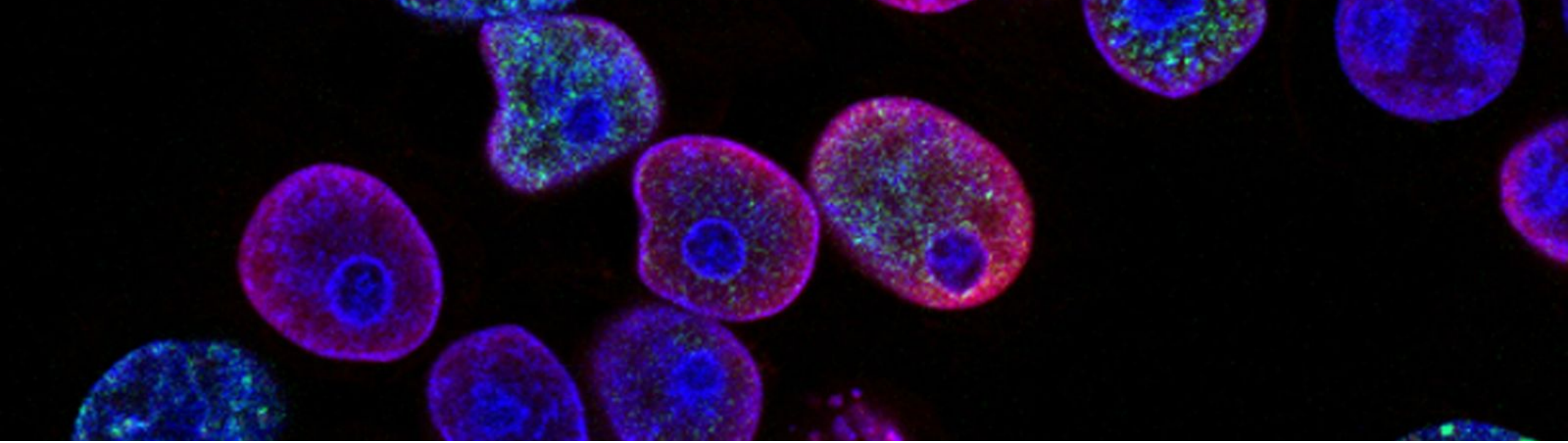


Breast Cancer Classification

-by Mugdha Paithankar

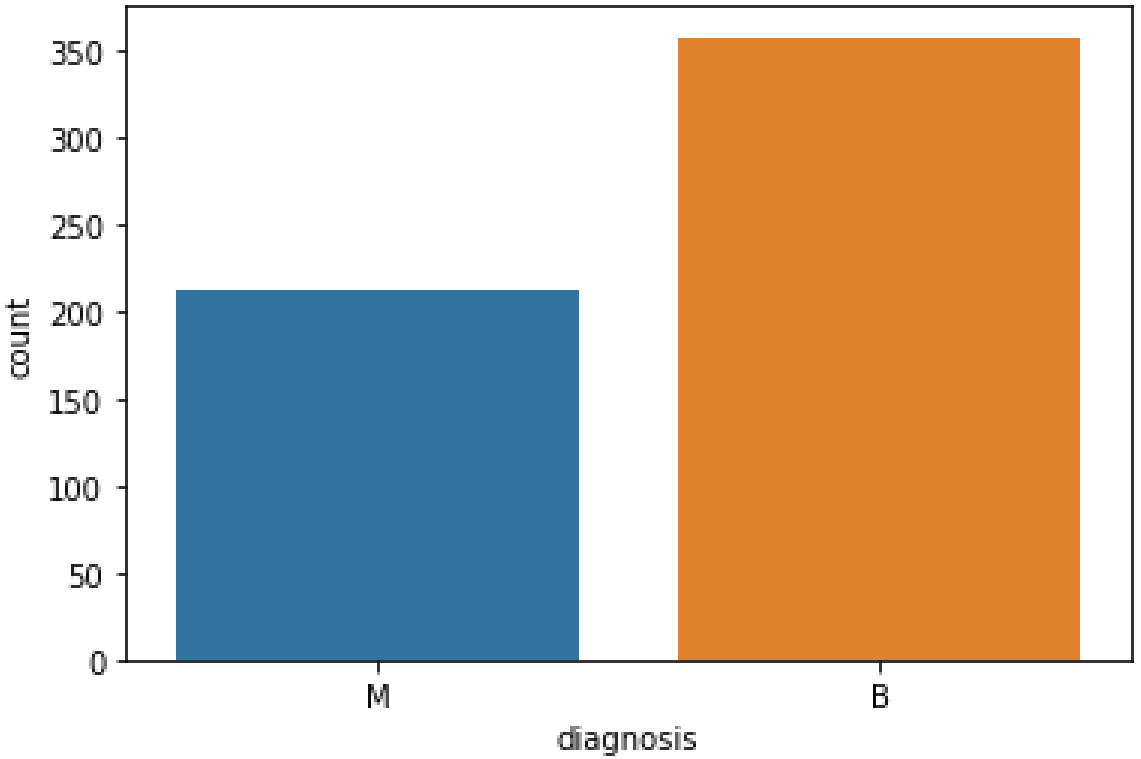
Overview

- 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.
- A correct diagnosis of BC and classification of tumors into malignant or benign groups is a subject of much ongoing research!

Project Procedure

- Get the data!
- Data Wrangling
- Exploratory Data Analysis
- Statistical Testing
- Machine Learning
- Conclusion
- Future Directions

Data



- 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.

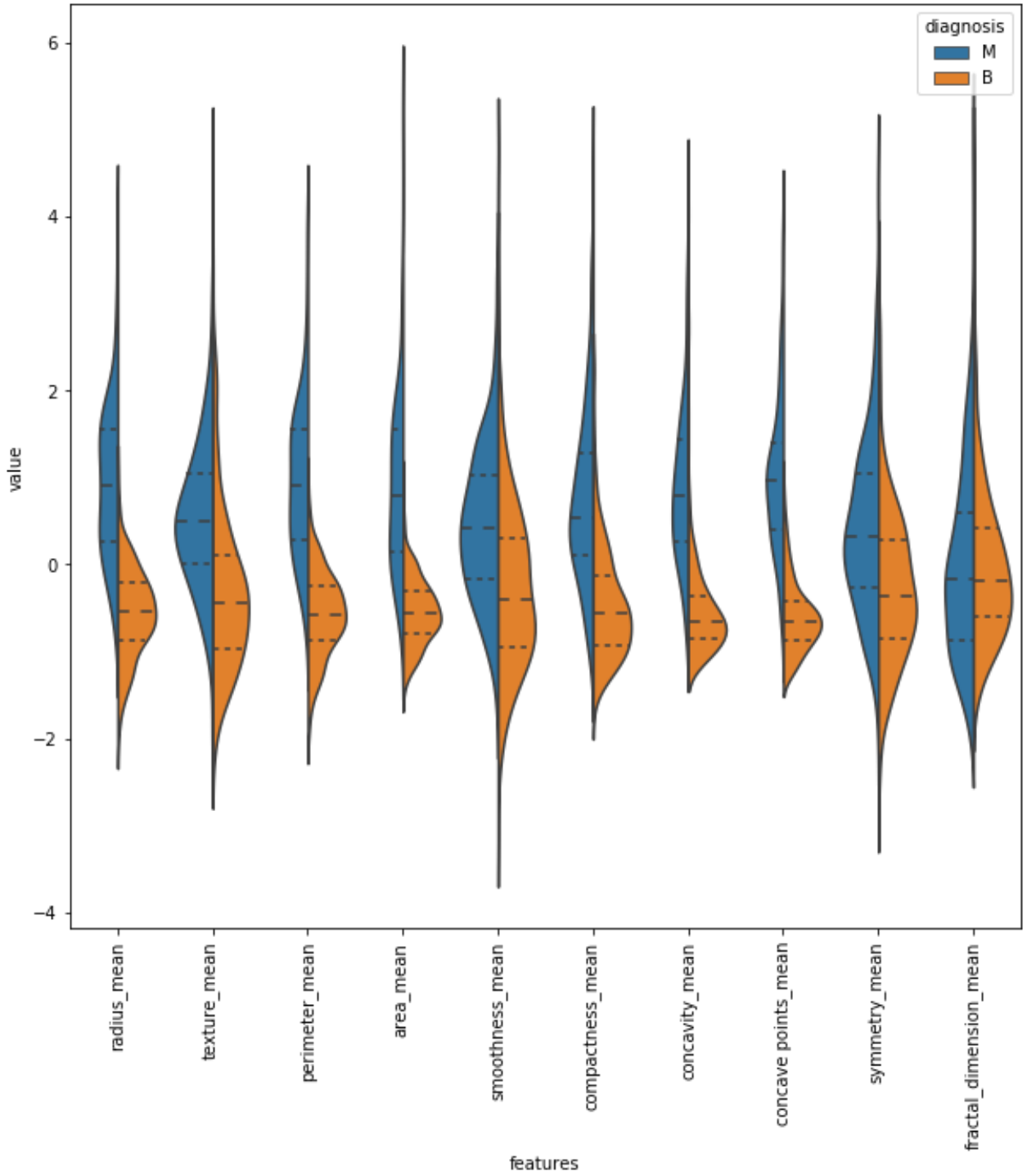
Data Description

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.

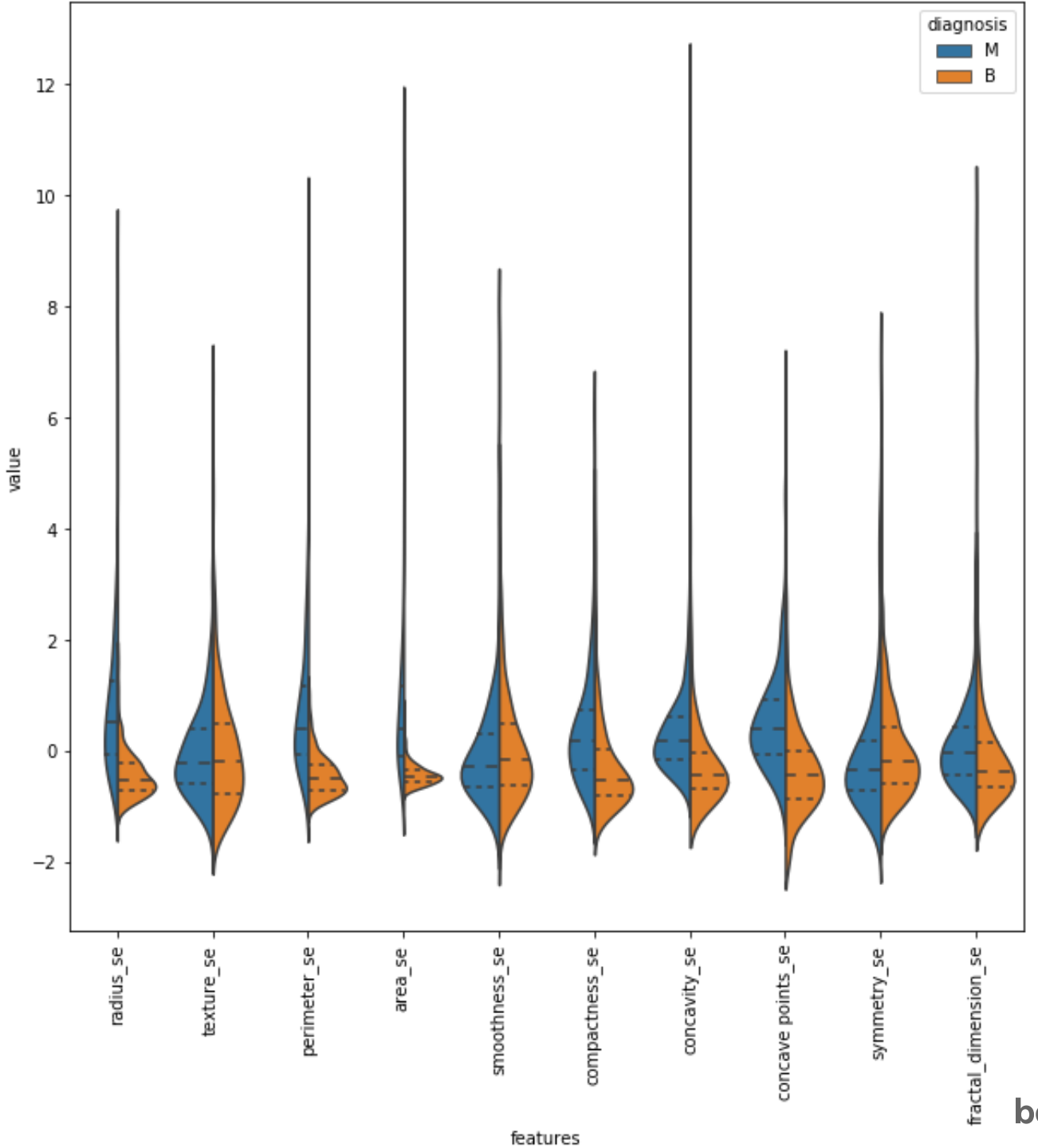
- tumor radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area — 1.0)
- concavity (severity of concave portions of the contour)
- concave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

Onto EDA!

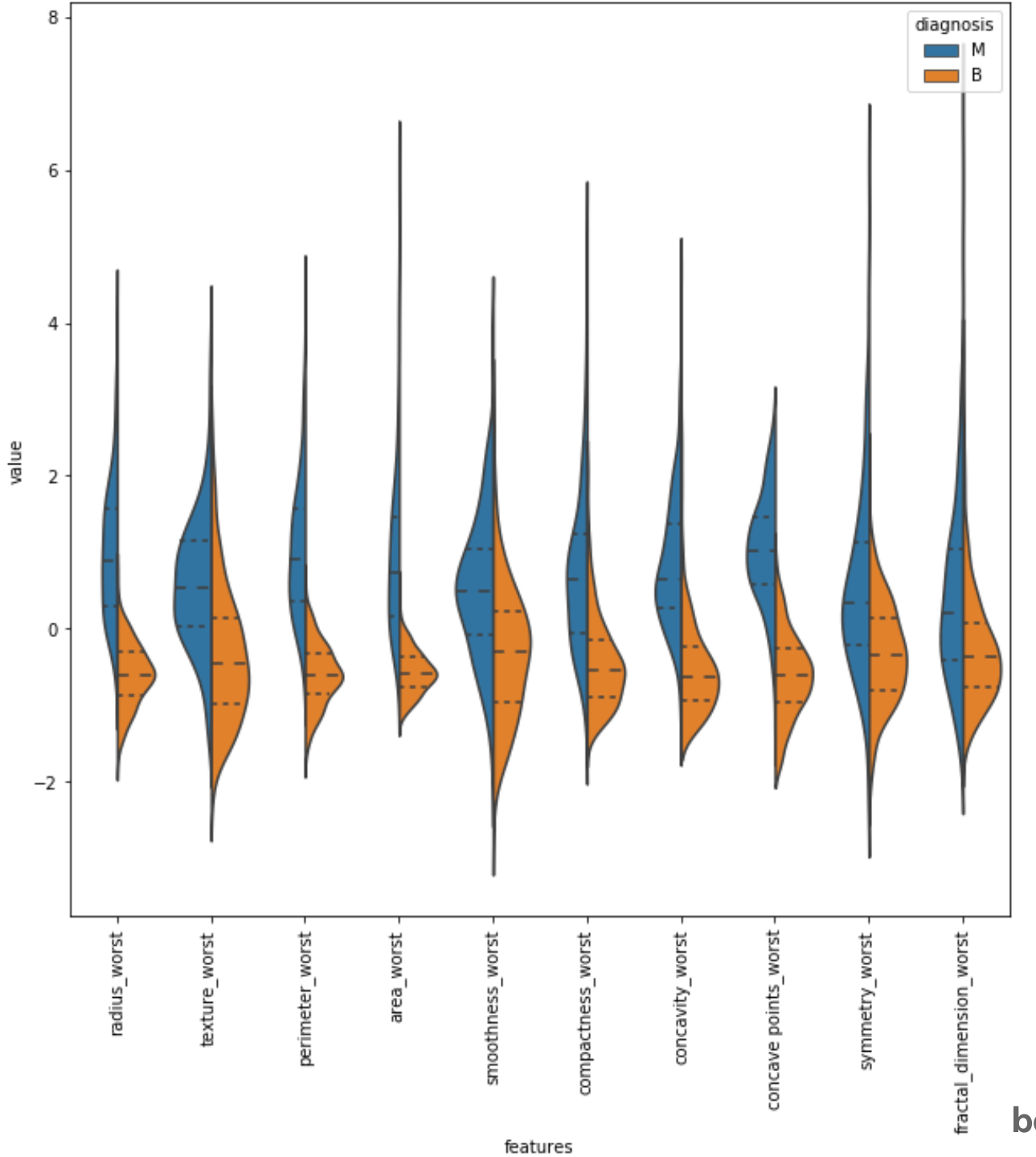
Violin plots for all the means



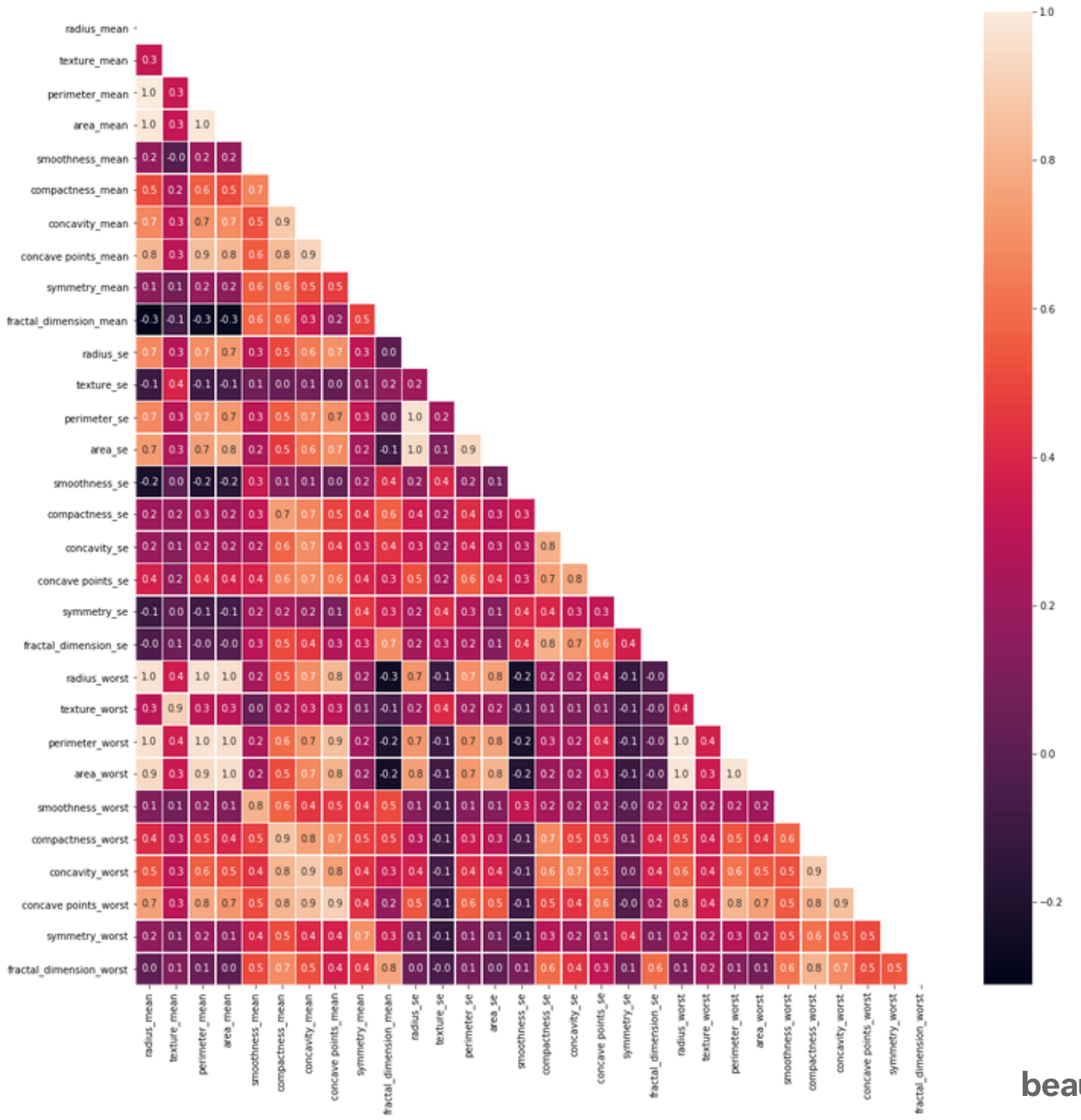
Violin plots for all the standard errors



Violin plots for all the worst dimensions

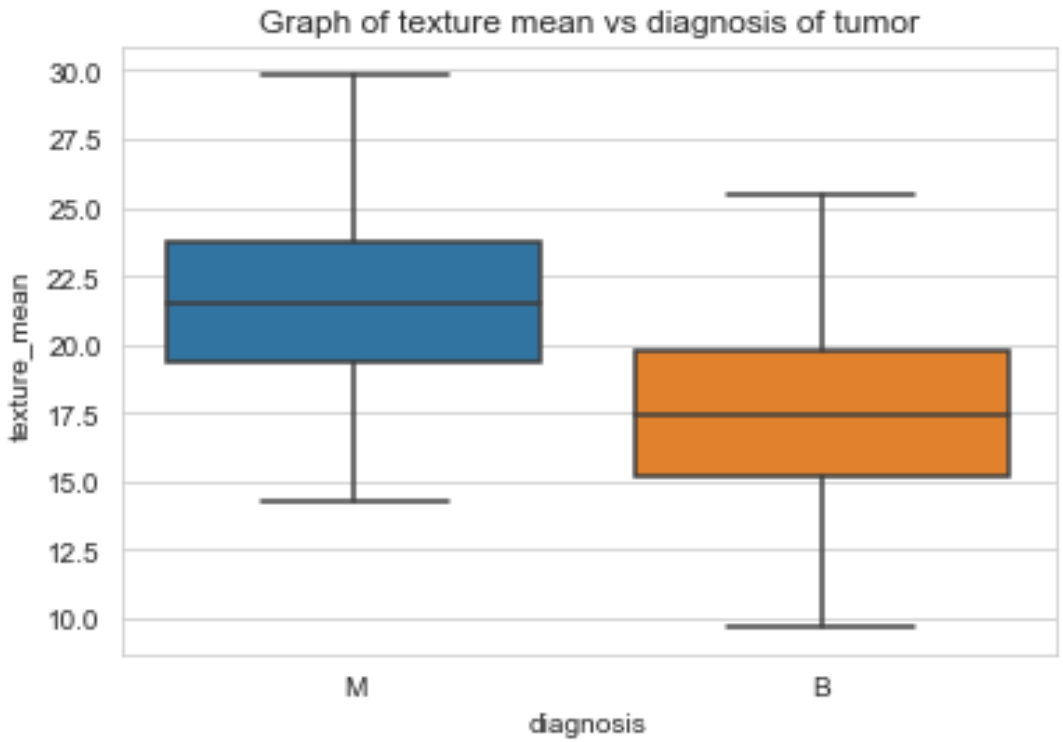


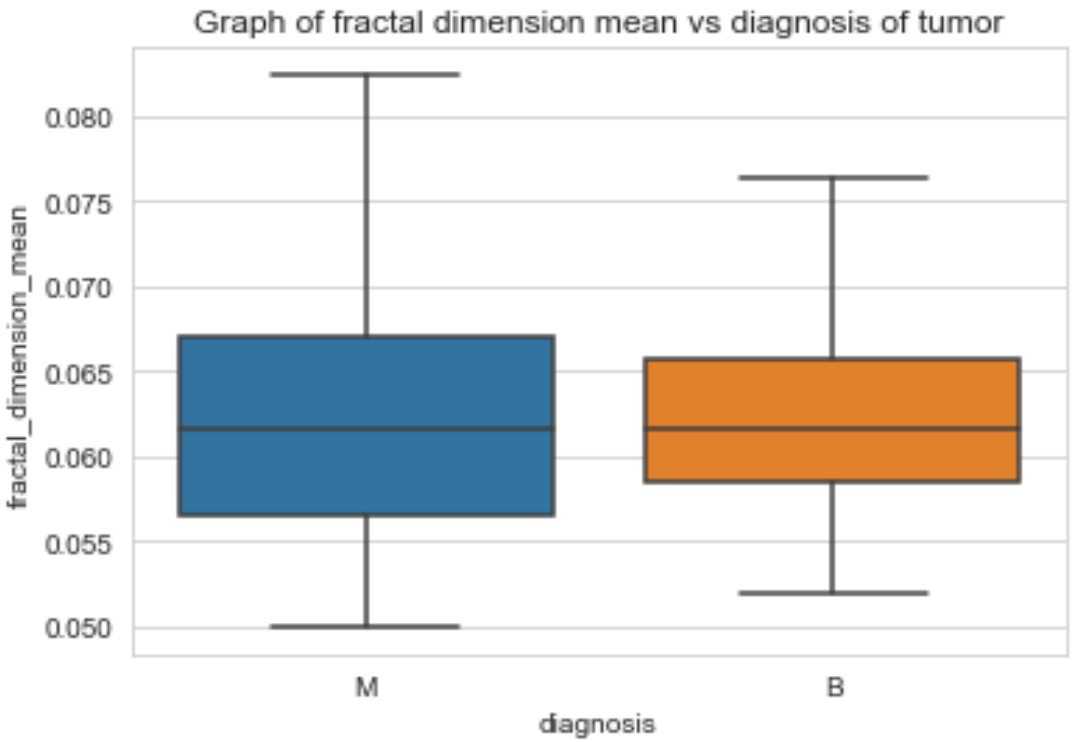
Multi collinearity check via correlation matrix



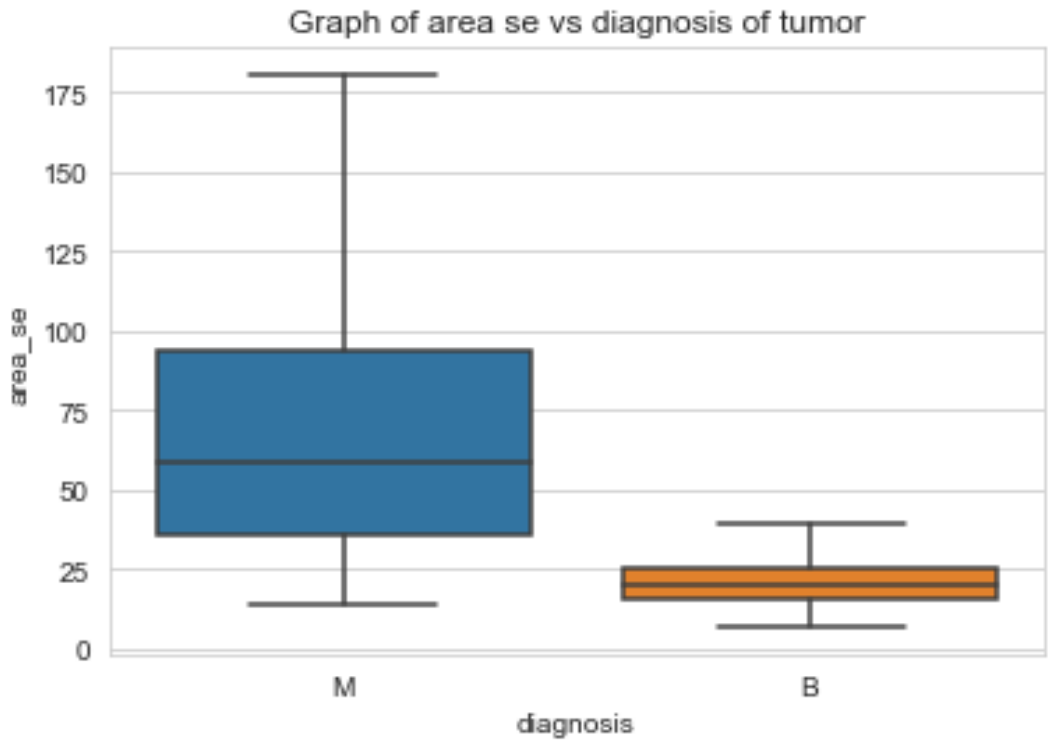
Few insights from the huge correlation matrix

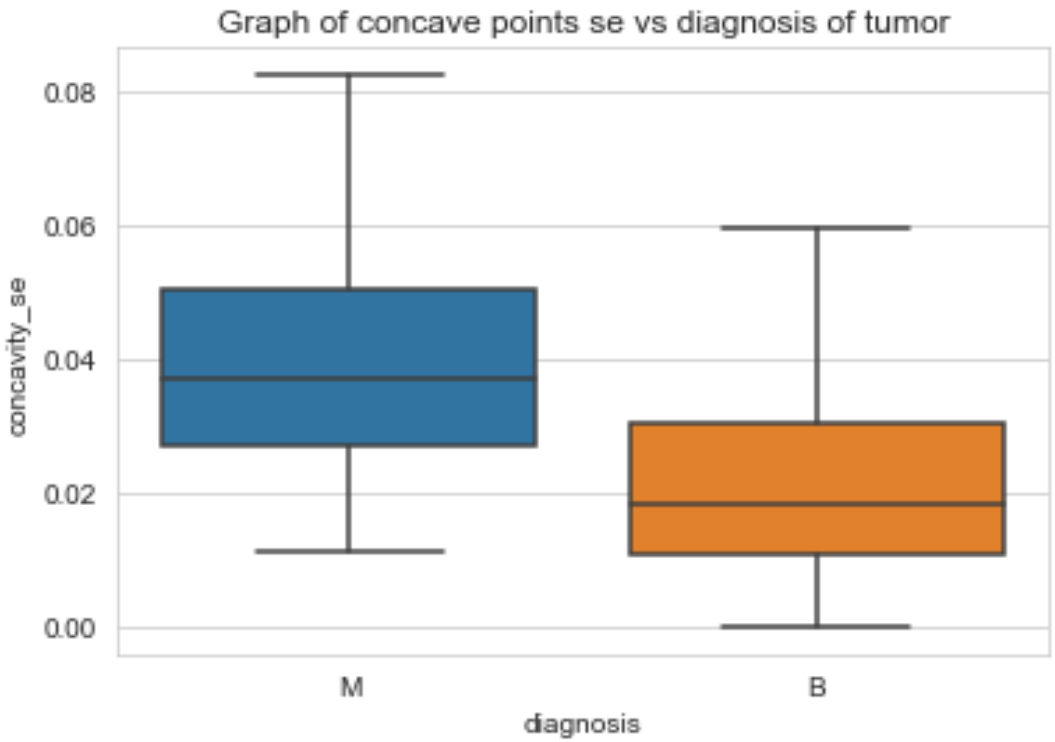
- The means, std errors and worst dimension lengths of compactness, concavity and concave points of tumors are highly correlated among each other (correlation > 0.8).
- The std errors and worst dimensions of radius, perimeter and area of tumors have a correlation of 1.
- texture_mean and texture_worst 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.



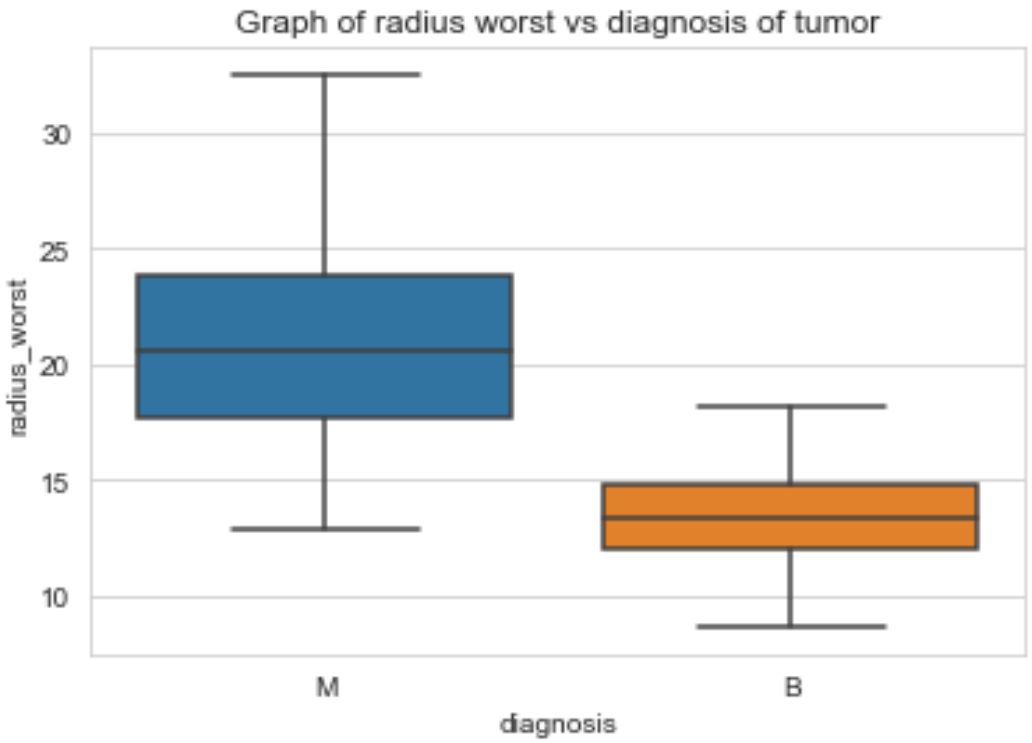


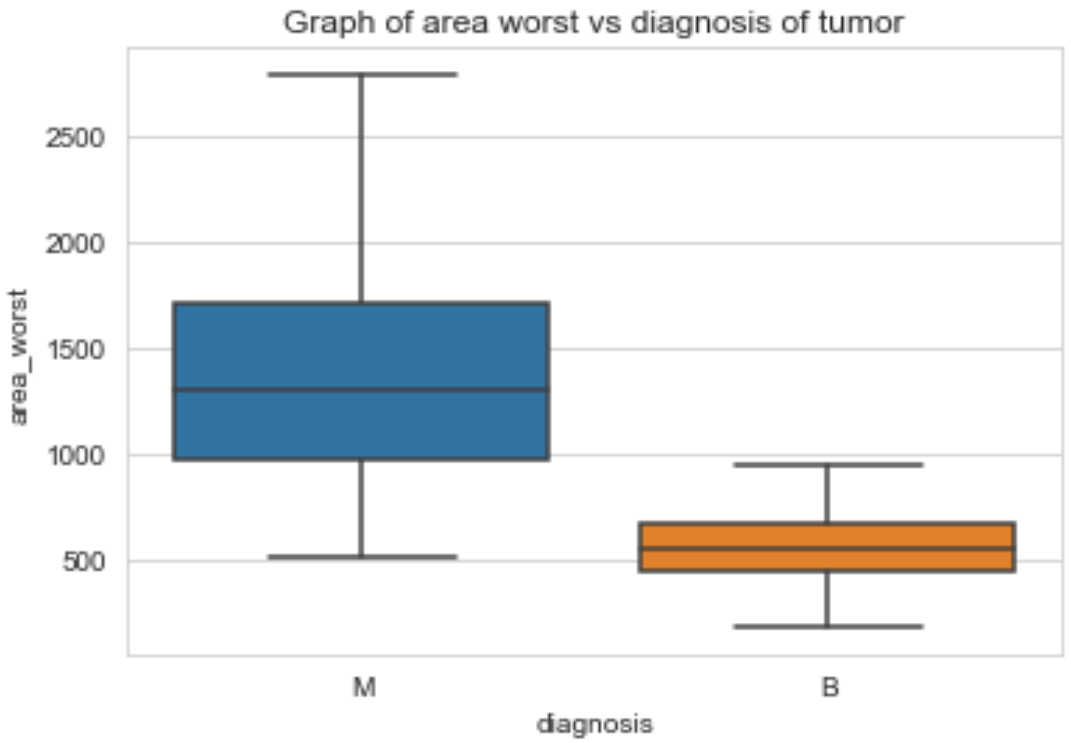
Texture means, for malignant and benign tumors vary by about 3 units. The distribution looks similar for both the groups. Malignant tumors tend to have a higher texture mean compared to benign.
Fractal dimension means are almost the same for malignant and benign tumors. The IQR is widers for malignant 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.



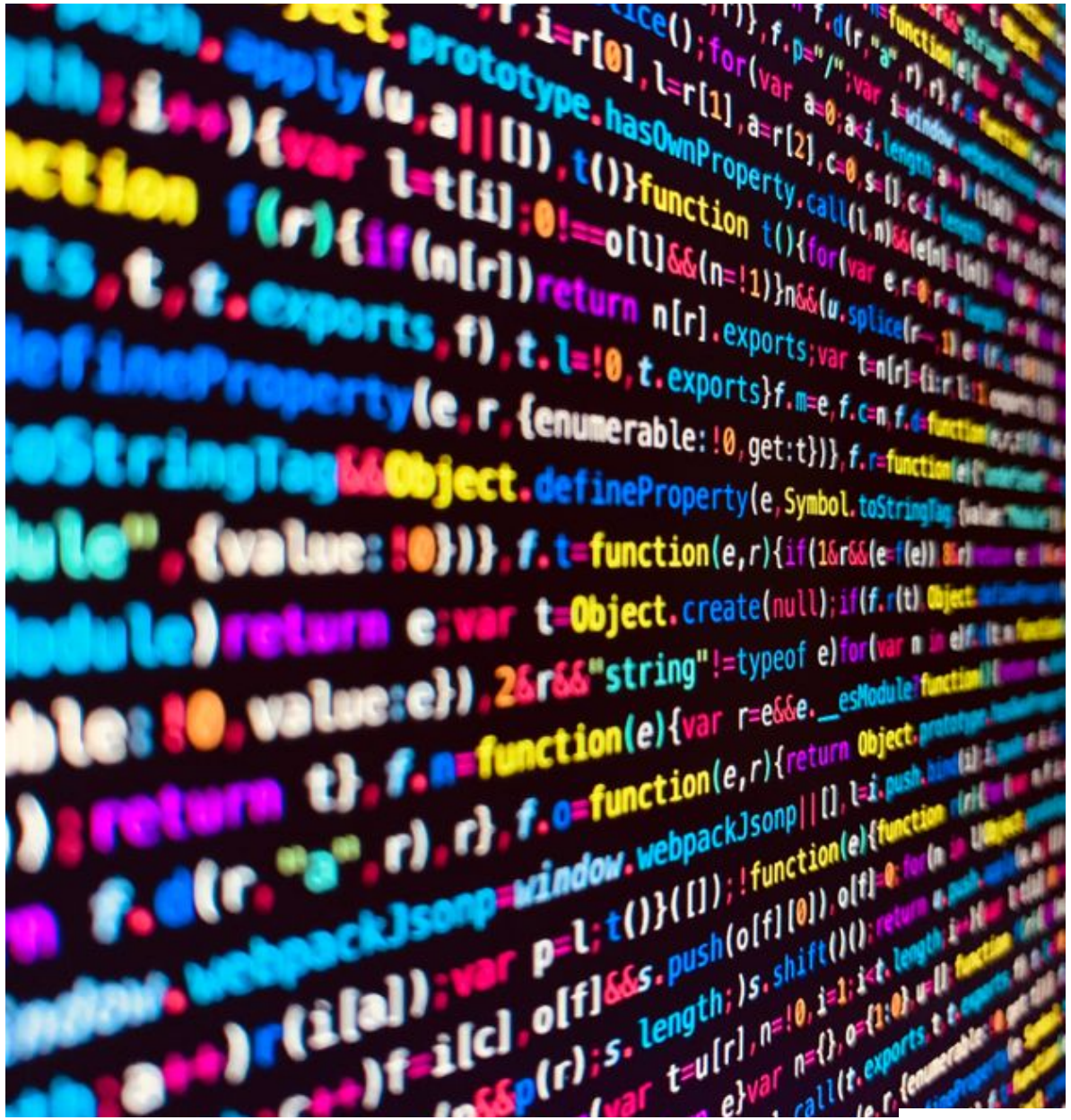


Malignant groups have a wider range of values for radius worst compared to benign groups. Malignant tumors have a higher radius worst compared to benign groups. Malignant tumors tend to have a higher value of area_wosrst mean and wider IQR range. Because of noticeable differences between B and M tumors, this could be a good feature for classification.

Statistical Analysis: t test

Feature	t-statistic	p-value
texture mean	10.86720108000000	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	25.7215903	2.8288477e-97

Onto machine learning!



A brief overview on the ML methodology followed

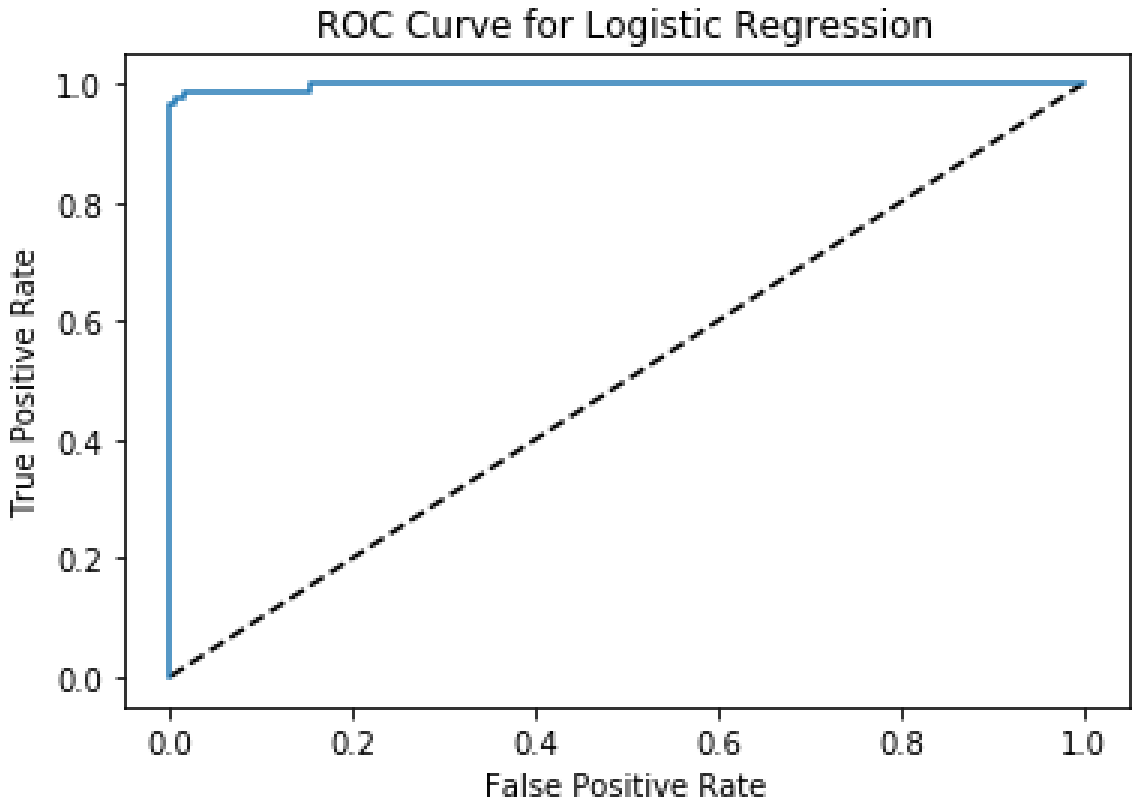
- **Data manipulation:** sklearn's LabelEncoder was used to convert the categorical dependent variable (M or B) of the diagnosis column to a numeric data type.
- **Train Test Split**: sklearn's train_test_split was used to split the dataset into training and test sets. 40% of the data was reserved for 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 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 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

Discussion

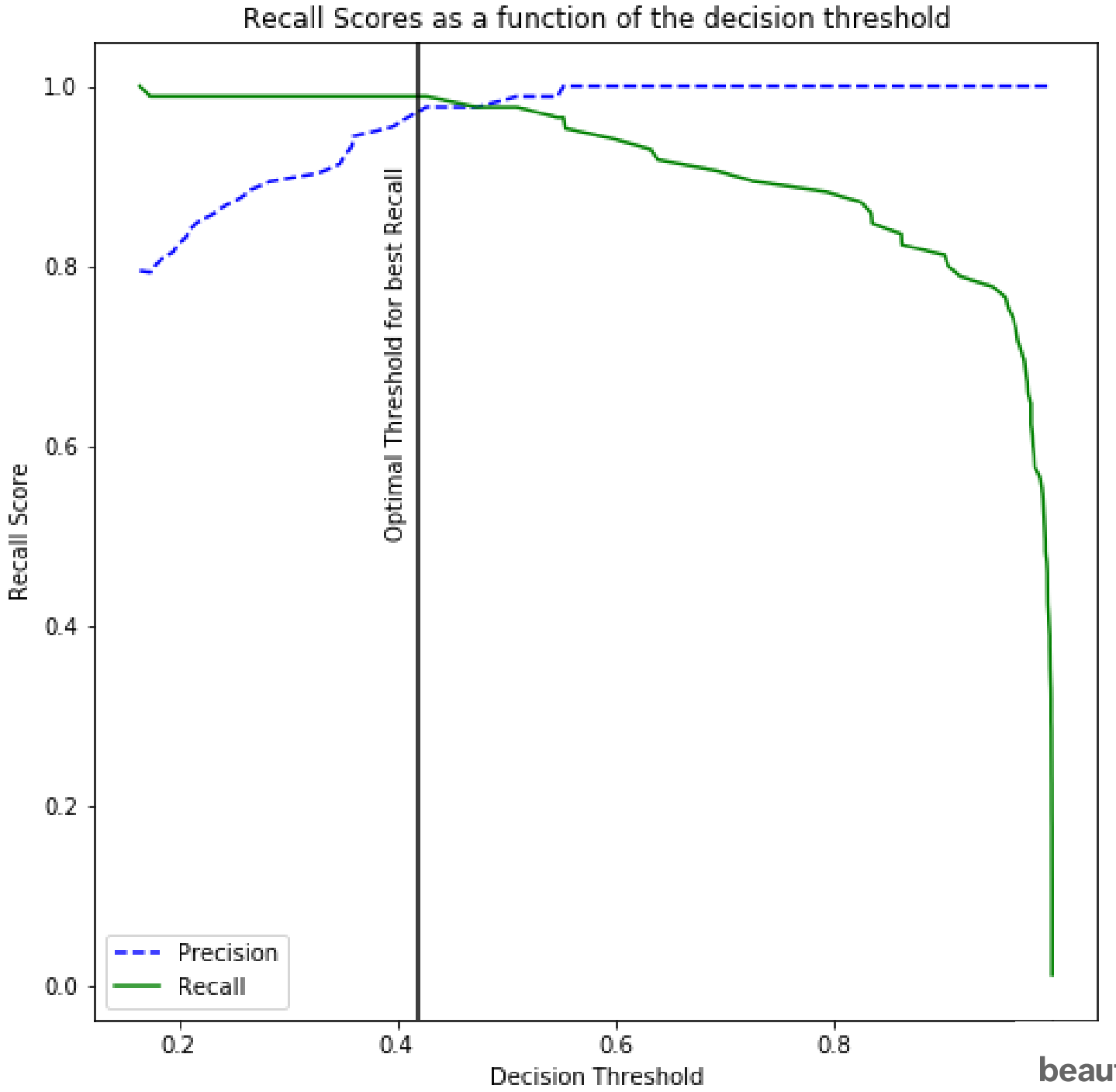
- The Logistic Regression model with l2 as penalty, C = 0.591 and threshold set to 0.42 had an AUC score of 0.99. The model misclassified only 1 tumor as FN and 2 tumors as FPs.
- From the point of view of the patient's health, classifying a malignant tumor as benign is worse than classifying a benign tumor as malignant.
- Therefore, I was focused on getting the least number of FNs for the models, which meant maximizing 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.

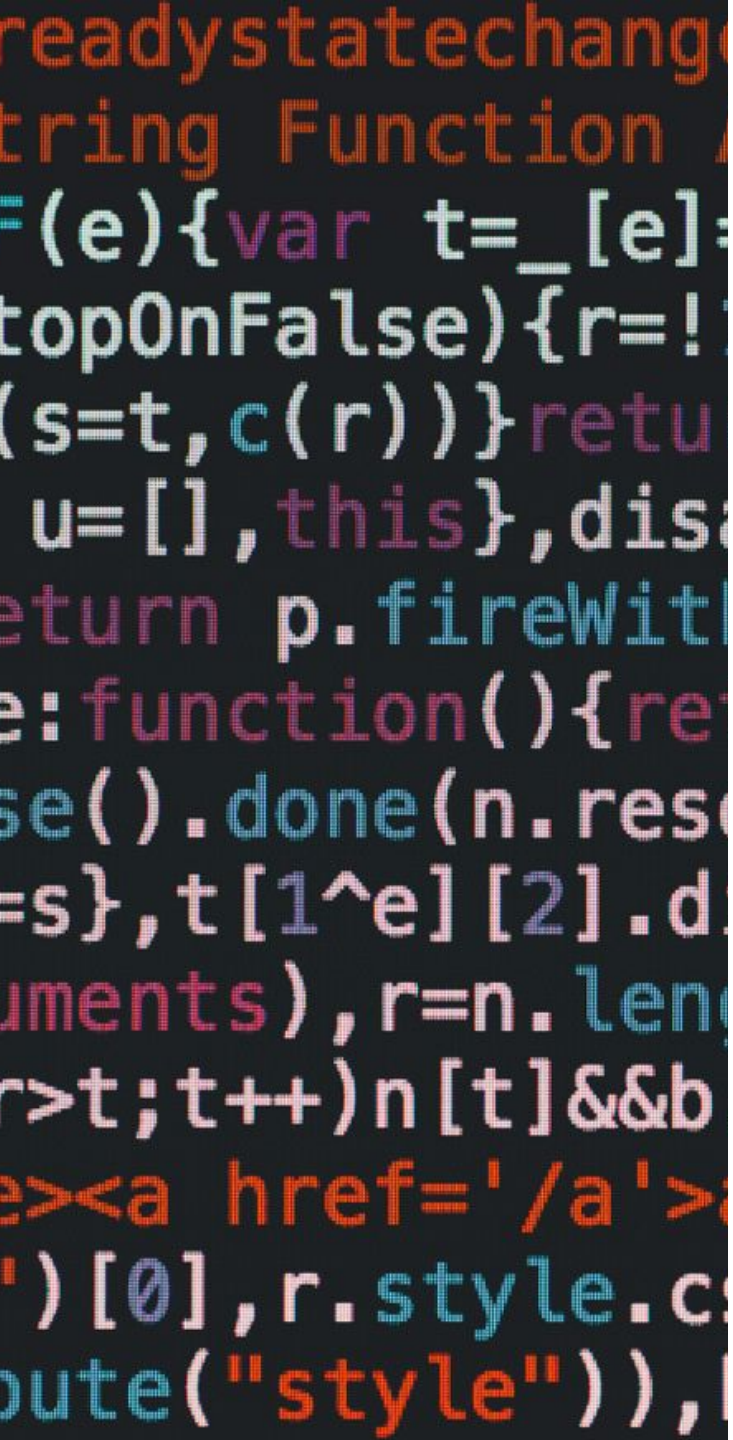


	precision	recall	f1-score	support
0	0.99	0.99	0.99	143
1	0.98	0.99	0.98	85
accuracy			0.99	228
macro avg	0.98	0.99	0.99	228
weighted avg	0.99	0.99	0.99	228

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

A threshold of 0.42 was chosen for maximum recall.

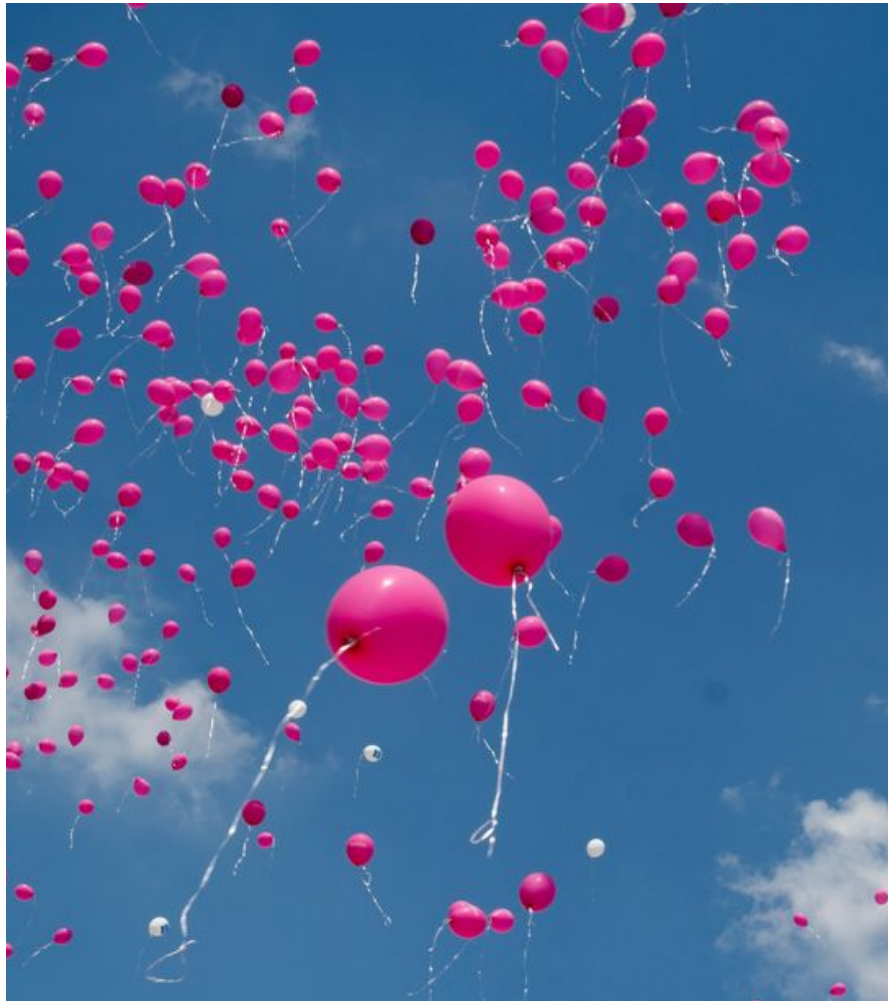




Conclusion

- To conclude my project, I used a variety of algorithms in order to correctly classify tumors as malignant or benign. PS: multicollinearity was not a problem for prediction!
- 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.

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.
- 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!

Thank you!
Questions?