DATA WAREHOUSING AND MINING

22AD3104A

STUDENT ID:

STUDENT NAME:

ACADEMIC YEAR: 2024-25

Table of Contents

A.Y. 2024-25 LAB CONTINUOUS EVALUATION

S.No	*Date*	*Experiment Name*	*Pre-Lab (5M)*	*In-Lab (25M)*			*Post-Lab (10M)*	*Viva Voce (5M)*	*Total (50M)*	*Faculty Signature*
				Program/ Procedure (10M)	*Data and Results(10M)*	*Analysis & Inference(10M)*				
1.		*Basic Statistical Descriptions*								
2.		*To implement data pre-processing techniques*								
3.		*To implement principle component analysis*								
4.		*Classification using Decision Trees*								
5.		*Classification using K Nearest Neighbor*								
6.		*Classification using Bayesian Classifiers*								
7.		*Classification usingBack propagation*								
8.		*Association Rule Mining - Apriori*								
9.		*Implementation of K-Means Clustering*								
10.		*Classification: SupportVector Machine (SVM)*								
11		*Rule Based Classification*								
12.		*Outliers detection*								

Experiment #	<TO BE FILLED BY STUDENT>	Student ID	<TO BE FILLED BY STUDENT>
Date	<TO BE FILLED BY STUDENT>	Student Name	<TO BE FILLED BY STUDENT>

Lab#1: Basic Statistical Descriptions

Date of the Session:_____/_____/___ *Time of the Session______to______*

Pre-lab

BASIC STATISTICAL DESCRIPTIONS OF DATA

Basic statistical descriptions provide the analytical foundation for data pre processing. It can be used to identify properties of the data and highlight which data values should be treated as noisy or outliers.

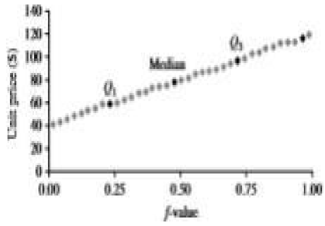
Answer the following Questions

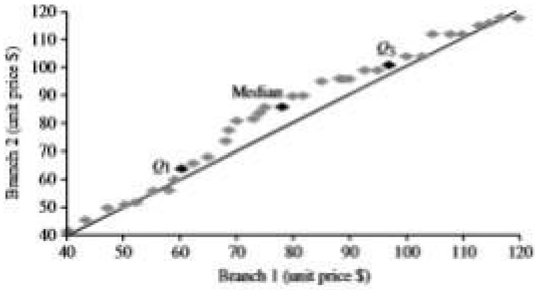
1. *What are the various ways to measure the central tendency of data?*

2. *What are the several ways of measuring the dispersion of data?*

3. *What is IQR (inter quartile range)?*

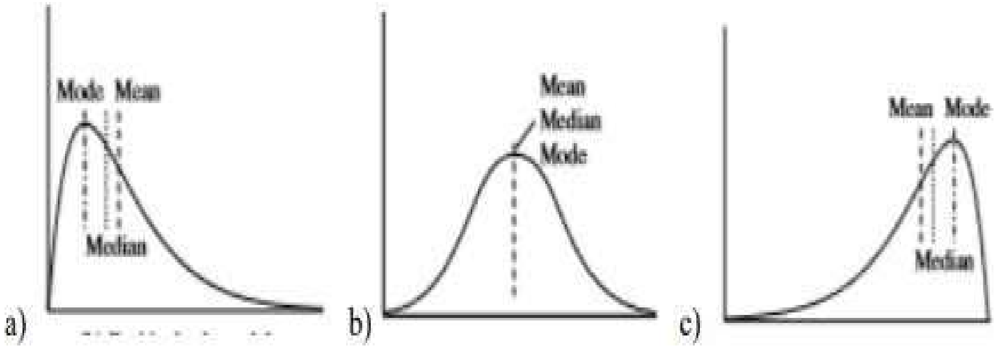
4. *Observe the following diagrams; identify the quantile and q-q plot? Define how the q-q-plotis different from quantile plot?*





5. *What are the items involved Five number summary?*

6. *Identify the symmetric data , positively skewed data and negatively skewed data from the below graphs?*



In-lab

1. *Given a dataset "cars" for analysis it includes the variables speed and distance.(Download the dataset from lms)*

   a) What are the average speed and the distance of the cars?
   b) What is the median and midrange of the data?
   c) Find mode of the data and comment on the data modality (i.e, unimodal or bimodal)?
   d) What are the variance and the standard deviation of the data?
   e) Find the five number summaries of the data?
   f) Show the histogram and box plot of the data?

Writing space of the Problem :(For Student's use only)

Post-lab

1. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results. (Download the dataset from lms)

a) Find the maximum and the minimum percentage of the fat and age of the adults who visited the hospital.

b) Calculate mean, median and midrange of the age.

c) Find the first quartile and third quartile of the data.

d) Draw a scatter plot and q-q plot based on these two variables.

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *Difference between symmetric data and skewed data.*
2. *What are the most widely used forms of quartiles?*
3. *Variance and Standard deviation fall under what category of measuring data?*
4. *What do low and high standard deviations indicate?*
5. *Based on what condition, two variables are said to be correlated?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Experiment #	<TO BE FILLED BY STUDENT>	Student ID	<TO BE FILLED BY STUDENT>
Date	<TO BE FILLED BY STUDENT>	Student Name	<TO BE FILLED BY STUDENT>

Lab#2: To implement data pre-processing techniques.

Date of the Session:_____/____/____Time of the Session ______________________________to______

Pre Lab

DATA PREPROCESSING:

Databases are exceedingly helpless to noisy, missing, and inconsistent data because of their commonly enormous size (frequently a few gigabytes or more).Low-quality information willprompt low-quality mining results. Pre-processing helps to get a quality data. The steps involved in data pre-processing are data cleaning, data integration, data reduction, data transformation.

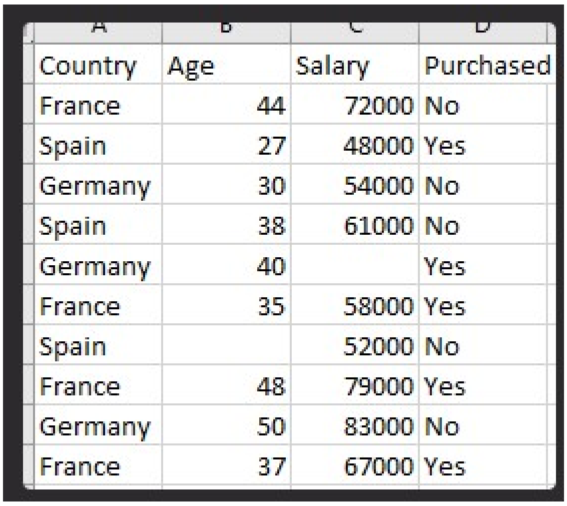
Match the following:

i. Data cleaning	*a. Reduced representation of data*
ii. Data integration	*b. x_{old}/ x_{max}*
iii. Data reduction	*c. deal with missing values and noisy data*
iv. Normalization	*d. works to remove noisy data*
v. Data transformation	*e. $(x_{old}-x_{min})/(x_{max}-x_{min})$*
vi. Decimal scaling	*f. merging of data from multiple data stores*
vii. Minmax normalization	*g. scale the data values in specified range*
viii. Z score normalization	*h. convert data into appropriate forms*
ix. Smoothing	*i.(x_{old}-mean)/standard deviation*

1. Mention any two methods that deal with missing values and noisy data.
2. Mention two techniques that are applied to obtain a reduced dataset.
3. Using min-max normalization, transform the value 35ontothe range [0.0,1.0].
4. Using z-score normalization, transform the value 35, where the standard deviation is 12.94years.
5. Using normalization by decimal scaling, transform the value 35.

In-lab

1. Given a data set “data” for analysis it includes the attribute Country, purchased item, age, Salary.



A	B	C	D
Country	Age	Salary	Purchased
France	44	72000	No
Spain	27	48000	Yes
Germany	30	54000	No
Spain	38	61000	No
Germany	40		Yes
France	35	58000	Yes
Spain		52000	No
France	48	79000	Yes
Germany	50	83000	No
France	37	67000	Yes

a. Identify number of missing values in a given dataset
b. Drop the tuples that have missing values in the attributes.
c. Check the data type of age , ifit is not an integer then convert into integer.
d. Normalize the salary using simple feature scaling.
e. Categorize the salary into low, high, medium bins.
f. Turn the categorical values into numerical.

Writing space of the Problem :(For Student’s use only)

2. Suppose john is working as a manager at Nuclear Power Corporation of India and have been charged with analyzing the Nuclear power station construction data. He carefully inspects the company's database identifying and selecting the attributes (cost, date, t1, t2 and cap) to be included in the analysis.(Download the dataset from lms)

a. He noticed that several values of the attributes for various tuples have no recorded value.

b. He observed that data type of year is recorded in float instead of integer type.

c. He wants to normalize all the data (variables)in equal weights.

d. Finally, he wants to know if there are any outliers present in cost of the construction. You immediately set out to perform this task.

***Hint**: missing values can be solved by replacing with mean)*

Writing space of the Problem :(For Student's use only)

Post-lab

1. *Data(13,5,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70)*
2. *Use smoothing by bin means to smooth the above data, using a bin depth of*
3. *Illustrate your steps.*

Comment on the effect of this technique for the given data. Also Plot a histogram.

Writing space of the Problem :(For Student's use only)

2. Use these two methods below to normalize the following group of data: 200, 300, 400,600and1000.

a. min-max normalization by setting min =0 and max=1

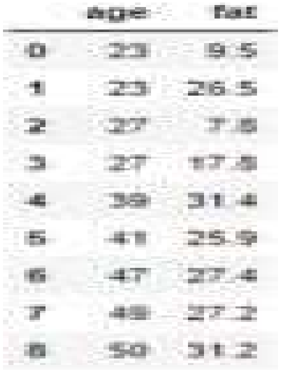
b. z-score normalization

c. z-score normalization using the mean absolute deviation of standard deviation

d. Normalization by simple feature scaling.

Writing space of the Problem :(For Student's use only)

3.a. Normalize the two variables(age, fat) based on z-score normalization
b. Calculate the correlation matrix. Are these two variables positively or negatively correlated?



	age	fat
0	23	9.5
1	23	26.5
2	27	7.8
3	27	17.8
4	39	31.4
5	41	25.9
6	47	27.4
7	49	27.2
8	50	31.2

Writing space of the Problem :(For Student's use only)

VivaVoce:-

1. *What are the factors that comprising data quality?*
2. *What do you mean by noise in the dataset?*
3. *What are outliers in the dataset?*
4. *What is discretization?*
5. *What is the difference between lossy and lossless in data reduction?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#3: To implement principle component analysis

Date of the Session:_____/____/_____ ***Time of the Session_____to______***

Pre-lab:-

Principal Component Analysis:

Principal Component Analysis is a method of extracting important variables from large set of variables available in a dataset. Suppose that the data to be reduced consist of tuples or data vectors described by n attributes or dimensions. Principal components analysis (PCA; also called the Karhunen-Loeve, or K-L, method) searches for k n-dimensional orthogonalvectors that can best be used to represent the data, where k ≤ n. The original data are thus projected onto a much smaller space, resulting in dimensionality reduction.

1. *What are principal components?*
2. *Mention the steps to construct principal components?*

In-lab:-

1. *Suppose that you are given a small 3x2 matrix, you have to calculate Principal Component Analysis without using pca () function?*
   Matrix :([3, 5], [4, 2], [1, 6])

Writing space of the Problem :(For Student's use only)

2. *Calculate the principal component analysis for the matrix given in Q1 using PCA?*

Post-lab:-

1. *The Iris dataset is a classic dataset in statistics, often used for testing and benchmarking algorithms. How does Principal Component Analysis (PCA) transform the high-dimensional Iris dataset into a lower-dimensional space, and what can be inferred about the dataset from the visual representation of the first two principal components?*
   (Download the dataset from lms)

Writing space of the Problem :(For Student's use only)

VivaVoce:-

1. *Why PCA is preferable, mention the two primary reasons?*
2. *Is there any loss of data if we use PCA?*
3. *PCA is an unsupervised technique, will you agree with it? Why?*
4. *What are the applications of PCA?*
5. *Define covariance matrix?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of___________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#4: Classification using Decision Trees.

Date of the Session:_____________/___/____ *Time of the Session:___to_____*

Pre-lab:-

1. *What are the attribute selection measures in modelling a decision tree and write the respective equations for each of them.*
2. *What do you mean by entropy in a decision tree? How is it calculated?*
3. *What is Information gain and how does is matter in a Decision Tree?*
4. *List out the parameters involved in Decision Tree Classifier and export_graphviz and try to understand the role of each parameter.*

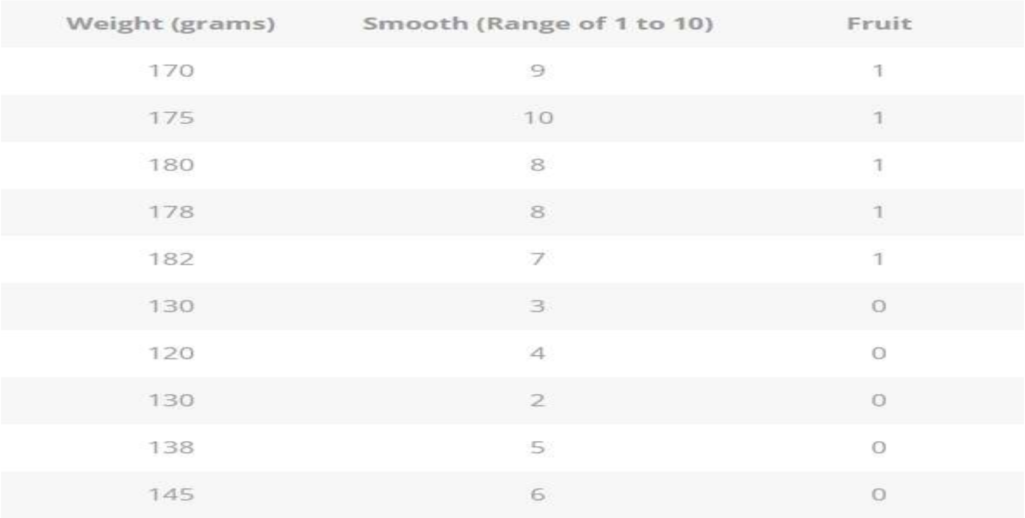
5. Match the following:

1. ID3	*a. GAIN RATIO*
2. CART	*b. INFORMATION GAIN*
3. C4.5	*c. GINI INDEX*

In-lab:-

1. *Implement the decision tree algorithm on the given data which has weight and smoothness as the segregating criteria for the fruit apple and orange. Apple is represented by the number '1'andorangeby'0'.Constructa decision tree and apply the prediction measures for The given data to obtain the types of fruits.*

Weight	*Smooth*	*Fruit*
180	*7*	*?*
140	*8*	*?*
150	*5*	*?*



Weight (grams)	Smooth (Range of 1 to 10)	Fruit
170	9	1
175	10	1
180	8	1
178	8	1
182	7	1
130	3	0
120	4	0
130	2	0
138	5	0
145	6	0

Fruit dataset

https://drive.google.com/file/d/1qoMDjozHHELVn5tFAJxp8mMw0Ggt-BVX/view?usp=sharing

Convert the trained decision tree classifier into graphviz object. Later; we use the converted graphviz object for visualization. To visualize the decision tree, you just need to open the .txt file and copy the contents of the file to paste in the graphviz web portal graphyiz web portal address: http://webgraphviz.com

Writing space of the Problem :(For Student's use only)

2. *Below given is the diabetes dataset.*

(Ref: *https://drive.google.com/file/d/1PJizP39JPh_T-5dQVcUVCfswrPSxT734/view?usp=sharing)*

Make sure to install the scikit-learn package and other required packages.

1. Find the correlation matrix for the diabetes dataset?
1. *Split the dataset into train_set and test_set for modeling and prediction. Divide the dataset in such a way that the trained dataset constitutes 70 percent of the original dataset and the rest of the part belongs to the test dataset.*
2. *Produce a decision tree model using*
   a. *Gini index metric*
   b. *Entropy and Information gain metric on the trained dataset*
   using the Decision Tree Classifier function.
3. *Apply the prediction measures on the test dataset.*
4. *Define a function named accuracy_score by interpreting the difference between the predicted values and the test set values.*
5. *Display the accuracy in terms of*
   a) *using the accuracy_score function*
   b) *Fraction Number of correct predictions.*
   c) *Print the confusion matrix of the test dataset.*
6. *Calculate the following values manually after obtaining the confusion matrix*
   a. *Accuracy*
   b. *Error rate*
   c. *Precision*
   d. *Recall (sensitivity)*
   e. *F1Score*
   f. *Specificity*

   Compare the two results (obtained from two kinds of metrics) and state which method ismoreaccurateforthisdataset. Convertthetraineddecisiontreeclassifierinto graphviz object. Later, we use the converted graphviz object for visualization.

10. *Plot ROC curve and calculate AUC*
11. *Plot recall vs precision curve*

Writing space of the Problem :(For Student's use only)

Post-lab:-

1. *What is the C4.5 algorithm and how does it work? State the differences between ID3 and C4.5.*
2. *Differentiate between over-fitting, over-fitting and over-fitting loss? Why does it occur during classification?*
3. *Explain the concept to fpruning and why it is important. Differentiate between pre-pruning and post-pruning.*

VivaVoce:-

1. *What is the difference between supervised and unsupervised machine learning?*
2. *What is a confusion matrix?*
3. *Which of the following is true about training and testing error in such case?*
    a. *The difference between training error and tester or increases as number of observations increase.*
    b. *The difference between training error and test error decreases as number of observations increase.*
    c. *The difference between training error and test error will not change*
4. *What is the difference between classification and clustering?*
5. *What are Recommender Systems?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#5: Classification using K Nearest Neighbour.

Date of the Session:____/___/____ ***Time of the Session:_______to________***

Pre-requisite:

***In LMS: Find the file named** "Concept of k-Nearest-Neighbor.doc". Read the specified document and answer the below questions.*

Pre-lab:-

1. *State whether the given statement is true or false with supported reasoning.*

   a. *k-Nearest-Neighbor is a simple algorithm that stores all available cases and classifies the new case based on dissimilarity measure.*
   b. *The value of 'k'in k-nearest-neighbor algorithm helps to check the no. of training sets labels to assign the most common label for the testing set.*

2. *List the industrial uses of k-nearest-neighbor algorithm in the real world.*

-

3. *Write an algorithm for k-nearest-neighbor classification given k, the nearest number of neighbors, and n, the number of attributes describing each tuple.*

4. *Compare the advantages and disadvantages of eager classification (e.g., decision tree, Bayesian, neural network) versus lazy classification (e.g., k-nearest-neighbor, case-based reasoning).*

5. *Give the distance methods that are most commonly used in k-nearest-neighbor algorithm.*

In-lab:-

Perform the following Analysis:

Step-by-step process to compute k-nearest-neighbor algorithm is:

1. *Determine parameter k=no. of nearest neighbors*
2. *Calculate the distance between the test sample and the training samples.*
3. *Sort the distance and determine nearest neighbors based on the k^{th} minimum distance.*
4. *Gather the category of nearest neighbors.*
5. *Use simple majority of the category of nearest neighbors as the prediction value of testing sample.*

Dataset:

*Suppose we have the following "**Student Data Set**" dataset which consists of 1^{st} yearCGPA,2^{nd} year CGPA, Category(**C**: CRT, **NC**: Non-CRT)as parameters.*

Std.No	***1^{st} yearCGPA***	***2^{nd}yearCGPA***	***Category***
1	*8.5*	*8.5*	*C*
2	*8.2*	*9*	*C*
3	*7.5*	*7.6*	*C*
4	*5.5*	*4.5*	*NC*
5	*9.2*	*9*	*C*
6	*7.8*	*7.3*	*C*
7	*7.3*	*7.4*	*NC*
8	*7.9*	*7*	*NC*
9	*10*	*6*	*C*
10	*6.8*	*7.1*	*NC*
11	*6.5*	*7.1*	*NC*
12	*7.2*	*7.3*	*NC*

When a new student comes only with 1^{st} year CGPA and 2^{nd} year CGPA as information predict the category of that new student(whether he belongs to CRT or Non-CRT)by Euclidean distance measure, where Euclidean distance between 2 points or tuples, say$X_1=(x_{11},x_{12},\ldots\ldots\ldots,x_{1n})$ and $X_2=(x_{21},x_{22}, ,x_{2n})$, is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}.$$

Tests ample:

1^{st} year CGPA and 2^{nd} year CGPA of the new student are 8.4 and 7.1 respectively.
(Consider k=3)

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. *Predict the Category of student with **1st year CGPA and 2nd year CGPA as 7.3 and 7.1 respectively** using the Manhattan measuring technique formula with k=3 **(Manually)**. Note: The Manhattan distance between two tuples (or points) a and b is defined as* $\sum i|ai-bi|$
2. *By considering the above **Student Data Set ,,** predict the Category of the new student having **1st year CGPA and 2nd year CGPA as 8.4 and 7.1 respectively,** by implementing the python code using Manhattan distance measure in order to find nearest neighbors fork=3 and check whether the output is same for both the measuring techniques or not.*

VivaVoce:-

Refer Page no: 423 ,424, 425 in Han J & Kamber M, "Data Mining: Concepts and Techniques", Third Edition, Elsevier, 2011

1. *k-nearest-neighbor is a______________lazy learning algorithm.*
2. *How can the distance be computed for attributes that are not numeric, but nominal (or categorical) such as color?*
3. *List some techniques used to speed up the classification time.*
4. *If the value of a given attribute A is missing in tuple X_1 and/or in tuple X_2, the difference is always______________*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab #6: Classification using Bayesian Classifiers

Date of the Session:____/___/____ ***Time of the Session:_______to________***

Pre-lab:-

1. *Match the following*

Column A	*Column B*
a. Naïve Bayesian Classification	*a. Values are continuous*
b. Bayesian belief network	*b. Attributes condition all y dependent*
c. Gaussian distribution	*c. To avoid zero probability*
d. Laplace estimator	*d. Attributes condition all y independent*

2. *Explain Baye's theorem and write its derived formulae.*

3. *Suppose we have continuous values for an attribute in a data set then how to calculate probability.*

4. Let us assume
p (age=youth/buys_car =yes) =0.222,
p (income=medium/buys_car) =0.444 and
p (buys_car=yes) =0.643 then
Find the probability of p(x/buys_car=yes), where x= (income=medium, age=youth).

5. While implementing Naïve Bayesian classifier, suppose we have encountered a zero probability then we should add one count to each of the probability to avoid zero probability. What is this estimation is called?

In-lab:-

1. Consider the given table named "Weather_cond.csv" consisting of attributes Temperature Humidity, Windy and a class label named "Outcome". Depending on the weather conditions you have to choose whether to play cricket or not.

a. Unlike conventional function, write a python function to split the dataset into training set and test set. Assume test size length as 0.33.

b. Write a python function to calculate mean and standard deviation for each numerical attribute in the data set.

c. Calculate the number of priors for the given data set after splitting into training and test sets using python.

Writing space of the Problem: (For Student's use only)

2. The problem is comprised of 100 observations of medical details for Pima Indian's patients. The records describe instantaneous measurements taken from the patient such as their age, the number of times pregnant and blood workup. All patients are women aged21 or older. All attributes are numeric, and their units vary from attribute to attribute. Each record has a class value that indicates whether the patient suffered an onset of diabetes within 5 years of when the measurements were taken (1) or not(0).This is a standard data set that has been studied a lot in machine learning literature. A good prediction accuracyis70%-76%.

Implement a python code to find the accuracy for given data set named "Diabetes.csv" based on train set and test set. Take test size length as 0.4.

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. *Consider the given table that specifies loan classification problem.*

Tid	*Home Owner*	*Marital status*	*Annual Income*	*Defaulted Borrower*
1	*Yes*	*Single*	*125K*	*No*
2	*No*	*Married*	*100K*	*No*
3	*No*	*Single*	*70K*	*No*
4	*Yes*	*Married*	*120K*	*No*
5	*No*	*Divorced*	*95K*	*Yes*
6	*No*	*Married*	*60K*	*No*
7	*Yes*	*Divorced*	*220K*	*No*
8	*No*	*Single*	*85K*	*Yes*
9	*No*	*Married*	*75K*	*No*
10	*No*	*Single*	*90K*	*Yes*

   a. *Compute the class conditional probability for each categorical attribute.*
   b. *Predict the class label value for test record X = (Home Owner=No ,Marital Status=Married, Income=$120K)*

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *Explain the difference between a Validation Set and a Test Set?*
2. *What are the three types of Naïve Bayes classifier?*
3. *How many terms are required for building a Bayes model?*
4. *What is training test and testing set?*
5. *What are the advantages of Naive Bayes?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)

Evaluator's Observation

Marks Secured:________out of___________

Full Name of the Evaluator:

Signature of the Evaluator Date of Evaluation:

Lab #7:Classification using Back propagation

Date of the Session:____________/___/____ *Time of the Session:___to_____*

Pre-lab:-

***In LMS**: Find the file named **"Han J & Kamber M, Data Mining Concepts andTechniques.doc"**.*

Read the specified document from Pg.No:398–404 and answer the below questions.

1. State whether the given statement is True/False.

a. Back propagation is neural network learning algorithm.

b. Back propagation learns by iteratively processing a data set of training tuples, comparing the network's prediction for each tuple with the actual known target value.

2. What is the objective of Back propagation?

3. Explain about Multilayer Feed-Forward Neural Network with diagram.

4. *How does Back propagation work?*

5. *Consider the following table.*

Input	***Desired Output***	***Model Output***	***Absolute Error***	***Square Error***
0	*0*			
1	*2*			
2	*4*			

Predict the Model Output by considering the initial value of weight as ***3****. Find the Absolute Error and Square Error. Use the Back propagation algorithm to update the weight and try to minimize the* ***square error*** *as much as possible.*

Hint:

i. ***Model Output=*** *W*I(x)(W=weight, I=Input,x=indexthatiteratesfrom0to length(Input))*

ii. ***Absolute Error =*** *mod(Model Output-Desired Output)*

iii. ***SquareError=*** *(Absolute Error)^2*

Writing space of the Problem: (For Student's use only)

In-lab:-

Analysis:

The following steps will provide the foundation that you need to implement the Back propagation algorithm and apply it to your own predictive modeling problems:

1. *Initialize Network.*
2. *Forward Propagate.*
   i. *Neuron Activation.*
   ii. *Neuron Transfer.*
   iii. *Forward Propagation.*
3. *Back Propagate Error.*
   i. *Transfer Derivative*
   ii. *Error Back propagation*
4. *Train Network.*
   i. *Update Weights.*
   ii. *Train Network.*
5. *Test Network.*

Dataset:

*Suppose we have the following "**Results Dataset**" which consists of GPA's of some students that they had scored in two internal tests. And, it also consists of anotherattribute named 'Qualified', which holds a character (Q/NQ), representing the student qualification for final examination.*

S.No	***Test– 1***	***Test– 2***	***Qualified***
1	*8.5*	*8.5*	*Q*
2	*8.2*	*9.0*	*Q*
3	*3.5*	*5.0*	*NQ*
4	*5.5*	*4.5*	*NQ*
5	*9.2*	*9.0*	*Q*
6	*7.8*	*7.3*	*Q*
7	*8.0*	*3.1*	*NQ*
8	*10*	*7.0*	*Q*
9	*4.5*	*6.0*	*NQ*
10	*6.8*	*7.1*	*Q*
11	*5.1*	*4.1*	*NQ*
12	*4.2*	*5.3*	*NQ*

Problem: *Train a network on above **"Results Dataset"** by applying Back propagation algorithm.*

a. *Initializing a network with all weights and biases. (Consider weights in range-0.5 to +0.5, biases=1, Learning Rate = {0.5, 0.7, 1})*
b. *Training the network according to the Dataset. (Consider both Activating Functions– Sigmoid Function and Tanh Function)*
c. *Back propagating the errors.*

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. *Use the network which is trained on the above **"Results Dataset"** and test whether it is trained with 100% accuracy or not. And, predict the result (qualified for final examination or not)of a new entry which contains 5.9and 5.9 GPA'softest-1 and test-2 respectively.*

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *What are the general tasks that are performed with back propagation algorithm?*
2. *What kind of real-world problem scan neural networks solve?*
3. *What is a gradient descent?*
4. *Why is zero initialization not a recommended weight initialization technique?*
5. *How are artificial neural networks different from normal networks?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:________out of___________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#8: Association Rule Mining -Apriori

Date of the Session:_____________/___/____Time of the Session:________________________________to______

Pre-lab:-

1. *Define what is Apriori algorithm.*

2. *What is association mining?*

3. *What is the need of association mining?*

4. *What is minimum support and minimum confidence?*

5. *Consider the market basket transactions given in the following table. Let min-sup=40%and min_conf=40%*

Transaction ID	*Items Bought*
T1	*A,B,C*
T2	*A,B,C,D,E*
T3	*A,C,D*
T4	*A,C,D,E*
T5	*A,B,C,D*

a. *Find all the frequent item sets using apriori algorithm.*
b. *Obtain significant decision rules.*

In-lab:- For the following given transaction dataset, perform following operations :

a. Generate rules using Apriori algorithm by using below dataset.

shrimp	*almonds*	*avocado*	*vegetables mix*	*green grapes*	*whole wheat flour*	*yams*	*cottage cheese*
burgers	*meatballs*	*eggs*					
chutney							
turkey	*avocado*						
mineral water	*milk*	*energy bar*	*whole wheat rice*	*Green tea*	*eggs*		
Low fat yogurt							
whole wheat pasta	*french fries*						
soup	*light cream*	*shallot*					
frozen vegetables	*spaghetti*	*green tea*					
french fries							
eggs	*Pet food*						
cookies							
turkey	*burgers*	*mineral water*	*eggs*	*cooking oil*			
spaghetti	*champagne*	*cookies*					
mineral water	*salmon*	*eggs*					
mineral water							
shrimp	*chocolate*	*chicken*	*honey*	*oil*	*Cooking oil*	*Low fat yogurt*	
turkey	*eggs*						
turkey	*Fresh tuna*	*tomatoes*	*spaghetti*	*mineral water*	*Black tea*	*salmon*	*eggs*
meatballs	*milk*	*honey*	*french fries*	*Protein bar*			
redwine	*shrimp*	*pasta*	*pepper*	*eggs*	*chocolate*	*shampoo*	
rice	*sparkling water*						
spaghetti	*mineral water*	*ham*	*body spray*	*pancakes*	*Green tea*		
burgers	*grated cheese*	*eggs*	*pasta*	*avocado*	*honey*	*white wine*	*toothpaste*
eggs							

parmesan cheese	*spaghetti*	*soup*	*avocado*	*milk*	*Fresh bread*		
ground beef	*spaghetti*	*mineral water*	*milk*	*eggs*	*Black tea*	*salmon*	*frozen smoothie*
sparkling water							
mineral water	*eggs*	*chicken*	*chocolate*	*French fries*			
Frozen vegetables	*spaghetti*	*yams*	*mineral water*				

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. *Same as In-lab question generate rules on below dataset.*

Citrus fruit	semi-finished bread	margarine	ready soups							
tropical fruit	yogurt	coffee								
Whole milk										
pip fruit	yogurt	cream cheese	meat spreads							
Other vegetables	Whole milk	condensed milk	Long life Bakery product							
Whole milk	butter	yogurt	rice	abrasive cleaner						
rolls/buns										
Other vegetables	UHT-milk	rolls/buns	bottled beer	liquor(appetizer)						
Pot plants										
Whole milk	cereals									
tropical fruit	Other Vegetables	white bread	bottled water	chocolate						
Citrus fruit	Tropical fruit	whole milk	butter	curd	yogurt	flour	Bottled water	dishes		
beef										
frankfurter	rolls/buns	soda								
chicken	tropical fruit									
butter	sugar	fruit/vegetable juice	Newspapers							
fruit/vegetable juice										
packaged fruit/vegetables										
chocolate										
specialty bar										
other vegetables										
butter milk	pastry									
Whole milk										

tropical fruit	*cream cheese*	*processed cheese*	*detergent*	*newspapers*						
tropical fruit	*Root vegetables*	*Other vegetables*	*Frozen dessert*	*rolls/buns*	*flour*	*sweetspreads*	*salty snack*	*waffles*	*candy*	*bathroom cleaner*
bottled water	*canned beer*									
yogurt										
sausage	*rolls/buns*	*soda*	*chocolate*							
other vegetables										
brown bread	*soda*	*fruit/vegetable juice*	*canned beer*	*newspapers*	*shopping bags*					

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *Who proposed Apriori algorithm in which year?*
2. *What is frequent itemset?*
3. *Why do we convert dataset into list?*
4. *What is the formula for support, confidence and lift?*
5. *How they get the name as Apriori?*

(For Evaluator’s use only)

Comment of the Evaluator (if Any)	*Evaluator’s Observation* *Marks Secured:________out of___________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#9: Implementation of K-Means Clustering

Date of the Session:____/___/____ ***Time of the Session:_______to________***

Pre-Requisites:

Data pre-processing
Basics of plotting techniques
Various clustering techniques

Pre-lab:-

1. Matchthefollowing.

Parameters	*Application*
1. pch	*a. To set orientation of axis labels*
2. col	*b. No. of plots per row and column*
3. mfrow	*c. To set plot color*
4. lwd	*d. Plotting symbol*
5. las	*e. To setline width*

2. List out various parameters and attributes in KMeans clustering.

3. Into how many types does clustering divided into and name them.

4. List out various applications of clustering.

5. *Describe Euclidean distance and Manhattan distance in brief with its derived formula.*

6. *List out basic steps involved in KMeans clustering.*

In-lab:-

1. *The given dataset comprises of 150 data entries of different countries around the world. It is a report on world happiness, a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves tobe, with a focus on the technologies, social norms, conflicts and government policies that have driven those changes. The records contains various attributes of each countrythatincludespositive_effect, negative_effect, corruption, freedom, health life expectancy etc. The data frame includes categorical variables, numerical values and their values vary from country to country.*

   Implementapythoncodeusingscikit-learntodisplayaK-meansclusteringplotforgivendataframe named "world_happiness_report.csv".

Writing space of the Problem: (For Student's use only)

2. *The given dataset named "Student_performance" consists of 150 data entries of students in an institution that displays the performance of a student. It consists of various attributes such as gender, ethnicity, test_preparation, math_score, reading_scoreetc. Perform the K means clustering for the given dataset taking an appropriate number of centres based on mean and standard deviation for the data entries. Analyze the cluster plot and give a brief note based on results obtained.*

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. This lab module aims to build an analysis on customers of a shopping mall. It consists of 150 observations of customers consisting details that include gender,age, annual_income, spending_score etc. Based on the two parameters annual_income and spending_score, try to build a analysis on customers through cluster graphs

Apply k means clustering on the given data set named "Mall_customers" marking number of clusters based on mean and standard deviation of any two attributes of your choice and implement the K-means iteratively till the centroids get stabilized

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *K-means is which type of algorithm.*
2. *In K-means clustering algorithm what is the criteria used by the data points to get separated from one cluster to another.*
3. *What are the basic steps in KMeans clustering?*
4. *What does K refer in K-means algorithm - K refers to k no. of clusters.*
5. *How is K-means algorithm is different from KNN algorithm*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:________out of__________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#10: Classification: Support Vector Machine (SVM)

Date of the Session:____/___/____ ***Time of the Session:_______to________***

Pre-lab:-

1. *What is SVM?*

2. *When do we use SVM?*

3. *What is maximum marginal hyper plane and what is the equation of separating hyperplane?*

4. *What are the two cases of SVM?*

5. *What are the equations for point that lies above these parating hyperplane and below these parating hyperplane?*

In-lab:-

1. *Below is the data of the employees in the company. The data shows whether employee purchased software or not. Take x co-ordinate as age and y co-ordinate asestimated_salary.Now, Considerthefollowing datasetandperformthebelowoperations:*

User ID	Gender	Age	Estimated Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1
15714658	Male	48	41000	1
15599081	Female	45	22000	1
15705113	Male	46	23000	1
15631159	Male	47	20000	1
15792818	Male	49	28000	1
15633531	Female	47	30000	1
15744529	Male	29	43000	0

a. *Import the data set into python*
b. *Split the dataset set into training and testing sets*
c. *Apply feature scaling on training and test sets*
d. *Fit SVM to the training set*
e. *Visualize the training set results*
f. *Visualize the test set results.*

Post-lab:-

1. *Below dataset represents the bank transactions of KVB bank for a hour. Consider x co-ordinate as Balance and y co-ordinate as Trtn_amt. Perform following operations on given dataset:*

S. No	*Transaction _ID*	*Balance*		*Trtn _amt*	*Suc or not*
1	*3467*	*98687.36*		*500*	*0*
2	*4801*	*8510.47*		*100*	*0*
3	*2093*	*2475.3*		*200*	*1*
4	*9933*	*37743.25*		*1000*	*0*
5	*7178*	*2705.95*		*600*	*0*
6	*1093*	*60314*		*750*	*1*
7	*3708*	*812129.5*		*280*	*1*
8	*3804*	*8076.25*		*140*	*0*
9	*3192*	*42323.14*		*310*	*1*
10	*3666*	*47045.25*		*2500*	*0*
11	*8598*	*96171.25*		*6900*	*0*
12	*8743*	*608581.8*		*8520*	*1*
13	*9302*	*586057.3*		*410*	*1*
14	*6127*	*4587.5*		*750*	*0*
15	*7502*	*43597.75*		*250*	*0*

a. Import the data set into python
b. Split the dataset set into training and testing sets
c. Apply feature scaling on training and test sets
d. Fit SVM to the training set
e. Visualize the training set results
f. Visualize the test set results.

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *What are the advantages of SVM?*
2. *How many types of machine learning's are there and in which type this svm fall under?*
3. *What are the turning parameters in SVM?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#11: Rule Based Classification

*Date of the Session:*_____________/___/______ *Time of the Session:*____*to*______

Pre-requisite:

Refer Page no: 355-363 in Han J & Kamber M, "Data Mining: Concepts and Techniques", Third Edition, Elsevier, 2011

Pre-lab:-

1. What is rule-based classification in data mining?

2. Briefly explain about the building classification rules.

3. When to stop building a rule?

4. List some aspects of sequential covering.

5. What are the characteristics of rule-based classifier?

6. Define coverage and accuracy.

In-lab:-

1. *Implement a simple python code for rule-based classification on **"All Electronics Customer"** database (Download the dataset from LMS)*

RID	*age*	*income*	*student*	*Credit_rating*	*Class:buys computer*
1	*youth*	*high*	*no*	*fair*	*no*
2	*youth*	*high*	*no*	*excellent*	*no*
3	*middle_aged*	*high*	*no*	*fair*	*yes*
4	*senior*	*medium*	*no*	*fair*	*yes*
5	*senior*	*low*	*yes*	*fair*	*yes*
6	*senior*	*low*	*yes*	*excellent*	*no*
7	*middle_aged*	*low*	*yes*	*excellent*	*yes*
8	*youth*	*medium*	*no*	*fair*	*no*
9	*youth*	*low*	*yes*	*fair*	*yes*
10	*senior*	*medium*	*yes*	*fair*	*yes*
11	*youth*	*medium*	*yes*	*excellent*	*yes*
12	*middle_aged*	*medium*	*no*	*excellent*	*yes*
13	*middle_aged*	*high*	*yes*	*fair*	*yes*
14	*senior*	*medium*	*no*	*excellent*	*no*

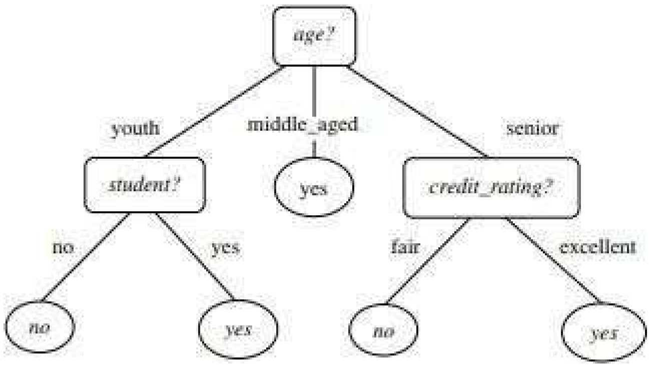
a. *Calculate accuracy, coverage and print the RID values when the following rules are satisfied:*

- ***Rule R1**: if the age of the person is in the category of "youth" and he/she is a student then the person purchases the computer.*
- ***Rule R2**: if age of the person is in the category of "middle_aged" , income is either medium or high and with excellent Credit_rating then the person buys a computer*
- ***RuleR3:** if age of the person is in the category of "senior" and he/she is a student then purchases a computer.*
- ***Rule R4**: if age of the person is in the category of "senior" , income ishigh, he/sheisastudentandwith Credit_rating fair thenpurchasesacomputer.*

Writing space of the Problem: (For Student's use only)

Post-lab:-

1. *Extract possible classification rules from the given decision tree.*



2. *Write the sequential covering algorithm used in rule induction.*

3. *Difference between Decision tree and rule based classification.*

VivaVoce:-

1. *Rule-Based classifier classifies records by using a collection of________rules.*
2. *Most rule-based classification systems use which strategy?*
3. *Difference between class-based ordering and rule-based ordering.*
4. *Briefly explain the below terms in your own words:*
   a. *Mutually exclusive*
   b. *Exhaustive*
5. *Name the terms that define the following statements:*
   a. *Fraction of records that satisfy only antecedent of a rule.*
   b. *Fraction of records that satisfy both antecedent and consequent of a rule.*

(For Evaluator's use only)

Comment of the Evaluator (if Any)		*Evaluator's Observation* *Marks Secured:_________out of____________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Lab#12: Outliers Detection

Date of the Session:_____________/___/____ *Time of the Session:____to______*

Pre-lab:-

1. *What do you mean by an outlier? What are the main causes for outliers?*

2. *What are the important methods for outlier detection?*

3. *Why is outlier detection necessary in data analysis?*

4. *How do we calculate z-score?*

5. *Consider the below dataset which comprises of the income (in thousands) of 15 people in an organisation.*

   [45, 51, 63, 48, 67, 48, 56, 2, 62, 59, 44, 61, 99, 46, 52]

   What do you observe from the above data? Is there any significant difference between the incomes of few employees? If so, what could be the reason of it?

In-lab:-

1. The dataset Boston house prices consists of 9 attributes CRIM, ZN, INDUS, LSTAT, NOX, RM, DIS, RAD, TAX. The description of each attribute

- *CRIM per capita crime rate by town*
- *ZN proportion of residential land zoned for lots over25,000 sq.ft.*
- *INDUS proportion of non-retail business acres per town*
- *NOX nitricoxides concentration(parts per 10 million)*
- *RM average number of room per dwelling*
- *DIS weighted distances to five Boston employment centres*
- *RAD index of accessibility to radial highways*
- *TAXfull-valueproperty-taxrateper$10,000*

Boston dataset: *https://drive.google.com/file/d/1YVYWQWPKsLX1UM-0XCnGCwD1NIi7_uIv/view?usp=sharing*

a. *Using boxplot detect which columns have outliers*
b. *Implement scatter plot between INDUS and TAX and inspect the outliers*
c. *Apply z_score outlier detection method on Boston dataset considering threshold =3*
d. *Print any five z_score values of the outliers.*
e. *Remove all the outliers obtained from the dataset and refashion the dataset.*
f. *Apply interquantile range (IQR) outlier detection on the dataset and print IQR values of each columns.*
g. *Calculate lower_bound and upper_bound and print Boolean values wherein the outliers are represented as TRUE.*
h. *Removealltheoutliersproducedbyinterquartilerangemethodandrefashionthedataset.*

Writing space of the Problem: (For Student's use only)

2. *Consider the iris dataset. It includes three iris species with 50 samples each as well as some properties about each flower. The columns in this dataset are:*
    - *Sepal Length Cm*
    - *Sepal Width Cm*
    - *Petal Length Cm*
    - *Petal Width Cm*
    - *Species*

https://drive.google.com/file/d/1HEEMrAQqAynHdM5TmK0G-mD5Qr0OW2J8/view?usp=sharing

Import the csv file and use the box plot method to visualize the outliers considering the 4 properties of a flower. You will notice that one of the properties has outliers.

1. *Considering the range of the outliers from the visualisation, display the observations which have outliers.*
2. *Implement a DBSCAN model fitting on the dataset taking epsilon value as 0.8 and minimum samples value as 19.*
3. *Print the counter values using the counter function on the model labels.*
4. *Considering the values obtained from the model labels print the outliers of the data.*
5. *Draw a scatter plot between petal length and sepal width to visualise the outliers.*

Writing space of the Problem: (For Student's use only)

Post-lab:-

Consider the following student dataset https://drive.google.com/file/d/1edmKnHjXkTyHT6gSYhwLw9rTpzoy1Ciq/view?usp=sharing

Which consists of student details of two schools in a town?

i. Find the students who have taken more number of leaves than the average number of absences by implementing a z_score function taking mean and standard deviation into account.

ii. Find the number of students who got least and highest score in the subject G1 considering threshold =2.5

iii. Apply box plot for the above two instances.

2. Can we find outliers for categorical values? Explain.

3. A sugar factory weighs every sugar packet in the weighing machine before packing them into cartons. As per the guidelines of the factory, the standard weight of each sugar packet should be 60 grams. It has been observed that during the final weighing of the packets, few of them gave an anomalous weight due to malfunctioning of weighing machines.
Consider the below dataset which comprises of weights of the packets.https://drive.google.com/file/d/1JkdkQ3j-J93DCfZa3kUjDycEtRzShk6V/view?usp=sharing

a. Find those anomalous weights by plotting a histogram

b. In the range 0to1, consider the lower_bound = 0.1 & upper_bound =0.9 and find the outliers using the quantile method.

c. Segregate the outliers from in lines using "loc" method to get the values of "true_index". Also obtain values of "false_index".

.

d. Now find the median from the values obtained in "true_index"

.

e. Replace all the outliers with median.

Writing space of the Problem: (For Student's use only)

VivaVoce:-

1. *Is it good to remove an outlier from the dataset all the time?*
2. *What the applications of outlier detection.*
3. *What the different types of outliers?*
4. *Are outliers just side products of some clustering algorithms?*
5. *What is the difference between noise and anomaly?*

(For Evaluator's use only)

Comment of the Evaluator (if Any)	*Evaluator's Observation* *Marks Secured:________out of___________* *Full Name of the Evaluator:* *Signature of the Evaluator Date of Evaluation:*

Experiment #	<TO BE FILLED BY STUDENT>	Student ID	<TO BE FILLED BY STUDENT>
Date	<TO BE FILLED BY STUDENT>	Student Name	<TO BE FILLED BY STUDENT>