Bansilal Ramnath Agarwal Charitable Trust's

Vishwakarma Institute of Information Technology, Pune-48

(An Autonomous Institute Affiliated to Savitribai Phule Pune University)

Department of Computer Engineering

Data Science and Machine Learning Problem Statements

Suggested List of Assignments

1. Perform the following operations using Python on a data set : read data from different formats(like csv, xls),indexing and selecting data, sort data, describe attributes of data, checking data types of each column. (Use Titanic Dataset).

2. Perform the following operations using Python on the Telecom_Churn dataset. Compute and display summary statistics for each feature available in the dataset using separate commands for each statistic. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles).

3. Perform the following operations using Python on the data set House_Price Prediction dataset. Compute standard deviation, variance and percentiles using separate commands, for each feature. Create a histogram for each feature in the dataset to illustrate the feature distributions.

4. Write a program to do: A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. (Implement step by step using commands - Dont use library) Use this dataset to build a decision tree, with Buys as the target variable, to help in buying lipsticks in the future. Find the root node of the decision tree.

5. Write a program to do: A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. (Use library commands)
According to the decision tree you have made from the previous training data set, what is the decision for the test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?

6. Write a program to do: A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. (Use library commands)
According to the decision tree you have made from the previous training data set, what is the decision for the test data: [Age > 35, Income = Medium, Gender = Female, Marital Status = Married]?

7. Write a program to do: A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. (Use library commands)
According to the decision tree you have made from the previous training data set, what is the decision for the test data: [Age > 35, Income = Medium, Gender = Female, Marital Status = Married]?

8. Write a program to do: A dataset collected in a cosmetics shop showing details of customers and whether or not they responded to a special offer to buy a new lip-stick is shown in table below. (Use library commands)
According to the decision tree you have made from the previous training data set, what is the decision for the test data: [Age = 21-35, Income = Low, Gender = Male, Marital Status = Married]?

9. Write a program to do the following: You have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16, 0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering with initial centroids as m1=P1 =Cluster#1=C1 and m2=P8=cluster#2=C2. Answer the following 1] Which cluster does P6 belong to? 2] What is the population of a cluster around m2? 3] What is the updated value of m1 and m2?

10. Write a program to do the following: You have given a collection of 8 points. P1=[2, 10] P2=[2, 5] P3=[8, 4] P4=[5, 8] P5=[7,5] P6=[6, 4] P7=[1, 2] P8=[4, 9]. Perform the k-mean clustering with initial centroids as m1=P1 =Cluster#1=C1 and m2=P4=cluster#2=C2, m3=P7 =Cluster#3=C3. Answer the following 1] Which cluster does P6 belong to? 2] What is the population of a cluster around m3? 3] What is the updated value of m1, m2, m3?

11. Use Iris flower dataset and perform following :

1. List down the features and their types (e.g., numeric, nominal) available in the dataset. 2. Create a histogram for each feature in the dataset to illustrate the feature distributions.

12.Use Iris flower dataset and perform following :

1. **Create a box plot for each feature in the dataset.**
2. **Identify and discuss distributions and identify outliers from them.**

13. Use the covid_vaccine_statewise.csv dataset and perform the following analytics.

a. Describe the dataset

b. Number of persons state wise vaccinated for first dose in India

c. Number of persons state wise vaccinated for second dose in India

14. Use the covid_vaccine_statewise.csv dataset and perform the following analytics.

A. Describe the dataset.

B. Number of Males vaccinated

C.. Number of females vaccinated

15. Use the dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Use the Seaborn library to see if we can find any patterns in the data.

16. Use the inbuilt dataset 'titanic'. The dataset contains 891 rows and contains information about the passengers who boarded the unfortunate Titanic ship. Write a code to check how the price of the ticket (column name: 'fare') for each passenger is distributed by plotting a histogram.

17. Compute Accuracy, Error rate, Precision, Recall for following confusion matrix (Use formula for each)

True Positives (TPs): 1	**False Positives (FPs): 1**
False Negatives (FNs): 8	**True Negatives (TNs): 90**

18. Use House_Price prediction dataset. Provide summary statistics (mean, median, minimum, maximum, standard deviation) of variables (categorical vs quantitative) such as- For example, if categorical variable is age groups and quantitative variable is income, then provide summary statistics of income grouped by the age groups.

19. Write a Python program to display some basic statistical details like percentile, mean, standard deviation etc (Use python and pandas commands) the species of 'Iris-setosa', 'Iris-versicolor' and 'Iris-versicolor' of iris.csv dataset.

20. Write a program to cluster a set of points using K-means for IRIS dataset. Consider, K=3, clusters. Consider Euclidean distance as the distance measure. Randomly initialize a cluster mean as one of the data points. Iterate at least for 10 iterations. After iterations are over, print the final cluster means for each of the clusters.

21. Write a program to cluster a set of points using K-means for IRIS dataset. Consider, K=4, clusters. Consider Euclidean distance as the distance measure. Randomly initialize a cluster mean as one of the data points. Iterate at least for 10 iterations. After iterations are over, print the final cluster means for each of the clusters.

22. Compute Accuracy, Error rate, Precision, Recall for the following confusion matrix.

Actual Class\Predicted class	cancer = yes	cancer = no	Total
cancer = yes	90	210	300
cancer = no	140	9560	9700
Total	230	9770	10000

23. With reference to Table , obtain the Frequency table for the attribute age. From the frequency table you have obtained, calculate the information gain of the frequency table while splitting on Age. (Use step by step Python/Pandas commands)

Age	Income	Married	Health	Class
Young	High	No	Fair	No
young	High	No	Good	No
Middle	High	No	Fair	Yes
Old	Medium	No	Fair	Yes
Old	Low	Yes	Fair	Yes
Old	Low	Yes	Good	No
Middle	Low	Yes	Good	Yes
Young	Medium	No	Fair	No
Young	Low	Yes	Fair	Yes
Old	Medium	Yes	Fair	Yes
Young	Medium	Yes	Good	Yes
Middle	Medium	No	Good	Yes
Middle	High	Yes	Fair	Yes
Old	Medium	No	Good	No

24. Perform the following operations using Python on a suitable data set, counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa), identifying missing values and filling in the missing values.

25. Perform Data Cleaning, Data transformation using Python on any data set.