PROJECT BUSINESS
INTELLIGENCE

PAYMENT METHODS

18, rue de l'Usine - ZI Aéroport
Charguia II - 2035 Ariana
Tél. : +216 71 941 541 (LG)
Fax. : +216 71 941 889
e-mail : contact@esprit.ens.tn
www.esprit.ens.tn

Academic Year:2022/ 2023

Elaborated By: Eya Goutet
Maryam Gadri
Ayoub Mabrouk
Nour elhouda Makkari
Syrine Baklouti
Yassine Mestiri

TABLE OF CONTENTS

TABLE OF FIGURES

figure 3 :CAH_result

INTRODUCTION

In the context of the 4th year of studying Business Intelligence, students are presented with diverse real-world subjects and associated problems that they may encounter in their future professional careers. These problems serve as the focus of group work, which is supervised by a professional and multiple teachers. The purpose of this group work is to assess the students' ability to collaborate as a team on a professional subject that may be unfamiliar to them, while utilizing the knowledge and skills acquired during their four years of study at ESPRIT.

The subject on which our group has been working is commissioned by a company which is a technology solutions provider that works with banks to enhance their operational efficiency and customer experience. Its solutions include payment management, customer data management and customer relationship management, as well as IT support and infrastructure services.
We are interested in the processing of means of payment, specifically checks, since they require preci1.

CHAPTER I :
BUSINESS UNDERSTANDING

1. BUSINESS UNDERSTANDING

Description

Our company is a leading technology solutions provider that specializes in offering innovative solutions to the banking industry. With a strong focus on improving operational efficiency and enhancing customer experience, Sibtel collaborates with banks to deliver cutting-edge solutions that streamline their processes and drive growth.

It offers a wide range of technological solutions, including software applications, data analytics, business intelligence tools, and consulting services. These solutions are designed to help banks optimize their operations, reduce costs, and enhance their competitive advantage in the ever-evolving banking landscape.

2. BUSINESS OBJECTIVE

01. Divide payments into groups according to periods

- DMO (1): Segment results into k homogeneous groups.

02. Analyze bank payments

- DMO(1): Segment results into uniform groups.
- DMO(2): Calculate the Sum of payment for each bank in both cases.

03. Identify banks with the largest transactions

- DMO (1): Target banks with maximum foreign exchange.

04. Reducing the risk of fraud and financial loss by identifying high-risk banks/agencies.

- DMO (1) : Identifying unusual or suspicious transactions that could indicate fraud or fraudulent behavior.

05. Develop a predictive model that accurately estimates the value of a variable based on a given set of variables

- DMO (1) : Performing a comparative analysis of two predictive algorithms to determine their effectiveness.

CHAPTER II:
DATA UNDERSTANDING

3. DATA UNDERSTANDING

We have at our disposal a dataset of 5,830,771 transactions arranged in 13 descriptive variables:

- **9 qualitative variables :**

a- RIBBeneficaire : this is an encrypted field presenting the RIB of each beneficiary bank b-BanqueBeneficaire : an encrypted field with the name of each beneficiary bank.
b- Agency Beneficiaire : an encrypted field with the name of each recipient agency.
c- RIBPayeur : this is an encrypted field presenting the RIB of each payer (customer or bank)
d- BanquePayeur : This is an encrypted field with the name of each Payor bank.
e- AgencePayeur : This is an encrypted field with the name of each Payor Agency.
f- RejetCCE : this is a coded field presenting ....(we are missing data)
g- MotifRejet : this is a coded field with the reason for rejection.
h- Etat : this is a coded field with the status of each transaction (accepted or rejected)

- **4 quantitative variables :**

a- idTypelot
b- NuméroLigne
c- Montant which varies between 660626625608 and 1021254566206
d-Date from 01/12/2022 to 30/12/2022

CHAPTER III:
DATA ANALYTICS

3. DATA ANALYTICS

- Profiling report :

The ProfileReport (dataSet_Name) command returns an interactive report on the dataset containing for example:
The number of rows, columns, number of duplicate rows, percentage of fields missing, Interactions between variables etc...

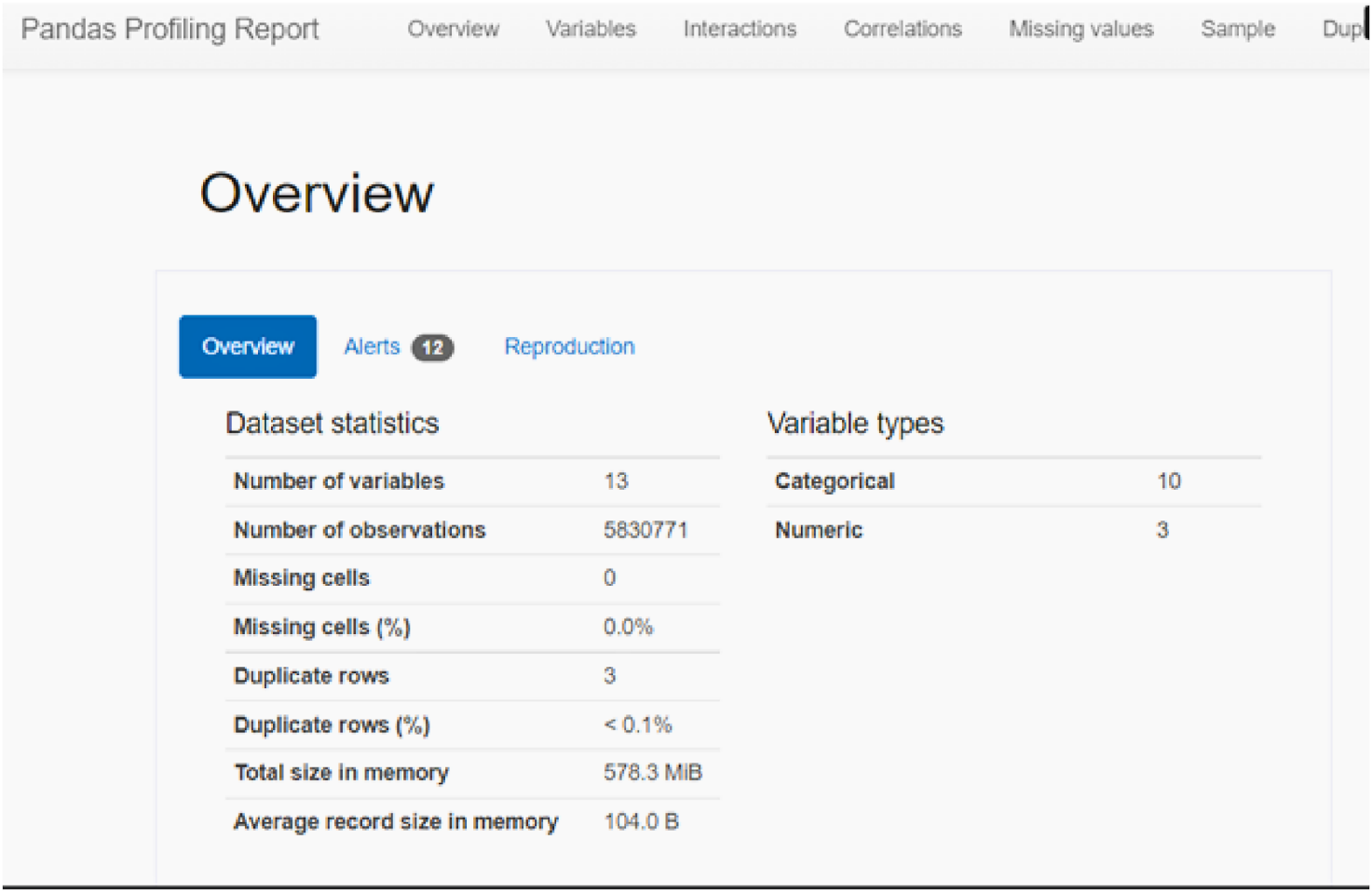

figure 1 : ProfileReport

=> 3 duplicate lines corresponding to 3 payments with the same informations.

We will eliminate duplicate rows.
payment_data = payment_data.drop_duplicates()

It is useful to understand the data still it doesn’t offer many features. Pandas profiling is the solution to this problem. It offers report generation for the dataset with lots of features and customizations for the report generated. It has advanced use cases and integrations that can prove useful to create stunning reports out of the data frames! It delivers an extended analysis of a DataFrame.

- CAH Results:

figure 2 :CAH_with_materialization_of_the_5_species

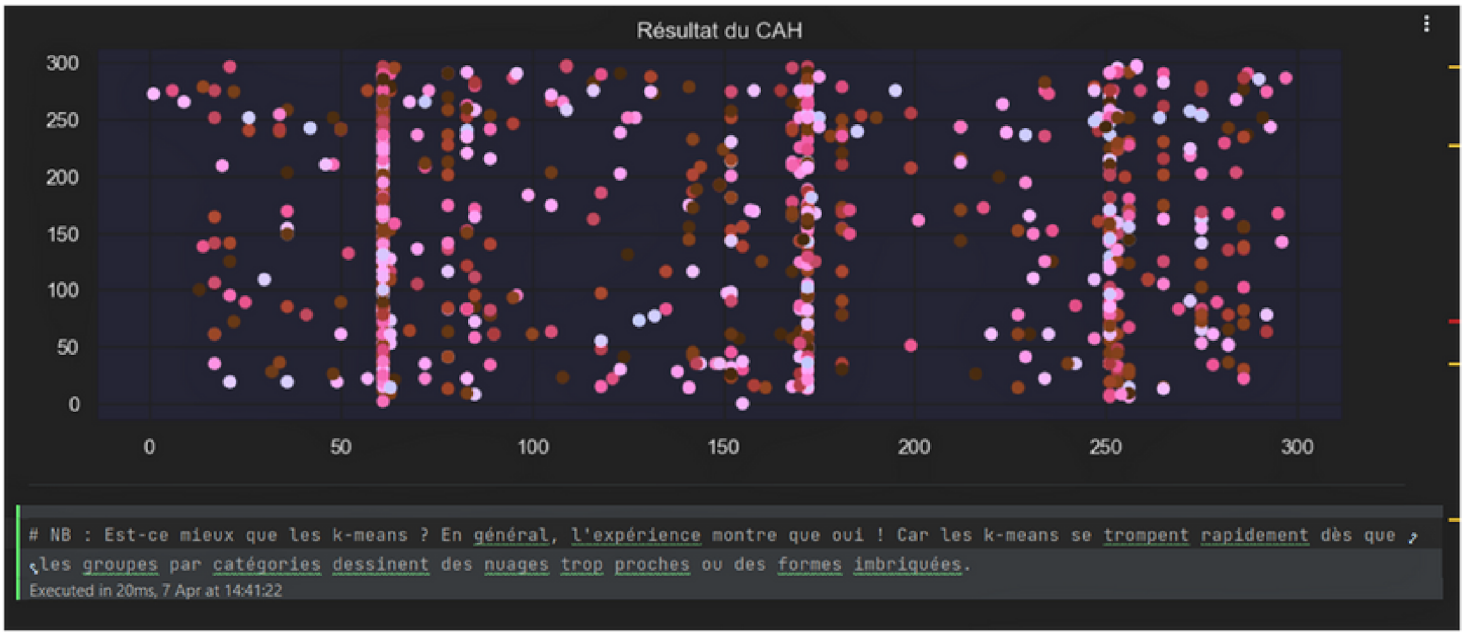

figure 3 :CAH_result

ne pas mettre

Here is the associated dendrogram.it was created with a final subdivision of 5 groups. If you cut the dendrogram above, the final groups would be fewer, but the level of similarity would be reduced. If you cut the dendrogram lower, the level of similarity would be higher, but the final groups would be more numerous.

- Payments per day :

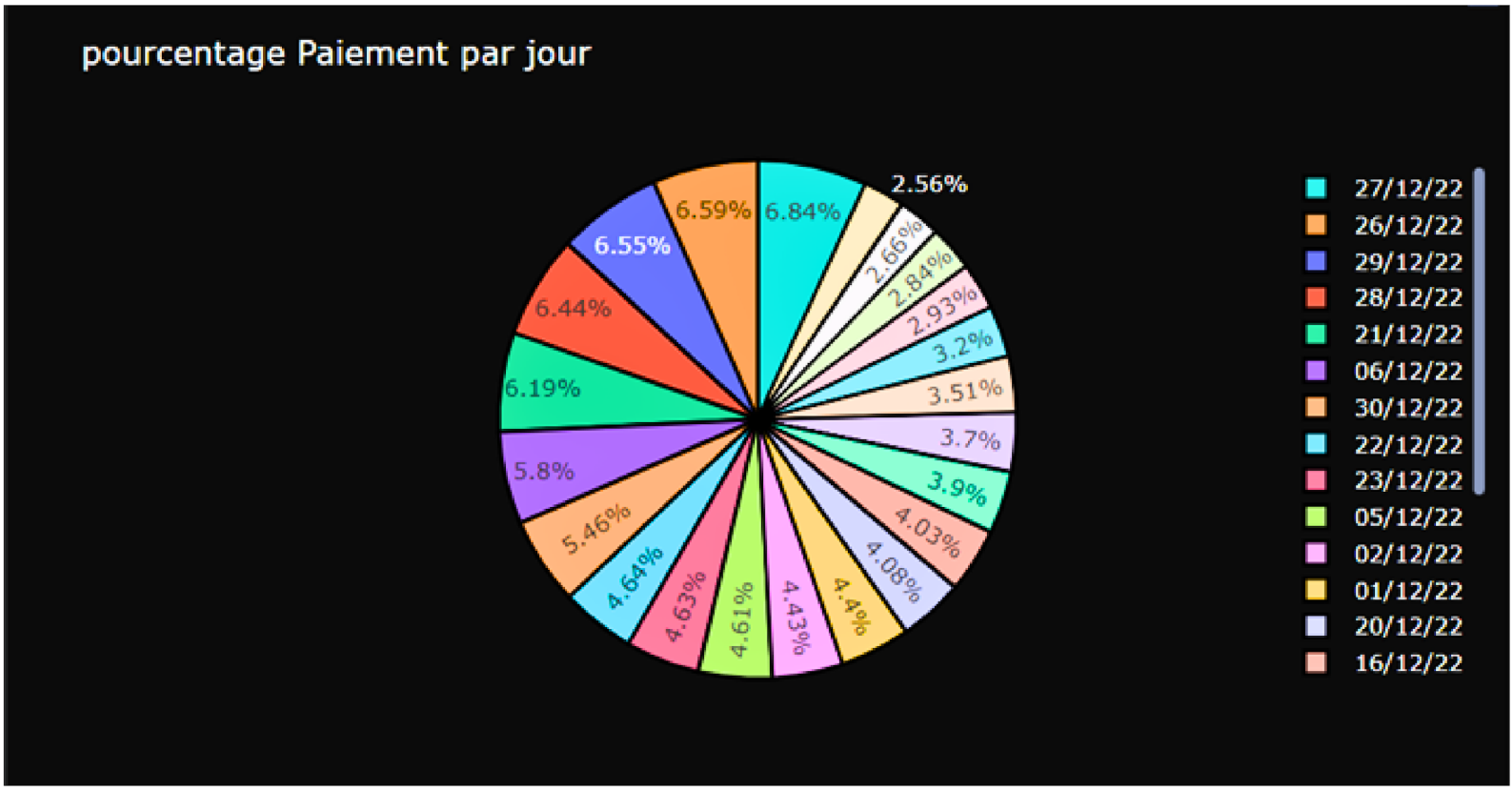

figure 4 : Payments_per_day

We noticed that the payments done in the last period of the month (from 26 -> 30 december) represent particularly the quarter of the sum of all payments. This is due to the fact that in that period of the month , employee's salaries are being paid.

- Correlation matrix :

figure 5 : Correlation matrix

This matrix presents the correlation between variables .
each color represents a range of values .
for example the orange color represents a value between 0.6 and 0.8.

- Total payments received by bank :

figure 6 : Total_payments_received_by_bank

- Total payments paid by bank :

figure 7 : Total_payments_paid_by_bank

In these two figures , we noticed that the majority of banks that have a great sum of payed payments also have a great sum of received payments. For example BxBQm37... sum of received payments = 2.8T and the paid one = 2.9T which means great exchanges in general .

- Sum of amounts per day :

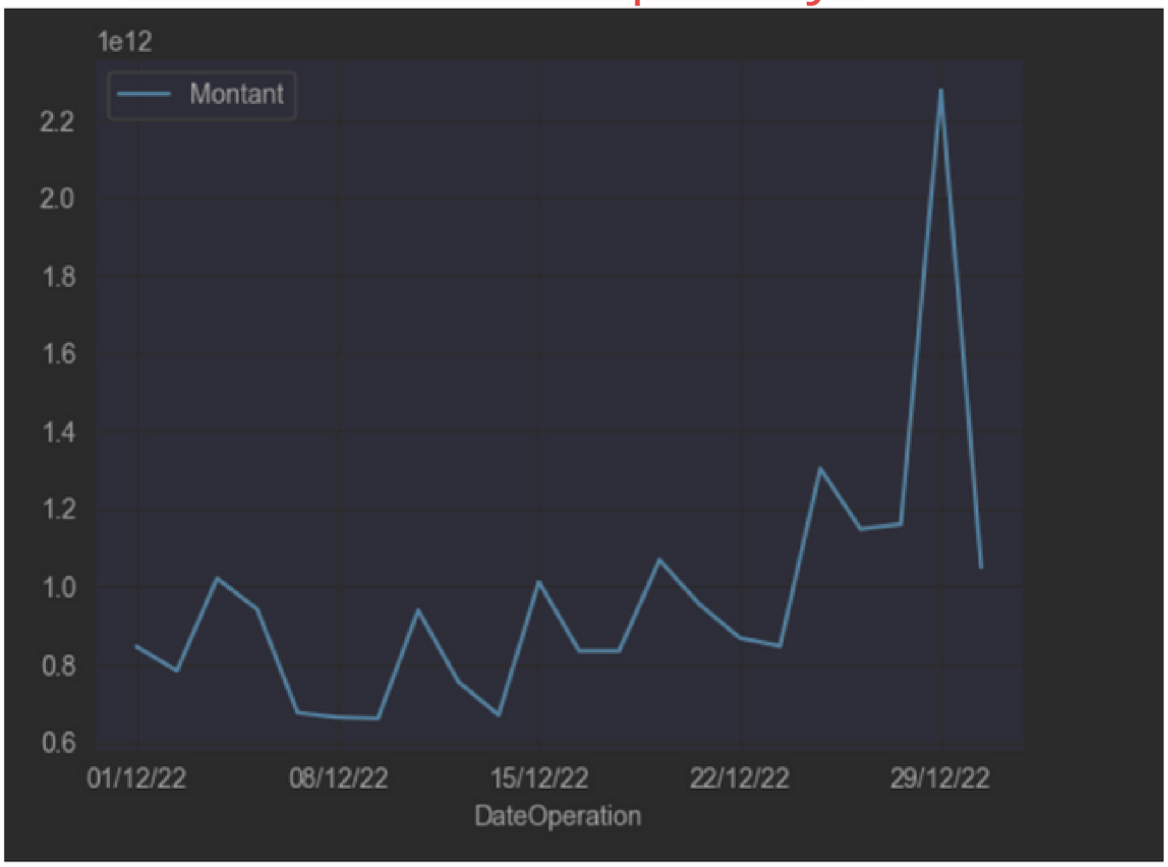

figure 8: Sum_of_amounts_per_day

the differnce between this chart and the chart in figure 2 , is its form . in fig 2 we presented the payment of each date, and in this one we presented the evolution by date of payments in a curve

- Handling categorical variables :

```
paiement_dataEncoded=paiement_data
labelencoder = LabelEncoder()

paiement_dataEncoded['RIBBeneficiaire'] = labelencoder.fit_transform(paiement_data['RIBBeneficiaire'])
paiement_dataEncoded['BanqueBeneficiaire'] = labelencoder.fit_transform(paiement_data['BanqueBeneficiaire'])
paiement_dataEncoded['AgenceBeneficiaire'] = labelencoder.fit_transform(paiement_data['AgenceBeneficiaire'])
paiement_dataEncoded['RIBPayeur'] = labelencoder.fit_transform(paiement_data['RIBPayeur'])
paiement_dataEncoded['BanquePayeur'] = labelencoder.fit_transform(paiement_data['BanquePayeur'])
paiement_dataEncoded['AgencePayeur'] = labelencoder.fit_transform(paiement_data['AgencePayeur'])
paiement_dataEncoded['DateOperation'] = labelencoder.fit_transform(paiement_data['DateOperation'])
paiement_dataEncoded['NumeroLigne'] = la
```

B:/4Bi6/PI/PythonTuto/PythonTutoProject/PaiementVersionJupyter2.ipynb

```
labelencoder: LabelEncoder = LabelEncoder()

paiement_dataEncoded.head(10)
```

Executed in 20s, 9 Apr at 23:54:21

Out 20 — 10 rows × 13 columns pd.Dataframe

	RIBBeneficiaire	BanqueBeneficiaire	AgenceBeneficiaire	RIBPayeur	BanquePayeur	AgencePaye
0	1598099	4	127	299966	19	
1	1788732	28	63	252782	5	
2	214432	28	208	1367	15	
3	673412	28	239	495416	3	
4	1311825	24	78	241852	18	
5	1261333	4	97	208081	22	
6	1460919	4	251	169078	5	
7	277358	25	247	226207	4	

figure 9 : Handling_categorical_variables(Label Encoding)

Label Encoding is a popular encoding technique for handling categorical variables. In this technique, each label is assigned a unique integer based on alphabetical ordering. -> ch 3

CHAPTER IV:
DATA PREPARATION

- Elbow Method :

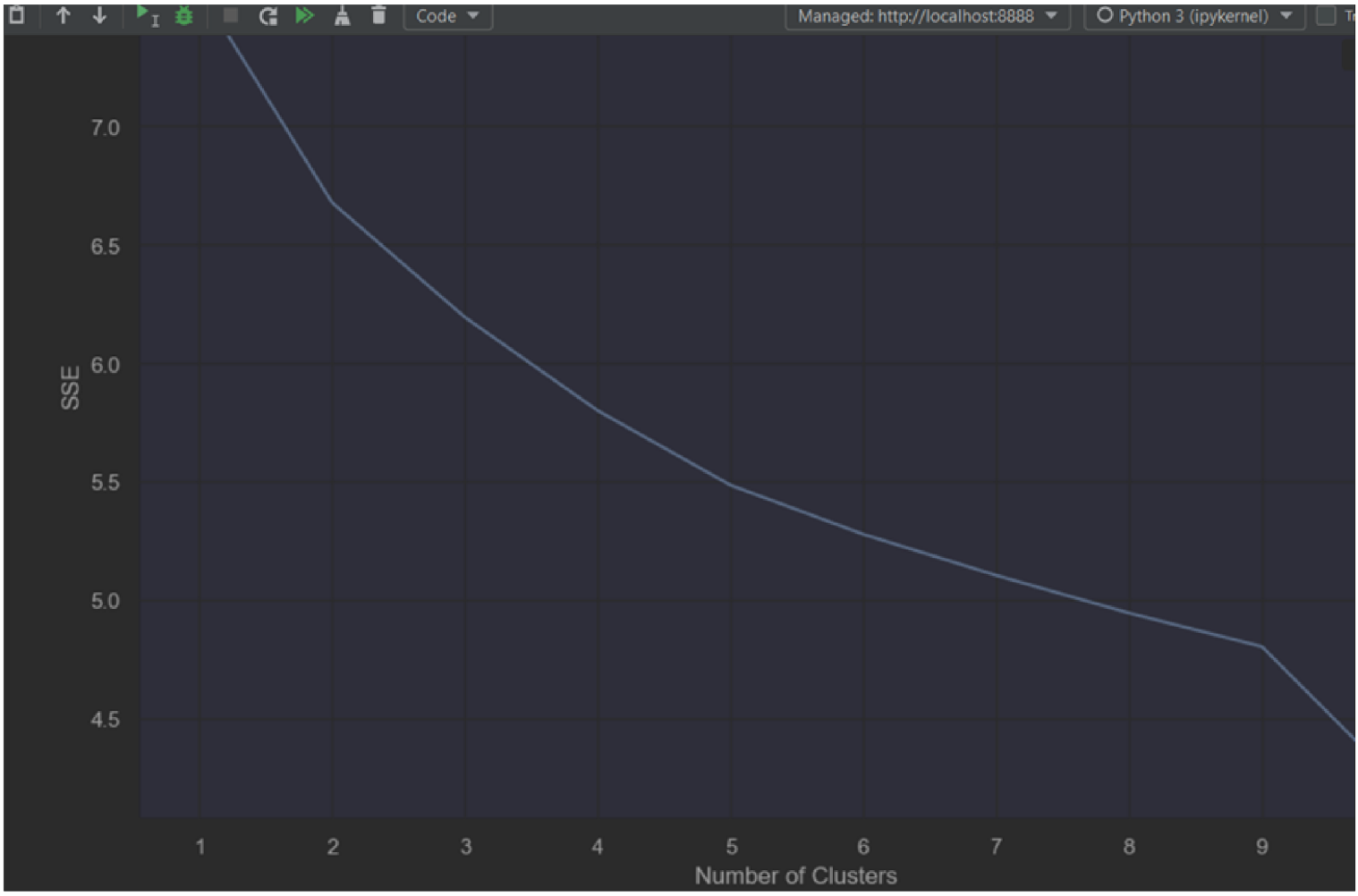

figure 10 : Elbow_Method

This method was utilized to simplify the identification of the ideal number of clusters.
we obtained k=5 from this graph.

- K-means Results :

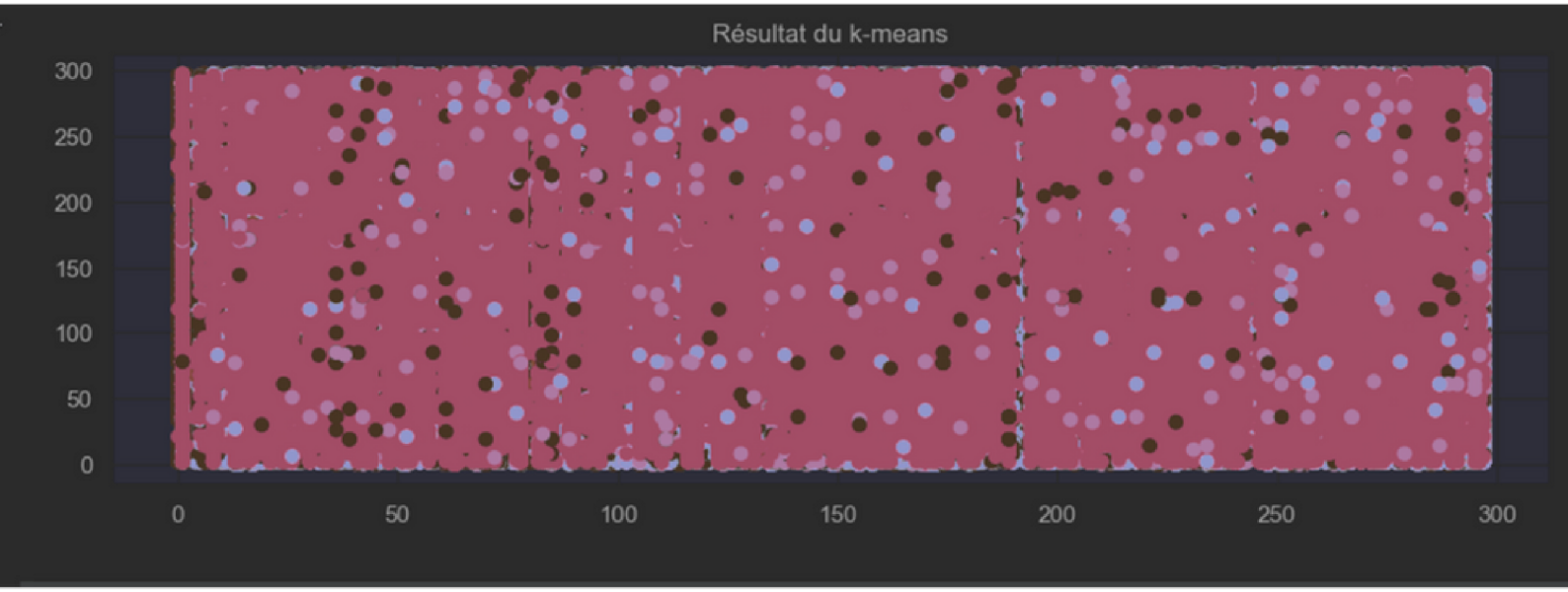

figure 11 : K-means_Result

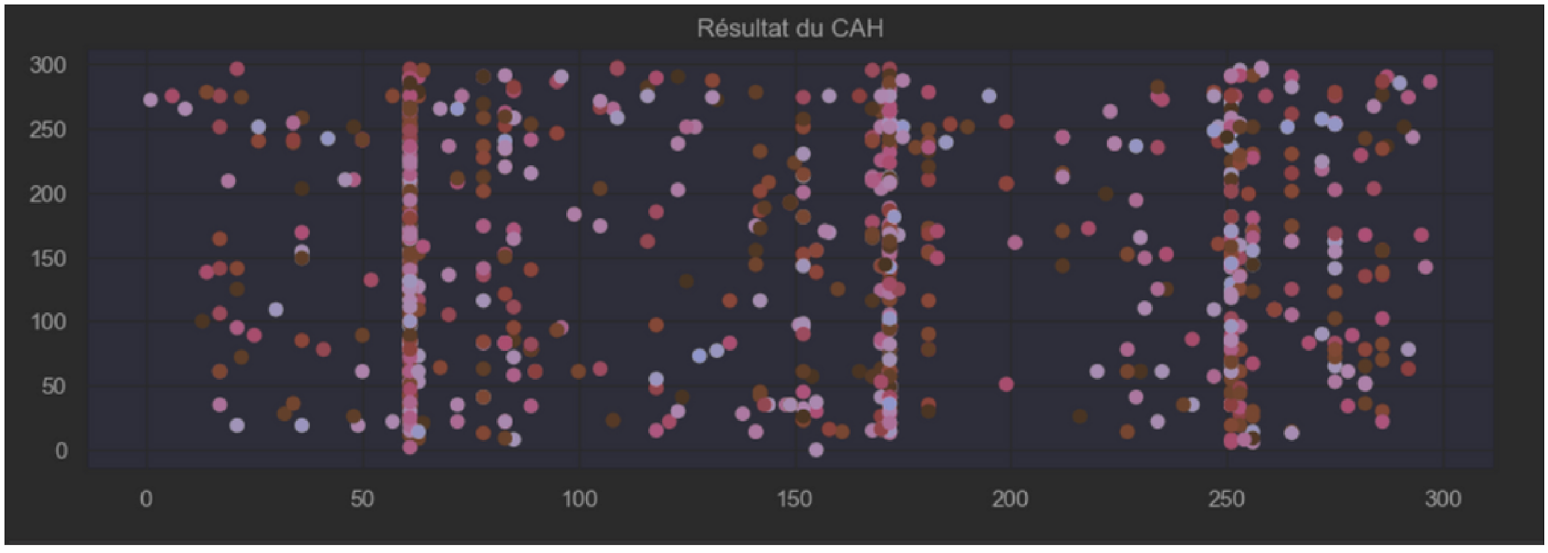

figure 12 : CAH_Result à ne pas mettre

- Comparison:

Is CAH better than k-means? In general, the experience shows that yes! This is because k-means can quickly make mistakes when category groups form clouds that are too close or overlapping shapes.

CHAPTER V:
DATA EVALUATING

- PCA :

The PCA algorithm is used to simplify and analyze complex datasets, reduce the dimensionality of data, and identify the most important features that capture the most variation in the data.

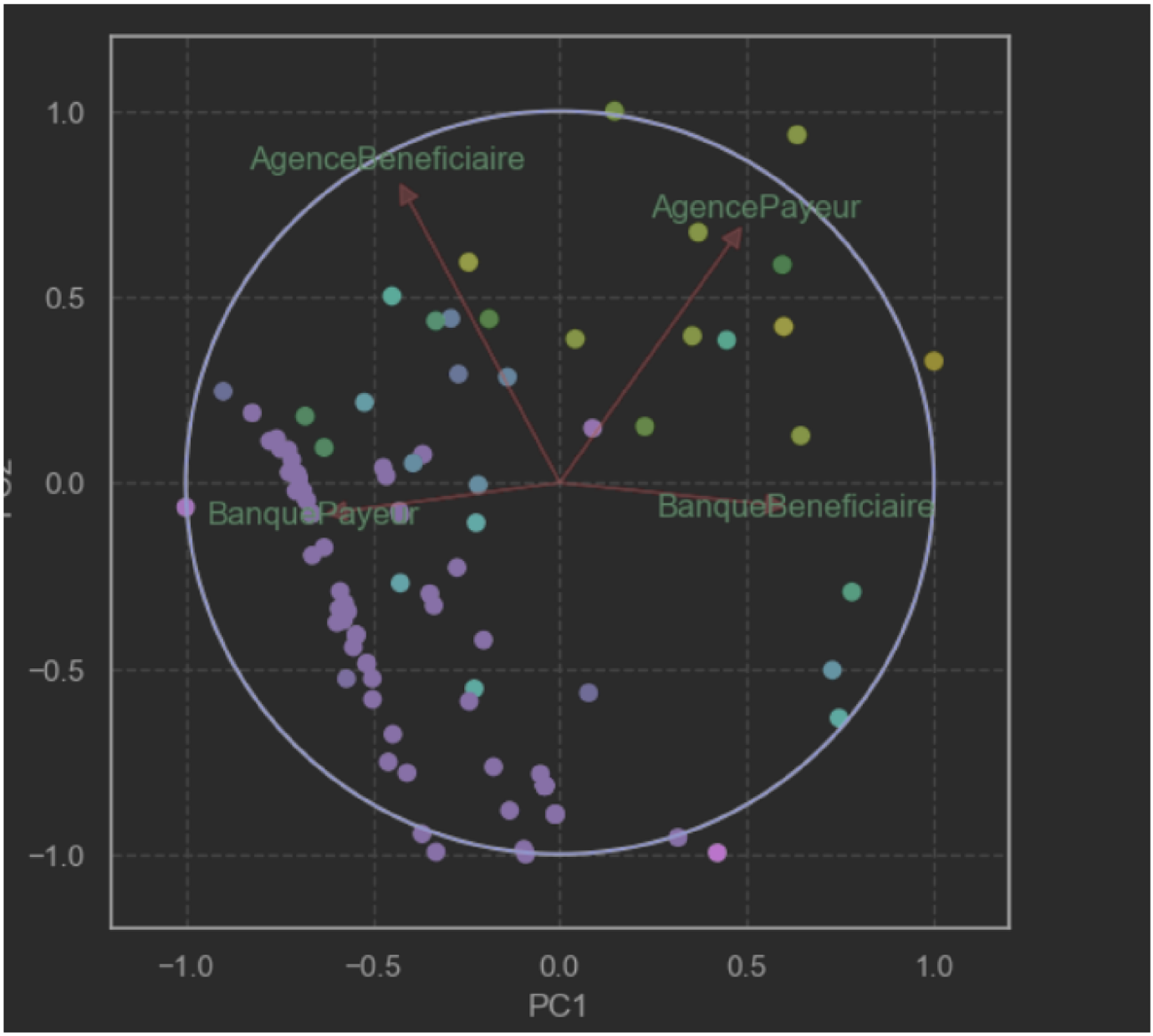

figure 12 : PCA_Result

the result of our PCA shows that the most efficient variables are "BanqueBeneficiaire" and "BanquePayeur" because they are the most tangent to the axes.

- RandomForestClassifier :

Random Forest Classifier is a supervised learning algorithm used for classification tasks. It is an ensemble method that uses multiple decision trees to make predictions.
The algorithm works by creating multiple decision trees based on a randomly selected subset of the data and features.

```
scores = pd.DataFrame(sc,columns=['classifier','imputer strategy', 'rescaler', 'score'])
scores.sort_values(by=['score'],ascending=False).head()
```

Executed in 17ms, 30 Apr at 19:03:10

5 rows × 4 columns pd.Dataframe

	classifier	imputer strategy	rescaler	score
0	DecisionTreeClassifier()	median	RobustScaler()	0.83
1	DecisionTreeClassifier()	median	MinMaxScaler()	0.83
16	RandomForestClassifier()	mean	MinMaxScaler()	0.83
15	RandomForestClassifier()	mean	RobustScaler()	0.83
14	RandomForestClassifier()	most_frequent	StandardScaler()	0.83

figure 13 : Random Forest Classifier Result

the score is high : 0.83
a higher score for any of these metrics indicates better performance of the random forest classifier.

- SVM :

Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression analysis. The algorithm works by finding the best boundary (hyperplane) that can separate the data into different classes.

Some advantages of SVM include : Effective in high-dimensional spaces, Robust to noise, Versatile, Efficient memory usage

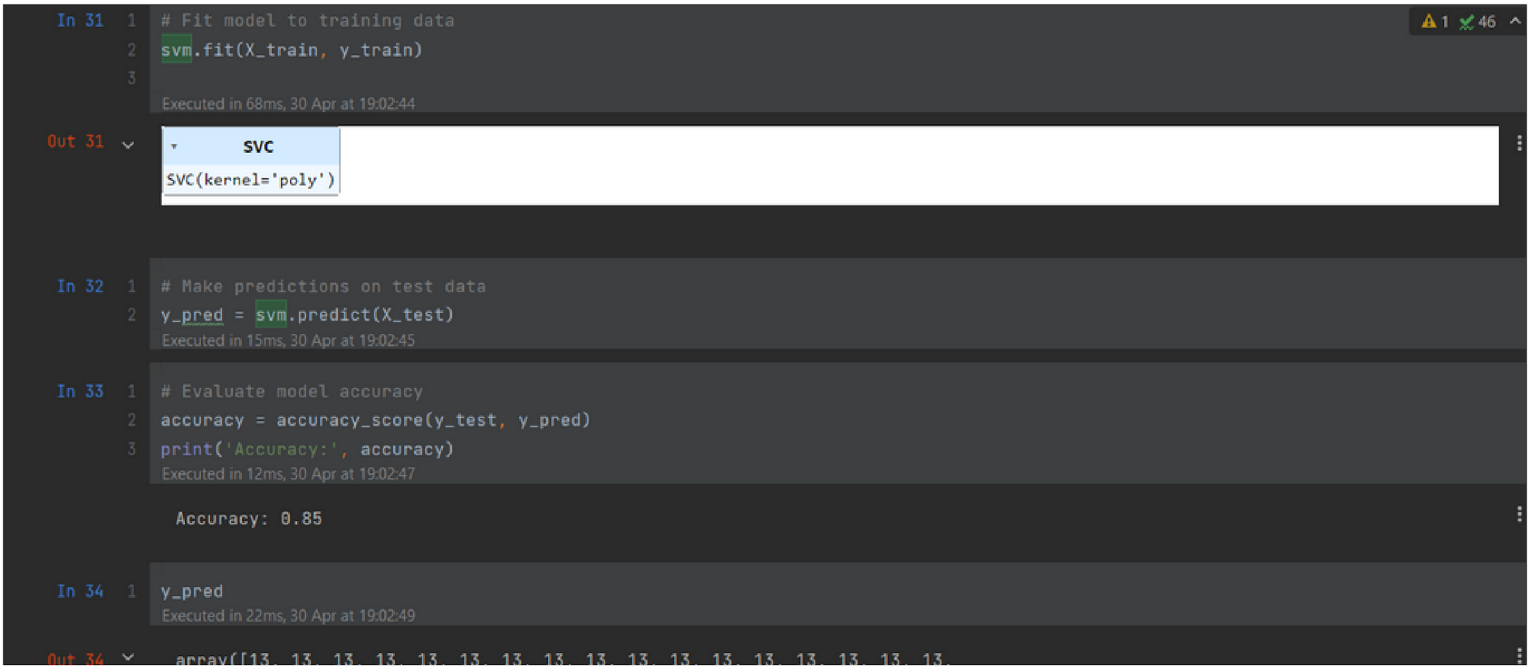

figure 14 : SVM Result

the accuracy is high : 0.85
a higher accuracy indicates better performance

- Polynomial Regression :**dep 1**

Polynomial regression, abbreviated E(y |x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y. It usually corresponded to the least-squares method. According to the Gauss Markov Theorem, the least square approach minimizes the variance of the coefficients.

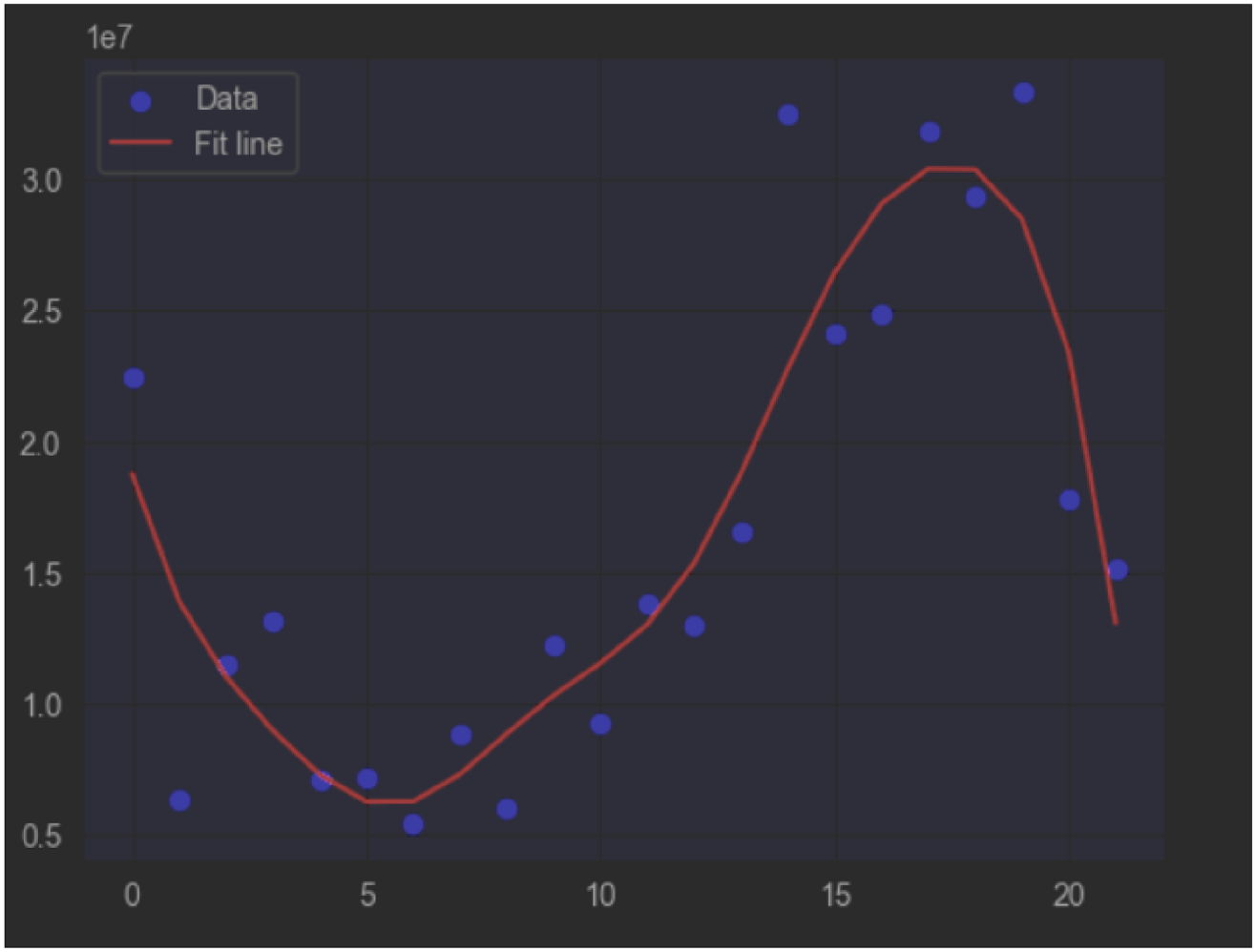

figure 15 : Polynomial Regression Result (Prediction Of Sum Of Amounts Per Day)

The scatter plot is close to sinusoidal curve with some outliers that we will eliminate later.

- KNN :

KNN (K-Nearest Neighbors) is a machine learning algorithm used for classification and regression tasks. It is a non-parametric algorithm that makes predictions based on the k closest training examples in the feature space .

```
# Create KNN and Random Forest classifiers
knn = KNeighborsClassifier()
rf = RandomForestClassifier()

# Fit the classifiers on the training data
knn.fit(X_train, y_train)
rf.fit(X_train, y_train)

# Predict the probabilities for the test data
y_knn_prob = knn.predict_proba(X_test)[:, 1]
y_rf_prob = rf.predict_proba(X_test)[:, 1]

# Compute the ROC curves and AUC scores
fpr_knn, tpr_knn, thresholds_knn = roc_curve(y_test, y_knn_prob)
auc_knn = roc_auc_score(y_test, y_knn_prob)

fpr_rf, tpr_rf, thresholds_rf = roc_curve(y_test, y_rf_prob)
auc_rf = roc_auc_score(y_test, y_rf_prob)

# Plot the ROC curves
plt.plot(fpr_knn, tpr_knn, label='KNN (AUC = {:.2f})'.format(auc_knn))
plt.plot(fpr_rf, tpr_rf, label='Random Forest (AUC = {:.2f})'.format(auc_rf))

# Plot the random line
plt.plot([0, 1], [0, 1], 'r--')
```

figure 16: KNN algorithm use

- ROC :

ROC (Receiver Operating Characteristic) is a graphical plot used to visualize and evaluate the performance of a binary classification model. The plot is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) for different threshold values.
In a binary classification problem, we have two classes: positive and negative.

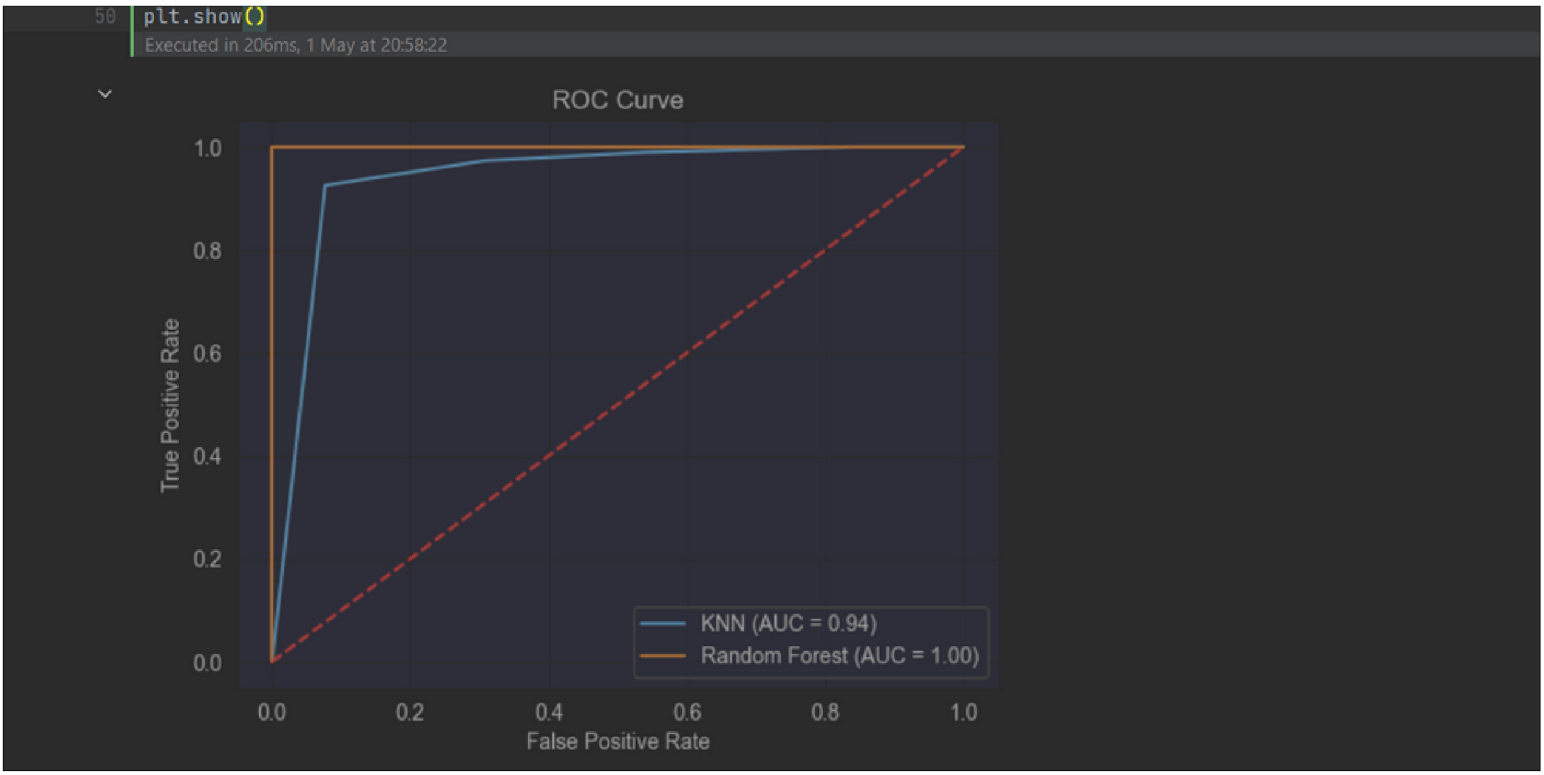

figure 17 : ROC Result

We don't have Binary data in our dataFrame , so we used a test : if (Montant > 50000) . Results were saved into a new column and we explored that binary column to make our ROC .
AUC values are high which means a good performance for the two compared Algorithms : KNN and Random Forest

CHAPTER VI:
DEPLOYMENT

- Predict Sum Of Paiment By Date :

We used the regression algorithm to predict the Sum of all payments in the data by choosing a date

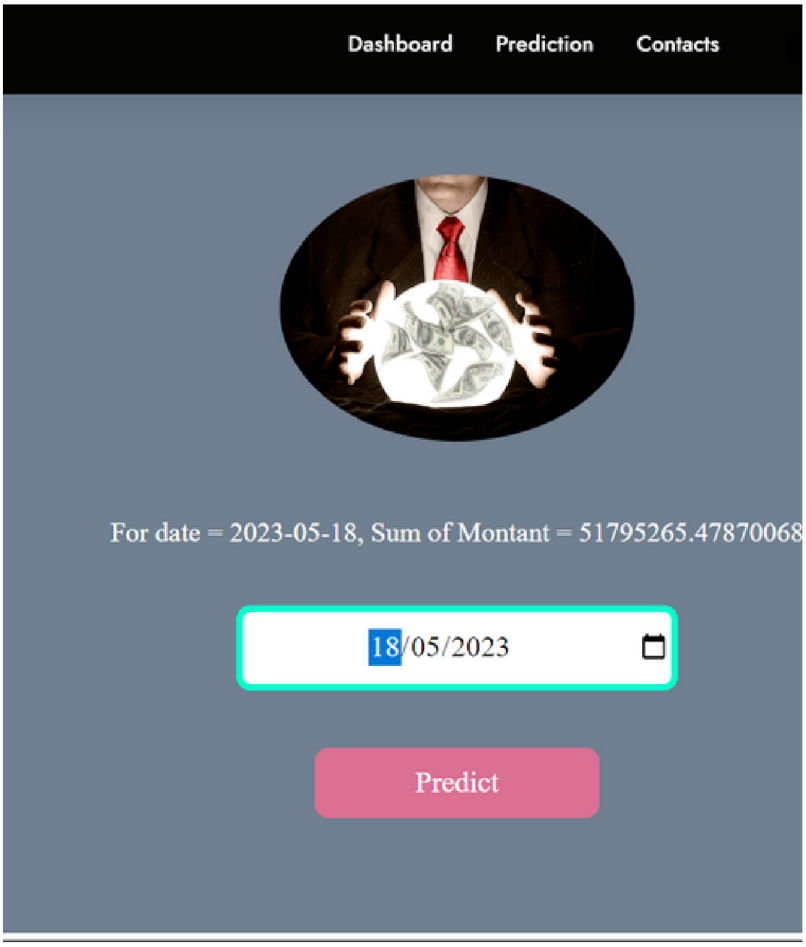

figure 18 : Sum Of Payments By Date Prediction

- Predict Cheque Situation :

In some cases , the cheque of the client is blocked and RejetCCE = -1 , so we tried to guess it's situation by training a prediction model

figure 19 : Cheque Situation Prediction

CONCLUSION

In conclusion, this project provided our group with a unique opportunity to apply our Business Intelligence knowledge and skills to a real-world problem, while also developing our teamwork and collaboration abilities. We focused on the processing of means of payment, specifically checks, and worked on a subject commissioned by a technology solutions provider that partners with banks to enhance their operational efficiency and customer experience.

Through the utilization of various data mining techniques, we were able to analyze and interpret the data to identify key insights and propose effective solutions to improve the payment processing system. These solutions have the potential to enhance operational efficiency and customer experience for the technology solutions provider's clients in the banking industry.

Overall, this project allowed us to successfully achieve our objective of assessing our ability to collaborate and apply our Business Intelligence knowledge and skills to solve complex real-world problems. In addition, this project equipped us with valuable experience and skills that we can apply in our future professional careers, allowing us to contribute to the success of our future employers and the industry as a whole.