CIS 635 Knowledge Discovery and Data Mining

Final Project Report for Streamflow Data Analysis

Data Alliance

1. Introduction

This project focuses on the application of data mining techniques and time series analysis to predict streamflow patterns, a critical component in hydrological studies. The goal is to achieve an in-depth understanding of the streamflow dataset and employ predictive models for accurate forecasting. In this project, we evaluated an ARIMA model performance for streamflow time-series data and also calculated the RMSE score. The objective is to apply different classification models to assess their performances through accuracy scores and evaluate the accuracy of each model to determine the most suitable classification algorithm for our streamflow dataset utilizing a comprehensive streamflow dataset starting from 1967, we applied machine learning models and time series forecasting methods to predict future flow rates. This project aims to extract valuable insights and enhance forecasting accuracy. The methodology encompasses data loading, missing values handling, exploratory data analysis (EDA), and the application of various classification models.

2. Related Work

Prior research has shown various approaches in streamflow prediction, ranging from hydrological modeling to machine learning and time series analysis. Our project extends these studies, employing advanced techniques to improve prediction accuracy.

3. Methods

3.1 Our methodology involved a step-by-step approach to ensure a thorough examination of the streamflow dataset. We loaded the data into a Pandas data frame, addressed missing values through careful imputation, and conducted exploratory data analysis (EDA) to visualize patterns and trends. The data was preprocessed to guarantee its integrity, and the time series was made stationary for effective modeling.

For classification tasks, we split the dataset into training and testing sets and applied a range of algorithms, including KNN, Decision Trees, Random Forest, and AdaBoost. Accuracy scores were systematically assessed to evaluate the performance of each model.

3.1.1 Data Collection: Our dataset comprises daily streamflow records, including the year, month, day, and flow rate (Q in m^3/s). The dataset is available at [link to dataset](https://yong-zhuang.github.io/gvsu-cis635/_downloads/ac180a42f06404d9ccbdcd704750ff8e/streamflow.csv).

3.1.2 Data Loading: We loaded the streamflow data into a Pandas DataFrame.

3.1.3 Missing Values Analysis: We first identified the missing values in the dataset and handled those by using the imputer method

2.1.4 Exploratory Data Analysis (EDA): We performed different exploratory data analysis operations to visualize and understand the dataset. We plotted different graphs such as temporal patterns and trend analysis of streamflow, average monthly streamflow rates.

3.1.5 Descriptive Statistics: We computed statistics of the dataset such as min, max, mean, count, and standard deviation to understand data distribution.

3.1.6 Data Preprocessing: We did preprocessing of the dataset including attribute classification and duplicate removal.

3.1.7 Time Series Stationarity: The time series dataset was made stationary to eliminate trends and seasonality, which is vital for accurate forecasting. We validated it by performing a Dicky-fuller test.

3.1.8 Trend Estimation and Elimination: Analysing and removing trends from the time series data.

3.1.9 Classification Models: We divided the dataset into training and testing and applied different classification models such as KNN Classifier, Decision Tree Classifier, AdaBoost Classifier, and random forest classifier. We calculated the accuracy of each model.

3.1.10 Forecasting: Lastly, we employed time series forecasting methods to predict future streamflow.

4. Results and Discussion

Through EDA, preprocessing, and time series analysis, we gained valuable insights into the streamflow data's behavior and underlying patterns. Implementing time series stationarity and trend elimination techniques prepared the data for effective forecasting. The preliminary results from our forecasting models show promising directions for accurately predicting future streamflow.

The culmination of our streamflow data analysis project brings forth insightful results, showcasing the performance of various classification models and the inclusion of time series forecasting with the ARIMA model. The inclusion of ARIMA enriches our analysis by offering a time-dependent forecasting perspective, complementing the classification models that operate on a categorical outcome.

Classification Model Performance

4.1 KNN Classifier: 29.13% Accuracy

4.2 Decision Tree Classifier: 34.83% Accuracy

4.3 Random Forest Classifier: 34.53% Accuracy

4.4 Adaboost Classifier: 10.21% Accuracy

These accuracy scores provide a comparative understanding of the strengths and weaknesses of each classification model. Decision Tree and Random Forest models demonstrate robust performance, while Adaboost exhibits lower accuracy.

In addition to classification models, we employed the Autoregressive Integrated Moving Average (ARIMA) model for time series forecasting. The Root Mean Square Error (RMSE) score for the ARIMA model is 17888.

The diversity in model performances highlights the nuanced nature of streamflow prediction. While classification models provide accuracy in categorical outcomes, the ARIMA model contributes by capturing temporal dependencies and trends in the time series data. The coexistence of classification and time series forecasting methods broadens the applicability of our analysis.

5. Conclusion

The project has successfully integrated data mining techniques with time series analysis to develop a foundation for accurate streamflow prediction. In conclusion, this project provides a comprehensive exploration of streamflow data, from data loading to forecasting. The accuracy scores showcase the efficacy of various classification models in predicting streamflow outcomes. Insights gained contribute not only to hydrological understanding but also to informed decision-making in water resource management.

6. Data and Software Availability

The data and tools used in this project are Python 3.7 Google Colab Notebook. It is available under the Data Alliance Git repository.

7. References

1. Shang, S., Tang, C., & Zhao, L. (2014). Streamflow prediction with small sample data: A comparison of relevance vector machine and support vector machine. Water Resources Management, 28(8), 2227-2246.
2. Shabri, A., & Samsudin, R. (2006). Rainfall–runoff modeling using the hybrid of wavelet transform and neural network. Journal of Hydrology, 329(3-4), 356-367.
3. Shrestha, P., Sulis, M., & Taormina, R. (2020). Streamflow forecasting using hybrid machine learning and deep learning models: A review. Water, 12(1), 32.