Finial Project

Warfarin Dose Estimation

Group 10

Josna Pitta

Sai Rudh Reddy Velapati

Saint Louis University

Prof. Jie Hou, PhD

May 14, 2023

Abstract:

The experiment is performed to find the dose of warfarin medication. The IWPC dataset is used to perform the experiment. The experimental flow chart describes the flow of the experiment conducted. A description of the dataset is provided. Various data preprocessing steps are performed to clean the raw data. Used simple imputation technique to handle the missing values. Label Encoder is used to convert the textual data to numeric values performed normalization. Different Machine learning models are used for training. ML models used are Linear, Lasso, Decision Tree, SVR, and MLP regressor. Performed cross-validation to find the optimum values of the hyperparameters used for model building. The comparison table representing all evaluation matrices is provided for all the datasets for all models. Discussed the future extensions and possibilities to resolve the current issues of the experiment conducted.

Introduction:

Warfarin is a medication that has been used for several decades as an anticoagulant to prevent and treat thromboembolism in patients with various medical conditions. Thromboembolism is a condition where blood clots form in the blood vessels, obstructing the flow of blood and potentially leading to serious health complications. Despite its widespread use and low cost, determining the optimal dose of warfarin can be challenging for clinicians due to its narrow therapeutic index which refers to the small difference between the effective and toxic doses of a medication, and also has high variability among patients in dose requirements. Patients starting warfarin therapy are at an increased risk of overdosing during the initial weeks of therapy, which can render them susceptible to thromboembolism or increase the risk of bleeding.

Traditionally, clinicians use trial-and-error dosing procedures to determine the therapeutic dose for patients receiving warfarin and require frequent follow-up visits to adjust the dose. However, this approach can lead to increased utilization of healthcare services and potential harm to patients. Therefore, researchers have been investigating the relationship between warfarin dose requirements and critical factors to improve dosing accuracy.

In this project, we aim to study the warfarin dose-response relationship using machine learning techniques to determine the therapeutic warfarin dose during the initiation period. By utilizing machine learning algorithms, we hope to develop a more accurate and efficient way of determining the therapeutic dose, reducing the need for frequent follow-up visits, and ultimately improving patient outcomes.

Related work:

Found some notable research work on using machine learning models to estimate warfarin dose using the IWPC dataset. As mentioned in [1], the dropping of null records reduces the size of the dataset, so it is suggested to perform an imputation on the dataset to handle missing values. Performed feature scaling on the dataset to reduce the magnitude of the features. As per the research paper [1],[2],[3],[4] Mean Absolute Error is widely used to study warfarin dose estimations. According to the results in [1] for different models, the MAE reported was around 8.5 to 10 approximately. The mentioned algorithms are used in this project to verify the results mentioned in [1]. The stacked ensembles are considered the top performers among all on the IWPC data, with the Stacked SV with an MAE of 8.55. Taking the experiments and methods as a reference from [1], we have designed the experiment.

Dataset:

The IWPC dataset is downloaded from the official PharmGKB website.The downloaded dataset is loaded into a notebook using the pandas library. The description of the dataset is mentioned below:

```
Data columns (total 11 columns):
 #   Column                                                                  Non-Null Count  Dtype
---  ------                                                                  --------------  -----
 0   Gender                                                                  5696 non-null   object
 1   Race (Reported)                                                         5194 non-null   object
 2   Age                                                                     5658 non-null   object
 3   Height (cm)                                                             4554 non-null   float64
 4   Weight (kg)                                                             5413 non-null   float64
 5   Diabetes                                                                3283 non-null   float64
 6   Simvastatin (Zocor)                                                     3861 non-null   float64
 7   Amiodarone (Cordarone)                                                  4182 non-null   float64
 8   Therapeutic Dose of Warfarin                                            5528 non-null   float64
 9   INR on Reported Therapeutic Dose of Warfarin                            4968 non-null   float64
 10  VKORC1 genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T  4046 non-null   object
dtypes: float64(7), object(4)
```

We can observe the presence of missing values in all the features. So Simple Imputation is applied for the dataset to handle the missing values. Performed label Encoding to convert textual categorical values to numeric values. Calculated cook's distance to identify the anomalies and removed from the dataset as part of the data preprocessing steps. The data information after removing the anomalies are given below.

```
Int64Index: 5400 entries, 0 to 5699
Data columns (total 10 columns):
 #   Column                                                                  Non-Null Count  Dtype
---  ------                                                                  --------------  -----
 0   Height (cm)_impute                                                      5400 non-null   float64
 1   Weight (kg)_impute                                                      5400 non-null   float64
 2   INR on Reported Therapeutic Dose of Warfarin_impute                     5400 non-null   float64
 3   Race (Reported)_impute                                                  5400 non-null   int64
 4   Age_impute                                                              5400 non-null   int64
 5   Gender_impute                                                           5400 non-null   int64
 6   Diabetes_impute                                                         5400 non-null   int64
 7   Simvastatin (Zocor)_impute                                              5400 non-null   int64
 8   Amiodarone (Cordarone)_impute                                           5400 non-null   int64
 9   VKORC1 genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T_impute  5400 non-null   int64
```

Feature Description:

Height: It is the height of the patient in centimeters which is stored as a float value in the dataset.

Weight: It gives the weight of the patient in kilograms which is stored as a float value in the dataset.

Age: The age is represented as some predefined ranges using string datatype.

Race: It gives the information about patient's ethnicity. It has predefined categories.

INR on Reported Therapeutic Dose of Warfarin: International Normalized Ratio is used to indicate the blood's ability to clot.

Gender: It specifies the gender of the patient, and the dataset has two categories that are male and female.

Diabetes: It is represented with 0 and 1. 0 indicates the absence of diabetes and 1 represents the presence of diabetes.

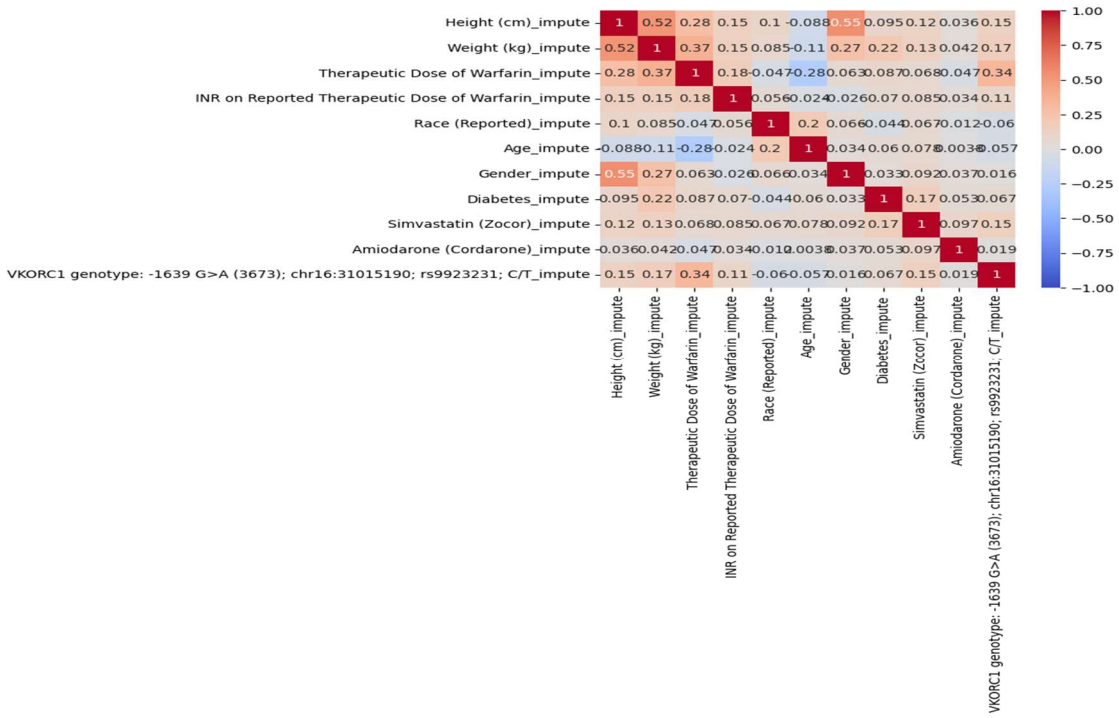
Simvastatin (Zocor): It is a medication used to lower bad cholesterol. It is represented with 0 and 1. 0 indicates the absence and 1 represents its presence.

Amiodarone (Cordarone): It is a medication used to treat heart rhythm problems. It is represented with 0 and 1. 0 indicates the absence and 1 represents its presence.

VKORC1 genotype: -1639 G>A (3673); chr16:31015190; rs9923231; C/T: This feature represents the possible genotypes and has three categories namely A/A, A/G, G/G.

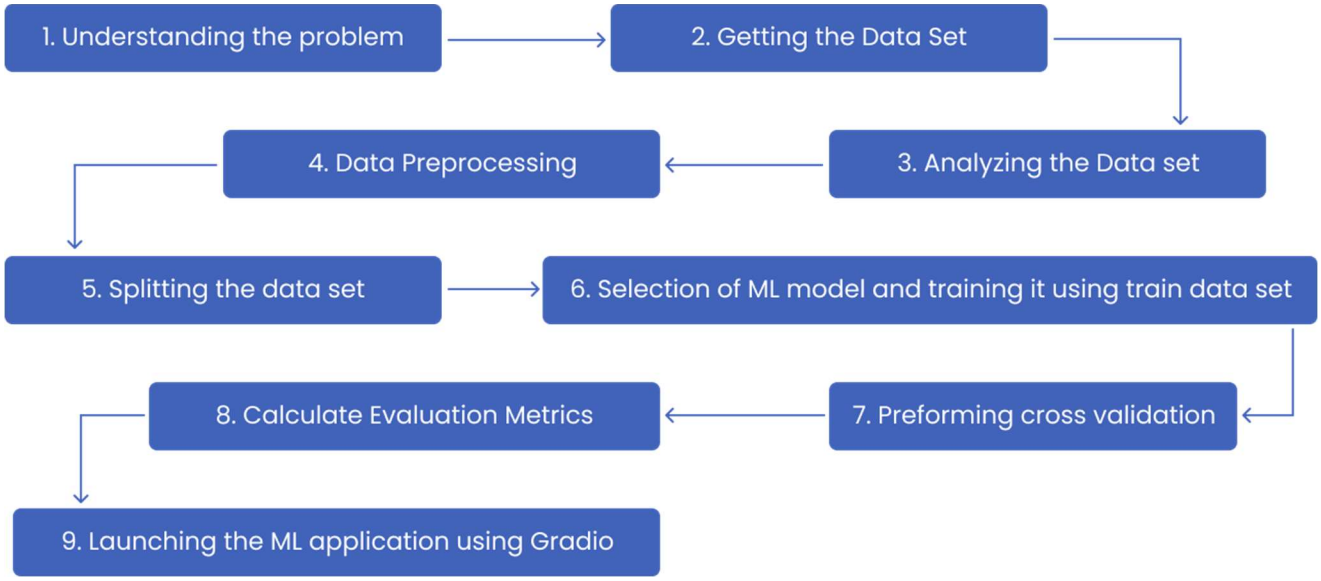
Therapeutic Dose of Warfarin: It is considered as a target variable and it represents the dose given to the patient in mg/week.

The correlation graph, we can observe that weight and height are highly correlated when compared to other features. Even gender and height are also correlated with each other.



Methodology:

Experimental Flow chat



1. *Understanding the problem:* Need to predict the dose of warfarin from the patient data. We identified the above problem as a regression problem as we need to predict a continuous value.
2. *Getting the Data Set:* Downloaded the dataset from the official PharmGKB website: https://www.pharmgkb.org/downloads. The IWPC dataset is used for the project.
3. *Analyzing the Data set*: Analyzing the total number of rows, features, and descriptions about the data from the metadata available in the data set. Observed the rows with presence of null values in the data set.
4. *Data Preprocessing Steps*: For handling missing values using a Simple Imputer, textual categorical data is converted into numeric values using a Label Encoder. Identified anomalies present in the data and removed those rows from the dataset. Cook's distance is used to identify the anomalies present in the dataset.
5. *Splitting the data set*: After the preprocessing the data is split into a training data set, test dataset, and validation set. Performed feature scaling using the MinMaxScaler method available in sklearn.preprocessing library.
6. *Selection of ML model and training it using train data set*: Trained different ML models using a train data set.ML models used in this project are Liner regression, Lasso regression, Decision Tree Regressor, SV regression, and MLP Regressor.
7. *Preform cross-validation*: Performed the Cross-validation using to get the optimal parameters and the Best Model. Passed parameter dictionary containing different values to the GridSearchCV and the mean absolute error is selected as an evaluation metric.
8. *Calculate Evaluation Metrics:* Calculated evaluation metrics available for regression problems that is mean square error, r2 square error and mean absolute error.

9. *Launch Web Application using Gradio*: Created a web application for the developed project using Gradio

Algorithms:

Linear Regression: The linear regression algorithm is the fundamental regression algorithm. Used in this experiment to compare it with other algorithms. In this method, a linear relationship is established between the dependent variable and the input features. A function is determined in linear regression, where the predicted value can be calculated using the input features and the slope of the line.

Lasso regression: It is also known as L1 regularization. It is the same as linear regression, but an extra term is added to the loss function. The parameter alpha defines the amount of regularization. The addition of an extra term reduces the magnitude of coefficients which helps in reducing the model's complexity. When alpha is zero, it is similar to linear regression.

Decision Tree Regressor: In this method, the trees are built by splitting the given data. The feature values are selected and the standard deviation is calculated, based on the threshold value, the data is split into root node and leaf nodes. If the leaf node has more than one point then the average of all points of that leaf node is calculated and used as the final value. When new data is given to the tree, The values travel across the tree and reach the node they belong and the final value is the average of the node.

SV Regressor: This method finds the function which represents the relation between input features and the targeted continuous variable to achieve minimum error. This is achieved by finding a hyperplane that best fits the data points. SVR can also handle non-linear relationships by using different kernel functions. This approach helps to reduce the prediction error and allows SVR to

handle non-linear relationships between input variables and the target variable using a kernel function.

MLP Regressor: it is a neural network-based algorithm, where we build a network with specific nodes and layers. The input features are given to the network, it passes through the nodes and it learns the optimum values for the weights and bias to minimize the error. The new value can be predicted using the optimum weights identified while training.

Results:

After model training and cross-validation is performed to find the optimum parameters for the models. The cross-validation is performed on the basis of MAE. The evaluation metrics are calculated for all the datasets. Considered MSE, MAE and r2 score as the evaluation metric for this regression problem.

Linear algorithm:

The evaluation results for linear regression on different datasets are tabulated.

	Train Set	Validation Set	Test Set
MAE	8.45	9.03	8.78
MSE	118.43	136.09	128.33
R2 Score	0.32	0.29	0.25

Lasso regression:

The optimum parameter for lasso regression is alpha which is set to 0.001 obtained from the cross-validation results. Alpha represents the constant that handles regularization.

	Train Set	Validation Set	Test Set
MAE	8.45	9.03	8.78
MSE	118.43	136.04	128.33
R2 Score	0.32	0.29	0.25

Decision Tree Regressor:

The optimum parameter for decision Tree regression 'criterion': 'absolute_error', 'max_depth': 5, 'min_samples_split': 20.

	Train Set	Validation Set	Test Set
MAE	8.20	9.14	8.56
MSE	119.30	146.44	129.65
R2 Score	0.32	0.23	0.24

SV regressor:

The optimum parameter for SV regression is C= 4 and kernel as ‘rbf’, obtained from the cross-validation results.

	Train Set	Validation Set	Test Set
MAE	8.12	8.78	8.37
MSE	112.66	133.35	119.68
R2 Score	0.35	0.30	0.30

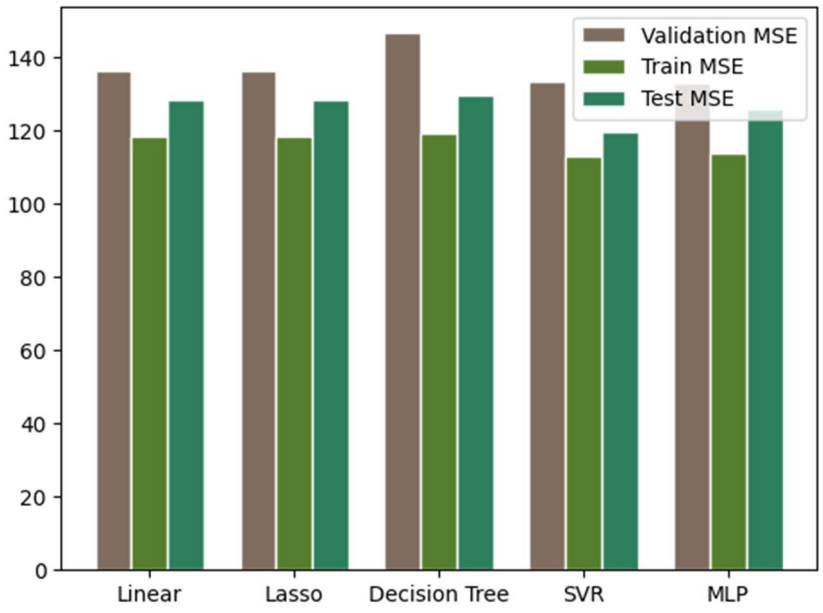
MLP Regressor:

The optimum parameters for MLP Regressor are 'activation': 'relu', 'alpha': 0.1, 'hidden_layer_sizes': (100, 100), 'max_iter': 100.

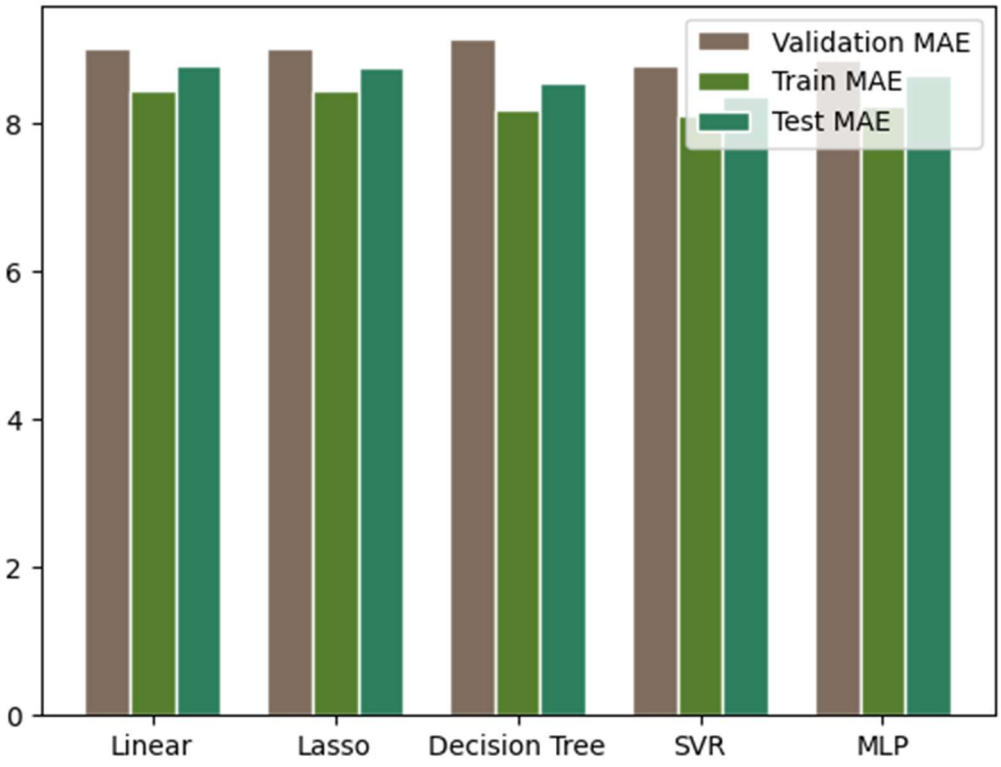
	Train Set	Validation Set	Test Set
MAE	8.24	8.87	8.66
MSE	113.58	132.89	125.86
R2 Score	0.35	0.30	0.26

Comparison Graphs:

This is the plot to compare the MSE of all sets for all models.



This is the Comparision Graph for the MAE of all datasets for all models.



Discussions:

By looking at the comparison graphs, the performance of all models is almost similar. But SV Regressor performed well when compared with other models. We can observe that the training evaluation metric are low when compared to test data, so there is an overfitting issue in this experiment. The model can be more accurate in the future by using different techniques for handling data. We can use some predictive models for handling the missing data which will improve the dataset quality. Using Ensemble techniques can also be a promising extension for this experiment. A stacked generalization approach can be used to improve warfarin dose estimation.

Overfitting is observed in this experiment which can be handled in the future. Performed all the required operations to get the best performance but was not able to achieve the ideal values. More complex techniques can be applied to get better performance.

Course Summary:

Learned about various machine learning models. Gained knowledge of the machine learning pipeline and how to tackle problems using machine learning methods. The course involved various techniques to resolve common issues like overfitting and underfitting. Learned about various Python libraries used for machine learning and improved Python programming skills.

Contribution:

Josna Pitta: Project Coding

Sai Rudh Reddy: Documentation and worked on Visualization.

References

[1] Gianluca Truda, Patrick Marais

Evaluating warfarin dosing models on multiple datasets with a novel software framework and evolutionary optimization

URL https://www.sciencedirect.com/science/article/pii/S1532046420302628?via%3Dihub

[2] Rong Liu, Xi Li ,Wei Zhang, Hong-Hao Zhou

Comparison of nine statistical model-based warfarin pharmacogenetic dosing algorithms using the racially diverse international warfarin pharmacogenetic consortium cohort database

PLoS One, 10 (8) (2015), 10.1371/journal.pone.0135784

URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0135784

[3] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi

Revisiting warfarin dosing using machine learning techniques

Comput. Math. Methods Med., 2015 (2015), 10.1155/2015/560108

URL https://www.hindawi.com/journals/cmmm/2015/560108

[4] Zhiyuan Ma , Ping Wang, Zehui Gao, Ruobing Wang, Koroush Khalighi

Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose

PLoS One, 13 (10) (2018), Article e0205872, 10.1371/journal.pone.0205872

URL http://dx.plos.org/10.1371/journal.pone.0205872