**STATISTICS WORKSHEET 7**

Answer 1. B) 0.135

Answer 2. D) 0.53

Answer 3. C) 0.745

Answer 4. B) 0.577

Answer 5. C) 0.6

Answer 6. C) 0.33

Answer 7. C) 0.33

Answer 8. B) 0.22

Answer 9. A) 0.66

Answer 10. A) 0.33

Answer 11. C) 0.5

Answer 12. A) 0.166

Answer 13. D) 0.25

Answer 14. D) 0.06

Answer 15. D) ¾

**SQL WORKSHEET 7**

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. The primary key is selected from the

A. Composite keys B. Candidate keys C. Foreign keys D. Determinants

**Answer.** B. Candidate Keys

2. Which is/are correct statements about primary key of a table?

A. Primary keys can contain NULL values.

B. Primary keys cannot contain NULL values.

C. A table can have only one primary key with single or multiple fields.

D. A table can have multiple primary keys with single or multiple fields.

**Answer**. B. Primary keys cannot contain NULL values, C. A table can have only one primary key with single or multiple fields.

Q3 to Q10 have only one correct answer. Choose the correct option to answer your question.

3. Which SQL command is used to insert a row in a table?

A. Select B. Create C. Insert D. Drop

**Answer.** C. Insert

4. Which one of the following sorts rows in SQL?

A. SORTBY B. ALIGNBY C. ORDERBY D. GROUPBY

**Answer.** C. ORDERBY

5. The SQL statement that queries or reads data from a table is

A. QUERY B. READ C. SELECT D. QUERY

**Answer**. C. SELECT

6. Which normal form is considered adequate for relational database design?

A. 1NF B. 2NF C. 3NF D. 4NF

**Answer**. C. 3NF

7. SQL can be used to

A. Create database structures only

B. Modify database data only

C. All of the above can be done by SQL

D. Query database data only

**Answer**. C. All of the above can be done by SQL.

8. SQL query and modification commands make up

A. DDL B. DML C. HTML D. XML

Answer. B. DML

9. The result of a SQL SELECT statement is a(n).

A. File B. Table C. Report D. Form

**Answer**. B. Table

10. Second normal form should meet all the rules for

A. 1 NF B. 2 NF C. 3 NF D. 4 NF

**Answer**. 2NF

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What are joins in SQL?

**Answer**. A join command in SQL is used to combine records from more than one table based on a common field between them.

12. What are the different types of joins in SQL?

**Answer**. Different types of joins are:

Right join - returns all rows from the right table, even if there are no matches in the left table.

Left join - returns all rows from the left table, even if there are no matches in the right table.

Inner join – returns rows when there is match in both the tables.

Full join - returns rows when there is a match in one of the tables.

Self-join - It is used to join a table to itself as if the table were two tables, temporarily renaming at least one table in the SQL statement.

Cartesian join - returns the Cartesian product of the sets of records from the two or more joined tables.

13. What is SQL Server?

**Answer**. SQL is a relational database management system developed by Microsoft that has been supporting business applications for multiple decades.

14. What is primary key in SQL?

**Answer**. A Primary key is a field in a table that uniquely identifies each record in a database table. Primary key must contain unique values. A table can have only one primary key, which may consist of single or multiple fields.

15. What is ETL in SQL?

**Answer**. ETL stand for **E**xtract, **T**ransform and **L**oad, is a process in which data is collected from various sources and transform the data depending on business rules/needs and load the data into a destination database.

**MACHINE** **LEARNING** **ASSIGNMENT** **- 7**

1. Which of the following in sk-learn library is used for hyper parameter tuning?

A) GridSearchCV() B) RandomizedCV() C) K-fold Cross Validation D) All of the above

**Answer.** A. GridSearchCV

2. In which of the below ensemble techniques trees are trained in parallel?

A) Random forest B) Adaboost C) Gradient Boosting D) All of the above

**Answer**. A. Random Forest

3. In machine learning, if in the below line of code: sklearn.svm.SVC (C=1.0, kernel='rbf', degree=3) we increasing the C hyper parameter, what will happen?

A) The regularization will increase

B) The regularization will decrease

C) No effect on regularization

D) kernel will be changed to linear

Answer. B) The regularization will decrease

4. Check the below line of code and answer the following questions: sklearn.tree.DecisionTreeClassifier(\*criterion='gini',splitter='best',max\_depth=None, min\_samples\_split=2) Which of the following is true regarding max\_depth hyper parameter?

A) It regularizes the decision tree by limiting the maximum depth up to which a tree can be grown. B) It denotes the number of children a node can have.

C) both A & B

D) None of the above

**Answer**. C. both A and B

5. Which of the following is true regarding Random Forests?

A) It's an ensemble of weak learners.

B) The component trees are trained in series

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

D)None of the above

**Answer**. A. It is an ensemble technique.

6. What can be the disadvantage if the learning rate is very high in gradient descent?

A) Gradient Descent algorithm can diverge from the optimal solution.

B) Gradient Descent algorithm can keep oscillating around the optimal solution and may not settle.

C) Both of them

D) None of them

**Answer**. C) Both of them

7. As the model complexity increases, what will happen?

A) Bias will increase, Variance decrease B) Bias will decrease, Variance increase C) both bias and variance increase D) Both bias and variance decrease.

**Answer**. B) Bias will decrease, Variance increase

8. Suppose I have a linear regression model which is performing as follows: Train accuracy=0.95 and Test accuracy=0.75 Which of the following is true regarding the model?

A) model is underfitting B) model is overfitting C) model is performing good D) None of the above

**Answer**. B) model is overfitting

Q9 to Q15 are subjective answer type questions, Answer them briefly.

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.

**Answer.** Entropy is 0.97 bits

Gini index = A/A+B = 0.4

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

**Answer.** A random forest is simply a collection of decision trees whose results are aggregated into one final result. Their ability to limit overfitting without substantially increasing error due to bias is why they are such powerful models. Random forests overcome several problems with decision trees, including:

* 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.

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

**Answer.** It is very common in a dataset to have columns with different units like one column can be age in years other can weight in kilograms. Also, we can have columns of salary which can range from 20,000 to 1,00,000 and experience in years which can range from 0-30 years. Income here is very large as compared to experience. When we feed these features to the model as is, there is every chance that the income will influence the result more due to its larger value. But this doesn’t necessarily mean it is more important as a predictor. So, to give importance to both Experience, and Income, we need feature scaling.

Two most common techniques used for this purpose are- Normalization and Standardization.

Normalization is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.

Standardization is another scaling technique where the values are centred around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation. It is also known as StandardScaler.

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

**Answer.** Gradient descentis an optimization algorithm used to minimize the cost function in machine learning algorithms like Logistic Regression, SVM, Neural Networks etc. If features are on different scale, certain weights are updated faster than others in Gradient Descent. However, feature scaling helps in causing Gradient Descent to converge much faster as standardizing all the variables on to the same scale

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?

**Answer.** Accuracy Metric is one the simplest and widely used metric to measure the performance of a classification predictive model. The reason for its wide use is because it is easy to calculate, easy to interpret, and is a single number to summarize the model’s capability. However, accuracy metric fails to perform on an imbalanced dataset as it gives misleading conclusions. In an imbalanced dataset getting an accuracy score of 90 or 99 are trivial as model might have considered the less numbered observation as error or outliers and could have ignored them in the prediction.

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

**Answer.** F1-score or F-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=

15. What is the difference between fit(), transform() and fit\_transform().

**Answer**.  The fit() method is used to fit the transformer like MinMaxSCaler to the input data and perform the required computations to the specific transformer we apply.

Now the transform() method of sklearn transformers, will transform the input data into some transformed spaced. The output is usually an array matrix with equal number of samples as the input data. The transformation will be performed based on the parameters that were computed during fit.

This fit\_transform() method is basically the combination of fit method and transform method, it is equivalent to fit().transform().This method performs fit and transform on the input data at a single time and converts the data points.If we use fit and transform separate when we need both then it will decrease the efficiency of the model so we use fit\_transform() which will do both the work.