Supervised Classification of Malignant Tumours

Index

The Objective: To create a machine learning model that can accurately differentiate between malignant and benign tumours. As such, this model will focus on prediction, and will be beneficial by reducing the number of samples that have to be sent to labs for blood testing.

The Data: a data set containing 570 tumours, either labelled Benign or Malignant, with corresponding physical measurements of each.

- Contained 31 columns: of which, the 30 features are continuous data (physical measurements of the tumour), and the target is binary categorical (B/M).
- Contained no missing or repeated values.
- Each feature contained outliers in both the M and B classes.
    - Models were tested with outliers:
        - Removed
        - Retained
    - Outliers were not modified as multiple features displayed a degree of multicollinearity.
- The classes were imbalanced with roughly 63% of cases being benign and the other 37% being malignant.

- Label encoding was utilised to convert B (0) and M (1).
- Two columns were removed before training.
    - ID column – Not relevant to training.
    - “Unknown: 32” column – Contained no values.
- All remaining features were passed to the models, and either recursively removed, if allowed by the model, or manually selected for removal based on permutation feature importance.

- This analysis aims to produce the simplest model that has a sensitivity greater than 99% whilst minimising False Positives.
    - Due to the nature of target (cancerous tumours) False Negatives could be highly damaging, and potentially fatal.
    - False Positives are less important as these samples would still undergo further testing to confirm the tumour is malignant.
- All remaining features were passed to the models, and either recursively removed, if allowed by the model, or manually selected for removal based on permutation feature importance.

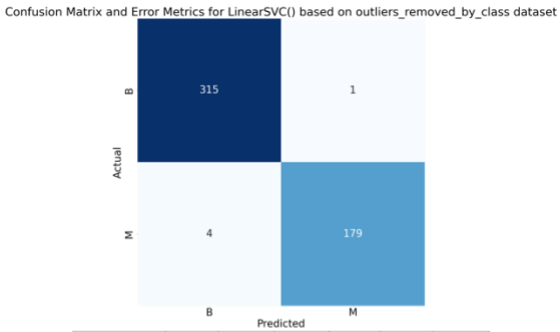
Classifier Models: Both individual and ensemble methods were used on a variety of different datasets to find the best possible model for malignant tumour identification.

1. Logistic Regression
2. SVM
3. KNN
4. Decision Trees
5. Bagging
6. Random Forests
7. StackingClassifier

The models were tested with selected features and the full 30 features, in general the models performed as well if not better with the full 30, except for non-linear SVC; therefore, the full 30 features were used to train the final models. Use of all 30 features was chosen due to improved recall and so that the training data was consistent across models and could then be used to create a stacking classifier. The models with the highest recall were then stacked, the best result of all models tested is on the following page.

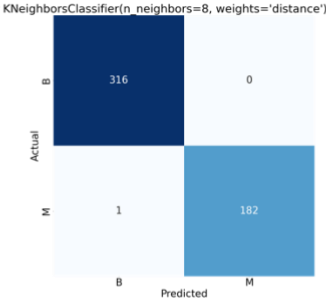
Results:

	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
LogisticRegression	10.0	30.0	0.98998	0.99444	0.97814	0.98623	0.98749



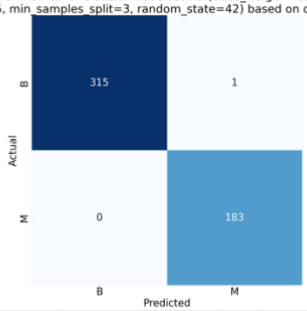
	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
LinearSVC	10.0	30.0	0.98998	0.99444	0.97814	0.98623	0.98749

Confusion Matrix and Error Metrics for KNeighborsClassifier(n_neighbors=8, weights='distance') based on outliers_removed_by_class dataset



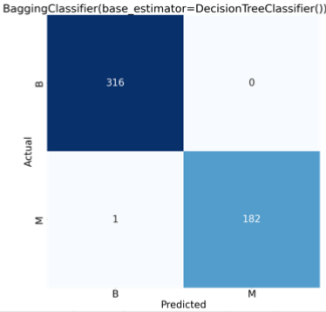
	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
KNeighborsClassifier	10.0	30.0	0.998	1.0	0.99454	0.99726	0.99727

Confusion Matrix and Error Metrics for DecisionTreeClassifier(class_weight='balanced', criterion='entropy', max_depth=25, min_samples_split=3, random_state=42) based on outliers_removed_by_class dataset



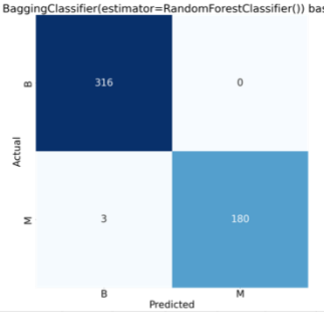
	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
DecisionTreeClassifier	10.0	30.0	0.998	0.99457	1.0	0.99728	0.99842

Confusion Matrix and Error Metrics for BaggingClassifier(base_estimator=DecisionTreeClassifier()) based on outliers_removed_by_class dataset



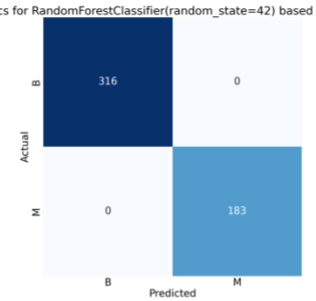
	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
BaggingClassifier	10.0	30.0	0.998	1.0	0.99454	0.99726	0.99727

Confusion Matrix and Error Metrics for BaggingClassifier(estimator=RandomForestClassifier()) based on outliers_removed_by_class dataset



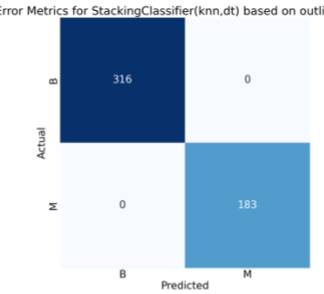
	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
BaggingClassifier	10.0	30.0	0.99399	1.0	0.98361	0.99174	0.9918

Confusion Matrix and Error Metrics for RandomForestClassifier(random_state=42) based on outliers_removed_by_class dataset



	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
RandomForestClassifier	10.0	30.0	1.0	1.0	1.0	1.0	1.0

Confusion Matrix and Error Metrics for StackingClassifier(knn,dt) based on outliers_removed_by_class dataset



	Stratified splits	Features	Accuracy	Precision	Recall	F1 Score	AUC-ROC
StackingClassifier(knn,dt)	10.0	30.0	1.0	1.0	1.0	1.0	1.0

Model Recommendations: Based on the above results stacking the optimised KNN and decision trees together and using logistic regression as the final estimator is the simplest model able to identify all cases in the outliers removed by class dataset correctly, and even when outliers were retained only misidentified 1 case as B when it was M. As this model pertains to human health the model must be highly accurate and relatively interpretable, and as such I believe that the stacking of self-interpretable models provides the best compromise. However, as previously stated, this model must be highly trusted as it would be working in the healthcare sector, and as such perhaps a model such as Decision Trees alone would be a better, this would be dependent on how consumers view, and how much they trust, machine learning algorithms.

Key Findings: Multiple models were able to identify malignant tumours with a recall of 98% or greater, the more complex methods (RandomForests and Stacking) were able to identify all cases with 100% accuracy; this indicates that machine learning may be an excellent method of identifying malignant tumours. However, this dataset was relatively

small with less than 1,000 instances and it may be coincidence that the models performed so well on this dataset.

Future Work: The method chosen to deal with outliers was simplistic and as such a useful next step in the analysis of this data would be the use of more sophisticated methods of dealing with outliers. Additionally, the use of LIME on the false negatives in the outliers retained dataset may provide new information that may help refine the model. Lastly, I would like to test this model on a much larger dataset and ensure that the high level of accuracy shown by the model generalises well to data outside the dataset; primarily I am concerned by just how well the model performed and want to ensure that data leakage did not occur during training.