Experiment 1

Page No.
Date

Dr. D. Y. PATIL EDUCATIONAL FEDERATION Varale Campus

Title : DATA WRANGLING I

PROBLEM STEMENT :-

1. Locate open source data from web.
2. Provide clear description of the data
3. Load the Dataset into pandas data frame.
4. Data preprocessing.
5. Data formating and data normalization.
6. Turn categorical.

Objective :

1. To learn and concepts of python libraries.
2. To learn and understand data science.
3. To understand and practice.

Prerequisite :-

1. Basic python programing.
2. Concept of Data Preprocessing, Data Formatting.

Theory :-

1. Introduction to Big data :-

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing technique. Big data involves data produced by diffrent devices and applications.

Page No.
Date

Dr. D. Y. PATIL EDUCATIONAL FEDERATION
Varale Campus

2. Introduction to Dataset.

A dataset is a collection of records, similar to relational database table. Records are similar to table rows, but the columns can contain not only string

3. Python Libraries for Data science.

a. Numpy :-

One of the most fundamental packges in python. Numpy is general purpose array processing package. It provides high-Performance multidimentional array objects and tools to work with arrays.

What can do Numpy.'

1. Basic array operations: add, multiply
2. Advanced array operations: stack and
3. Work with DateTime
4. Basic slicing and Advanced indexing.

b. Pandas :-

- Pandas is an open-source package that provides high-performance, easy to use data

What can you do with Pandas:

1. Indexing, manipulating, remaning.
2. Update, Add, Delete columns.
3. Impute missing files.

Page No.
Date

Dr. D. Y. PATIL EDUCATIONAL FEDERATION Varale Campus

C: scikit learn :-

-Introduced to the world as Google summer code project. Scikit learn is robust machine learning library for python. It features various features like SVM's, random forests.

What can you do scikit learn?

1. Classification : Spam detection.
2. Clustering : Drug response, stock price.
3. Dimensionality reduction.
4. ~~Dimension~~ Model selection.

Description of Dataset :-

The iris dataset was used in R.A. Fisher's clasic 1936 paper. Measurements in Taxonomic problems, and can also be found on UCI machine Learning Repository. It includes three iris species with 50 sample each and as well as some properties about each flower.

The sample - 150

The columns in this dataset are:

1. Id
2. SepalLengthCm
3. SepalWidthCm.
4. PetalLengthCm.
5. Species.

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

NAAC

5. Pandas Data frame for Load Dataset

1. The dataset is downloads from UCI ~~we~~

2. Now read csv file as a Dataframe in python from path where you saved the same. The Iris Dataset is stored in .cs stands for comma seprated value. It is easier load .csv files in panda data frame and perform various analytical operations on it.
Load Iris.csv into a pandas data frame
Syntax:
iris = Pd.read_csv (csv_url, header=Non

3] Read the dataset from UCI machi Learning Repository link and specify columns name to use.

	Dataframe Function	Description
1	Dataset.head (n)	Return first n rows.
2.	dataset. ~~ind~~tail (n)	Return last n rows.
3.	dataset. index	The index of data
4.	dataset. columns.	The column lab of the dataset
5.	dataset. shape	Return a tuple representing the dimen. of dataset.
6.	dataset. dtype	Return the dtypes of dataset.
7.	dataset [column name]	Read data column wise.
8.	dataset. iloc [5]	Purely integer-location based indexing for selection.
9.	dataset [0:3]	selecting via [], which slices the rows.
10.	dataset. iloc [:, :n]	a subset of first n columns of original data
11.	dataset. iloc [:m, :n]	a subset of the first m rows & first n columns.

- missing values in data frame as a whole.
syntax: Dataframe. isnull()
- missing values across each columns
syntax: Dataframe. isnull. any()
- Count of missing values across each column using isna() & isnull()
~~Syntax~~ :- dataframe. isnull(). sum(). sum()

7. Pandas fundtions for Data ~~frame~~ ing and Normalization :-

The transforming data stage is about converting the data set into a format that can be analyzed or modeled ~~effectively~~.

Dataframe Function	Description	Output
① df.dtypes	To check the data type	df.dtypes
② df['Petal length (cm)'] = df['Petal length (cm)'].astype ("int")	To change the data type	df.dtypes

B) Data normalization :- Mapping all the ... values onto a uniform scale is involved ... data normalization. It is also known a... Min-max scaling.

Algorithm:

step 1: ~~Import~~ pands and sklearn libra...

step 2: Load Iris dataset in dataframe object df.

```
iris = load_iris()
df = pd.DataFrame(iris.data, columns = iris.features_names)
```

step 3: Print dataset

```
df.head()
```

step 4: create minimum & maximum Processor project.

step 5: x-scaled = min_max_scaler.

```
fit_transform(x)
```

step 6: view the dataframe

```
df_normalized.
```

8. Panda function for handling Categorical variables:

- Categorical variables have values that describe 'quality' or 'characteristics' of a data unit.
- Categorical features refers to string type data and can be easily understood by human beings. Therefore, categorical data must be translated into numerical data that can be understood by machine.

Label Encoding : Label Encoding refers to converting the labels into a numeric from so as to convert them into machine readable form. It is an important processing step for structured dataset in supervised learning.

eg:

Height
Tall
Medium
short

Height
0
1
2

- Label Encoding on iris dataset :-
- For iris dataset target column is species.
- PreProcessing. LabelEncoder :
- fit_transform(y):

Page No.
Date

Algorithm :-

step 1 : Import pandas and sklearn library for preprocessing.
from sklearn import Preprocessing

step 2 : Load the iris dataset in dataframe object df.

step 3 : observe the unique value for the species column.
df['species'].unique()

step 4 : define lable_encoder object known how to understand as lables
label_Encoder = Preprocessing.LabelEncode

step 5 : Encode labels in column.
df['species'] = label_Encoder.fit_transform (df['species

step 6 : observe the unique values for the species column.
df['species'].unique()

Conclusion :-

In this way we have explored the functions of Python library for Da Preprocessing, Data wrongling tec and How to handle missing values on Iris dataset.

Experiment 2

Page No.
Date

Dr. D. Y. PATIL EDUCATIONAL FEDERATION Varale Campus

Title :- DATA HRANGLING II

Problem statement :-
Create an 'Academic Performance' dataset of students and perform the following operations using python.
1. Scan all variables for missing values and inconsistencies.
2. Scan all numeric variables for outliers.
3. Apply data transformations on least one of the variables.

objective :-
Students should be able to perform data wrangling operation using python on any open source dataset.

PREREQUISITE :-
1. Basic python programing.
2. Concept of Data preprocessing, Data formating, Data normalization.

THEORY :-
1. Creation of dataset using MS Excel :-
The dataset created as "CSV" format.
- The name of dataset is studentsPerformance
- The features of dataset are: Math score Reading score, Writing score, Placement score.

- Number of instances : 30
- The response variable is: Placement_offer_co
- Range of values.
Math_score [60-80], club_join_date [2 01
Syntax :- RAND'BETHEN (bottom, up)

2. Identification and Handling of NULL :-
Missing data can be occur when no info is provided for one or more items or for while unit. Missing data is very big probl in real life scenarios. Missing data can also refer to as NA.
In pandas missing data is represented
1. None
2. NaN
Pandas treat None and NaN as essentially interchangable for indicating missing or NULL values. To faciliate this conve
- isnull()
- notnull()
- dropna()
- fillna()
- replace()

1. Checking for missing values using isnull() and notnull()
- checking for missing values using isnull()

Algorithm :-
Step 1: Import pandas and numpy in order to check missing value in Pandas dataframe.
import pandas as pd
import numpy as np
Step 2 : Load the dataset in dataframe object df
df = Pd.read_csv('student_performance.csv')
Step 3 : Display the dataframe
Step 4 : Use isnull() function to check null values in dataset.
df.isnull().
Step 5 : To create a series true for NaN values for specific columns.

- checking for missing values using notnull()

1. Algorithm
step 1 : Import pandas and numpy in order to check missing values in Pandas dataframe.
import pandas as pd
import numpy as np
step 2 : Load the dataset in dataframe
df = Pd.read_csv (student_performance.csv)
step 3 : Display dataframe.
df

Page No.
Date

3. Identification and Handling of outlie
One of the most imp steps as part of data preprocessing is detecting and dealing th outliers as they can negatively affect the statistical analysis and the training pr of machine learning algorithm.

1. What are outliers:-
We all have heard the idiom "odd one out" which means something unusual in comparison to the others in group.

2. Why do they occur?
An outlier may occur due to the variab in data, or due to experimental error human error.

3. Why do they affect?
- In statistics we have three measure of tendencies mean, median, mode.

4. Detecting outliers:-
If our dataset is small we can detect the outlier by at the dataset.
But what if we have huge dataset.
Below are some techiques:-
- Boxplots
- Scatterplots.
- Z score.
- Inner Quartile Range.

Page No.
Date

Detecting outliers using Boxplot:-
It captures the summary of data effectly and efficiently with only a simple box and whiskers. Boxplot summarizes sample data using 25th, 50th & 75th percentile.

Algorithm:-
Step 1: import pandas as pd.
import numpy as np
step 2: Load the dataset
df = pd.read_csv("demo.csv")
step 3: Display the data frame.
step 4: Select the columns for boxplot.
col = ['math score', 'reading score', 'writing score', 'Placement score']
step 5: He can now print the outliers foreach column with refernce
Print(np.where(df['math score']>90)

Handling of outliers:-
For removing the outlier, one must follow the same process of removing an entry from the dataset using its exact position in the dataset because in all the above methods of detecting the outliers end result is the list of all those data items.

Page No.

Date

Conclusion :- In this way we have explored the functions of the Python libring for data identifying and he the outlies. Data transformations T niques are explored with the purp of creating the new variable an reducing the skewness from datas

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Experiment NO - 3

Title :- Measure of Central Tendancy AND Variablity.

Problem Statement :-
Perform the following operations on any source dataset.
1. Provide summary statistics (mean, median, minimum, S.D) for a dataset with numeric variables.
2. Write a python program to display some basic statistical details like percentile, mean.

Objective :- To analyze and demonstrate knowledge of statistical data analysis technique for decision-making.

PREREQUISITE :-
① Basic of python programming.
② Concept of statistics mean, median, minimum, maximum, S.D.

THEORY :-
The data are summarized in some, but not all ways. We choose descripitive that are either most often reported covered in introducey courses.
Central Tendancy :-
- Mean
- Median
- Mode.

Page No.
Date

- Dispersion :-
- standard deviation.
- Minimum.
- Maximum.

Central Tendency :-

The mean :- $\bar{x} = \frac{\Sigma x}{N}$

Here, Σ represents summation.
x represents Observation.
N represents the no. of observ

Mean = $\Sigma fx / \Sigma f$
Where $\Sigma f = N$

The Median :-
If the total no. of obs (n) is ana o number, then formula given below.
Median = $\left(\frac{n+1}{2}\right)^{th}$ obs.

If the total no. of obs (n) is even number, the formula given below:
Median = $\frac{\left(\frac{n}{2}\right)^{th} \text{obs} + \left(\frac{n}{2}+1\right)^{th} \text{obs}}{2}$

- When data presented in the form of frequency distribution.

Page No.
Date

Find median class, the total Σf.
The median class consist of class in which $(n/2)$ is present.

Median = $l + \left[\frac{\frac{n}{2} - C}{f}\right] \times h$

The mode :-
The mode is most frequent occuring obs. Consider case where data is continous and the value mode be computed using following steps:
a) determine modal class that possesing maximum frequency.
b) calculate mode using formula

Mode = $l + \left[\frac{f_m - f_1}{2f_m - f_1 - f_2}\right] \times h$

Standard Deviation :-

$$S = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}}$$

Steps for calculating S.D :-
1. Step 1 : Find Mean.
2. Step 2 : Find each scores deviation from the mean.

Page No.
Date

step 3 : Square each deviation from the mean.

Step 4 : Find the Sum of Squares.

step 5 : Find the variance

step 6 : Find square root of variance.

Use of describe function
To display statistical details like percentile, mean, S.D etc.

Conclusion :- In this way we have explored the function of python library for calculate. Data statistics.

Page No.
Date

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

Experiment - 4.

Title - Data analytics - I

Problem statement :- Create a Linear Regression model using PYTHON/R to predict home price prediction using Boston Housing Dataset.

Objective :- Students should be able to data analysis using linear regression using python for any open-source dataset

Prerequisite :-
1. Basic of python programming
2. Concept of Regression

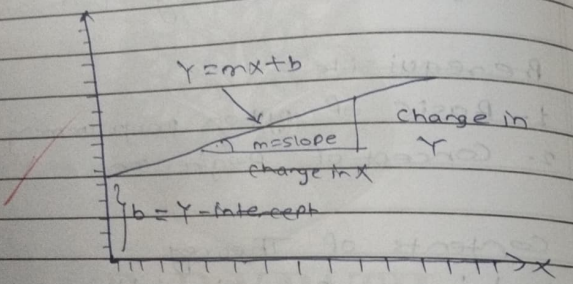
Contents of Theory :-
1. Linear Regression : Univariate & Multivariate.
2. Least Square Method for Linear regression
3. Measuring performance of LR.
4. Example of LR.
5. Training dataset & Testing dataset.

1. Linear Regression :-
It is machine learning algorithm based on supervised learning. It target predicted values on basis of independent variable.
- It is preferd to find out relationship betn forecasting and variables.

- Linear regression is popular because its cost function is mean squared error which is equal to average squared difference between observations actual and predict values.

equation of line like:

$$Y = m * X + b + c$$



Multivariate Regression :-

- It concerns the study of two or more predictor variables.
- Usually transformations of the original features into polynomial features.
- A simple linear model $Y = a + bx$ in original feature will be transformed into polynomial feature a linear regression applied it and, it something like.

$$Y = a + bX + cX_2$$

- If high degree value used in the curve becomes over-fitted

Dr.D.Y.PATIL EDUCATIONAL FEDERATION Varale Campus

2) List Square Method of Linear Regression

- Linear regression involves establishing linear relationship bern dependent & independent variables. Such relationship is portrayed in the form of eqn.
- A simple linear mode is is one of which involves only one dependent and one independent variable.
- However, for a simple univariate linear model, it can be denoted by the regression eqn.

3. Measuring Performance of Linear Regression. Mean square Error :-

- The mean squared Error (MSE) represent the error of estimator or Predictive model created based on given set of obs in the sample. Two or more regression models created using give sample data

$$MSE = \frac{1}{n} \Sigma (y - \bar{y})^2$$

An MSE of zero(0) represent the fact that the predictor is a perfect Predictor.

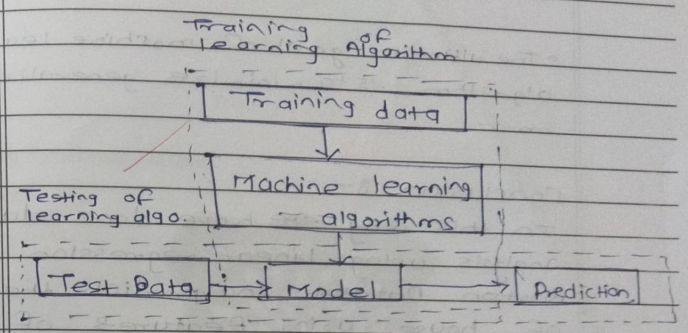
ii) Interpretion of regression line.

Interpretion 1
For an increase in value of x by 0.5
units there is an increase in value of y
in one unit.

Interception 2 :-
Even if X=0 value of independent variable
it is expected that value of Y is 26
score in XII standard (Yi) is 0.64
units depending on score in X standard.

Machine learning algorithm has two ph
1. Training and 2. Testing.

- The input of the training phase is training data, which is passed to any machine learning algo and machine learning model is generated as output of training phase.
- The input of testing phase is test d which is passed to the machine lea model and prediction is done.



Training & testing phase in machine learning :-

a) Training phase :-
- Training dataset is provided as input.
- Training dataset is dataset having attributes & class labels.

b) Testing phase :-
- Testing dataset is provided as input to this phase.
- Test dataset for which class label is unknown.
- A test dataset used for which class label

c) Generalization :-
- Generalization is the prediction of the future based on the past system.
- It needs to generalize beyond the train data to some future data.

Page No.
Date

Dr. D. Y. PATIL EDUCATION SOCIETY Varale Campus

- The ultimate goal of machine learning algorithms is to minimize generalization error.

Conclusion :-

In this way we have done data analysis using linear regression for Boston Dataset and predict the price of house using features of the Boston Dataset.

Continous Assesment of student :

[3]
[4]
[4]
[5]