Final Project (Option 1): Build End-to-End ML pipeline for Warfarin Dosing Prediction

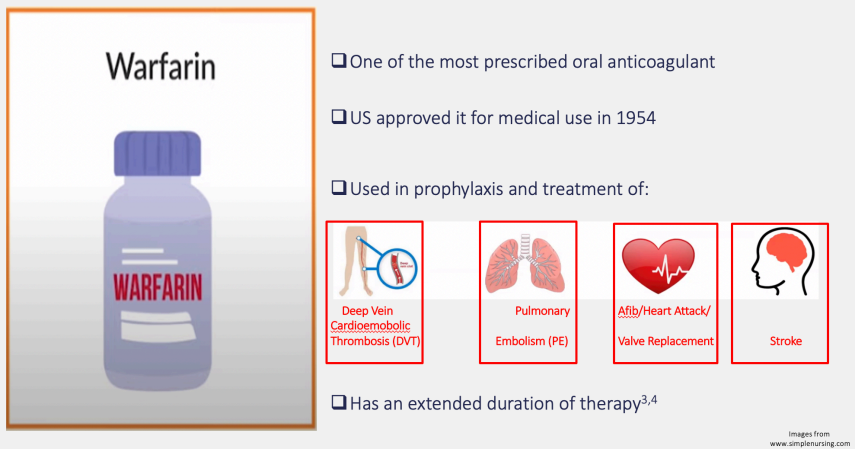
5/12/2025

Add Comment

Details

CSCI4750/5750 Course Project (Option 1): Build End-to-End ML pipeline for Warfarin Dosing Prediction

The selected course project (option 1) for Spring 2025 will focus on the machine learning in precision medicine applications: warfarin dose prediction



A. Problem Introduction

Warfarin is a low-cost and effective anticoagulant prescribed widely for decades in clinical use. Warfarin can prevent and treat thromboembolism in patients with many medical conditions (i.e., cardiac valve replacement, atrial fibrillation, and joint replacement surgeries). However, determining the optimal dosage of warfarin therapy is challenging for clinicians due to its narrow therapeutic index and high inter- and intra-individual variability among patients in dose requirements. Patients starting warfarin therapy are most likely to overdose during the initial weeks of therapy. Studies have shown that improper determination of the therapeutic dose will render patients susceptible to thromboembolism or increase the risk of bleeding. Therefore, in traditional therapy, clinicians still 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, unexpectedly increasing the utilization of health care services [1].

Researchers have already identified the relationship between the warfarin dose requirements and certain critical factors in many warfarin research.

In this project, we will study the warfarin dose-response relationship using machine learning techniques to determine the therapeutic warfarin dose during the initiation period.

[1] International Warfarin Pharmacogenetics Consortium. "Estimation of the warfarin dose with clinical and pharmacogenetic data." *New England Journal of Medicine* 360.8 (2009): 753-764.

B. Dataset Access

The dataset meets the following criteria for machine learning use in this class:

(1) the raw dataset is freely accessible online,
(2) the raw dataset requires data cleaning (i.e., missing values, a mix of categorical and numeric features),
(3) the dataset supports regression analysis, classification analysis, and clustering analysis,
(4) the related analysis & results are accessible from the online literature

Learning outcome I: one of the goal in this project is to practice how to start the machine learning project from the original, uncleaned, mixed (i.e., missing data, text values, numeric, categorical) dataset.

The dataset comes from the literature (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722908/), which provides the following description:

"*The International Warfarin Pharmacogenetics Consortium comprises 21 research groups from 9 countries and 4 continents. The research groups contributed clinical and genetic data for a total of* ***5700 patients*** *who were treated with warfarin. These data were curated (i.e., collected, formatted, and subjected to quality control) by staff at the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB,* ***www.pharmgkb.org*** *(****http://www.pharmgkb.org****)) and by members of the consortium.*"

The data can be downloaded from the official PharmGKB website: **https://www.pharmgkb.org/downloads (https://www.pharmgkb.org/downloads)** , please follow steps below to derive the original data:

- Step 1: direct to **https://www.pharmgkb.org/downloads (https://www.pharmgkb.org/downloads)**
- Step 2: search keyword: International Warfarin Pharmacogenetics Consortium (IWPC)
- Step 3: Click "IWPC Data Set" to download the data, which is excel format

International Warfarin Pharmacogenetics Consortium (IWPC)

- Data from *Estimation of the warfarin dose with clinical and pharmacogenetic data* (PMID:19228618):
  - IWPC Data Set
  - Ethnicity Data Set
  - Dosing Algorithm

C. Dataset, Feature variables and Target Output

The excel file comprises of all features and target warfarin dose label for all patients, with sample columns as follows:

PharmGKB Subject ID	PharmGKB Sample ID	Project Site	Gender	Race (Reported)	Race (OMB)	Ethnicity (Reported)	Ethnicity (OMB)	Age	Height (cm)	Weight (kg)	Indication for Warfarin Treatment
PA135312261	PA135312629	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	60 - 69	193.04	115.7	7
PA135312262	PA135312630	1	female	White	White	not Hispanic or Latino	not Hispanic or Latino	50 - 59	176.53	144.2	7
PA135312263	PA135312631	1	female	White	White	not Hispanic or Latino	not Hispanic or Latino	40 - 49	162.56	77.1	7
PA135312264	PA135312632	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	60 - 69	182.24	90.7	7
PA135312265	PA135312633	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	50 - 59	167.64	72.6	7
PA135312266	PA135312634	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	40 - 49	177.80	104.3	7
PA135312267	PA135312635	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	70 - 79	167.64	84.8	7
PA135312268	PA135312636	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	40 - 49	187.96	99.8	7
PA135312269	PA135312637	1	female	White	White	not Hispanic or Latino	not Hispanic or Latino	60 - 69	160.02	106.6	7
PA135312270	PA135312638	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	60 - 69	177.80	93.9	7
PA135312271	PA135312639	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	60 - 69	180.34	142.9	7
PA135312272	PA135312640	1	female	White	White	not Hispanic or Latino	not Hispanic or Latino	70 - 79	167.64	61.2	7
PA135312273	PA135312641	1	male	White	White	not Hispanic or Latino	not Hispanic or Latino	70 - 79	172.72	81.6	7

Target INR	Estimated Target INR Range Based on Indication	Subject Reached Stable Dose of Warfarin	Therapeutic Dose of Warfarin	INR on Reported Therapeutic Dose of Warfarin
2.5	NA	1	49.00	2.60
2.5	NA	1	42.00	2.15
2.5	NA	1	53.00	1.90
2.5	NA	1	28.00	2.40
2.5	NA	1	42.00	1.90
2.5	NA	1	49.00	2.88
2.5	NA	1	42.00	NA
2.5	NA	1	71.00	2.50
2.5	NA	1	18.00	2.53
2.5	NA	1	17.00	2.40
2.5	NA	1	97.00	NA
2.5	NA	1	49.00	2.20
2.5	NA	1	42.00	2.10

Note: The description of each column can be found in tab 'Metadata' of excel file.

Important: In this project we will only include the following features variables in data analysis, and exclude the remaining columns in excel file:

	Data Columns				
Variable	**Column Name**	**Description**	**Data Type**	**Unit Of Measure**	**Unit Of Measure Type**
Feature 1	Gender	Male, Female or not known = -99	character		
Feature 2	Race (Reported)	Self-reported information.	character		
Feature 3	Age	Binned age reported in years (0 - 9, 10 - 19, 20 - 29, 30 - 39, 40 - 49, 50 - 59, 60 - 69, 70 - 79, 80 - 89, 90+)	character		
Feature 4	Height (cm)	Reported in centimeters	number	cm	Height
Feature 5	Weight (kg)	Reported in kilograms	number	kg	Mass
Feature 6	Diabetes	yes = 1, not present = 0 or not known = NA	character		
Feature 7	Simvastatin (Zocor)	yes = 1, not present = 0 or not known = NA	character		
Feature 8	Amiodarone (Cordarone)	yes = 1, not present = 0 or not known = NA	character		
Feature 9	Target INR	Target International Normalized Ratio or NA	character	mg/week	Mass per Unit Time
Feature 10	INR on Reported Therapeutic Dose of Warfarin	International Normalized Ratio on the Therapeutic Dose of Warfarin Reported Above	number		
Feature 11	Cyp2C9 genotypes	*1, *2, *3, *4, *5, *6, *7, *8, *9, *10, *11, *12, or *13 (see https://www.pharmgkb.org/do/serve?objId=PA126&objCls=Gene for specifics of named alleles)	character		
Feature 12	VKORC1 genotype: -1639 G>A (3673);	A/A, A/G, G/G or NA	character		

	chr16:31015190; rs9923231; C/T				
Target	Therapeutic Dose of Warfarin	Dose given in milligrams/week	number	mg/week	Mass per Unit Time

Note: Make sure the column 'Therapeutic Dose of Warfarin' should not be used as features when training machine learning algorithms.

D. Regression or Classification

The dataset supports the following machine learning problems

(1) Regression: The raw output variable ('Therapeutic Dose of Warfarin') is numeric variable. If you directly use the variable as the output of your machine learning algorithms, you can train regression models on features for warfarin dose prediction.

(2) Classification: as proposed in this **paper (https://www.hindawi.com/journals/cmmm/2015/560108)** , the raw output variable ('Therapeutic Dose of Warfarin') can also be converted into binary classes. The first class contains patients who require doses of **>30mg/wk (high required dose (HRD))** and the second class contains the patients who need **doses of ≤30mg/wk (low required dose (LRD))**.

Learning outcome II: In this project, you are required to develop a machine learning pipeline for either regression problem, or classification problem, or both if you have time.

E. Related Publications

The following publications include research results of warfarin dose prediction using statistical and machine learning algorithms. You can read the details and discuss with team members to figure out the best workflow to apply ML algorithms on the dataset. You can refer to any of the papers to preview (1) how the ML framework was developed, (2) how the results are evaluated, and (3) how the results are visualized.

- Paper 1: Linear regression: **https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722908 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722908)**
- Paper 2: Treat as classification problem using Machine learning algorithms: **https://www.hindawi.com/journals/cmmm/2015/560108 (https://www.hindawi.com/journals/cmmm/2015/560108)**

- Paper 3: Ensemble technique: **https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205872 (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205872)**

> Note: The papers are only for reference. You should also refer to any other related publications or resources. You do not need to reproduce all analyses conducted in the papers. You should design any framework that you can understand and are capable to carry out the experiments.

F. Project Requirement

F.1. General instructions: Everyone needs to build a machine learning framework for the task above. Please carefully read the following materials to note the expected outcomes from each task. You can form a team (**up to 2 students**) to work on the project.

Please sign up your group in

F.2. You have many flexible rooms to decide how you will use this data and develop algorithms to accomplish the task. **However, the grading will be influenced according to the experiment design of the completed work.** (Found the detailed grading rubric at the bottom of the page)

Learning outcome III: I would expect to see the following analysis should be included in your final report, including but not limited to:

> **Question 1: How the data is preprocessed? How the data is loaded into python to start the ML pipeline?**

Question 2: How is the missing data handled (i.e., imputation)? Do any of variables have multicollinearity issues?

Question 3: how to design an ML framework that applies multiple ML methods on this dataset and benchmarks their performance;

- The experiments must include the standard steps including data visualization, data diagnosis, cross-validation, parameter tuning, and results evaluation.
- For graduate students, at least four ML algorithms (at least three traditional ML algorithms, and at least one deep learning algorithm) should be selected from the following topics in our textbook. Higher scores will be assigned to teams that conduct more comprehensive analysis or more effort put into different methods.

    a. chapter 3: classification

    b. chapter 4: regression problem (i.e., Linear, Lasso, Ridge regression)

    c. chapter 5: support vector machine

    d. chapter 6: decision trees

    e. chapter 7: ensemble learning and random forests

    f. chapter 10: artificial neural networks

    g. or other machine learning methods (i.e., KNN, LDA, QDA, Naive Bayes)

- For undergraduate students, at least two ML algorithms (at least one traditional ML algorithms, and at least one deep learning algorithm) should be selected from the above topics in our textbook. Your machine learning pipeline must also include the standard components of the ML pipeline, including data visualization, data diagnosis, cross-validation, parameter tuning, and results in the evaluation.

Expected results from this step should be organized according to a similar format:

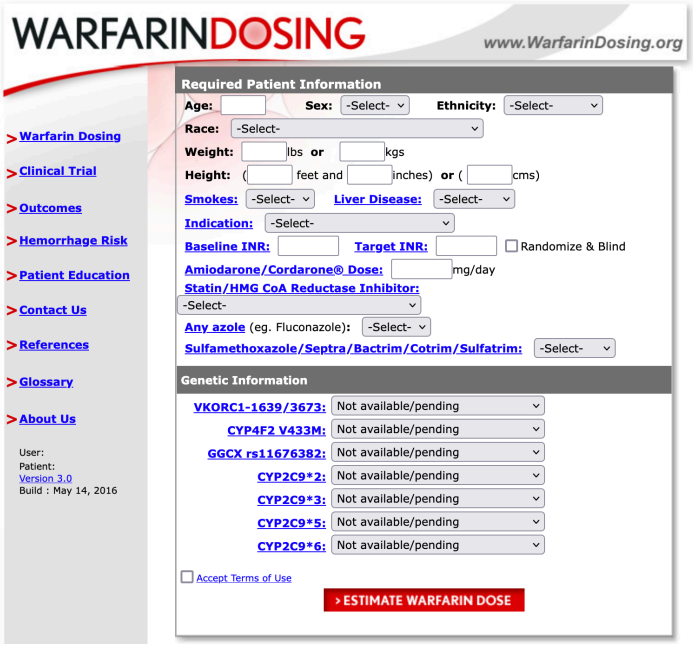
- Data descriptions, analysis, and visualization
- Method descriptions, data processing workflows, normalization, model selection, evaluation metrics, etc.
- Method comparison Table 3, 4 in paper (**https://www.hindawi.com/journals/cmmm/2015/560108** **(https://www.hindawi.com/journals/cmmm/2015/560108)**)
- If regression, report MAE, MSE, and R-square

- If classification, report accuracy, precision, recall, F1-score and ROC curves (AUC-ROC)
- Learning curves
- Other tabular or graphical visualizations

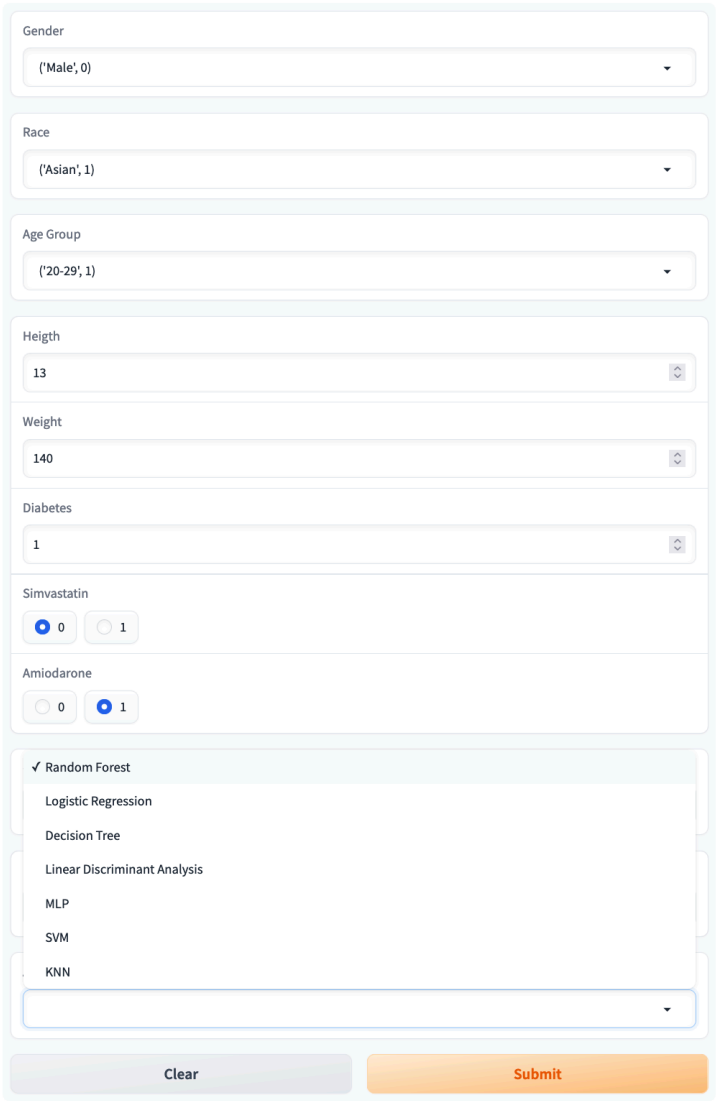
Question 4: Bonus points will be assigned if any feature selection algorithms or feature dimension reduction analysis can be applied for performance comparison.

Question 5: How to build web applications for your ML models?

Goal: We need to build a similar web application to deploy the ML models developed in this project using Gradio + HuggingFace.



A sample gradio/HuggingFace output:



G. (40%) Final Report

A report should include a title, author name, an abstract, an introduction, method description, results, conclusion, and references if any.

Every group needs to submit one report that contains the following parts:

- **Introduction**: a simple introduction to this ML topic. It may start with the importance of the topic, an appropriate review of current research on this topic, and an overview of your report sections. (around half to one page)
- **Method**:
  - Overview of your experimental pipeline (i.e. flowchart)
  - Describe how the data is obtained and processed.
  - Describe how the data is visualized.
  - Describe how the multicollinearity is assessed.
  - Describe any other techniques used such as missing value imputation, feature encodings

    - Clearly describe input features and outputs for ML methods.
    - Describe the machine learning methods you choose to study the topic, including a description of algorithms.
    - Describe how the ML model/results will be evaluated.
- **Results**:
    - model selection or parameter tuning using cross-validation
    - results on training/validation/testing
    - any other evaluation results in table or graph (i.e. learning curve, significant test, ROC)
    - detailed discussion on your results as well as possible future work.
    - Describe what parts go right, and what parts go wrong in this project.
- **Contribution:** the roles and responsibilities of team members should be clarified and documented. Every team member is expected to contribute equally to the project.
- **Course summary:** summarize what you have learned in this course.

- **Following the academic integrity polic**y (**https://www.slu.edu/arts-and-sciences/student-resources/academic-honesty.php (https://www.slu.edu/arts-and-sciences/student-resources/academic-honesty.php)**), all written work will be verified by 'ithenticate plagiarism checker software'.

H. (40%) Source code (i.e. Jupyter notebook or python files):

- ***Readability*** of the source code, including well-defined pipeline and appropriate inline comments.
- ***Contains*** all machine learning methods used in the experiments.
- ***A similarity*** check will be performed. Largely overlap with online tutorials or other submissions without your own coding may decrease your grades significantly.

I. (20%) 20~25 min Presentation (each group needs to make one presentation recording):

Step 1. Open zoom (myslu -> zoom), find the following buttons

 Screen Shot 2022-05-22 at 5.28.23 PM.png

Step 2. Click "share screen" button

Step 3. Click "record" button

Step 4. Present the following items:

a. (5 points) The presentation should contain **a short intro to your problem, a brief description of your methods, results from your experiments, and any output visualization from your machine learning model.**

b. (5 points) give a demo to show **how you processed the missing data, and cleaned the data for machine learning analysis**.

c. (5 points) give a demo to show **how to build machine learning pipeline for warfarin dosing prediction**. You need explain your code structures and run the codes without the errors.

d (5 points) summarize what you have learned in this homework.

Step 5. Stop recording

Step 6. Submit the recording link as following format in the text box of submission page

Zoom recording link:XXXXXXXXX
Passcode: XXXXXX

(do not download recording video)

J. Requirement/Grading

The final project will account for 15% of your final grade. Higher credits will be assigned if your submission contains the following contents:

1. Evaluating the effects of different pre-processing techniques on the performance. The potential pre-processing techniques including feature scaling approaches, missing value imputations, outlier removals, dimension reductions, feature encoding methods

2. Hyper-parameter tuning analysis is performed for different ML methods and discussed in report.

3. The comparison of different ML methods on the dataset is summarized and discussed in report.

The final project submission will be due on **May 12**.

The instructor will post another submission page for students to submit Project Files (your report, recorded presentation, source codes)

View Rubric

Grading Rubric for Final Project/Presentation/Codes (2) (1)

Criteria	Ratings		Pts
Presentation: Introduction / Motivation / Background view longer description	**4 pts** **Full Marks**	**0 pts** **No Marks**	/ 4 pts
Presentation: Overall organization of presentation view longer description	**4 pts** **Full Marks**	**0 pts** **No Marks**	/ 4 pts
Presentation: Novelty and critical thinking of the project view longer description	**4 pts** **Full Marks**	**0 pts** **No Marks**	/ 4 pts
Presentation: Clear discussion of results view longer description	**4 pts** **Full Marks**	**0 pts** **No Marks**	/ 4 pts
Presentation: Presentation skills view longer description	**4 pts** **Full Marks**	**0 pts** **No Marks**	/ 4 pts
Final Report: Introduction view longer description	**5 pts** **Full Marks**	**0 pts** **No Marks**	/ 5 pts
Final Report: Related work view longer description	**5 pts** **Full Marks**	**0 pts** **No Marks**	/ 5 pts
Final Report: Dataset view longer description	**10 pts** **Full Marks**	**0 pts** **No Marks**	/ 10 pts
Final Report: Methods view longer description	**10 pts** **Full Marks**	**0 pts** **No Marks**	/ 10 pts
Final Report: Results view longer description	**10 pts** **Full Marks**	**0 pts** **No Marks**	/ 10 pts
Final Report: Discussion view longer description	**10 pts** **Full Marks**	**0 pts** **No Marks**	/ 10 pts

Grading Rubric for Final Project/Presentation/Codes (2) (1)

Criteria	Ratings		Pts
Source Codes view longer description	**30 pts** **Full Marks**	**0 pts** **No Marks**	/ 30 pts

Total Points: 0