Classification of Breast Cancer Tumours

Problem Identification

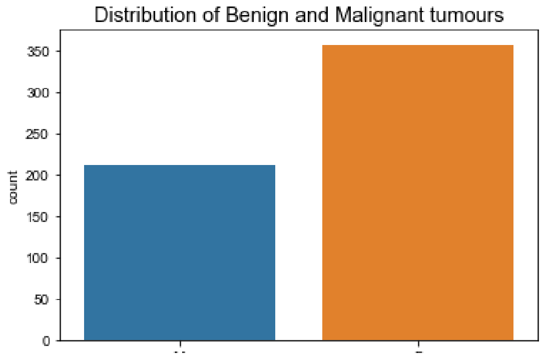
The Wisconsin Breast Cancer dataset gives us 30 attributes of tumours of the breast and has labels for malignant and benign tumours. The aim of this project is to develop a classification model which will:

1. Correctly classify tumours as malignant or benign with an accuracy score of 90% or above.
2. Have a higher false positive rate compared to a false negative rate.
3. Correctly classify tumours with reduced number of features.

As medical diagnosis rarely depends on one type of testing, a 90% accuracy score and a higher false positive rate alongside alternative metrics should provide medical professionals with a reliable model that reduces number of patients who are falsely classified with benign tumours. Furthermore, as Covid-19 has stretched healthcare systems across the globe with backlogs that are predicted to last for the coming few years, models which rely on fewer features might help speed up the classification process reducing additional pressures on medical staff.

Data Wrangling

The Wisconsin dataset is clean to start with, and there wasn't any need for data wrangling. It is important to note that this data set has more benign cases than malignant, this unevenness in the dataset can affect the performance of the classification model. As the malignant and benign values are recorded as strings: 'M' and 'B', feature engineering was used to convert these to 1 and 0 respectively.



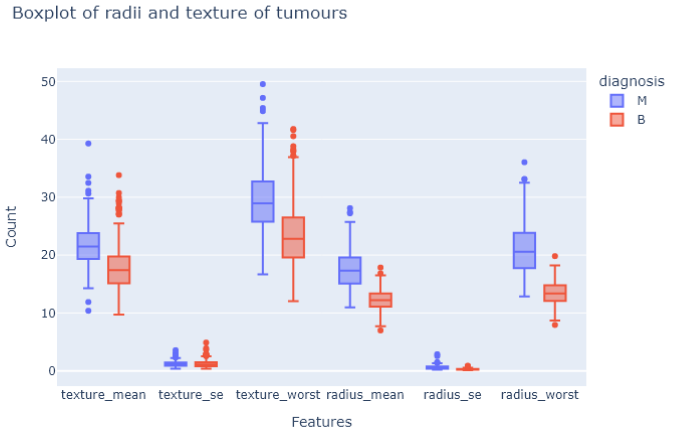
EDA

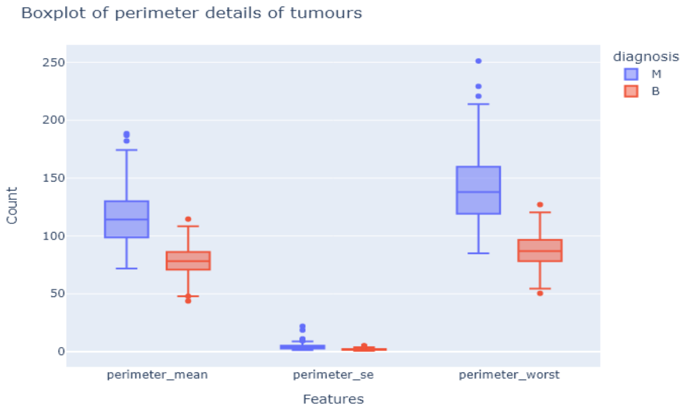
During exploratory data analysis, the need for scaling the data set became apparent as the scales of the attributes vary greatly. A few samples that stood out when trying to identify outliers were:

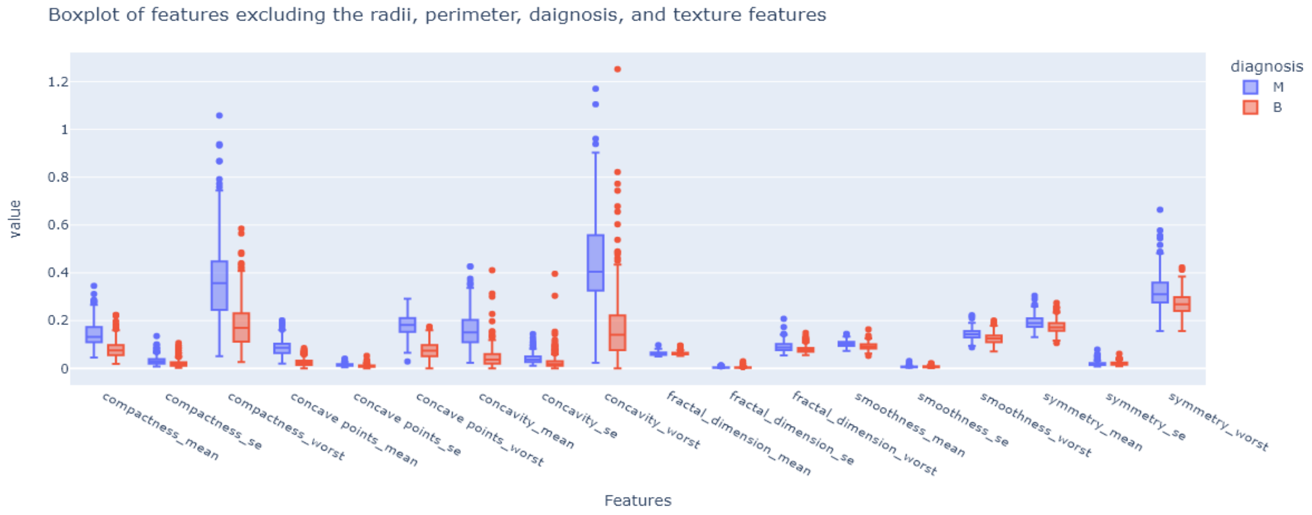
- ID 68: this is a benign sample with the worst concavity value as 1.252.
- ID 152: another benign sample with a very high concavity standard error.

When these two samples were looked at closely, their other features were not remarkable, so these two samples were left in the dataset.

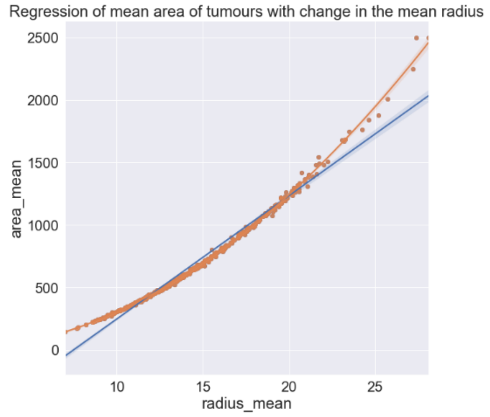
The boxplots also highlight the varying ranges for each feature. It's interesting to note that for benign tumours, the range of values for each feature is almost always smaller than its malignant counterparts. This may be due to the unbalanced dataset favouring benign tumours, or the difference in range may be an important feature to highlight to domain experts.

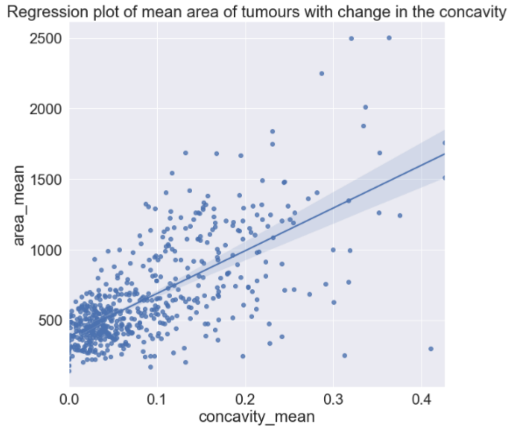






A closer look at the features highlighted right skewness for area, compactness, concave points, and concavity, whilst texture, symmetry, smoothness, perimeter, and fractal dimensions showed more symmetric distributions. When two or more features show strong linear correlation, it can be reasonable to drop one or more of these features, the reasoning being that if one feature can help predict another feature, then by dropping one of these features we can reduce the complexity of our classification model.





As outlined earlier, our goal is to reduce the number of features in the classification model, so linear correlation between the features were explored extensively. It was however in vain. Correlations between area, compactness concave point, concavity, perimeter, and radius of the tumours were highlighted as features with high associations with each other. Attributes related to texture, smoothness, fractal dimensions, and symmetry showed only weak correlations with other features. Deeper exploration showed that although some of these features had high Pearson Correlation Coefficients, they didn’t exhibit linear correlation. Examples of this included mean area, mean radius, and the mean perimeters of the tumours. Mean area and mean concavity had a Pearson correlation coefficient of 0.69, however the scatterplot indicated otherwise. To further investigate this possible linear correlation, a linear regression model for area and concavity, and a multiple linear regression model for concavity and area, without area, radius and perimeter values were developed:

- Linear Regression Model for concavity mean and area mean- R^2 = 0.439
- Multiple Regression model- R^2 = 0.997

At this point, we couldn’t reduce any features with confidence, however it did give some possibility that some features were correlated enough that we might be able to reduce the number of features at a later point.

Model Development and Evaluation

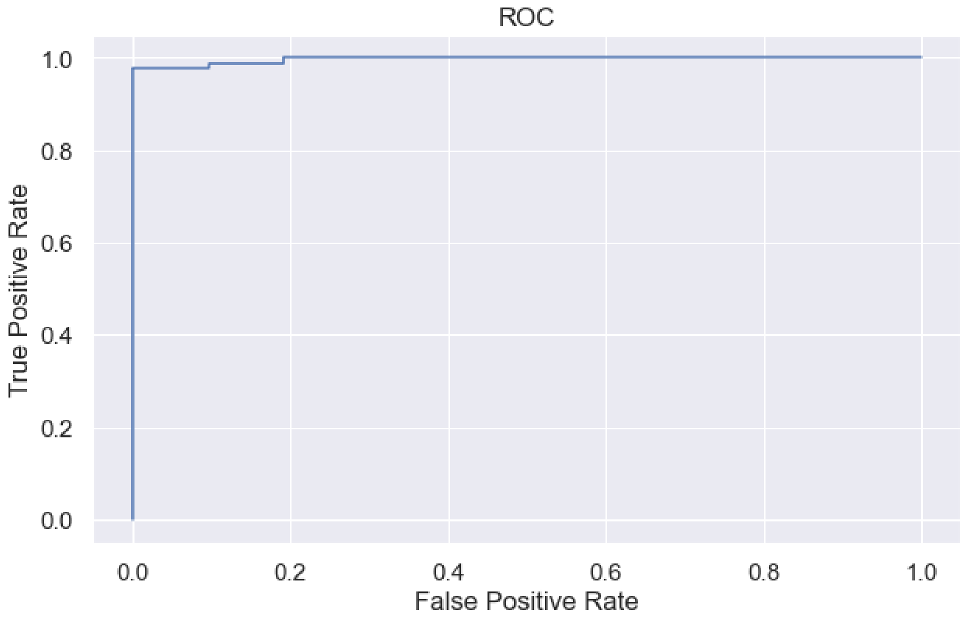
75% of the dataset is used for the training set. For scaling, the standard scaler was used, and in cases where appropriate, Scikit Learn’s Power Transformer was used for logarithmic scaling as our EDA highlighted several features that showed right skew. Each of the models was evaluated with its confusion matrix, classification report, and Matthews Correlation Coefficient (MCC). The evaluation features used were: high values for the area under the roc, high f-1 scores, and low precision scores for malignant tumours. This will allow us to select a model with high level of accuracy whilst still allowing us to favour a higher false positive rate. The table below summarises the top 5 models developed (a csv file summarizing all the models can be made available).

Model	Scaled	Parameters	Features Removed	MCC	Precision for M	F1-Score for M	AOC	Important Features > 1%
KNN	Standard Scaler	Leaf size = 1, n_neighbors = 3, p = 2	N/A	0.970	0.98	0.99	0.999	N/A
KNN	Standard Scaler	n_neighbors = 1	N/A	0.970	0.99	0.99	0.985	N/A
KNN	Standard Scaler	n_neighbors = 4	N/A	0.985	1.00	0.99	0.5	N/A
Random Forest Classifier	No	n_estimators = 23, max_depth=4	N/A	0.940	0.98	0.98	0.997	area worst, concave points worst, concave points mean
Random Forest Classifier	No	n_estimators = 23, max_depth=4	perimeter worst, area se, perimeter mean, radius se	0.955	0.98	0.98	0.998	N/A

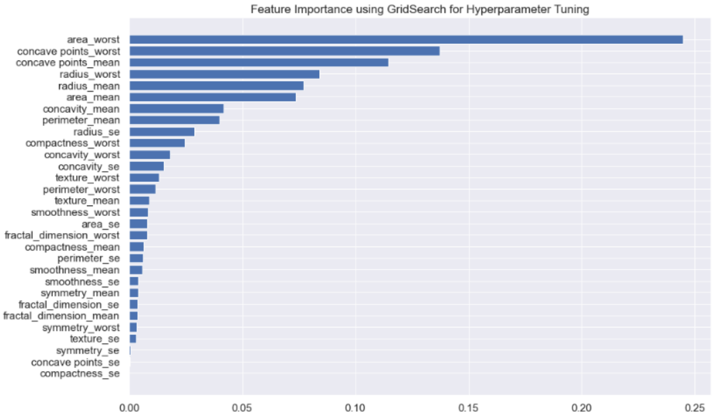
Model Selection

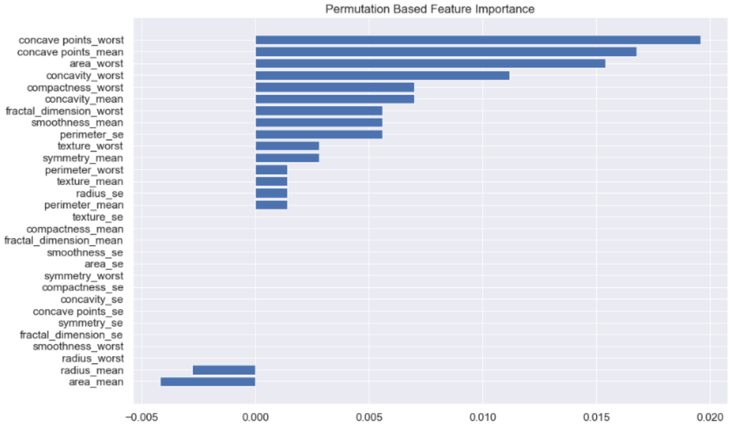
The best model therefore is a KNN model with leaf size = 1, n neighbors = 2, p =2. The confusion matrix and classification report for this model are:

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



It’s important to highlight that a random forest classifier with n_estimators = 23 and max_depth = 4 is also quite useful, as this highlighted the important features as area_worst, concave points_worst, concave points_mean. When feature importance was identified using permutations, these three features were again identified as most important.





Dropping features of lower than 1% feature importance values did show a slight improvement to both the MCC score and the area under the roc, and the confusion matrix is below:

$$Confusion\ Matrix = \begin{bmatrix} 51 & 1 \\ 2 & 89 \end{bmatrix}$$

The KNN model is slightly superior in the sense that it helps us achieve our aim of selecting a model with a higher false positive rate than false negative whilst maintaining high levels of accuracy.

Future Work

It would be ideal to test this model further on sets of different breast cancer data to identify overfittings and underfittings of this model. After speaking to medical experts, it might also be worth considering testing this model with cancerous tumours in general to see if breast cancer tumour classifications have commonalities with any other types of malignant tumours.