Department of Computer Engineering

Lab Manual

Final Year Semester VIII

Subject: Applied Data Science

EVEN SEMESTER

Institutional Vision, Mission and Quality Policy

Our Vision

To foster and permeate higher and quality education with value added engineering and technology programs by providing all facilities in terms of technology and platforms for all-round development with social awareness for youths.

Our Mission

- To become a pivotal center of service to Industry, academy, and society with the latest technology by providing facilities for advanced research and development programs on par with international standards.
- To produce engineering and technology professionals who are innovative and inspiring thought leaders, adept at solving problems faced by our nation and world by providing quality education.

Our Quality Policy

ज्ञानधीनं जगत् सर्वम ।

Knowledge is supreme.

Our Quality Policy

It is our earnest endeavour to produce high quality engineering professionals who are innovative and inspiring, thought and action leaders, competent to solve problems faced by society, nation and world at large by striving towards very high standards in learning, teaching and training methodology.

Our Motto: If it is not of quality, it is NOTRAIT!

Departmental Program Educational Objectives (PEOs)

1. **Learn and Integrate**
   To provide Computer Engineering students with a strong foundation in the mathematical, scientific and engineering fundamentals necessary to formulate, solve and analyze engineering problems and to prepare them for graduate studies.

2. **Think and Create**
   To develop an ability to analyze the requirements of the software and hardware, understand the technical specifications, create a model, design, implement and verify a computing system to meet specified requirements while considering real-world constraints to solve real world problems.

3. **Broad Base**
   To provide broad education necessary to understand the science of computer engineering and the impact of it in a global and social context.

4. **Techno-leader**
   To provide exposure to emerging cutting edge technologies, adequate training & opportunities to work as teams on multidisciplinary projects with effective communication skills and leadership qualities.

5. **Practice citizenship**
   To provide knowledge of professional and ethical responsibility and to contribute to society through active engagement with professional societies, schools, civic organizations or other community activities.

6. **Clarify Purpose and Perspective**
   To provide strong in-depth education through electives and to promote student awareness on the life-long learning to adapt to innovation and change, and to be successful in their professional work or graduate studies.

Departmental Program Outcomes (POs)

PO1: Engineering knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

PO2: Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

PO3 : Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

PO4: Conduct investigations of complex problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

PO5: Modern tool usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

PO6: The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

PO7: Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

PO8: Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

PO9: Individual and team work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

PO10 : Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

PO11 : Project management and finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

PO12 : Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Program Specific Outcomes: PSO

PSO1: To build competencies towards problem solving with an ability to understand, identify, analyze and design the problem, implement and validate the solution including both hardware and software.

PSO2: To build appreciation and knowledge acquiring of current computer techniques with an ability to use skills and tools necessary for computing practice.

PSO3: To be able to match the industry requirements in the area of computer science and engineering. To equip skills to adopt and imbibe new technologies.

Index

Sr. No.	Contents	Page No.
1.	List of Experiments	
2.	Experiment Plan and Course Outcomes	
3.	Mapping of Course Outcomes – Program Outcomes and Program Specific outcome	
4.	Study and Evaluation Scheme	
5.	Experiment No. 1	
6.	Experiment No. 2	
7.	Experiment No. 3	
8.	Experiment No. 4	
9.	Experiment No. 5	
10.	Experiment No. 6	
11.	Experiment No. 7	
12.	Experiment No. 8	
13.	Experiment No. 9	
14.	Experiment No.10	

List of Experiments

Sr. No.	Experiment Name
1.	Formulate Problem statement as a Case study and Explore the descriptive and inferential statistics on the same case study dataset.
2.	Perform data cleaning techniques w.r.t. Case Study.
3.	Explore data visualization techniques for your Problem statement.
4.	Implement and explore performance evaluation metrics for Data Models.
5.	To generate synthetic data using SMOTE technique.
6.	Outlier detection using density/distance based method.
7.	To Implement time series forecasting w.r.t. Case study.
8.	Illustrate data science lifecycle for selected case study.
9.	Content Beyond Syllabus: Introduction to Rapidminer

Course Objective, Course Outcome & Experiment Plan

Course Objective:

1	1. To explore various stages in the data science lifecycle.
2	2. To understand data preparation, exploration and visualization techniques.
3	To model and evaluate different supervised/unsupervised learning techniques

Course Outcomes:

CO1	Understand fundamental concept of descriptive and inferential statistics on the dataset.
CO2	Apply data cleaning techniques.
CO3	Explore data visualization techniques.
CO4	Iimplement and explore performance evaluation metrics for Data Models.
CO5	Generate synthetic data.
CO6	Apply distance based method to detect Outlier detection and Implement time series forecasting

Experiment Plan:

Module No.	Week No.	Experiments Name	Course Outcome	Weightage
1.	W1	Formulate Problem statement as a Case study and Explore the descriptive and inferential statistics on the same case study dataset.	CO1	10
2.	W2	Perform data cleaning techniques wrt Case Study.	CO2	10
3.	W3	Explore data visualization techniques for your Problem statement.	CO3	10
4.	W4	Implement and explore performance evaluation metrics for Data Models.	CO4	10
5.	W5	To generate synthetic data using SMOTE technique.	CO5	10
6	W6	Outlier detection using distance based method.	CO6	03
7.	W7	To Implement time series forecasting w.r.t. Case study.	CO6	02
8.	W8	Illustrate data science lifecycle for selected case study.	CO6	05
9.	W9	**Content Beyond Syllabus: Introduction to Rapidminer**		

CO-PO & PSO Mapping

Mapping of Course outcomes with Program Outcomes:

Subject Weight	Course Outcomes	Contribution to Program outcomes											
		1	2	3	4	5	6	7	8	9	10	11	12
Practical 80%	Understand fundamental concept of descriptive and inferential statistics on the dataset.		1	1	1	3	1			2			1
	Apply data cleaning techniques		1	2	2	2				1		1	1
	Explore data visualization techniques.		1	2	2	2				1		1	1
	Implements and explore performance evaluation metrics for Data Models.		1	2	2	2				1		1	1
	Generate synthetic data.		1	2	2	2				1		1	1
	Apply distance based method to detect Outlier detection and Implement time series Forecasting		1	2	1	1		1	1	1		1	1

Mapping of Course outcomes with Program Specific Outcomes:

Course Outcomes		Contribution to Program Specific outcomes		
		PSO1	PSO2	PSO3
CO1	Understand fundamental concept of descriptive and inferential statistics on the dataset.	3	3	2
CO2	Apply data cleaning techniques	3	3	3
CO3	Explore data visualization techniques.	3	3	3
CO4	Implements and explore performance evaluation metrics for Data Models.	3	3	3
CO5	Generate synthetic data.	3	3	3
CO6	Apply distance based method to detect Outlier detection and Implement time series Forecasting	2	2	3

Study and Evaluation Scheme

Course Code	Course Name	Teaching Scheme			Credits Assigned			
		Theory	Practical	Tutorial	Theory	Practical	Tutorial	Total
CEHDL501	Foundation of Data Science Lab	--	02	--	--	01	--	01

Course Code	Course Name	Examination Scheme		
		Term Work	Practical & Oral	Total
CEHDL501	Foundation of Data Science Lab	25	25	50

Term Work:

The Term work Marks are based on the weekly experimental performance of the students, Oral performance and regularity in the lab.

Students are expected to be prepared for the lab ahead of time by referring the manual and perform the experiment under the guidance and discussion. Next week the experiment write-up to be corrected along with oral examination.

End Semester Examination:

End of the semester, there will be oral evaluation based on the Theory and laboratory work.

Applied Data Science Lab

Experiment No. : 1

Formulate Problem statement as a Case study and explore the descriptive and inferential statistics on the same case study dataset.

Experiment No.1

1. **Aim:** Formulate Problem statement as a Case study and Explore the descriptive and inferentia statistics on the same case study dataset
2. **Objectives:**
   - To study the descriptive and inferential statistics
   - To analyse the real time dataset

   Outcomes: Students will be able to analyze statistical parameters for their dataset

3. **Hardware / Software Required: Python**
4. **Theory:**

Identify case Study.
Student can take any case study related to Data Science of their choice.
Following are suggested case studies.
Customer Segmentation
Fraud Detection
House Price prediction
Product Recommendation
Stock price prediction
Weather prediction

Problems statement should consist of following points.
1. Purpose– it includes the introduction and purpose of case study.
2. Scope –it includes the detailed scope and working of the case study.
3. Functionalities-it includes the detailed functions included of working model.

Note: Students has to write above mentioned things according to their case study topic.

Descriptive Statistics:
Descriptive Statistics describes the characteristics of a data set. It is a simple technique to describe, show and summarize data in a meaningful way.
There are three major types of Descriptive Statistics.

1. Frequency Distribution

Frequency distribution is used to show how often a response is given for quantitative as well as qualitative data. It shows the count, percent, or frequency of different outcomes occurring in a given data set. Frequency distribution is usually represented in a table or graph. Bar charts, histograms, pie charts, and line charts are commonly used to present frequency distribution. Each entry in the graph or table is accompanied by how many times the value occurs in a specific interval, range, or group.

These tables of graphs are a structured way to depict a summary of grouped data classified on the basis of mutually exclusive classes and the frequency of occurrence in each respective class.

2. Central Tendency

Central tendency includes the descriptive summary of a dataset using a single value that reflects the center of the data distribution. It locates the distribution by various points and is used to show average or most commonly indicated responses in a data set. Measures of central tendency or measures of central location include the mean, median, and mode. Mean refers to the average or most common value in a data set, while the median is the middle score for the data set in increasing order, and mode is the most frequent value.

3. Dispersion:

A measure of variability identifies the range, variance, and standard deviation of scores in a sample. This measure denotes the range and width of distribution values in a data set and determines how to spread apart the data points are from the center. The range shows the degree of dispersion or the difference between the highest and lowest values within the data set. The variance refers to the degree of the spread and is measured as an average of the squared deviations. The standard deviation determines the difference between the observed score in the data set and the mean value. This descriptive statistic is useful when you want to show how to spread out your data is and how it affects the mean.

Inferential Statistics

Inferential Statistics helps to draw conclusions and make predictions based on a data set. It is done using several techniques, methods, and types of calculations. Some of the most important types of inferential statistics calculations are:

1. Regression Analysis

Regression models show the relationship between a set of independent variables and a dependent variable. This statistical method lets you predict the value of the dependent variable based on different values of the independent variables. Hypothesis tests are incorporated to determine whether the relationships observed in sample data actually exist in the data set.

2. Hypothesis Tests

Hypothesis testing is used to compare entire populations or assess relationships between variables using samples. Hypotheses or predictions are tested using statistical tests so as to draw valid inferences.

3. Confidence Intervals

The main goal of inferential statistics is to estimate population parameters, which are mostly unknown or unknowable values. A confidence interval observes the variability in a statistic to draw an interval estimate for a parameter. Confidence intervals take uncertainty and sampling error into account to create a range of values within which the actual population value is estimated to fall.

7. Conclusion:

We have analyze the data set using descriptive and inferential statistics.

8. Viva Questions:

- What is descriptive Statistics?
- What is inferential Statistics

References:

1. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
2. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 2

Perform data cleaning techniques w.r.t. Case Study.

Experiment No.2

1. **Aim:** Perform data cleaning techniques w.r.t. Case Study.
2. **Objectives:**
   - To study the data cleaning techniques
   - To analyze the real time dataset

   Outcomes: Students will be able to apply different data cleaning techniques on given dataset

3. **Hardware / Software Required: Python**

4. **Theory:** Data cleaning, data cleansing, or data scrubbing is the act of first identifying any issues or bad data, then systematically correcting these issues. If the data is unfixable, you will need to remove the bad elements to properly clean your data. Unclean data normally comes as a result of human error, scraping data, or combining data from multiple sources. Multichannel data is now the norm, so inconsistencies across different data sets are to be expected. Here are few effective data cleaning techniques:

1. Remove duplicates
2. Remove irrelevant data
3. Standardize capitalization
4. Convert data type
5. Clear formatting
6. Fix errors
7. Language translation
8. Handle missing values

1. Remove Duplicates

When you collect your data from a range of different places, or scrape your data, it’s likely that you will have duplicated entries. These duplicates could originate from human error where the person inputting the data or filling out a form made a mistake.Duplicates will inevitably skew your data and/or confuse your results. They can also just make the data hard to read when you want to visualize it, so it’s best to remove them.

2. Remove Irrelevant Data

Irrelevant data will slow down and confuse any analysis that you want to do. So, deciphering what is relevant and what is not is necessary before you begin your data cleaning. For instance, if you are analyzing the age range of your customers, you don’t need to include their email addresses.

Other elements you’ll need to remove as they add nothing to your data include:

Personal identifiable (PII) data

URLs

HTML tags

Boilerplate text (for ex. in emails)

Tracking codes

Excessive blank space between text

3. Standardize Capitalization

Within your data, you need to make sure that the text is consistent. If you have a mixture of capitalization, this could lead to different erroneous categories being created. It could also cause problems when you need to translate before processing as capitalization can change the meaning. For instance, Bill is a person's name whereas a bill or to bill is something else entirely. If, in addition to data cleaning, you are text cleaning in order to process your data with a computer model, it’s much simpler to put everything in lowercase.

4. Convert Data Types

Numbers are the most common data type that you will need to convert when cleaning your data. Often numbers are imputed as text, however, in order to be processed, they need to appear as numerals. If they are appearing as text, they are classed as a string and your analysis algorithms cannot perform mathematical equations on them.The same is true for dates that are stored as text. These should all be changed to numerals. For example, if you have an entry that reads September 24th 2021, you’ll need to change that to read 09/24/2021.

5. Clear Formatting

Machine learning models can’t process your information if it is heavily formatted. If you are taking data from a range of sources, it’s likely that there are a number of different document formats. This can make your data confusing and incorrect. You should remove any kind of formatting that has been applied to your documents, so you can start from zero. This is normally not a difficult process, both excel and google sheets, for example, have a simple standardization function to do this.

6. Fix Errors

Errors as avoidable as typos could lead to you missing out on key findings from your data. Some of these can be avoided with something as simple as a quick spell-check. Spelling mistakes or extra punctuation in data like an email address could mean you miss out on communicating with your customers. It could also lead to you sending unwanted emails to people who didn’t sign up for them. Other errors can include inconsistencies in formatting. For example, if you have a column of US dollar amounts, you’ll have to convert any other currency type into US dollars so as to preserve a consistent standard currency. The same is true of any other form of measurement such as grams, ounces, etc.

7. Language Translation

To have consistent data, you’ll want everything in the same language. The Natural Language Processing (NLP) models behind software used to analyze data are also predominantly monolingual, meaning they are not capable of processing multiple languages. So, you'll need to translate everything into one language.

8. Handle Missing Values

When it comes to missing values you have two options:

1. Remove the observations that have this missing value
2. Input the missing data

The real-world data often has a lot of missing values. The cause of missing values can be data corruption or failure to record data. The handling of missing data is very important during the preprocessing of the dataset as many machine learning algorithms do not support missing values.

1. Deleting Rows with missing values
2. Impute missing values for continuous variable
3. Impute missing values for categorical variable
4. Using Algorithms that support missing values
5. Prediction of missing values
6. Imputation using Deep Learning Library — Datawig

5. Conclusion:

We have learnt about data cleaning techniques.

6. Viva Questions:

1. What data cleaning?
2. What are different techniques used in data cleaning?

References: References:

1. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
2. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : Data visualization techniques

Experiment No.3

1. **Aim:** Explore data visualization techniques for given problem statement.
2. **Objectives:**
    - To study the visualization techniques.
    - To analyze the real time dataset.
3. **Outcomes:** Students will be able to apply visualization techniques on real time data.
4. **Hardware / Software Required: Python**
5. **Theory:** Data visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data. Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose. If you’ve ever stared at a massive spreadsheet of data and couldn’t see a trend, you know how much more effective a visualization can be. The uses of Data Visualization as follows.

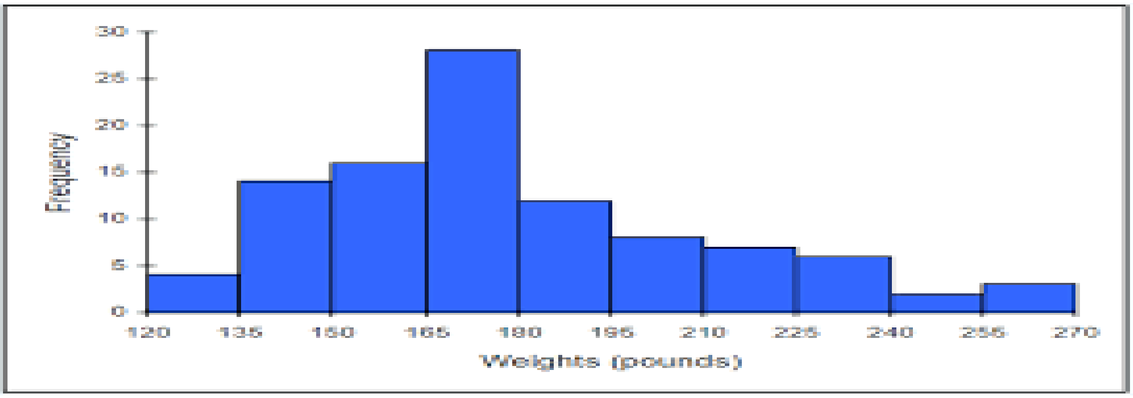
1. Powerful way to explore data with presentable results.
2. Primary use is the pre-processing portion of the data mining process.
3. Supports the data cleaning process by finding incorrect and missing values.
4. For variable derivation and selection means to determine which variable to include and discarded in the analysis.
5. Also play a role in combining categories as part of the data reduction process.

Data Visualization Techniques:

- Histograms
- Box plot
- Scatter plot
- Bubble chart
- Density Chart
- Distribution Chart

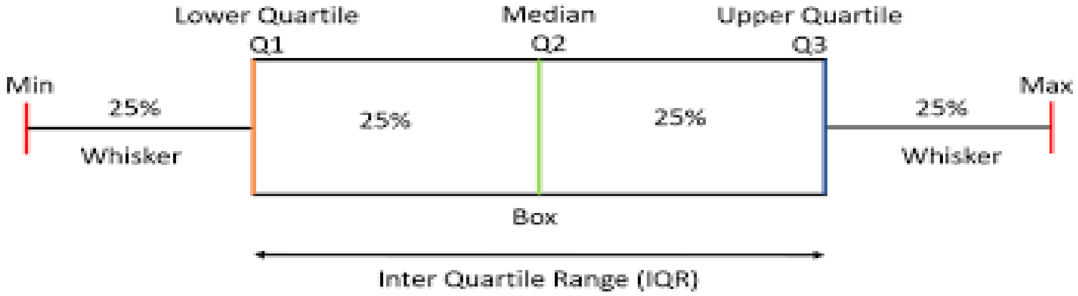
Histograms

A histogram is a graphical display of data using bars of different heights. In a histogram, each bar groups numbers into ranges. Taller bars show that more data falls in that range. A histogram displays the shape and spread of continuous sample data. It is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc. It is an accurate representation of the distribution of numerical data, it relates only one variable. Includes bin or bucket- the range of values that divide the entire range of values into a series of intervals and then count how many values fall into each interval. Bins are consecutive, non- overlapping intervals of a variable. As the adjacent bins leave no gaps, the rectangles of histogram touch each other to indicate that the original value is continuous.



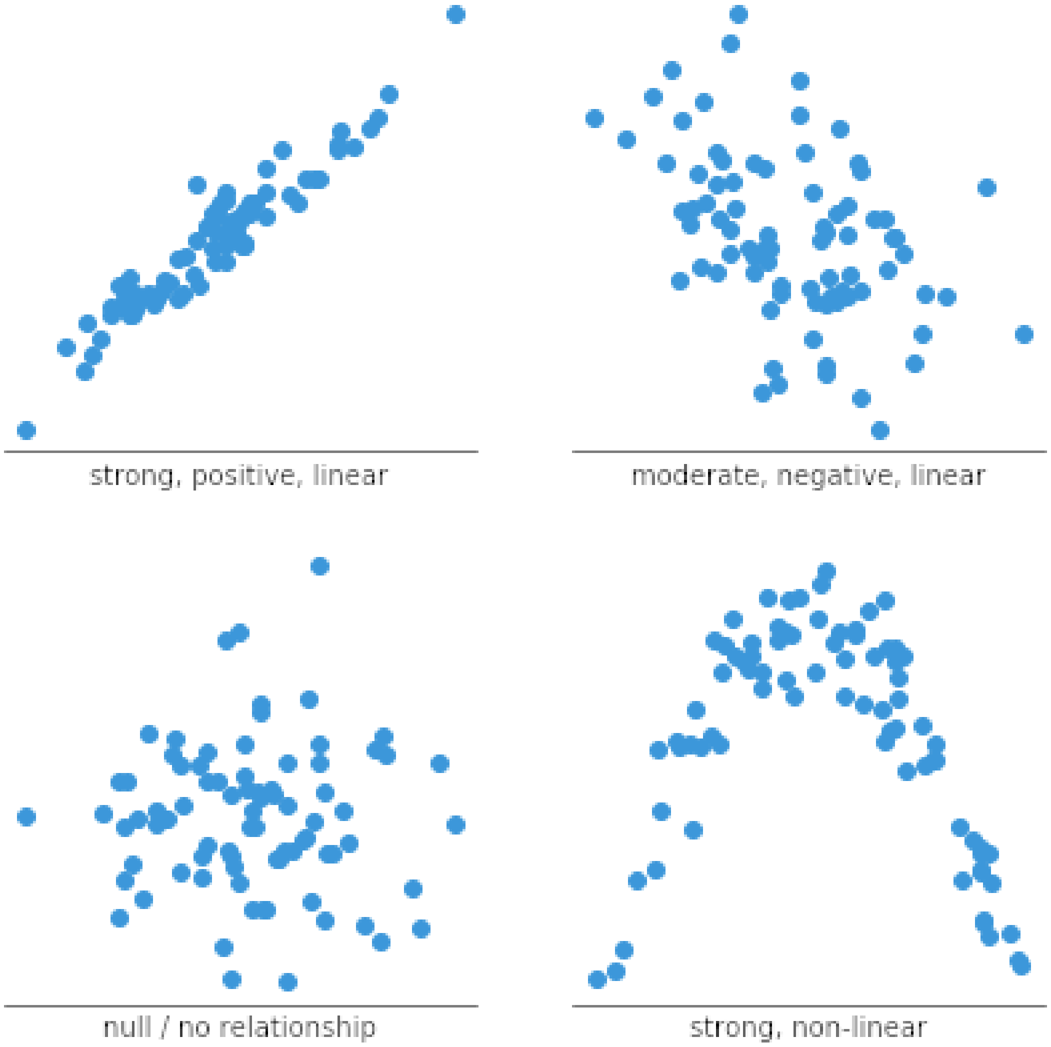
Box Plots

A boxplot is a standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It can tell you about your outliers and what their values are. It can also tell you if your data is symmetrical, how tightly your data is grouped, and if and how your data is skewed.A box plot is a graph that gives you a good indication of how the values in the data are spread out. Although box plots may seem primitive in comparison to a histogram or density plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets. For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode).



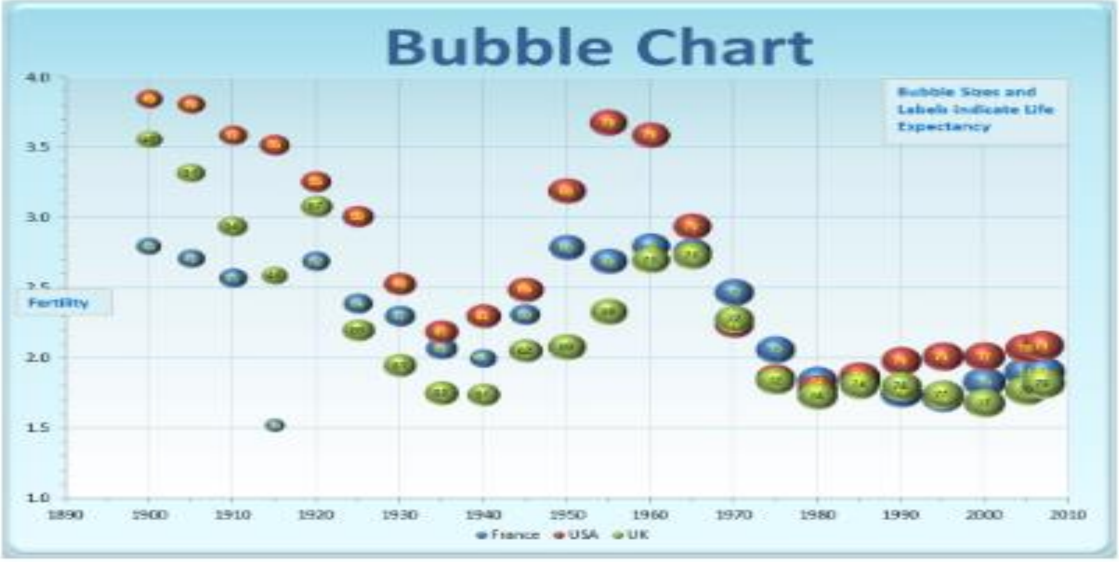
Scatter plot

You need to have information on the variability or dispersion of the data.A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables. Scatter plots' primary uses are to observe and show relationships between two numeric variables. The dots in a scatter plot not only report the values of individual data points, but also patterns when the data are taken as a whole.Identification of correlational relationships are common with scatter plots. In these cases, we want to know, if we were given a particular horizontal value, what a good prediction would be for the vertical value. You will often see the variable on the horizontal axis denoted an independent variable, and the variable on the vertical axis the dependent variable. Relationships between variables can be described in many ways: positive or negative, strong or weak, linear or nonlinear.



Bubble Chart

A bubble chart is a variation of a scatter chart in which the data points are replaced with bubbles, and an additional dimension of the data is represented in the size of the bubbles. Just like a scatter chart, a bubble chart does not use a category axis — both horizontal and vertical axes are value axes. In addition to the x values and y values that are plotted in a scatter chart, a bubble chart plots x values, y values, and z (size) values.



Conclusion: We have learnt about data visualization techniques.

7. **Viva Questions:**
    1. What data visualization?
    2. What are different data visualization techniques?

References: References:

3. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
4. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 4

Implement and explore performance evaluation metrics for Data Models.

1. **Aim:** Implement and explore performance evaluation metrics for Data Models
2. **Objectives:**
   - To study the implementation of different data models using machine learning algorithm
   - To analyze the evaluation metrics for data models

Outcomes: Students will implement different machine learning algorithm and able to evaluate different data models

3. **Hardware / Software Required: Python**

Theory: Steps Involved in Data Science Modelling

The key steps involved in Data Science Modelling are:

- Step 1: Understanding the Problem
- Step 2: Data Extraction
- Step 3: Data Cleaning
- Step 4: Exploratory Data Analysis
- Step 5: Feature Selection
- Step 6: Incorporating Machine Learning Algorithms
- Step 7: Testing the Models
- Step 8: Deploying the Model

Step 6 Incorporating Machine Learning Algorithms

This is one of the most crucial processes in Data Science Modelling as the Machine Learning Algorithm aids in creating a usable Data Model. There are a lot of algorithms to pick from, the Model is selected based on the problem. There are three types of Machine Learning methods that are incorporated:

1) Supervised Learning

It is based on the results of a previous operation that is related to the existing business operation. ious patterns, Supervised Learning aids in the prediction of an outcome. Some of the Supervised Learning Algorithms are:

- Linear Regression
- Random Forest
- Support Vector Machines

2) Unsupervised Learning

This form of learning has no pre-existing consequence or pattern. Instead, it concentrates on examining the interactions and connections between the presently available Data points. Some of the Unsupervised Learning Algorithms are:

- KNN (k-Nearest Neighbors)
- K-means Clustering
- Hierarchical Clustering
- Anomaly Detection

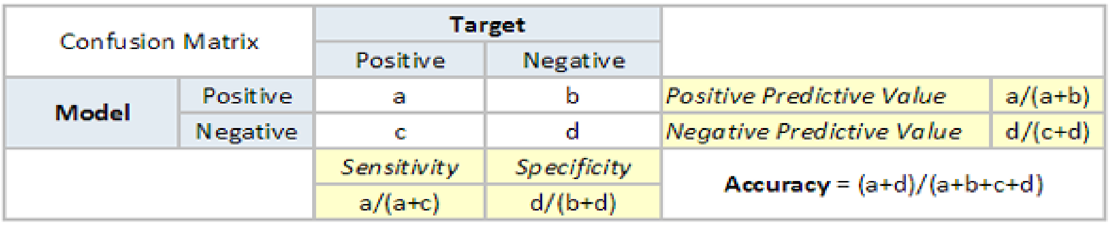
Evaluation metrics quantify the performance of a machine learning model. It involves training a model and then comparing the predictions to expected values.

1. Confusion Matrix

A confusion matrix is an N X N matrix, where N is the number of predicted classes. For the problem in hand, we have N=2, and hence we get a 2 X 2 matrix. It is a performance measuremen for machine learning classification problems where the output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. It is extremely useful for measuring precision-recall, Specificity, Accuracy, and most importantly, AUC-ROC curves.

Here are a few definitions you need to remember for a confusion matrix:

- **True Positive:** You predicted positive, and it’s true.
- **True Negative:** You predicted negative, and it’s true.
- **False Positive: (Type 1 Error):** You predicted positive, and it’s false.
- **False Negative: (Type 2 Error):** You predicted negative, and it’s false.
- **Accuracy:** the proportion of the total number of correct predictions that were correct.
- **Positive Predictive Value or Precision:** the proportion of positive cases that were correctly identified
- **Negative Predictive Value:** the proportion of negative cases that were correctly identified.
- **Sensitivity or Recall:** the proportion of actual positive cases which are correctly identified.
- **Specificity:** the proportion of actual negative cases which are correctly identified.
- **Rate:** It is a measuring factor in a confusion matrix. It has also 4 types TPR, FPR, TNR, and FNR.



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Recall

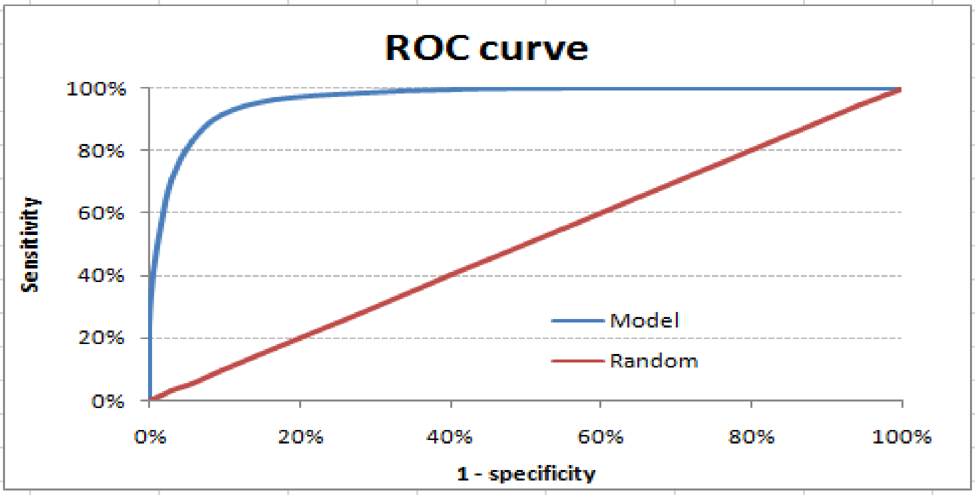
$$\textit{Recall} = \frac{TP}{TP + FN}$$

Precision

$$\textit{Precision} = \frac{TP}{TP + FP}$$

F-measure

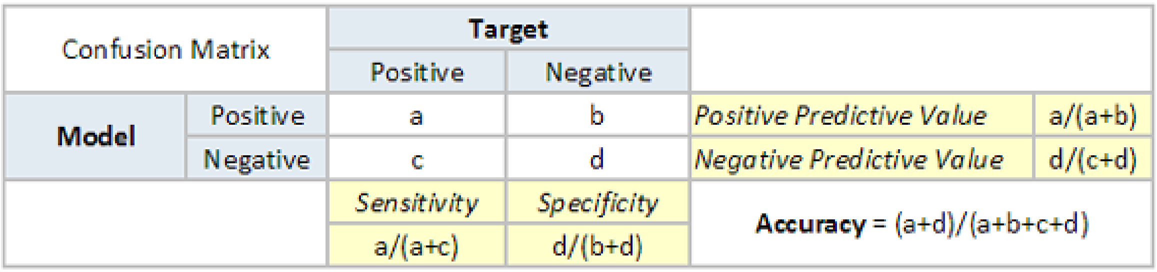
$$\textit{F - measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$



2.

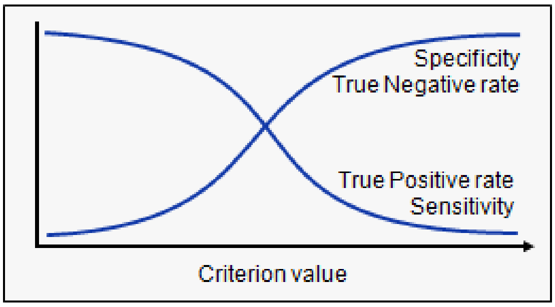
Area under the ROC Curve (AUC – ROC)

This is again one of the popular evaluation metrics used. The biggest advantage of using the ROC curve is that it is independent of the change in the proportion of responders. This statement will get clearer in the following sections.Let's first try to understand what the ROC (Receiver operating characteristic) curve is. If we look at the confusion matrix below, we observe that for a probabilistic model, we get different values for each metric.



Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

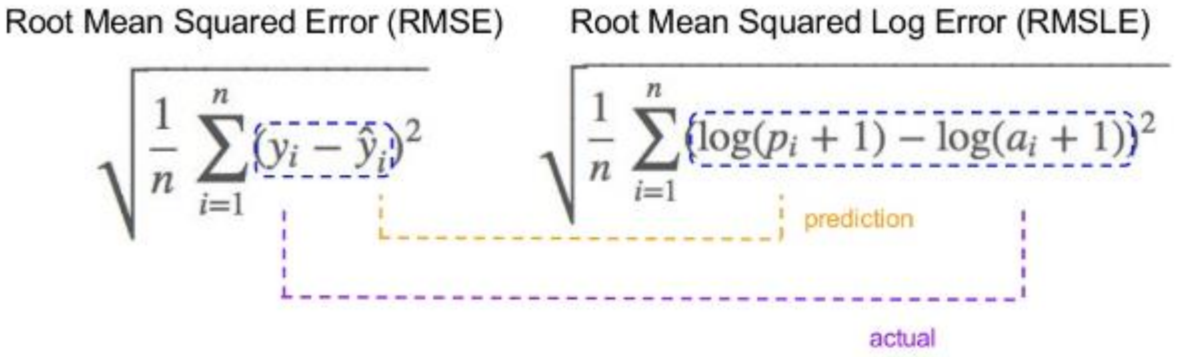
Hence, for each sensitivity, we get a different specificity. The two vary as follows:



The ROC curve is the plot between sensitivity and (1- specificity). (1- specificity) is also known as the false positive rate, and sensitivity is also known as the True Positive rate. Following is the ROC curve for the case in hand.

3.Root Mean Squared Logarithmic Error

In the case of Root mean squared logarithmic error, we take the log of the predictions and actual values. So basically, what changes are the variance that we are measuring? RMSLE is usually used when we don’t want to penalize huge differences in the predicted and the actual values wher both predicted, and true values are huge numbers.


Root Mean Squared Error (RMSE)
Root Mean Squared Log Error (RMSLE)
prediction
actual

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

1. If both predicted and actual values are small: RMSE and RMSLE are the same.
2. If either predicted or the actual value is big: RMSE > RMSLE
3. If both predicted and actual values are big: RMSE > RMSLE (RMSLE becomes almost negligible)

Conclusion: We have learnt how to implement and explore performance evaluation metrics for Data Models (Supervised/Unsupervised Learning)

8. **Viva Questions:**
    1. What are different Supervised/Unsupervised Learning techniques in data science?
    2. Explain confusion matrix?

References: References:

1. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
2. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 5

Use SMOTE technique to generate synthetic data.

1. **Aim:** Use SMOTE technique to generate synthetic data. (To solve the problem of class imbalance
2. **Objectives:**
   - To use SMOTE technique to generate synthetic data

   Outcomes: Students will able to use SMOTE technique to generate synthetic data

3. **Hardware / Software Required: Python**

4. **Theory**: Imbalanced classification involves developing predictive models on classificatio datasets that have a severe class imbalance. The challenge of working with imbalanced datase is that most machine learning techniques will ignore, and in turn have poor performance on, th minority class, although typically it is performance on the minority class that is most importan.

One approach to addressing imbalanced datasets is to oversample the minority class. Th simplest approach involves duplicating examples in the minority class, although these exampl don't add any new information to the model. Instead, new examples can be synthesized fro the existing examples. This is a type of data augmentation for the minority class and is referre to as the **Synthetic Minority Oversampling Technique**, or **SMOTE** for short.

SMOTE works by selecting examples that are close in the feature space, drawing a line betwee the examples in the feature space and drawing a new sample at a point along that line. Ste involved in SMOTE

1. Identify the minority class vector.
2. Decide the number of nearest numbers (k), to consider.
3. Compute a line between the minority data points and any of its neighbors and place a synthetic point.
4. Repeat step 3 for all minority data points and their k neighbors, till the data is balanced.

Specifically, a random example from the minority class is first chosen. Then *k* of the nearest neighbors for that example are found (typically *k=5*). A randomly selected neighbor is chose and a synthetic example is created at a randomly selected point between the two examples in feature space.

we can use imbalanced-learn Python library, which can be installed via pip as follows:

imbalanced-learn

imbalanced-learn is a python package offering a number of re-sampling techniques commonl used in datasets showing strong between-class imbalance. It is compatible with scikit-learn and is part of scikit-learn-contrib projects.

Conclusion: We have learnt how to use SMOTE technique to generate synthetic data. (To solve th problem of class imbalance)

9. **Viva Questions:**

1. What do you mean by imbalanced data?
2. Explain Synthetic Minority Oversampling Technique.

References: References:

3. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
4. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 6

Outlier detection using density based method.

1. **Aim:** Outlier detection using density based method.
2. **Objectives:**
   - To learn Outlier detection using density based method

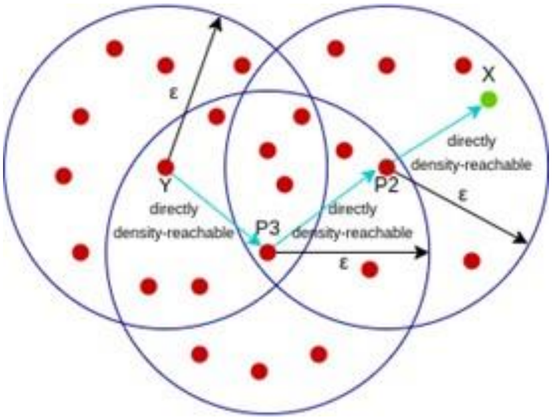
   Outcomes: Students will able to implement outlier detection using density based method.

3. **Hardware / Software Required: Python**

4. **Theory**: The DBSCAN is density fundamental cluster formation. Its advantage is that it candiscover clusters with arbitrary shapes and size. The algorithm typically regards clusters as dense regions of objects in the data space that are separated by regions of low-density objects.The algorithm has two input parameters, radious Eps and MinPts. According to the above definitions, it only needs to find out all the maximal density connected spaces to cluster the data points in an attribute space. And these density-connected spaces are the clusters. Everyobject not contained in any cluster is considered noise and can be ignored. Explanation of DBSCAN steps:

1. DBSCAN requires two parameters: Eps and MinPts. It starts with an arbitrary starting point that has not been visited. From the starting point the neighbor pointswithin distance Eps of the starting point.
2. A cluster is formed when the number of neighbors is greater than or equal to MinPts.The starting point and its neighbors are added to this cluster and the starting point is marked as visited. The algorithm then repeats the evaluation process for all the neighbors' recursively.
3. If the number of neighbors is less than MinPts, the point is marked as noise.
4. If all points within reach are visited then the algorithm proceeds to iterate through theremaining unvisited points in the data set. Data set is generated from Rapid Miner using the data generated block.

Working of DBSCAN clustering is as shown in below figure. A point X is directly density-reachable from point Y w.r.t epsilon, minPoints if, if there is a chain of points p1,p2, p3,…, pn and p1=**X** and pn=**Y** such that pi+1 is directly density-reachable from pi



As seen in the diagram, **X** is density-reachable from **Y** with **X** being directly densityreachable from **P2**, **P2** from **P3,** and **P3** from **Y.** But, the inverse of this is not valid.

DBSCAN Algorithm:

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a base algorithm for density-based clustering. It can discover clusters of different shapes andsizes from a large amount of data, which is containing noise and outliers.

The DBSCAN algorithm uses two parameters:

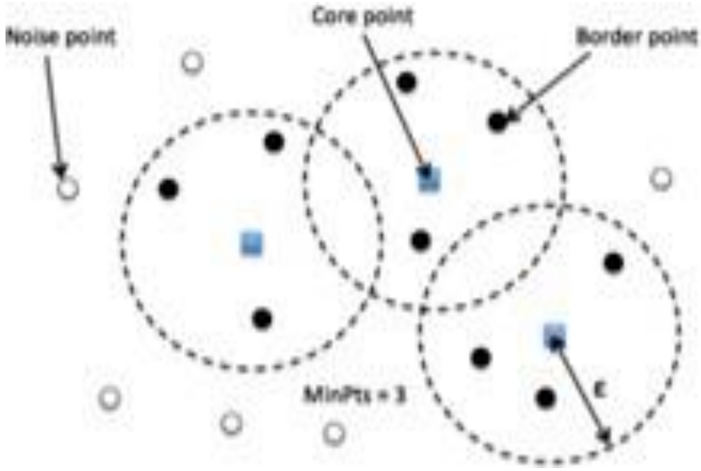
- minPts: The minimum number of points (a threshold) clustered together for a regionto be considered dense.
- eps **(ε):** A distance measure that will be used to locate the points in the neighborhoodof any point.

These parameters can be understood if we explore two concepts called DensityReachability and Density Connectivity.

- Reachability in terms of density establishes a point to be reachable from another ifit lies within a particular distance (eps) from it.
- Connectivity, on the other hand, involves a transitivity based chaining-approach todetermine whether points are located in a particular cluster. For example, p and q points could be connected if p->r->s->t->q, where a->b means b is in the neighborhood of a.

There are three types of points after the DBSCAN clustering is complete:

- **Core** — This is a point that has at least *m* points within distance *n* from itself.
- **Border** — This is a point that has at least one Core point at a distance *n*.
- **Noise** — This is a point that is neither a Core nor a Border. And it has less than *m* points within distance *n* from itself. All these points are as shown in figure.



Algorithmic steps for DBSCAN clustering:

- The algorithm proceeds by arbitrarily picking up a point in the dataset (untilall points have been visited).
- If there are at least ‘minPoint’ points within a radius of ‘ε’ to the point thenwe consider all these points to be part of the same cluster.
- The clusters are then expanded by recursively repeating the neighborhoodcalculation for each neighboring point.

Conclusion: We have learnt how to use DBSCAN algorithm for outlier detection.

10. Viva Questions:

1. What is DBSCAN clustering?
2. How DBSCAN eliminates noise?

References: References:

5. S.C. Gupta, V. K. Kapoor ―Fundamentals of Mathematical Statistics‖, S. Chand and Sons, New Delhi.
6. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 7

To implement time series forecasting w.r.t. Case study.

1. **Aim:** To Implement time series forecasting w.r.t. Case study.
2. **Objectives:**
   - To study the time series forecasting

   Outcomes: Students will study time series forecasting for Case study

3. **Hardware / Software Required: -Python**

4.Theory: Time series data is an important source for information and strategy used in various businesses. From a conventional finance industry to education industry, they play a major role in understanding a lot of details on specific factors with respect to time. Time series forecasting is basically the machine learning modeling for Time Series data (years, days, hours…etc.)for predicting future values using Time Series modeling .This helps if your data in serially correlated.

Stationarity

This is a very important concept in Time Series Analysis. In order to apply a time series model, it is important for the Time series to be stationary; in other words all its statistical properties (mean,variance) remain constant over time. This is done basically because if you take a certain behavior over time, it is important that this behavior is same in the future in order for us to forecast the series. There are a lot of statistical theories to explore stationary series than non-stationary series.

Making The Time Series Stationary

There are two major factors that make a time series non-stationary. They are:

• Trend: non-constant mean

• Seasonality: Variation at specific time-frames

The basic idea is to model the trend and seasonality in this series, so we can remove it and make the series stationary. Then we can go ahead and apply statistical forecasting to the stationary series. And finally we can convert the forecasted values into original by applying the trend and seasonality constrains back to those that we previously separated.

Trend

The first step is to reduce the trend using transformation, as we can see here that there is a strong positive trend. These transformation can be log, sq-rt, cube root etc . Basically it penalizes larger values more than the smaller. In this case we will use the logarithmic transformation.

There is some noise in realizing the forward trend here. There are some methods to model these trends and then remove them from the series. Some of the common ones are:

• Smoothing: using rolling/moving average

• Aggression: by taking the mean for a certain time period (year/month)

Smoothing:

In smoothing we usually take the past few instances (rolling estimates) We will discuss two methods under smoothing- Moving average and Exponentially weighted moving average.

Decomposing:

Here we model both the trend and the seasonality, then the remaining part of the time series is returned.

Forecasting a Time Series

Now that we have made the Time series stationary, let’s make models on the time series using differencing because it is easy to add the error , trend and seasonality back into predicted values .

We will use statistical modelling method called ARIMA to forecast the data where there are dependencies in the values.

Auto Regressive Integrated Moving Average(ARIMA) — It is like a liner regression equation where the predictors depend on parameters (p,d,q) of the ARIMA model .

Let me explain these dependent parameters:

• p : This is the number of AR (Auto-Regressive) terms . Example — if p is 3 the predictor for y(t will be y(t-1),y(t-2),y(t-3).

• q : This is the number of MA (Moving-Average) terms . Example — if p is 3 the predictor for y(t) will be y(t-1),y(t-2),y(t-3).

• d :This is the number of differences or the number of non-seasonal differences .

Combining all of the three types of models above gives the resulting ARIMA(p,d,q) model.

The ARIMA methodology is a statistical method for analyzing and building a forecasting model which best represents a time series by modeling the correlations in the data. Owing to purely statistical approaches, ARIMA models only need the historical data of a time series to generalize the forecast and manage to increase prediction accuracy while keeping the model parsimonious.

5.Conclusion: We have studied Time series forcasting

6.Viva Questions:

1. What is trend and Seasonality?
2. Explain ARIMA model.

7.References: References:

1. Vijay Kotu ,Bala Deshpande , —Data Science: Concepts and Practice —, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 8

Illustrate data science lifecycle for selected case study.

1. **Aim:** Illustrate data science lifecycle for selected case study
2. **Objectives:**
    - To study the data science lifecycle.

Outcomes: Students will study different data science lifecycle phases for selected case study

3. **Hardware / Software Required: -**

4.Theory: The data science process is a systematic approach to solving a data problem. The methodical discovery of useful relationships and patterns in data is enabled by a set of iterative activities collectively known as the data science process.

The standard data science process involves:

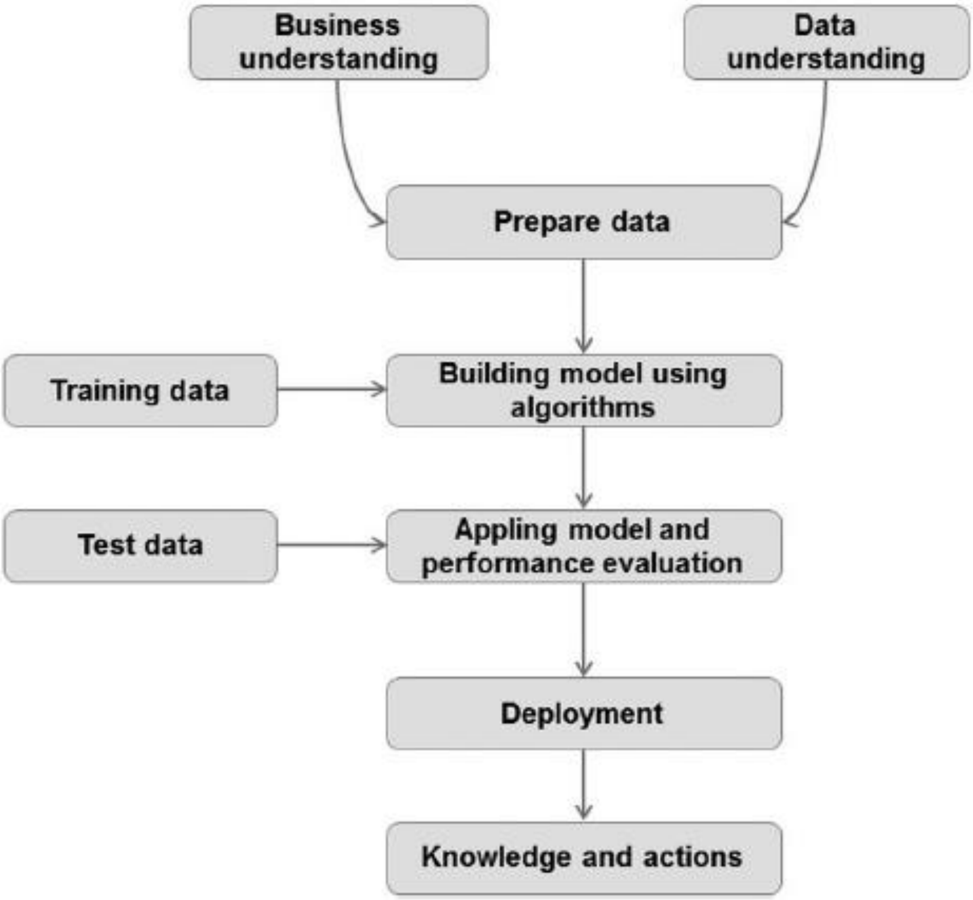
(1) Understanding the problem

(2) Preparing the data samples

(3) Developing the model

(4) Applying the model on a dataset to see how the model may work in the real world

(5) Deploying and maintaining the models.



1. Prior Knowledge
   - Existing subject matter and contextual information that is already known
2. Data Preparation
   - Clean the data and make it ready to suit a data science task
3. Modeling
   - Process of building representative models that can be inferred from the sample dataset which can be used for either predicting (predictive modeling) or describing the underlying pattern in the data (descriptive or explanatory modeling)
4. Application
   - model deployment to deal with: assessing model readiness, technical integration, response time, model maintenance, and assimilation
5. Knowledge
   - The data science process starts with prior knowledge and ends with posterior knowledge, which is the incremental insight gained.

5.Conclusion: We have different phases in data science life cycle.

6.Viva Questions:

3. What are different phase in data science life cycle?
4. Explain data Preparation phase.

7.References: References:

2. Vijay Kotu ,Bala Deshpande , ―Data Science: Concepts and Practice ―, Morgan and Kaufmann publisher (Elseveir)

Applied Data Science Lab

Experiment No. : 9

Content beyond Syllabus: Introduction to Rapidminer

1. **Aim:** To study and visualize the real time dataset using Rapid Miner
2. **Objectives:**
   - To study the visualization tool Rapid Miner
   - To analyse the real time dataset

   Outcomes: Students will be able to apply Rapid Miner tool to visualize realtime data

3. **Hardware / Software Required: Rapid Miner Studio**

4. **Theory:** Rapid Miner is a platform for data scientists and big data analysts to quickly analyze their data. Rapid Miner has taken a huge leap in the AI community since it is most popularly used by non-programmers and researchers. The platform provides a vastnumber of options in terms of plugins and data analysis techniques. The idea behind RapidMining tool is to create one place for everything. Starting from providing multiple datasets to model deployment

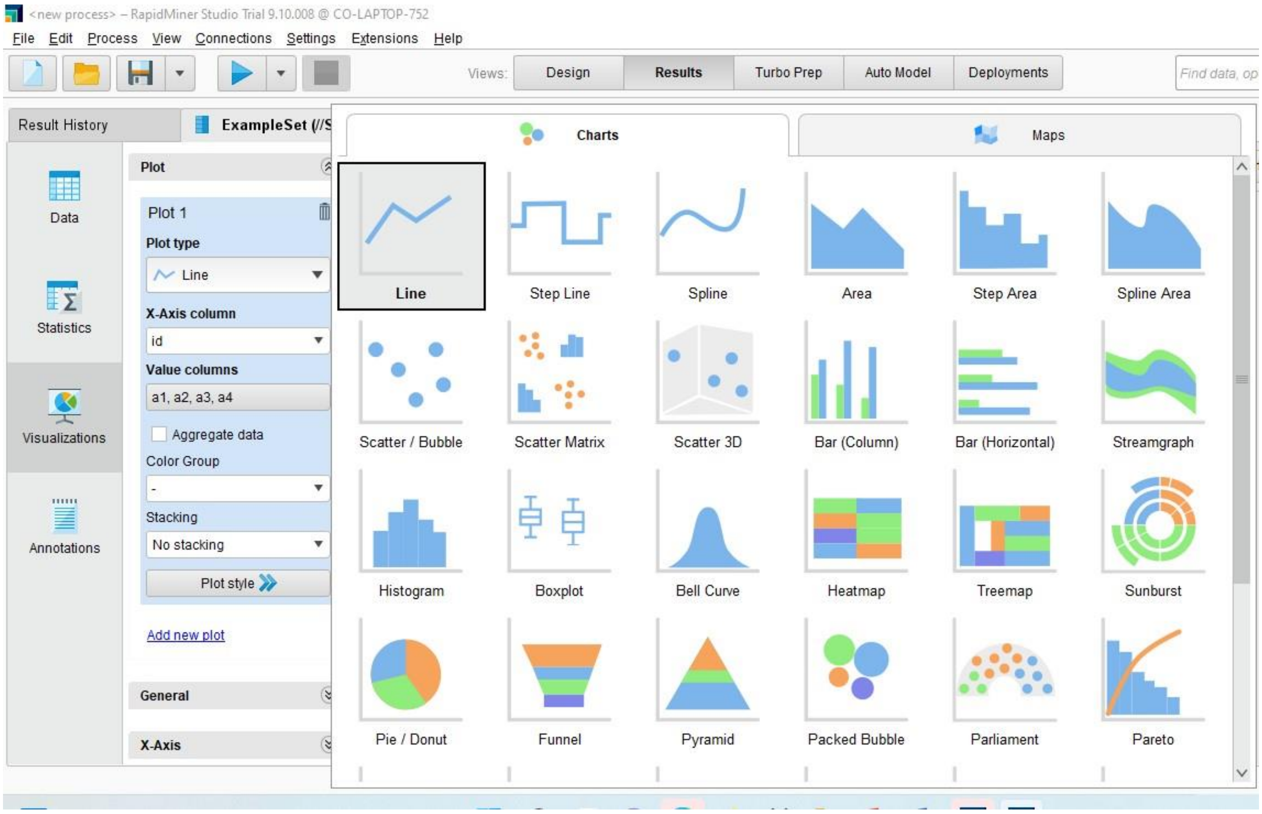
   Some of the facilities of this platform are:

   - Rapid Miner provides its own collection of datasets but it also provides options to set up a database in the cloud for storing large amounts of data. You can storeand load the data from Hadoop, Cloud, RDBMS, NoSQL etc. Apart from this, you can load your CSV data very easily and start using it as well.
   - The standard implementation of procedures like data cleaning, visualization, pre-processing can be done with drag and drop options without having to write evena single line of code.
   - Rapid Miner provides a wide range of machine learning algorithms in classification, clustering and regression as well. You can also train optimal deeplearning algorithms like Gradient Boost, XGBoost etc. Not only this, but the toolalso provides the ability to perform pruning and tuning.
   - Finally, to bind everything together, you can easily deploy your machine learningmodels to the web or to mobiles through this platform. You just need to create user interfaces to collect real-time data and run it on the trained model to serve atask.

   A Step-by-Step Guide to Using Rapid Miner

   - The first step is to download the rapid miner tool in your local system. You can download the tool from official site. Download the ‘Rapid Miner Studio’ option andselect the operating system type of your system. Once done, wait for the downloadto complete and set up your account in the studio.
   - After creating your account you will see this screen in front of you.
   - To load some data, click the green button. Then, click on Samples folder->data. Once you have navigated to this folder you can see a list of datasets. We have pickedthe Iris dataset. You can also load your own dataset either from your local system or from a database by clicking on the Import data option.

- For visualization purposes of the data, you can click on the result button, drag and drop your dataset and you will be able to see few options as shown below. To the left click on the visualization button. Here you can play around with data visualization and see how to points are related to each other. There are a plethora ofvisualization types available as shown below.
-



Plot

Plot 1

Plot type: Line

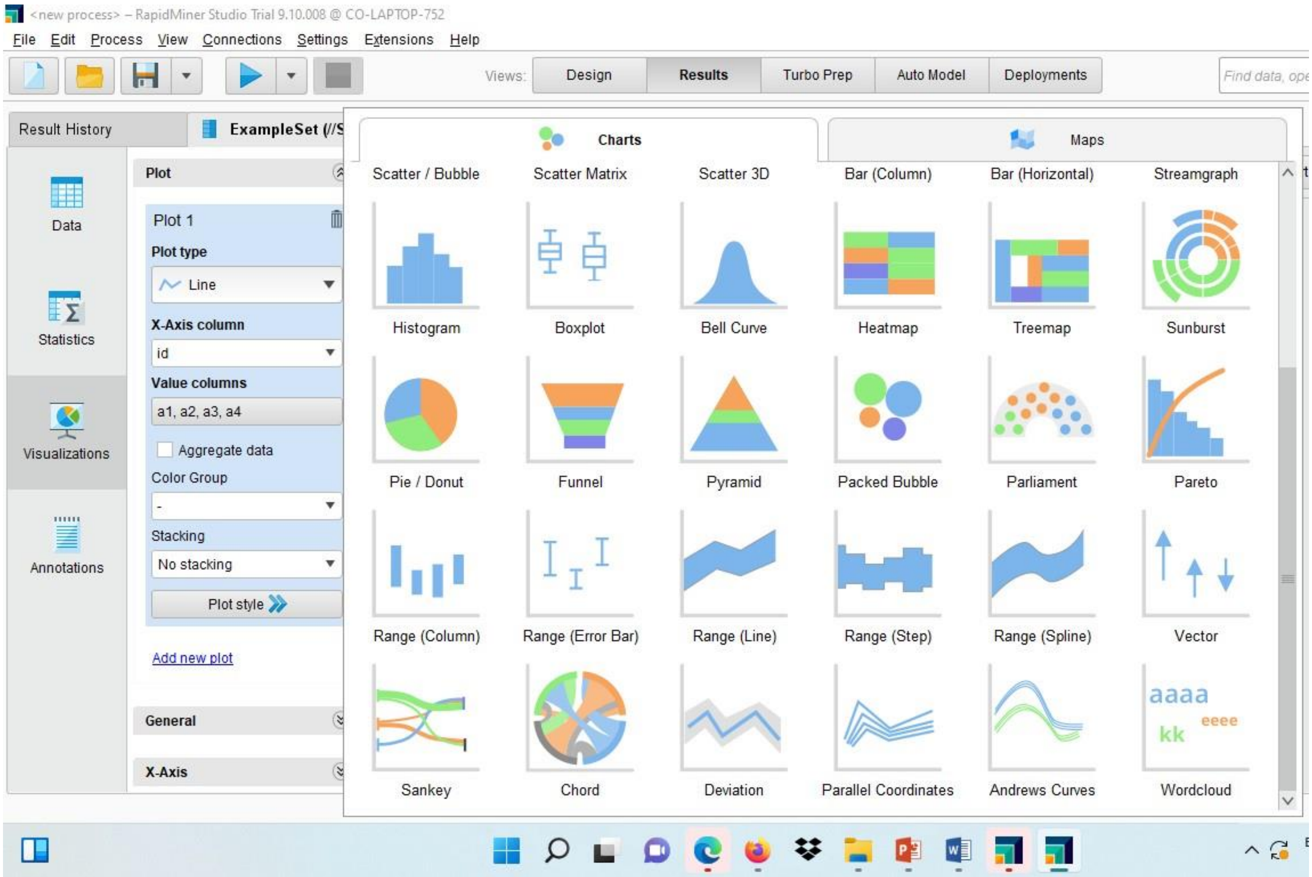
X-Axis column: id

Value columns: a1, a2, a3, a4

□ Aggregate data

Color Group: -

Stacking: No stacking



Plot type: Line

X-Axis column: id

Value columns: a1, a2, a3, a4

□ Aggregate data

Color Group: -

Stacking: No stacking

9. Conclusion:

We have learnt about working of Rapid miner and different visualization graphs in it

10. Viva Questions:

- What is data visualization?
- What are different applications of data visualization?

References:

3. http://www.rapidminer.com
4. Vijay Kotu ,Bala Deshpande , —Data Science: Concepts and Practice —, Morgan and Kaufmann publisher (Elseveir)