Date : / /

DAA - Assignment - 1

Title :- Download the iris flower dataset into a dataframe using python & perform following:
i) How many features are there & what are their types (e.g. numeric, nomical)
ii) Compute & display summary statistics for each feature available in the dataset (e.g. minimum values, maximum values, mean, range, standard deviation, variance, percentiles.
iii) Data visualization - create a histogram for each feature distribution plot the histogram.
iv) Create a boxplot for each feature in the dataset all of the boxplots should be combined into a signal plot compare distributions & identify outliers.

Objectives :- i) To understand python commands
ii) To understand Data visualization.

Outcomes :- i) Understand the data visualization & perform the operations for minimum, maximum, mean, Range values.

Software :- Jupyter Notebook [Web application]

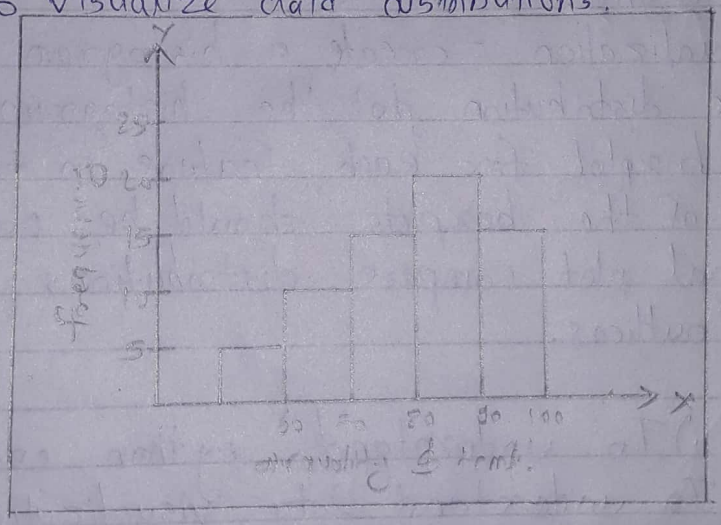
Hardware :- 512 MB RAM, 500 GB HDD.

Theory :-

- **Data Visualization :-**

i) Histogram :- A histogram is a graphical representation that organizes a group of data points into user-specified range.

- A histogram is a bar graph-like representation of data that buckets a range of outcomes into columns along x-axis.
- The y-axis represents the number count or percentage of occurence in the data for each column & can be used to visualize data distributions.

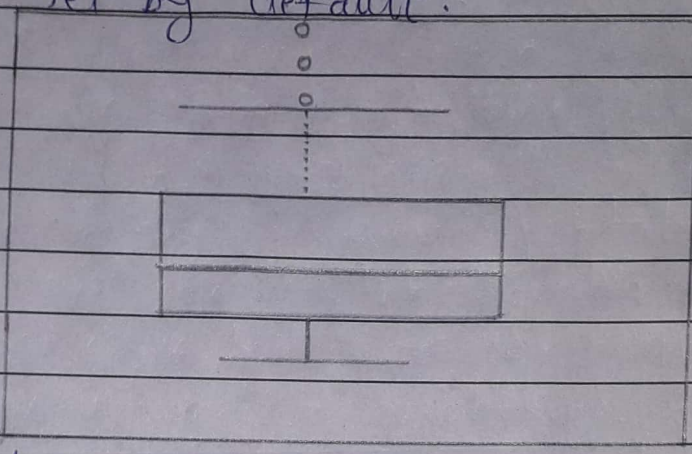


ii) Box plots :- A box plot or box & whiskers plot is a graphical summary of a distributions.

- The box in the middle indicate hinges [close to the first & third qualities] & median.
- The lines shows largest & smallest observation that fall within the distance
- A boxplot can often give a good idea of the data distribution & is often more useful to compare distributions side by side as it is more compact than histogram.

Date : / /

- Thus use of boxplot functions of to calculate quick summary for all the function variables in out set by default.



- Commands :-

1) Dataframe - head() fn. :→ Used to get first n rows.

2) Data frame - info() fn. :→ Used to print concise summary of a data frame.

3) Data frame - shape - stores no. of rows & column.

4) Data frame - describe → Used for calculating some statistical data like percentile mean & std. for numerical values of series.

5) Data frame - dtype → describe how the bytes in the fixed size book of memory corresponding to an array should be interpreted.

- Conclusion :- From this assignment, we understood python commands &, data visualization.

Page No. 1

Date : / /

DAA Assignment - 2

Title :- Download pima Indian Diabetes dataset, Use 'Naive Bayes' algorithm for classification.

i) Load the data from CSV file & split it into training & test datasets.

ii) Summarize the properties in the training dataset so that we can calculate probabilities & make predictions.

iii) Classify samples from a test dataset & a summarized training dataset.

Objective :-

i) To understand naive baye's algorithm.

ii) To understand classified samples from a test dataset & summarized training dataset.

Outcomes :-

i) Understand Naive Baye's algorithm.

ii) Understand classified, summerized & training Dataset.

Theory :-

- Naive Bayes :-

i) Naive Bayes is a probablistic classification method based on Bayes theorem with a few weaks.

ii) Bayes theorem provides the relationship among

the probabilities of 2 events with their conditional probabilities.

iii) Bayes law is named after the english mathematician Thomas Bayes.

iv) A Naive Bayes classifier considers that the existance or non-existance of specific feature of a class is not related to the Existence or non-existence of other features.

v) As the implementation of Naive Bayes classifiers is easy and can execute efficiently even without previous knowledge of the data they are considered as the frequently used algorithms for classifying text documents.

vi) A logical classification of object can be done on the bases of its attributes like shape, color.

- Summarize data & classify samples from a ~~da~~test dataset & a summarized training dataset:-

i) The naive bayes model is comprised of a summary of a data in the training dataset. This summary is then used when making predictions.

ii) The summary of the training data gathered involves the mean & the standard deviations for each attributes by class value.

iii) We can split the preparation of this summary data down into the following sub-tasks.

a) Seperate data by class
b) Calculate mean
c) Calculate standard deviation
d) Summarize dataset
e) Summarize attributes by class.

Date : / /

- Handle data :-

i) The first thing we require to do is load our data file. The data is in .csv format not including any quotes or a header lines.

ii) We can open the file with the open function & read the data lines using the reader function in the csv module.

iii) We also require to change the attributes that were landed as strings into numbers that we can work with them.

Conclusion :- I have studied Naive Bayes algorithm for classification & classify the training and summarized data.

Page No. 1

Date : / /

DA Assignment - 3

Title :- Bigmart Sales analysis.

Problem statement :- Bigmart sales analyses for data comprising of transaction records of a sales store. The data has 8523 rows & 12 column.
Predict the sales of a store.

Objective :- To apply different regression techniques to find/predict the sales of a store.

Outcome :- The outcomes are :-
- To learn to preprocess tabular data.
- To apply different regression technique.

Theory :- The data scientists of Bigmart have collected 2013 scales data for 1559 products across to different stores in diff. cities.

Data :-
- Item Identifier - Unique product ID
- Item weight - Weight of product
- Item_fat_content : Whether the product is low fat or
- Item_visibility :- The % of total display area of all products in a store allocated to the particular product.

- Item_Type :- The category to which the product belongs
- Item_MRP :- Maximum Retail price of product.
- Outlet_Identifier :- Unique store Id
- Outlet_size :- The size of the stores interms of ground area covered.
- Outlet_location_Type :- The type of city in which the store is located.
- Outlet_Type :- Whether the outlet is just a grocery store or some sort of supermarket.
- Item_outlet_sales :- Sales of the product in the particular store.

* The different steps involved are :-
- Data exploration :- looking at the catagorical & continuous feature summerize & making inferences about the data.
- Data cleaning :- Inputting missing values in the data & checking for outlets.
- feature Engineering :- Modifying existing variables & creating new ones for analysis.
- Model Building :- Making predictive modles on the data.

* Since we are dealing with continuous values as our target value (Item_outlet_sales), This would come under regression problem.

* Algorithms :-

(1) linear Regression :- It is a linear approach to modelling the relationships between a scalar response (or dependent variable) & one or more explanatory variables (or independent variables).

(ii) Random forests :-

Random forests or random forest tree are an ensemble learning method for classification, regression & other tasks that operate by constructing a multitude of decision trees at training time & outputting the class that is the mode of the classes are mean average prediction of the individual trees.

* Different libraries used :-
- numpy
- pandas
- scikit learn

- We split the train data into training & validation at 70:30 ratio.

Analysis :-

Algorithm	Validation store	Test Score
1) linear regression	1148.49	1277.805
2) Random forest Regression	1138.185	1226.34

Evaluation Metric :- Root Mean squared errors.

Conclusion :- We have, thus build a machine learning model to predict outlet sales using BigMart Dataset.

Page No. 1

Date : / /

DA Assignment - 4

Title :- Trip history analysis : Use trip history dataset that is from a bike sharing service in the united states, The data is provided quarter wise from 2010 on wards. Each file has 7 columns, predict the class of user sample Test dataset available here.
(https://www.capitalbikeshare.com/trip-history-data)

Objective :-

1) To understand dataset

Outcomes :-

1) understand the dataset.

Theory :-

- Hypothesis Generation :-

Before discovering the data to recognize the relationship between variables Id recommend you to focus on hypothesis generation first. now this might sound counter intuitive for solving a data science problem but if there is one thing I have learnt over years, it is this.

- Here, is some of the hypothesis which I thought could influence the demand of bikes. :-

1) Hourly Trend :- Early morning & late evening can have different trend & low demand during 10:00 to 4:00 am. There must be high demand during office timing.

2) Daily Trend :- Registered users demands more bikes on week days as compared to holidays or weekend.

3) Time :- Total demand should have higher contribution of registered user as compared to casual because registered user base should increase over time.

- Understanding Data set :-

a) The datasets demonstrates hourly rental data for two years (2011 & 2012).

b) The training dataset is for the first 19 days of each month.

c) In the training dataset they have detachly given bike order by recorded, casual users & sum of both is given as count.

* Conclusion :-

I have studied about the understand dataset, on the training dataset as sample of or perform the dataset operation.