Parkinson's Disease Progression Prediction

Authors: Anirudh Gupta, Devansh Singh

Summary: The dataset aims to use protein abundance data to predict the course of Parkinson's disease (PD). The dataset contains mass spectrometry data at the peptide and protein level, clinical data of patients with Parkinson's disease, and supplemental clinical data. The protein data is derived from cerebrospinal fluid (CSF) samples of several hundred patients who contributed multiple samples over multiple years while they also took assessments of PD severity.

This dataset is designed to help predict the progression of Parkinson's disease using protein abundance data from cerebrospinal fluid (CSF) samples. The dataset includes mass spectrometry measurements of peptides and proteins, clinical data of patients with Parkinson's disease, and supplemental clinical data. The peptide and protein measurements are taken from CSF samples of several hundred patients who were assessed for PD severity over multiple years. The data could provide valuable insights into the proteins involved in PD and help advance PD research.

We plan to use data preprocessing, feature engineering and supervised learning algorithms such as Random Forest, Gradient Boosting, and Neural Networks. The performance of the models is evaluated using metrics such as R2-score, RMSE, accuracy, F1 score, and AUC-ROC.

Proposed Plan / Research Plan: The dataset contains time series data that will help us predict the course of Parkinson's disease. Being multivariate data, we will first check to see if the data is stationary or not, if not we'll have to make the data static (making mean and variance constant). We'll use the process of differencing to make the data stationary. Further, we'll decompose the data into 4 components: a) Seasonality b) Trend c) Cycle d) Errors. We'll implement different ARIMA and SARIMA models based on different p,d, and q (hyperparameters) values. Our goal would be to select the model which has the most optimal metrics values.

Preliminary results: In the exploratory data analysis of the dataset, it was found that there are 248 patients and 1,113 visit IDs, indicating that each patient visited the hospital an average of 4.48 times. On average, each patient has 4 to 5 records in the dataset. It was observed that there are multiple peptides associated with a single protein, and therefore, it is recommended to merge the peptides dataset with the protein dataset. Additionally, the dataset consists of 227 unique UniProt IDs and 968 different types of peptides.

Research Complexity: The research complexity of the proposed plan depends on the specific implementation details and the complexity of the chosen models. Using time series analysis techniques such as ARIMA and SARIMA models to analyze longitudinal data can be challenging and requires a strong understanding of statistical concepts and programming skills. Additionally, feature engineering and model

selection can also add to the complexity of the research. The size of the dataset and the complexity of the protein abundance data may also pose challenges in terms of data preprocessing and model training. Overall, the proposed plan involves advanced statistical analysis techniques and may be considered moderately complex.

References:

https://www.kaggle.com/competitions/amp-parkinsons-disease-progression-prediction/data