Data Glacier

Your Deep Learning Partner

Healthcare – Drug Persistence

Groupe name: Ehiafarin

Country: Iran

Groupe member: Alireza Ehiaei

Email: arh.ehiaei@yahoo.com

Specialization: Data Science

EDA Presentation and proposed modeling techniques

October-2021

Table Of Contents

01 Introduction

02 Machine learning models

03 Problem Statement

04 Classification Models

05 Recommendation to technical users

Introduction

The objective of this document is to provide short exploratory data analysis (EDA) on what is machine learning (ML) model, what was the problem of the research that is a predicting problem in nature, and how ML models can be used to solve the problem.

Predicting problem can be treated intuitively (that is, through visualization) or rigorous (that is, statistical or machine learning-based) analysis.

Machine learning models

Machine learning [1]

In general, a learning problem considers a set of n samples of data and then tries to predict properties of unknown data. This problem can be either:

1. supervised learning, in which the data comes with additional attributes that the goal is predicting them, including the prediction models of:

- **classification**: samples belong two or more classes and we want to learn from already labeled data how to predict the class of unlabeled data.
- **regression**: if the desired output consists of one or more continuous variables, then the task is called regression.

[1] - https://scikit-learn.org/stable/tutorial/basic/tutorial.html

Machine learning models

Machine learning

2. unsupervised learning, in which the training data consists of a set of input vectors x without any corresponding target values and the goal in such problems may be:

- **Clustering**: discovering groups of similar examples within the data.
- **Density estimation**: determining the distribution of data within the input space, or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization.

Machine learning models

Therefore, supervised Learning is applied when we have a labelled data set (the output variable/dependent variable). For example, a data set which contains the size of the house (independent variable) and corresponding house price (dependent variable). We can predict the house price of new data points with respect to the size of the house. Another example is, determining if a tumor is harmful or not harmful when we already have a list of tumors which are harmful or not. In supervised learning we know the problem statement and have all the necessary features to get answer.

Machine learning models

But in unsupervised Learning, we do not have labelled data. We do not have any output variable. We do not know the problem statement. It is applied when we need to find a structure in the data set and extract meaningful insights out of it. For example, a data set of Walmart containing its customer's buying pattern.

Classification and regression algorithms are used when dealing with a supervised learning problem and clustering algorithms are used when dealing with unsupervised learning.

Machine learning models

Three major classification of ML algorithms are:

1) Classification Algorithms - Naive Bayes Classification, Decision Tree, Random Forest, kNN, Support Vector Machine (SVM), Neural Networks, etc.

2) Regression Algorithms - Linear Regression, Logistic Regression, Lasso Regression, etc. (Note: Although Logistic Regression has Regression in its name, it is essentially a classification algorithm).

3) Clustering Algorithms - K-Means Clustering, Fuzzy C Means, Mixture of Gaussian, etc.

Machine learning models

Since the problem explained later in this research is a classification problem, some clarifications about classification is necessary.

Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.

The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.

Classification Terminologies In Machine Learning

Classifier – It is an algorithm that is used to map the input data to a specific category.

Classification Model – The model predicts or draws a conclusion to the input data given for training, it will predict the class or category for the data.

Feature – A feature is an individual measurable property of the phenomenon being observed.

Binary Classification – It is a type of classification with two outcomes, for eg – either true or false.

Multi-Class Classification – The classification with more than two classes, in multi-class classification each sample is assigned to one and only one label or target.

Multi-label Classification – This is a type of classification where each sample is assigned to a set of labels or targets.

Initialize – It is to assign the classifier to be used for the

Train the Classifier – Each classifier in sci-kit learn uses the fit(X, y) method to fit the model for training the train X and train label y.

Predict the Target – For an unlabeled observation X, the predict(X) method returns predicted label y.

Evaluate – This basically means the evaluation of the model i.e classification report, accuracy score, etc.

Problem Statement

Drug persistency defined as "the extent to which a patient acts in accordance with the prescribed interval, and dose of a dosing regimen. Measuring patient persistency with drug therapy provides valuable information for healthcare decision makers concerning the effectiveness of a drug in a routine practice setting, which epidemiologists call the population based setting, as opposed to the trial- or clinic-based setting.

One of the challenge for all pharmaceutical companies is to understand the persistency of drug as per the physician prescription. To solve this problem ABC pharma company approached an analytics company to automate this process of identification.

Problem Statement

Drug persistency detection can be identified as a classification problem, this is a binary classification since there can be only two classes i.e has persistency or does not have persistency. The classifier, in this case, needs training data to understand how the given input variables are related to the class. And once the classifier is trained accurately, it can be used to detect whether persistency is there or not for a particular patient.

Data insight

There are 3424 rows and 69 coloumnsin which except columns ‘DexaFreqDuring Rx’ and ‘Count Of Risks’ which are integer, other coloumnshave string type. From 3424 cases, there are 1289 drug persistent patients and 2135 non persistent patients.

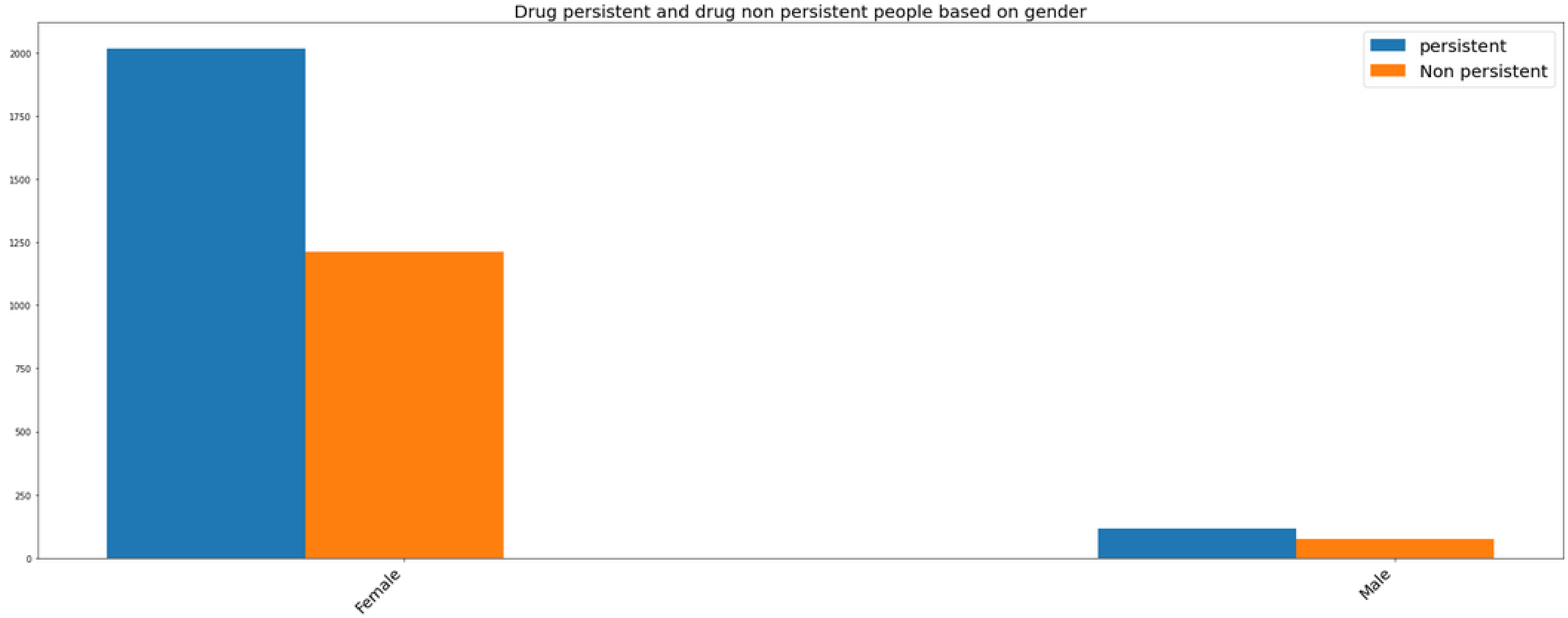
The aim is predicting drug persistency of a patient. And there is Persistency-Flag column as target column, with two labels of Persistent and Non-Persistent.o, It is a binary classification problem.

Data analysis

The features in this research are categorical variables. These variables are typically stored as text values which represent various traits. Some examples include gender, age_bucket, risk_low_calcium_intake, risk_vitamin_D_insufficiency. We can do some explanatory analysis to find insights for relationship among these variables by using statistical and visualizing methods

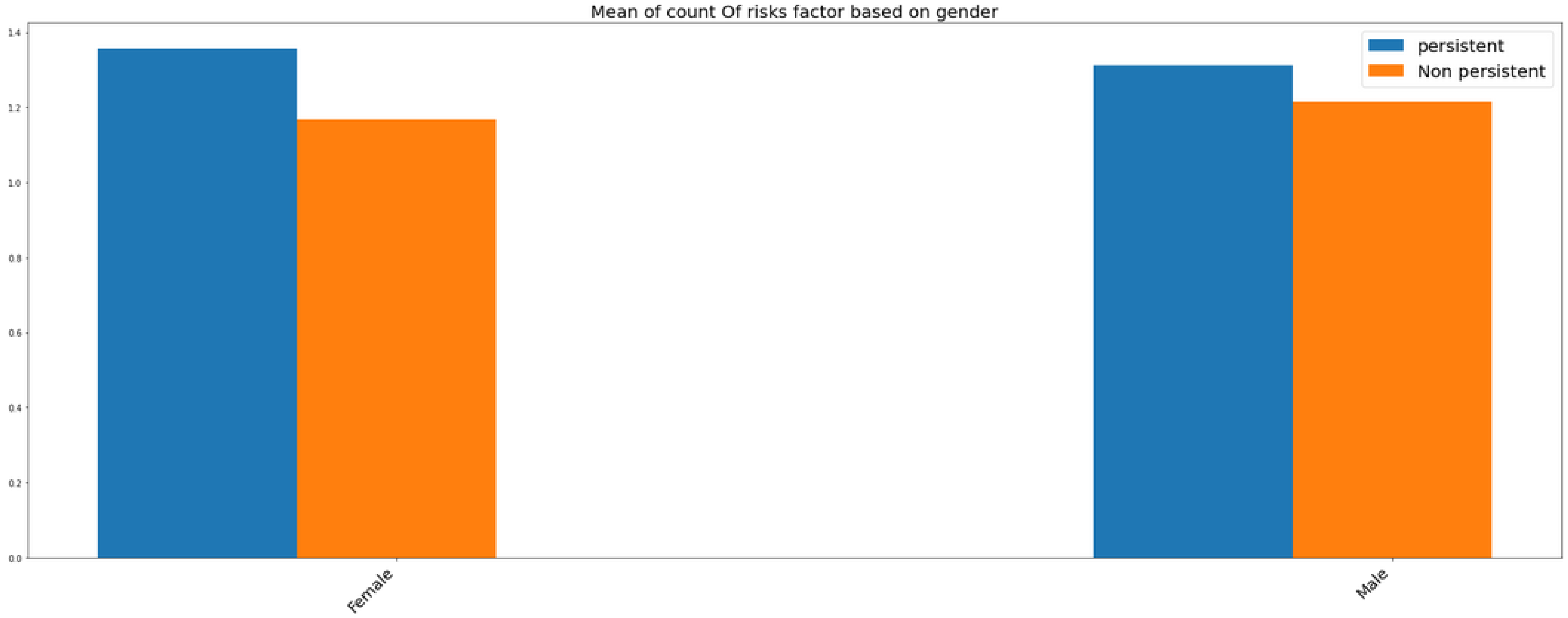
Data analysis

Most of the data is related to female:



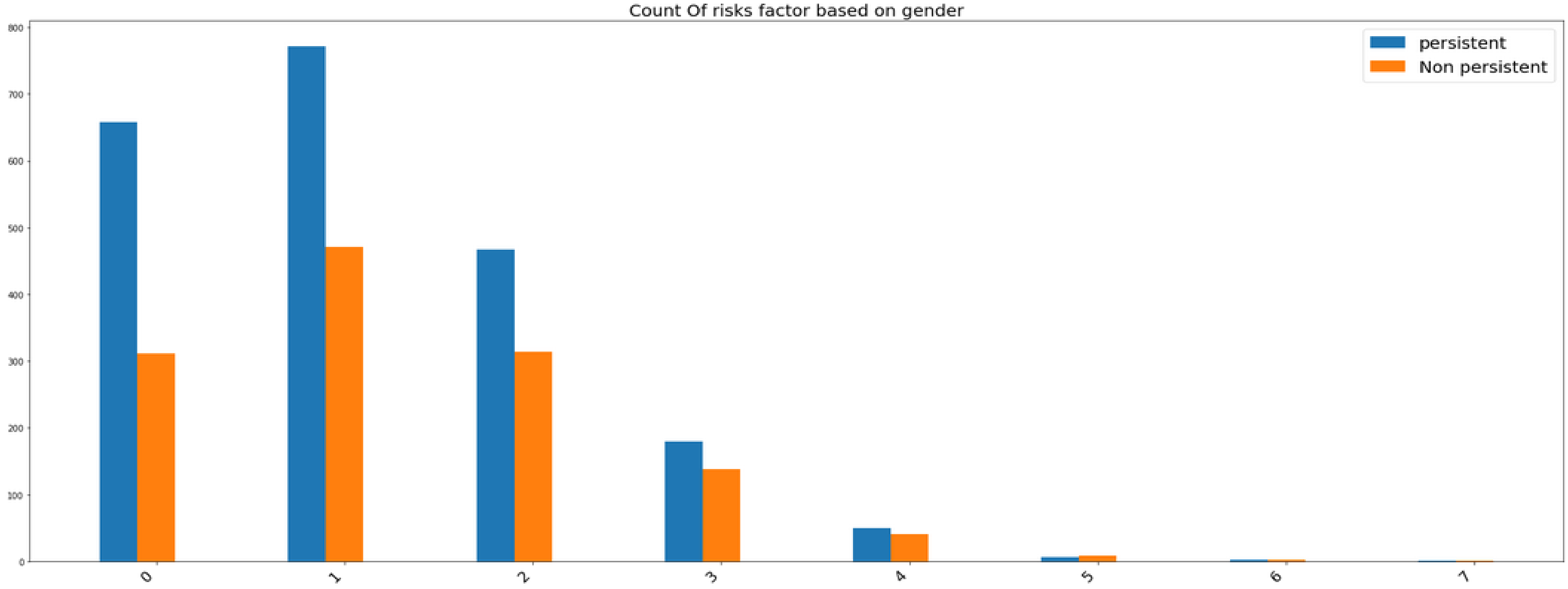
Data analysis

Average of counts of risk factors is same in comparison of female and male, and both has higher average in their drug persistent groupe.



Data analysis

Interestingly, people with one risk factor are accounts most subset of people in both groups of persistent and non persistent people.



Machine learning model

Regardless of what the value is used for, the challenge is determining how to use this data in machine learning. Many machine learning algorithms can support categorical values without further manipulation but there are many more algorithms that do not. Therefore, we converted to numbers to be able to use in different methods.

Machine learning model

In this research different supervised machine learning algorithm from Scikit-learn is used to create a model predicting the targeted variable. Scikit-learn is a free software machine learning library for the Python programming language.

The dataset has missing values in different columns that have been treated appropriately to create a robust model.

There is no unique rule to handle missing values in a specific manner. One can use various methods on different features depending on how and what the data is about. Indeed, each dataset needs some specific approaches to handle missing values.

Machine learning model

After dealing with missing data, all columns except the persistence one are assigned to a data frame called X as input variables, and the column persistence is added to the data frame as the y variable. Then with three different classification methods explained below predicting models are developed and fitted across X and y.

Finally, the model can be used in order to predict whether a patient given a set of information of features in X, is persistent in drug administration or not?

Classification Models

In machine learning, classification is a supervised learning concept which basically categorizes a set of data into classes. It can be either a binary classification problem like our problem or a multi-class problem too. There are a bunch of machine learning algorithms for classification in machine learning.

We used random forest, logistic regression, and neural network algorithms to create prediction models. Finally, a combined model so-called ensemble model is developed.

Random Forest

Random decision trees or random forest are an **ensemble learning method** for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees.

A random forest is a meta-estimator that fits a number of trees on various subsamples of data sets and then uses an average to improve the accuracy in the model's predictive nature. The sub-sample size is always the same as that of the original input size but the samples are often drawn with replacements.

Random Forest

Advantages and Disadvantages

The advantage of the random forest is that it is more accurate than the decision trees due to the reduction in the over-fitting. The only disadvantage with the random forest classifiers is that it is quite complex in implementation and gets pretty slow in real-time prediction.

Random Forest

In order to create a prediction model using random forest technics, we determined all columns except the first one as explanatory variables, and the first one as target variable that is the persistency flag.

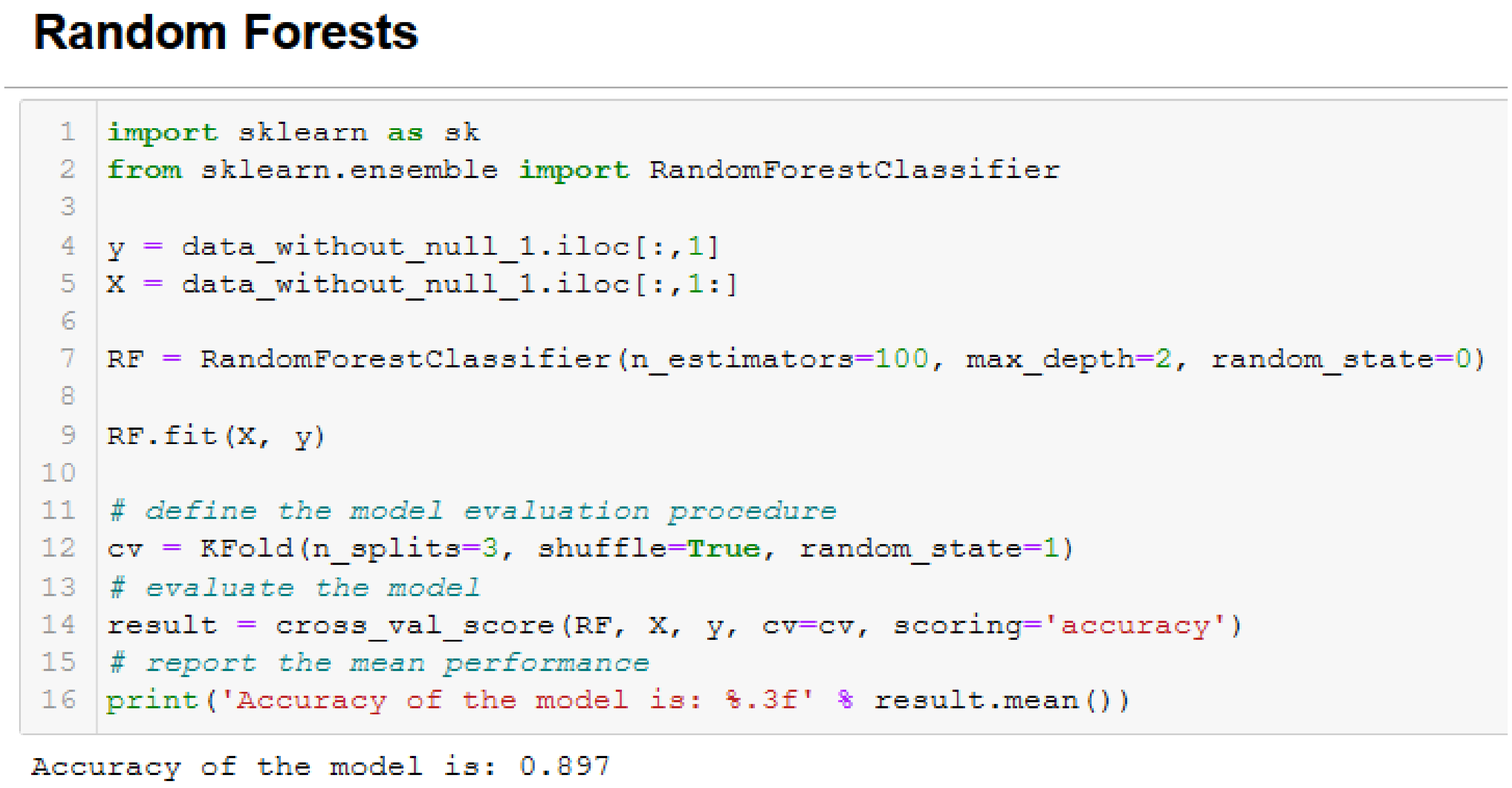
Most of the data including all columns are determined as training data and the built model using them can be evaluated by using different methods such as k-Fold Cross-Validation on the remain of data so-called test data set.

Random Forest

k-Fold Cross-Validation is used to determine the accuracy of the model (below the accuracy is 0.89). It is common to evaluate machine learning models on a dataset using k-fold cross-validation.

The k-fold cross-validation procedure divides a limited dataset into k non-overlapping folds. Each of the k folds is given an opportunity to be used as a held-back test set, whilst all other folds collectively are used as a training dataset. A total of k models are fit and evaluated on the k hold-out test sets and the mean performance is reported.

Random Forest



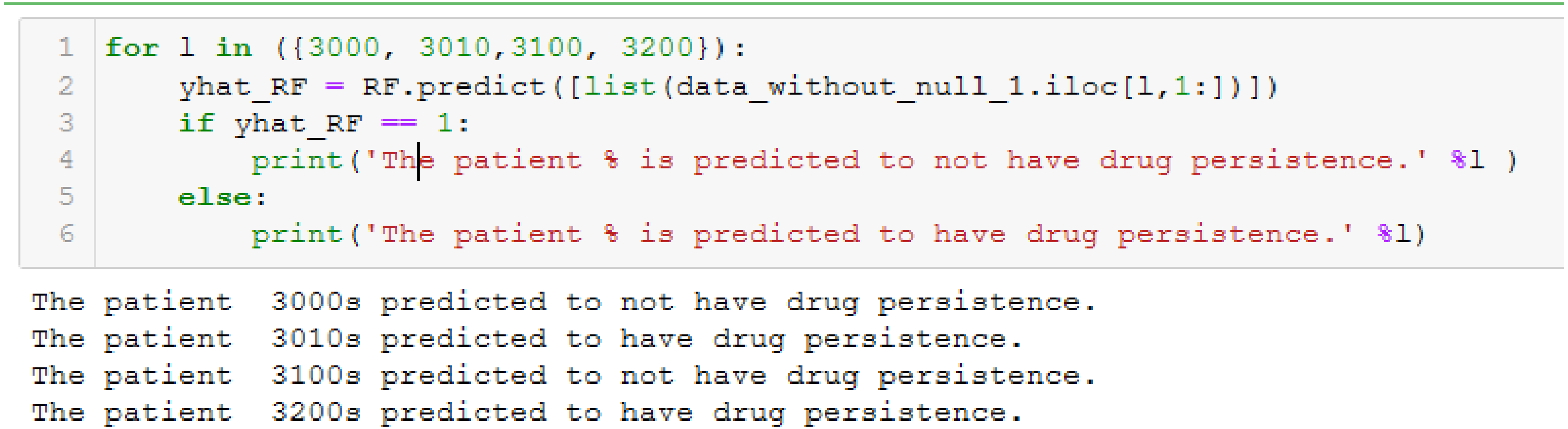
Random Forests

```python
import sklearn as sk
from sklearn.ensemble import RandomForestClassifier

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

RF = RandomForestClassifier(n_estimators=100, max_depth=2, random_state=0)

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
```

Random Forest

Then, the application of the developed model is to predict the drug persistency of a patient. For example, assume we want to predict whether a patient with information mentioned in the rows 3000, 3010, 3100, and 3200:



```
for l in ({3000, 3010,3100, 3200}):
    yhat_RF = RF.predict([list(data_without_null_1.iloc[l,1:])])
    if yhat_RF == 1:
        print('The patient % is predicted to not have drug persistence.' %l )
    else:
        print('The patient % is predicted to have drug persistence.' %l)
```

```
The patient  3000s predicted to not have drug persistence.
The patient  3010s predicted to have drug persistence.
The patient  3100s predicted to not have drug persistence.
The patient  3200s predicted to have drug persistence.
```

Neural Networks

The last algorithm used to create prediction model is neural network . A neural network consists of neurons that are arranged in layers, they take some input vector and convert it into an output. The process involves each neuron taking input and applying a function which is often a non-linear function to it and then passes the output to the next layer.

In general, the network is supposed to be feed-forward meaning that the unit or neuron feeds the output to the next layer but there is no involvement of any feedback to the previous layer.

Weighings are applied to the signals passing from one layer to the other, and these are the weighings that are tuned in the training phase to adapt a neural network for any problem statement.

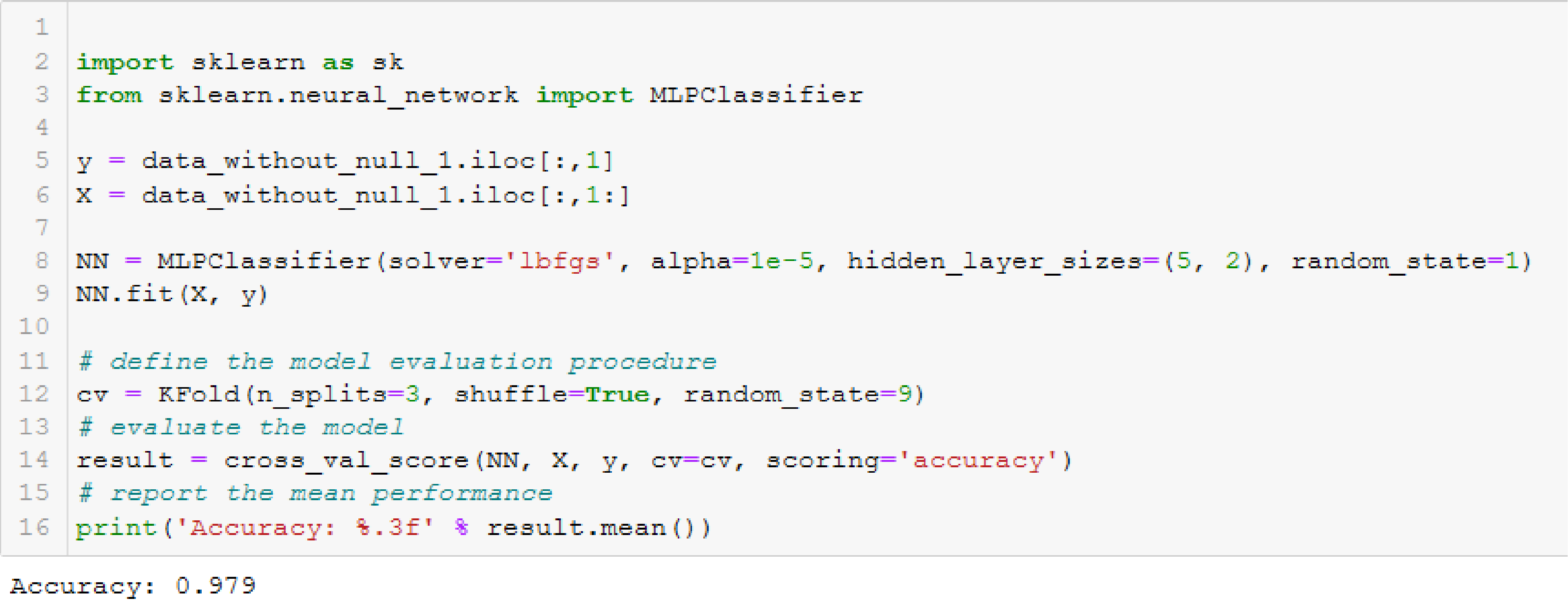
Neural Networks

Advantages and Disadvantages

It has a high tolerance to noisy data and able to classify untrained patterns, it performs better with continuous-valued inputs and outputs. The disadvantage with the artificial neural networks is that it has poor interpretation compared to other models.

Neural Networks

It seems neural network is a better technic in the our case:

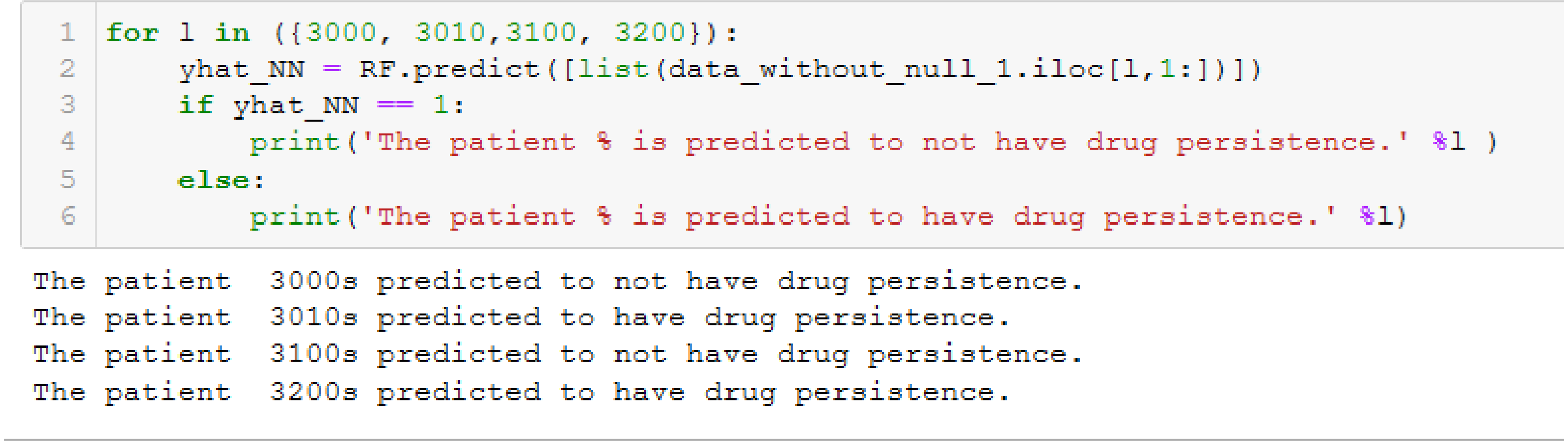


```python

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
```

Neural Networks

We can test result of this model that is similar to prediction of random forest model:



```
for l in ({3000, 3010,3100, 3200}):
    yhat_NN = RF.predict([list(data_without_null_1.iloc[l,1:])])
    if yhat_NN == 1:
        print('The patient % is predicted to not have drug persistence.' %l )
    else:
        print('The patient % is predicted to have drug persistence.' %l)
```

```
The patient  3000s predicted to not have drug persistence.
The patient  3010s predicted to have drug persistence.
The patient  3100s predicted to not have drug persistence.
The patient  3200s predicted to have drug persistence.
```

Logistic Regression

It is a classification algorithm in machine learning that uses one or more independent variables to determine an outcome. The outcome is measured with a dichotomous variable meaning it will have only two possible outcomes.

The goal of logistic regression is to find a best-fitting relationship between the dependent variable and a set of independent variables. It is better than other binary classification algorithms like nearest neighbor since it quantitatively explains the factors leading to classification.

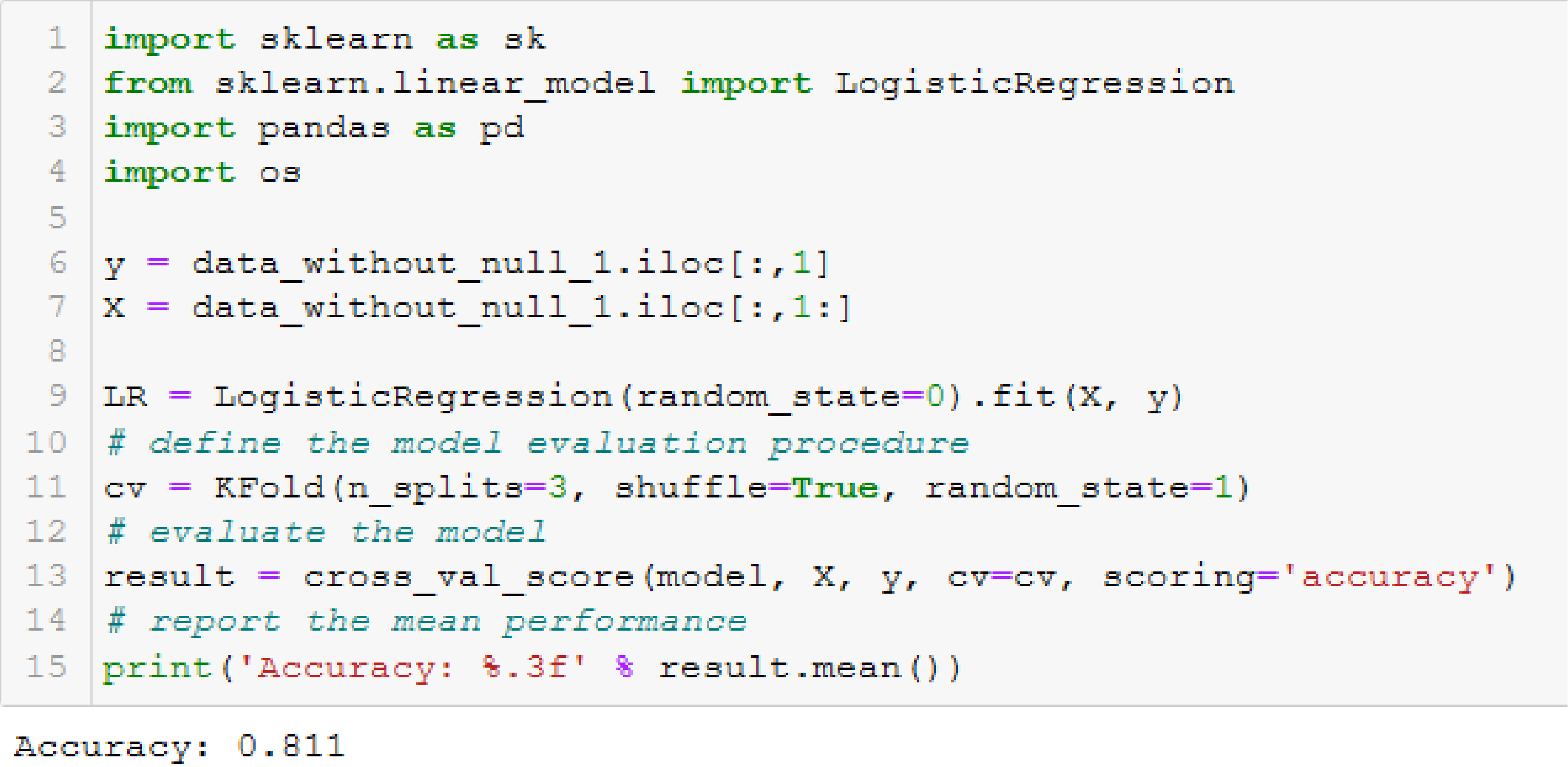
Logistic Regression

Advantages and Disadvantages

Logistic regression is specifically meant for classification, it is useful in understanding how a set of independent variables affect the outcome of the dependent variable.

The main disadvantage of the logistic regression algorithm is that it only works when the predicted variable is binary, it assumes that the data is free of missing values and assumes that the predictors are independent of each other.

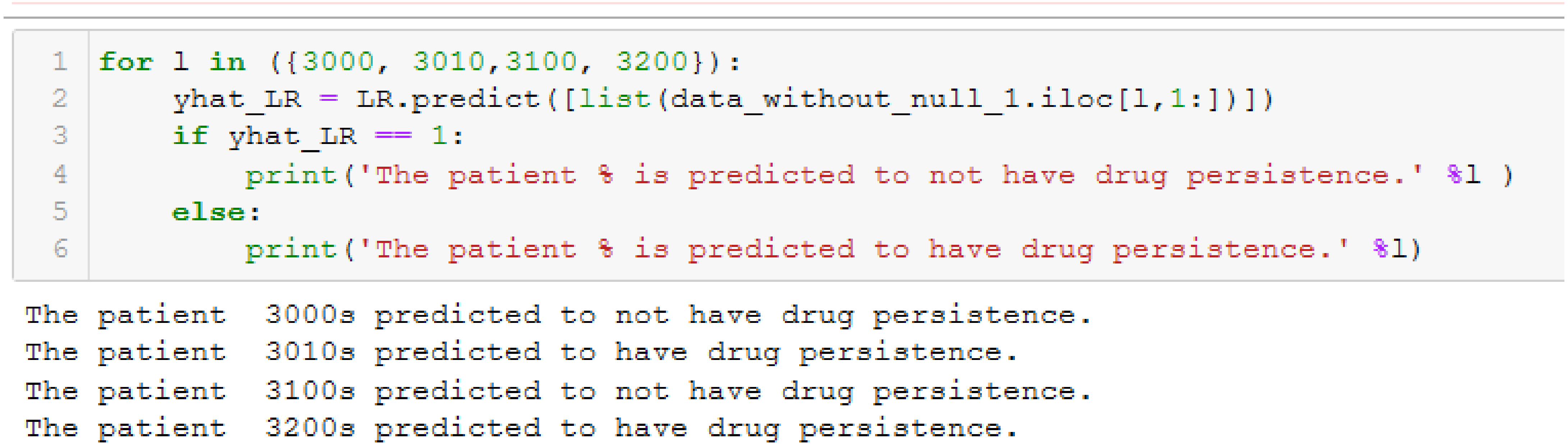
Logistic Regression



```
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
```

Logistic Regression



```
for l in ({3000, 3010,3100, 3200}):
    yhat_LR = LR.predict([list(data_without_null_1.iloc[l,1:])])
    if yhat_LR == 1:
        print('The patient % is predicted to not have drug persistence.' %l )
    else:
        print('The patient % is predicted to have drug persistence.' %l)
```

```
The patient  3000s predicted to not have drug persistence.
The patient  3010s predicted to have drug persistence.
The patient  3100s predicted to not have drug persistence.
The patient  3200s predicted to have drug persistence.
```

Recommendation to technical users

Also, other machine learning methods can be recommended to professional users. For example, support vector machine.

The support vector machine is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to. It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces. The only disadvantage with the support vector machine is that the algorithm does not directly provide probability estimates.

Thank You