CSCI-B 365

Introduction to Data Analysis and Mining

Fall 2024

PROJECT

Due: 11.59 PM Sunday, November 26th, 2024

DESCRIPTION

In this project, you will select a dataset from a well-known source, conduct a comprehensive analysis, and apply various data mining techniques. You will analyze the data, describe its attributes, and implement methods such as decision trees, rule-based classification, k-nearest neighbors (KNN), naive Bayes, artificial neural networks (ANN), support vector machines (SVM), and ensemble methods. Finally, you will interpret and comment on the results of your analysis.

PURPOSE

The purpose of this project is to provide hands-on experience with data mining techniques, enhance your ability to analyze and interpret data, develop skills in Python programming for data analysis, and foster critical thinking regarding model selection and evaluation.

REQUIREMENTS

1. **Dataset Acquisition**

You will begin by choosing a dataset from recognized sources, such as the UCI Machine Learning Repository, Kaggle, or government open data portals. You should provide a brief overview of the dataset, including the number of instances and features.

2. **Data Analysis**

You will perform exploratory data analysis (EDA) to understand the dataset's structure and describe the attributes, including data types, summary statistics, and potential issues like missing values or outliers.

3. **Data Preprocessing**

Next, you will need to clean the data as necessary, addressing missing values and normalizing or standardizing features. After preprocessing, you should split the dataset into training and testing sets.

4. **Implementation**

You are required to implement one of the data mining techniques from the following list: Decision Trees, Rule-Based Classification, K-Nearest Neighbors (KNN), Naive Bayes, Artificial Neural Networks (ANN), Support Vector Machines (SVM), and Ensemble Methods (e.g., Random Forest, Boosting). Comparison of the selected method with another method could get up to 20% bonus.

QUESTIONS THAT NEED TO BE ANSWERED

- What challenges did you face during data cleaning and preprocessing?
- How did you choose which models to implement?
- What were the key factors that influenced the performance of your models?
- How did you interpret the results of your chosen models?
- What insights can you draw from the data and your analysis?

NOTES

You are expected to prepare a one-page presentation answering the questions as well as presenting the results. You are highly encouraged to use Python and relevant libraries (e.g., pandas, NumPy, scikit-learn, Matplotlib, seaborn).