1. **Build a decision tree – an *rpart* tree and report confusion table, balanced accuracy as well as the tree diagram?**

Diagram

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
|  | B | M |
| B | 79 | 10 |
| M | 5 | 48 |
| Sensitivity | 79/79+10 = 88.76% | |
| Specificity | 48/(48+5) = 90.57% | |

In the decision tree the majority of splits in the first node were based on the variable perimeter worst. This variable gave us the most information gain and lower the entropy of the model before the next split. At each node the variable that contributed to the most information gain is displayed and the split value i.e., if true it goes to the left. At the leaf nodes we have the output/classification of the datapoints. We want the bins to be as pure as possible meaning only data of one type being in each bin or as minimal as possible. This is so we can be confident in correctly identifying if the tumours are malignant or benign based on the morphology data.

Graphical user interface, text

Description automatically generated

Lets take node 4 for example. That is a datpoint were classified as benign in this leaf there is a 98% probability that this was the correct call. The p(node) also shows how much of the data is in this leaf i.e., just over half. This could be an indication of a relationship between the values of perimeter\_worst and concace.points\_worst being less than the cut off points and tumours being accurately classified as benign.

A balance in the sensitivity and specificity are needed when deciding on acceptability. High sensitivity and low specificity would lead to high false negatives and overtreatment of those patients with benign tumours. While the opposite would lead to false negatives and thus delay diagnosis and treatment. In terms of cancer screening, we ideally want to keep the false negatives low thus we would prioritise sensitivity at the risk of overtreating some patients. This tree has a good balance on the two but if we were to prioritise sensitivity, we may have to adjust some of the decision criteria to improve the sensitivity.

1. **Build a random forest with at least 10 trees and report confusion table and balanced accuracy. Is it different compared to the one from the tree above? Compare and interpret the confusion tables from both models – the decision tree and the random forest.**

|  |  |  |
| --- | --- | --- |
| Ntree= 10 | | |
|  | B | M |
| B | 82 | 7 |
| M | 8 | 45 |
| Sensitivity | 82/(82+7) = 92.13% | |
| Specificity | 45/(45+8) = 84.9% | |

|  |  |  |
| --- | --- | --- |
| Ntree= 500 | | |
|  | B | M |
| B | 83 | 6 |
| M | 6 | 47 |
| Sensitivity | 83/(83+6) = 93.25% | |
| Specificity | 47/(47+6) = 88.67% | |

Two random forests were built, one with ten trees and the other with 500. The accuracy of the forest didn’t really improve drastically with the addition of more decision trees. However, the sensitivity has increased from the singular tree which is something that would be considered when deciding on a model for cancer screening. The improved sensitivity could be from stricter criteria in the splits.

1. **Find important features from both models. Compare and interpret on the similarities/differences of important features/covariates from the tree (generated from Q1 of this section) and random forests (generated from Q1 of this section)?**

In the tree model the variables importance were listed as:

Text

Description automatically generated with medium confidenceTimeline

Description automatically generated with medium confidence

While both models use different methods of calculating the importance both are based on the Gini index. In both perimeter\_worst appeared as one of the top 3 variables in the decisions.

Index:

Q1.

Graphical user interface, text

Description automatically generated

Text

Description automatically generated with medium confidence