Breast Cancer Classification

by Mugdha Paithankar

OVERVIEW

Breast cancer (BC) is one of the most common cancers among women in the world today. An early diagnosis of BC can greatly improve the prognosis and chance of survival for patients. An accurate classification of benign tumors can prevent patients from undergoing unnecessary treatments. Thus, the correct diagnosis of BC and classification of tumors into malignant or benign groups is a subject of much ongoing research. With the use of advanced machine learning algorithms, I plan to build a model which accurately classifies tumors as Benign or Malignant based on certain features.

DATA

The dataset has been obtained from Kaggle. It contains 596 rows and 32 columns of tumor shape and specifications. The tumor is ultimately classified as benign or malignant based on its geometry and shape. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. There are 212 malignant and 357 benign tumors in the dataset.

The features of the dataset include:

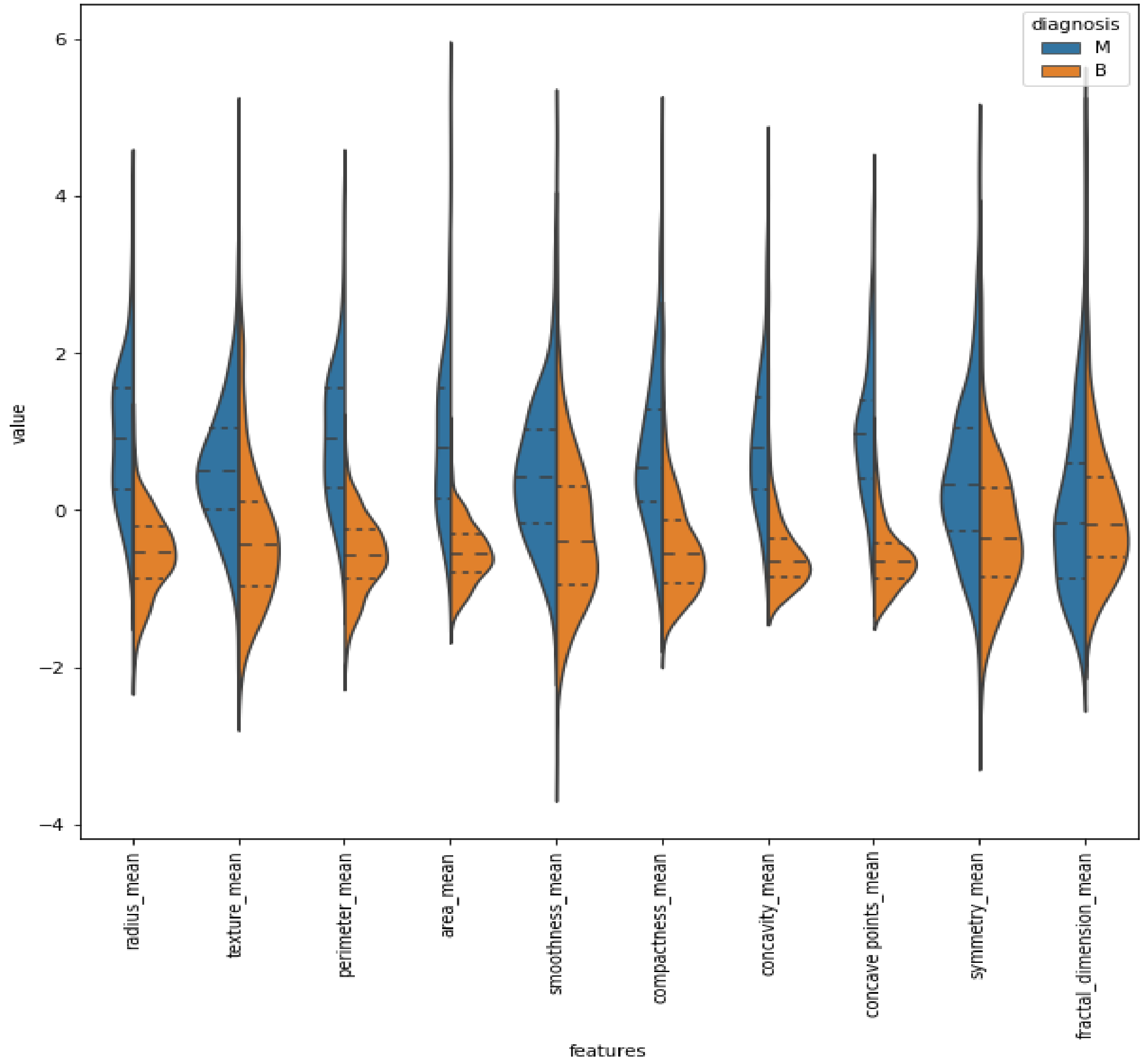
1. tumor radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area — 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features.

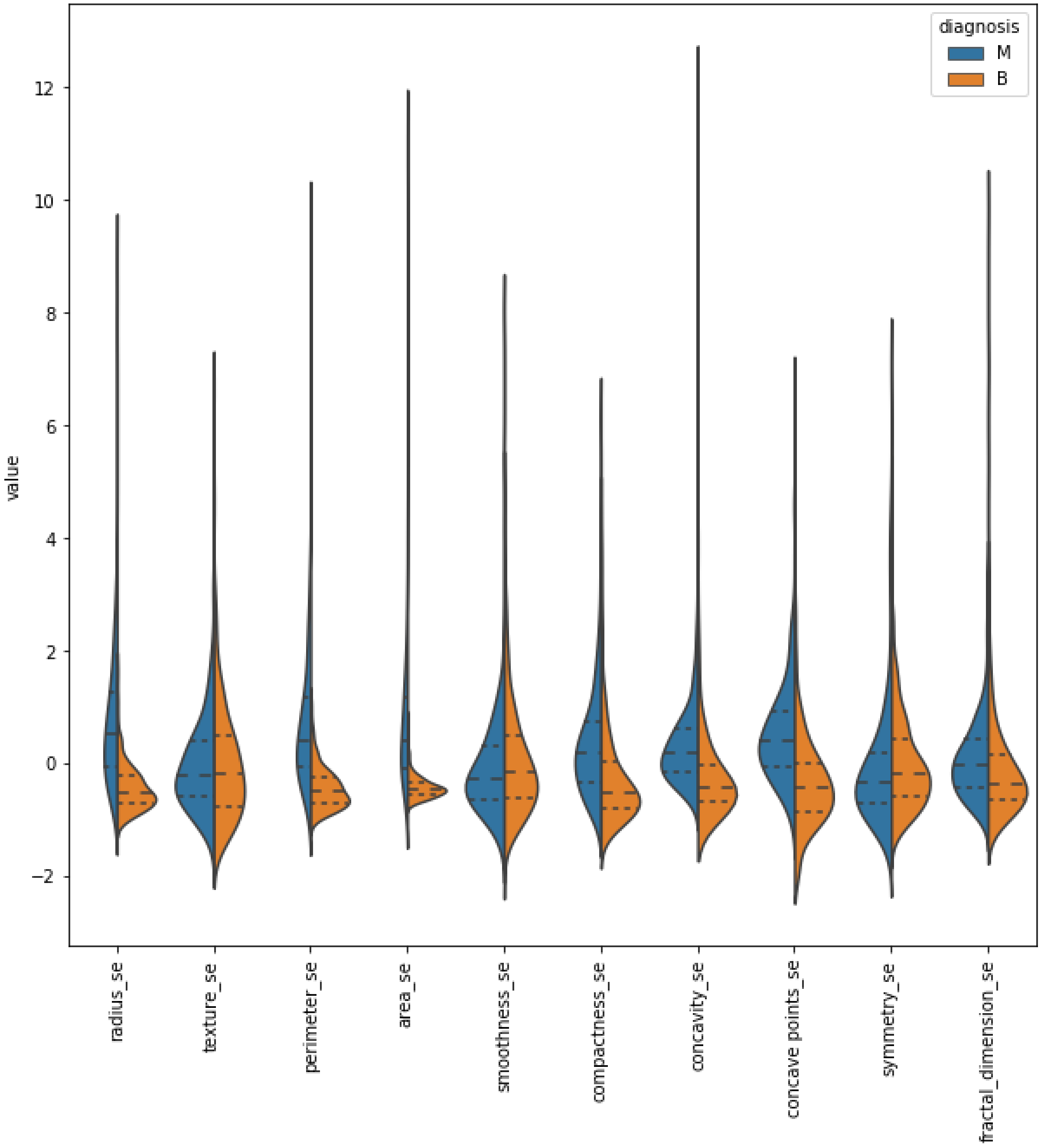
EXPLORATORY DATA ANALYSIS

Violin Plots

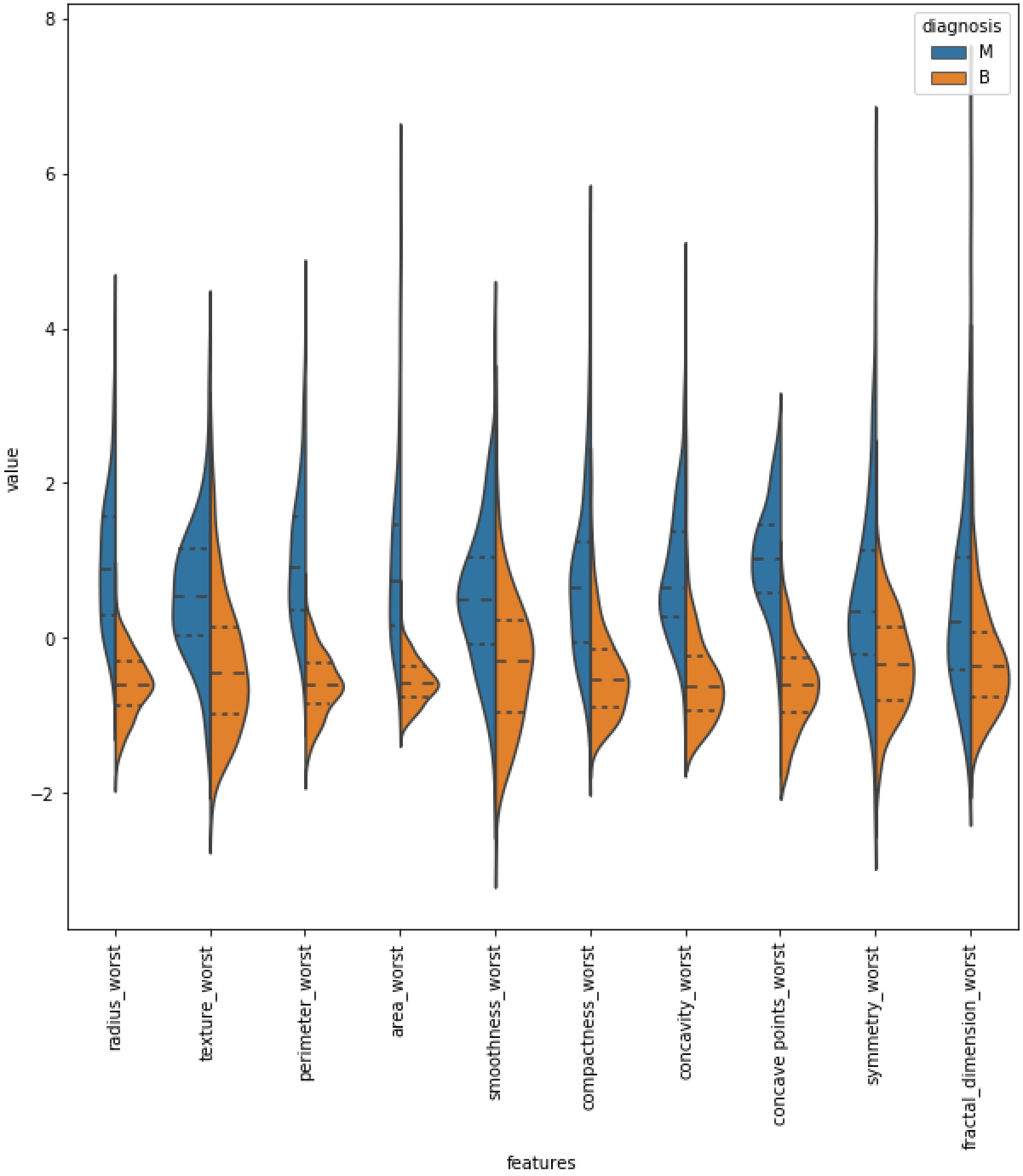
The dataset did not have any missing or null values. So I moved on to EDA and visualization. Violin plots were plotted to visualize the 30 features in order to get some insights into the distribution pattern of the features, their mean, std deviation or variance. A violin plot shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared. The mean, standard error and worst dimensions of the ten features were plotted separately in each series of violin plots.



The median of texture_mean for Malignant and Benign looks separated, so it might be a good feature for classification. For fractal_dimension_mean, the medians of the Malignant and Benign groups are very close to each other which might not be good for classification. smoothness_mean seems to have the highest range of values.



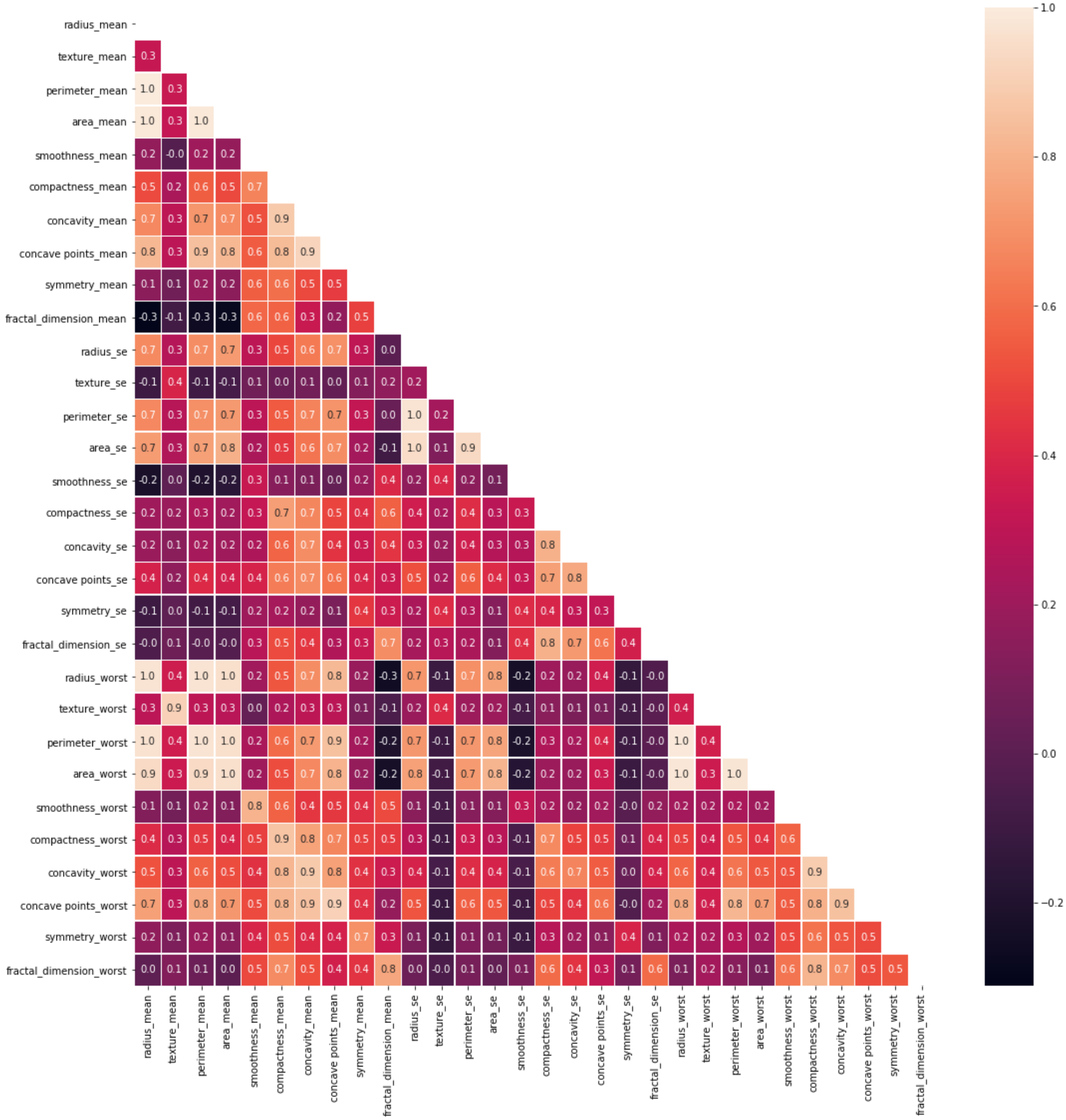
The medians for almost all Malignant or Benign don't vary much for the standard error features above, except for concave points_se and concavity_se. smoothness_se or symmetry_se have a very similar distribution which could make classification using this feature difficult. The shape of the violin plot for area_se looks warped. The distribution of data points for benign and malignant in area_se looks very different and varys the most.



area_worst look well separated, so it might be easier to use this feature for classification! Variance seems highest for fractal_dimension_worst. concavity_worst and concave_points_worst seem to have a similar data distribution. Let us check if these features are correlated.

Correlation Matrix

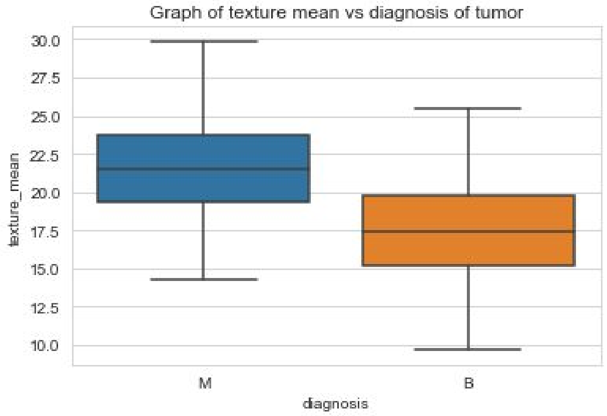
A correlation matrix was plotted to check for correlation and multicollinearity between the features.

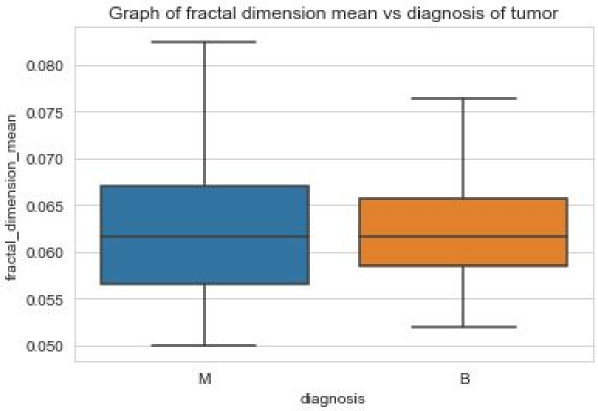


The means, std errors and worst dimension lengths of compactness, concavity and concave points of tumors are highly correlated amongst each other (correlation > 0.8). The std errors of radius, perimeter and area of tumors have a correlation of 1! The worst dimension of radius, perimeter and area also have a correlation of 1. texture_mean and texture_worst have a correlation of 0.9. area_worst and area_mean have a correlation of 1. The mean and worst dimension of radius, perimeter and area have a correlation of 1.

STATISTICAL ANALYSIS

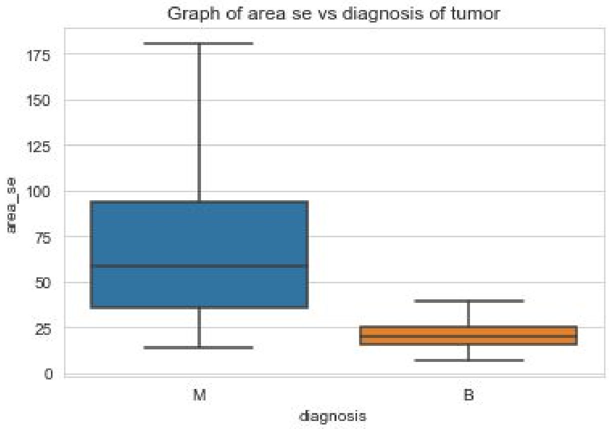
Box Plots

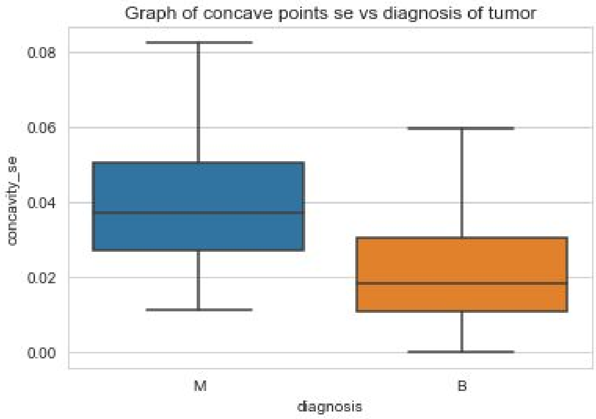




Texture means, for malignant and benign tumors vary by about 3 units. The distribution looks similar for both the groups. Malignant tumors tend to have a higher texture mean compared to benign.

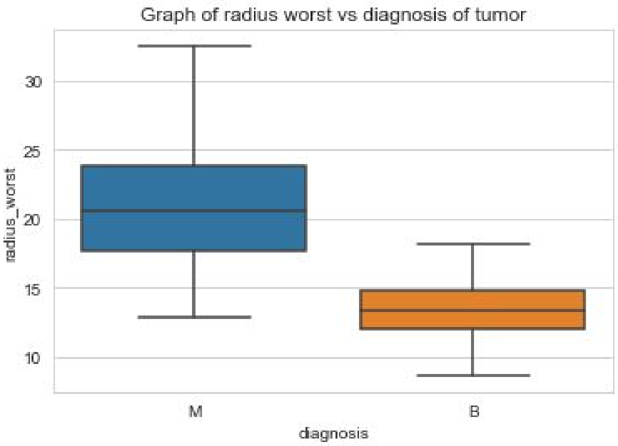
Fractal dimension means are almost the same for malignant and benign tumors. The IQR is widers for malignant tumors.

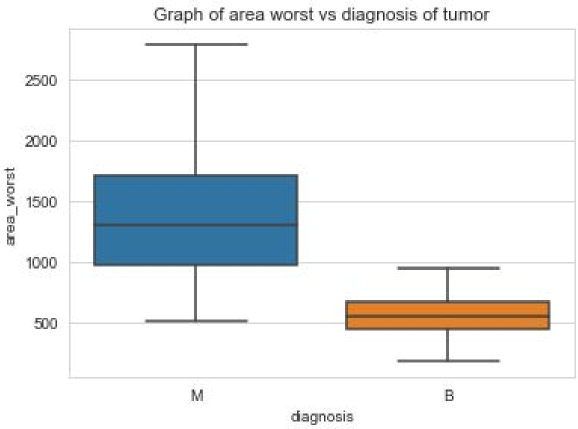




Malignant groups have a distinctly wider range of values for area se. The distribution range is very narrow for benign groups. This might be a good feature for classification.

Standard error (se) of Concave points has a higher mean and IQR for malignant tumors. The distribution looks somewhat similar for both tumor types.





Malignant groups have a wider range of values for radius worst compared to benign groups. The IQR is wider for the same. Malignant tumors have a higher radius worst compared to benign groups.

Similar to area_se, area_worst has very different data distribution for malignant and benign tumors. Malignant tumors tend to have a higher value of mean and wider IQR range. The area_worst for benign tumors varies only between 10-1000, whereas the range is 500-2500< for malignant tumors. Because of noticeable differences between B and M tumors, this could be a good feature for classification.

t-test

Features	**t-statistic**	**p-value**
texture mean	10.86720108	4.05863605e-25
fractal dimension mean	-0.30571113	0.7599368
area se	15.6093429	5.89552139e-46
concave points se	6.24615734	8.26017617e-10
radius worst	29.33908156	8.48229192e-116
area worst	25.72159026	2.8288477e-97

Except for fractal dimension mean, the p value and t statistic is statistically significant for all the features in the table above. The t statistic for the fractal dimension mean, is negative so if there is a difference between the M and B samples, it will be in the negative direction, meaning M tumor samples have lesser means than B tumor samples. However the value of t statistic is very small and p value > 0.01, this means we cannot reject null hypothesis. The difference in means for fractal dimension_mean samples of M and B tumors might not be statistically significant.

From the correlation matrix it was clear that there are quite a few features with very high correlations. So I dropped one of the features, from each of the feature pairs which had a correlation greater than 0.95. 'perimeter_mean', 'area_mean', 'perimeter_se', 'area_se', 'radius_worst', 'perimeter_worst', 'area_worst' were amongst the features that were dropped.

MACHINE LEARNING

The main aim of this project was to classify tumors as benign or malignant and I did so with the help of machine learning algorithms. I used sklearn's Logistic Regression, Support Vector Classifier, Decision Tree and Random Forest for this purpose.

Methodology

- **Data manipulation:** sklearn's LabelEncoder was used to convert the categorical dependent variable (M or B) of the diagnosis column to a numeric data type.
- **Train Test Split:** sklearn's train_test_split was used to split the dataset into training and test sets. 40% of the data was reserved for testing purposes. The dataset was stratified in order to preserve the proportion of target as in the original dataset, in the train and test datasets as well.
- **Feature Scaling:** sklearn's RobustScaler was used to scale the features of the dataset. The centering and scaling statistics of this scaler are based on percentiles and are therefore not influenced by a few number of very large marginal outliers.
- **Training and Testing:** The scaled dataset was then trained and tested using Logistic Regression, SVC, Decision Tree and Random Forest algorithms. An initial classification report and confusion matrix were printed to check the model performance.
- **Hyperparameter tuning:** Each model's parameters were tuned using GridSearchCV in order to improve the model performance.
- **Custom Thresholding:** Finally, a custom threshold was set instead of the default 0.5 threshold value, to try and improve the model performance further.

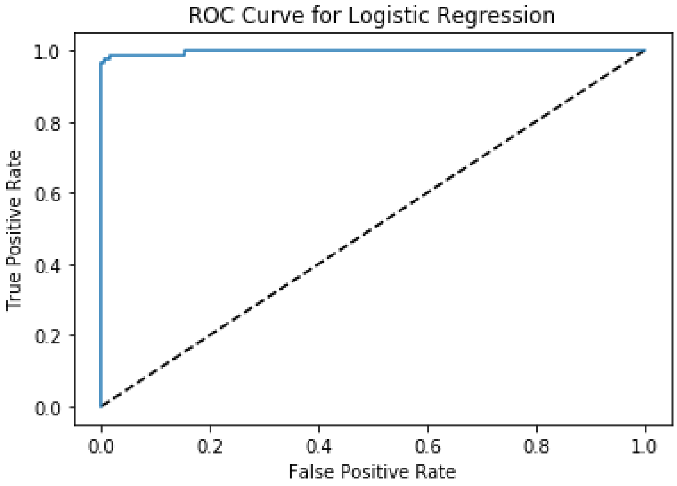
Result

Overall, the Logistic Regression performed the best, followed by the SVM model. The Logistic Regression model with l2 as penalty, C = 0.591 and threshold set to 0.48 had an AUC score of 0.99. The model misclassified only 1 tumor as FN and 2 tumors as FPs. From the point of view of the patient's health, classifying a malignant tumor as benign is worse than classifying a benign tumor as malignant. Therefore, I was focused on getting the least number of FNs for the models, which meant maximising the recall value. In order to do so, I used the fbeta score function with a beta > 1 while grid searching, in order to focus on getting more recall from the model.

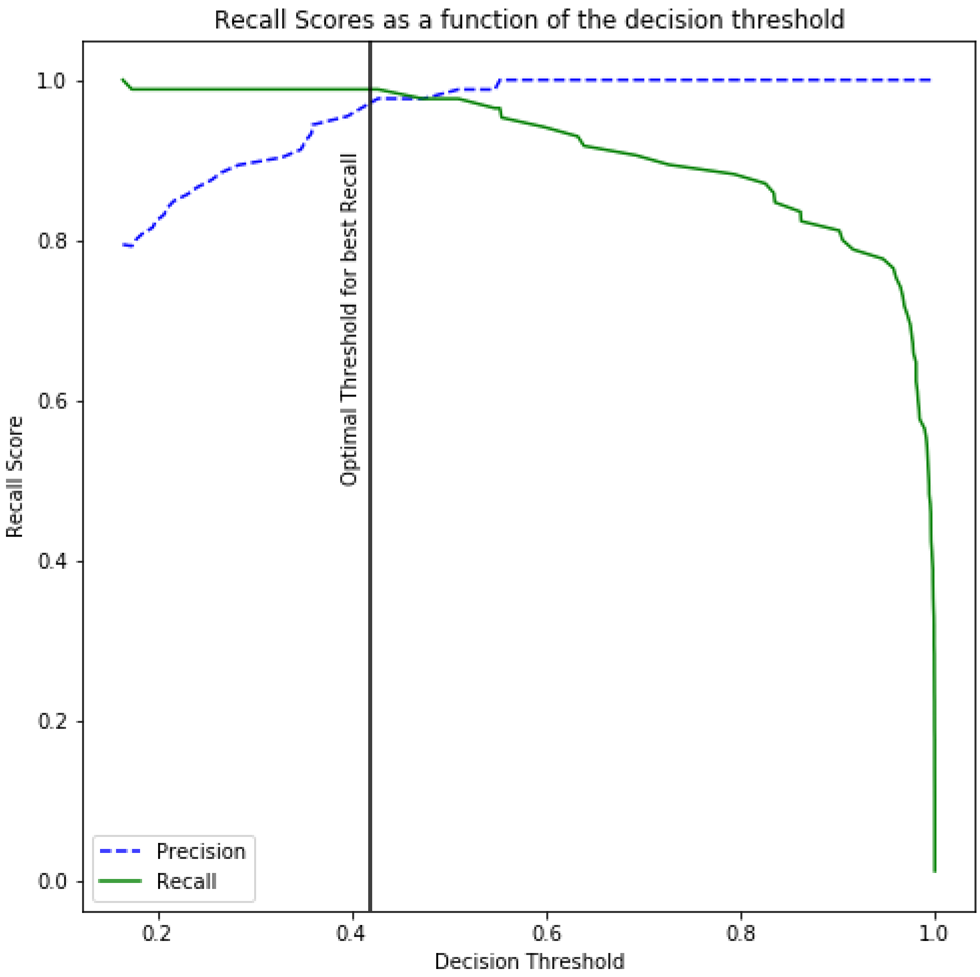
A summary of all the models is shown below.

Model type	Initial values	Hyperparameters	Final values
Logistic Regression	FN: 2 FP: 1	Best Penalty: l2 Best C: 0.591	FN: 1 FP: 2
SVC	FN: 4 FP: 2	C:0.07100000000000001 kerne': linear	FN: 3 FP: 0
Decision Tree	FN: 5 FP: 14	max_depth: 3 max_features: 0.4 min_samples_leaf: 0.06	FN: 4 FP: 14
Random Forest	FN: 6 FP: 4	max_depth: 15 max_features: 10 min_samples_split: 3 n_estimators: 100	FN: 2 FP: 4

As seen from the table above, the AUC score for the logistic regression model is 0.9980 and it has a minimum number of misclassifications for the positive class. The SVC and random forest models are a close second with only 3 and 2 FNs. Recursive Feature Elimination was applied to the random forest model to check for any redundant features.



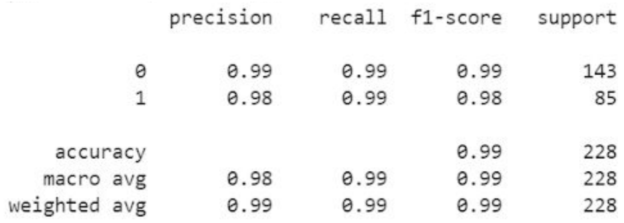
ROC curve for the best performing model of Logistic Regression



Graph of recall scores VS thresholds

CONCLUSION

To conclude my project, I used a variety of algorithms in order to correctly classify tumors as malignant or benign. Among all the algorithms tried out, the Logistic Regression and Support Vector Classifier gave maximum accuracies and minimum misclassifications for the positive class. The goal was to maximise recall values so as to avoid misclassifications of FN type.



```
              precision    recall  f1-score   support

           0       0.99      0.99      0.99       143
           1       0.98      0.99      0.98        85

    accuracy                           0.99       228
   macro avg       0.98      0.99      0.99       228
weighted avg       0.99      0.99      0.99       228
```

Classification reports for logistic regression and SVM classifiers

As can be seen from the classification report displayed above, both the models performed exceedingly well. The recall scores were 0.99 and 0.96 respectively.

FUTURE DIRECTIONS

- **Get more data!** This dataset had only 569 rows. More data would improve the viability of these models in a real world scenario.
- **More feature selection/reduction techniques** The dataset has a large number of correlated features. For this project I was only able to try the correlation matrix heatmap and RFE on random forest but other techniques like PCA can be tried out too. The scikit-learn library provides the SelectKBest class that can be used with a suite of different statistical tests to select a specific number of features.
- **Diversify the dataset** This is a diagnostic dataset. But details about the patient, like his age, family and medical history can also be included to get a more real world understanding of who is most at a risk of developing breast cancer.
- **Try to reduce the FNs to 0?** As much as I tried, I was only able to get the FNs to 1. Getting them to 0 without a lot of compromise on precision would be great!