**EAS 509**

**Statistical Learning and Data Mining – 2**

**Project - 1**

**Team members:**

Harin Rishabh; UBIT: harinris

Tejaswini Kankanala; UBIT: tkankana

Vaishnavi Malalur Rajegowda; UBIT: vmalalur

**Data Ingestion**

The dataset is loaded from an Excel file, specifically from the second sheet, using the **readxl** library. This approach ensures that the analysis is based on the most relevant data provided.

1. **Importing Required Libraries**: We start by loading the **readxl** library, which is essential for reading Excel files in R. This library provides a set of functions that make it straightforward to import data from both **.xls** and **.xlsx** formats directly into R data frames, without the need for intermediate file format conversions.

2. **Specifying the File and Sheet**: The data is ingested from an Excel file named "Mine\_dataset.xls". Excel files often contain multiple sheets, each potentially holding different sets of data. We specify not only the file name but also the exact sheet to use (**sheet=2**), ensuring the correct dataset is loaded. This is important in projects where multiple datasets are maintained within the same file, possibly for organizational convenience.

**Preprocessing**

Data preprocessing is a crucial step in the data analysis and machine learning pipeline, essential for transforming raw data into a clean and usable format. This process improves the quality of data and, consequently, the accuracy and effectiveness of the analytical models that rely on it.

* **Renaming Columns**: The columns are renamed for clarity to "Voltage", "Height", "Soil\_Type", and "Landmine\_Type".

By renaming columns: You ensure that each column's purpose is immediately apparent, reducing confusion and errors during data analysis. Data manipulation becomes easier, as scripts and commands can more intuitively reference column names that accurately describe the data they hold.

* **Type Conversion**: "Landmine\_Type" is converted into a categorical factor to facilitate analysis.

Type conversion is essential for ensuring that data types align with their intended analysis functions. Converting to a factor allows the use of statistical models that handle categorical variables effectively, facilitating analyses like chi-square tests, ANOVAs, regression models, and more, which are dependent on understanding the variable type correctly.

* **Normalization Correction**: "Soil\_Type", which had been accidentally normalized, is converted back to a categorical variable with six distinct types using the cut() function.

This step is critical because: It restores the categorical nature of "Soil\_Type", ensuring that subsequent analyses treat it appropriately as a qualitative variable rather than quantitative. Correct categorization is essential for performing any statistical tests or data visualizations that depend on grouping or segmenting the data based on soil type.

**Data Cleaning and Transformation**

Handling outliers and removing duplicates are crucial tasks that ensure the integrity and usability of the dataset for analysis and modeling. Here’s a detailed look into these two specific operations:

* **Handling outliers:**

Outliers are data points that deviate significantly from other observations and can distort statistical analysis and predictive modeling. Identifying and addressing outliers is essential to improve the overall dataset quality.

**Method:** Interquartile Range (IQR)The IQR method is a robust technique used to identify outliers by focusing on the statistical distribution of the data.

Handling outliers and removing duplicates are foundational steps in data cleaning that significantly impact subsequent data analysis and modeling phases. By cleaning the dataset in this manner, you ensure that the data more accurately represents the real-world phenomena being studied, thereby improving the reliability and validity of your conclusions and predictions.

**Data Visualization**

Data visualization plays a pivotal role in data analysis by providing a visual representation of the data, which can help uncover underlying patterns, trends, and anomalies that might not be evident through raw data analysis alone.

**1. Histograms:**

Histograms are a type of bar chart that represents the frequency distribution of numerical data by dividing the data range into bins or intervals. Each bar in a histogram represents the tabulated frequency at each interval/bin.

**Uses and Benefits**: Understanding Distribution: Histograms provide insights into the underlying distribution of data—whether it is normal, skewed, or bimodal, etc.

**Identifying Outliers**: Large gaps or an unusually long tail in a histogram might suggest outliers. **Evaluating Modality**: The presence of multiple peaks can indicate multiple groups or modes in the data.

A screenshot of a graph

Description automatically generated

These histograms depict the frequency distributions of two variables: "Height" appears uniformly distributed across its range, while "Voltage" is left-skewed with a higher concentration of values in the lower range.

**2.Boxplots:**

Boxplots summarizes the minimum, first quartile (Q1), median (second quartile, Q2), third quartile (Q3), and maximum.

**Uses and Benefits:**

**Visualizing Spread and Central Tendency**: Boxplots show the median, quartiles, and extremes at a glance, providing a concise statistical summary of the distribution.

**Detecting Outliers:** Points that appear outside the whiskers (typically 1.5 times the IQR from the quartiles) are often considered outliers. Comparing Distributions: They are especially useful for comparing distributions across groups.

A diagram of a box plot

Description automatically generatedA diagram of a box plot

Description automatically generated

The "Height" boxplot shows a relatively symmetrical distribution with no visible outliers, while the "Voltage" boxplot indicates a right-skewed distribution with several outliers above the upper whisker.

A graph showing different colored rectangular shapes

Description automatically generated

The boxplot illustrates the distribution of "Height" across six different "Soil\_Type" categories, each color representing a unique type, indicating variability in height measurements within and across the soil types.

**3.Bar Plots**:

Bar plots are used to display the frequency or proportion of categorical data, with each bar representing a category.

**Uses and Benefits**:

**Comparative Analysis**: Easily compare different categories to see which are most or least common.

**Clarity in Categorical Data Representation**: Provides a clear visualization of how various categories stack up against each other.

A graph of a number of soil types

Description automatically generated

The bar chart displays the distribution of six soil types, showing a relatively even distribution across the categories with no significant variations in frequency.

**4. Scatter Plots:**

Scatter plots display values for typically two variables for a set of data, with the data points plotted on a two-dimensional graph. When enhanced by ggplot2 in R, scatter plots can include additional variables through aspects like color, shape, and size of the data points, and even grouping (faceting).

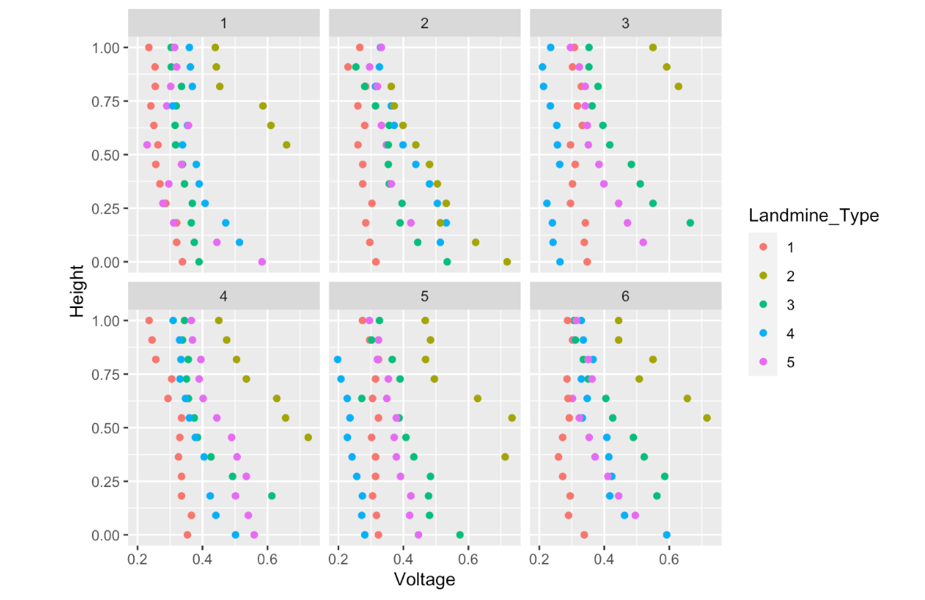
**Uses and Benefits**:

**Relationship Assessment**: Scatter plots are ideal for observing the relationships or associations between two continuous variables.

**Identifying Correlations**: Trends in the scatter plot can indicate correlation types and strengths between variables.

**Highlighting Clusters**: Natural clustering of data points can be visually identified.

**Faceted Plotting with ggplot2**: Using ggplot2 for faceted scatter plots allows for the exploration of how relationships between variables change across different subgroups within the data, handled by facet\_wrap or facet\_grid functions.



The scatter plot matrix shows the relationship between "Voltage" and "Height" for different "Soil\_Type" categories, colored by "Landmine\_Type," indicating varied distributions and potential patterns within each soil type.

**Machine Learning Models Implementation**

**1.Naive Bayes:**

Naive Bayes classifiers are probabilistic models based on applying Bayes' theorem with strong independence assumptions between the features. They are particularly known for their simplicity and effectiveness in large datasets.

A screenshot of a computer

Description automatically generated

**2. Support Vector Machines (SVM):**

SVMs are a set of supervised learning methods used for classification, regression, and outliers detection. They are effective in high-dimensional spaces and in situations where the number of dimensions exceeds the number of samples.

A screenshot of a computer

Description automatically generated

**3. K-Nearest Neighbors (KNN):**

KNN algorithm is a non-parametric method used for classification and regression. A data point is classified by a majority vote of its neighbors, with the data point being assigned to the class most common among its k nearest neighbors.

A table of numbers with black text

Description automatically generated

A screenshot of a computer

Description automatically generated

**4.Logistic Regression:**

logistic regression is used for binary classification problems. It estimates the probability that a given instance belongs to a particular class. Its simplicity lies in interpretability and ability to handle leaner relationships well.

A screenshot of a computer

Description automatically generated

**5.Decision Trees:**

Decision trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

A screenshot of a computer

Description automatically generated

**6. Random Forests:**

Random forests are an ensemble learning method for classification, regression, and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.

A screenshot of a computer

Description automatically generated

**Performance Evaluation:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Model** | **Naive Bayes**  **Classifier** | **Support vector classifier** | **K-nearest neighbor**  **(KNN)** | **Logistic regression** | **Decision trees** | **Random Forests** |
| **accuracy** | **0.49206** | **0.38095** | **0.26984** | **0.46031** | **0.42857** | **0.47619** |

Implementing a variety of machine learning models is a best practice in data science as it allows for comparison across different approaches, leading to a more robust selection of the best model for a given problem based on the performance metrics.

**Results:**

The dataset consisted of 4 columns. We had 2 continuous value columns, height and voltage measurement.

Soil type was a categorical variable that consisted of 6 types. The target variable was mine type. We ran several classification algorithms to classify the mines into the predetermined labels.

The data had no missing values. The height column had some outliers which were removed to improve accuracy. The data was split into test and train samples.

Overall, the data was clean but too limited to achieve reasonable accuracy. After removing outliers we had around 300 observations which is too less for 3 features of which one was categorical. All our models give a max accuracy of around 45%.

**References :**

Data Source - <https://archive.ics.uci.edu/dataset/763/land+mines-1>