Question 8

PRUNE set to TRUE:
Training data results:
Precision: 0.8087, Recall: 0.6885, Accuracy: 0.8819
Test data results:
Precision: 0.7364, Recall: 0.6230, Accuracy: 0.8526
PRUNE set to FALSE:
Training data results:
Precision: 0.9821, Recall: 0.9153, Accuracy: 0.9748
Test data results:
Precision: 0.6352, Recall: 0.6095, Accuracy: 0.8181

Question 8a

We prune the branches when the chi-square value is not greater than the threshold meaning that the two variables are independent. If we do not prune our tree, then our model might overfit the training data with extra branches that are not meaningful. This will result in higher precision, accuracy and recall for the training results and lower for the test results. Generally, the results for the training set should be better than the results for the test set. The experimental values that I obtained from running the model on the dataset with and without the pruning matches the expectations.

Question 8b I might exclude race and sex from the list of features because the car company that is trying to show these ads might face lawsuits for being sexist or racist if the feature list somehow got leaked to the public. I think we will still end up with a good enough model without those two fields and the risk of lawsuits is not worth a small decrease in accuracy.

Question 8c

Let:
True positive = a, False positive = b, False negative = c, True negative = d
$Accuracy = \frac{a+d}{a+b+c+d} = 0.9$
$Precision = \frac{a}{a+b} = 0$
$Recall = \frac{a}{a+c} = 0$
so there are no true positives
$Accuracy = \frac{d}{b+c+d} = 0.9$
$10d = 9b + 9c + 9d$
$d = 9(b+c)$
Possible if the number of True positives is 0 and the number of true negatives is nine times the sum of false positives and false negatives.