**Level 3 Assignment**

1. Naïve bayes is so naïve because it makes assumption that all attributes are independent of each other and naive model generalizes strongly that each attribute is distributed independently of any other attribute. Since the assumptions made as such that all of the features in a data set are equally important and independent and as these assumptions are rarely true in the real world Naïve bayes is so 'naïve'.

2. We use library rpart,

   Tree <- rpart(Private ~ ., method='class', data =train)

   Tree = The decision tree
   Private= The target variable in data
   '.' = use all variables
   Method='class' = indicates that we are classifying
   Data= train = passing in the train data

3. Since no training or learning of the model is required and such all of the work happens when the prediction is being made hence Knn is called as lazy algorithm.

4. Naïve bayes works well for categorical features indeed by calculating Posterior probability
   $P(c|x) = (((P(x|c)\ P(c)\ )/\ P(x))$ where

   $P(c|x)$ = posterior probability
   $P(c)$ = prior probability of the class
   $P(x|c)$ = likelihood of predictor
   $P(x)$= prior probability of predictor

   If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a zero probability hence to overcome this, we use the smoothing technique.

   For numerical data we use normal distribution. Example:- Bell curve

5. Build a confusion matrix and calculate the accuracy. Then calculate the false negative rate.

6. High validation error means that the model has overfitted. Usually RF has a very high accuracy on the training population, because it uses many different characteristics to make a prediction and because of the same reason, it sometimes over fits the model on the data. The classifier had minimized training data patterns to an extent and these are not available in the test data i,e the unseen data and when we ran on the test data it couldn't find those patterns hence the error rate went high.

   To avoid this condition we should choose lesser number of trees.

7. For decision trees these are the ways of handling overfitting:
   a) Don't grow the trees to their entirety
   b) Pruning
   c) Setting constraints on the size of the tree

   The same applies to a forest of trees.

   Pruning is one of the most popular techniques and here is how it work:

   a) First make the decision tree to a large depth.

   b) Then start removing leaves from bottom which are giving us negative returns when compared from the top.

   Several other parameters that you can use to tune your forests are:
   nodesize- minimum size of terminal nodes
   maxnodes-maximum number of terminal nodes
   mtry-number of variables used to build each tree

8. Statistical hypothesis testing is used to determine whether the result of a data set is statistically significant. This test provides a p-value, representing the probability that random chance could explain the result. In general, a p-value of 5% or lower is considered to be statistically significant.

   We can also implement by using 'arules' package.

9. The tree decides where to split by using algorithms such as:-

a) **Gini Index:** If two items are selected from a population at random then they must be of same class and probability for this is 1 if population is pure.

Steps to Calculate Gini for a split:

- Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
- Calculate Gini for split using weighted Gini score of each node of that split

  Which is having the higher Gini score the spit will take place there.

b) **Chi- Square**: The statistical significance between the differences between sub-nodes and parent node is calculated.

Steps to Calculate Chi-square for a split**:**

- Calculate Chi-square for individual node by calculating the deviation for Success and Failure.
- Calculated Chi-square of Split using Sum of all Chi-square of success and Failure of each node of the split.

Formula used: Chi-Square= ((Actual – Expected)^2 / Expected)^1/2

c) **Information Gain**: A measure to define the degree of disorganization in a system is known as Entropy.

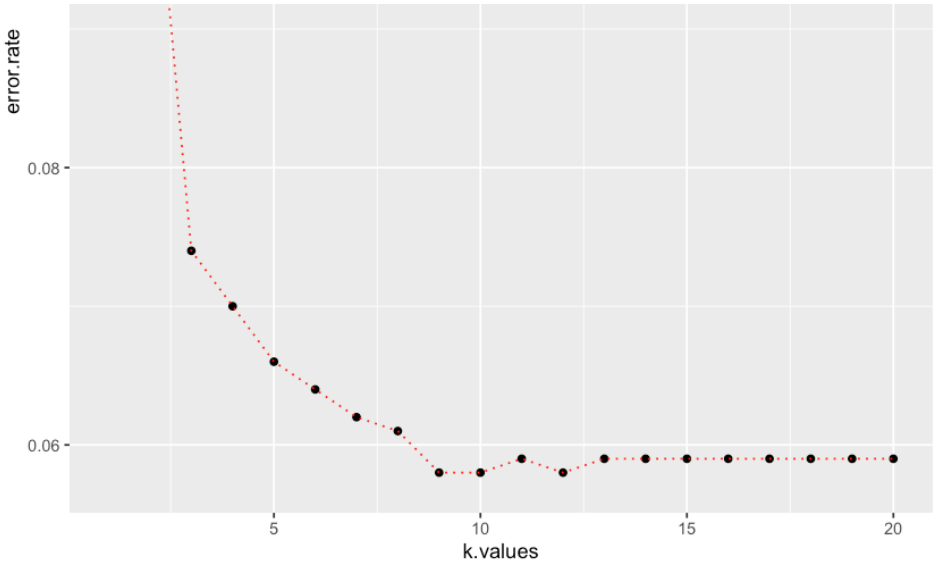
Entropy = -plog2p – qlog2q

p and q are the probability of success and failure respectively in that node. Entropy chooses the split which has lowest entropy compared to parent node and other splits. The lesser the entropy, the better it is.

Steps to calculate entropy for a split:

- Calculate entropy of parent node
- Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

10. Using Elbow method we can plot out the various error rates for the K values. We should see an "elbow" indicating that we don't get a decrease in error rate for using a higher K. This is a good cut-off point for choosing K value.




I have plotted a scatter plot-k-value vs error rate, as we can see the error rate is lowest for k=9 i,e 0.058 and then increases to 0.059 and then becomes stable.