WORKSHEET 6 MACHINE LEARNING

1) D

2) A

3) A

4) A

5) C

6) A

7) B

8) D

9) Gini index = 1- (p(A)2 + p(B)2)

= 1-((0.4)2+(0.6)2)

= -1

Entropy = -(p(A) log2(p(A)) + p(B) log2(p(B)))

= -(0.4*log2(0.4) + 0.6*log2(0.6)))

= 0.97

10) Random forest algorithm avoids and prevents overfitting by using multiple trees.This gives accurate and precise results when compared to decision tree algorithm which are more prone to overfitting and can't guarantee optimal trees.

11) We need to scale all numerical features in a dataset so that they are all on the same scale, as this helps the model to assign equal importance to all features and make predictions without bias.

Min-Max Normalization: This technique re-scales a feature or observation value with distribution value between 0 and 1.

Standardization: It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

12)

a) It makes the training faster
b) It prevents the optimization from getting stuck in local optima
c) It gives a better error surface shape
d) Weight decay and Bayes optimization can be done more conveniently

13) Accuracy is not good metric for imbalanced dataset because, suppose we have an imbalanced dataset and a badly performing model which always predicts for the majority class. This model would receive a very good accuracy score as it predicted correctly for the majority of observations, but this hides the true performance of the model which is objectively not good as it only predicts for one class.

14) It is a metric used to evaluate the performance of a Machine Learning model. It combines precision and recall into a single score. The

accuracy metric computes how many times a model made a correct prediction across the entire dataset.

F-measure formula:

F-score = 2 * (precision * recall) / (precision + recall)

Accuracy in making positive predictions is measured by a recall, while identifying all positive occurrences in the data is quantified by precision. The F-score ranges from 0 to 1, with higher values indicating better performance.

15) The fit() method is used to compute the mean and std dev for a given feature to be used further for scaling. The transform() method is used to perform scaling using mean and std dev calculated using the . fit() method. The fit_transform() method does both fits and transform.