Baderia Global Institute of Engineering and Management

Department of Computer Science and Engineering

Name of the Student: ______________________________

Enrolment Number: ______________________________

Semester: **VI**

Department: **Computer Science and Engineering**

Subject: **Data Analytics Lab Manual**

Data Analytics Lab Manual

Contents

Vision and Mission of the Institute

Vision of Institute

Transforming lives by providing professional education with excellence.

Mission of Institute

1. Quality Education: Providing Education with quality and shaping up Technocrats and budding managers with a focus on adapting to changing technologies.
2. Focused Research & Innovation: Focusing on Research and Development and fostering Innovation among the academic community of the Institution.
3. People Focused: Accountable and committed to institutional operations for effective functioning by Faculty members, Staff and Students.
4. Holistic Learning: Focus on conceptual learning with practical experience and experiential learning with strong Industrial connections and collaborations.
5. Service to Society: Providing Technical and Managerial services to society for betterment of their quality of life with best of the skills, compassion and empathy.

Vision and Mission of the Department

Vision of the Department

Transforming the lives of graduates by providing excellent education in the field of Computer Science & Engineering.

Mission of the Department

The department strives to:

1. Create student centric learning ambience so as to produce graduates who are well informed about latest technological trends and advancement in the world of computing, technology and research.
2. Produce professionals who are capable to work in diversified fields, find workable solutions to complex problems with awareness and concern for society and environments.
3. Continuously upgrade faculty through training so that they function effectively.
4. Encourage industry institute collaborations through consultancies and research, helping students to have conceptual learning.

Program Outcomes

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

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

3. 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 public health and safety, and cultural, societal, and environmental considerations.

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

5. 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.

6. 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.

7. 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.

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

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

10. Communication: Communicate effectively on complex engineering activities with the engineering community and with the 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.

11. 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.

12. 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.

Course Outcomes

After completing the course, student should be able to:

CO1. Apply foundational principles of data analytics by employing concepts of statistics and probability.

CO2. Evaluate the significance of data processing techniques through an explanation of their relevance and necessity in handling information.

CO3. Apply data analytics techniques using R, MATLAB, and Python, demonstrating knowledge through executing relevant commands and scripts.

CO4. Analyze and synthesize information gathered through data analytics techniques, evaluating their applicability in real-life scenarios, and proposing solutions based on the results.

LABORATORY REGULATIONS AND SAFETY RULES

The following Regulations and Safety Rules must be observed in all concerned laboratory locations.

1. It is the duty of all concerned parties who use any electrical laboratory to take all reasonable steps to safeguard the HEALTH and SAFETY of themselves and all other users and visitors.
2. Make sure that all equipment is properly working before using them for laboratory exercises. Any defective equipment must be reported immediately to the Lab. Instructors or Lab. Technical Staff.
3. Students are allowed to use only the equipment provided in the experiment manual or equipment used for senior project laboratory.
4. Power supply terminals connected to any circuit are only energized with the presence of the Instructor or Lab. Staff.
5. Students should keep a safe distance from the circuit breakers, electric circuits or any moving parts during the experiment.
6. Avoid any part of your body to be connected to the energized circuit and ground.
7. Switch off the equipment and disconnect the power supplies from the circuit before leaving the laboratory.
8. Observe cleanliness and proper laboratory housekeeping of the equipment and other related accessories.
9. Wear proper clothes and safety gloves or goggles required in working areas that involves fabrications of printed circuit boards, chemicals process control system, antenna communication equipment and laser facility laboratories.
10. Double check your circuit connections specifically in handling electrical power machines, AC motors and generators before switching “ON” the power supply.
11. Make sure that the last connection to be made in your circuit is the power supply and first thing to be disconnected is also the power supply.
12. Equipment should not be removed, transferred to any location without permission from the laboratory staff.
13. Software installation in any computer laboratory is not allowed without the permission from the Laboratory Staff.
14. Computer games are strictly prohibited in the computer laboratory.
15. Students are not allowed to use any equipment without proper orientation and actual hands on equipment operation.

List of Experiments

S. No.	Experiment	Date of Completion	Date of Checking	Sign
1.	The probability that it is Friday and that a student is absent is 3%. Since there are 5 working days, the probability that it is Friday is 20%. What is the probability that the student is absent given that the day is Friday? Apply Baye's theorem.			
2.	Create two random datasets and determine the Pearson's Correlation between them. Also demonstrate that Null Hypothesis will be rejected or accepted based on this correlation coefficient.			
3.	Using a dataset compare multiple ROC curves in a single plot and automatically displays the AUC for each model as well.			
4.	Using a dataset plot confusion matrix in one line of code			
5.	Use ColumnTransformer to apply different preprocessing to different columns: ● select from DataFrame columns by name ● passthrough or drop unspecified columns			
6.	Using a dataset visualize bar plot, histogram, box plot and scatter plot using R.			
7.	Using a dataset implement linear regression using R and determine the p value.			
8.	Using a dataset demonstrates how to import data, perform a basic analysis, trend the results, and export the results to another text file using MATLAB.			

Experiment 01

Aim- The probability that it is Friday and that a student is absent is 3%. Since there are 5 working days, the probability that it is Friday is 20%. What is the probability that the student is absent given that the day is Friday? Apply Baye's theorem.

Platform- Google Colaboratory

Theory-

Bayes Theorem

Bayes theorem is a theorem in probability and statistics, named after the Reverend Thomas Bayes that helps in determining the probability of an event that is based on some event that has already occurred. Bayes theorem has many applications such as Bayesian interference, in the healthcare sector - to determine the chances of developing health problems with an increase in age and many others.

Bayes theorem, in simple words, determines the conditional probability of an event A given that event B has already occurred. Bayes theorem is also known as the Bayes Rule or Bayes Law. It is a method to determine the probability of an event based on the occurrences of prior events. It is used to calculate conditional probability. Bayes theorem calculates the probability based on the hypothesis.

Bayes theorem states that the conditional probability of an event A, given the occurrence of another event B, is equal to the product of the likelihood of B, given A and the probability of A. It is given as:

$$\boxed{P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}}$$

P (A) = how likely A happens (Prior knowledge) - The probability of a hypothesis is true before any evidence is present.

P (B) = how likely B happens (Marginalization) - The probability of observing the evidence.

P (A/B) = how likely A happens given that B has happened (Posterior) -The probability of a hypothesis is true given the evidence.

P (B/A) = how likely B happens given that A has happened (Likelihood) - The probability of seeing the evidence if the hypothesis is true.

Result- The result has been verified under experimental error conditions.

Experiment 02

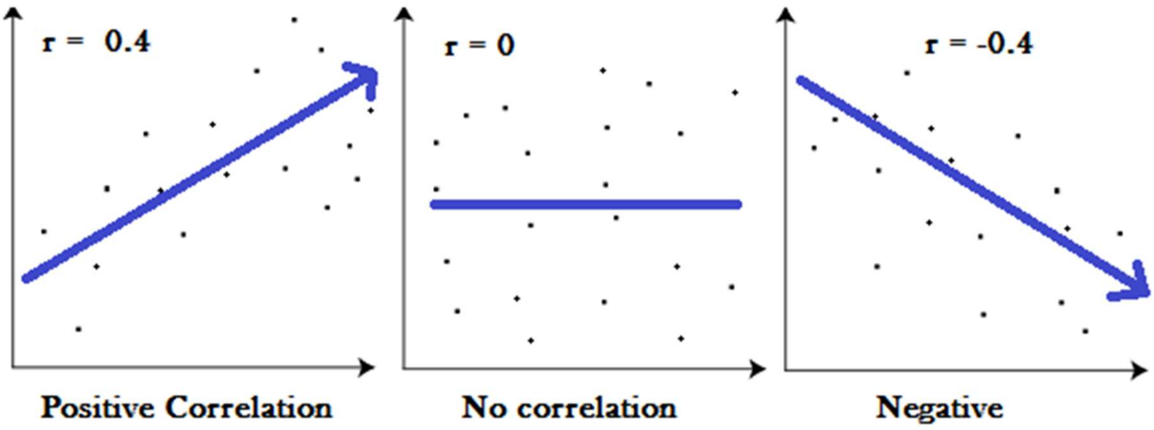
Aim- Create two random datasets and determine the Pearson's Correlation between them. Also demonstrate that Null Hypothesis will be rejected or accepted based on this correlation coefficient.

Platform- Google Colaboratory

Theory-

Correlation coefficients are used to measure how strong a relationship is between two variables. Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.

- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.

- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

The absolute value of the correlation coefficient gives us the relationship strength. The larger the number, the stronger the relationship. For example, |-.75| = .75, which has a stronger relationship than .65.

There are several types of correlation coefficient, but the most popular is Pearson's. Pearson's correlation (also called Pearson's *R*) is a correlation coefficient commonly used in linear regression. The full name is the **Pearson Product Moment Correlation (PPMC).** Mathematically, it is given as,

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Result- The result has been verified under experimental error conditions.

Experiment 03

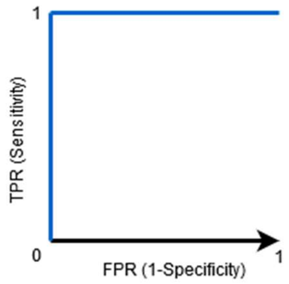
Aim- Using a dataset compare multiple ROC curves in a single plot and automatically displays the AUC for each model as well.

Platform- Google Colaboratory

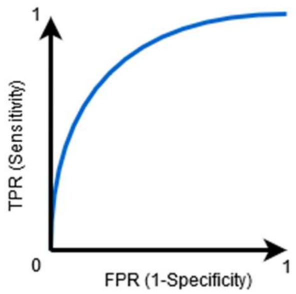
Theory-

The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

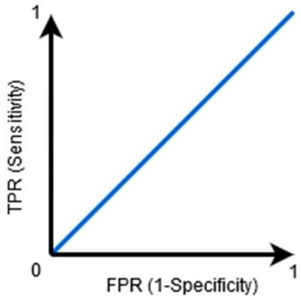
The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.



When AUC = 1, then the classifier is able to perfectly distinguish between all the Positive and the Negative class points correctly. If, however, the AUC had been 0, then the classifier would be predicting all Negatives as Positives, and all Positives as Negatives.



When 0.5<AUC<1, there is a high chance that the classifier will be able to distinguish the positive class values from the negative class values. This is so because the classifier is able to detect more numbers of True positives and True negatives than false negatives and false positives.



When AUC=0.5, then the classifier is not able to distinguish between Positive and Negative class points. Meaning either the classifier is predicting random class or constant class for all the data points.

So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes.

Result- The result has been verified under experimental error conditions.

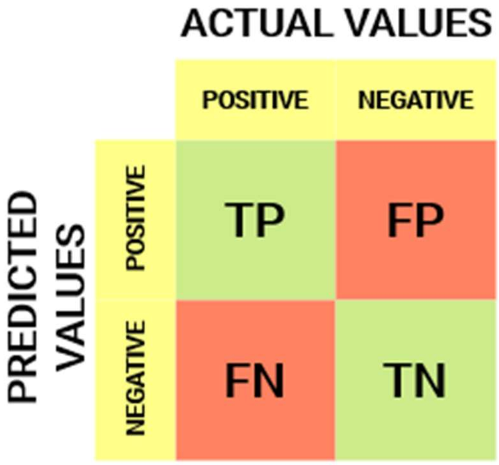
Experiment 04

Aim- Using a dataset plot confusion matrix in one line of code

Platform- Google Colaboratory

Theory-

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Sensitivity / True Positive Rate / Recall

Sensitivity tells us what proportion of the positive class got correctly classified.

$$Sensitivity = \frac{TP}{TP + FN}$$

A simple example would be to determine what proportion of the actual sick people were correctly detected by the model.

False Negative Rate

False Negative Rate (FNR) tells us what proportion of the positive class got incorrectly classified by the classifier.

$$FNR = \frac{FN}{TP + FN}$$

A higher TPR and a lower FNR is desirable since we want to correctly classify the positive class. False Negative is also known as the **Type II error**

Specificity / True Negative Rate

Specificity tells us what proportion of the negative class got correctly classified.

$$Specificity = \frac{TN}{TN + FP}$$

Taking the same example as in Sensitivity, Specificity would mean determining the proportion of healthy people who were correctly identified by the model.

False Positive Rate

FPR tells us what proportion of the negative class got incorrectly classified by the classifier.

$$FPR = \frac{FP}{TN + FP} = 1 - Specificity$$

A higher TNR and a lower FPR is desirable since we want to correctly classify the negative class. False Positive is also known as the **Type 1 error**

Result- The result has been verified under experimental error conditions.

Experiment 05

Aim- Use ColumnTransformer to apply different preprocessing to different columns:

- select from DataFrame columns by name
- passthrough or drop unspecified columns

Platform- Google Colaboratory

Theory-

It is important to prepare data prior to modeling.

This may involve replacing missing values, scaling numerical values, and one hot encoding categorical data. Data transforms can be performed using the scikit-learn library; for example, the SimpleImputer class can be used to replace missing values, the MinMaxScaler class can be used to scale numerical values, and the OneHotEncoder can be used to encode categorical variables.

The ColumnTransformer is a class in the scikit-learn Python machine learning library that allows you to selectively apply data preparation transforms.

For example, it allows you to apply a specific transform or sequence of transforms to just the numerical columns, and a separate sequence of transforms to just the categorical columns.

To use the ColumnTransformer, you must specify a list of transformers.

Each transformer is a three-element tuple that defines the name of the transformer, the transform to apply, and the column indices to apply it to. For example:

- (Name, Object, Columns)

Setting *remainder='passthrough'* will mean that all columns not specified in the list of "*transformers*" will be passed through without transformation, instead of being dropped. A ColumnTransformer can also be used in a Pipeline to selectively prepare the columns of your dataset before fitting a model on the transformed data.

Result- The result has been verified under experimental error conditions.

Experiment 06

Aim- Using a dataset visualize bar plot, histogram, and box plot and scatter plot using R.

Platform- R Studio

Theory-

Barplots: Whenever we have variables that contain categorical values, variables with limited numeric values, we can use the bar charts to present a visual chart based on those. This chart creates a bar for distinct grouping values of the variable on the X-axis and then plots their frequencies on the Y-axis.

Histogram: When you have to represent a single variable in a way that the probability distribution of that univariate data comes visible, you prefer the histogram as a graphical representation. In R, we have hist () function that does the task for us.

Boxplot: Sometimes, some situations lead you towards a conclusion that requires additional information other than the measures of central tendency (mean, median, and mode). There is a box plot visualization which helps us to get information beyond measures of central tendency associated with the data you are working on. In R, we have a function named boxplot () which comes as a part of base R.

Scatterplots: Scatterplots are important when we wanted to deal with relationships (present if any) among the two numeric variables. The scatterplots allow us a way to look at the relationship between two numeric variables and give a glimpse of what sort of relationship they both could have (negative relationship: increase in one variable shows decrease in other etc.) They are very useful in day to day life of a data scientist. To generate a scatterplot, we have the plot () function in R, which does the work for us.

Result- The result has been verified under experimental error conditions.

Experiment 07

Aim- Using a dataset implement linear regression using R and determine the p value.

Platform- R Studio

Theory-

A linear regression is a statistical model that analyses the relationship between a response variable (often called y) and one or more variables and their interactions (often called x or explanatory variables).

A linear regression's equation looks like this:

y = B0 + B1x1 + B2x2 + B3x3 +....

Where B0 is the intercept (value of y when x=0)

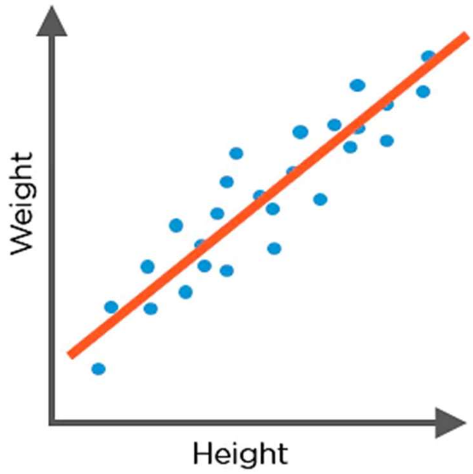
B1, B2, B3 are the slopes

And x1, x2, x3 are the independent variables

Creating a Linear Regression in R.

Not every problem can be solved with the same algorithm. In this case, linear regression assumes that there exists a linear relationship between the response variable and the explanatory variables. This means that you can fit a line between the two (or more variables). Linear regression is a form of statistical analysis that shows the relationship between two or more continuous variables. It creates a predictive model using relevant data to show trends. Analysts typically use the "least square method" to create the model. There are other methods, but the least square method is the most commonly used.

Below is a graph that depicts the relationship between the heights and weights of a sample of people. The red line is the linear regression that shows the height of a person is positively related to its weight.



The value of m & c for the best fit line, y = mx+ c can be calculated using these formulas:

$$m = \frac{((n^{*}\Sigma\,(x\ *\ y))\ \ -(\Sigma(x)^{*}\Sigma(y))}{((n^{*}\ \Sigma(x^{2}))-(\Sigma(x)^{2})}$$

$$c = \frac{((\Sigma(y)^{*}\Sigma(x^{2}))\text{-}(\Sigma(x)^{*}\Sigma(x^{*}y)}{((n^{*}\Sigma(x^{2}))\text{-}(\Sigma(x)^{2})}$$

Result- The result has been verified under experimental error conditions.

Experiment 08

Aim- Using a dataset demonstrates how to import data, perform a basic analysis, trend the results, and export the results to another text file using MATLAB.

Platform- MATLAB

Theory-

MATLAB provides functions and GUIs to perform a variety of common data-analysis tasks, such as plotting data, computing descriptive statistics, and performing linear correlation analysis, data fitting, and Fourier analysis. Typically, the first step to any data analysis is to plot the data. After examining the plot, you can determine which portions of the data to include in the analysis. You can also use the plot to evaluate if your data contains any features that might distort or confuse the analysis results, and then process your data to work only with the regions of interest.

The first step in analysing data is to import it into MATLAB. The MATLAB Programming documentation provides detailed information about supported data formats and the functions for bringing data into MATLAB. The easiest way to import data into MATLAB is to use the MATLAB Import Wizard, as described in the MATLAB Programming documentation. With the Import Wizard, you can import the following types of data sources:

- Text files, such as .txt and .dat

- MAT-files

- Spreadsheet files, such as .xls

- Graphics files, such as .gif and .jpg

- Audio and video files, such as .avi and .wav

After you import data into MATLAB, it is a good idea to plot the data so that you can explore its features. An exploratory plot of your data enables you to identify discontinuities and potential outliers, as well as the regions of interest

Result- The result has been verified under experimental error conditions.