Data Glacier

Healthcare – Drug Persistency

Groupe name: Ehiafarin
Country: Iran
Groupe member: Alireza Ehiaei
Email: arh.ehiaei@yahoo.com

Specialization: Data Science
Model Selection and Model Building

September-2021

Table Of Contents

01 Introduction

02 Problem Statement

03 Logistic Regression

04 Neural Networks

05 Ensemble Modeling

Introduction

In previous research, we developed a machine learning model to predict the drug persistency of patients based on a high number of features. The number of features and their relationships is reasons that why traditional methods could not have the accuracy of machine learning methods.

But the question is which machine learning method would be the best one to use for a specific data type?

Introduction

For example, we used the random forest method from the grope of classification algorithms, but one can choose another machine learning method such as Logistic regression from the regression algorithms, or for example, K-Means clustering, from clustering algorithms.

The aim of this research is to use another ML method than random forest and compare their results, then finding a way to use combined results of these methods. This is achieved in machine Learning by a technique called Ensemble Learning. Before starting a new model, reminding the data set and problem.

Problem Statement

The dataset is related to drug treatment, and the data set includes 3424 rows (patients) and 69 columns which one of which is drug persistency as the target variable that has the value of persistence or non-persistence.

Therefore, the problem is a binary classification problem. It can be analyzed using different machine learning models from different groups. We continue by developing a Logistic regression model and the model of a neural network. First, reminding the random forest model.

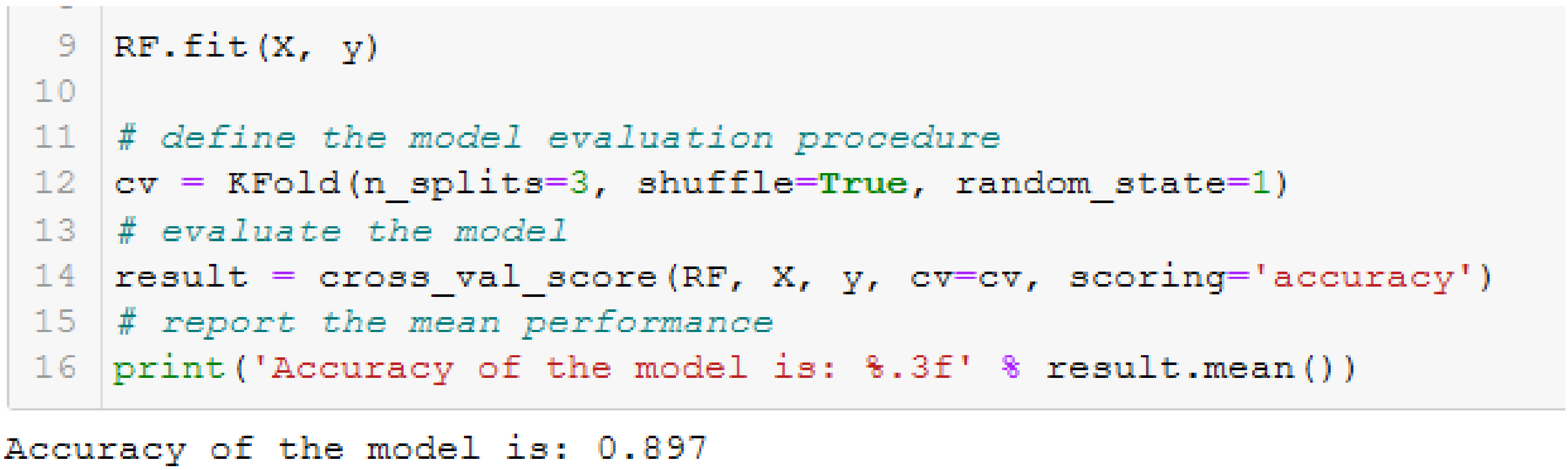
Random Forests

In previous practice, we used the random forest method. We remind its result here to make ensemble modeling. Random forest is a supervised machine learning algorithm that is used in classification and regression problems.

It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Random Forests

By fitting the random forest model, and use it to predict drug persistency, we find the accuracy of this method on our data is 0.89.



```
RF.fit(X, y)

# define the model evaluation procedure
cv = KFold(n_splits=3, shuffle=True, random_state=1)
# evaluate the model
result = cross_val_score(RF, X, y, cv=cv, scoring='accuracy')
# report the mean performance
print('Accuracy of the model is: %.3f' % result.mean())
```

```
Accuracy of the model is: 0.897
```

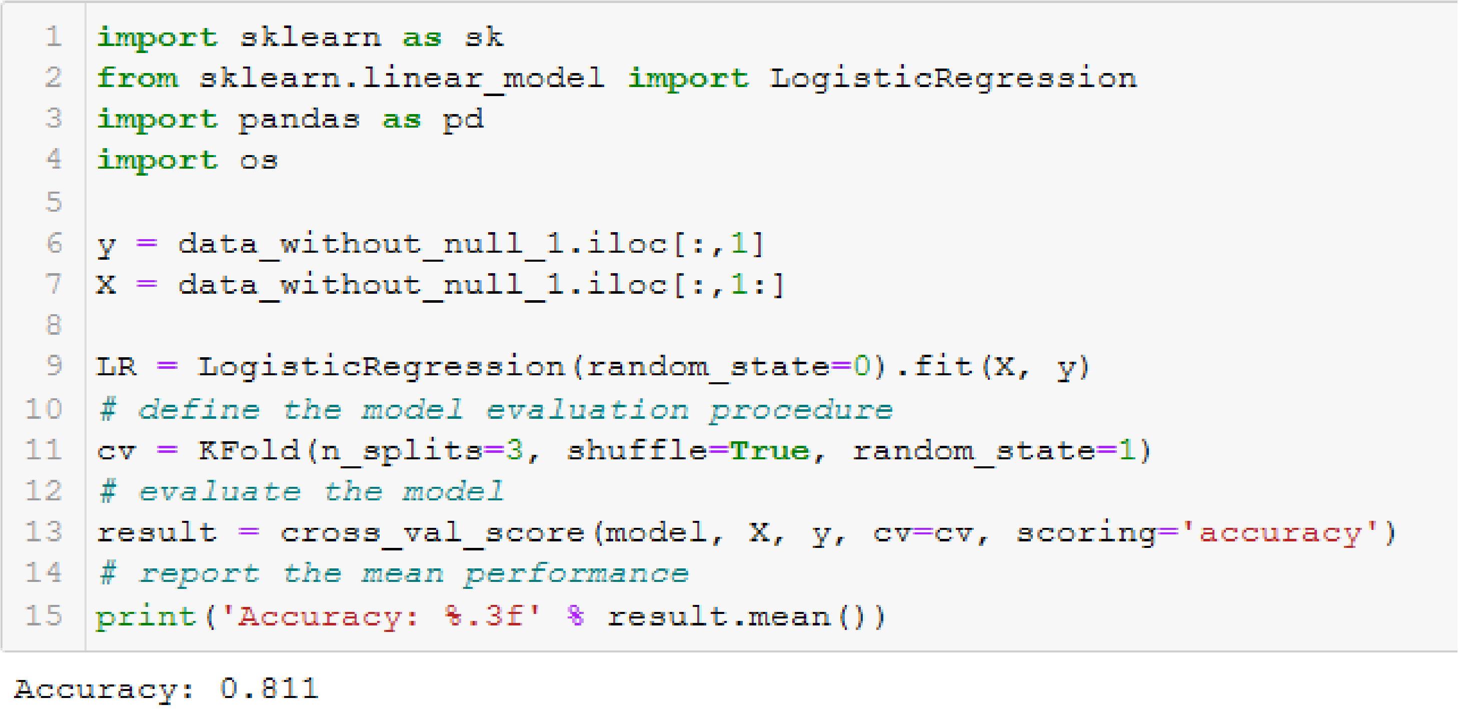
Logistic Regression

Logistic Regression is a type of Generalized Linear Model (GLM) that uses a logistic function to model a binary variable based on any kind of independent variable.

Logistic regression is used to obtain an odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial.

Logistic Regression

Logistic regression is applied to predict the categorical dependent variable, and our problem is predicting a binary target variable. So, logistic regression can be a good choice to solve our classification.



```
import sklearn as sk
from sklearn.linear_model import LogisticRegression
import pandas as pd
import os

y = data_without_null_1.iloc[:,1]
X = data_without_null_1.iloc[:,1:]

LR = LogisticRegression(random_state=0).fit(X, y)
# define the model evaluation procedure
cv = KFold(n_splits=3, shuffle=True, random_state=1)
# evaluate the model
result = cross_val_score(model, X, y, cv=cv, scoring='accuracy')
# report the mean performance
print('Accuracy: %.3f' % result.mean())
```

```
Accuracy: 0.811
```

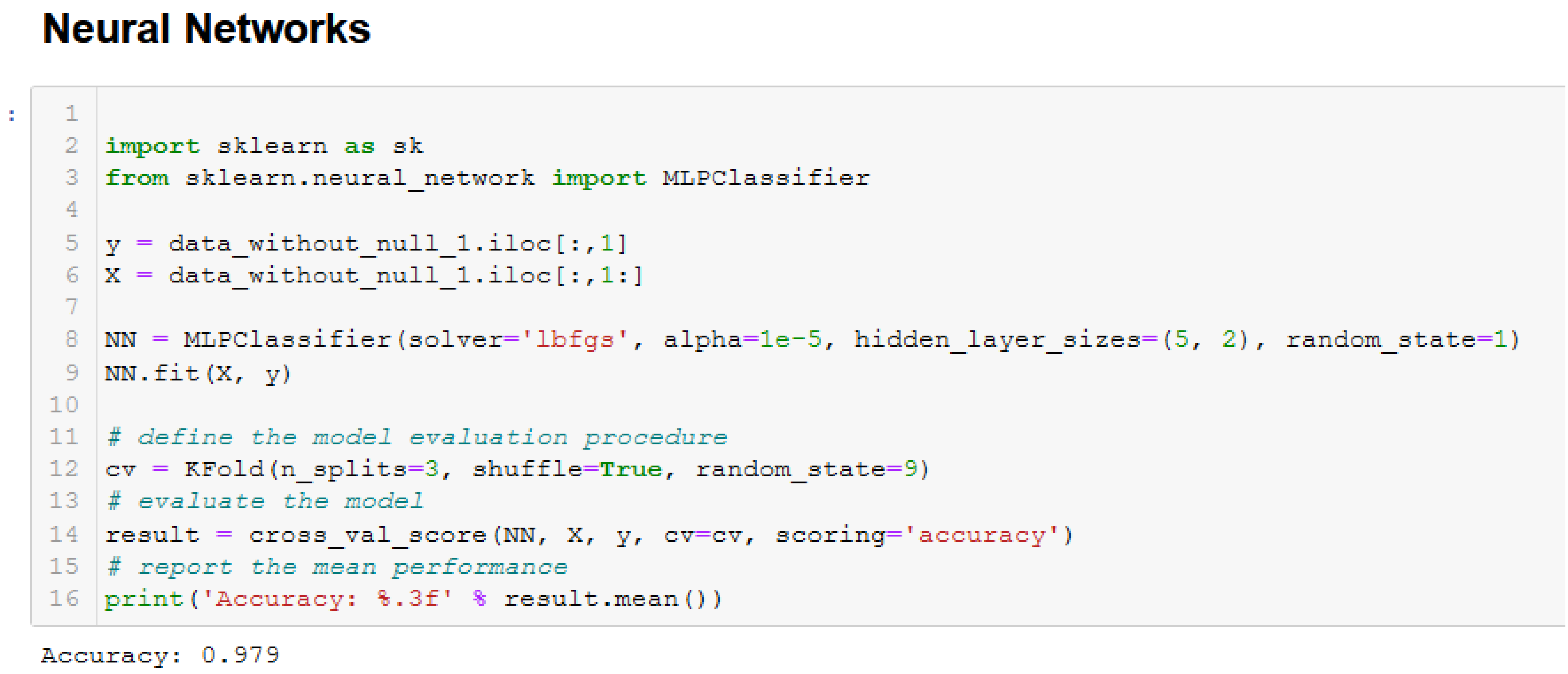
Neural Networks

Neural Networks are a machine learning algorithm that involves fitting many hidden layers used to represent neurons that are connected with synaptic activation functions. These essentially use a very simplified model of the brain to model and predict data.

They consists of an artificial network of functions, called parameters, which allows the computer to learn, and to fine tune itself, by analyzing new data. Each parameter (neurons) is a function which produces an output, after receiving one or multiple inputs and then passed to the next layer of neurons several times and produce further outputs until every layer of neurons have been considered, and the terminal neurons have received their input. Those terminal neurons then output the final result for the model.

Neural Networks

Neural Networks are well known techniques for classification problems.
They can also be applied to regression problems.



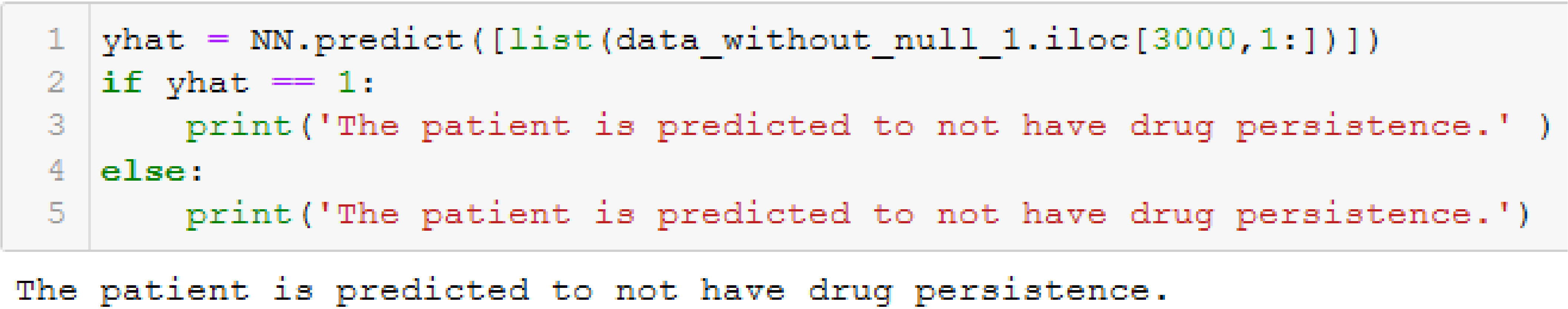
Neural Networks

```python

import sklearn as sk
from sklearn.neural_network import MLPClassifier

y = data_without_null_1.iloc[:,1]
X = data_without_null_1.iloc[:,1:]

NN = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5, 2), random_state=1)
NN.fit(X, y)

# define the model evaluation procedure
cv = KFold(n_splits=3, shuffle=True, random_state=9)
# evaluate the model
result = cross_val_score(NN, X, y, cv=cv, scoring='accuracy')
# report the mean performance
print('Accuracy: %.3f' % result.mean())
```

```
Accuracy: 0.979
```

Prediction

As it is shown, for the patient number 3000 the prediction is not to be drug persistence. The question arises that which model is better?



```
yhat = NN.predict([list(data_without_null_1.iloc[3000,1:])])
if yhat == 1:
    print('The patient is predicted to not have drug persistence.' )
else:
    print('The patient is predicted to not have drug persistence.')
```

```
The patient is predicted to not have drug persistence.
```

Ensemble Modeling

An ensemble is a supervised learning technique for combining multiple weak learners/ models to produce a strong learner .

It is possible to combine multiple models of same ML algorithms, but combining multiple predictions generated by different algorithms would normally lead to better predictions.

For example, the predictions of a random forest, a KNN, and a Naive Bayes may be combined to create a stronger final prediction set as compared to combining three random forest model.

Ensemble Modeling

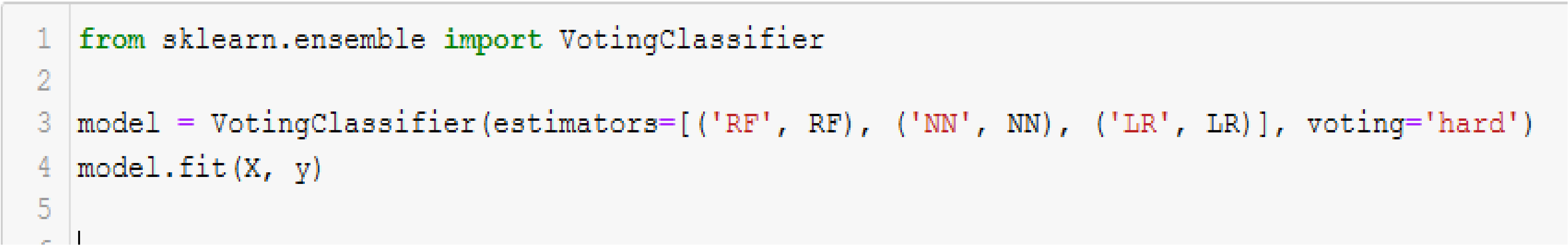
Therefore, the key to creating a powerful ensemble is model diversity leading to two major benefits of Ensemble models: better prediction and more stable model.

We have three ML models, random forest, logistic regression and neural networks which belong to different groups of classification and regression algorithms. So, they are diverse enough in nature that lead to increase performance of ensemble model we are going to develop using them.

Ensemble Techniques; Max Voting

The max voting method is generally used for classification problems. In this technique, multiple models are used to make predictions for each data point.

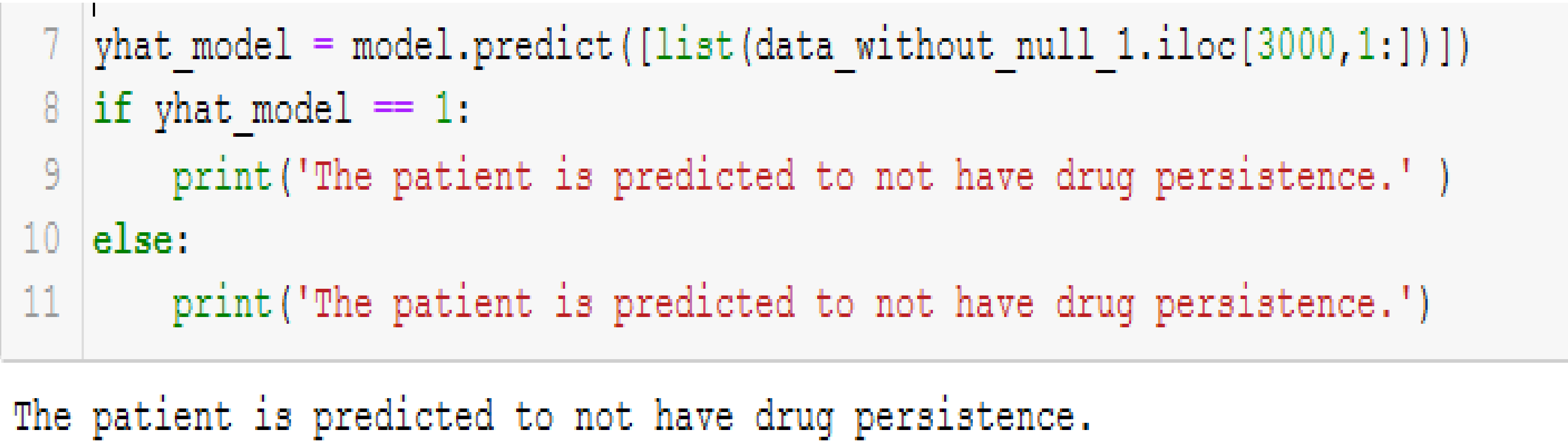
The predictions by each model are considered as a ‘vote’. The predictions which we get from the majority of the models are used as the final prediction.



```python
from sklearn.ensemble import VotingClassifier

model = VotingClassifier(estimators=[('RF', RF), ('NN', NN), ('LR', LR)], voting='hard')
model.fit(X, y)

```

Ensemble Techniques; Max Voting

Finally, the ensemble model is ready to predict the target variable.



```
yhat_model = model.predict([list(data_without_null_1.iloc[3000,1:])])
if yhat_model == 1:
    print('The patient is predicted to not have drug persistence.' )
else:
    print('The patient is predicted to not have drug persistence.')
```

```
The patient is predicted to not have drug persistence.
```

Data and code link

Data and code are uploaded at:

https://github.com/Alireza-Ehiaei/Data_Sciences/tree/main/Drug_Persistency1/Combined_ML_model

Thank You