Problem: Flower

Link to the GitHub page:

https://github.com/samuelppd/DataScience

Description of the algorithm:

The algorithm is quite easy.

At the beginning, I import all the libraries I use in the project.

Then I load the CSV file into a DataFrame named df.

After that, I explore, and I analyse the data. I didn't had to clean the data because the dataset isn't complex.

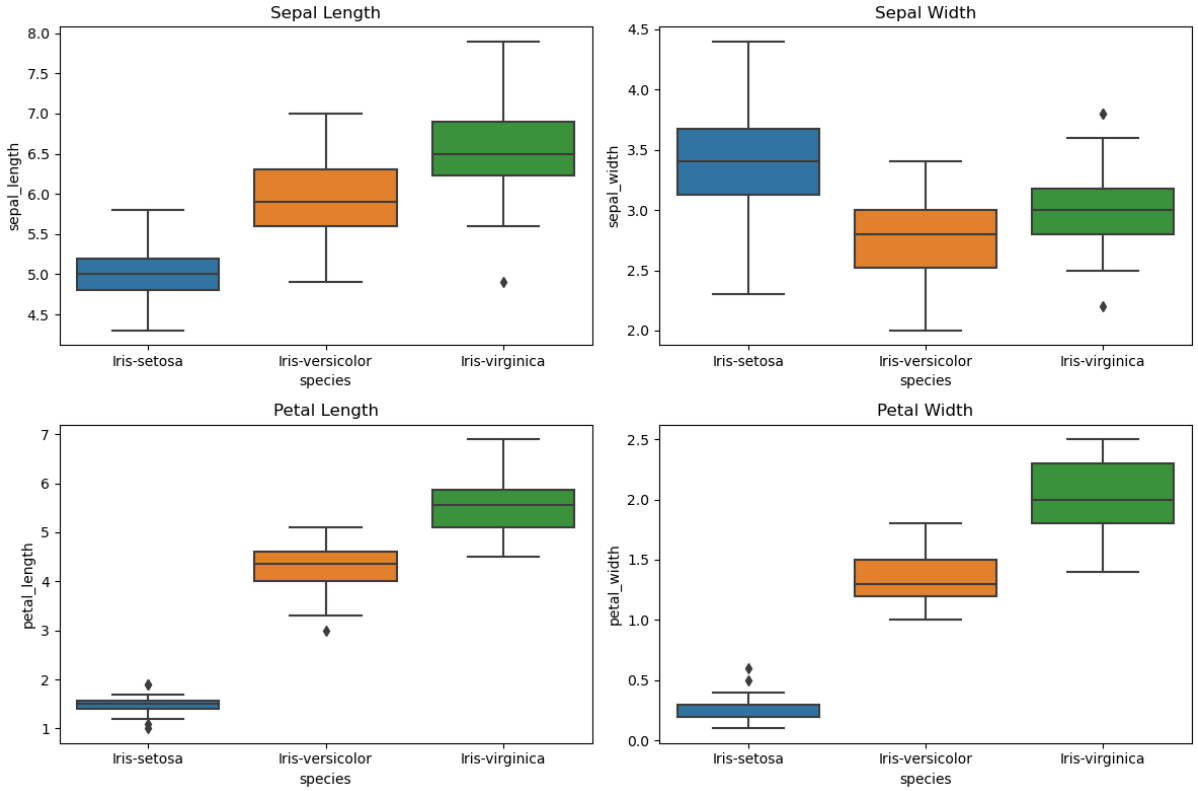
I made a representation of the four values in relation to the species in four graphs, by using boxplots.

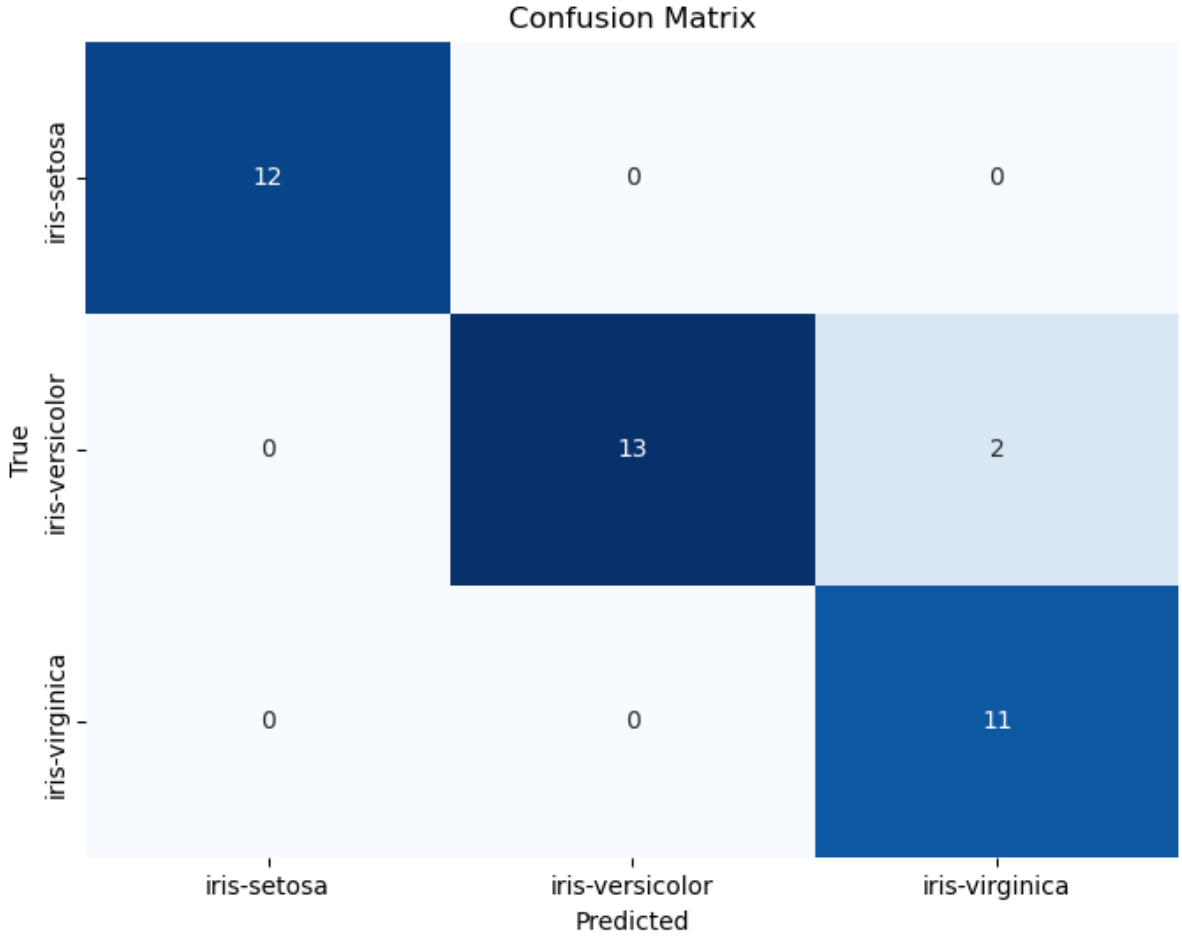
The next step is the beginning of the random forest test. I started by separate the values. I used 25% of the data to train the algorithm and 75% to test it.

Then I created and trained the model using RandomForestClassifier and fit.

Finally, I had the results of the random forest test. I viewed the importance of the different values, the accuracy of the model, and the confusion matrix with the associated graph.

Graphs:


Sepal Length
sepal_length
Iris-setosa
Iris-versicolor
Iris-virginica
species
Sepal Width
sepal_width
Petal Length
petal_length
Petal Width
petal_width


Confusion Matrix
iris-setosa
iris-versicolor
iris-virginica
True
12
0
0
0
13
2
0
0
11
iris-setosa
iris-versicolor
iris-virginica
Predicted

Problem: Credit

Link to the GitHub page:

https://github.com/samuelppd/DataScience

Description of the algorithm:

This algorithm is more complex.

At the beginning, I import all the libraries I use in the project.

Then I load the first CSV file into a DataFrame named df1. I removed the column OCCUPATION_TYPE because there was missing data. I also converted the columns DAYS_BIRTH and DAYS_EMPLOYED into AGE and YEARS_EMPLOYED.

I loaded the second CSV file into a DataFrame named df2. I removed all the values "X," and for each client, I associated the percentage of approved credits. To improve the model, I rounded this percentage to 0%, 25%, 50%, 75%, and 100%.

With the two datasets that had ID as the index, I merged them into a single dataset named df.

After that, I explored and analyzed the data. I created a representation of some values in relation to the CREDIT_RATE in four graphs, using different types of plots.

The next step is the beginning of the random forest test. I started by separating the values. I used 25% of the data to train the algorithm and 75% to test it.

Then I created and trained the model using RandomForestClassifier and fit.

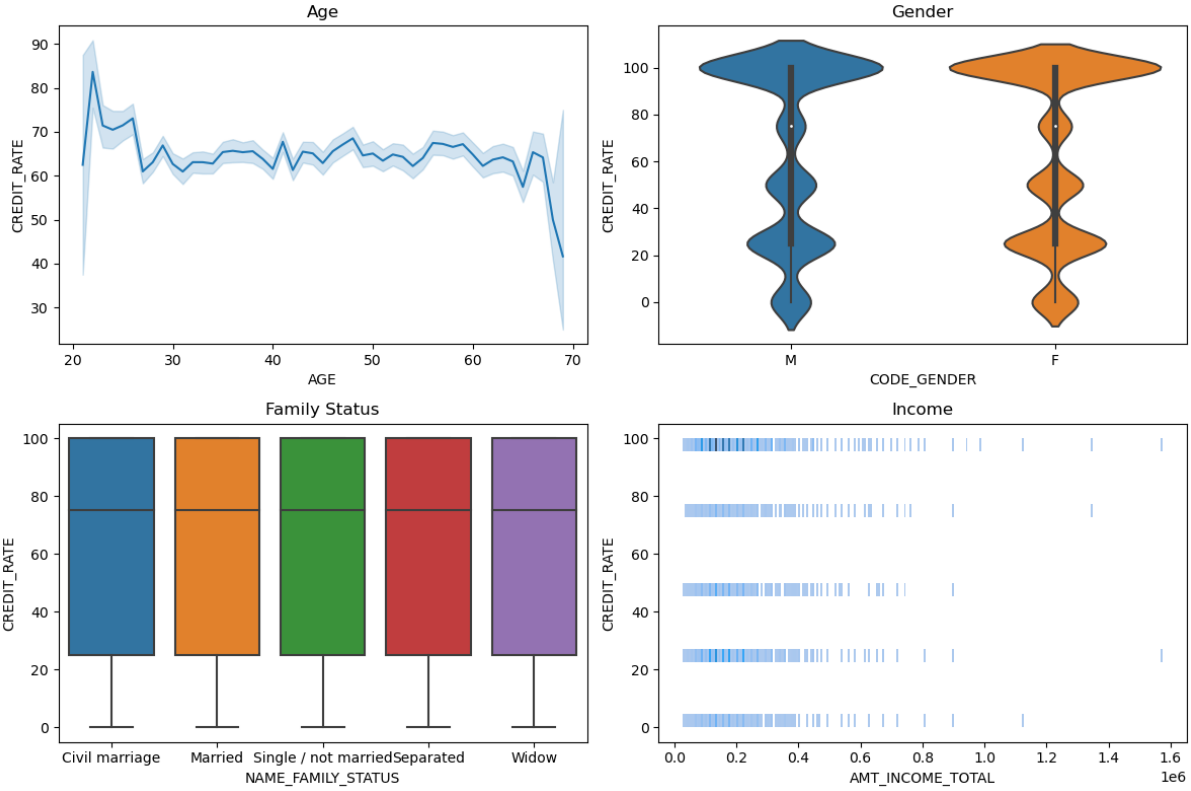
Finally, I obtained the results of the random forest test. I examined the importance of the different values, the accuracy of the model, and the confusion matrix with the associated graph.

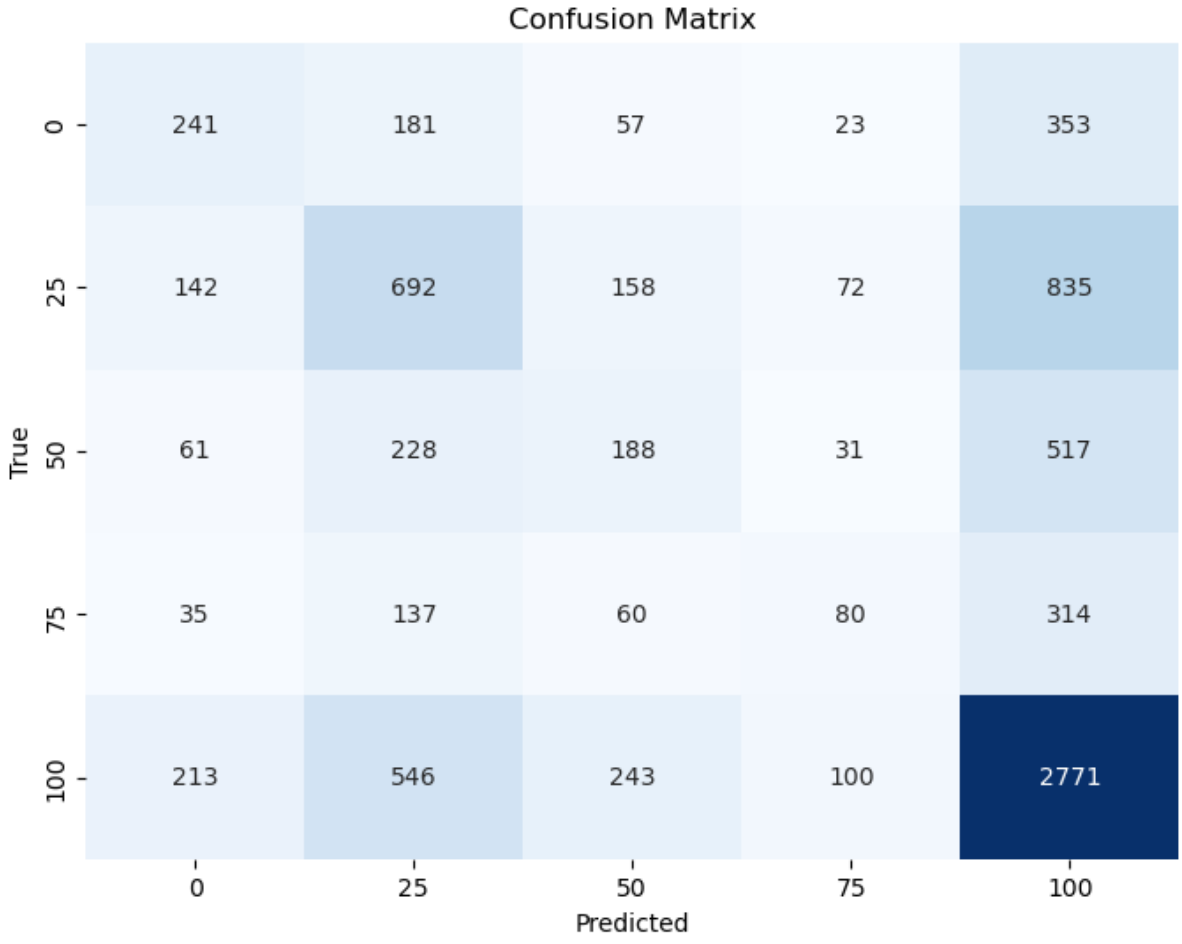
The accuracy of this model is not as great as the previous one, perhaps because the data I chose are not as representative as those for the flowers. Another possibility is that the percentage of credit accepted is more complex to train and contains more options than for the flower species.

Graphs:

Correlation:

	CNT_CHILDREN	AMT_INCOME_TOTAL	FLAG_MOBIL	FLAG_WORK_PHONE	FLAG_PHONE	FLAG_EMAIL
CNT_CHILDREN	1.000000	0.037617	nan	0.048964	-0.016430	0.020126
AMT_INCOME_TOTAL	0.037617	1.000000	nan	-0.033096	0.018423	0.092916
FLAG_MOBIL	nan	nan	nan	nan	nan	nan
FLAG_WORK_PHONE	0.048964	-0.033096	nan	1.000000	0.314704	-0.032959
FLAG_PHONE	-0.016430	0.018423	nan	0.314704	1.000000	0.013613
FLAG_EMAIL	0.020126	0.092916	nan	-0.032959	0.013613	1.000000
CNT_FAM_MEMBERS	0.888412	0.025476	nan	0.063232	-0.004784	0.018313
AGE	-0.340715	-0.070278	nan	-0.178377	0.025158	-0.106743
YEARS_EMPLOYED	0.231205	0.166273	nan	0.243625	0.010000	0.086152
CREDIT_RATE	0.002155	0.022897	nan	-0.009390	0.012946	-0.001036


Age
CREDIT_RATE
AGE
Gender
CREDIT_RATE
CODE_GENDER
M
F
Family Status
CREDIT_RATE
Civil marriage
Married
Single / not married
Separated
Widow
NAME_FAMILY_STATUS
Income
CREDIT_RATE
AMT_INCOME_TOTAL
1e6


Confusion Matrix
True
Predicted
0
25
50
75
100
241 181 57 23 353
142 692 158 72 835
61 228 188 31 517
35 137 60 80 314
213 546 243 100 2771