Machine Learning

1. A, B
2. A
3. A
4. A
5. C
6. C
7. B
8. B
9.

   Class A = 60% , Class B = 40%
   Gini index = $1 - \Sigma(Pi)^2$
   Gini index = $1 - (0.6)^2+(0.4)^2$
   Gini index = 1 - (0.36) + (0.16)
   Gini index = 1- 0.52
   Gini index = 0.48

   Entropy = $- \Sigma$ Pi log2(Pi)
   Entropy = - [0.4*log2(0.4) + 0.6*log2(0.6)]

10. Advantages of Random Forest:
    1. It is one of the most accurate algorithm for classification problem.
    2. It good for large data.
    3. It can handle thousands of variables.
    4. It is less prone to overfitting than decision tree.
    5. It is work well with booth categorical and numerical data and no scaling or transformation of variable is not required usually.
    6. It can handle linear and non linear relationship well.
    7. It is not influenced by outliers.

11. Scaling the data is pre processing step. As we know most of supervised and unsupervised learning methods make decision according to data and often algorithms calculate distance between the data points to make better ifereces out of the data.
    In machine learning if values of feature are closer to each other there are chances for algorithms to get trained well and faster. If data is not

scaled then there chances of prediction of algorithms will biased or inaccurate.

Followings are technique for scaling:

- Standardization: in standardization calculate the z value for each of data point and replace with z values. This best to use when feature data is normally distributed.
- Normalization: normalization change your observation so they can be described as a normal distribution. the mean and median are same, there are more values closer to the mean value.

12. Advantages of Gradient descent:
    - Gradient descent can be used with various function and handle non linear regression problem.

13. In case of highly imbalanced dataset for a classification problem accuracy is not good metrics to measure the performance of the model because data is imbalanced so there are chances of model will be biased. In this case other metrics like precision, recall and F1 score also to understand the performance of model.
    Accuracy alone not good metrics to analyze performance of model but with other metrics for imbalance data has to consider to estimate model performance.

14. F score is used to evaluate binary classification problem. F score or f1 score is harmonic mean of precision and recall. It is used to measure the test accuracy.
    F Score =2*1/(1/precision)+(1/recall)

15. fit():the stored the data for transformation and it doesn't return anything.
    transform(): transform use stored data in by fit function and transform the data.
    fit_transform():it is combination of fit and transform method.