# Data Mining and Data Visualization

# Lab Manual

**303108304**



Department of Information technology Parul Institute of Engineering and Technology

Faculty of Engineering and Technology Parul University

Session 2024-25

*Published by:*

**Faculty of Engineering and Technology Department of Information technology Parul University**

EXPERIMENT NO. 1

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to implement various data mining algorithms. |
| **Outcome:**  The students will be able to apply different data mining techniques on data sets like data pre-processing, classification and regression, clustering, association rule mining, feature selection etc. |
| **Problem Statement:**  To study WEKA tool. |
| **Background Study:**  **1) Introduction:**   * Weka is a collection of machine learning algorithms for data mining tasks. * It contains tools for regression, clustering, association, data processing, classification & visualization. * Weka results in a quicker development of machine learning models. * It provides facility to work with big data comfortably.     **2) Installation Guide:**   * Go to any browser and Search for “Weka tool download”. * There you will find many official site of Weka or. * Go to https://sourceforge.net/projects/weka/ and download .exe file. * After downloading, install the .exe file and complete installation process. * Launch Weka tool. * After installing weka tool open it and we can see interface of weka tool:   **Advantage:**   * Testing new ideas. * Easy to learn and solve data mining problems. * Different algorithms should be applied on data set to quickly see which gives the most accurate results. * Useful in reading multiple formats from multiple sources. * Independent Software that can run on most machines that run java.      * **Disadvantage:** * Some of the users are disappointed about the graphics quality * It can only handle small datasets * Alternative algorithms are not available * The user interface needs enhancement, and the designs and looks of the tool look old. * Users are complaining that it is lagging sometimes and slow in uploading * It doesn’t have a feature to change the numeric variable to a categorical variable |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What is the use of WEKA tool? 2. Which are the different data mining algorithms available in WEKA? 3. How do you integrate WEKA with other tools and languages? 4. How do you evaluate model performance? |

EXPERIMENT NO. 2

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to perform pre-processing of a database. |
| **Outcome:**  The students will be able to apply different pre-processing techniques like handling missing data, resampling of database, merging nominal values, replacing missing values etc. |
| **Problem Statement:**  Perform Pre-processing on a dataset using Weka Tool. Apply various filters and discuss the effect of each filter applied. |
| **Background Study:**  **What is data Pre-processing?**   * Data pre-processing is a crucial step in the data analysis and machine learning pipeline. It involves the cleaning and transformation of raw data into a format that can be effectively and accurately utilized for analysis or model training. * The quality of the data directly influences the performance and reliability of the final analysis or machine learning model. The main goals of data pre-processing are to address issues such as missing values, outliers, noise, and inconsistencies in the data, and to prepare it for further analysis or modeling. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What is data pre-processing? 2. What is the necessity of data pre-processing? 3. How do you handle missing values? 4. Which algorithms are necessary for this task? |

EXPERIMENT NO. 3

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to perform association rule mining on a database. |
| **Outcome:**  The students will be able to apply different association rule mining techniques like apriori algorithm, filtered associator, FPgrowth. |
| **Problem Statement:**  Perform Association rule mining using WEKA tool. |
| **Background Study:**  **Introduction:**   * Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis. * Market Based Analysis is one of the key techniques used by large relations to show associations between items. It allows retailers to identify relationships between the items that people buy together frequently. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What should be the support and confidence thresholds? 2. Which is the best algorithm for association rule mining? 3. How to make rules using apriori algorithm? |

EXPERIMENT NO. 4

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to perform classification algorithms on a database. |
| **Outcome:**  The students will be able to apply different classification techniques like J48filter, BayesNET, Logistics, PART etc. |
| **Problem Statement:**  Perform Classification with WEKA tool. |
| **Background Study:**  **Introduction:**   * The model or classifier is derived by the algorithm using the training dataset. A decision tree, mathematical formula, or neural network can all be used as the derived model.      * When unlabeled data is fed into a classification model, it should identify the class to which it belongs. you can perform classification tasks efficiently using the WEKA tool.      * Additionally, WEKA offers extensive documentation, tutorials, and online forums where you can find further guidance and support. The algorithm is given the set of input data and the corresponding outputs. There are thus a set number of options. There may occasionally be more than two classes to categorize. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What type of learning is used for classification? 2. Which are the different classification models? 3. Which is the best algorithm for classification? |

EXPERIMENT NO. 5

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to perform clustering algorithms on a database. |
| **Outcome:**  The students will be able to apply different clustering techniques like Simple Kmean filter, Density based clustering, etc. |
| **Problem Statement:**  Perform Clustering with WEKA tool. |
| **Background Study:**  **Introduction:**  Clustering the method of converting group of abstract objects into classes of similar objects. Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters. It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms. Data objects of a cluster can be considered as one group. We first partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups. The over-classification main advantage is that it is adaptable to modifications, and it help layout important characteristics that differentiate between distinct groups. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What type of learning is used for clustering? 2. Which are the different clustering models? 3. Which is the best algorithm for clustering? |

EXPERIMENT NO. 6

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of WEKA tool to perform binning method to smooth out noise from a database. |
| **Outcome:**  The students will be able to apply binning method to smooth out the noise in dataset to make it more stable and easier to handle. |
| **Problem Statement:**  Perform Binning techniques to smooth out the noise in the dataset. |
| **Background Study:**  **Binning**:   * Binning, also known as discretization or bucketing, is a data preprocessing technique used in data mining. It involves dividing a continuous variable into a set of smaller intervals or bins and replacing the original values with the corresponding bin labels. * It involves grouping a continuous variable into a smaller number of intervals, or "bins", thereby converting it into a categorical variable. * Binning in data mining can be useful in various scenarios, such as reducing the noise in the data, improving the accuracy of predictive models, and making the data easier to understand and interpret. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. Which are the different binning methods? 2. How many bins should be created? |

EXPERIMENT NO. 7

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of linear regression algorithm using python language. |
| **Outcome:**  The students will be able to apply linear regression analysis and discover the best line to fit two attributes. |
| **Problem Statement:**  Write a python program for linear regression analysis on the given dataset. |
| **Background Study:**  Theory:  Regression defines a type of supervised machine learning approaches that can be used to forecast any continuous-valued attribute. Regression provides some business organization to explore the target variable and predictor variable associations. It is an essential tool to explore the data that can be used for monetary forecasting and time series modeling.  **Linear Regression** − Linear regression includes discovering the “best” line to fit two attributes (or variables) therefore that one attribute can be used to predict the other. Multiple linear regression is an advancement of linear regression, where higher than two attributes are included and the record are fit to a multidimensional area.  For example, the equation is  Y = a + b\*X + e.  Where,  a defines the intercept  b defines the slope of the regression line  e defines the error  X and Y define the predictor and target variables, accordingly. If X is create up of higher than one variable, defined as multiple linear equations. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:** |

EXPERIMENT NO. 8

|  |
| --- |
| **Objective(s):**  To familiarize students with importing and loading a csv file and using python language. |
| **Outcome:**  The students will be able to load csv file and generate charts from it to visualize the data using python libraries. |
| **Problem Statement:**  Use python libraries to generate chart from csv data. |
| **Background Study:**  Theory:  **Pandas:** Pandas is a Python library designed for data manipulation and analysis. It provides data structures like Series and Data Frames, making it easy to work with labeled and relational data. Pandas is built on top of NumPy and is commonly used for tasks such as cleaning, merging, and analyzing datasets. It is an essential tool for data analysts, scientists, and engineers working with structured data in Python  **Matplotlib:** Matplotlib is a powerful Python library for generating high-quality static, animated, and interactive visualizations. It offers a wide range of tools through its pyplot submodule, allowing developers to easily create charts, graphs, and other graphics. Matplotlib is often compared to MATLAB due to its similarity in usage via the pyplot interface, which consists of a series of commands that modify the state of a figure, making each call build upon the previous ones. This approach simplifies the creation process while still offering flexibility when needed. Matplotlib supports numerous formats and backends, enabling it to adapt to diverse needs and environments.  **Pyplot**: Pyplot is a user-friendly interface for Matplotlib, providing a concise way to generate visualizations without requiring deep knowledge of the underlying architecture. It follows a stateful design, meaning that operations accumulate until either a new figure is opened or the plt.close() function is invoked. Pyplot offers a collection of functions that allow users to perform common tasks such as creating figures, adding plots, setting titles, and showing the resulting graphic. It is particularly useful for rapid prototyping and experimentation because it closely mimics the MATLAB environment. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:** |

EXPERIMENT NO. 9

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of different data visualization tools. |
| **Outcome:**  The students will be able to learn about the different data visualization tools used in data mining like Power BI, Tableau, Google Data Studio. |
| **Problem Statement:**  Study different Data visualization tools. |
| **Background Study:**  **Theory:**  Data and information visualization is the practice of designing and creating easy-to- communicate and easy-to-understand graphic or visual representations of a large amount of complex quantitative and qualitative data and information with the help of static, dynamic or interactive visual items. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. Which are the different data visualization tools? 2. Which is the most efficient data mining and visualization tool? 3. How do you create a dashboard using Power BI? |

EXPERIMENT NO. 10

|  |
| --- |
| **Objective(s):**  To familiarize students with the working of Power BI tool for data visualization. |
| **Outcome:**  The students will be able to do hands on working of Power BI tool by loading a dataset and creating a dashboard using different plots and graphs. |
| **Problem Statement:**  Case Study: Interactive Data Analytics with Power BI. |
| **Background Study:**  **What is Power BI?**  Power BI is a robust tool for interactive data analytics and visualization, offering a range of features that streamline the process of transforming raw data into actionable insights.  **What is Power BI used for?**  There are various tools and techniques for analytics and machine learning in the fascinating and extensive realm of data science. Power BI is a high-level, all-in-one solution for data analytics in data science. Data science aids in the discovery of relevant and productive trends and insights. It involves analyzing the data and assists us in identifying entirely new features in it. Business intelligence is sifting through data to extract meaningful organizational ideas and insights. BI enhances and strengthens the business infrastructure to get desired or projected results.  Many data sciences and analysis tasks can be automated with Power BI, eliminating the need for spreadsheets and static presentation tools. One of Power-most BI's most impressive features is its ability to create stunning visualizations. The software is packed with excellent and eye-catching visualization templates. The integration of Power BI into Data Science holds great importance for businesses. This allows for smooth and effective data visualization, which plays a vital role in an organization’s success. |

|  |
| --- |
| **Algorithm (Student Work Area):** |
| **Code (Student Work Area):** |
| **Question Bank:**   1. What is a Power BI dashboard? 2. Difference between Power BI and Tebleau? 3. How do you create a dashboard using Power BI? |