Classification of Breast Cancer Tumours

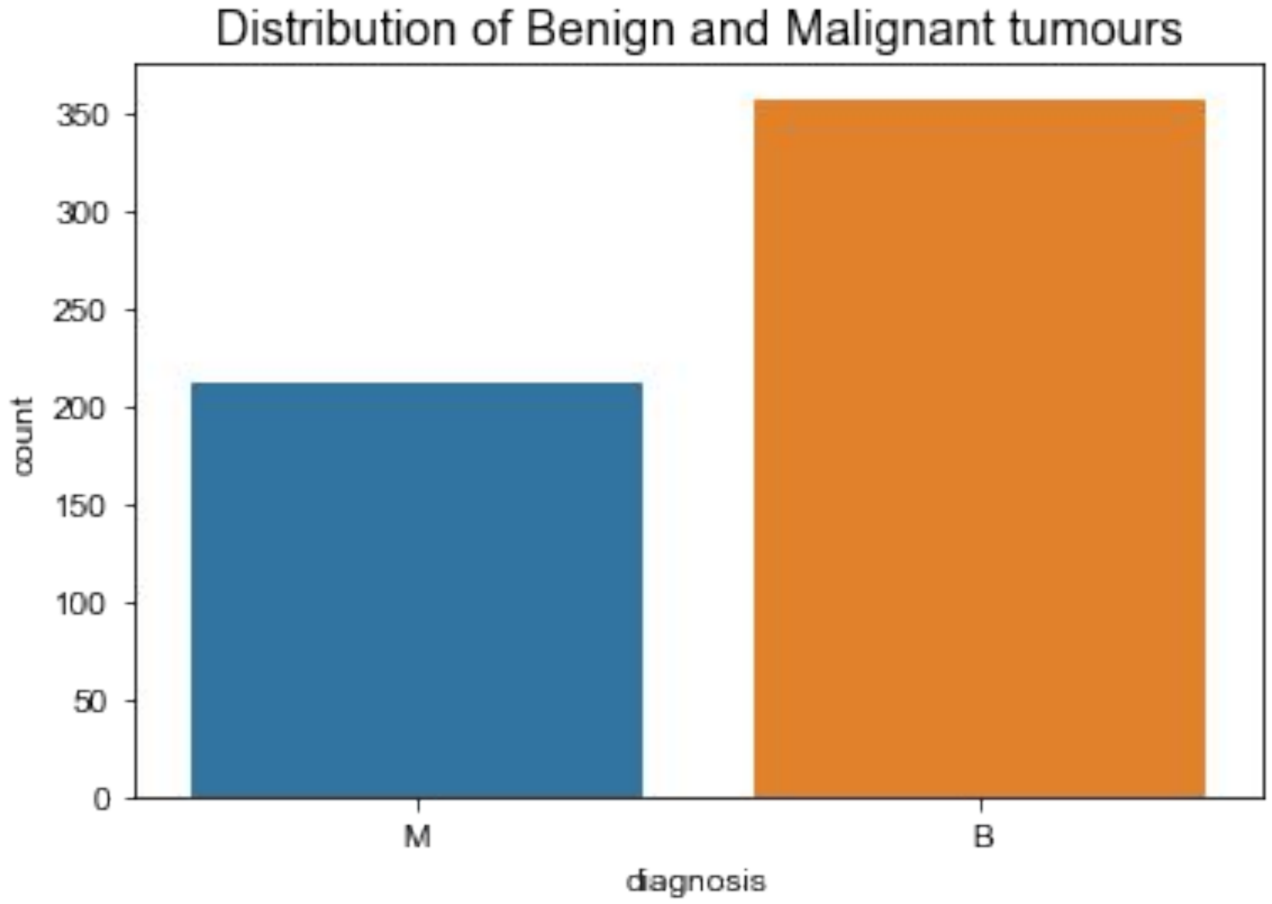
Indrani Banerjee

Data

- Dataset had no missing values and consists of:
  30 features
  Diagnosis Label: M or B
  569 samples

- Uneven Dataset:
  357 Benign Samples
  212 Malignant Samples

- Feature Engineering:
  Malignant Tumours = 1



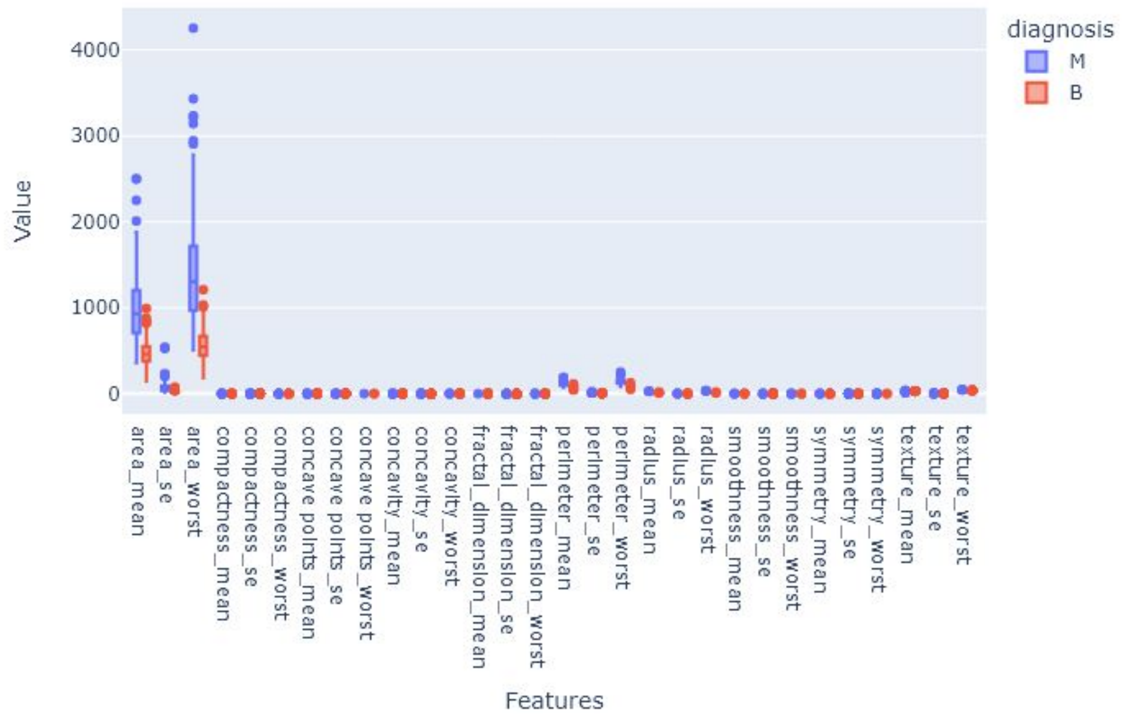
EDA: Outliers I

Top boxplot: all features

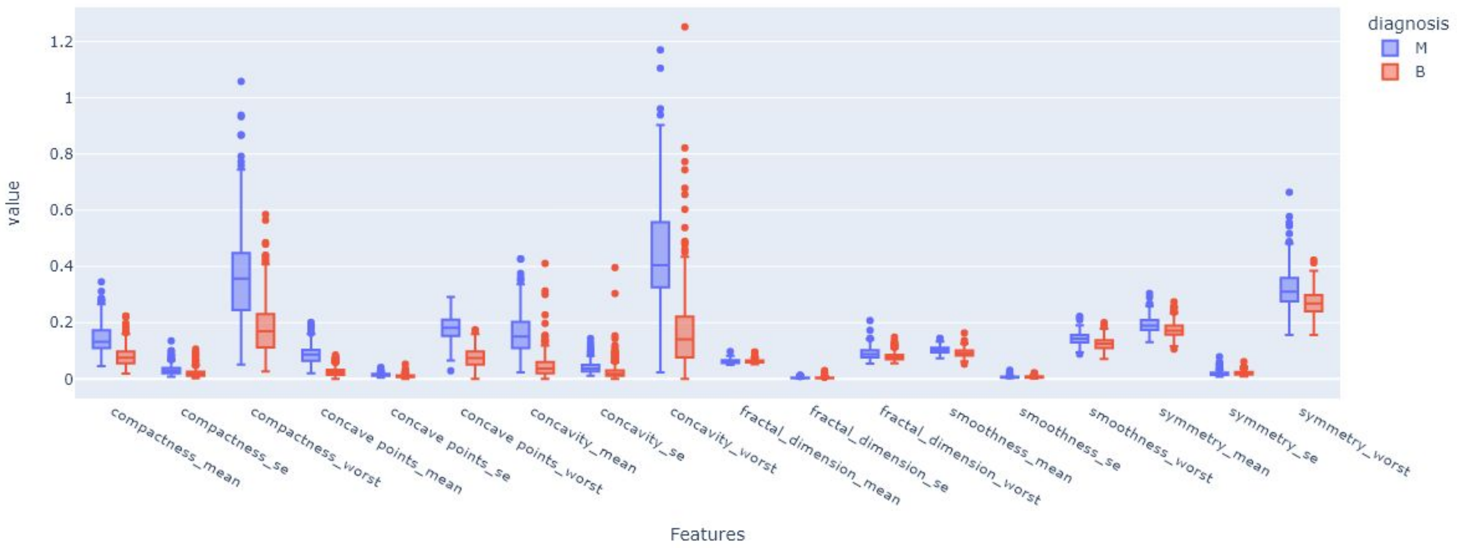
- Highlights the importance of scaling
- ‘area_mean’ and ‘area worst’: values vary greatly between malignant and benign tumours

Bottom boxplot: omission of radius, perimeter and texture features

- Quite a few outliers for all features
- Range of values for malignant tumours are greater than the benign counterpart



Boxplot of features excluding the radii, perimeter, daignosis, and texture features



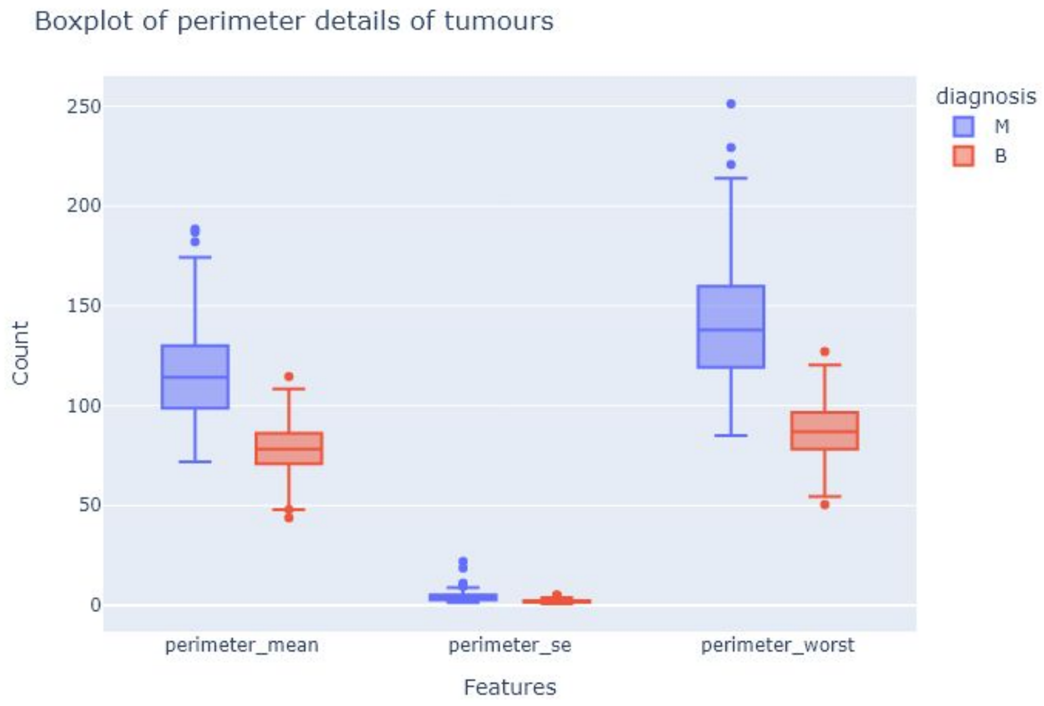
EDA: Outliers II

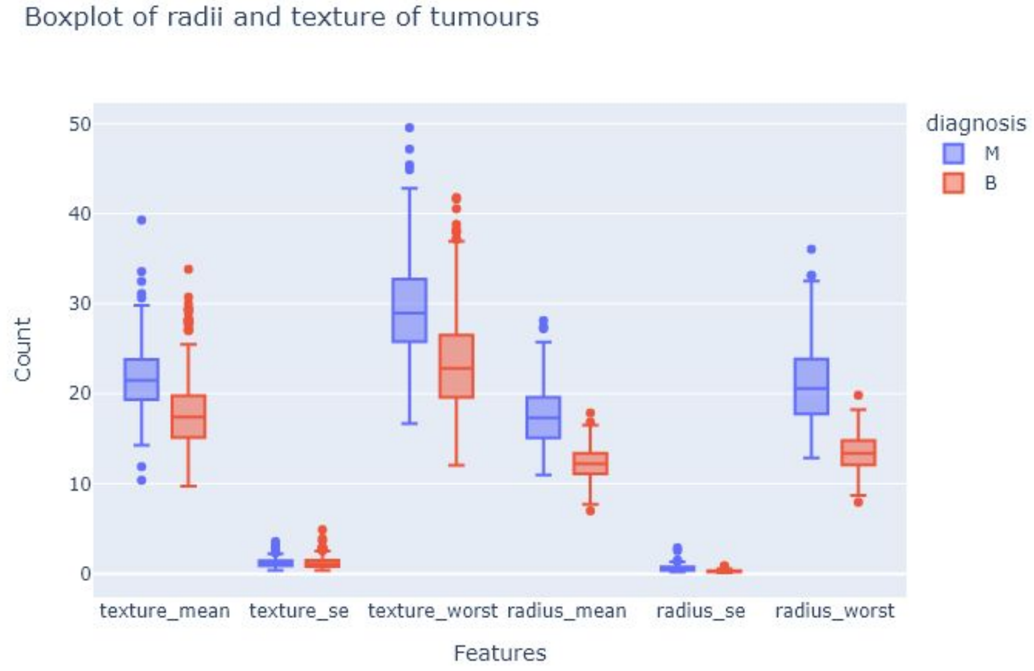
Top boxplot: perimeter

- Echoes greater range for malignant tumours than benign
- Malignant tumours have higher number of outliers

Bottom boxplot: radius and texture

- Mean texture and 'texture worst' have similar values for both malignant and benign tumours
- Values for the radius of tumours have more pronounced difference between the two diagnosis





EDA: Area, Concavity, and Compactness

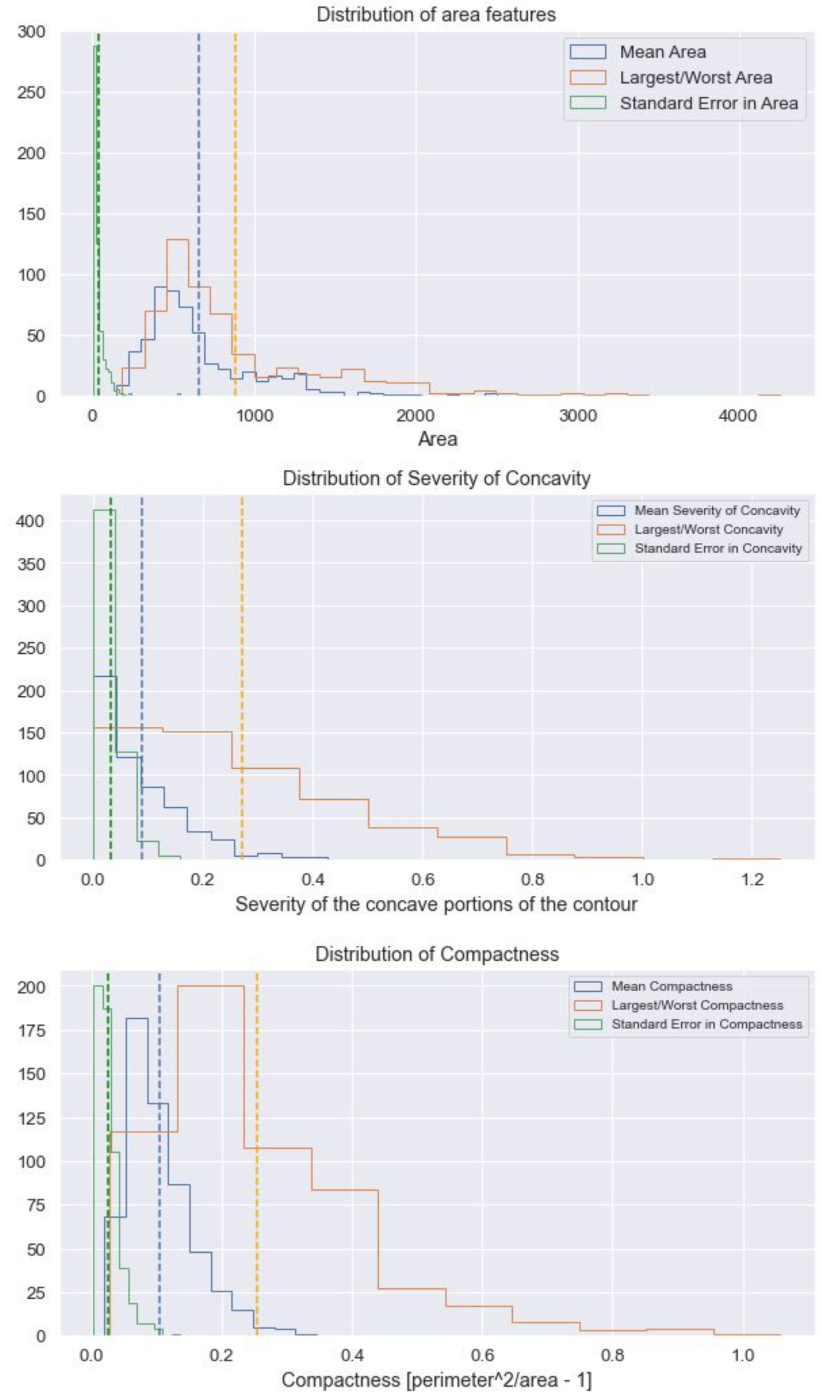
- Area, concavity, and the compactness measures all the samples are positively skewed (right skewed).

- This means that at least 9 out of the 30 features show skewness:

    □ logarithmic scaling is a possible option because of the skewness

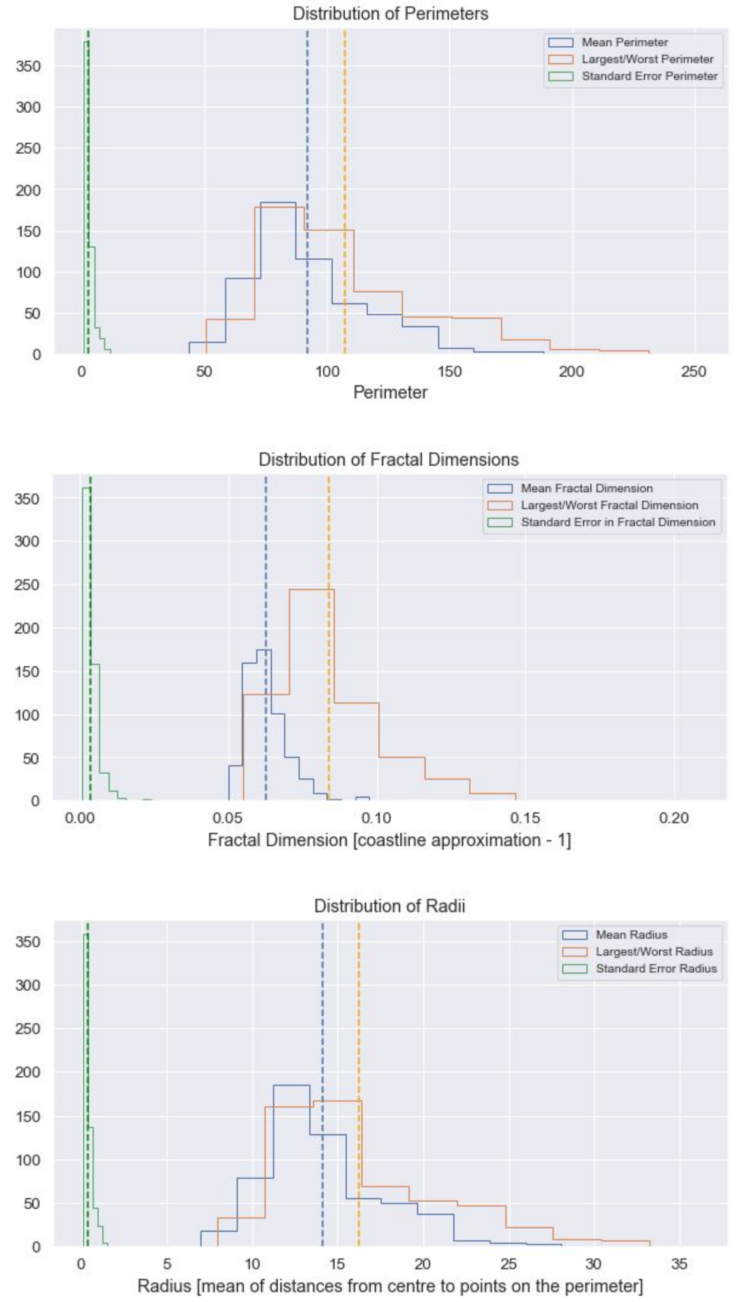
- They have a long tail which illustrates the large range of values we observed in the boxplots.

- The maximum y values range from 200 to 400 confirming again the need to scale these values before modelling.



EDA: Perimeter, Fractal Dimensions, and Radius

- Perimeter, fractal dimensions, and the radius shows somewhat symmetrical distributions
- The same observations are notes in terms of the large range of values
- As these 9 features don't shows significant skew, standardization is a good scaling option

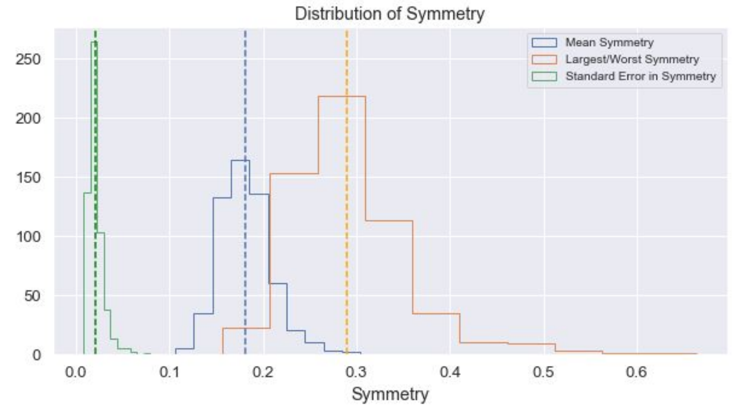


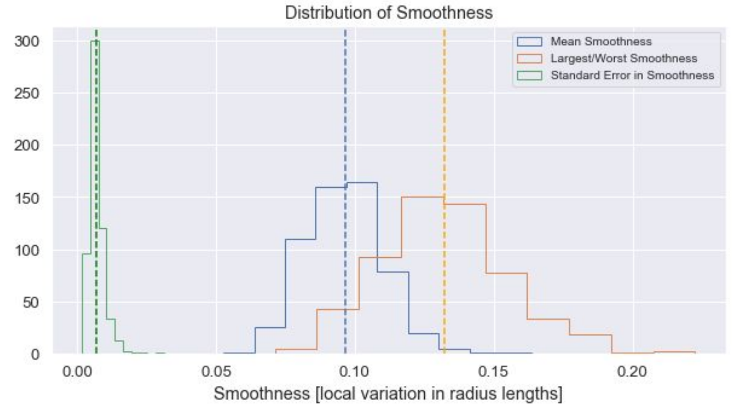
EDA: Symmetry, Smoothness, and Texture

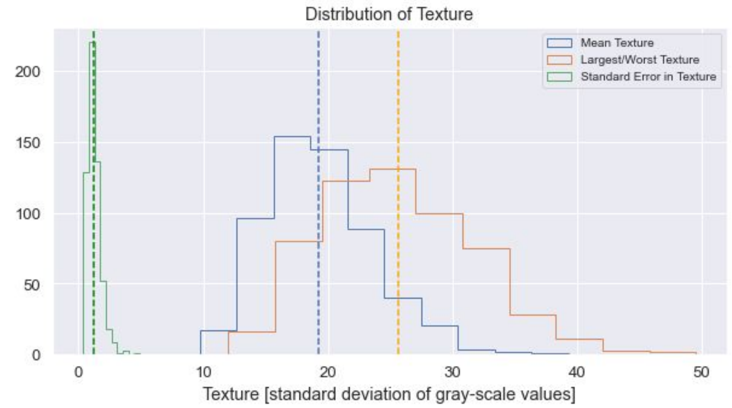
- Symmetry, smoothness and texture show very similar distributions to the perimeter, fractal dimensions, and radius: symmetrical spreads

- The same observations are notes in terms of the large range of values

- As 18 of the 30 features show symmetrical spreads, the StandardScaler is used predominantly for scaling.





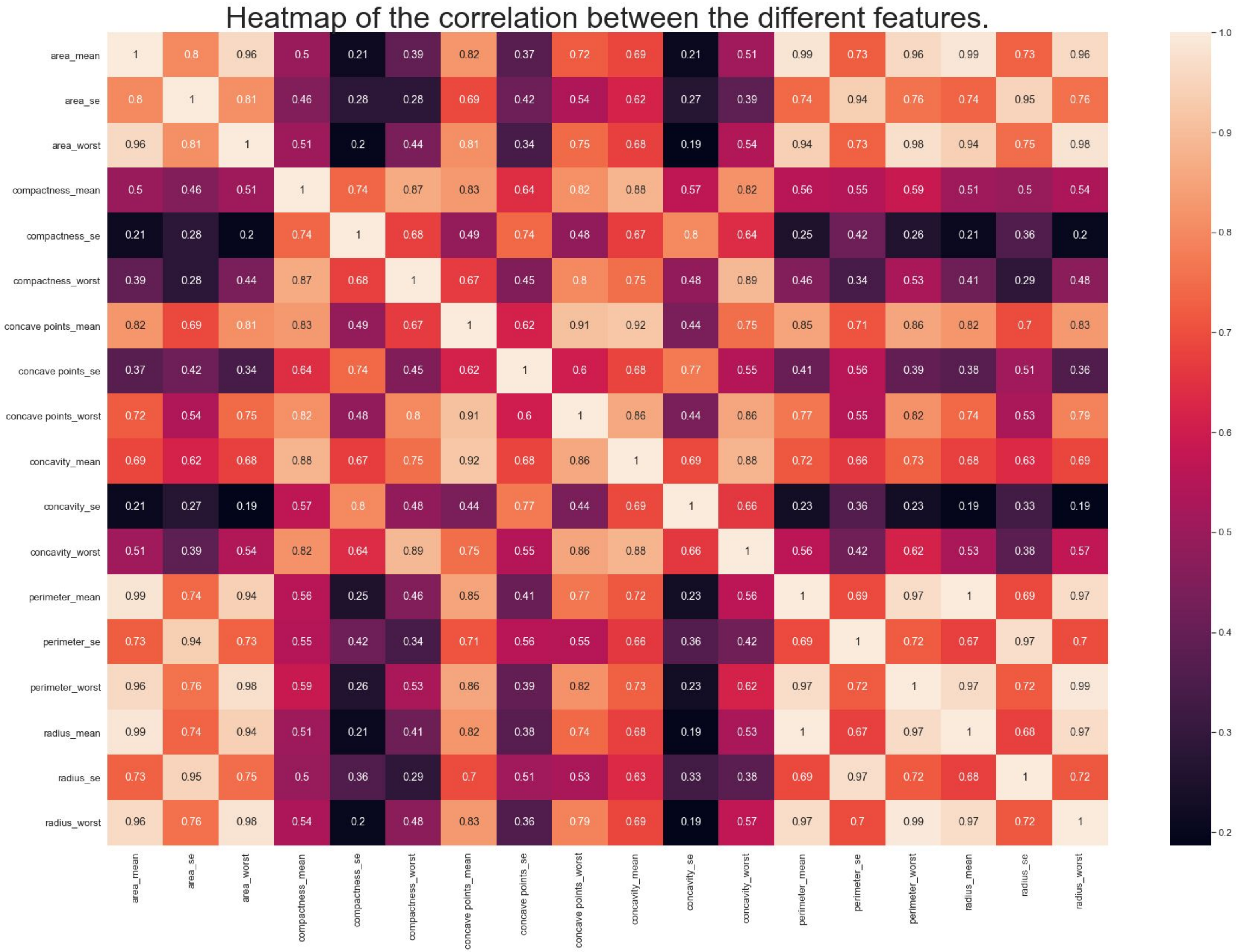


EDA: Correlation I

- The following features were dropped as they had low correlations
  1. Fractal Dimensions
  2. Smoothness
  3. Texture

- The linear correlations between features in the hopes of reducing the number of features

- Aim: identify two or more features that are perfectly correlated and drop one or more of these



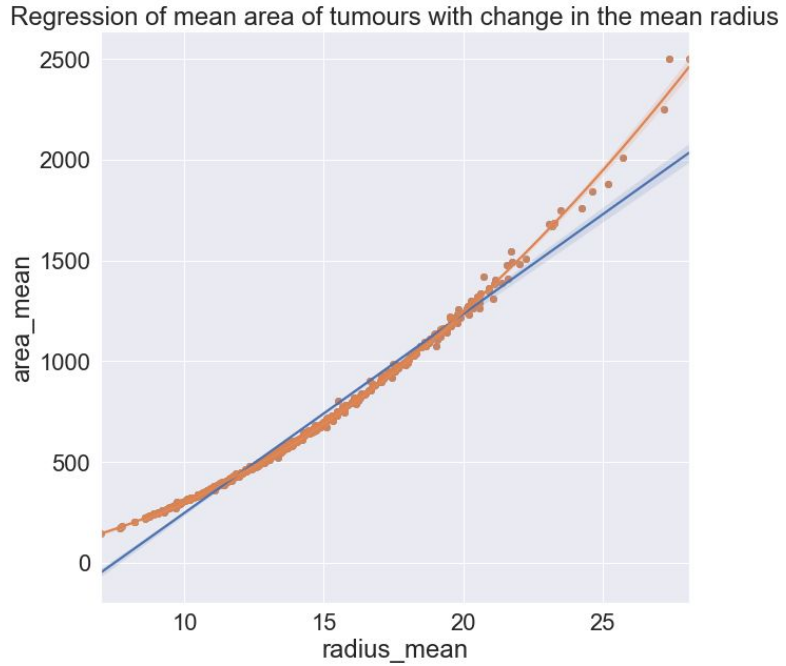
EDA: Correlation II

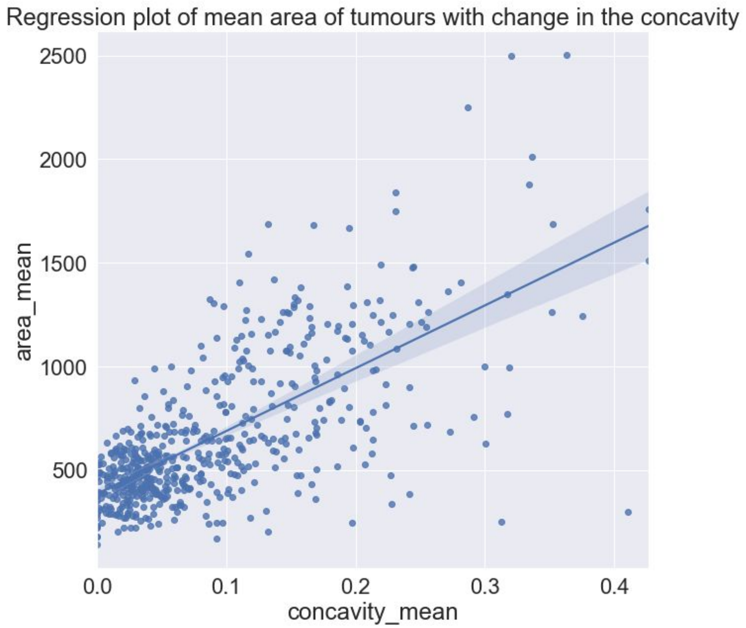
- Pair plots and the heatmap helped identify linear correlations between features such as:
  1. Mean area
  2. Mean perimeter
  3. Mean concavity
  4. Mean radius

- We notice that even features with high Pearson correlation coefficients, when plotted they show polynomial or even an exponential regression:
  - Blue line illustrates a linear fit
  - Orange line shows a much better fit with a non-linear regression model

- The pair plots also highlight that the ‘worst’ features of the mean, radius, perimeter, concavity, and compactness didn’t any clear linearity: they showed similar spread to area mean vs concavity mean

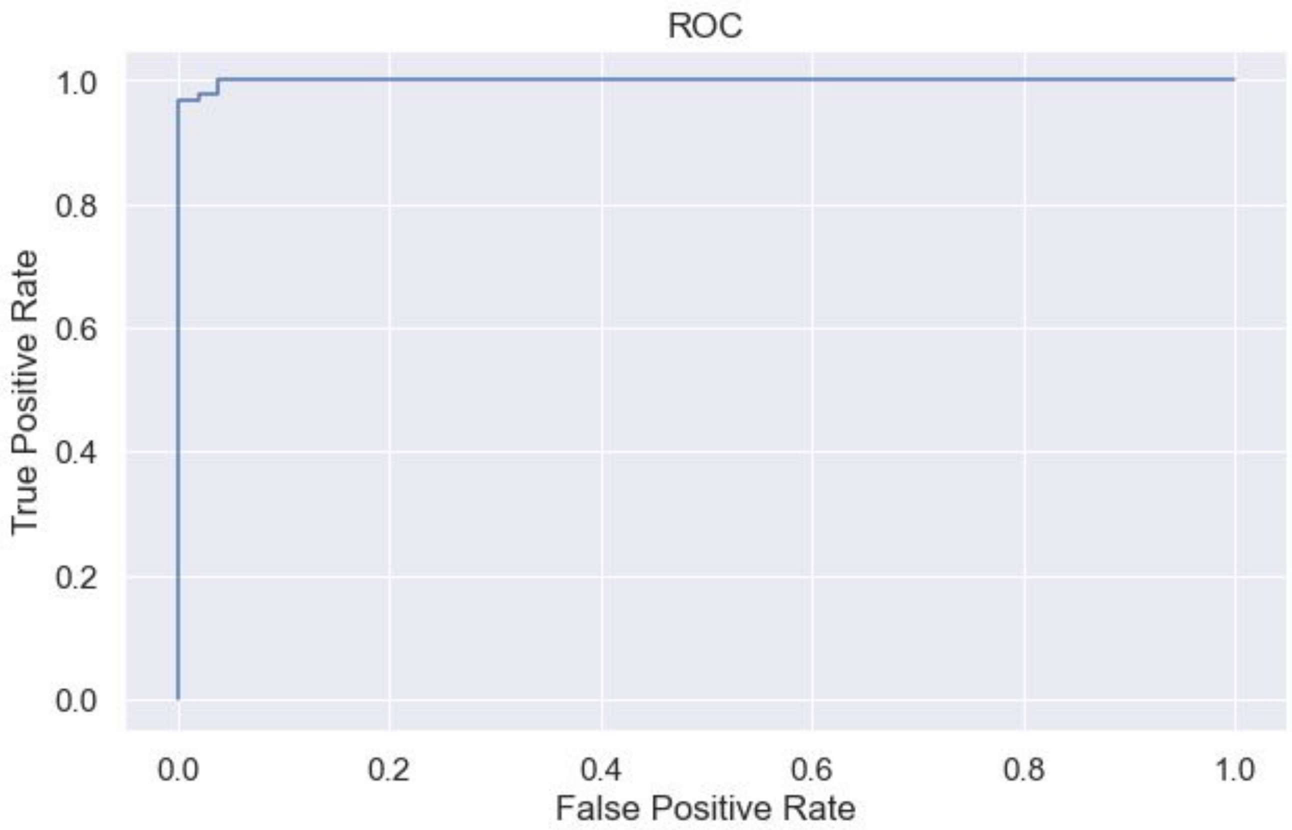
- Conclusion: Linear Regression Model for mean area gave $R^2 = 0.439$ so it was decided to not drop any features at this time.





Modelling: Logistic Regression Model

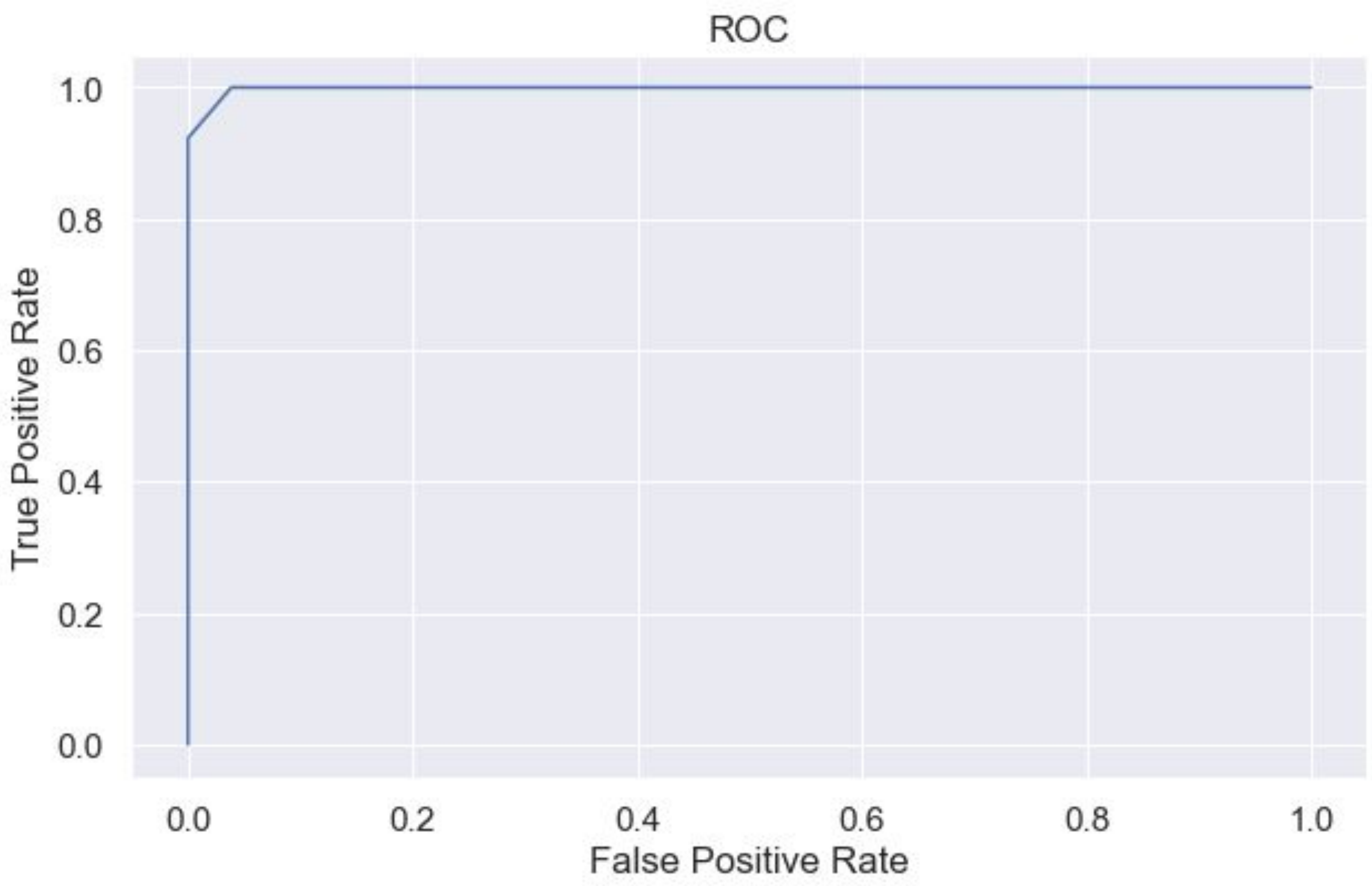
- Two logistic regression models were created with and without scaling: scaled model performed best.
- Table (on the right) summarizes parameters of the best model
- Ridge Regression was determined most suitable through GridSearch: this implies that perhaps quite a few of the parameters have similar influence on the classification



Scaling	Parameters	MCC	AOC
Standardized	C = 0.01 Penalty = l2 Solver = liblinear	0.955	0.999

Modelling: K Nearest Neighbor

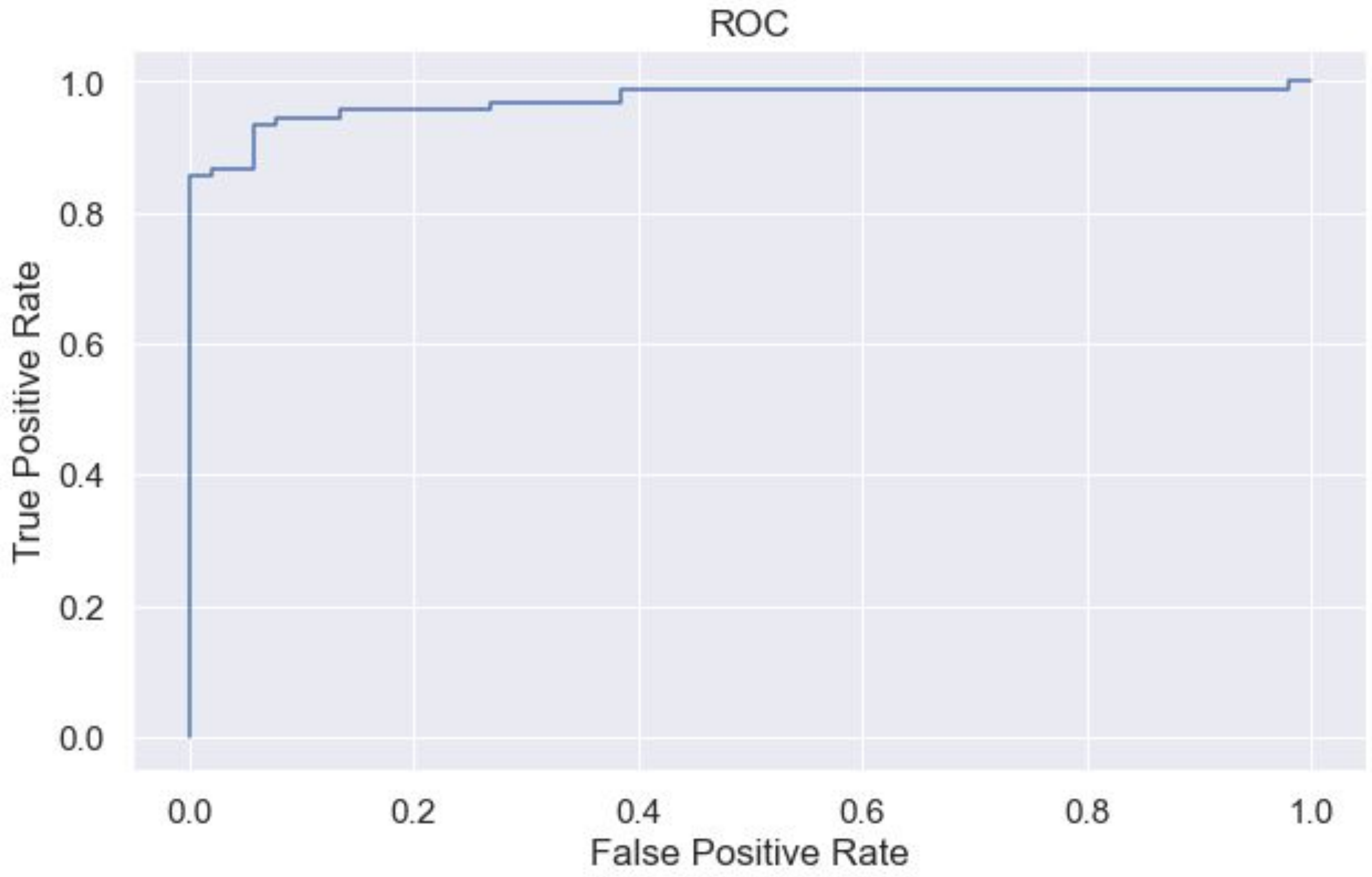
- 7 K Nearest Neighbor models were developed with varying number of neighbor values.
- StandardScaler was used to standardize the dataset
- GridSearch with 10 fold cross validation determined 3 as the best number of neighbors, and the optimal hyperparameters are stated in the table on the right.



Scaling	Parameters	MCC	AOC
Standardized	Leaf size = 1 N neighbors = 3 p = 2	0.97	0.999

Modelling: SVC

- 4 Support Vector Classifier models were developed with and without scaling.
- The best model was determined for non-scaled data using Grid Search with 10-fold cross validation.
- Logarithmic scaling and standardization was used but neither of them yielded a model with high area under the ROC.



Scaling	Parameters	MCC	AOC
none	C = 1 Gamma = 0.0001 Kernel = rbf	0.849	0.971

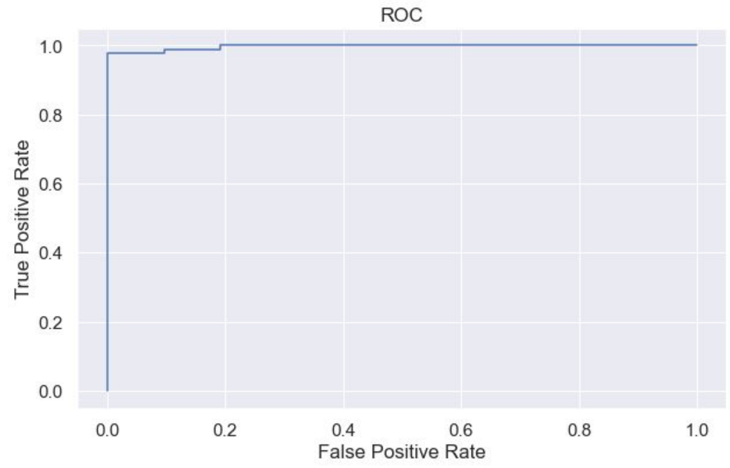
Modelling: Random Forest Classifier

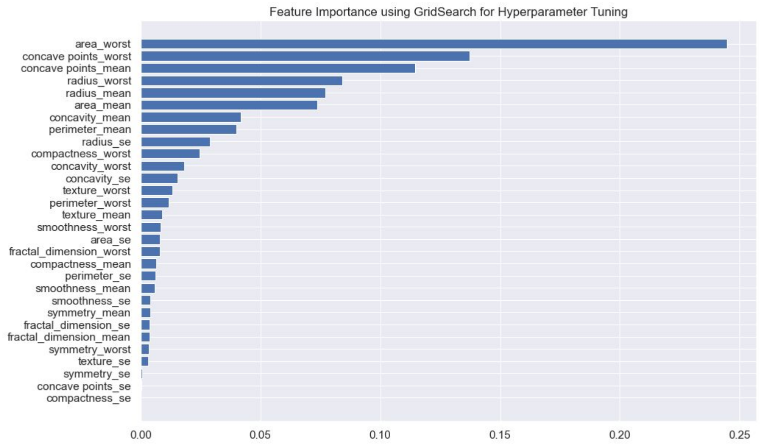
- 6 different Random Forest Classifiers were developed, and hyperparameters were tuned using Grid Search and Bayesian Optimization

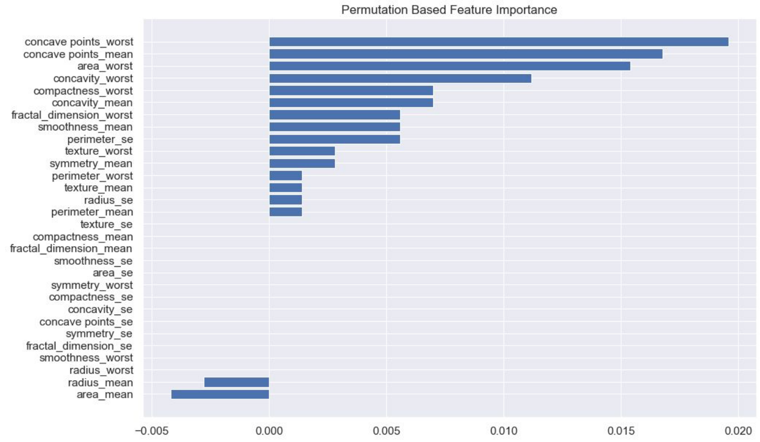
- Feature Importance were identified and the models were redone by dropping the features with less than 1% importance:
  - MCC improved from 0.940 to 0.954
  - AOC improved from 0.997 to 0.998
  - False positive rate fell below false negative rate

- All features were decided to be kept for the best model as it ensured higher false positive rate than false negative (details below)

Parameters	MCC	AOC
n estimators = 23 Max depth = 4	0.940	0.997







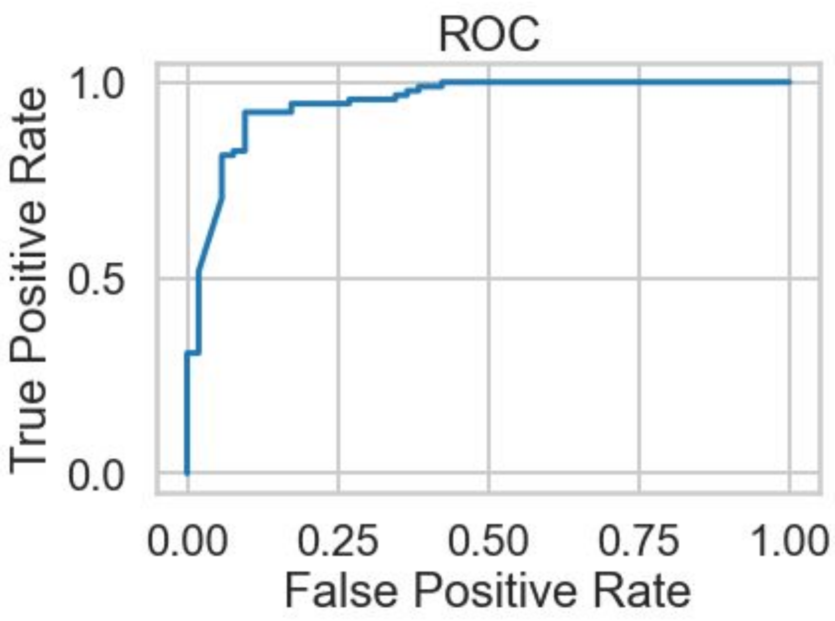
Modelling: Gradient Boosting Classifier

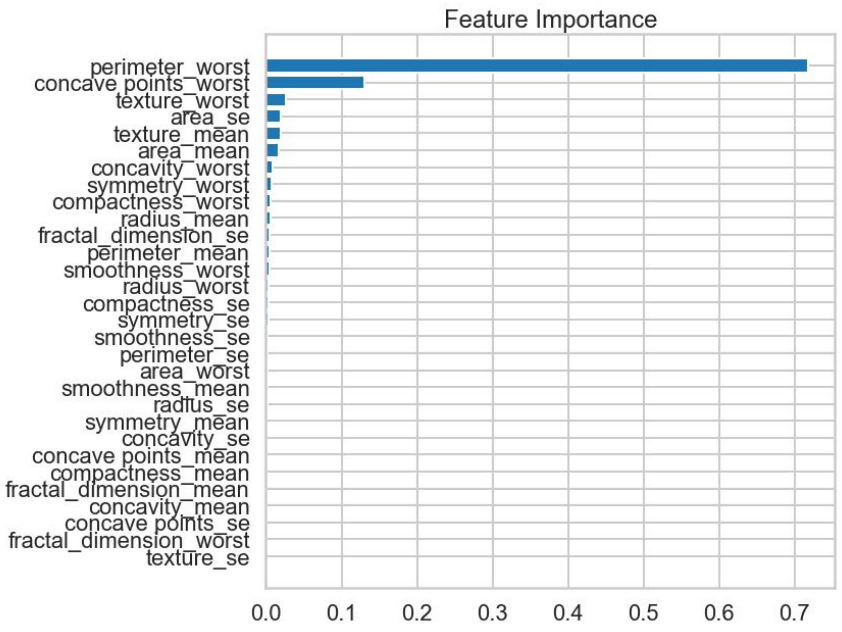
- Hyperparameter tuning had mixed results:
  - improved the MCC from 0.949 to 0.95
  - AOC dropped from 0.995 to 0.949

- Concave points ‘worst’ comes up as an important feature which concurs the findings from the random forest classifier

- Perimeter worst is also identified as an important feature

Parameters	MCC	AOC
learning_rate = 0.088, max_depth = 8, n_estimators = 471, subsample = 0.9533	0.956	0.949





Modelling: XGBoost

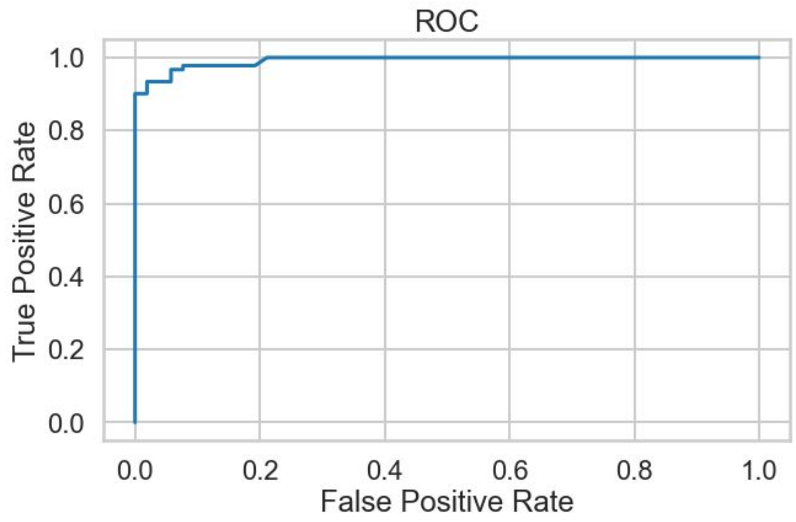
- Hyperparameter tuning had mixed results:
    - Reduced from 0.955 to 0.894
    - AOC dropped from 0.997 to 0.992

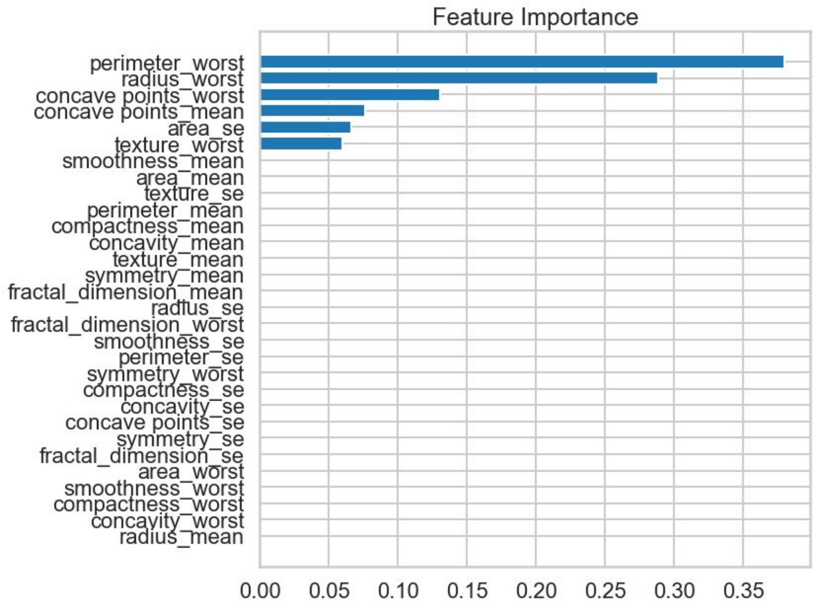
- Hyperparameter tuned model had 0.98 recall, and also had higher false positive than false negative rates, so deemed the best model (details below)

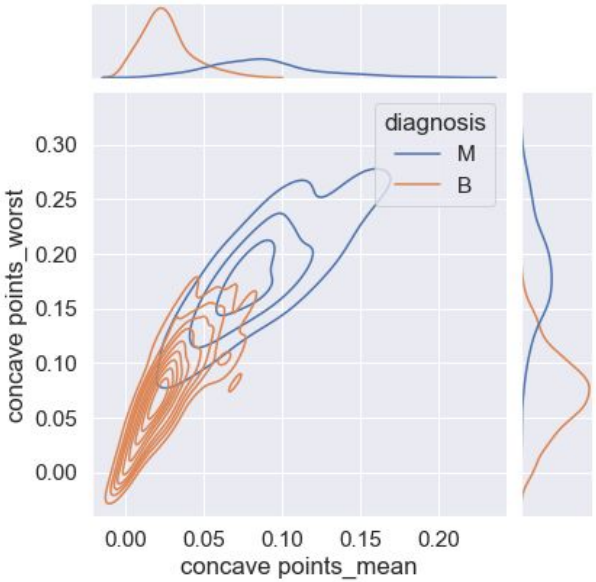
- Radius, concave points and perimeter are again identified as important features.

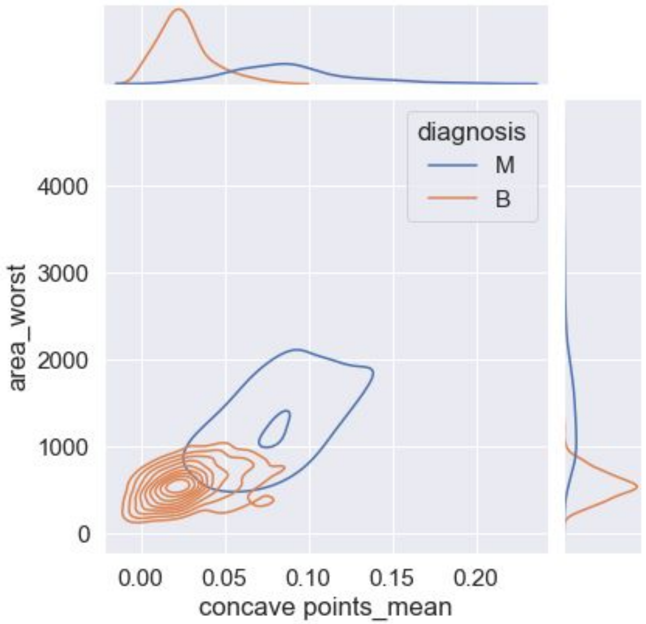
- Perimeter worst is also identified as an important feature

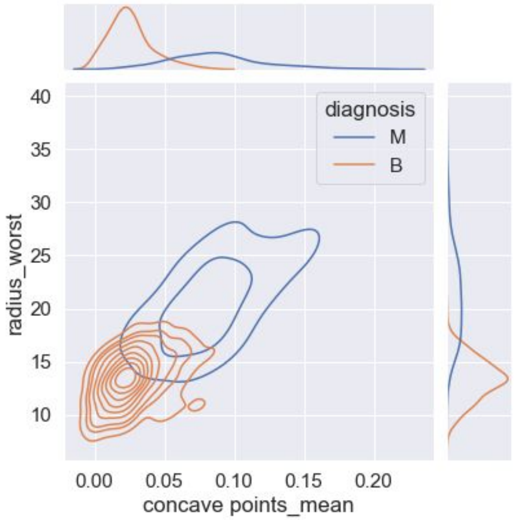
Parameters	MCC	AOC
reg lambda= 3, reg alpha = 0.5, objective = 'reg:squarederror', N estimators= 500, min child weight= 15, max depth= 10, learning rate = 0.1, gamma = 3, colsample_bytree= 0.8	0.894	0.992

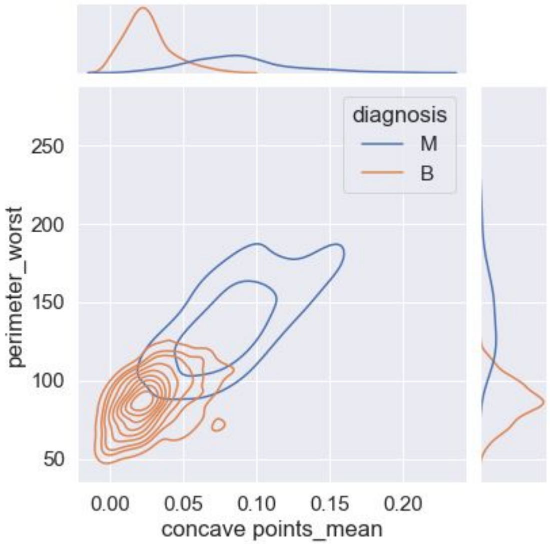












Important Features

- Mean Concave Points are plotted the four other important features (four plots on the left)
- the bunched-up orange contours illustrate that benign tumours have clear range of 'safe' values
- from the blue, spreadout contour lines it can be inferred that malignant tumours can vary in sizes: the lower range of malignant values are higher than the means of benign values for each feature

Conclusions

- Best Model: KNN with StandardScaler
- Parameters: Leaf size = 1 ; N neighbors = 3 ; p = 2
- Confusion Matrix

$$\begin{bmatrix} 50 & 2 \\ 0 & 91 \end{bmatrix}$$

Classification Report	Precision	Recall	F1 Score	Support
Benign	1.0	0.96	0.98	52
Malignant	0.98	1	0.99	91
Accuracy			0.99	143
Macro Avg	0.99	0.98	0.98	143
Weighted Avg	0.99	0.99	0.99	143