DSBDAL-1

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Practicle No. 1

Title :- Data Wrangling. I.

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

Aim :-

Data Wrangling :- To perform the following operations using python on any open source dataset (eg. data.csv)

1. Import all the required python libraries.
2. Locate an open source data from the web provider a clear description of the data & its source.
3. Load the dataset into panda's data frame.
4. Data preprocessing : Check for missing values in it using pandas. isnull() describe function to get some initial statistics. Provide variable descriptions, types of variables, etc. Check the dimensions of the data frame.
5. Data formatting & Data Normalization : Summation of the type of variables by checking the data types of data frame.
6. Turn categorical variables into quantitative variables in python. In addition to the codes & o/p explain every operation that you do in the above steps & explain everything that you do to import / read / scrape the data set.

Requirements: 1. Basic of python programming
2. Concept of data preprocessing, data form
Data Normalization & Data cleaning.

Theory:
Data wrangling in python.
- Data wrangling is the process of gathering, collecti
& transforming raw data into another format for
better understanding, decision-making, accessing & ana
in cell time. Data wrangling is also known as Data
mugging.

Importance of Data Wrangling.
-. Data wrangling is a very important step. The be
example will explain its importance as:
Book selling website want to show top selling books o
diffrent domains, according to user preference. For
example a new user search for motivational book.
which sell the most or having a high rating, etc. Bu
on their website, there is plenty of raw data from cli
users. Here the concept of data mugging or data wra
is used. As we know Data is not wrangled by syst
This process is done by Data Scientists.
- Data wrangling in python is a crucial topic for da
science & data analysis. Pandas framework of python
used for data wrangling. Pandas is an open source
specifically developed for data analysis & data scien
The process like data sorting, Data filteration,
Data grouping, etc.

Data wrangling in python deals with the below functionalities.

1. Data Exploration: In this process, the data is studied, analysed & understood by visualizing representation of data.
2. Dealing with missing values: most of the dataset having a vast amount of data contain missing values of NaN. they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having NaN value.
3. Rephasing Data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.
4. Filtering Data: Sometimes datasets are comprised of unwanted rows or columns which are required to be removed or filtered.
5. Other: After dealing with the raw dataset as per our requirements & then it can be used for required purpose like data analysing, machine learning, data visualization, model training, etc.

Below is an example which implements the above functionalities on a raw datasets.

- Data exploration, here we assign the data & then we visualize the data in a tabular format.

- CSV file / Dataset :- titanic.csv

- Required python libraries.
  Numpy. Pandas. matplotlib. seaborn.

- Required syntax.

1. Load the dataset using Pandas data frame.
   df = pd.read_csv (titanic.csv)
   df

2. For showing top results.
   df.head()

3. For showing bottom results.
   df.tail()

4. Calculating Null values.
   df.isnull().sum()

5. Calculating Null Values in 'Age' & 'Cabin' column.
   df['Age'].isnull().sum()
   df['Cabin'].isnull().sum()

6. Get some initial statistics.
   df.describe()

7. Getting some information about dataset.
   df.info()

8. Finding Data types
   df.dtypes

9. Finding Dimensions of data frame.

```
df.shape.
```

10. Making impute function for filling Null values.

```
def imput_age (costs):
    Age = cots [0]
    PClass = cots [1]

    if pd.isnull (Age):
        if Pclass==1:
            return 37
        elif Pclass ==2:
            return 29
        else:
            return 24
    else:
        return Age.
```

- Displaying two-dimensional data in grid format where the color intensity represents value.

```
sns.heatmap()
sns.heatpmap (df.isnull(), yticklabels = False, cbar=fal
        cmap = 'viridis')
```

- Applying impute function & droping column.

cabin respectively

```
df['Age'] = df [['Age', 'Pclass']].apply (impute_age,ax
df.drop ('cabin', axis = 1, inplace = True)
```

- df.drop_duplicates()

- Data type conversions.
  df['Age'] = df['Age'].astype('int')
  df['Age'] = df['Age'].round(0).astype('int')

- converting categorial variables to Quantitative variables
  cat = pd.get_dummies(df, columns = ['sex'])

- Female = 0 & male = 1
  cat['sex_female']
  cat['sex_male']

- df.columns.

Conclusion :- Hence we have throughly studied how to perform the following operations on python on any open source dataset. (eg. data.csv).

1. Import all required libraries.
2. locate an open source data from the web, provide clear decription of data & its source.
3. load the dataset into pandas dataframe.
4. Data preprocessing.
5. Data formatting & data normalization
6. Turn categorial variables into Quantitative variables in python.

Q1. Explain Dataframe with suitable examples.

→ - Data frames are data displayed in a format as a table. Data frames can have diffrent types of data inside it.
- Example: lets have data of training pulse duration

	Training	Pulse	Duration
1.	Strength	100	60
2.	Stamina	150	30
3.	Other	120	45

Q2. What is the limitations of label encoding methods?

→ - limitations of label encoding method:-
It converts data in machine readable form, but it assigns a unique number to each class of priority issues in training of data sets. A label with a high value may be considered to have high priority than the label having lower value.

Q3. What is need of data normalization?

→ - The main objective of database normalization is to eliminate redundant data, minimize data modification errors & simplify the query process.

Q4. What are diffrent techniques for handling the missing data?

→ 1. Mean or median imputation
2. Multivariate Imputation by chained equations.
3. Random forest.

Q5. What is meant by data preprocessing?

→ 1. Checking of NULL values using pandas isnull() fu
2. By using describe() function to get some initial stati
3. Provide variable discription.
4. Type of variables.
5. checking the dimension of data frame.

Q6. What is meant by data wrangling?

→ 1) Data exploration.
2) Dealing with missing values.
3) Data reshaping.
4) Filtering data.

Q7. Importance of data wrangling?

→ - Improve data usability.
- converts data into compatible format.

Q8. What is meant by data wragling process?

→ Cleaning, organizing & enriching row data so that can be used for decision making process.

Q9. Use of pandas libraries.

→ Pandas is an open source library, used for:

1. Data cleaning.
2. Data Normalization
3. Data Visualization.
4. Data Inspection.
5. Data fill
6. Merges & joins
7. Statistical analysis.
8. loading & saving data

Q10.
→

Q11.
→

Q12

Q13

Q10. Uses of numpy library.

→ Numpy is numerical python library.

Uses:

1. Working with arrays.
2. Working in domain of linear algebra.
3. Fourier transformation.
4. Working with matrices. It is also open source library.
5. Working in vector - vector multiplication.

Q11. Uses of matplotlib library.

→ Matplotlib is used for data visualization & graphical plotting library (histogram, scatter, bar, charts, plots).

Q12. Uses of seaborn library.

→ Seaborn library is used for making statistical graphics in python.

Q13. Diffrence between matplotlib & seaborn.

→ Matplotlib.

- It is a python library used to plot various graphs with the help of additional libraries, like numpy & pandas.
- Matplotlib creates simple graphs, including histograms, bargraphs, pie-charts, scatterplots, lines & others visual representation of data.
- mainly used to plot 2D graphs of arrays.
- It is uses syntax that is relatively complicated & extensive.

Seaborn:

- It is also a python library that utilizes matplotlib pandas & numpy to plot graphs.
- It is a superset of matplotlib library.
- It has relatively simple syntax.
- eg. seaborn.barplot (xaxis, yaxis) syntax for bar graph. seaborn is more comfortable with pan data frames
  It prevents overlapping with the help of default

Q14. How to install any library in python programming.

→ pip install package-name.

eg: pip install seaborn..

DSBDAL-2

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Practicle No.2.

Title :- Data wrangling - II

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

Aim :- To create an "Academic Performance" dataset of students & perform the following using python.

1. Scan all variables from missing values & inconsistencies. If there are missing values &/or inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outliers. If these are outlier use any of the suitable techniques to deal with them.
3. Apply data transformations on atleast one of the variables. The purpose of the transformation should be one of the following reasons: to change the scale for better understand of the variable, to convert a non-linear relation into a linear one or to decrease the skewness & convert the distribution into a normal distribution.
4. Reason and document your approach properly.

Prerequisite :- 1. Basics of programming.
2. concept of data preprocessing, Data formatting, Data Normalization & Data Cleaning.

Theory :- Detailed explaination of exploratory data analy using Iris dataset.
For complete code please visit:
https://github.com/Naidu-Bharyal/Exploratory-Data-Analysis-on-Iris-Dataset.

CSV file/ Dataset - Academic performance

- Required libraries
  - Pandas, Numpy, matplotlib, Seaborn, math.

- Functions used.
  - Create dataset by using roll-no, marks, name, gr

- df = pd.dataframe({"roll-no": rollno, "name": name "Marks": marks, "Grade": grade})

- df.info()

- df.describe()

- df.datatypes

- df.columns , df.read-csv("academic.csv")

- df.isnull().sum()

- first-outlier
- second-outlier

- df.loc[15] = first_outlier
- df.loc[16] = second_outlier
- sns.countplot()
- sns.boxplot()
- df = df.drop()

Scaling the marks column :- import minmax scaler from sklearn preprocessing. scalar = minmax.scalar()
df[(marks)] = scalar.fit_transfer(df[['marks']]).

Conclusion :- Hence we have throughly studied the / how to perform the following operations using python on any open source dataset. (eg. data.csv).

1. Import all the required python libraries.
2. Locate an open source data from the web. Provide a clear description of data & its source.
3. Load the dataset into pandas data frame.
4. Data preprocessing : Check ~~for~~ for missing values in the data using pandas isnull(), describe() function to get some initial statistics. check the dimension of the data frame.
5. Data formatting & data normalization : Summarize the type of variables by checking the datatype of the variables in the dataset. If variables aren't in correct datatype.

apply proper type conversion.
6. Turn categorial variables into quantitative variable in python. In addition to the codes & o/p, explain every operation than you do in the above steps & e everything that you do to import/ read/ scrape the d

Q1. What is exploratory data analysis?
→ 1. Exploratory data anylysis is a task of analysing t data using simple tools from statastical, some plott tools, linear algebra.
2. Exploratory data analysis is a crucial step before you jump to maintaine learning or modeling of your d By doing this you can get to know whether the sele features are good enough to model, are all the feat required, are these any correlations based on which we either go back to the data preprocessing step or more to modeling.
3. Once exploratory data anylysis is complete & insights are drawn, its features can be used for supervised & unsupervised machine learning modeling.

Q2. Importance of EDA.
→ Many Data scientists will be in a hurry to get to the machine learning stage, some either entirely skip explora process or do a very minimal job. This is a mistake with may many implications, including generating inaccurate models, generating accurate models but on the wrong data not creating the right type of variables in data preperati

and using resources inefficiently because of realizing only after generating models that perhaps the data is skewed or has outliers, or has too many missing values or finding that some values are inconsistent.

Q. which parameters used to create "Academic performance" datasets.

→ "roll.no", "name", "marks", "grade".

Q. which library is used for scaling the marks column.

→ import minmax scalar from sklearn preprocessing

scalar = minmaxscalar()

df['marks'] = scalar.fit_transform[df['marks']]

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.
Accredited by NAAC

Practicle No.:3.

Title:- Descriptive statastics - measures of central Tendancy of variability.

Aim:- Descriptive statastics - To measure of central tendancy & variability.

- Perform the following operations on any open source dataset.

1. Perovide summary statistics for a dataset with numeric variables grouped by one of the quantitative variable. For example, if your categorial variable is age, groups & quantitative variable in income, then provide summary statastics of income grouped by the age group. Create a list that contains numeric value for each element to the categorial variable.
2. Write a python program to display some basic statastical details like percentile, mean, standard, derivation, etc. of the species of 'Iris-setosa', 'Iris-versicolor' & Iris.csv dataset.
3. Provide the codes with o/p & explain everything that you do in this step.

Objectives: Students should be able to perform the statastical operations using python on any open source dataset.

Requirements: 1. Basics of python programming.
2. Concept of Data preprocessing, Data formatting, Data normalization & cleaning.

Theory :-

Introduction :- Descriptive statastics is the building block of data science. Advance analytics is often incomplete without analysing descriptive statastics of the key metr In simple terms, descriptive statastics can be defined the measures of the central tendancy & these measu can be broken down further into the measures of cen tendancy & the measures of dispersion.

Measures of central tendancy include mean, median & while the measures of variability include standard de variance & the interquartile range. In this guide, y will learn how to compute these measures of descrip statastics & use them to interpret the data.

We will cover the topics given below.

1. Mean 2. Median 3. Mode 4. Standard Deviati 5. Variance 6. Interquartile range 7. Skewness.

we will begin by loading the dataset to be used in guide..

Data : In this guide, we will be using fictitious data loan applications containing 600 observations & 10 var as described below :

1. Marital_status : whether the applicant is marrie ("Yes") or not ("No").
2. Dependents : Number of dependants of the applica
3. Is_graduate : whether the applicant is graduate. ("Yes") or not ("Not").

4. Income : Annual Income of the applicant (in USD).
5. Loan-amount : Loan amount (in USD) for which the application was submitted.
6. Term_months : Tenure of the loan (in month).
7. Credit-score : Whether the applicants credit score was good ("satisfactory") or not ("Not satisfactory").
8. Age : The applicants age (in years).
9. Sex : Whether the applicant is female or male.
10. Approval_status : Whether the loan applicantion was approved.

- CSV file / Dataset - Iris Dataset.

- Required libraries.
  - Pandas, Numpy, Matplotlib, Seaborn, Sklearn.

- Functions used

- df = pd.read_csv ("Iris-data-set.csv")

- df.head()

- df.shape()

- df.info()

- df.describe()

df.isnull().sum()

plt.show()

- mean = grouped_df.mean()
- median = grouped_df.median()
- min = grouped_df.min()
- max = grouped_df.max()
- std = grouped_df.std().
- df.skew()

- Do all operations on each column. Also Draw boxpl for each column.

Conclusion: In this guide, you have learned about th fundamentals of the most widely used descriptive sta & their calculations with python. we covered the follou topics.

1. mean 2. median 3. mode 4. Standard Deviatio 5. Variance 6. Interquartile range 7. skewness.

It is important to understand the usage of these stata & which one to use, depending on the problem statement & the data.

To learn more about data preperation & building machine models using python's scikit-learn library, please refer to the following guides.

1. Scikit machine learning.
2. Ensemble modeling with scikit-learn.

Q.1 Which term include under measures of central tendancy?

→ Mean, Median, Mode.

Q2. Which terms include under measures of variability?

→ Standard Deviation, variance & interquartile range.

Q3. Describe Iris Dataset.

→ Iris is a collection of instruments, materials, stimuli, data & data coding & analysis tools used for research into languages, signed language learning, etc. It contains four features (length & width of sepals & petals) of 50 samples of three species of the Iris. This dataset contains five columns such as Petal length, Petal width, Sepal length, Sepal width & species type & rows being the samples.

Q4. Explain mean.

→ mean represents the arithmatic average of the data.

Q5. Describe measures of central tendancies.

→ It describe the centre of the data & often represented by mean, median, mode.

Q6. Explain median.

→ median represents the 50th percentile or the middle of the data, that seperates the distribution into two

Q7. Explain mode.

→ It represents the most frequent value of a variable the data.

Q8. Explain Standard Deviation.

→ It is a measure that is used to quantify the amount variation of a set of data values from its mean. A standard deviation for a variable indicates that the points tend to the close to its mean & vice versa.

Q9. Explain Variance.

→ It is square of standard deviation.

Q10. Explain Interquartile Range (IQR).

→ measure of statistical dispersion. It is calculated as the di between upper quartile (75th percentile) & the lower quartil IQR is very important measure for identifying outliers & c visualize using boxplot.

Q12. Explain term outliers.

→ outliers are always given wrong direction for your expected re outliers always talk about extremities too small or too large.

Q13. What is skewness.

→ It is used to measure of symmetry or lack of symmetry.

Practicle No 4.

Title:- Data Analytics I

Objectives:- Student should be able to data anylysis using linear regression using python for any open source dataset.

Aim:- Create a linear regression model using python IR to predict home prices using Boston Housing dataset. (https://www.kaggle.com/c/boston-housing). The Boston housing dataset contains information about various houses in Boston through diffrent parameters. There are objective is to predict the value of prices of the house using the given features.

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

Theory:-
Boston Housing with linear regression with this data our objective is create a model using linear regression to predict the houses prime.

The data contains the following columns:-
1. 'crim' : per capital crime rate by town.

2. 'zn' : Properties of residential land zoned for

lots over 25000 sq. fit.

3. 'inclus' : proportion of non-retail business acers/ town.

4. 'chas' : charles river dumy variable C=1 if tract bounds river i 0 otherwise).

5. 'nox' : nitrogen oxides concentration (parts per 10 million).

6. 'rm' : average number of ~~roms~~ rooms per dwelling.

7. 'Age' : Proportion of owner-occupied unites built prior to 1940.

8. 'dis' : weighted means of distances to five Boston employment centers.

9. 'tax' : full-value property-tax rate per $10000.

10. 'rad' : index of accessibility to radial highways

11. 'ptratio' : pupil-teacher ratio by town.

Conclusion :-

Thus the we learn about how to analysis data using ~~learn~~ linear regression using python.

Practicle No. 5

Title :- Data Analytics II

Objectives :- Students should be able to data analysis using logistic regression using python for any open source dataset.

Aim :- Data Analytics II
i) Implement logistics regression using python/R to perform classification on social-Network_Ads.csv dataset.
ii) Compute confussion matrix to find TP, FP, TN, FN Accuracy, Error rate, precision.

Requirements :-
i) Basic of python programming.
ii) Concept of Regression.

Theory :-

Logistic Regression : Social Network - Ads
- This project will be a walkthrough of a simple logistic regression model in an attempt to strategies a basic ad-targetting compaing for a social media ~~n~~ network.
- One of our sponsers advertisements seems to be particularly successful among our older welthier users but seemingly less so with your younger onces.

- we'd like to implement an appropriate model so that we know who our target audience is for this specific advertisement thus maximize our click-through rate.
- we'd like to show younger users this ad with a lower probability.
- our dataset contains some information about all of our users in social network..
- The last column of the dataset is a vector of booleans describing whether or not each individual ended up clicking on adv (0 = false, 1 = True)
- let's import the libraries & dataset & establish which variable are dependent or independent.
- worried about how the users Age & estimated salary effect the decision on click or not click on adv. extracting relevant vectors.
  Independent variables (X).
  Dependent variables (Y).

- Split data into two set: train to learn m/c for & test set for m/c to execute on.
- To get the most accurate results a common tool within m/c learning models is to apply feature scaling. It is also knowns as the data normalization & is generally per-formed during the data preprocessing step.
- Scikit learn, again has a helpful library called standardize scalar that will quickly preprocess the data for us in this manner.

- Now we are ready to build our logistic regression model. We create an object of the logistic regression () class.
- We then fit the classifier to the training set with the apply named fit() method so that it can understand correlation between x & y.
- lastly we will test the classifiers predictive power on the test set.
- The logistic regression predict() method will give us a vector of prediction for our dataset x-test.
- we can form the first ten values for y-pred that only this eight & tenth individuals within the index are predicted to click ~~and~~ on the adv.
- We can start en inferences about relationships now.
- We can think of 0 as the center of our normal distribution & consider any value in y-test.
- We'll print out the 20 values adjacent to each other to get a sneak-peak of how close our classifier come to a 100% prediction rate.
- When you run the cell below our mode has 19/20 or 95% predication rate.
- This is great start but we'll have to look at the rest of the data as well.
- Now, we can use confusion matrix to evaluate exactly how accurate out logistic regression model is.

- This graph help us see the clear correlation between the dependent & independent variab
- Intuitively, this graph makes alot of sens because user can quickly tell that about 80- of the observations have been correctly ident
- Now let's map the test set result to visuali where our confusion matrix came from.
- The best x-intercept is probably clear to l closer to 1 that it is 2 & the y-intercept likely between 243.

Conclusion:-

The confusion matrix tells us that were 89 correc prediction & 11 incorrect ones, meaning the m overall accomplished an 89% accuracy rating.

- This is very good & there are many ways to improve the model by parameter turning & sampl size increasing, but those topics are outside the scope of this project.

Practicle No. 6.

Problem Statement: Data Analytics III

1) Implement simple Naive Bayes classification also using python / R as iris.csv dataset.
2) Compute confusion matrix to find TP, FP, TN, FN accuracy, Error rate, Precesion, Recall on the given dataset.

Theory :-

Naive Bayes is a statistical classification technique based on the Bayes Theorem & one of the simplest supervised learning algorithm. The naive Bayes classifier is a quick, accurate & trust worthy method, especially on large datasets.

Simple formula of Bayes Theorem :

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}$$

Where, P(A) & P(B) are two independent events & P(B) is not equal to zero.

- P(A/B) is the conditional probability of event A occuring given that B is true.
- P(B/A) is the conditional probability of event B occuring given that A is true.

- P(A) & P(B) are the probabilities of A & B ocuring independently of another.

What is Naive Bayes classification?

- The Naive Bayes classification algorithm is a probabilitic classifier & it belongs to supervi learning.
- It is based on probability model that inco strong independent assumption.
- Therefore, they are considered naive.
- Another assumption made by the naive Ba classifier is that all the predictions have equal effect on the outcome.
- The Naive Bayes classification has the followi diffrent types-
- multinominal Naive Bayes method is a common Bayesian learning approach is natural language processing.
- Using the Bayes theorem, the progrom estimate the tag of a text, such as an email.
- It access the likehood of each tag for ag sample & returns the tag with the highest possibility.
- Bernouli Naive Bayes is a part of the family of Naive Bayes.
- It is only take binary values. There may be multiple features, but each is assumed to be a binary-valued variable.

- Therefore, this class requires samples to be represented as binary-valued Feature vectors.

Gaussian Naive Bayes :-
- It is a variant of Naive Bayes that follows guassian normal distribution & supports data.
- To build a simple model using guassian navice naive bayes, we assume the data is characterized by a gaussian distribution with no convariance between parameter.
- This model may be fit by calculating the mean & standard deviation of the points within each label.

Conclusion :-
- In this, we use Iris flower dataset to implement simple Naive Bayes classification algorithm. Use sepal-length sepal width as input & class is an output.

Practicle No-7

Problem Statement:- Text Analytics.
1) Extract sample document & apply following document preprocessing methods:
   a. Tokenization
   b. Pos Tagging.
   c. stop words removal.
   d. Stemming and Lemmatization.
2) Create representation of document by calculating Term Frequency and Inverse Document Frequency.

Theory:-
Step 1:-
- Extract sample document ie. extracting text from Doc file.
- Here we will extract text from the Doc file using docx module.

A) Tokenization.
   1) Tokenization with NLTK:
   - NLTK stands for Natural Language Toolkit.
   - This is suite of libraries & the program from statistical natural language processing for english written in python.

B) Pos Tagging:-
   - Pos Tagging is a process to make up the words in text format for a particular part of speech

based on its definition & context.
- It is responsible for text reading a language
assigning some specific token to each other
- It also called grammatical tagging.

Steps involved in the pos tagging example:
1) Tokenize text (word-tokenize)
2) Apply pos-tag to above step that is NLTK-p
tag (tokenize text).

NLTK Pos tags examples are as belong below:

Abbreviation	Meaning
CC	Coordinating Conjuction
CD	Cardinal digit
DT	determiner
EX	Existential there
FW	Foreign word.
IN	Preposition conjunction
JJ	This NLTK Pos tag
JJR	objective.

TF-IDF.
from sklearn.feature extraction.text import
TFidvectorovizer. vectorizer = Tf idvectorizer(

List of the document preprocessing methods.
1) Tokenization 2) Pos tagging 3) Stop words
4) Stemming 5) Lemmatization.

CSV file / dataset:
Required Libraries.
import nltk
nltk.download("punkt")
nltk.download("stopwords")
nltk.download("wordnet")
nltk.download("averaged_perception_tagger")

1. Tokenization
- from nltk import word_tokenize, sent_tokenization
from corpus.

2. POS tagging
- from nltk import pos_tag
tokens = word_tokenize.

3. Stop-word Removal.
- from nltk.corpus import stopwords stop_words_set()
stemming.

4. from nltk stem import portest zemmer stemmer =
poster stemmer()

5. Lemmatization
- from nltk stem import wordnet lemmatize
together the diffent inflected forms of a word.
- Lemmatization is similar to stemming but it
brings context to words. so it links words
with similar meaning to one word.

Application of lemmatization are :
1) Used in comprehesive.
2) Used in compact indexing..

Example of lemmatization
→ rocks : rock
→ corpard : corpus
→ better : good..

Formula :
tf (t,d) : count of a t in a / number of words in d..

Conclusion :- Pune District Education Association
IF IDF is the inverse of the document frequency which measures the informative o term t.

Practicle No 8.

Problem statement :- Data Visualization I

1. Use the inbuit dataset 'titanic'. The dataset contains 891 rows & containes information about the passangers who boarded the unfortunate titanic ship. Use the seaborn library to see if we can find any patterns in the data.

2. Write a code to how the price of the ticket (column name : 'Fare') for each passager is distributed by plotting a histogram.

Theory :-

Seaborn is a library mostly used for the statistical plotting in python. It is built on top of matplotlib & provides beutiful defaults styles & color patterns to make statistical plots more attractive.

Diffrent categories of plot in seaborn :

1) Relational plots :-
This plot is used to understand the relation between two variables.

2) Categorial plots :
- This plot deals with categorized variables & how they can be visualized.

3) Distributed plots:
This plot is used for examining univariat
bivariate distribution.

4) Regression plots:
The 2 regression plots in seaborn are
primirarly intended to add a visual guid
that helps to emphasizes patterns in
dataset during exploratory data analysis.

5) Matrix plots:
A matrix plots is an array of scatter

6) Multi-plot-grids.
It is useful approach is to draw multi
instances of the same plot an efficient
subset of the dataset.

The dataset consists of 891 rows & 12 col
1) Passanger ID
2) Survived
3) Pclass
4) Name
5) Sex
6) Age
7) SibSp
8) Parch
9) Ticket
10) Fare
11) Cabin
12) Embarted.

The seaborn library is built on top of matplotlib & offers many advanced data visualization capabilities.

- Features :
  - The titanic dataset has roughly the following type of features.

- Categorial / Nominal :
  - Variables that can be divided into multiple categories but having no order or priority.
  Eg: Embarked (C = Cherbrug ;
  Q = Queenstown ;
  S = Southampton)

- Binary :
  - A subtype of categorial features, where the variable has only two categories.
  Eg. Sex (male / Female).

- Ordinal.
  - They are similar to categorial feature but they have an order (ie. can be started)
  eg.. Pclass (1,2,3).

- Continous :
  - They can take up any value between the minimum & maximum values in the column..
  - eg. Age, Force.

• count:
They represent the count of a variable
Eg. Sibsp, parch.

• useless.
They dont contribute to the final outcom
of an ML model.
eg. Here, passangerID, Name, Cabin & Ti

• Distribution plots:
Distribution plots as the name suggests
type of plot that shows the statistical
distribution of data.
This distplot() shows the histogram distri
of data for a single column. The column n
is passed as a parameter to the distplot()
function.
We can see that most of the tickets
have been sowed between 0-50 dollers
The line that you see represents the
kernel density estimation. You can remove
the line by passing false as the paran
for the kde attribute as shown below:

```
sns.distplot(dataset['fare'].kde = false
```

```
sns.distplot(dataset['fare'].kde = false,
                                bins = 10
```

- Here we set the no. of bins to 10. In the
o/p you will see data distributed in 10 bins
shown below in the output.

Histogram :- Histogram are visualization tools that represent the distribution of set of continous data.
In a histogram, the data is divided into a set of intervals or bin & the count of data points that fall into each bin. These bins may or may not be equal in width but are adjacent (with no gaps)

Syntax :-
seaborn.histplot(data x, y, hue, stat, bins, bin-width, discrete, kde, log-scale).

Parameters :-
Data :
- Input data in the form of Dataframe or Numpy array.

x, y. [optional] :
- key of the data to be positioned on the x & y are respectively.

hue (optional) :
- Semantic data key which is mapped to determine the etc color of plot elements.

stat (optional) :
- count, frequency, density or probability.

Return:-
- This method returns the matplotlib axes with the plot drawns on it.

Conclusion:-
Here we defined the patterns using the seaborn library and also drawn a histogram.

Practicle No-9.

Problem statement :- Data Visualization II.

1. Use the inbuilt dataset 'titanic' as used in the above problem. Plot a box plot for distribution of age with respect to each other along with information about whether they survived or not.

2. Write observations on the inference from the above statistics.

Theory :-

What is Data Visualization ?

- Data visualization represents the text or numerical data in visual format which makes it easy to group the information the data express.
- We humans remember the picture more easily than resonable readable text, so python provides us various libraries for data visualization like matplotlib, seaborn, ploty.

Exploratory Data analysis :

Creating Hypothesis testing various to business assumptions while dealing with any ML learning problems statements in very important & this is about EDA stat, bing, bin width, discrete kde, long-scale.)

- parameters :
  Data : input data in the form of data forms o
  Numpy array.

- x, y (optional) : key of the data to be positioned
  on the x & y axes respectively..

- hue (optional) : semantic data key which is
  mapped to determine the ~~color~~ color.

- stat (optional) : count, frequency, etc.

1) Return : This method returns the matplotlib axe
   with plot ~~drawn~~ on it.

2) Distplot : Distplot is also known as the second
   Histogram because it is a slight.

3) Boxplot : Boxplot is a very interesting plot tha
   individually plots 0 - 5 numbers summary.

- median middle value is series after sorting..

Categorial :
1) Heatmap
2) Clustermap
3) Boxplot

sns.boxplot (data ('sex'), data ("age"))

sns.boxplot (data ('sex'), data ('age') data ('survived'))
plt.show()

- Now along with gender I also want to know whether the customer was a smoker or not so we can do this.

sns. scatterplot (tips [('total-bill")], tips ["tip"] hue = tips ["sex"], style = tips [" smoker"])
plt.show()

Numerical & categorial :

Bar plot :
- It is a simple plot which we can use to plot categorial variable on the x-axis & numerical variable on y-axis & numerical explore the relationship between both the variable.

sns.barplot (data [pclass], data ['Age'])
plt.show()

sns.barplot (data ['plclass'], data ['fare'], hue = data ['sex'])
plt.show()

Bivariate / multivariate Analysis:

- We have study about various plot to explore categorial & numerical data.
- And when we analyze more than 2 variable together then it is known as multivariate ana

Numerical:

Scatter plot:

- To plot the relationship both two numerical variables, scatter plot is a simple plot to d
- Relation ship between the total bill & tip sc
scatterplot [tips & {"total_bill"}. tips ["tip
- multivariable analysis with scatter plot:
- we can also plot 3 variable or a variable relationship with scatter plot.
- Suppose we want to find the seperate male & female with the total bill & tip pro
(Iris set as a, Iris virginica & Iris versicolor).

- The random forest classifier is then trained the training set & used to make prediction on testing set finally. The model is accuracy is calculated using the accuracy - score function from scikit - learn library & printed out to the console.
- The accuracy score represents the proportion o correctly classified instance out of all instances in the testing set.

Conclusion :- we have learn the data visualization techniques on 'titanic' dataset using various python libraries.

Practicle No. 10.

Problem statement:- Data visualization III

Download the Iris flower dataset or any other dataset into a Dataframe.
(eg. https://archive.ics.uci.edu/dataset/iris)
Scans the dataset & given the inference as:

1) List down the features & their types (Example numeric, nominal) available in dataset.

2) Create a histogram for each feature in the dataset to illustrate the feature the dataset distribution.

3) Create a box plot for each feature in the dataset to illustrate the feature the dataset distribution.

4) Compare distribution & identify outliers.

Theory:-

The Iris dataset contains feature of 50 samples of three species of iris.
(Iris setosa, iris virginica & iris versicolor)

These measure were used to make or create linear discriminant model to classify the species.

- The iris dataset is often used as the training dataset in machine learning & the machine

learning algorithms are trained to classify the different iris species based on the characteristics of the dataset.

- This is the often used as an example of multiclass classification.

The dataset is first split into training & testing test, where to of the data is used for training & 30% of for testing..

The random forest classifier is then trained on the training set & used to make prediction on the testing set.

- Finally, the model is accuray is calculated using the accuray score function from the skicit learn library & printed out to the conso

- The accuray score represents proportion of correctly classified instances out of all instan is the testing set.

- An accuracy score of 100% would indicate tha the model has classified all instance correctly while an accuray score of 0% would indicate that the model has classified all instances correctly while an instances incorrectly..

Q. Why Iris dataset is so popular?

- The iris dataset is popular dataset for severa reasons.

- First it is simplest of easy to understand dataset.
- The feature of the dataset are well understood & the classification of the species is well defined.
- This makes the dataset easy to work with & understand even for beginers in data science.
- Additionally, the iris dataset is a well established & widely used the dataset.
- This makes it easier for beginers in data science to find useful information & example of using the dataset.

Conclusion :-

- This iris dataset is a well-known & widely used dataset in data science. It is used for exploratory data analysis.

Practical No. 11

Problem statement:- Write a code in Java for a single simple word count the number of occurence of each word in given input sets using the hadoop, map reduce framework on local standalone setup.

Theory:-
- Map reduce word count is a framework which splits the chunk of data; slots the map reduce tasks.
- A file system stores the o/p & i/p of jobs.
- Re-execution of failed task of the framework & monitoring them is the task of the framework.

Map-function:
Create & process the import data takes in data elements converts it into a set of other data where the breakdown of individuals element into these tuples is done.
No API contract requiring a certain number of outputs.

Reduce Functions:
Mappers output is posted into the reduction process the data into something usable every singel mapper is passed into the reduced Functions dataset.

- This makes it easier for begineers in da science to find useful information & example of using the datasets.

Conclusion: The Iris dataset is a well-known & widely used dataset in data science.
- It is used for exploratory data analysis
- An accuracy score of 100% world indicate that the model has classified all instanc correctly while on accuracy score of 0% wo indicate that the model has classified a instances correctly while an instances incorre

Why Iris:
- Dataset Digital Education Association
- Dataset is so popular.
- The iris dataset is popular dataset for several reasons.
- First it simplest & easy to understand dataset
- The feature of the dataset are well unders & the classification of the species is well defined..
- This makes the dataset are well understood & the classification.
- Additionally, the iris dataset is a well-estabi & widely used the new o/p values are saved HDFS.

- A concept called streaming is used in a writing a code for word count.
- In python using mapReduce, let's look at the reducer code & how to execute that using a streamming a Jar file.

How mapreduce works?

- The mapreduce algorithm contains two important tasks, namely map & reduce.
- The map task takes a set of data & converts it into another set of data, where individual elements are broken down into tuples (key-value pairs).

MapReduce tasks takes the outputs from the map as an input & combines those data into a smaller set of tuples.

The reduce task is always performed after the map job.

1) Input phase:

Here we have a record reader that translates each record that in an input file & sends that parsed data to the mapper in the form of key value pairs.

Map:

Map is a user defined functions which takes a series of key value pairs & processes each & of them to generate zero or more key-value pairs.

Intermediate keys:

They key-value pairs generated by the mapper are known as intermediate keys.

Combiner:

A combiner is a type of local reducer that groups similar data from the map phase into identifiable sets.

Shuffle & sort:- The reducer tasks starts with the shuffle & sort step.

Reducer:

The reducer takes the grouped key-value paired data as input & ~~rules~~ runs a reducer function as each of them.

Conclusion:- Hence we studied about mapreduce and also about the map reduce log files.

Pune District Education Association's
College Of Engineering
Manjari (Bk.), Hadapsar, Pune-412307.

Practicle No.-12

Problem statement :- Design a distributed application using MapReduce which process a log file of systems.

Theory :-

Big data : Big data can be termed, as that closed load of data that can be hardly processed using the traditional data processing units.

Hadoop :-

- Hadoop is a big data framework designed and deployed by apache foundations. It is an open source software utility that works in the network of computer in parallel to find solution to big data process it using the mapreduce algorithm..

- Hadoop consists the master slave architecture
- The hadoop consists node & data node
- The name node is master and data node is slave
- The data is firstly divided into the sub blocks ie segments.
- Then that sumb blocks are transferred into name node. The name node transfer or give the infrastructure to the data node

- Then the data is stored into data nodes respectively.

- The big data is often characterized by the value three.

- Three large volume of data.
- The wide variety of data.
- The variety at which mush as the data.

- Examples of the how big data is used in organizations:

1) In the energy industry, big data helps oil and gas companies identify potential drilling location & monitor pipeline operations like wise help it to track the electrical grids.

2) Financial services firms users ~~beg~~ big data system for the risk management

3) manufacturers and transportation companies rely on the big data to manage their supply chains and optimize their delivery routes.

- Mapredued.

- It consist of two distinct token map & reduce. as the name mapreduce suggests the mapredue phase takes place after

mapper phase has been completed.
so, the first the map job where a block of data is read & processed to the produce key value pairs as intermediate outputs.

Conclusion :- Hence, we have learnt about the MapReduce log files.

Practicle No-13

Problem Statement: Write a simple program in Scala using apache spark framework.

Theory :-

Scala combines object-oriented and functional programming is one concise high-level languages Scala's static type help avoid bugs in complex application and its jvm & Javascript runtime.
- let you build high performance system with easy access to huge ecosystem of libraries.

- Scalar is a modern multi-paradigm programming language designed to express common programming pattern in a concise and type-safe.
- ex. scala has been created by martin addresly and he ~~realized~~ released the first version in 2003.

- Scala smoothly integrates the feature of object-oriented and functions language this is simple & reader friendly.

- Scala has thread based experience.
- Scala is statistically typed language.
- Scala can execute Java code.
- You can do concurrent and synchronized processing in scala..

- Companies using Scala :

1) LinkedIn
2) Twitter
3) Four square.
4) Netflix
5) Tumbler
6) The Gaurdian
7) Precog
8) Sony
9) AirBnB
10) Klout

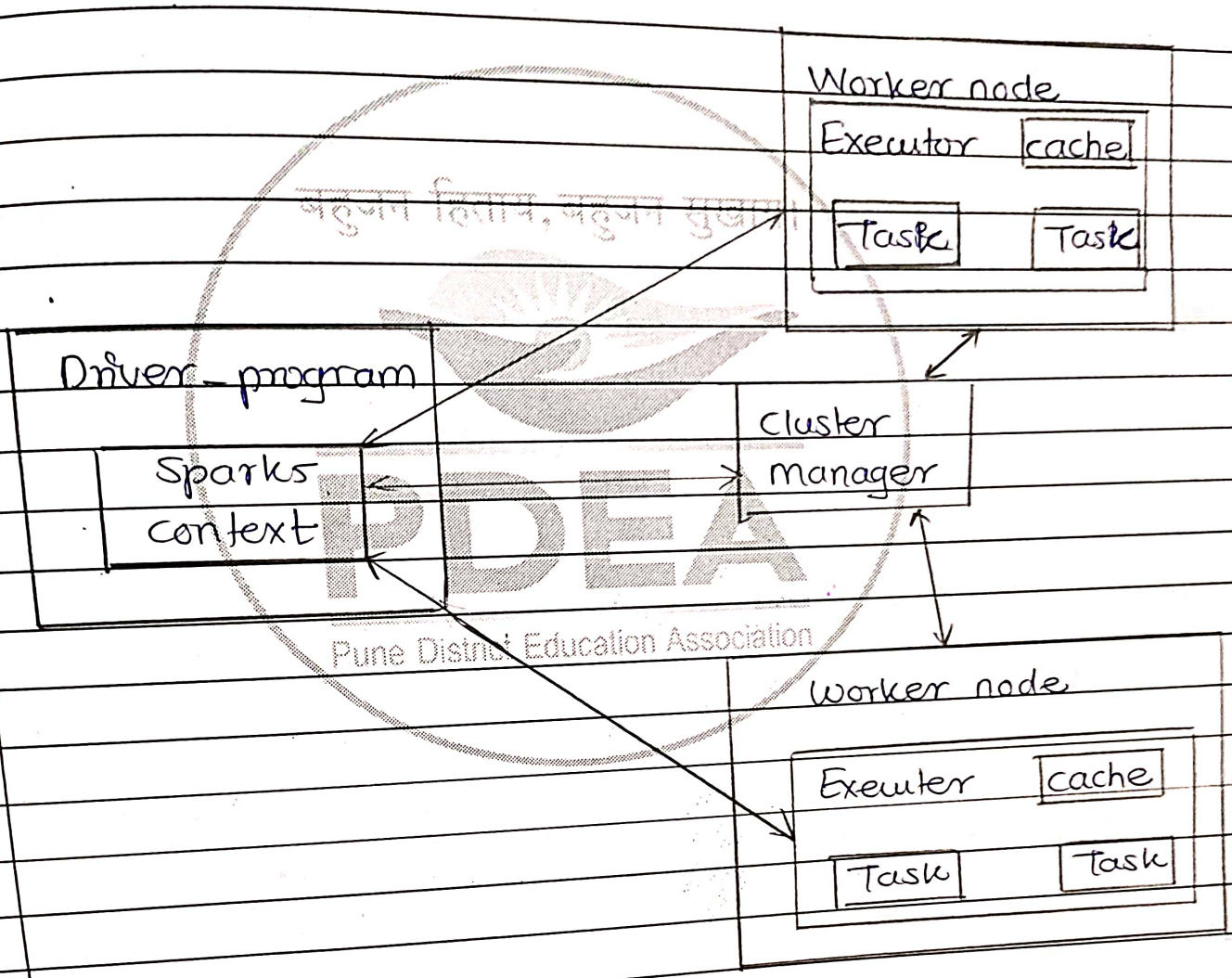
- Apache spark features :-

In memmory computation
- Distributed processing using paralized
- Fault - tolerent
- Immutable
- Lazy evaluation.
- Cache & persistance.
- Inbuilt - optimization when using data frames
- Supports ANSI SQL

- Program example.

```
object Hello world {
    def main (args : array [string]) {
    println ("Hello world")
    }
}
```

- Apache Spark Architecture :



Conclusion :-
Hence we're learnt about distributed application using log file of system..

Practicle No. 14

Case Study - I

Problem Statement :- Write down case study on global Innovations network and ~~ant~~ analysis (GNA) components of analytics plan are.

1) Discovery business problem framed.
2) Data.
3) Model planning analytic technique.
4) Result and key findings.

Theory :-

The Global Innovations Network Analysis (GINA) was study provided an example of how a team applied the data techniques analytics lifecycle to analyze innovation data at Emc to measure it and this team worked to look for ways to use advanced analytical methods to identify key innovations within the company GINA is a group of serial technologies located.

- The GINA team thought its approach would provide a means to share ideas globally & increase knowledge sharing among GNA members. who may be seperated geographically.

- It planned to create a data repository containing both structured & unstructured.

- Data to accomplish three main goals:

1) store formal & informal data.

2) Track research from global technologies.

3) mine the data for pattern and insights to improve the team's operator & the strategy.

1) Discovery :-
In the GINA projects discovery phase, the team began identifying data sources following person are involved in this phase.
i) Business users.
ii) BI analyst
iii) Data eng.
iv) Data scientist.

2) Data preperation :-
- It department to set up a new innovation sandbox to store & experiments on the data.

- The data scientists & data engineer began to notice that certain data needed conditioning & normalization.

3) Model Building.
- The team made a decision to initialize longitudinal study to begin tracking data points over time regarding people developing new intellectual property..

- The parameters related to the scope of the study included in the follows -

1) Identify
2) trace
3) Done.
4) Result - key finding.

- A key result indicator (KRI) is a metric that measures the quantitative result of business actions to help companies.
- Track progress & reach organization goals KPIs offer an overview of past performance helps to corporate management unity..

Conclusion :-

Hence, I studied an global innovation network and analysis (GINA).

Practicle No-15
Case Study - II

Problem Statement :- Write a case study to process data driven for digital marketing of health care system with hadoop ecosystem components as shown

HDFS : Hadoop distributed file system.
YARN : Yet Another Resource Negotiator MapReduce programming based data processing Spark. In memory data processing MQ HIGE query based processing of data services.

Theory :- Apache hadoop is an open source framework intended to make iteraction with big data easier.
- Hadoop ecosystem is a platform or a solve the big data problems.

HDFS :- Hadoop distributed file system.
- HDFS is the primary or major components of hadoop ecosystem and is responsible for storing large data sets of structured or unstructured data across various nodes and there by maintaining the metadata in the from or log files.

1) Name node
2) Data node.

YARN :-

Yet another resource negotiator, as the name implies YARN is the one. Who helps to manages the resources across the clusters in short it perfor the scheduling & resources allocation for the hadoop system.

- consists of three major components :-
  1) Resource manager.
  2) Nodes manager.
  3) Application manager.

MapReduce :-

By making the use of distributed & parallel ~~al~~ algorithms, Mapreduce makes it positive to carry over the processing to carry over the proce logic & helps to write applications which transform big data sets into a manageabe one..

MapReduce makes the use of distributed & para ~~algorith~~ two functions :
  1) mapReduce ()
  2) Reduce ()

Apache Spark :-

It is platform that handles all the process Interactive or Iterative real time processing graph conversion in and visualization.

- Spark is best world suited for real-time data whereas Hadoop is best suited for structure data or batch processing..

PIG -

PIG was basically developed by yahoo which makes on a big latin language, which is queen based language similar to SQL.

PIG does not work of executing the command & in the background all the activities of map.

HIVE -

- With the help of SQL methodology of interface HIVE performs reading & writing of large data sets.
- It is highly scalable as it allows runtime processing & batch processing both..
- All SQL datatypes are supported by the HIVE thus making the query processing easier.

Conclusion :-

I studied the case study to process data driven for digital marketing OR health care systems with the Hadoop ecosystem components.