UNIVERSITI MALAYA

Warfarin Dosage Predictor

Introduction

Preface

- Warfarin is the most widely used anticoagulant drug used for treating arterial and venous thromboembolism caused by blood clots.

Problem Statement

- Understanding the variability in Warfarin Dosages using Machine Learning (ML) techniques.
- Develop a ML model that can help predict Warfarin dosing

Importance

- Warfarin has a narrow therapeutic range and severe side effects at extreme concentrations.
- Precise dosing of warfarin is an important concern for clinicians

The Data

Source

- Source: IWPC - International Warfarin Pharmacogenetics Consortium
- Host Institution: PharmGKB

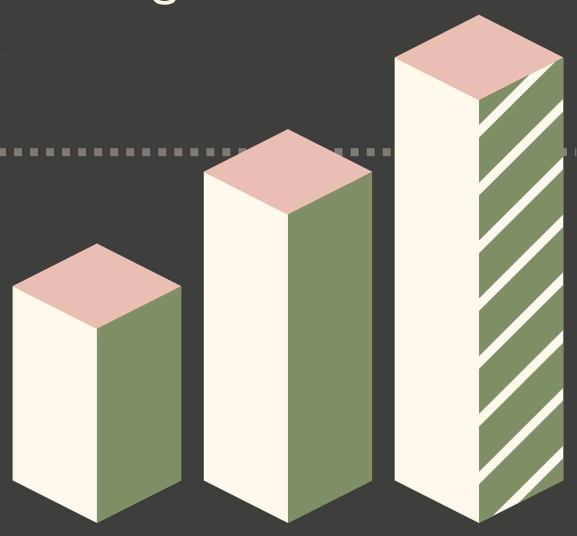
First Impressions

- Small sample size (6256 * 68)
- With missing data
- Mixture of Datatypes
- Useful features with clear target variable.

Methodology

1. Data Preprocessing techniques

- Feature Engineering - Encoding
  - Transform categorical features into numerical data
- Data Grouping
  - Organizing age and race into groups
- Data Filtering
  - Filter rare or extreme data points
- Data Imputation
  - Missing data points are imputed using Linear Regression / if-else conditions

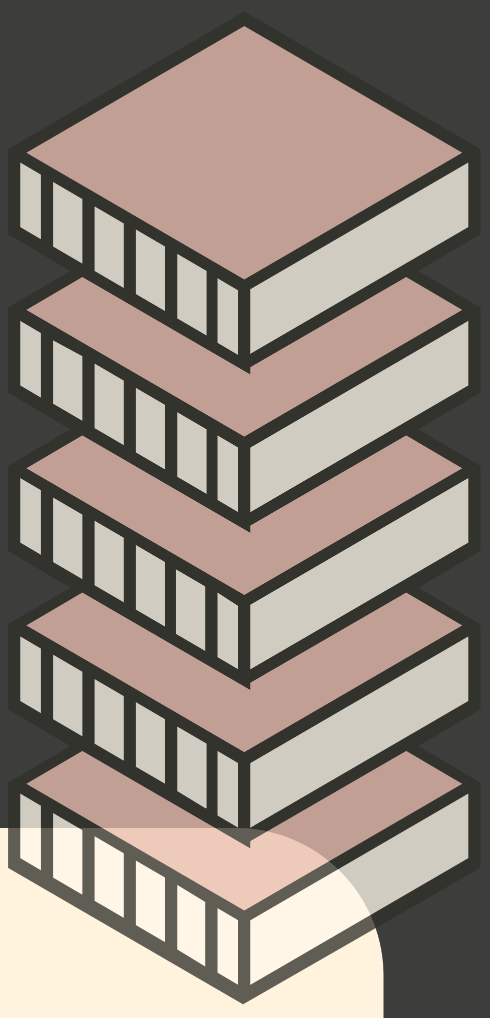


2. Data Analysis

- Analysis Methods
  - Bar graphs
  - Scatter plots
  - Descriptive statistical calculations
- Analysis Findings
  - Relationship between features and target variable
  - Comparison between feature groups
  - Highlight anomalies

3. Machine Learning Models

- Models
  - Linear Regression
  - Gradient Boosting Regressor
  - Linear Support Vector Regressor
  - Ridge Regressor
  - Support Vector Regressor
  - Multilayer Perceptron Regressor (Neural Network)
- Techniques
  - Stacking Ensemble
  - Resampling
  - Pipelining
- Performance Evaluation Metrics
  - Mean Absolute Error (MAE)
  - R-squared (R2)



Conclusion, Learning Outcome, Discussion

- Using imputation technique to estimate missing data helped maintain an adequate data sample size
- Basic Machine Learning (ML) models are able to perform well with resampling
  - Resampling helps to provide more accurate and robust results
- Stack Ensemble techniques and Pipelining ML Models are able to perform well without resampling
  - Prediction performance improved by leveraging the strengths of multiple models and capturing diverse perspectives on the data.
- With an R2 value ranging around 0.4 to 0.5, the models can be described to be moderately predictive and may be useful for exploratory analysis with room for improvement.

By: Low Ee Fei (s2149323)