MIT-WPU
।। विश्वशान्तिर्ध्रुवं ध्रुवा ।।

Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

School of Computer Engineering and Technology

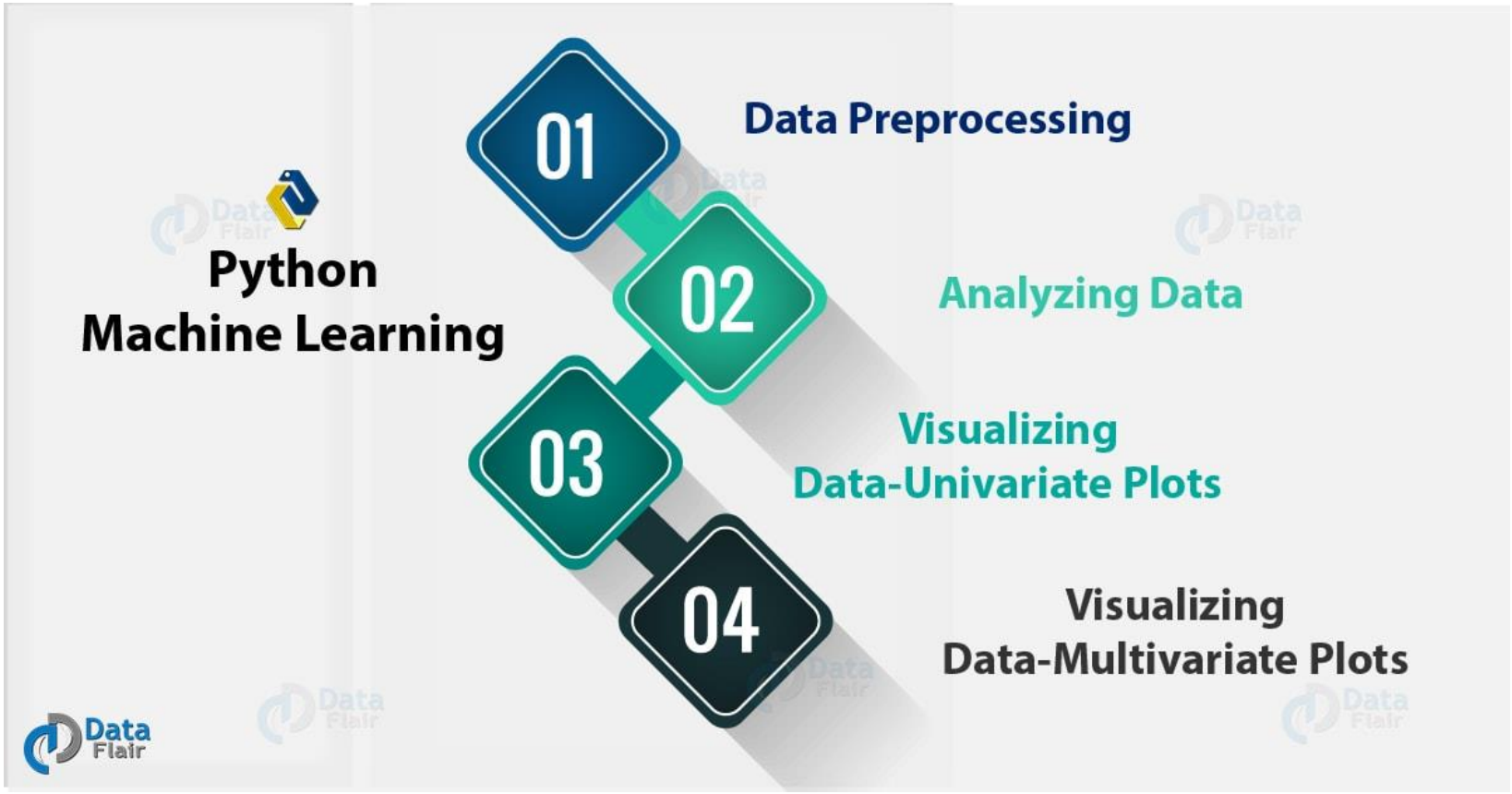
Big Data Analytics Laboratory

Lab Assignment-07

Write a python program to perform pre-processing on suitable dataset and illustrate various visualization techniques on suitable sample data. Analyze the same.

Index

- **Data Preprocessing steps in Python**
  - Importing the libraries.
  - Importing the Dataset.
  - Handling of Missing **Data**.
  - Handling of Categorical **Data**.
  - Splitting the dataset into training and testing datasets.
  - Feature Scaling.


Python
Machine Learning
01
Data Preprocessing
02
Analyzing Data
03
Visualizing
Data-Univariate Plots
04
Visualizing
Data-Multivariate Plots
Data
Flair

Step 1: Import Libraries

- Following are the key libraries that we will need to perform Assignment.
  - NumPy
  - SciPy
  - Pandas
  - SciKit-Learn
  - matplotlib
  - Seaborn
  - Bokeh
  - Altair
  - Plotly
  - ggplot
  - Eg: import pandas as pd

Step 2: Import the Dataset

- There are different file format commonly used to read data from
- .csv
- .xls
- .txt

```
data_csv=pd.read_csv('Iris_data_sample.csv')
```

dataset = pd.read_excel(‘age_salary.xls’)
dataset = pd.read_table(‘age_salary.txt’)

Methods for preprocessing data

- .head()
- .tail()
- .columns()
- .info()
- .describe()
- .dtypes()
- .index()
- fillna()
- dropna()
- isnull()
- isna()

- Demo Program

Methods description

A DataFrame is a 2-dimensional data structure that can store data of different types (including characters, integers, floating point values, factors and more) in columns.

df.attribute	description
dtypes	list the types of the columns
columns	list the column names
axes	list the row labels and column names
ndim	number of dimensions
size	number of elements
shape	return a tuple representing the dimensionality
values	numpy representation of the data

df.method()	description
head([n]), tail([n])	first/last n rows
describe()	generate descriptive statistics (for numeric columns only)
max(), min()	return max/min values for all numeric columns
mean(), median()	return mean/median values for all numeric columns
std()	standard deviation
dropna()	drop all the records with missing values

Introduction to Visualization

	description
distplot	histogram
barplot	estimate of central tendency for a numeric variable
jointplot	Scatterplot
regplot	Regression plot
pairplot	Pairplot

References

- https://data-flair.training/blogs/python-ml-data-preprocessing/
- Python for Data Analysis, Research Computing Services, Katia Oleinik (koleinik@bu.edu)
- https://blog.insightdatascience.com/data-visualization-in-python-advanced-functionality-in-seaborn-20d217f1a9a6