BREAST CANCER CLASSIFICATION AND PREDICTION

CREATED BY: -

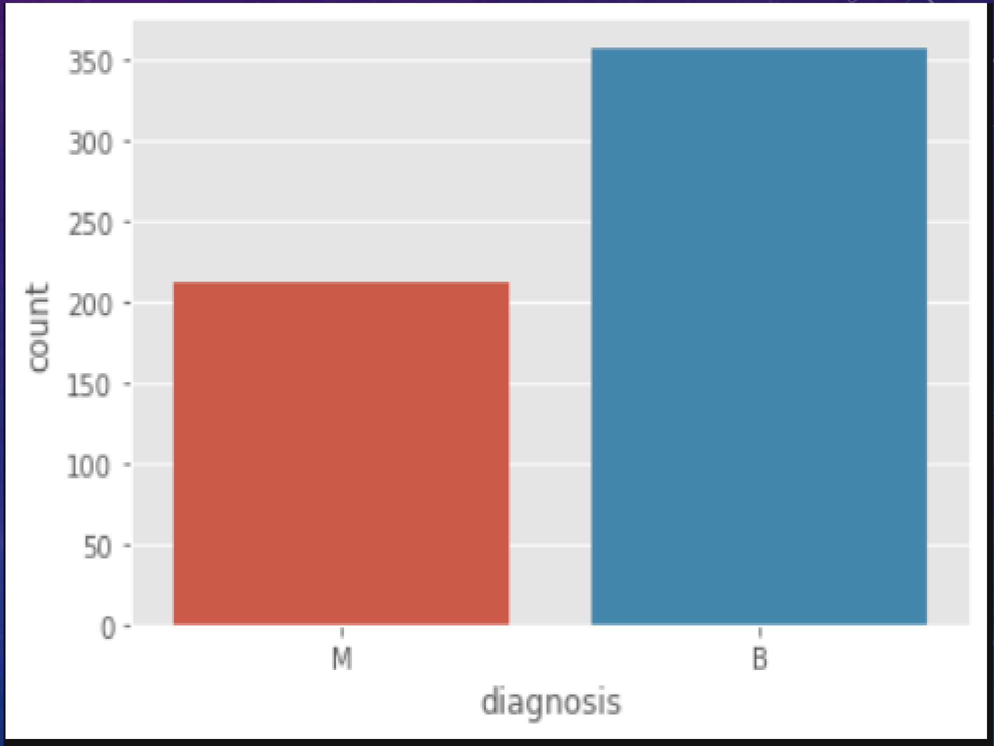
AYUSHYA VERMA

INFO

- Breast cancer (BC) is one of the most common cancers among women in the world today.
- A correct diagnosis of BC and classification of tumors into malignant or benign groups

DATA

- Obtained from Kaggle. It contains 596 rows and 32 columns of tumor shape and specification
- The tumor is classified as malignant or benign

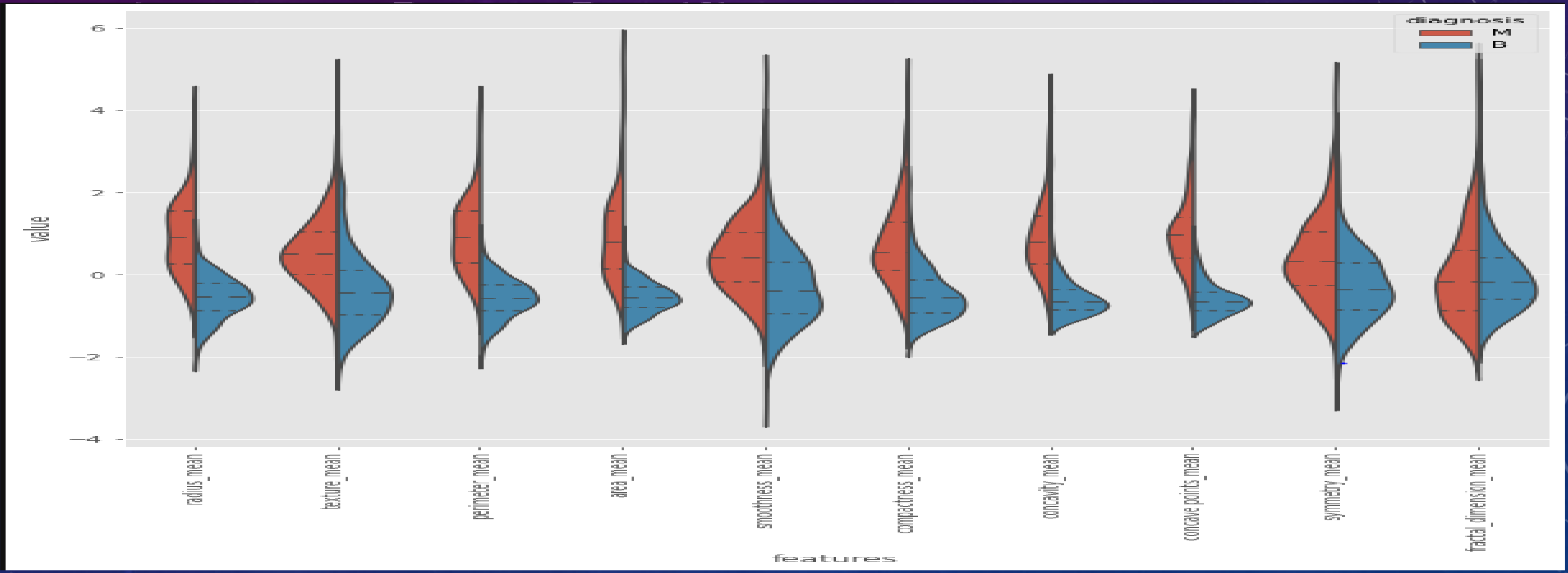


DATA DESCRIPTION

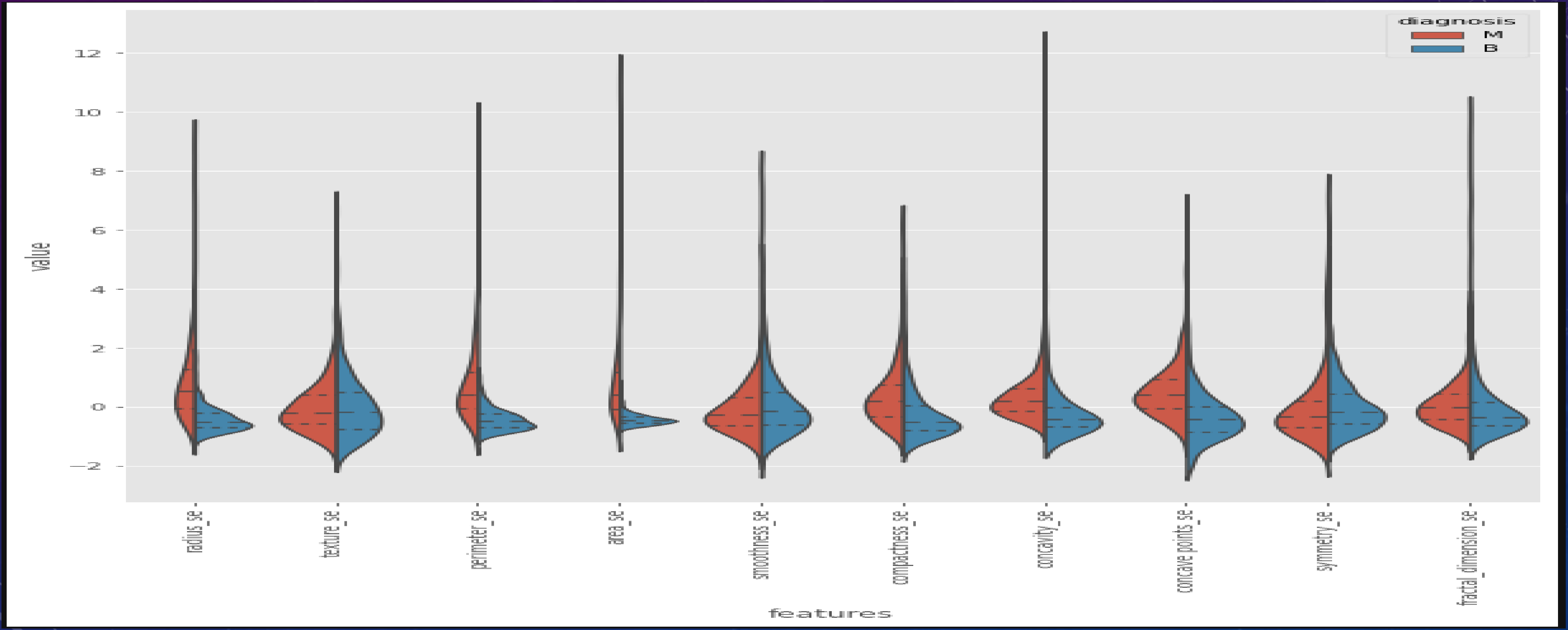
• tumor radius (mean of distances from center to points on the
perimeter)
• texture (standard deviation of gray-scale values)
• area
• smoothness (Local variation in radius lengths)
• compactness (perimeter2 / area — 1.0)
• concavity (severity of concave portions of the contour)
• concave points (number of concave portions of the contour)
• symmetry
• 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.

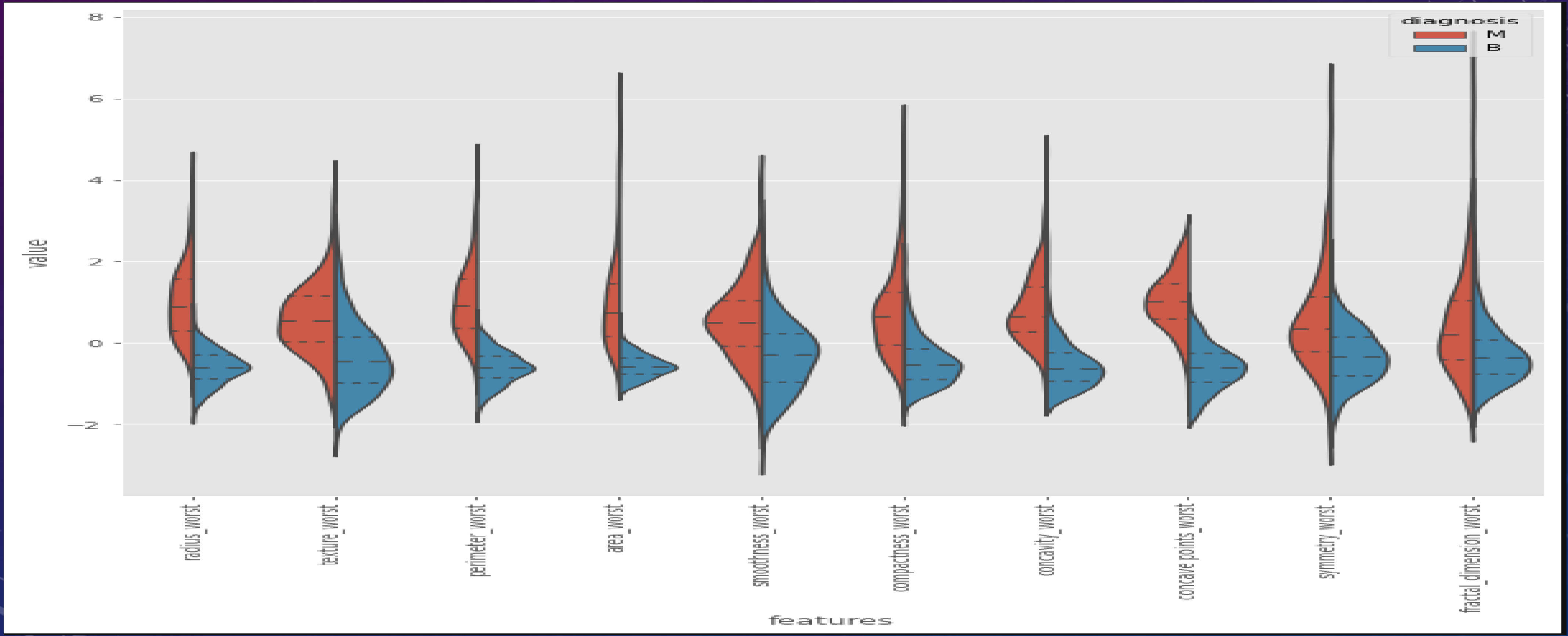
VIOLIN PLOTS FOR ALL THE MEANS



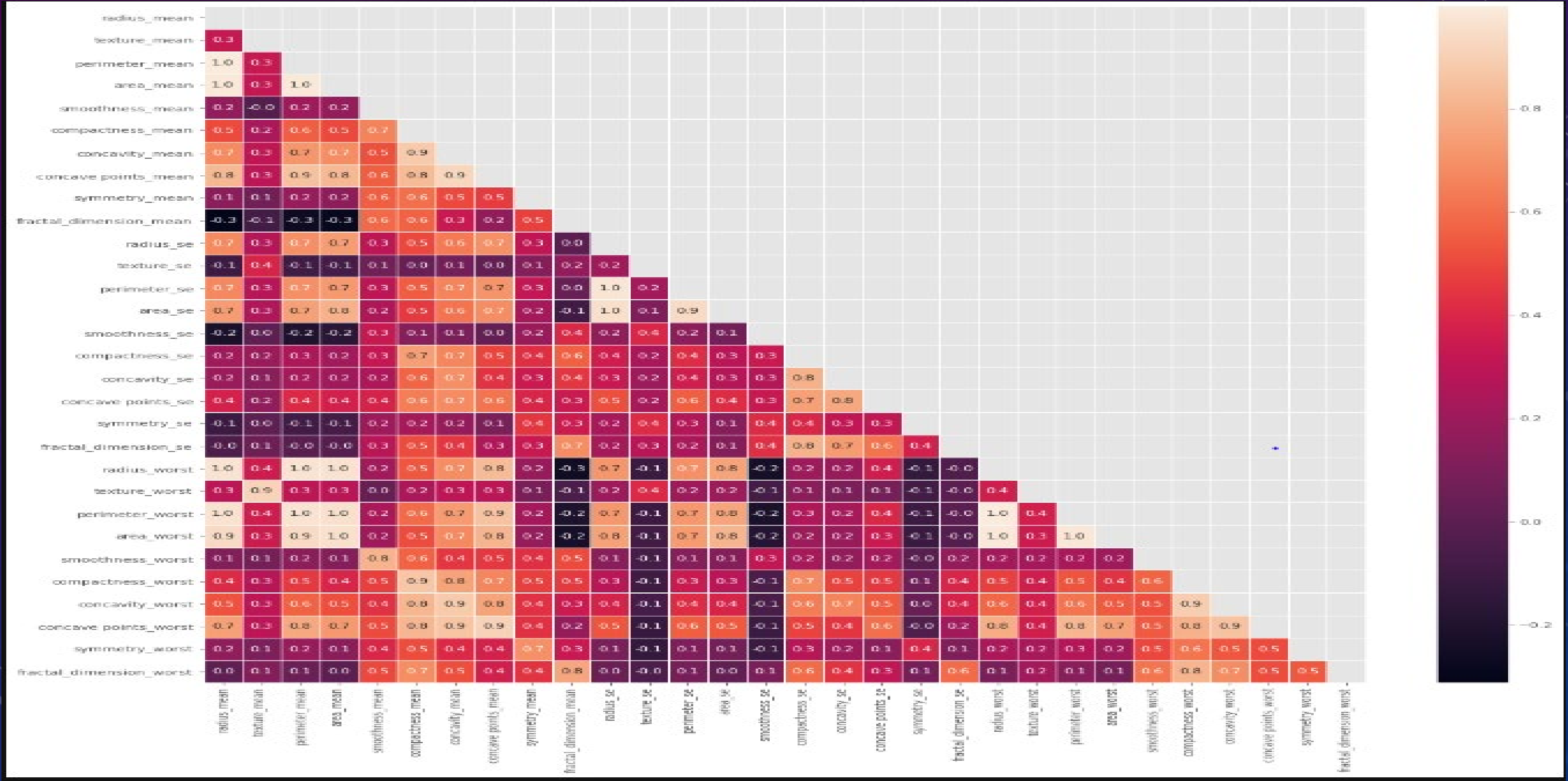
VIOLIN PLOTS FOR ALL THE STANDARD ERRORS

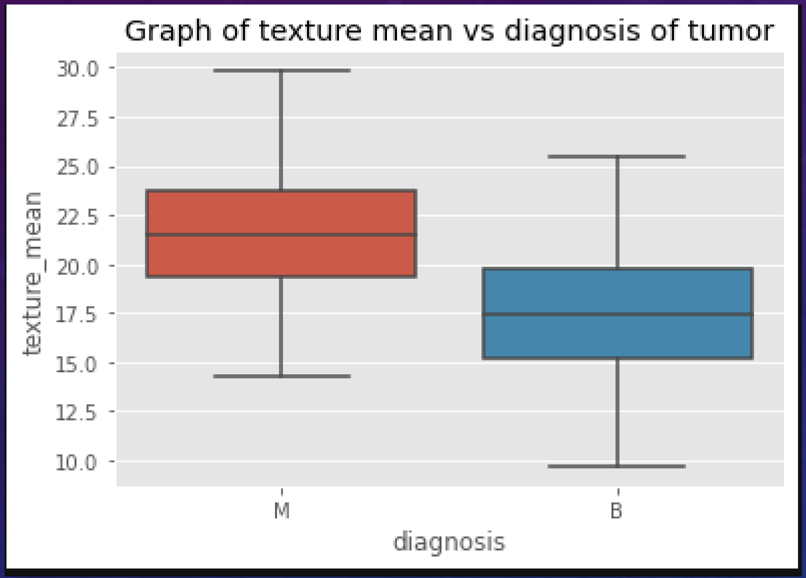


VIOLIN PLOTS FOR ALL THE WORST DIMENSIONS

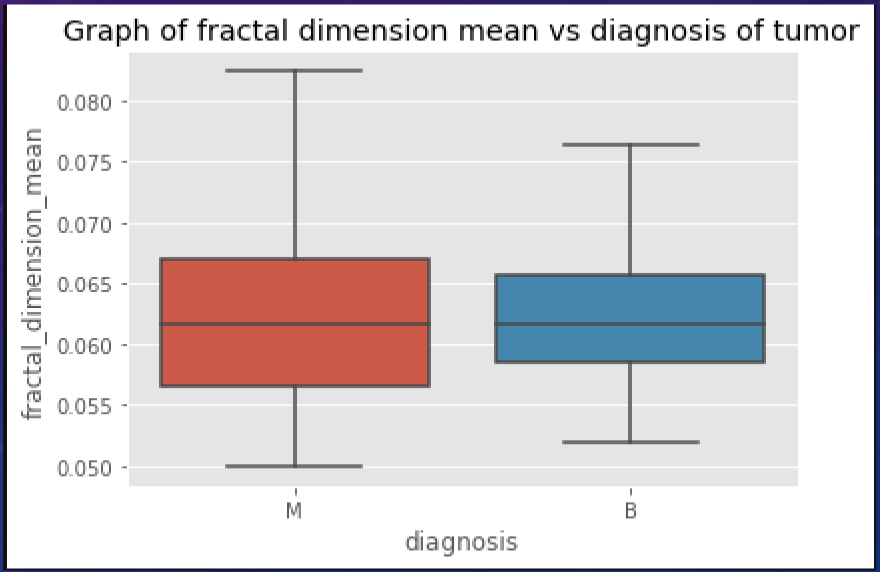


Multi collinearity check via correlation matrix

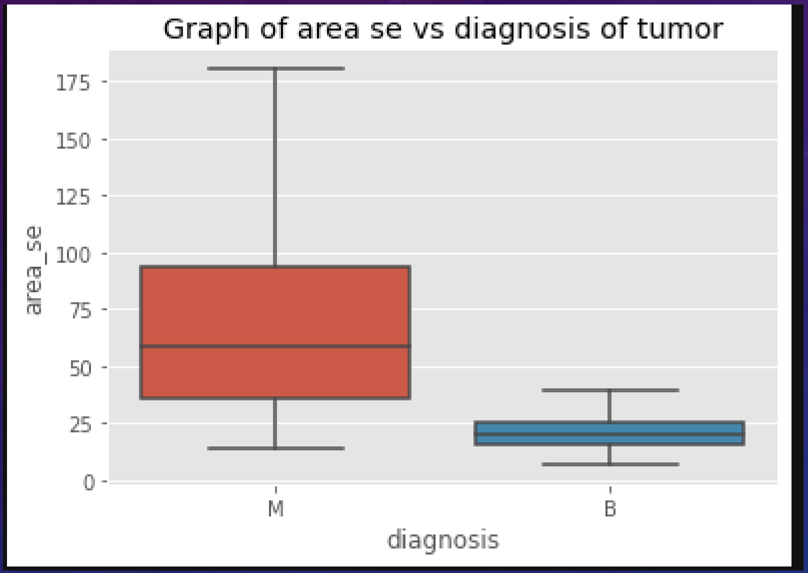


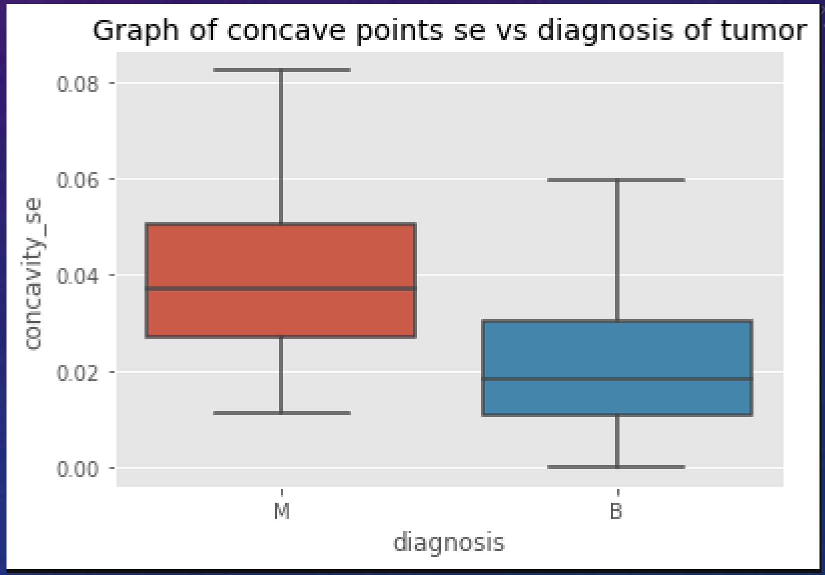


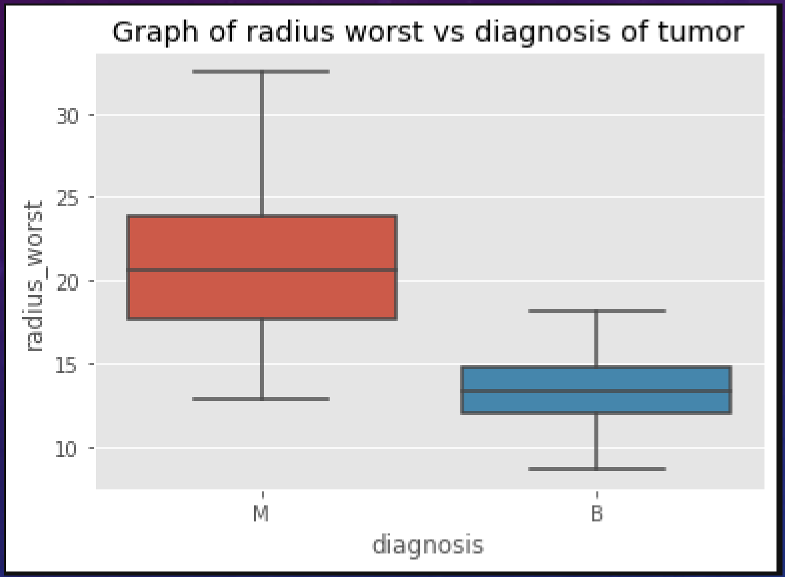
Graph of texture mean vs diagnosis of tumor

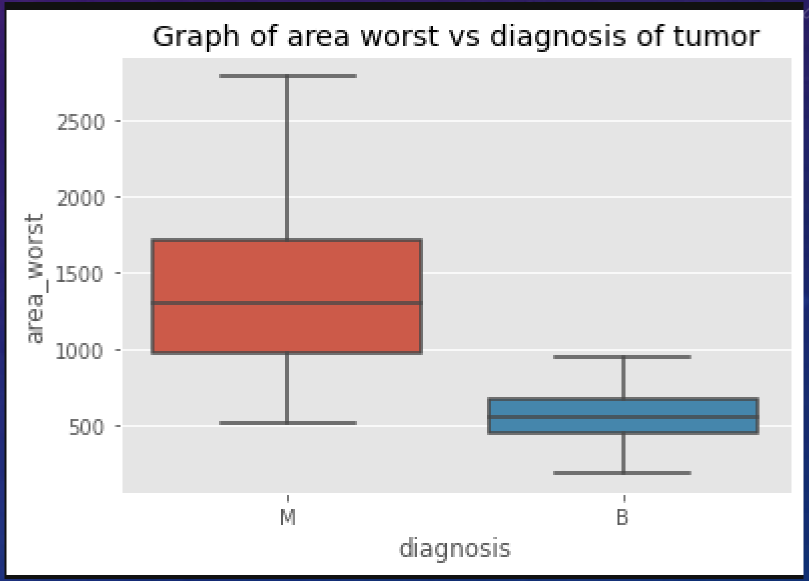


Graph of fractal dimension mean vs diagnosis of tumor


Graph of area se vs diagnosis of tumor
area_se
0
25
50
75
100
125
150
175
M
B
diagnosis


Graph of concave points se vs diagnosis of tumor
concavity_se
0.00
0.02
0.04
0.06
0.08
M
B
diagnosis


Graph of radius worst vs diagnosis of tumor
radius_worst
30
25
20
15
10
M
B
diagnosis


Graph of area worst vs diagnosis of tumor
area_worst
2500
2000
1500
1000
500
M
B
diagnosis

STATISTICAL ANALYSIS: T TEST

Feature	T-Static	P-value
Texture mean	10.86720108000000	4.05863605e-25
fractal dimension mean	-0.30571113	0.7599368
Area se	15.6093429	5.89552139e-46
Concave point se	6.24615734	8.26017617e-10
Radius worst	29.33908156	8.48229192e-116
Area worst	25.7215903	2.8288477e-97

ML 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.

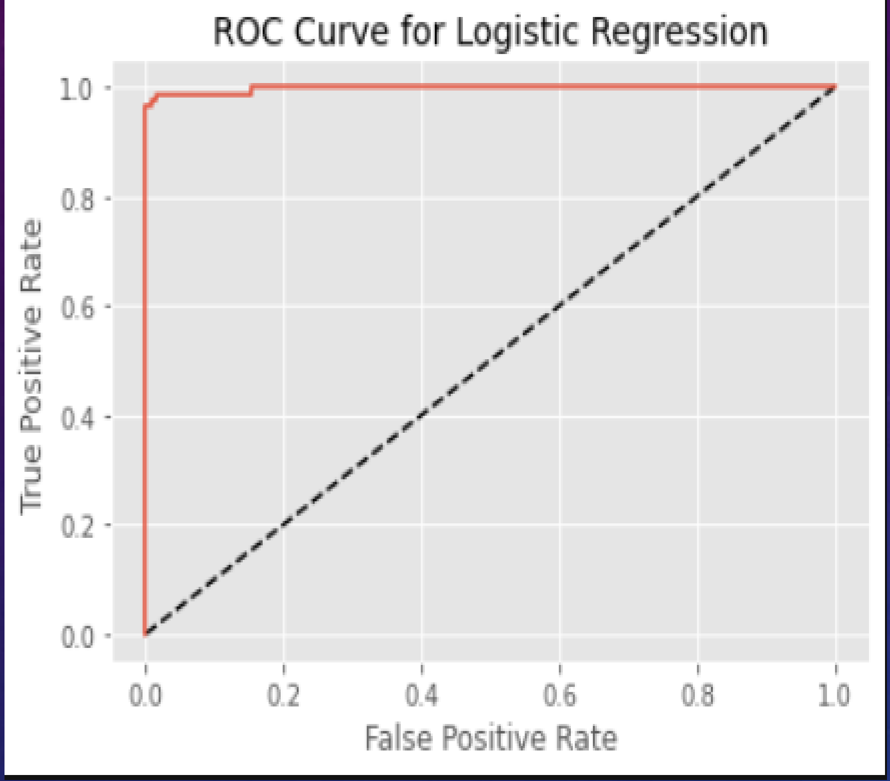
Hyperparameter tuning: Each model’s parameters were tuned using GridSearchCV in order to improve the model performance.

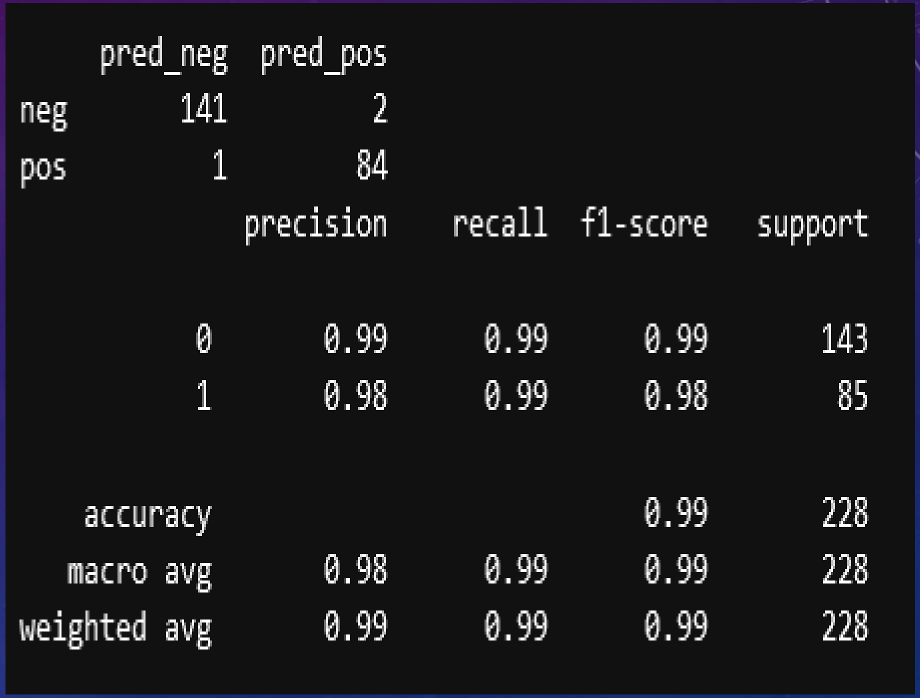
Custom Thresholding: Finally, a custom threshold was set instead of the default 0.5 threshold value, to try and improve the model performance further

SUMMARY OF ML RESULTS

Model Type	Initial Value	Hyperparameter	Final Value
Logistic Regression	FN:2 FP:1	Best Penalty: 12 Best C: 0.591	FN:1 FP:2
SVC	FN:4 FP:2	0:0.07100000000000001 kernel: 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

The AUC score for the Logistic regression model is 0.9980 and it has a minimum number of misclassifications for the positive class.

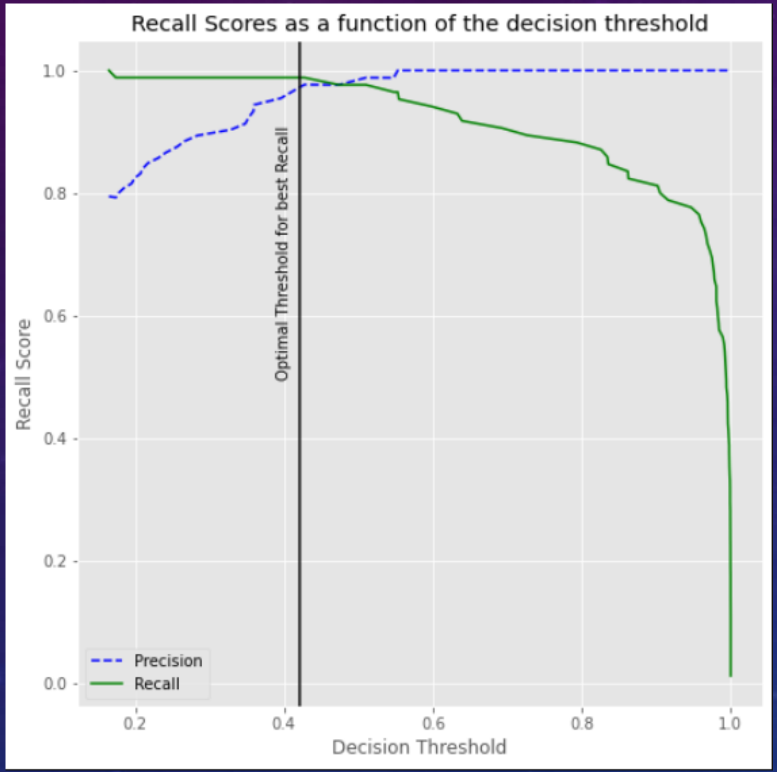




```
     pred_neg  pred_pos
neg       141         2
pos         1        84
              precision    recall  f1-score   support

           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
```



A threshold of 0.42 was chosen for maximum recall.

CONCLUSION

- 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 maximize recall values so as to avoid misclassifications of FN type
- Both the models performed exceedingly well. The recall scores were 0.99 and 0.96 for Logisitc Regression and SVC respectively.

THANK YOU