K. K. Wagh Institute of Engineering Education and Research, Nashik.

Department of Computer Engineering

Academic Year: 2020 – 2021 Semester: I
Class & Div: BE –A and B
Course Name & Code: Laboratory Practice - I (410246)
Teaching Scheme: Practical (04 Hrs / week)
Name of Faculty: Prof. J. R. Mankar, Prof. A. V. Taware and Prof. N. G. Sharma

ASSIGNMENT 03 (DATA ANALYTICS) SAMPLE ORAL QUESTIONS

1. What is the type of dataset that is used as input to the program?
   Capitalshare Bikers Dataset

2. How many attributes / features are there in the dataset?
   - Duration Start date
   - End date
   - Start station number
   - Start station
   - End station number
   - End station
   - Bike number
   - Member type

3. What is the type of each attribute?

```
Column                    Non-Null Count   Dtype
---  ------               --------------   -----
 0   Duration             115597 non-null  int64
 1   Start date           115597 non-null  object
 2   End date             115597 non-null  object
 3   Start station number 115597 non-null  int64
 4   Start station        115597 non-null  object
 5   End station number   115597 non-null  int64
 6   End station          115597 non-null  object
 7   Bike number          115597 non-null  object
 8   Member type          115597 non-null  object
```

4. How many features is numeric?
   8
5. How many features are nominal / categorical?
   1
6. Which function is used to display summary statistics in python?
   Df.describe()
7. What is decision tree?
   Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems.
   Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

We can represent any boolean function on discrete attributes using the decision tree.

8. What is the impact of using all features Vs selective features for model fitting?
   Using selective features increases accuracy decreases processing time

9. What is the purpose of using LabelEncoder?
   Categorical to numeric , long numeric to short numeric
10. What are the applications of decision trees?
11. **Decision trees are used for handling non-linear data sets effectively.The decision tree tool is used in real life in many areas, such as engineering, civil planning, law, and business.**

12. How to decide which classification algorithm is best suited for given problem?
    1)Size of dataset:
    if the training data is smaller or if the dataset has a fewer number of observations and a higher number of features like genetics or textual data, choose algorithms with high bias/low variance like Linear regression, Naïve Bayes, or Linear SVM.
    If the training data is sufficiently large and the number of observations is higher as compared to the number of features, one can go for low bias/high variance algorithms like KNN, Decision trees, or kernel SVM.

    2) Time :
    Algorithms like Naïve Bayes and Linear and Logistic regression are easy to implement and quick to run. Algorithms like SVM, which involve tuning of parameters, Neural networks with high convergence time, and random forests, need a lot of time to train the data.

    3) Accuracy

13. What is confusion matrix?
14. How to calculate accuracy?
15. How do we evaluate precision?
16. How do we evaluate recall?
17. How do we evaluate score?
18. How do we define split ratio?
19. What will be the impact of scaling on the output?
20. What is difference between fit(), transform() and fit_transform()?
    For All above refer previous assignment
21. On what basis do you decide whether there is need for scaling / transformation?
    On presence of outliers , if present scale/transform

Prof. J. R. Mankar, Prof. A. V. Taware and Prof. N. G. Sharma
Course Teacher