Worksheet Set 7 Solutions :

Machine Learning :

1. D) All of the above.
2. A) Random Forest
3. B) The regularization will decrease.
4. C) Both A and B
5. A) It is an ensemble of weak learners.
6. C) Both of them
7. B) Variance increases, Bias Decreases
8. C) Model is performing good
9. Percentage of Class A = 0.40
   Percentage of Class B = 0.60

   Entropy = - (0.40 * log2(0.40) - (0.60 * log2(0.60))
   = 0.971
   Gini Index = 1- ((0.4)*(0.4) + (0.6)*(0.6))
   = 0.48
10. Briefly, although 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.Robust to outliers.Works well with non-linear data. Lower risk of overfitting. Runs efficiently on a large dataset.Better accuracy than other classification algorithms.
11. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. MinMax Scaler and Standard Scaler are two techniques used for scaling the dataset
12. We can speed up gradient descent by scaling. 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.
13. In the framework of imbalanced data-sets, accuracy is no longer a proper measure, since it does not distinguish between the numbers of correctly classified examples of different classes. Hence, it may lead to erroneous conclusions.
14. In statistical analysis of binary classification, the F-score or F-measure is a measure of a test's accuracy. It is calculated from the precision and recall of the test, where the precision is the number of true positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of true positive results divided by the number of all samples that should have been identified as positive. Precision is also known as positive predictive value, and recall is also known as sensitivity in diagnostic binary classification.

The F1 Score = 2*((precision*recall) / (precision+recall)).

15. fit_transform() is used on the training data so that we can scale the training data and also learn the scaling parameters of that data. Here, the model built by us will learn the mean and variance of the features of the training set. These learned parameters are then used to scale our test data.The fit method is calculating the mean and variance of each of the features present in our data. The transform method is transforming all the features using the respective mean and variance.Now, we want scaling to be applied to our test data too and at the same time do not want to be biased with our model. We want our test data to be a completely new and a surprise set for our model. The transform method helps us in this case.

Statistics Solutions :

1. b) 0.135
2. d) 0.53
3. c) 0.745
4. b) 0.577
5. c) 0.6
6. b) 0.40
7. c) 0.33
8. b) 0.22
9. a) 0.66
10. a) 0.33
11. a) 0.33
12. b)0.34
13. d) 0.25
14. d) 0.06
15. d) ¾

SQL Solutions :

1. B) Candidate keys
2. B) and C)
3. C) Insert
4. C) Order By
5. C) Select
6. C) 3 NF
7. C) All of the above can be done by SQL
8. B) DML
9. B) Table
10. A) 1NF
11. A JOIN clause is used to combine rows from two or more tables, based on a related column between them.

12. The different types of the JOINs in SQL: (INNER) JOIN: Returns records that have matching values in both tables. LEFT (OUTER) JOIN: Returns all records from the left table, and the matched records from the right tablev. RIGHT (OUTER) JOIN: Returns all records from the right table, and the matched records from the left table . FULL (OUTER) JOIN: Returns all records when there is a match in either left or right table.
13. SQL Server is a database server by Microsoft. The Microsoft relational database management system is a software product which primarily stores and retrieves data requested by other applications. These applications may run on the same or a different computer.
14. The PRIMARY KEY constraint uniquely identifies each record in a table. Primary keys must contain UNIQUE values, and cannot contain NULL values. A table can have only ONE primary key; and in the table, this primary key can consist of single or multiple columns (fields).
15. ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources. It's often used to build a data warehouse.