SSE 2024 Data Analysis and Mining

Final Project Details

【重要】务必加入 qq 群：784831303。

Time flies and the semester is quickly coming to an end. This term's course projects were divided into two topics and students worked in a group to choose a topic to complete. The dataset used in the two topics is the same and can be found in the paper Chinese diabetes datasets for data-driven machine learning or directly download at here. The illustrative statements about the dataset can be found in Chinese diabetes datasets.

Topic 1: Blood Glucose Level Time Series Prediction

- **Domain Knowledge Understanding**: Gain in-depth understanding of glucose regulation mechanisms, factors influencing blood glucose levels (such as diet, exercise, medication, etc.), and the types and management of diabetes.
- **Data Collection & Preprocessing**: Gather relevant datasets, including personal health records, lifestyle habits, dietary diaries, etc. Perform data cleaning, handle missing and outlier values, and conduct necessary data transformations and normalization.
- **Feature Engineering**: Based on domain knowledge, select or construct features critical to glucose prediction. Explore correlations among features, perform feature selection or dimensionality reduction to enhance model efficiency.
- **Model Selection & Implementation**: Investigate and compare machine learning models suitable for time series glucose prediction. The choice of model can be based on methods from the latest papers or on classical and efficient methods. The direction of methods chosen is unlimited and our aim is to implement the chosen model and fine-tune parameters to optimize prediction performance.
- **Model Evaluation & Validation**: Assess the model's predictive accuracy and generalization capacity using techniques like cross-validation, AUC-ROC curves and mean squared error (MSE).
- **Results Analysis & Visualization**: Conduct a deep analysis of prediction outcomes, examining how different factors impact the accuracy of glucose predictions. Present prediction results and model performance clearly using graphs and visualization tools.

[Notes] You are required to predict the blood glucose level at 15, 30, 45 and 60 min. The project will be scored mainly according to the error value (e.g., MAE) of the prediction model. Of course, you can perform an evaluation analysis on the model by other means. Write down the analysis in the document. Part of the scores come from the documentation. It is also recommended to think more creative method, e.g., it is known that the dietary information will help on the accuracy of prediction model. Meanwhile, it is also possible to consider if transferring models from the OhioT1DM dataset.

Topic 2: Support Decision / Prescription Recommendation

- **Data Exploration & Preprocessing**: Conduct thorough exploratory data analysis on the dataset, encompassing data cleaning, handling of missing values, and anomaly detection to ensure data integrity.
- **Feature Engineering**: Based on medical expertise, select and engineer features that are indicative of complication risks and treatment effectiveness, potentially involving blood glucose levels, HbA1c, BMI, age, gender, treatment histories, etc.
- **Personalized Treatment Strategies**: Design algorithms to recommend the most suitable treatment plans tailored to individual patient differences, including medication choices, dosage adjustments, and lifestyle modifications.
- **Prescription Recommendation System**: Develop a smart prescription recommendation module that considers the patient's current condition, drug interactions, contraindications, and more, to provide safe and efficacious prescription suggestions.
- **Result Validation & Optimization**: Evaluate the system's performance using techniques like cross-validation and AUC-ROC analysis, iteratively refining model parameters and system functionalities based on feedback.

Submission Requirement

Submission date: 2024-6-14 23:59. Please submit your project details to Canvas.

The submission file is named as groupID_name_project.zip, where the name should be in Chinese. The zip file should include:1) a subdirectory named code, 2) a report document named report.pdf, 3) a member.txt file and 4) your project slides name groupID_slides.

- **The code subdirectory** includes the following contents: 1) Source code files, 2) The README file, which introduces the running environment and running steps.
- **The report document** should include a detailed process of data preprocessing, and difficulties in the implementation of the final project, experimental results, and a comparative analysis of the experimental results.
- **The member file** should list the student IDs and Chinese names of all group members on "members.txt" in main directory, and IDs should be in ascending order (the groupings are the same as for workshop). In addition, there should also be a description of what each person does.
- **The slides** should list the student IDs and Chinese names of all group members on the first page, which should be in ascending order. In addition, the person making the presentation should be identified.

Presentation

Presentation date: 2023-6-14 23:59

Presentation Time per Group: 10 min for presenting and 2 min for raising questions.