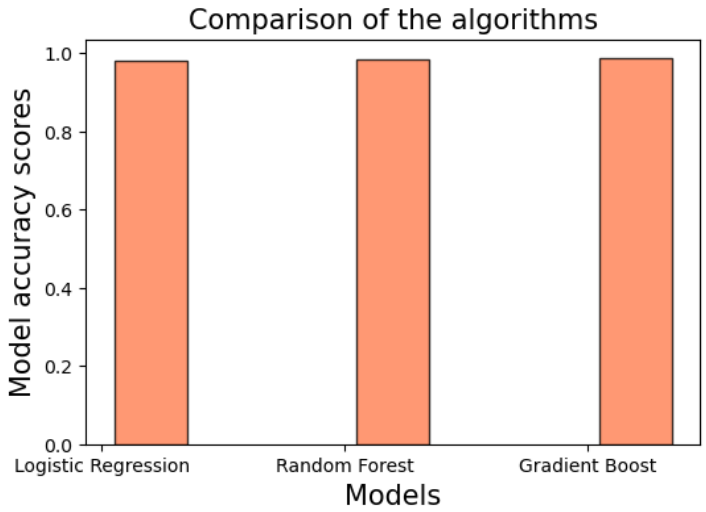

Figure 11. Bar chart of accuracy scores for the three models

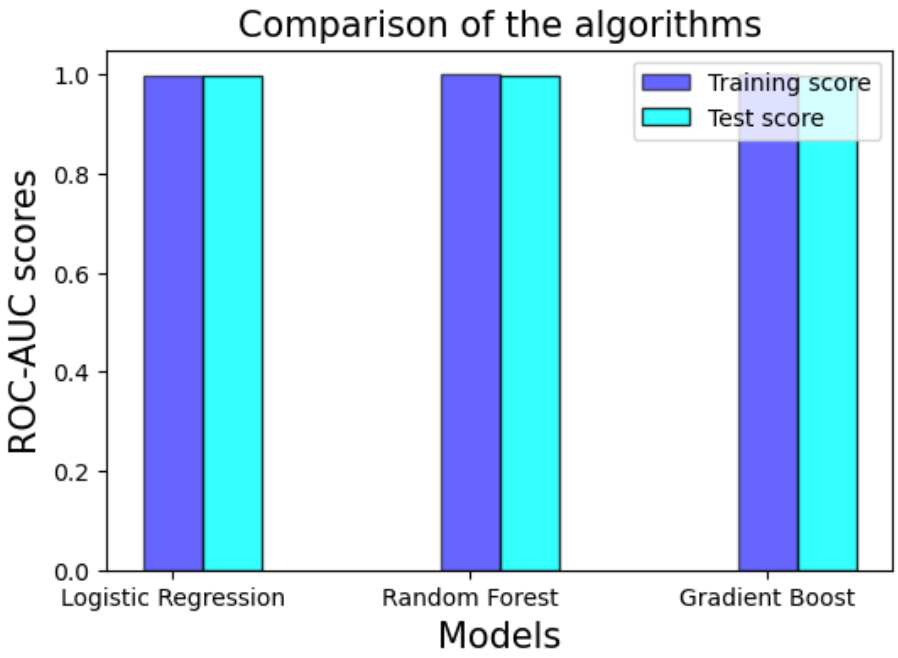

Figure 12. Bar chart of ROC-AUC scores for the three models

6. Conclusions

When the performance of the three models is compared, XGBoost edges out the Random Forest and Logistic Regression models slightly with better accuracy. Given the nature of the problem, it is highly essential to minimize the rate of false positives and false negatives as an incorrect diagnosis can significantly alter the patient's prognosis. The XGBoost model produces the least amount of false positives and false negatives, albeit by a very small margin hence it is the ideal choice for this classification problem.

However, if the model performance is observed on a grand scale, each model has unrealistically high scores which suggests the presence of data leakage and other fundamental issues with the dataset. The biggest issue is the absence of several key features that can directly influence the outcome of the target variable such as the time passed since initial diagnosis, tumor stage, etc. Furthermore, the dataset is static, i.e., the data for each patient is a one-time collection as opposed to multiple collections over a period of time. This prevents us from understanding the progression of the tumor and therefore comparing each tumor on an even scale. With the current dataset, it is possible that entries that are classified as malignant tumors are derived from individuals already diagnosed and in the later stages of cancer.

These fundamental issues with the dataset defeat the purpose of machine learning in this instance because the goal is to use ML models as a supplement to real-world situations, such as being able to predict potential malignant tumors before they are diagnosed by doctors or if they can't be identified. While we can fit models to this dataset and achieve high-accuracy results, the real-world utility is lacking.