Covid-19 Classification

Do you have Covid-19 ?

Let's find out!

Course:

Machine Learning

Masters:

Dr.Farzin Yaghmaee,
Mr.Amir Shokri

Student:

Havva askari

Student.ID:

40011920006

Havvaaskari0702@gmail.com

Abstract

This is a Covid-19 Data Set, collecting by student of semnan university. It contains 23 attributes including id. The response is to guess if a person has covid-19 or not.

1. Data

Attribute Information:

1 #:id
2 age
3 Sleep_problems
4 Headache
5 Diarrhea
6 Abdominal_pain
7 body_pain
8 Body_discoloration
9 Cough
10 Fever
11 Ague
12 Sore_throat
13 Fatigue
14 runny_nose
15 Chest_pain
16 Decreased_appetite
17 Vomit
18 Nausea
19 Sneezing
20 Shortness_of_breath
21 Loss_of_smell
22 Loss_of_taste
23 Urticaria

2. Features

This is a binary DataSet Regardless # , age. In this work we needed to change the values (yes,no) to (1,0) which is works for som algorithms, but some others neede to have int value so I changed them back again.

For age attribute because we didn't have some values, so I droped the feature. And also I didn't need the id number so I droped it either!

3. Models

The documentation uses the Find-s, C-E, Naïve Bayes , KNN algorithm, Kmeans clustering, Decision Tree, ID3; details are as follows:

3.1. Find-s

For this algorithm the final hypothesis is: ['?' '?']

Which is not a good one, because it takes every instance as yes!

3.2. Condidate Elimination

For this algorithm Final Specific_h is:

['?' '?']

Final General_h is:

[['?', '?']...] multiplied by 21 . one hypothesis for each feature. That also not a good hypothesis.

3.3. Naïve Bayes

Naive Bayes Algorithm Test Accuracy: 0.902834008097166

We put this dataset in rapidminer and result was a little different, the accuracy of naïve bayes in rapidminer was 89.7% that is less then ours!

3.4. K-NN

K-NN Algorithm Train Accuracy For k = 3 Is : 0.9378378378378378

K-NN Algorithm Test Accuracy For k = 3 Is : 0.9271255060728745

And we did this with rapidminer and result was :

Performance vector:

Root_mean_squared_error: 0.363 +/- 0.042(micro avarage:0.366 +/- 0.000)

Squared_error: 0.134 +/- 0.031 (micro avarage:0.134 +/- 0.028)

3.5. Kmeans

KMeans Algorithm Test Accuracy: -1016.7373300215618

And the result in rapidminer was:

Cluster 0 : 502 item

Cluster 1 : 283 item

Total number of items 785

That real number of 1 is 286!

That means accuracy is really high: 0.9961783439

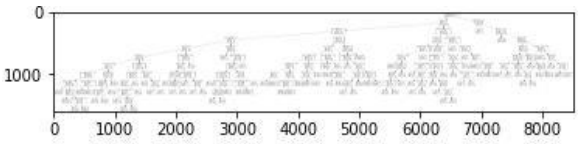
3.6. Decision Tree

We use two following decision tree:

3.6.1. Random

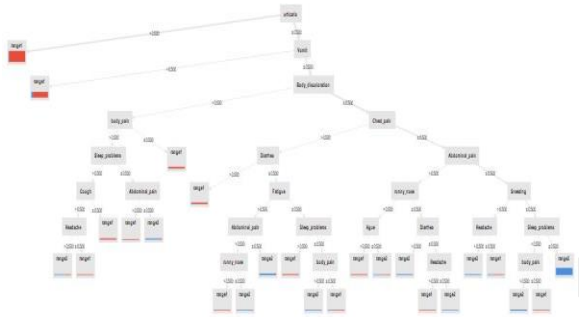
Decision Tree Algorithm Test Accuracy: 0.8987854251012146

And the accuracy in rapidminer was : 90.6% that is better then ours!



3.6.2. ID3

The result here was a tree that is:



The red and blue color shows the gain of each node.

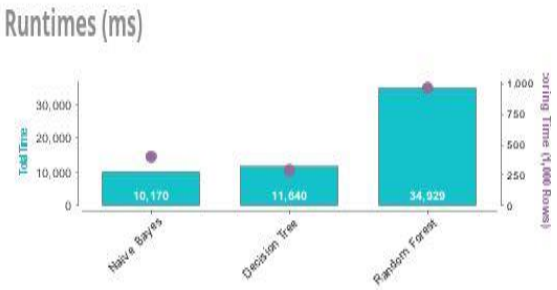
4. Result

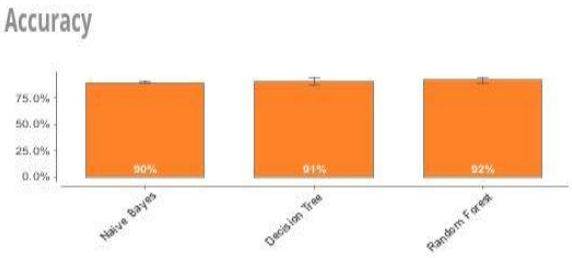
According to the results, the highest accuracy is 0.9961783439which is based on the KMeans algorithm.

Summary: You can see the summary in Table 1 that the Decision Tree algorithm is weak accuracy.

Naïve Bayes	0.902834008097166
K-NN	0.9271255060728745
KMeans	0.9961783439
Decision Tree	0.8987854251012146

Table 1. summary of accuracy





Model	Accuracy	Standard Deviation	Gains	Total Time
Naive Bayes	89.7%	± 1.2%	116	10 s
Decision Tree	90.6%	± 2.9%	122	12 s
Random Forest	92.4%	± 2.5%	128	35 s

5. Dicussion

This exercise was a great help to knowing better the classification and clustering algorithm. It make me to go for more learning in ML and python and rapidminer.

We found out if we change the input values, validation percentage and algorithm it'll change the result too.

6. References

https://www.kaggle.com/amirshnll/covid-patient-datasets

https://scikit-learn.org/stable/modules/classes.html?highlight=sklearn#module-sklearn.model_selection

https://academy.rapidminer.com/learning-paths/get-started-with-rapidminer-and-machine-learning

My project in kaggle:

https://www.kaggle.com/eveaskari/covid-project-eve

https://www.kaggle.com/eveaskari/find-s

https://www.kaggle.com/eveaskari/candidate-elimination

https://www.kaggle.com/eveaskari/id3-algorithm