Projet Machine-learning

Réalisé par
Ayedi Yassine
Bhar Hane
Dahmen Omar
Loukil Eya
Soffelgil Nour

Sommaire

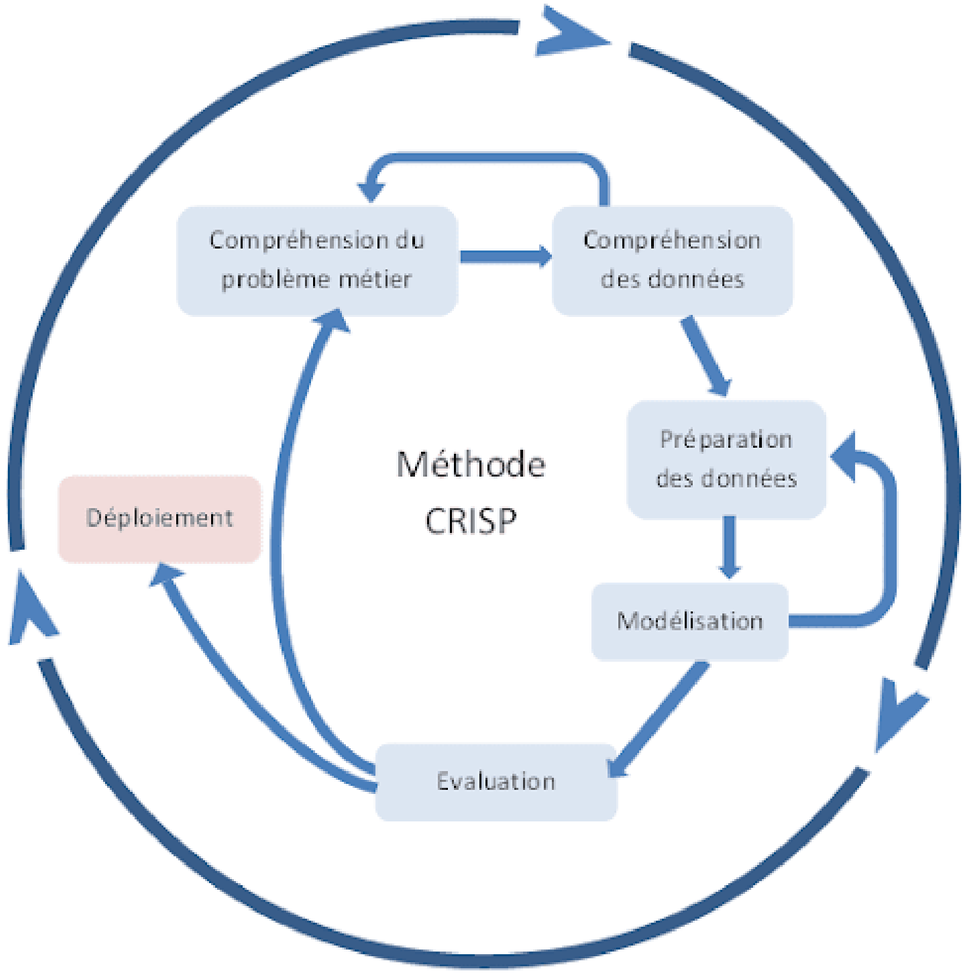
I Introduction

La méthodologie CRISP

Les années 60

6 étapes

Efficace



II La compréhension des données

1. Généralités

Problématique

Exploration des données

Description des données

Qualité des données

2. Notre cas

Dataset

Customer Churn

Variable cible

Churn

Problème

Fidélisation des clients

"C’est parce que **la perte de clients** pourrait causer **une perte critique** de revenus. Quant à le retenir **coûte de cinq à six fois** moins cher que de trouver **un nouveau client**."

III La préparation des données

1. Généralités

Sélection des données

Constructions de nouvelles données

Nettoyage des données

Ajout/ fusion données

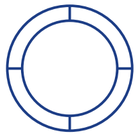
2. Notre cas

Total Charges

11 case manquantes

Churn Reseaon

5174 valeurs manquantes



Online

BackupInternetService

Variables catégorielles

CustomerID

State

Country

Variables supprimées

2. Notre cas

Encodage des variables

Détection de corrélation

Maintenant, on est prêt à affronter le cœur même de l'exploration de données : la modélisation.

IV La modélisation des données

1. Généralités

a Sélection de la technique de modélisation

b Création des modèles

c Evaluation des modéles

2. Notre cas

a Sélection de la technique de modélisation

- K-NN
- Naive Bayes
- SVM
- Régression logistique
- Random Forest
- Clustering K-means
- Arbre de décision

b Création des modèles

- Choix de K et la distance métrique
- Gaussien, binomiale et multinomiale
- Hyperplan
- Nombre de Cluster et d'itérations max
- Indice de GINI ou l’entropie

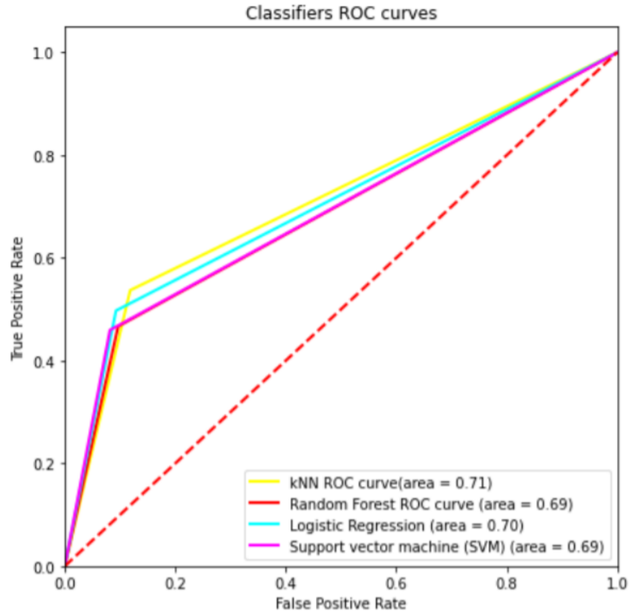
2. Notre cas

C Evaluation des modéles

- **Table de score**

	Model	Score
2	Aaccuracy SBS	77.74%
4	Aaccuracy SFBS	77.49%
3	Aaccuracy SFFS	77.47%
1	Aaccuracy SFS	77.29%
0	Aaccuracy NB	71.85%

- **Courbe ROC**

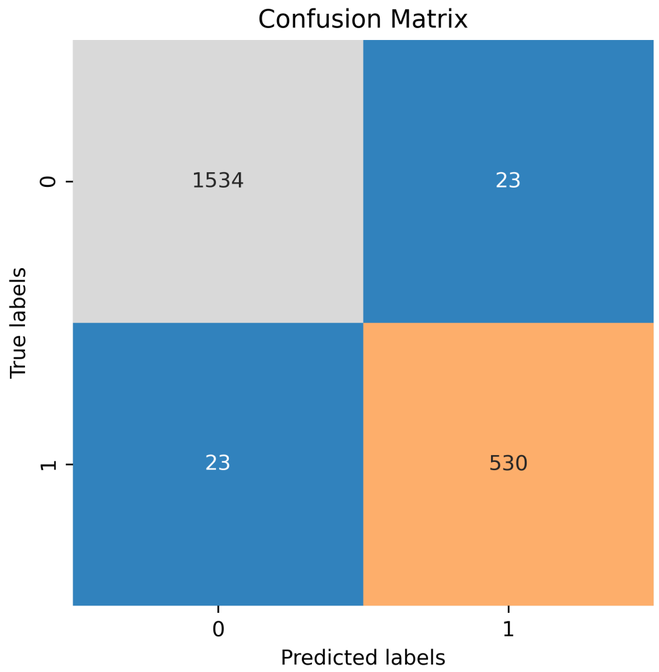


2. Notre cas

- **Rapport de classification**

	precision	recall	f1-score	support
0	0.98522800	0.98522800	0.98522800	1557
1	0.95840868	0.95840868	0.95840868	553
accuracy			0.97819905	2110
macro avg	0.97181834	0.97181834	0.97181834	2110
weighted avg	0.97819905	0.97819905	0.97819905	2110

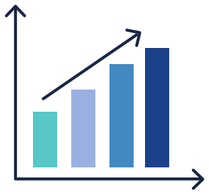
- **Matrice de confusion**



V La visualisation des données

1. Généralités

Les graphiques et les cartes



2. Notre cas

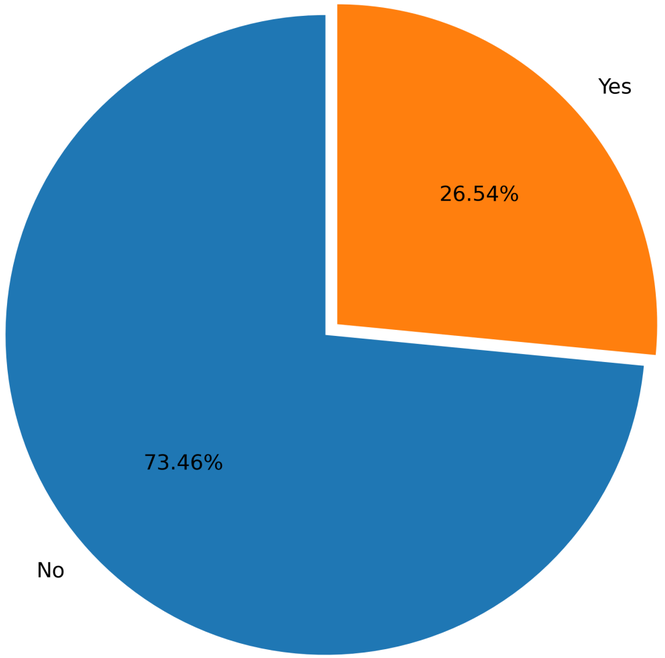


figure1: Customer Churn en pourcentage

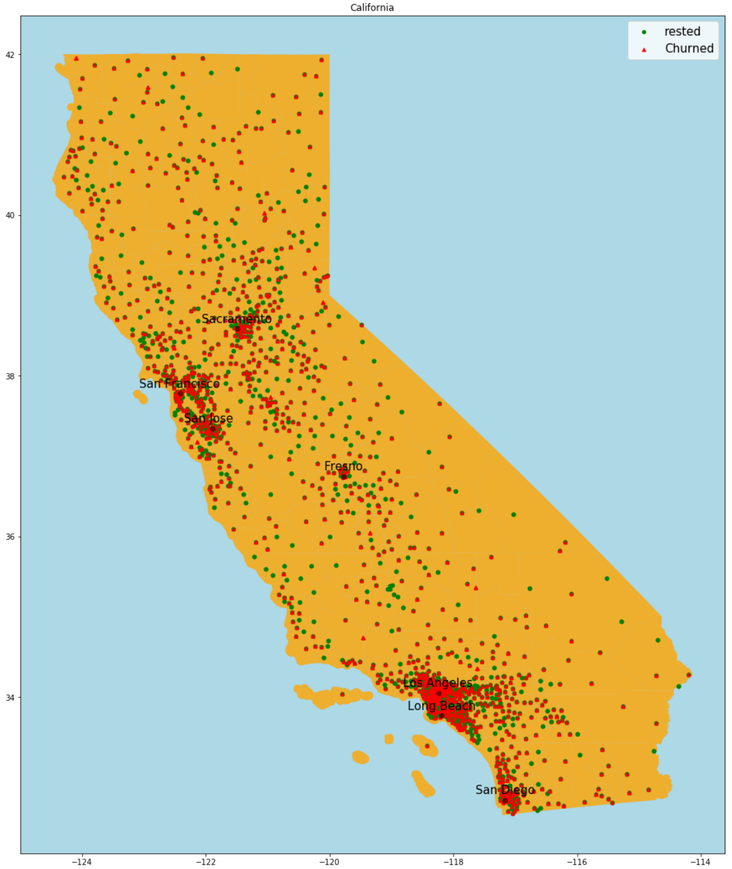


figure2:: Carte de Californie en fonction du Churn

VI Conclusion et perspectives

“You can have data without information, but you cannot have information without data.”
Daniel Keys Moran

Merci de nous avoir accordé votre attention!