University of Science and Technology of Hanoi

Classification II Report

Machine learning and Data mining II

Nguyễn Mạnh Hưng - 22BI13183

Nguyễn Trọng Minh - 22BI13304

University of Science and Technology of Hanoi

TABLE OF CONTENT

I. Decision Tree

1.Analyze the dataset

In this lab work, we chose 2 data sets both that are well-classified. The first data is used to predict whether one individual has heart problem or not. This data set is binary classified.

Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
67	0	3	115	564	0	2	160	0	1.6	2	0	7	Absence
57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence

Data about heart disease

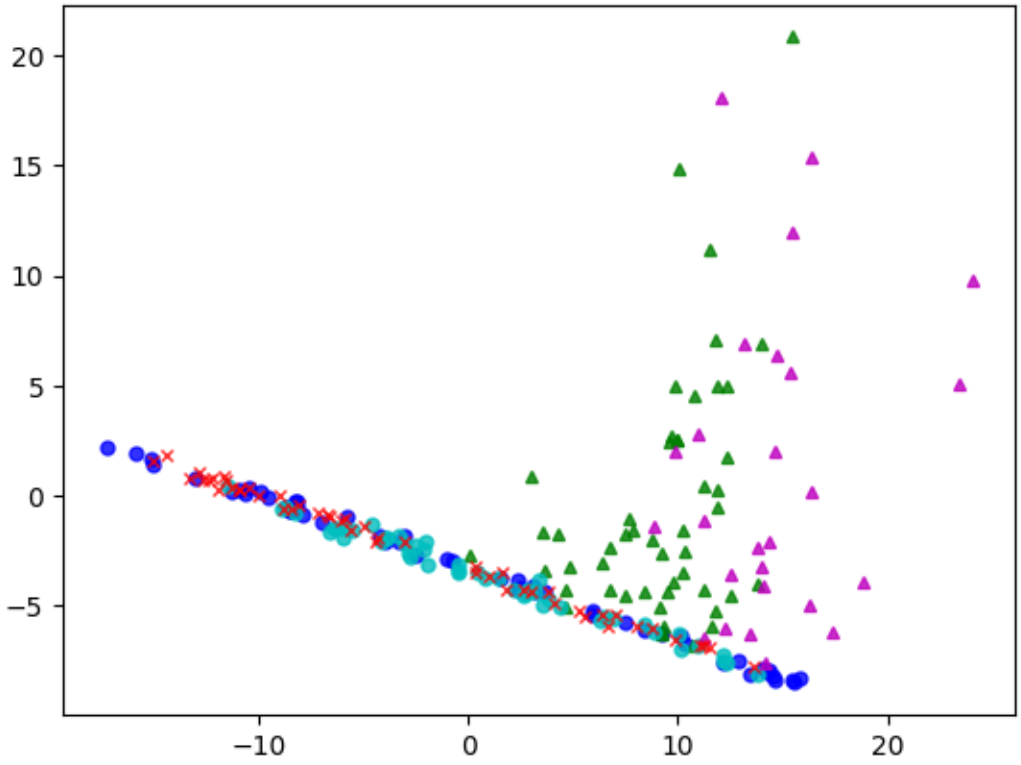
The other one is about predicting weather with 4 features, divided into 5 weather types.

date	precipitation	temp_max	temp_min	wind	weather
2012-01-01	0.0	12.8	5.0	4.7	drizzle
2012-01-02	10.9	10.6	2.8	4.5	rain
2012-01-03	0.8	11.7	7.2	2.3	rain
2012-01-04	20.3	12.2	5.6	4.7	rain
2012-01-05	1.3	8.9	2.8	6.1	rain
2012-01-06	2.5	4.4	2.2	2.2	rain
2012-01-07	0.0	7.2	2.8	2.3	rain
2012-01-08	0.0	10.0	2.8	2.0	sun
2012-01-09	4.3	9.4	5.0	3.4	rain

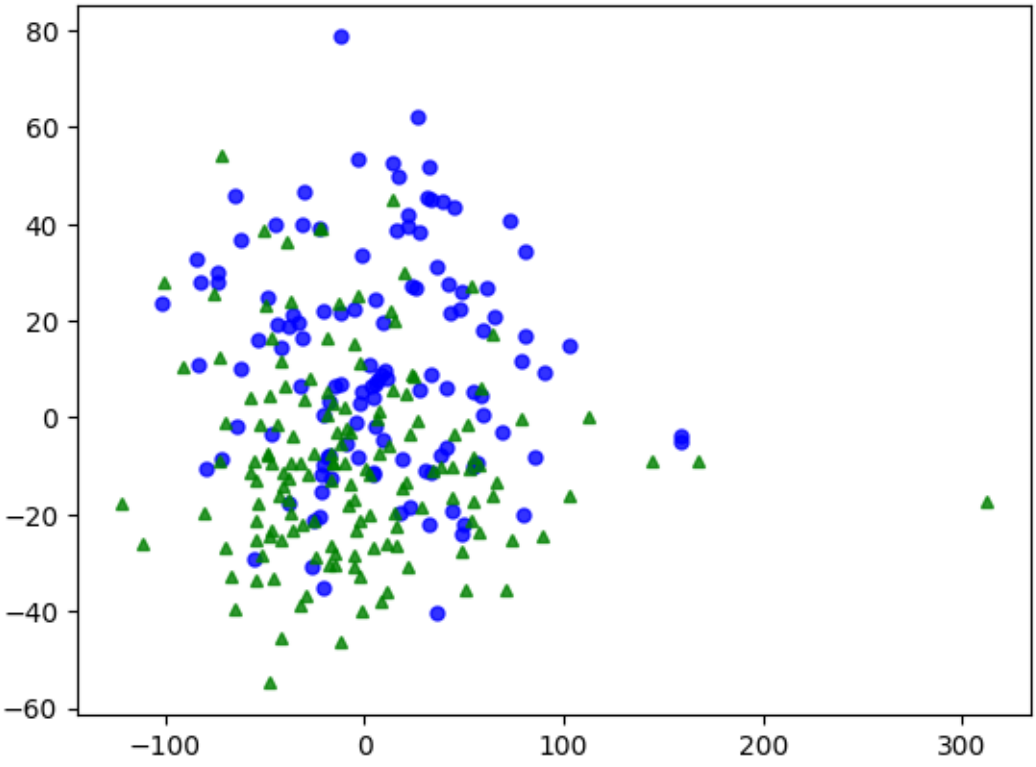
Data about weather

Both data sets are well-maintained so we actually do not need to clean the data much. We drop only data column in weather data set since it is just a time stamp. After that, we split data into training and testing set with 8-2 ratio.

If we take a look at the 2D plotted data, we can see that they are completely mess, and we cannot apply any linear classification. This is a good case of using tree classification.



2D graph of weather dataset



2D graph of heart disease dataset

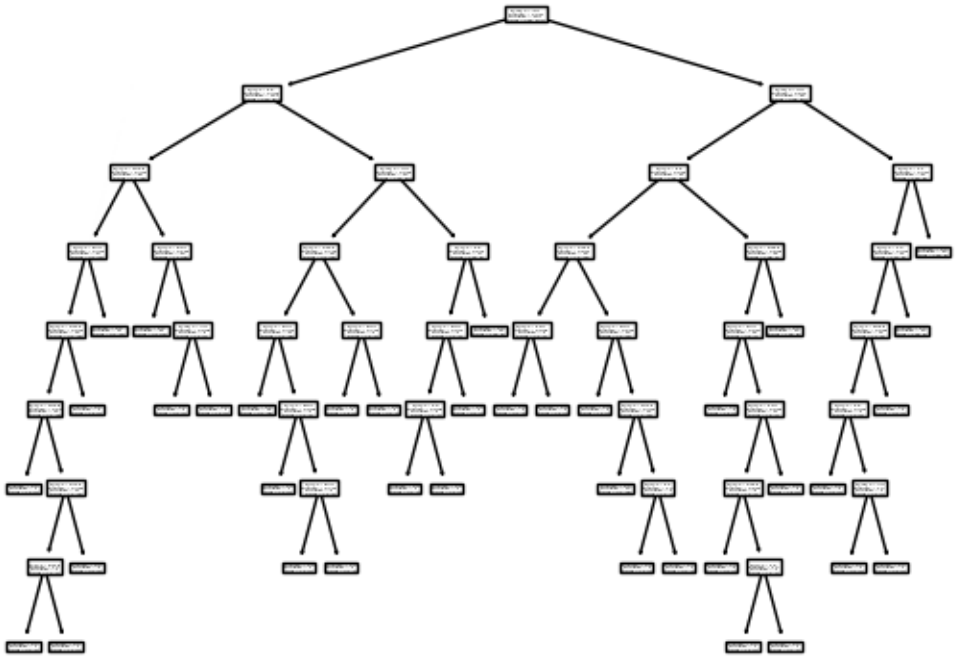
2.Decision tree

We start making tree with sklearn module. There are 3 criterions we can choose: Gini loss, entropy and log loss. Basically, entropy and log loss are the same, so we use entropy formular as our criterion to determine which value to split data into.

For the weather data, we came up with the root case is if precipitation is smaller than 0.15, since most of this feature is 0. After all, we got a tree with depth of 24 layers, and there are 222 leaves nodes.

Tree graph of weather data classification

With the heart disease data, there are much less data points, so that we got a smaller tree than the one above. The root case in this data set is if thallium level of patient is less than 4.5 . At the end, we got a tree with depth of 8 layers and 37 leaves nodes.

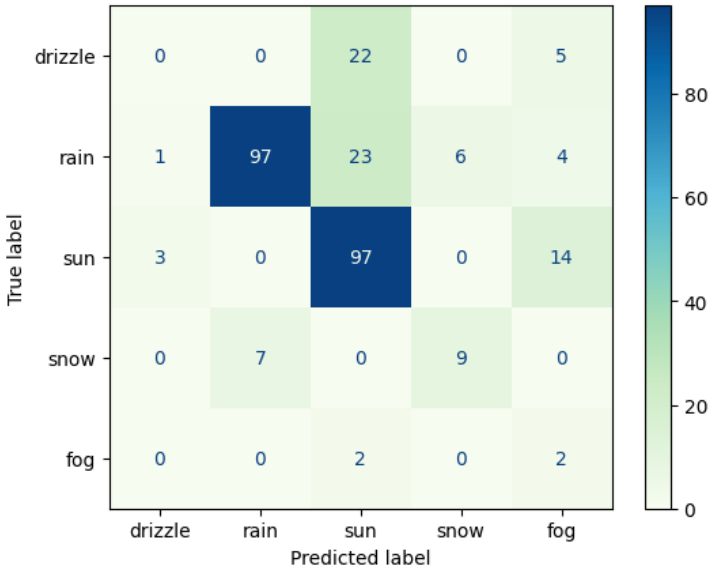


Tree graph of heart diseases data classification

3. Error calculation

To validate our tree classifier model, we can take a look at accuracy by compare predicted labels with true labels in testing data. This is a common way to check the error in supervised learning.

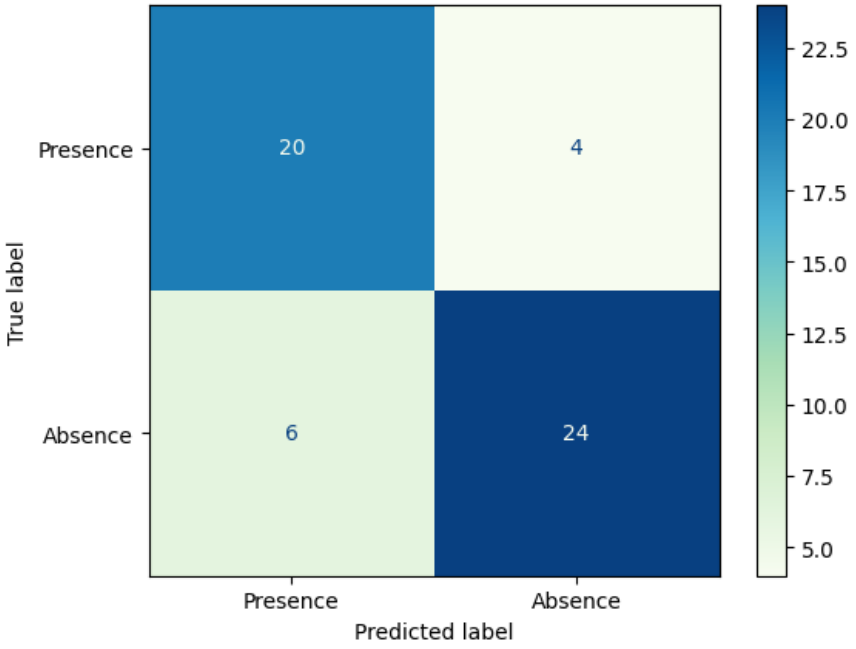
Fortunately, both of our models have quite a good result in prediction. The weather tree classifier has mean accuracy of 70,2% with the following confusion matrix.



Confusion matrix of weather prediction

We notice that most of the correctly predicted are rain and sun categories. This is because the majority of the data set are rain and sun categories, which led to poor accuracy in the rest of the data set.

On the other hand, the heart diseases data set contains only 2 categories and is more equally distributed (120 of absence and 150 of presence), it got a better accuracy of 81.5% mean accuracy.



Confusion matrix of heart disease prediction

II. Random Forest

1. Analyze the data

For the second part, we still chose the same datasets as the previous part, which are the weather dataset with the classification of 5 types and the heart diseases dataset to predict whether a person has heart diseases or not.

2. Random Forest

Before starting to train the model, we used cross-validation to create 100 training sets with Kfold and 1 testing set. Here, we chose 20% of the dataset will be the testing set.

We start making the Random Forest model. First of all, we built 100 Random Forest models with 100 trees and used the entropy formular as the criterion, then we started to train each of them with each training set. The next step is that we use the testing set to calculate the error of each Random Forest model by taking 1 minus the accuracy score of each model and putting it into a list called “Errors”. Finally, we convert the list into an array and calculate the mean error of the model.

- For the weather dataset: the mean error calculated is 18.17%
- For the heart disease dataset: the mean error calculated is 15.5%

3.Comparison

Regarding the weather dataset, we can see that the accuracy of the Decision Tree on the training set is 70.2% which means that the error calculated is 29.8% while the Random Forest model has the error calculated as 18.17%. As a result, we can see that the Random Forest works more accurately in classifying the dataset.

About the heart disease dataset, the accuracy of the Decision tree is 81.5% and the accuracy of the Random Forest is 84.5%, which means that the Random Forest model still works more accurately in classifying the dataset.