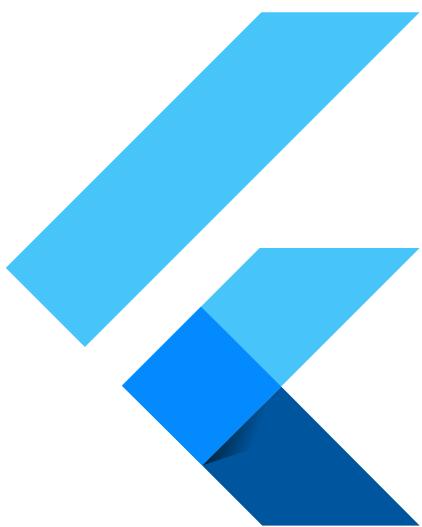


Project ML

Crédit BANK

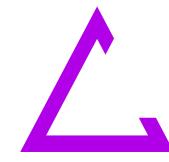


Groupe 7:

- Jérémi VESPUCE
- Yasmine Aarab
- Hounmenou Moise
- Diouf Sylvain
- Craysson TEDA TAKAM



Année : 2024



PLAN

- 1** Introduction
- 2** Conception du projet, Etablir un MLD
- 3** Processing (Traitement des données)
- 4** ML : Random Forest
- 5** Déploiement (Streamlit)
- 6** Conclusion

Introduction

- Ce projet consiste à développer un ou des modèles prédictifs à partir de ses données pour aider à évaluer les risques de défaut de remboursement des prêts.
- Eligibilité au prêt selon les critères
- A voir qui est-ce qui est susceptible au remboursement du prêt.

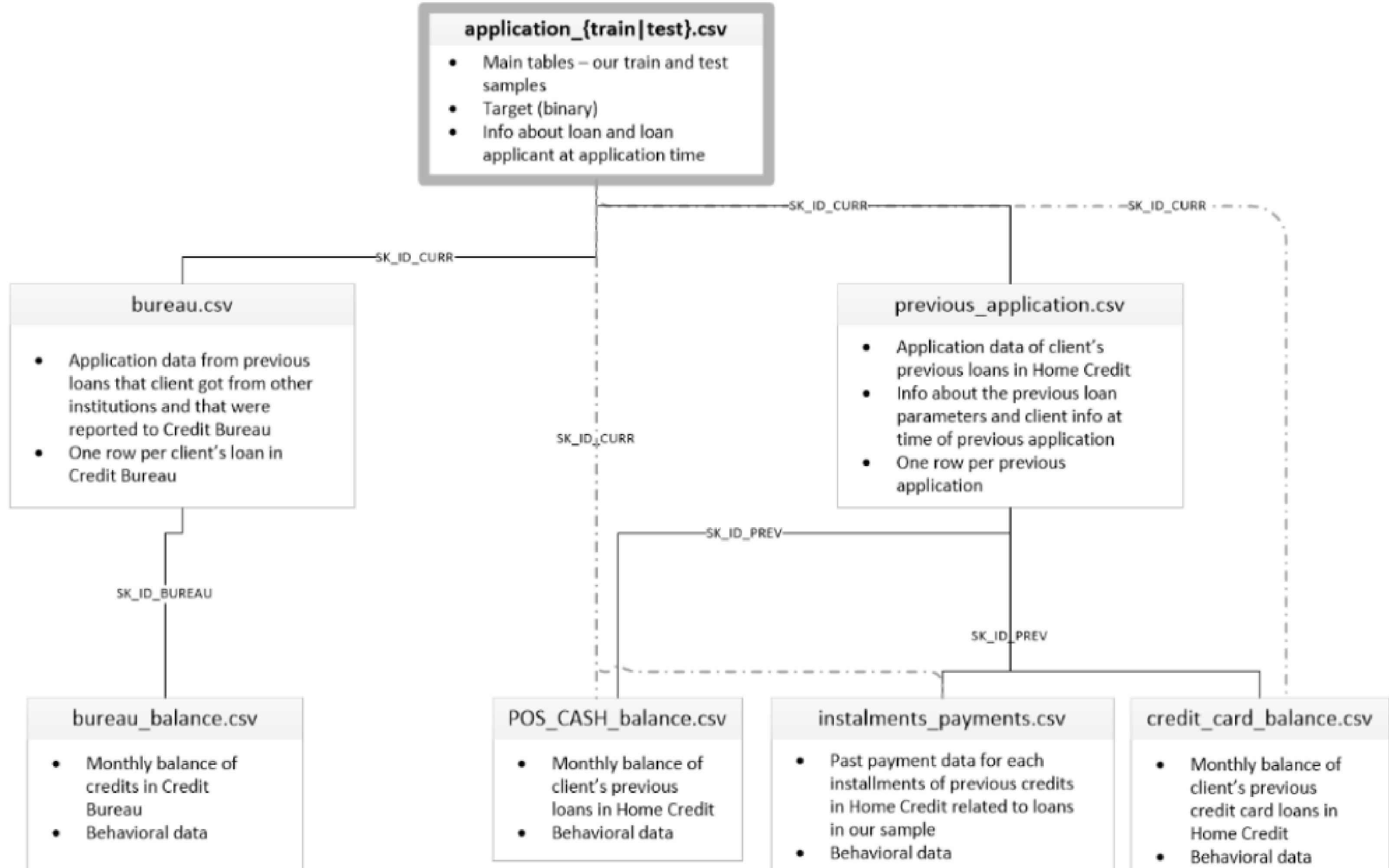
En principe, les résultats doivent être en mesure de définir les critères d'éligibilité pour effectuer un prêt ou non.

Conception du projet

La partie conception du projet consiste à identifier les grandes parties du projet dans le but de se fixer des objectifs sur un modèle à établir après une bonne compréhension du sujet.

Cette partie demande d'établir un MLD pour analyser la relation entre les tables.

MLD : Modèle Logique des données





Données utilisées

application_train

application_test

Feature_importances_

Processing: Nettoyage des données

Quelques
exemples :

Suppression des doublons

Filtrage des valeurs non valides

Remplacer les Caractères

Données catégorielles

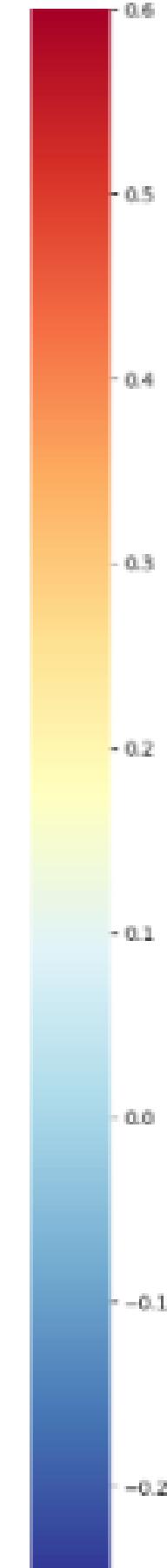
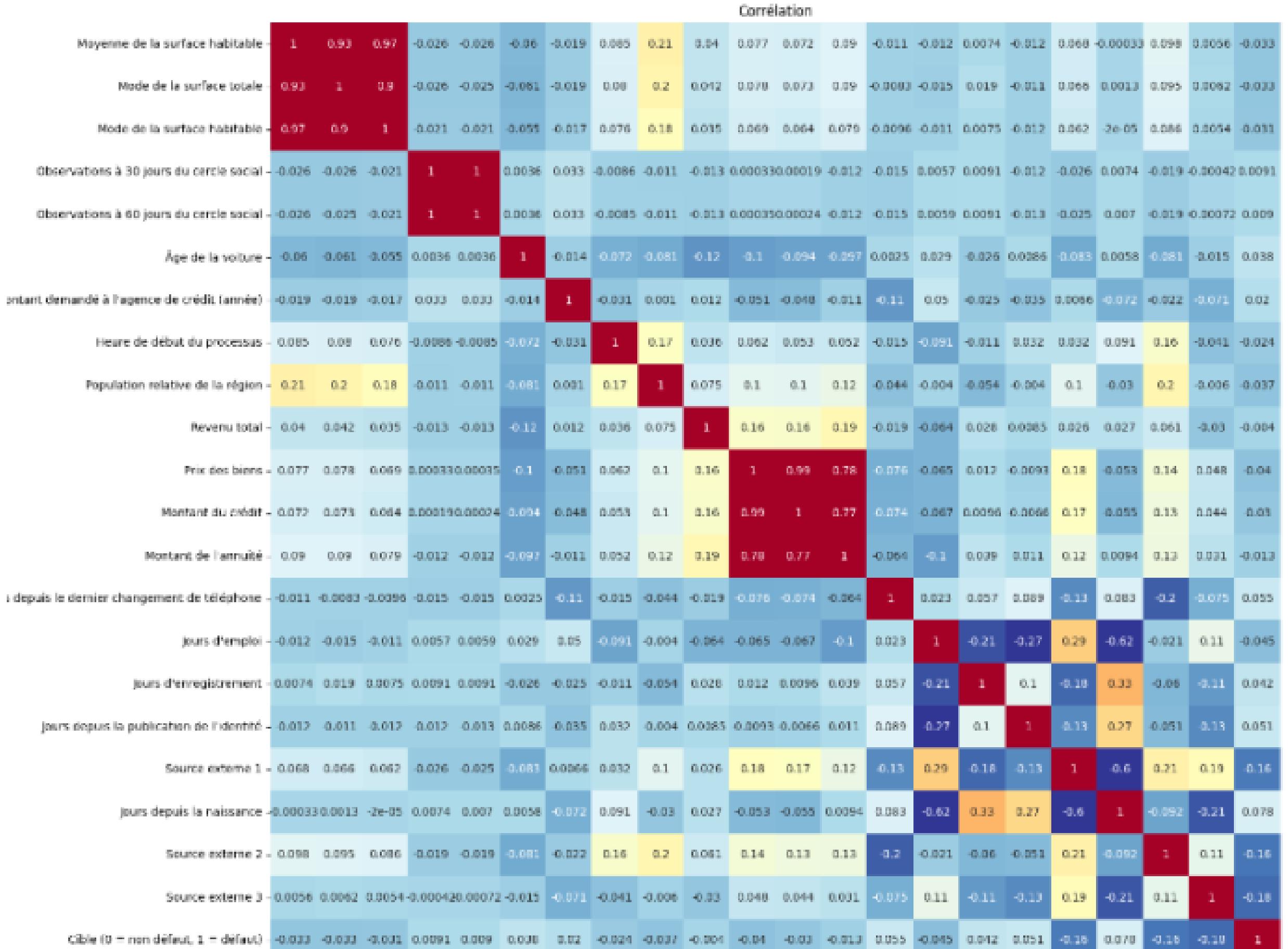
Corrélations

Remplacer les NAN par (mean)

Merge (concaténation) (tables)



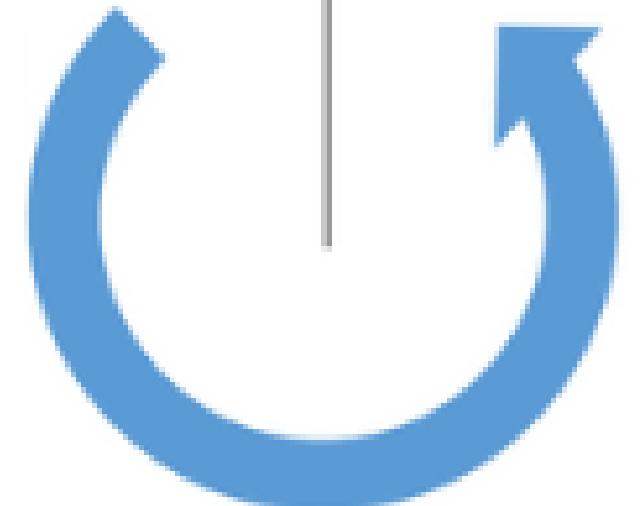
HeatMap: pour la Corrélation



Phases

1. Data Collection

kaggle



python

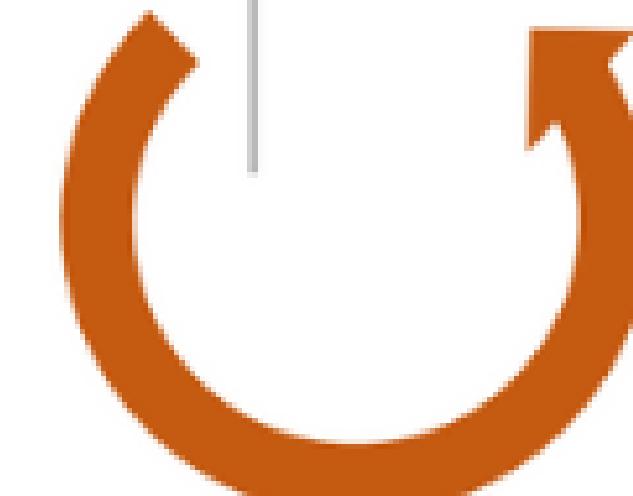


Pandas



2. Data validation

Suppression des doublons
Filtrage des valeurs non valides
Remplacement des valeurs manquantes
Transformation des données



3. Data Visualization

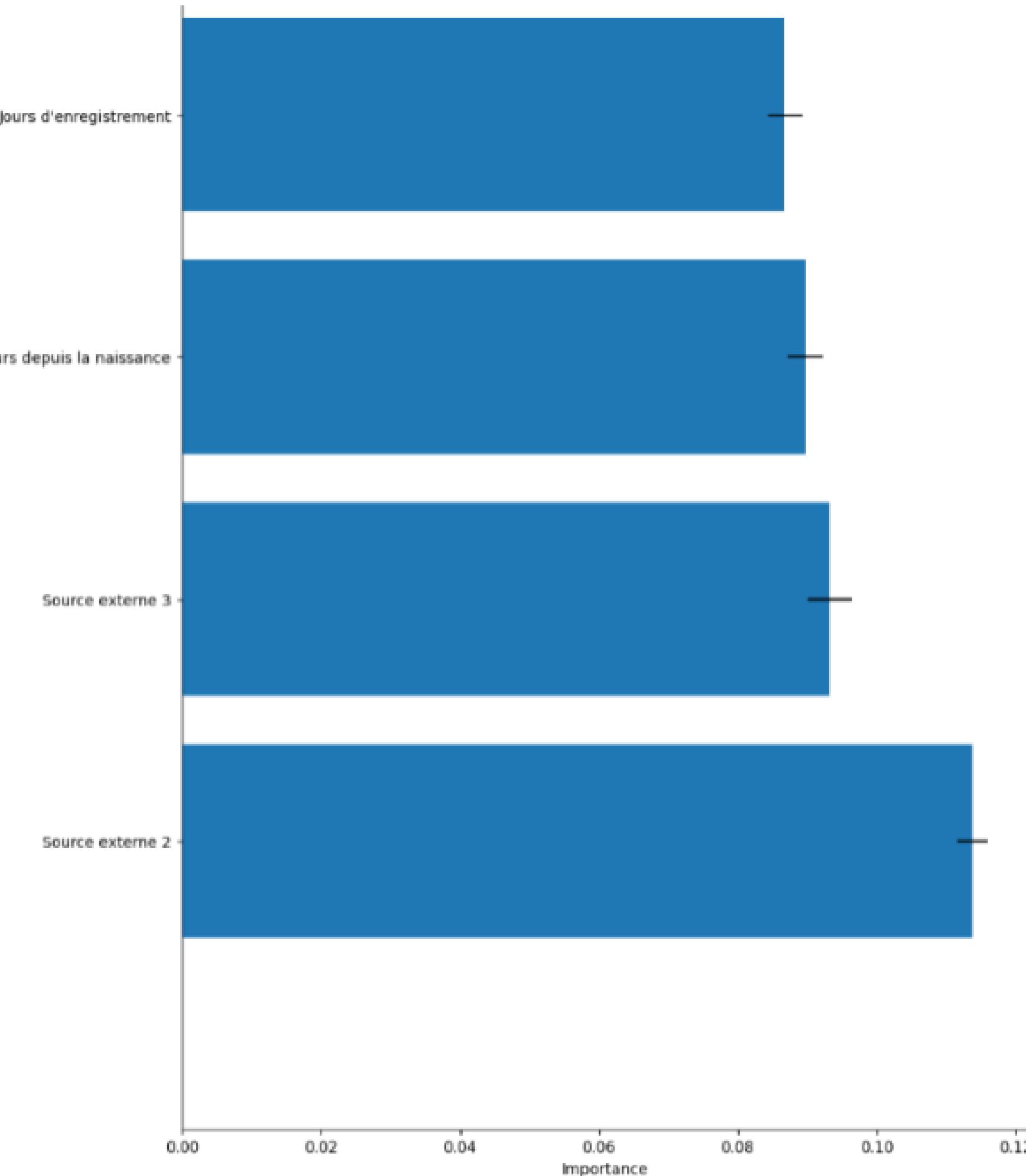
créer des visualisations interactives
basculer entre différentes sections
Affiche des statistiques descriptives et
des résultats de clustering.



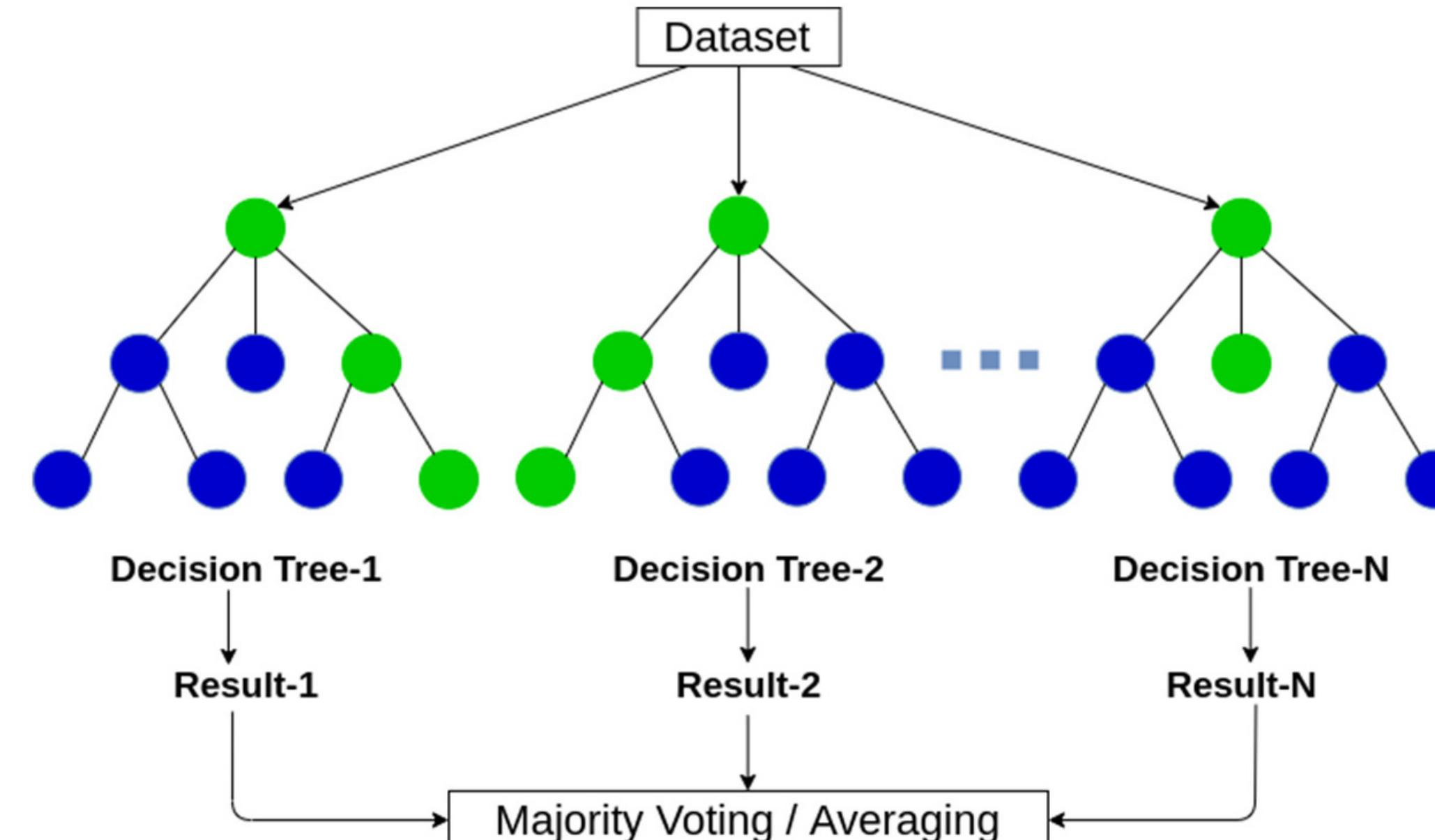
Streamlit



⚠ Quelques Variables choisies



Random Forest



Random Forest

Importance des variables (Après suppression des variables)

Corrélation

Mapage des variables catégorielles (remplacer par des valeurs numériques)

Remplacement des valeurs NaN par la moyenne

Normalisation

Modélisation

Random Forest

Accuracy au train

La précision (accuracy) est une métrique clé en classification qui

```
accuracy = accuracy_score(y_train, y_train_pred)
print(f"La précision est de : {accuracy:.2f}")
```

✓ 0.0s

La précision est de : 0.86

Recall au train

Le recall (rappel) est une métrique clé en classification qui mesure

```
recall = recall_score(y_train, y_train_pred)
print(f"Le recall est de : {recall:.2f}")
[277] ✓ 0.0s
... Le recall est de : 0.85
```

Random Forest

Recall au test

```
recall = recall_score(y_test, y_test_pred)  
print(f"Le recall est de : {recall:.2f}")
```

✓ 0.0s

Le recall est de : 0.67

F1 score au test

```
f1 = f1_score(y_test, y_test_pred)  
print(f"Le f1_score est de : {f1:.2f}")
```

✓ 0.0s

Le f1_score est de : 0.67

F1 score au train

Le f1_score est une métrique très utile pour évaluer les

```
f1 = f1_score(y_train, y_train_pred)  
print(f"Le f1_score est de : {f1:.2f}")
```

✓ 0.0s

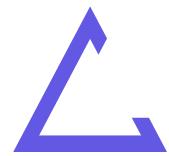
Le f1_score est de : 0.86



Conclusion

Ce projet nous a permis de :

- Réaliser et de comprendre les exigences et les critères pour faire un pret.
- Le modèle nous a permis de choisir les variables nécessaires pour savoir si une personnes est suceptible au remboursement ou non en se basant sur l'age



Perpectives Déploiement Streamlit ML XGBoost

- **Menu avec les différentes variables**
- **Afficher les critères pour un pret**
- **Visualisation Graphique (Scoring) sur le taux de remboursement selon l'âge (20-75)**
- **Etablir un autre modèle ensembliste (XGBoost) : pour comparer leur performance (RAndom Forest vs XGBoost)**



XGBoost

Streamlit



Thank you