## **MACHINE LEARNING**

## **WORKSHEET – 4**

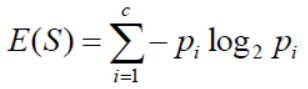
## **Solutions**

1. A) GridSearchCV()
2. A) Random forest
3. B) The regularization will decrease
4. C) both A & B
5. A) It's an ensemble of weak learners.
6. C) Both of them
7. B) Bias will decrease, Variance increase
8. B) model is overfitting
9. **Gini index** (a criterion to minimize the probability of misclassification):

Formula:  ****

**Gini Index**= 1-[(40/100)^2+(60/100)^2]= **0.48**

**Entropy** (a way to measure impurity):

Formula: 

*Where, E(s) is Entropy.*

**Entropy** = -[(40/100) \* log2 (40/100)] –[(60/100) \*log2 (60/100)]= **0.9709505944546686**

1. Advantages of Random Forests over Decision Tree are:

* Reduction in overfitting: by averaging several trees, there is a significantly lower risk of overfitting.
* Less variance: By using multiple trees, you reduce the chance of stumbling across a classifier that doesn’t perform well because of the relationship between the train and test data. As a consequence, in almost all cases, random forests are more accurate than decision trees.
* High predictive accuracy. **Decision trees** do not have same predictive accuracy

1. Scaling all numerical features in a dataset is used to treat skewed features and rescale them for modelling , when all features are in the same scale, it also helps algorithms to understand the relative relationship better. Two techniques used for scaling are :-

* Standard Scaler
* Min-Max Scaler

1. Feature scaling helps in causing Gradient Descent to converge much faster as standardizing all the variables on to the same scale. This is because θ will descend quickly on small ranges and slowly on large ranges, and so will oscillate inefficiently down to the optimum when the variables are very uneven.
2. In case of a highly imbalanced dataset for a classification problem, accuracy **is not a good** metric to measure the performance of the model. In a problem where there is a large class imbalance, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy. So, further performance measures are needed such as Confusion matrix and ROC-AUC Score.
3. The **F-score**, also called the F1-score, is a measure of a model’s accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into ‘positive’ or ‘negative’.

F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal.

The formula for the F1 score is: F1 = 2 \* (precision \* recall) / (precision + recall)

In the multi-class and multi-label case, this is the average of the F1 score of each class with weighting depending on the average parameter.

1. **Fit:** computes the mean and std (standard deviation) to be used for later scaling.

**Transform:** It’s uses a previously computed mean and std (standard deviation) to autoscale the data (subtract mean from all values and then divide it by std (standard deviation)).

**Fit Transform:** fit\_transform does both these process at the same time.