Project 3
Karthik Unnikrishnan

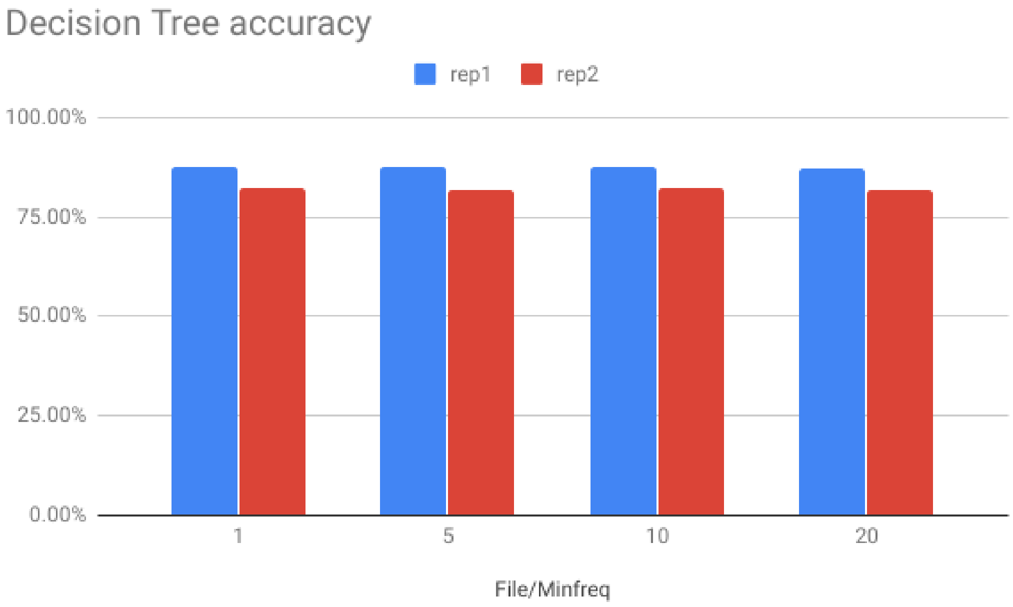
Decision tree implementation:

For the decision tree implementation, the following steps were followed:

1. Read the input file along with the value for minfreq
2. In order to determine the best split point at each node, compute the gini index for each possible split for every dimension
3. For optimizing the split selection process, each dimension is sorted in ascending order and the split point candidates are taken as the midpoint between two consecutive points
4. The splitting is done recursively for as long as the termination condition is not satisfied
5. In order to represent the decision tree in a file a data structure with the following properties for each node is used
    - nodeId: Unique identifier for a node
    - splitIndex: index at which the node is splitting
    - splitValue: value which is used to split the node
    - leftChild: nodeId of the left child node
    - rightChild: nodeId of the right child node

Accuracy:

File/Minfreq	1	5	10	20
rep1	87.5%	87.48%	87.54%	87.16%
rep2	82.08%	81.83%	82.26%	81.65%



Inference:

Based on the values of the accuracy for both files, the following observations can be made:

1. For rep1 the highest accuracy is for minfreq=5. Beyond this, the accuracy decreases, which indicates overfitting.
2. For rep1 the highest accuracy is for minfreq=10. Beyond this, the accuracy decreases, which indicates overfitting.
3. Increasing the model complexity by reducing minfreq does not seem to have a significant impact on the accuracy. Thus, a simpler model may be preferable
4. The accuracy values for the rep1 file are greater than those for rep2. This is as expected, due to the loss of information during PCA

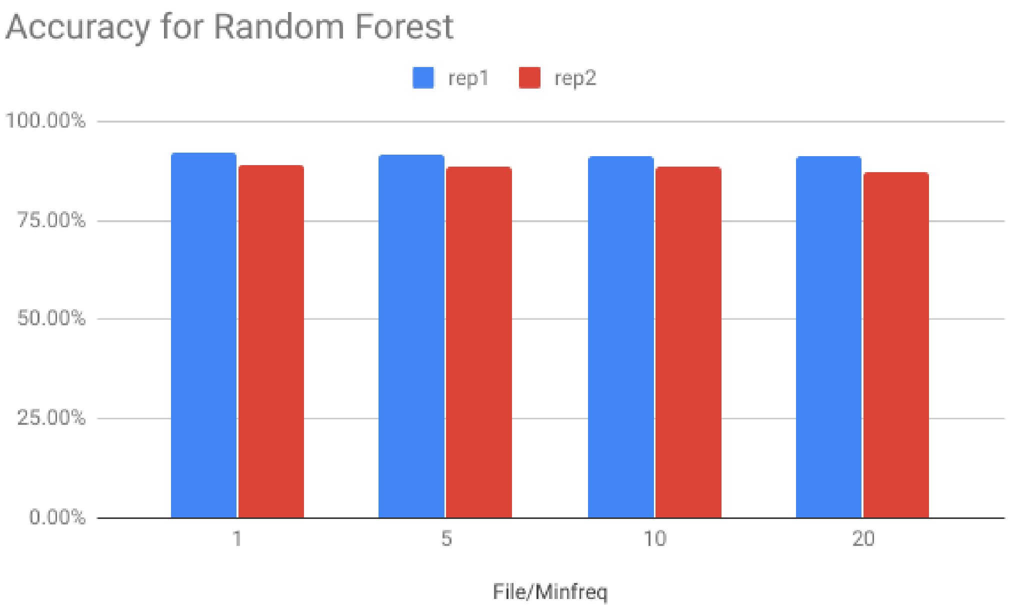
Random forest implementation:

As per the specifications, 100 samples of 40% the size of the training data were used to train 100 decision trees. The python function random.choices() was used for sampling with replacement

the decision trees generated previously are then used to generate 100 predictions and the label with the majority vote is taken as the label for the final predictions.

The accuracies for the different minfreq combinations are as follows:

File/Minfreq	1	5	10	20
rep1	92.06%	91.7%	91.13%	91.3%
rep2	89.06%	88.3%	88.44%	87.21%



Inference:

1. Sampling with replacement to build a random forest seems to have improved the accuracy of the predictions.