National University of Computer &
Emerging Sciences Peshawar Campus

AL 2002 – Artificial Intelligence Lab Project
Spring 2023

Deadline: April 25th,2023 **Total Marks: 40**

Note: Lab policy about plagiarism is as follows:

- Students must not share actual program code with other students.
- Students must be prepared to explain any program code they submit. (Viva)
- Students must indicate with their submission any assistance received.
- All submissions are subject to plagiarism detection. For report and code maximum 15% plagiarism is acceptable.
- Students cannot copy code from the Internet.

1 Project Description:

1.1 Objectives:

The objective of the project is to apply the different classification and clustering algorithms to the problem of classifying cyber-attacks in network traffic. This will help your retention of the material and significantly enhance the depth of your understanding. This will also develop your skills in reporting the performance. We encourage you to work in a group of (maximum) two students.

1.2 Dataset Description:

The dataset is available in the project folder. It comprises of two text files, the “Dataset.txt” file contains the complete dataset. Each column specifies the attributes of network traffic, out of which some or all may be considered as features for classification. The other file “Attack_types.txt” summarizes the possible attack types.

1.3 Scope of Work

For the given dataset, we want to develop classifiers for the prediction of attack type, given their attributes.

The scope of the project includes

- Formulation of the problem under consideration.
- Cleaning and pre-processing the data.
- Apply feature engineering (if needed).
- Implement the classification and clustering algorithm. (Details below)
- Report the performance of the algorithms and presentation of analysis/findings.

2 Expectations:

There are two components of assessment in the project:

- Report (20 marks)
- Code (20 marks)

Your report is expected to have the following sections and should be around 600 to 700 words.

1) Abstract (executive summary),

2) Introduction,

3) Data-preprocessing (extraction and cleaning),

4) Feature Engineering e.g., dimensionality reduction (if needed),

5) Use of the given classification and clustering algorithms

6) Comparison and Performance Evaluation (plots, tables etc.) Detailed comparison is expected for task 5 and 6.

7) Conclusions.

Implementation Details

For Implementation, you have to perform the following tasks:

1. Explore the dataset to understand the characteristics and distributions of different features.
2. Perform data preprocessing on the dataset by cleaning the data, handling missing values, outliers, and feature scaling.

3. The dataset provided to you contains 23 different classes (attack types). You need to convert it to 5-classes. (Hint: See the file "Attack_types.txt")
4. Identify the most relevant features for classification using technique such as correlation analysis.
5. The next step is to use the updated dataset for classification. Use the following algorithm to perform classification.

    a. Classification of Cyber Attacks Using Decision Tree Algorithm

    A decision tree algorithm will be used to develop a classification model. The model will be trained on the preprocessed dataset and its performance will be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score.

    b. Classification of Cyber Attacks Using K-Nearest Neighbors Algorithm

    The KNN algorithm will be implemented and trained using the preprocessed and selected features**.** Determine optimal value of k.

    c. Classification of Cyber Attacks Using Artificial Neural Networks (ANN)

    Choose an appropriate ANN model that suits the specific requirements of the task. Train the selected model on the preprocessed dataset to learn from the provided data. Evaluate the performance of the trained model using appropriate evaluation metrics such as accuracy, precision, recall, and F1-score. Optimize the performance of the ANN model by tuning its Hyperparameters such as the learning rate, and number of hidden layers.

6. Your last task includes classification followed by clustering. In this task, you will drop the column containing labels from the dataset and label the dataset using any clustering algorithm e.g., k-Means. Clustering will return label for each record. Visualize the results of the clustering using a scatter plot.

Rubrics for Assessment of Project

Component	Description	Marks
Project Report	Abstract, Introduction, Problem Formulation, Objectives, Dataset description, Data pre-processing, Classification and Clustering Algorithms, Comparison and Performance Evaluation, Conclusions	20
Data Pre-processing	Thorough cleaning and preparation of data	4
Exploratory Data Analysis and Visualization	Thorough and effective exploration of the dataset using appropriate visualization techniques	3
Feature Engineering	Effective use of dimensionality reduction techniques, if necessary	1
Ensemble Learning	Successful implementation classification algorithm(s)	6
Clustering Algorithms	Successful implementation and performance of K-Means/any other clustering algorithm	2
Comparison and Performance Evaluation	Detailed comparison of algorithms with appropriate plots, tables, etc.	4