WORKSHEET 7
MACHINE LEARNING

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

9. Suppose we have a dataset which have two classes A and B. The percentage of class A is 40% and percentage of class B is 60%. Calculate the Gini index and entropy of the dataset.

Gini index = 0.28. Entropy = 0.97

10.What are the advantages of Random Forests over Decision Tree?

Random forest algorithm avoids and prevents overfitting by using multiple trees. The results are not accurate. This gives accurate and precise results. Decision trees require low computation, thus reducing time to implement and carrying low accuracy.

11.What is the need of scaling all numerical features in a dataset? Name any two techniques used for scaling.

Scaling is required to rescale the data and it's used when we want features to be compared on the same scale for our algorithm. And, when all features are in the same scale, it also helps algorithms to understand the relative relationship better.

- MIN MAX SCALER
- STANDARD SCALER

12.Write down some advantages which scaling provides in optimization using gradient descent algorithm.

- It makes the training faster.
- It prevents the optimization from getting stuck in local optima.
- It gives a better error surface shape.
- Weight decay and Bayes optimization can be done more conveniently.
- It's also important to apply feature scaling if regularization is used as part of the loss function so that coefficients are penalized appropriately.

13.In case of a highly imbalanced dataset for a classification problem, is accuracy a good metric to measure the performance of the model. If not, why?

In case of a highly imbalanced dataset for a classification problem accuracy is not at all good metric to measure the performance of the model.

Achieving 90 percent classification accuracy, or even 99 percent classification accuracy, may be trivial on an imbalanced classification problem.

This means that intuitions for classification accuracy developed on balanced class distributions will be applied and will be wrong, misleading the practitioner into thinking that a model has good or even excellent performance when it, in fact, does not.

14.What is “f-score" metric? Write its mathematical formula.

In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive.

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

15.What is the difference between fit(), transform() and fit_transform()?

The fit(data) method is used to compute the mean and std dev for a given feature to be used further for scaling. The transform(data) 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