Machine Learning Worksheet-7

1. D) All of the above

2. A) Random forest

3. B) The regularization will decrease

4. A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown.

5. C) In case of classification problem, the prediction is made by taking mode of the class labels

predicted by the component trees.

6. C) Both of them

7. B) Bias will decrease, Variance increase

8. B) model is overfitting

9. **Gini index** = $1 – (0.4^2+0.6^2) = 0.48$

Entropy = $-(0.4*\log_2(0.4) +0.6*\log_2(0.6)) = 0.97$

10. Random forest reduces the overall variance of the model by implementing parallel decision trees, therefore, the risk of overfitting reduces.

11. Need for scaling numerical features is mainly due to following reasons:

- It increases computational speed of the algorithm.
- It may help algorithms implementing gradient descent to converge faster with feature scaling than without it.

Two technique used for feature scaling are:

- Min-Max Scaling
- Standard Scaling

12. Scaling may help algorithms implementing gradient descent to converge faster in compared to without it.

13. Accuracy is not a good metric for highly imbalanced dataset, as in the case of highly imbalanced dataset the model's **null accuracy** is usually highly due to always predicting the **most frequent** class, but the model's ability to classify the **minor** class cannot be analyzed in this case and therefore, model's performance cannot be correctly evaluated.

14. F1-score is the harmonic mean between **Precision** and **Recall**. In other words, **F1-score** conveys balance between **Precision** and **Recall** and is usually the best metrics for highly imbalanced datasets.

Mathematically,

F1-score = $\boldsymbol{2\times \frac{Precision \times Recall}{Precision + Recall}}$

15. In data pre-processing or transformation stage, the ***fit()*** method fits the training data and stores the statistics of the training data, whereas, ***transform()*** method transforms any given data using the same statistics of the fitted data (i.e., training data), and ***fit_transform()*** sequentially fits and transforms the given data.