ML Assignment 1

Accuracy Values Record

Data Set	H = Entropy P = No	H =Entropy P = REP	H = Variance P = No	H = Variance P = REP	Random Forests
C =300 D = 100	57.81%	59.31%	57.78%	58.29%	100%
C =300 D = 1000	64.98%	65.13%	64.78%	65.03%	100%
C =300 D = 5000	73.25%	73.62%	72.83%	73.39%	100%
C =500 D = 100	64.321%	64.321%	60.30%	61.81%	100%
C =500 D = 1000	68.98%	69.53%	66.83%	66.93%	100%
C =500 D = 5000	74.78%	75.13%	73.42%	73.58%	100%
C =1000 D = 100	69.35%	67.84%	68.34%	67.84%	100%
C =1000 D = 1000	79.94%	80.45%	79.94%	80.74%	100%
C =1000 D = 5000	83.81%	84.5%	83.18%	83.51%	100%
C =1500 D = 100	81.41%	87.44%	82.91%	87.94%	100%
C =1500 D = 1000	81.47%	87.94%	88.59%	88.61%	100%
C =1500 D = 5000	94.32%	94.47%	93.80%	93.97%	100%
C =1800 D = 100	90.95%	93.97%	90.95%	90.97%	100%
C =1800 D = 1000	97.40%	97.85%	96.20%	96.35%	100%
C =1800 D = 5000	98.43%	98.45%	97.70%	97.86%	100%

1. **Which impurity heuristic (Entropy/Variance) yields the best classification accuracy? How does increasing the number of examples and/or the number of clauses impact the (accuracy of the) two impurity heuristics. Explain your answer.**

   - **Entropy Heuristic yields the better classification accuracy when compared to variance heuristic**
   - **As the number of examples increases, the accuracy also increases in both the impurity heuristics because there will me more accurate classification of data when the number of examples is more.**

- **As the number of clauses increases, the accuracy also increases because the combinations of outcomes are more.**

Ex 1:

For c = 300, d = 100 accuracy : 57.81%
For c = 1800, d = 100 accuracy : 90.95%
This proves that as the number of clauses increases, the accuracy also increases

Ex 2:

For c = 300, d = 100 accuracy : 57.81%
For c =300, d = 5000 accuracy : 73.25%
This proves that as the number of examples increases, the accuracy also increases

2. **Which overfitting avoidance method (reduced error pruning/ depth-based pruning) yields the best accuracy? Again, how does increasing the number of examples and/or the number of clauses impact the (accuracy of the) two overfitting avoidance methods. Explain your answer.**

   Reduced error pruning is yields the good accuracy compared to depth based pruning. This is because in depth-based pruning, the decision tree is pruned to a particular depth whereas in reduced error pruning, decision tree is pruned based on the accuracy of the tree.

   Reduced error pruning gives the accuracy within linear time where as depth based pruning gives the solution in exponential amount of time.

3. **Are random forests much better in terms of classification accuracy than your decision tree learners? Why? Explain your answer.**

   Yes, random forests are much better in classification accuracy because in random forests, a forest of trees are taken and the most optimal decision tree is considered for classification of the class variable for the test data.
   There is a pool of decision trees with different classifiers and confusion matrix to determine the best suitable decision tree for the machine learning problem.