**SQL – Assignment 7**

1. B
2. B,C
3. C
4. C
5. C
6. C
7. C
8. B
9. C
10. A
11. Joins in SQL are used to query the results form multiple database tables after joining them on any matching column .
12. Inner join , left Join , right join , full outer join, cross join
13. SQL server is a software product with the primary function of storing and retrieving data as requested by other software applications—which may run either on the same computer or on another computer across a network.
14. Primary key is a constraint which when applied to any attribute ensures that the attribute will have no null values and contain unique values only.
15. **ETL** stans for Extracting the data , transforming the data as per business needs and storing the data in the data wareshouse.

**Statistics - Assignment 7**

1. B
2. D
3. C
4. B
5. C
6. **A**
7. **C**
8. **B**
9. **A**
10. **A**
11. **C**
12. **A**
13. **D**
14. **D**
15. B

**Machine Learning – Assignment 7**

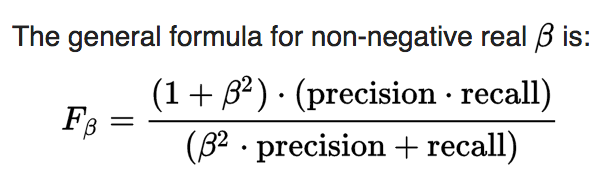
1. D
2. A
3. B
4. A
5. A
6. C
7. B
8. B
9. Entropy of dataset = -0.4log(0.4) – 0.6 log (0.6) = -0.4(-1.32) – 0.6 (-0.73)=0.52 +0.43= 0.96

Gini index= 1-[0.6\*0.6 + 0.4 \*0.4] = 0.48

1. Decision Trees are prone to overfitting whereas Random Forest removes this problem since it uses a voting mechanism to decide for the output hence combination of several decision trees reduces the overfitting problem.

If you wanted to put all your decision making in one decision tree, it can get very complicated for you to not only create it but also to maintain it.Random forests on the other hand normalize overfitting risk by analysing a random sub-sample of your data.

1. Most of the times, your dataset will contain features highly varying in magnitudes, units and range. But since, most of the machine learning algorithms use Eucledian distance between two data points in their computations, this is a problem. Hence scaling the data on one scale is good before feeding it to the model to ensure model receives same kind of data within a scale.
2. Since gradient descent is focused on convergence at global maxima therefore if we have value with larger range(large magnitude) the process for gradient descent will happen very slowly increasing the time .
3. IF dataset is imbalanced there are huge chances that model get trained on more percentage of one class or get validated on more percentage of one class and hence accuracy will be very high whereas model might be very poor.
4. F score is a metrics used where reduction of both False negative and false positive is very important. Its formula is as below:



Now if beta is ‘1’ the formula is called as F1 score or harmonic mean

F1 = 2(precision \* Recall)/(precision+Recall)

In case where FP is important to be reduced beta can be decreased below 1 (ex 0.5 ) and in cases where FN is important to be reduced beta can be increased from above 1 (ex 2).

1. Fit() method first fit the data to the object and transform() is used to transform the data in accordance with the object and return transformed values. Fit\_transform collectively does this task but if fit\_transform is used then the object cannot be utilized to transform any other data whereas if fit and transform are used separately then the object ca be used to transform any other data in accordance with the original data fitted to the object.