CIS 635 Knowledge Discovery and Data Mining

Final Project Progress Report

Data Alliance

TEAM MEMBERS:

- VAISHNAVI RASANE
- ROHITH ANUGOLU
- SANJANA DEVARUPALA
- ROHITA JAHNAVI JALA

PROGRESS SO FAR:

Completed Tasks:

1. **Data Loading**: Successfully loaded the streamflow data set into a Pandas data frame, providing a structured format for further analysis.
2. **Missing Values Analysis:** Identified the number of missing values in each column, laying the foundation for effective data imputation strategies.
3. **Exploratory Data Analysis (EDA)**: Conducted EDA to visualize the streamflow dataset, gaining insights into its distribution, trends, and potential outliers.
4. **Descriptive Statistics:** Identified and printed essential statistics for all the columns, providing a baseline understanding of the dataset.
5. **Data Preprocessing:** Implemented data preprocessing operations, including attribute classification, handling missing values, and removing duplicate data, ensuring data integrity.

Data:

We're using the streamflow dataset for this project.

https://yong-zhuang.github.io/gvsu-cis635/_downloads/ac180a42f06404d9ccbdcd704750ff8e/streamflow.csv

CHALLENGES:

We encountered quite some challenges during the implementation phase:

1. **Missing Values Handling**: Addressing missing values posed a challenge, and needed a careful evaluation of various imputation methods to ensure accuracy without introducing bias.

2. **Model Selection:** Selecting a suitable classification model for the dataset required a comprehensive review of available models to align with the project's objectives.

COLLABORATION:

We had weekly meetings every Monday before class from 5:00 PM to 5:45 PM. During these meetings, we discussed the project scope, identified the goals and objectives of the project, and divided the tasks among the team members. In every meeting, we also set goals for the upcoming meeting.

We're working together as a team and we assist each other with the code debugging and other project-related tasks.

NEXT STEPS:

1. Make the time series **Stationary**.

2. **Dataset Splitting:** Splitting the dataset into training and testing sets to facilitate model training and evaluation.

3. **Classification Model Training**: Applying different classification algorithms to the dataset, aiming to assess their performance through accuracy scores.

4. **Evaluation**: Systematically evaluating the performance of each model to determine the most suitable classification algorithm for our streamflow dataset.

This next phase will form the basis for subsequent analysis and optimization, moving us closer to achieving our project objectives.