**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: B.Tech\MBA.Tech\MBA.Tech AI

**Course: Machine Learning**

**Experiment No.02**

PART A

(PART A : TO BE REFFERED BY STUDENTS)

**A.1 Aim: To understand and implement d**ata exploration techniques using Pandas Library.

**Task 1:** **Perform Exploratory data analysis on Car dataset and write the inferences for each question.**

1. Read the Toyota.csv file into a DataFrame.
2. Explore size, shape, data types of each column in the dataset.
3. List down the columns of dataset
4. Find out ‘Fuel Type’ for the 4th row.
5. Find out value for second column for the 4th row.
6. Select all rows for column “Fuel Type”
7. Select all rows for columns “KM”, “HP” and “Automatic”
8. Display 1 to 5 rows for columns 2 to 4 (excluding row 5 and column 4)
9. Display the info of dataset and state your observations
10. Identify unique values for columns “KM”, “HP” and “Doors”
11. Create a new data frame, by replacing “??” with NAN
12. Replace the categorical values in the “Doors” column with its corresponding numeric value
13. Convert data types of columns “Doors”, “MetColor” and “Automatic” to int, and object
14. Identify the total number of null values in each column of the data set
15. Drop rows with null values
16. Identify total number of cars that runs on “Petrol”, “Diesel” or “CNG”
17. Identify mean of “KM” for the cars that runs on “Diesel”

**Task 2:**

**Perform one hot and label encoding on relationship column of “adults” dataset**

**A.2 Prerequisite:**

Python Programming, Pandas library

**A.3 Outcome:**

**After successful completion of this experiment students will be able to:**

* 1. Read different types of data files(csv, excel, text file etc.)
  2. Obtain metadata of given dataset
  3. Understand finding of null values and replacing null values
  4. Understand and implement class label encoding
  5. Understand and implement one hot encoding

**A.4 Theory:**

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA is to learn what our data can tell us. It generally starts out with a high level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modeling choices, such as by helping us decide which features to use.

**Pandas Library:**

**DataFrame** is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object. Like Series, DataFrame accepts many different kinds of input:

* Dict of 1D ndarrays, lists, dicts, or Series
* 2-D numpy.ndarray
* [Structured or record](http://docs.scipy.org/doc/numpy/user/basics.rec.html) ndarray
* A Series
* Another DataFrame

**Encoding:**

**One hot encoding:**

 One-hot encoding converts the categorical data into numeric data by splitting the column into multiple columns. The numbers are replaced by 1s and 0s, depending on which column has what value

**Label encoding:**

This approach is very simple and it involves converting each value in a column into a number.

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical.)***

|  |  |
| --- | --- |
| Roll No. | Name: |
| Class : | Batch : |
| Date of Experiment: | Date of Submission |
| Grade : |  |

**B.1 Task1**

**B.2 Task 2**

**B.4 Conclusion:**

*(Students must write the conclusion in their own words.)*