STATISTICS WORKSHEET-7

01. B) 0.135

02. D) 0.53

03. C) 0.745

04. B) 0.577

05. C) 0.6

06. C) 0.33

07. C) 0.33

08. B) 0.22

09. A) 0.66

10. A) 0.33

11. C) 0.5

12. A) 0.166

13. D) 0.25

14. D) 0.06

15. D) ¾

MACHINE LEARNING

01: (A) GridSearchCV()

02: (A) Random forest

03: (B) The regularization will decrease

04: (C) both A & B

05: (A) It's an ensemble of weak learners.

06: (C) Both of them

07: (B) Bias will decrease, Variance increase

08: (B) model is overfitting

09: Entropy is 0.97 bits || Gini index = A/A+B = 0.4

10: **ANSWER**:

Random forests consist of multiple single trees each based on a random sample of the training data. They are typically more accurate than single decision trees. Decision trees have a low bias / are non-parametric, they suffer from a high variance which makes them less useful for most practical applications.

By aggregating multiple decision trees, one can reduce the variance of the model output significantly, thus improving performance. While this could be archived by simple tree bagging, the fact that each tree is build on a bootstrap sample of the same data gives a lower bound on the variance reduction, due to correlation between the individual trees. Random Forest addresses this problem by sub-sampling features, thus de-correlating the trees to a certain extend and therefore allowing for a greater variance reduction / increase in performance.

11: **ANSWER**:

In many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

Two most common techniques of feature scaling are Normalization and Standardization.

12: **ANSWER**:

Gradient descent is 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: **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: **ANSWER**:

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'.

$F1=2 * \frac{*precision*recall}{precision+recall}$

15: **ANSWER**:

The fit () method is used to fit the transformer like Min-Max Scaler 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.

SQL WORKSHEET 7

01: (B) Candidate keys

02: (B) Primary keys cannot contain NULL values

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

03: (C) Insert

04: (C) ORDERBY

05: (C) SELECT

06: (C) 3NF

07: (C) All of the above can be done by SQL

08: (B) DML

09: (B) Table

10: (B) 2 NF

11: **ANSWER**

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

12: **ANSWER**

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: **ANSWER**

SQL Server is a relational database management system, or RDBMS, developed and marketed by Microsoft. SQL Server works exclusively on Windows environment for more than 20 years. SQL Server is built on top of SQL, a standard programming language for interacting with the relational databases. SQL server is tied to Transact-SQL, or T-SQL, the Microsoft's implementation of SQL that adds a set of proprietary programming constructs.

14: **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: **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.