Data Mining Project Report

Team Members:

- Amna Aiman (ID: i212743)

- Noor ul Huda (ID: i211357)

Project Overview

Our project focused on applying various data mining techniques to analyze and forecast time series data. We selected three datasets for our study: stock prices, energy consumption, and CO2 emissions. The project involved extensive preprocessing, model application, evaluation, and visualization, along with the development of a web-based frontend for displaying results.

Preprocessing

Before applying the models, we performed necessary preprocessing steps, including:

1. Handling Missing Values: Missing values in the datasets were filled using the mean of the respective columns.

2. Data Visualization: We visualized the original datasets to understand the underlying patterns and trends.

Models Applied

We applied eight different models to each dataset. Below is a brief explanation of each model:

1. ARIMA (AutoRegressive Integrated Moving Average):

- ARIMA is a popular statistical method for time series forecasting that combines autoregression, differencing, and moving average components.

2. ANN (Artificial Neural Network):

- ANN is a computational model inspired by the way biological neural networks work, useful for capturing complex patterns in time series data.

3. SARIMA (Seasonal ARIMA):

- An extension of ARIMA that includes seasonal components, making it suitable for data with seasonal patterns.

4. Exponential Smoothing:

- This model uses weighted averages of past observations to forecast future values, giving more weight to recent observations.

5. Prophet:

- Developed by Facebook, Prophet is an additive model that handles seasonality and holiday effects well, designed for business time series forecasts.

6. Support Vector Regression (SVR):

- SVR is a type of Support Vector Machine that supports linear and non-linear regression by transforming the data using kernel functions.

7. Long Short-Term Memory (LSTM):

- A type of recurrent neural network (RNN) that can learn long-term dependencies in sequential data, making it effective for time series forecasting.

8. Hybrid Models Integration:

- This approach combines multiple models to leverage their individual strengths, aiming for more accurate and robust forecasts.

Application and Evaluation

For each dataset, we split the work equally and applied the aforementioned models. We visualized the actual and forecasted values for each model to compare their performance. The evaluation metrics used included Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

Results

- Stocks Dataset:

- Each model provided a forecast for stock prices, and we compared their performance using MSE and RMSE.

- Energy Consumption Dataset:

  - Similar steps were followed for energy consumption data, evaluating how well each model predicted future energy usage.

  - CO2 Emissions Dataset:

  - For CO2 emissions, the models' forecasts were visualized and compared, highlighting the models that best captured the emission trends.

Frontend Development

We developed a frontend using HTML to display the visualizations and results. The web interface allows users to upload datasets, select models, and view forecasts along with the evaluation metrics.

Work Division

- Amna Aiman (i212743):

  - Focused on preprocessing, applying and evaluating ARIMA, ANN, and SVR, SARIMA, Exponential Smoothing, Prophet, LSTM, and Hybrid models.

- Noor ul Huda (i211357):

  - Handled preprocessing, applied and evaluated SARIMA, Exponential Smoothing, Prophet, LSTM, and Hybrid models.

- Developed initial versions of the frontend and integrated the visualizations.

Conclusion

This project provided valuable insights into the effectiveness of different time series forecasting models across various datasets. By combining statistical, machine learning, and hybrid approaches, we were able to identify the strengths and weaknesses of each model. The developed frontend serves as a useful tool for future analysis and model comparison.