1]A)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive:  Actual Yes | Negative:  Actual No |  |
|  | TP = 100 | FP = 10 |  |
|  | FN = 5 | TN = 50 |  |
|  | 105 | 60 |  |

**Accuracy** = TP+TN/N = 150/165 = 90.90%

**RECALL** = TP/TP+FN = 100/105 = 95.23%

**PRECISION** = TP/TP+FP = 100/110 = 90.90%

**F1-Score** = 2(Precision\*RECALL/Precision+RECALL) = 2((100/110)\*(100/105)/(100/110)+(100/105)) =

2((0.909090)\*(0.952380)/(0.909090)+(0.952380)) = 2(0.865799/1.86147) = 0.930231

**F1-Score =** 93%

1]B)

**1) K Nearest Neighbour:**

Is a supervised Learning algorithm.

Is a Predictive Model.

Class Variable (Predictive Variable) can be discrete or continuous.

KNN needs Supervised Data. Data which contains both Indicator variables and Predictive variables.

It is a non-parametric model. It will not assume anything on underlying data but learns from the data.

It is based on distance metric (Proximity Measures) such as Euclidean and Hamming.

Euclidean Distance:

According to Euclidean Distance, larger distance between data points represents “dissimilarity” whereas a small distance indicates “similarity” between observations.

This can be used only with continuous data values.

Hamming Distance:

This is also used to calculate Proximity Measures (Distance Metric)

This can be used with Categorical data values.

The distance is given as 0 otherwise it is 1.

This is compute intensive

**2) Naïve Bayes**

Is a supervised Learning algorithm.

Is used for Prediction

Is a probabilistic Model.

Class Variable (Predictive Variable) can be only discrete.

**Conditional Probability:**

If AA and BB are two events in a sample space SS, then the conditional probability of AA given BB is defined as

P(A|B)

P(A and B) = P(B|A).P(A) = P(A|B). P(B)

P(A|B) = P(A and B)/P(B)

**Bayes theorem:**  
P(A|B) = P(B|A)\*P(A)/P(B)

**Laplace Correction**

To avoid any Probabilistic query resulting in zero Laplace correction is introduced.

P(xi|c) = Nic+1/Nc+c

Where Nic = How many number of instances in which we have P(xi|c) in the dataset

Nc = Number of combinations of class

C= Number of instances of the class found.

Naïve Bayes assumes that there is no relation between Indicator variables.

This is not compute intensive.

**3) Decision Tree**

Is a supervised Learning algorithm.

Is used for Prediction

Works on Supervised Data

Class Variable (Predictive Variable) can be discrete or continuous.

It is a non-parametric model. It will not assume anything on underlying data but learns from the data.

It builds a model which will be in tree like structure composed of Features and relationships to predict.

Indicator variables are internal node and Predictor variable is leaf node.

Decision tree does pruning to eliminate irrelevant features and use simple direct short path to decision making.

Entropy is a measure of impurity. Which we will leverage to chose attributes for each node. The attribute that leads to clear decision can be calculated via entropy.

Even Gini Index can be leveraged for measuring the impurity.

Information Gain. The attribute with a high value for Information Gain will be placed at the root.

This is compute intensive.