

Projet Scoring



Academic Year
2020-2021

TABLE MATIERE

INTRODUCTION	3
2. IBM Master Plan	5
COMPREHENSION DU PROBLEME MÉTIER	7
1.Contexte du projet	7
2.Objectives Data Science	8
COMPRÉHENSION DES DONNÉES	9
1.Identification	9
2.Aperçu sur notre Dataset	11
3.Visualisation de la corrélation entre les variables	12
4.Visualisation du nombre de Classification Projet	13
5.Visualisation du nombre de Classification contrepartie	13
6.Visualisation du nombre de viabilité	13
PREPARATION DES DONNEES	14
1.Nettoyage des données	14
1.1 Vérification des valeurs manquantes	14
1.2 Vérification des lignes en double	15
1.3 Pourcentage des valeurs manquantes de chaque colonne	15
2.Imputation des données	16
3.Selection	17
MODELISATION DES DONNEES	18
1. KNeighborsRegressor(KNN Regressor)	21
1.1 Cas d'utilisation	21
2. KNeighborsClassifier(KNN Classifier)	22
3.RandomForestClassifier	22
3.1 Cas d'utilisation	23
DEPLOIEMENT	224
1-Logiciel de déploiement	24
2-Réalisation de l'application	25
CONCLUSION AND PERSPECTIVES	27

Introduction

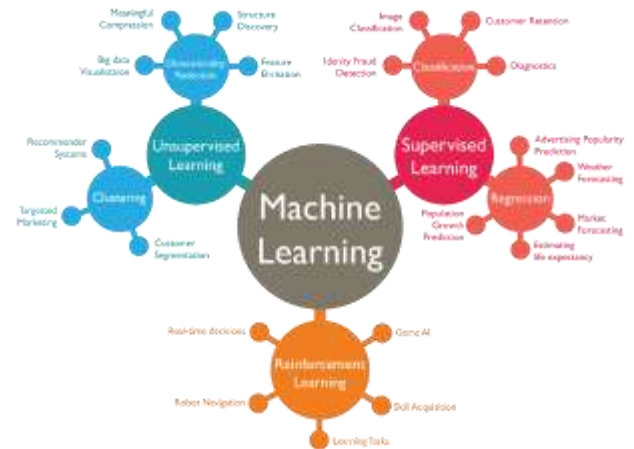
Les décisions basées sur des algorithmes deviennent prépondérantes dans bon nombre de domaines tels que le diagnostic médical, la justice prédictive, la reconnaissance faciale, la détection de fraudes, la recherche d'emploi, ou l'accès à l'enseignement supérieur. Le monde de la finance n'échappe bien évidemment pas à cette révolution de la science des données. L'intelligence artificielle (IA) et les techniques d'apprentissage automatique (Machine Learning ou ML par la suite) sont particulièrement utiles en matière de connaissance client, d'allocation d'actifs, de détection de blanchiment et de transactions illégales, de gestion des risques, ou d'améliorations des processus internes.

Même si par nature, le scoring de crédit est un sujet technique, il revêt une importance capitale au niveau économique et social. En effet, il conditionne l'allocation du crédit entre les agents économiques : quels ménages vont pouvoir accéder à la propriété, quelles entreprises vont pouvoir financer leurs investissements, quelles sociétés vont devoir déposer leur bilan et parmi celles-ci combien seront liquidées, etc. Dès lors, les modèles de scoring de crédit ont des implications majeures en termes de stabilité financière (provisions, capital réglementaire des banques), d'inclusion financière, d'emploi, et de croissance économique.

Le scoring de crédit fut historiquement l'un des premiers champs d'application des techniques de ML. Nous définissons dans cet article le ML comme un ensemble d'algorithmes destinés à résoudre des problèmes et dont la performance s'améliore avec l'expérience et les données sans intervention humaine. Ainsi, un algorithme de ML est un programme informatique qui permet de construire, et surtout d'améliorer de façon autonome un modèle de régression ou de classification. Dans le contexte du risque de crédit, nous considérerons principalement des modèles de classification. Un algorithme de classification vise à établir une fonction de lien (un modèle) entre une variable cible Y (de type binaire dans le cas du scoring de crédit, par exemple défaut ou non défaut) et un ensemble de prédicteurs ou caractéristiques X . Cette fonction de lien est révélée à partir d'un échantillon d'apprentissage. La méthode de classification est dite supervisée lorsque la variable cible Y est observée sur l'échantillon d'apprentissage. Une fois le modèle entraîné sur l'échantillon d'apprentissage, il est ensuite utilisé pour réaliser une prévision (classification) de la variable Y sur un

échantillon test à partir des observations des prédicteurs X , la distribution conditionnelle de Y sachant X étant supposée être la même dans les deux échantillons.

On peut distinguer deux grandes familles. Premièrement, les méthodes de classification supervisée dites individuelles visent à partitionner l'espace des prédicteurs afin de prévoir l'événement. Les plus utilisées dans le domaine du risque de crédit sont les arbres de classification, les machines à vecteurs de support (SVM), et les réseaux neuronaux. Deuxièmement, les méthodes d'ensemble combinent des prédictions issues d'un ensemble de plusieurs modèles de base en utilisant une règle de vote pour aboutir à la classification finale.



La phase d'approche analytique permet de limiter les algorithmes qui seront utilisés par la suite (modèle prédictif, clusters, modèle descriptif, modèle de classification etc...)

Exigences en matière de données (Data requirements):

Le contenu, les formats et les sources des données nécessaires à la collecte initiale des données. Lors de l'étape de collecte de données, nous extrayons les données de diverses sources, puis les regroupons dans des bases de données appropriées..

Collecte des données (Data collection):

Dans la phase initiale de collecte de données, les scientifiques des données identifient et rassemblent les ressources de données disponibles (structurées, non structurées, semi-structurées) pertinentes pour le domaine du problème.

Compréhension des données (Data understanding):

Après la collecte des données d'origine, les data scientists utilisent généralement des statistiques descriptives et des techniques de visualisation pour comprendre le contenu des données et évaluer la qualité des données.

Préparation des données (Data preparation):

Les données doivent être préparées à l'aide de nombreuses opérations telles que l'adressage des valeurs manquantes ou invalides et la suppression des doublons. Cette étape prend généralement près de 90 % du temps total du projet.

Modélisation (Modeling):

L'étape de modélisation comprend deux types qui dépendent de la compréhension métier que nous avons traitée dans un premier temps : descriptive et prédictive. Ces modèles sont basés sur l'approche analytique qui a été adoptée.

Evaluation (Evaluation):

L'évaluation de la précision d'un modèle est une partie essentielle du projet. C'est l'étape dans laquelle nous vérifions si le modèle que nous avons déjà généré répond ou non à la requête initiale.

Deploiement(Deployment):

Enfin, une fois validé, le modèle sera déployé et une phase de retour d'expérience sera lancée afin de le réévaluer d'un point de vue client.

II.COMPREHENSION DU PROBLEME METIER

Cette phase vise à déterminer précisément les données à analyser, à identifier la qualité des données disponibles et à faire le lien entre les données et leur signification d'un point de vue métier.

Entre fantasme et réalité, l'intelligence artificielle ne laisse pas indifférent. Utilisée dans de nombreux domaines, elle fait désormais partie de notre quotidien. Les banques n'échappent pas à la règle et utilisent intensivement l'intelligence artificielle. Chabots conversationnels, robots conseillers ou aides au prêt outils, l'intelligence artificielle transforme progressivement le quotidien des banquiers. Les banques sont également pionnières dans l'utilisation de l'intelligence artificielle avec des systèmes experts depuis les années 1980. C'est l'un des secteurs dans lesquels l'intelligence artificielle est la plus utilisée. Les systèmes experts (ou outils d'aide à la décision) sont l'une des formes les plus efficaces et éprouvées d'intelligence. 88% des banques les utilisent. Principalement utilisés dans l'octroi de crédit, ils sont capables de simuler le comportement d'un expert humain et d'analyser le risque d'octroi de crédit à particuliers, professionnels ou entreprises.



1.Contexte du projet

Les institutions financières se sont tournées vers la technologie, avec une augmentation de l'exploration de l'intelligence artificielle ces dernières années..

L'objectif de ce projet est de réaliser une application de notation de crédit en utilisant des données ainsi la restitution de tableaux de bord interactifs pour la prise de décision.

2.Objectives Data Science

L'objectif principal de la science des données est principalement divisé dans les domaines suivants : extraction de données, pré-traitement, analyse et déduction d'informations pour tirer des idées et des conclusions.

Dans ce projet, nous avons donc défini nos objectifs en science des données ci-dessous :

- Compréhension des sources de données
- Collecte des données
- Nettoyage et prétraitement des données
- Création de modèles utilisant machine learning

III.COMPREHENSION DES DONNEES

Cette phase vise à déterminer précisément les données à analyser, à **identifier** la qualité des données disponibles et à faire le lien entre les données et leur signification d'un point **de vue métier**.

1.Identification

Ci-dessous la liste des features trouvées au niveau de la base de données et leur description :

CODE_DEMANDE
NOTE_CONTRE_PARTIE :
CLASSIFICATION_CONTRE_PARTIE
NOTE_PROJET
CLASSIFICATION_PROJET
NOTE_VIABILITE
VIABILITE
RISQUE_ENG_PROMOTEUR
RISQUE_MANAGERIAL
RISQUE_PRODUCTION
RISQUE_MARCHE_PRODUIT
RISQUE_FINANCIER
REACTIVITE_COMPORTEMENT
RISQUE_REGLEMENTAIRE
RISQUE_STRAT_COMMERCIALE
QUALITE_ETUDE
PRISE_RISQUE_PROMOTEUR
RISQUE_INFRAST_LOGISTIQUE
APTITUDES_GESTION_ADM_FIN
APTITUDES_EXP_SECT_ACTIVITE
APTITUDES_EXP_ASPECTS_CCIAUX
RISQUE_TECH_COUVERTURE
RISQUE_MARCHE
RISQUE_APPRO
RISQUE_CONCURRENCE
RISQUE_PRODUIT
RISQUE_ASSISE_FI_SOUTIEN
CONCENTRATION_SOLVA_CLIENT
APTITUDE_REMBOURSEMENT
MOBILISATION_CREDIT

Note_Contraire_Partie	Classification_contre_partie
[0-39,70]	C5
[40.00-59.05]	C4
[60.00-79.00]	C3
[80.20-89.00]	C2
[90.00-100.00]	C1

Tableau 1: Relation entre 2 colonnes note_contre_partie et classification_contre_partie

Note_Projet	Classification_Projet
[0-39.70]	P5
[41.67-59.00]	P4
[60.00-79.00]	P3
[80.00-89.00]	P2
[90.00-99.00]	P1

Tableau 2: Relation entre 2 colonnes note_projet et classification_projet

Note_Viabilite	Viabilite
[28.00]	Non Crédible
[74.00]	Compromis
[107.23-114.00]	Acceptable avec attention
[123.20-134.00]	Bon
[143.00-159.25]	Très bon
[163.00-193.00]	Excellent

Tableau 3: Relation entre 2 colonnes note_viability et viabilité

2.Aperçu sur notre Dataset

	CODE_DEMANDE	NOTE_CONTRE_PARTIE	CLASSIFICATION_CONTRE_PARTIE	NOTE_VIABILITE	NOTE_PROJET	CLASSIFICATION_PROJET	RISQUE_ENG
0	99131083	48.25	C4	28.00	48.00	P4	
1	99131159	48.85	C4	28.00	50.00	P4	
2	99131144	49.90	C4	28.00	44.00	P4	
3	99131179	0.00	C5	NaN	0.00	NaN	
4	99131175	44.50	C4	28.00	54.00	P4	
...
1631	99191001	49.90	C4	28.00	43.00	P4	
1632	99184054	66.25	C3	127.25	61.00	P3	
1633	99191068	97.00	C1	159.00	75.00	P3	
1634	99191043	53.43	C4	28.00	42.67	P4	
1635	99192014	40.75	C4	28.00	38.00	P5	

```

CODE_DEMANDE                int64
NOTE_CONTRE_PARTIE          float64
CLASSIFICATION_CONTRE_PARTIE object
NOTE_VIABILITE              float64
NOTE_PROJET                  float64
CLASSIFICATION_PROJET        object
RISQUE_ENG_PROMOTEUR         float64
QUALITE_ETUDE                float64
REACTIVITE_COMPORTEMENT      float64
PRISE_RISQUE_PROMOTEUR       float64
RISQUE_MANAGERIAL            float64
APTITUDES_GESTION_ADM_FIN     float64
APTITUDES_EXP_SECT_ACTIVITE   float64
APTITUDES_EXP_ASPECTS_CCIAUX float64
RISQUE_PRODUCTION            float64
RISQUE_INFRAST_LOGISTIQUE     float64
RISQUE_TECH_COUVERTURE       float64
RISQUE_APPRO                 float64
RISQUE_REGLEMENTAIRE         float64
RISQUE_MARCHE_PRODUIT        int64
RISQUE_MARCHE                float64
RISQUE_CONCURRENCE           float64
RISQUE_PRODUIT               float64
RISQUE_STRAT_COMMERCIALE     float64
RISQUE_FINANCIER             int64
RISQUE_ASSISE_FI_SOUTIEN     float64
CONCENTRATION_SOLVA_CLIENT   float64
APTITUDE_REMBOURSEMENT       float64
MOBILISATION_CREDIT          float64
VIABILITE                    object

```



Ce jeu de données contient
1636 lignes et 31 variables
 dont **les variables cibles**
 sont : **Note_Projet**
Note_Contre_Partie
Classification_Projet
Classification_Contre_Partie
Note_viabilite
Viabilite

Notre jeu de données contient 3 attributs classés comme objet. Un attribut de type d'objet consiste en des données catégoriques qui placent chaque client dans un certain groupe.

Les autres types des données sont int64 et float64, ce qui signifie que les données de l'attribut peuvent être calculées.

3. Visualisation de la corrélation entre les variables

La matrice de corrélation indique les valeurs de corrélation, qui mesurent le degré de relation entre chaque paire de variables. Les valeurs de corrélation peuvent aller de -1 à +1. Si les deux variables ont tendance à augmenter et à diminuer en même temps, la valeur de corrélation est positive.

Lorsqu'une variable augmente tandis que l'autre diminue, la valeur de corrélation est négative.

Utilisez la matrice de corrélation pour évaluer l'importance et la direction de la relation entre deux variables. Une valeur de corrélation positive élevée indique que les variables mesurent la même caractéristique. Si les items ne sont pas fortement corrélés, ils peuvent mesurer différentes caractéristiques ou peuvent ne pas être clairement définies.

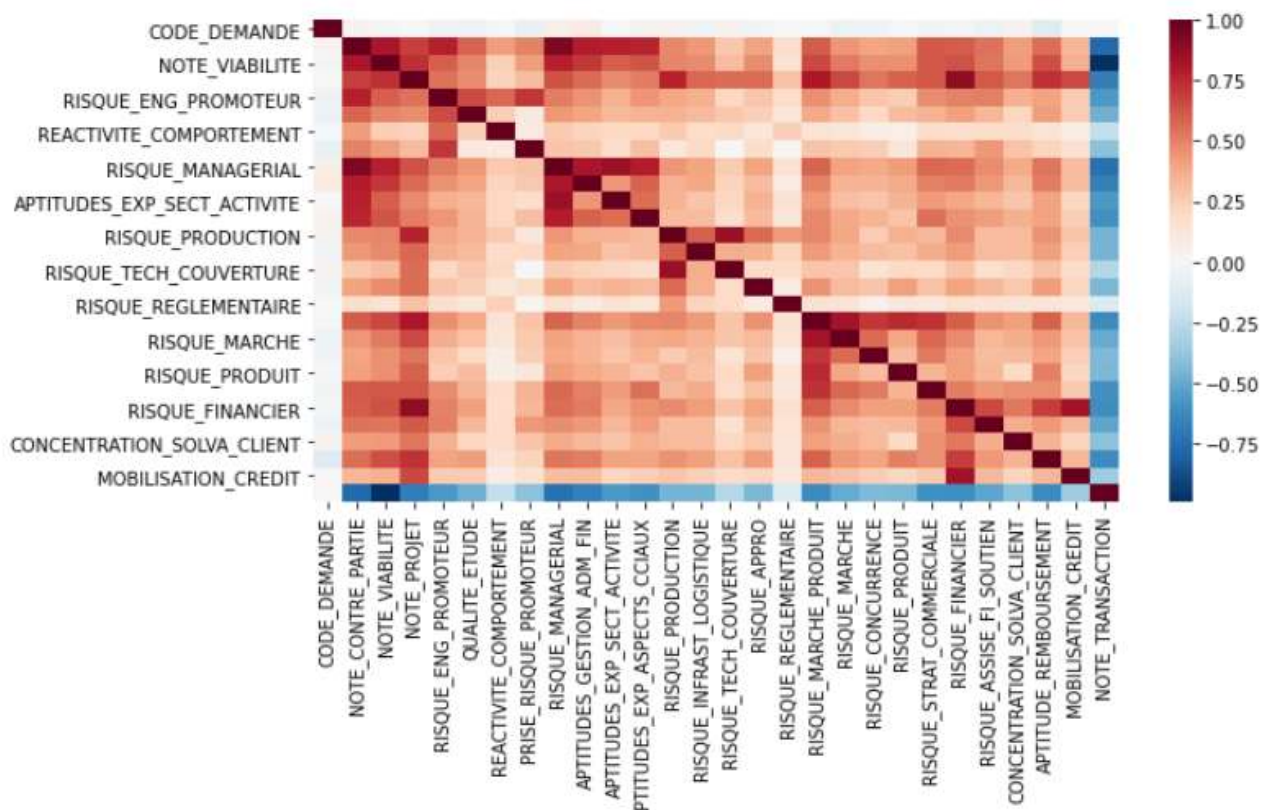
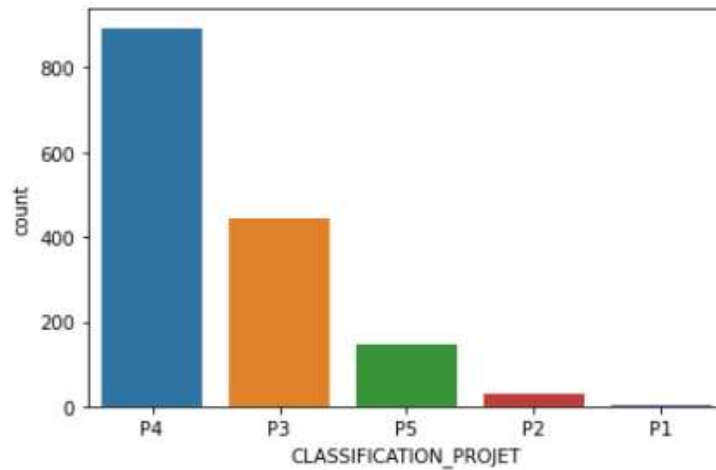
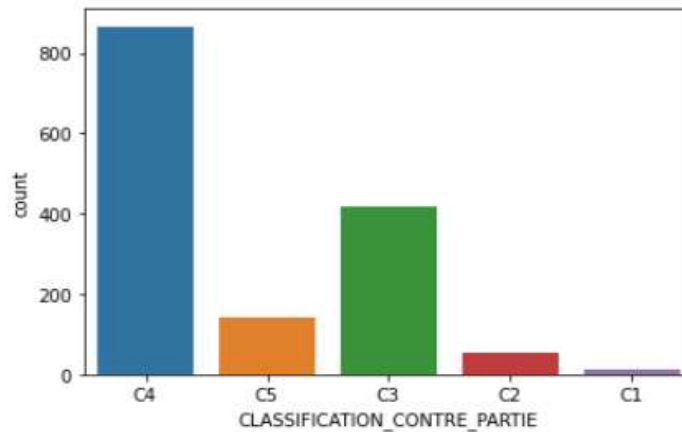


image 2 : Correlation des variables

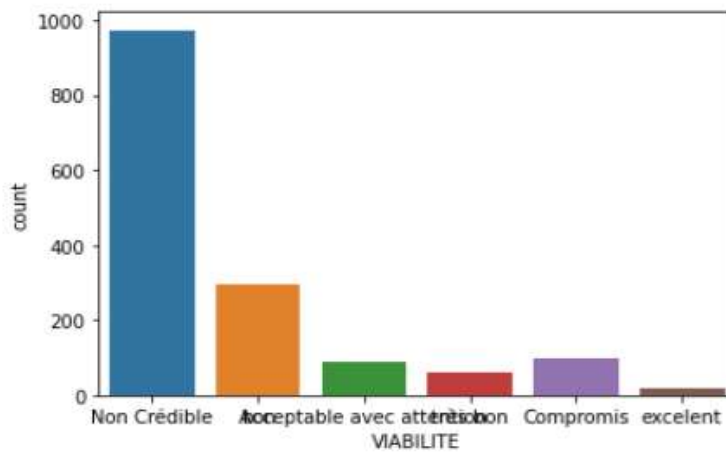
4. Visualisation du nombre de Classification Projet



5. Visualisation du nombre de Classification contrepartie



6. Visualisation du nombre de viabilité



IV. PREPARATION DES DONNEES

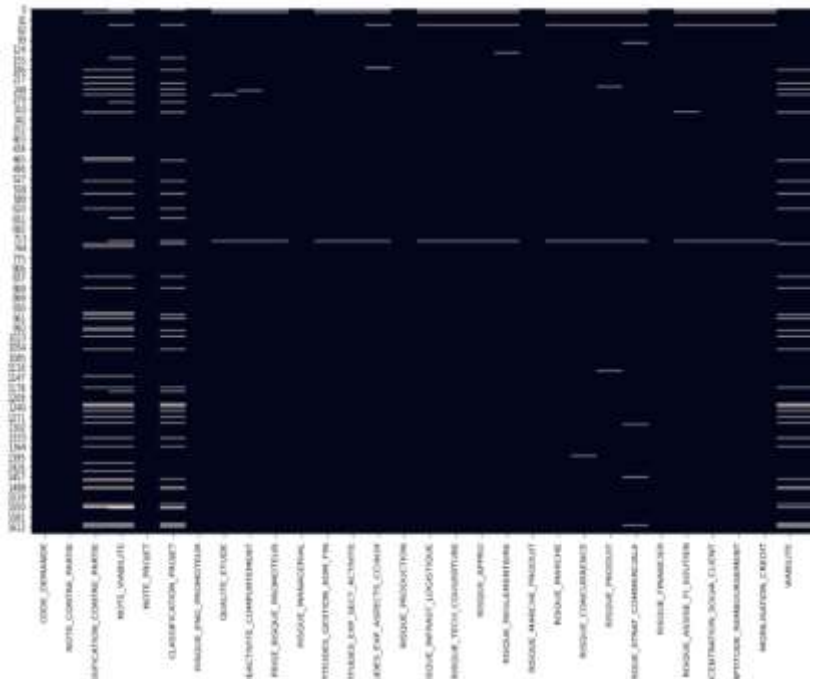
1. Nettoyage des données



Cette phase de préparation des données comprend des activités liées à la construction de l'ensemble de données à analyser. Il comprend ainsi la classification des données selon des critères choisis, le nettoyage des données, et surtout leur recodage pour les rendre compatibles avec les algorithmes qui seront utilisés.

1.1 Vérification des valeurs manquantes

```
CODE_DEMANDE 0
NOTE_CONTRE_PARTIE 0
CLASSIFICATION_CONTRE_PARTIE 142
NOTE_VIABILITE 161
NOTE_PROJET 0
CLASSIFICATION_PROJET 121
RISQUE_ENG_PROMOTEUR 0
QUALITE_ETUDE 6
REACTIVITE_COMPOTEMENT 10
PRISE_RISQUE_PROMOTEUR 7
RISQUE_MANAGERIAL 0
APTITUDES_GESTION_ADM_FIN 6
APTITUDES_EXP_SECT_ACTIVITE 6
APTITUDES_EXP_ASPECTS_CCIAUX 8
RISQUE_PRODUCTION 0
RISQUE_INFRAST_LOGISTIQUE 9
RISQUE_TECH_COUVERTURE 8
RISQUE_APPRO 8
RISQUE_REGLEMENTAIRE 13
RISQUE_MARCHE_PRODUIT 0
RISQUE_MARCHE 7
RISQUE_CONCURRENCE 8
RISQUE_PRODUIT 9
RISQUE_STRAT_COMMERCIALE 19
RISQUE_FINANCIER 0
RISQUE_ASSISE_FI_SOUTIEN 8
CONCENTRATION_SOLVA_CLIENT 7
APTITUDE_REMBOURSEMENT 7
MOBILISATION_CREDIT 8
VIABILITE 102
NOTE_TRANSACTION 102
dtype: int64
```



1.2 Vérification des lignes en double

```
sum(data.duplicated())==0
```

True

➡ Après la vérification on remarque qu'il n'y a pas des lignes dupliquées .

1.3 Pourcentage des valeurs manquantes de chaque colonne

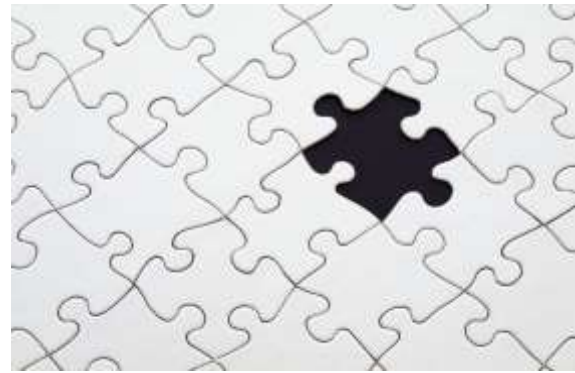
```
round(100*(data.isnull().sum()/len(data.index)),2)
```

CODE_DEMANDE	0.00
NOTE_CONTRE_PARTIE	0.00
CLASSIFICATION_CONTRE_PARTIE	8.68
NOTE_VIABILITE	9.84
NOTE_PROJET	0.00
CLASSIFICATION_PROJET	7.40
RISQUE_ENG_PROMOTEUR	0.00
QUALITE_ETUDE	0.37
REACTIVITE_COMPORTEMENT	0.61
PRISE_RISQUE_PROMOTEUR	0.43
RISQUE_MANAGERIAL	0.00
APTITUDES_GESTION_ADM_FIN	0.37
APTITUDES_EXP_SECT_ACTIVITE	0.37
APTITUDES_EXP_ASPECTS_CCIAUX	0.49
RISQUE_PRODUCTION	0.00
RISQUE_INFRAST_LOGISTIQUE	0.55
RISQUE_TECH_COUVERTURE	0.49
RISQUE_APPRO	0.49
RISQUE_REGLEMENTAIRE	0.79
RISQUE_MARCHE_PRODUIT	0.00
RISQUE_MARCHE	0.43
RISQUE_CONCURRENCE	0.49
RISQUE_PRODUIT	0.55
RISQUE_STRAT_COMMERCIALE	1.16
RISQUE_FINANCIER	0.00
RISQUE_ASSISE_FI_SOUTIEN	0.49
CONCENTRATION_SOLVA_CLIENT	0.43
APTITUDE_REMBOURSEMENT	0.43
MOBILISATION_CREDIT	0.49
VIABILITE	6.23

➡ On a remarqué que les pourcentages des colonnes sont tous inférieur à 10% ce qui nous permet d'utiliser l'imputation des données .

2.Imputation des données

L'imputation des données manquantes réfère au fait qu'on remplace les valeurs manquantes dans le jeu de données par des valeurs artificielles. Idéalement, ces remplacements ne doivent pas conduire à une altération sensible de la distribution et la composition du jeu de données.



```
note_reac = data1['REACTIVITE_COMPORTEMENT'].fillna(0)
note_reg = data1['RISQUE_REGLEMENTAIRE'].fillna(0)
note_strat = data1['RISQUE_STRAT_COMMERCIALE'].fillna(0)

qualite_etude=data1['QUALITE_ETUDE'].fillna(0)
prise_risque_prom=data1['PRISE_RISQUE_PROMOTEUR'].fillna(0)
risque_infra=data1['RISQUE_INFRASTR_LOGISTIQUE'].fillna(0)

aptitudes_gest_adm_fin=data1['APTITUDES_GESTION_ADM_FIN'].fillna(0)
aptitudes_exp_sect_act=data1['APTITUDES_EXP_SECT_ACTIVITE'].fillna(0)
aptitudes_exp_aspects_cciaux=data1['APTITUDES_EXP_ASPECTS_CCIAUX'].fillna(0)
risque_tech_couv=data1['RISQUE_TECH_COUVERTURE'].fillna(0)
risque_marche=data1['RISQUE_MARCHE'].fillna(0)
risque_appro=data1['RISQUE_APPRO'].fillna(0)
risque_concurrence=data1['RISQUE_CONCURRENCE'].fillna(0)
risque_prod=data1['RISQUE_PRODUIT'].fillna(0)

risque_assise_fi_soutien=data1['RISQUE_ASSISE_FI_SOUTIEN'].fillna(0)
risque_concentration_solva_client=data1['CONCENTRATION_SOLVA_CLIENT'].fillna(0)
risque_remboursement=data1['APTITUDE_REMBOURSEMENT'].fillna(0)
risque_mobilisation_credit=data1['MOBILISATION_CREDIT'].fillna(0)
```

➔ Pour résoudre ce problème on a appliqué la méthode “**fillna**” en effet on a remplacé les valeurs manquantes par 0 ou “NAN” afin d’éviter des lignes vides .

```
RISQUE_ENG_PROMOTEUR      0
RISQUE_MANAGERIAL         0
RISQUE_PRODUCTION         0
RISQUE_MARCHE_PRODUIT     0
RISQUE_FINANCIER          0
REACTIVITE_COMPORTEMENT   0
RISQUE_REGLEMENTAIRE      0
RISQUE_STRAT_COMMERCIALE  0
QUALITE_ETUDE             0
PRISE_RISQUE_PROMOTEUR    0
RISQUE_INFRASTR_LOGISTIQUE 0
APTITUDES_GESTION_ADM_FIN  0
APTITUDES_EXP_SECT_ACTIVITE 0
APTITUDES_EXP_ASPECTS_CCIAUX 0
RISQUE_TECH_COUVERTURE    0
RISQUE_MARCHE             0
RISQUE_APPRO              0
RISQUE_CONCURRENCE        0
RISQUE_PRODUIT            0
RISQUE_ASSISE_FI_SOUTIEN  0
CONCENTRATION_SOLVA_CLIENT 0
APTITUDE_REMBOURSEMENT    0
MOBILISATION_CREDIT       0
CLASSIFICATION_PROJET     0
dtype: int64
```

➔ Après imputation on a obtenu 0 valeurs manquantes dans chaque colonne.

3.Selection



La **sélection de feature** est un **processus** utilisé en apprentissage automatique et en traitement de données. Il consiste, étant donné des données dans un **espace de grande dimension**, à trouver un sous-ensemble de variables pertinentes.

```
data1=data.drop('CODE_DEMANDE', axis=1)
```

on a supprimé la colonne CODE_DEMANDE parce qu'elle n'a pas d'effet sur notre prédiction

👍 **Aperçu sur la base après avoir**

	NOTE_PROJET	RISQUE_ENG_PROMOTEUR	QUALITE_ETUDE	REACTIVITE_COMPORTEMENT	PRISE_RISQUE_PROMOTEUR	RISQUE_MANAGERIAL	A
0	48.00	24.25	3.0	3.0	1.0	24.0	
1	50.00	27.25	3.0	3.0	2.0	21.6	
2	44.00	23.50	2.0	3.0	2.0	26.4	
3	0.00	0.00	NaN	NaN	NaN	0.0	
4	54.00	20.50	2.0	3.0	1.0	24.0	
...
1631	43.00	23.50	2.0	3.0	2.0	26.4	
1632	61.00	30.25	3.0	3.0	3.0	36.0	
1633	75.00	37.00	4.0	3.0	4.0	60.0	
1634	42.67	19.83	2.0	1.0	3.0	33.6	
1635	38.00	16.75	1.0	3.0	1.0	24.0	

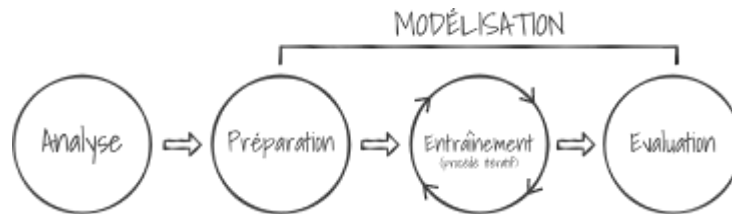
1636 rows x 24 columns

image 3 : nouvelle data

Après la suppression du CODE_DEMANDE ainsi les colonnes à prédire , notre nouvelle dataset contient 1636 lignes et 24 colonnes.

V. MODELISATION DES DONNEES

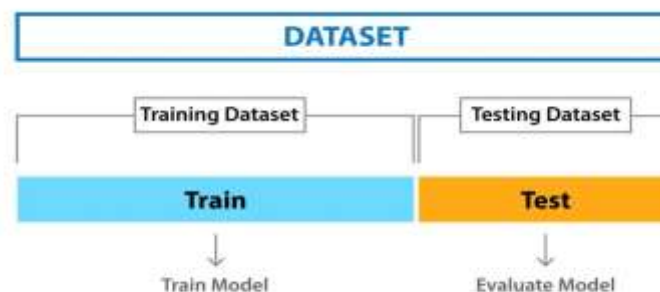
C'est la phase de **Data Science** proprement dite. La **modélisation** comprend le choix, le paramétrage et le test de **différents algorithmes** ainsi que leur enchaînement, qui constitue **un modèle**. Ce processus est d'abord descriptif pour générer de la connaissance, en expliquant pourquoi les choses se sont passées. Il devient ensuite prédictif en expliquant ce qu'il va se passer, puis prescriptif en permettant d'optimiser une situation future.



L'entraînement et le test de l'ensemble de données seront utilisés dans cette étude pour modéliser les données à l'aide de l'algorithme proposé.

Le rapport est de **80-20** pour **KNeighborsClassifier** et **Random Forest**, **60-40** pour **KNeighbors Regressor**.

Basé sur la théorie, la technique évaluera les modèles prédictifs en divisant l'original échantillon dans un ensemble d'apprentissage pour entraîner le modèle et un ensemble de test pour l'évaluer. La moyenne de tout le processus est produit comme résultat du modèle.



1- KNeighborsRegressor(KNN Regressor)

L'algorithme K-NN (K Nearest Neighbor) est une méthode d'apprentissage supervisé. Il peut être utilisé aussi bien pour la régression que pour la classification. Son fonctionnement peut être assimilé à l'analogie suivante "dis moi qui sont tes voisins, je te dirais qui tu es...".

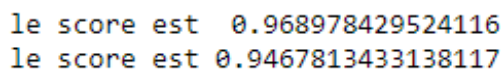
Pour effectuer une prédiction, l'algorithme K-NN ne va pas calculer un modèle prédictif à partir d'un Training Set comme c'est le cas pour la régression logistique ou la régression linéaire. En effet, K-NN n'a pas besoin de construire un modèle prédictif. Ainsi, pour K-NN il n'existe pas de phase d'apprentissage proprement dite. C'est pour cela qu'on le catégorise parfois dans le Lazy Learning. Pour pouvoir effectuer une prédiction, K-NN se base sur le jeu de données pour produire un résultat.

1.1 Cas d'utilisation

L'algorithme KNN Regressor peut être utilisé pour les exemples de cas suivants :

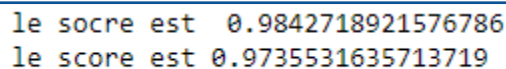
- La prédiction des prix des actions.
- Les systèmes de recommandation.
- L'analyse du risque de crédit.
- Planification prévisionnelle des voyages

En effet on a appliqué l'algorithme **K Neighbors Regressor** pour prédire **les notes de projet, contrepartie , viabilité** , ensuite on a entraîné nos trois modèles et finalement on a calculé les scores et on a obtenu **des scores d'entraînement égale à respectivement 0.96 , 0.98 , 0.76** et **des scores de test égale à respectivement 0.94,0.97,0.60** .



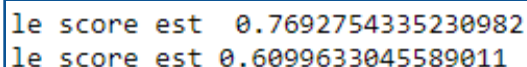
```
le score est 0.968978429524116
le score est 0.9467813433138117
```

image 4 : Score du modèle note de projet



```
le socre est 0.9842718921576786
le score est 0.9735531635713719
```

image 5 : Score du modèle note contrepartie



```
le score est 0.7692754335230982
le score est 0.6099633045589011
```

image 6 : Score du modèle note viabilité

2- KNeighborsClassifier(KNN Classifier)

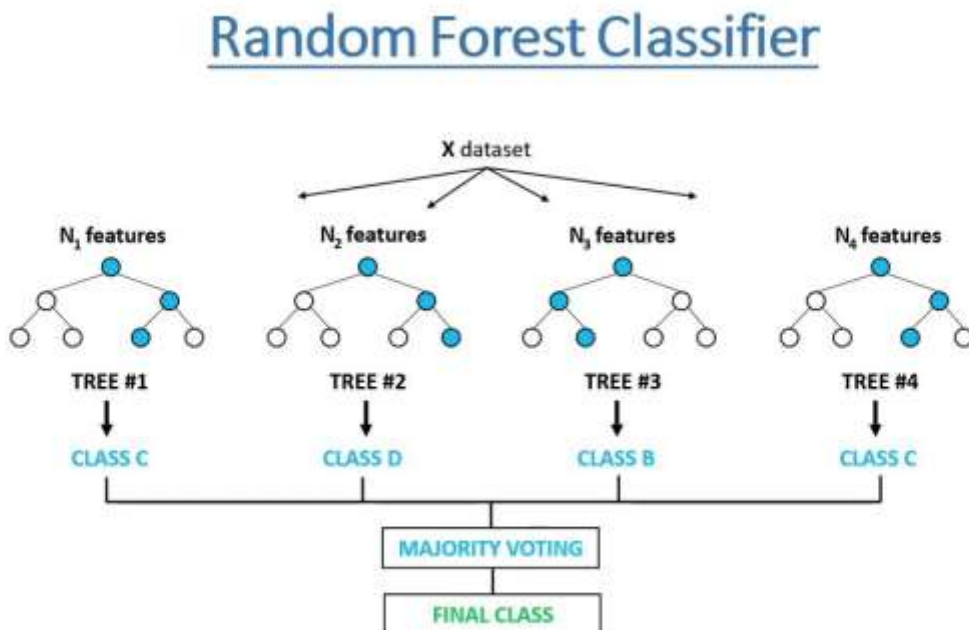
Comme on a travaillé avec KNeighbors Regressor pour prédire les notes de projet , contrepartie , viabilité maintenant on va utiliser KNeighbors Classifier afin de prédire la **Classification_Projet** , on a entraîné notre modèle et finalement on a calculé le score et on a obtenu comme **score d'entraînement 0.87 et comme score de test 0.82**

```
le score est 0.8730886850152905
le score est 0.8201219512195121
```

image 7 : Score du modèle Classification Projet

3-RandomForestClassifier (Forêt d'arbres décisionnels)

Les forêts d'arbres décisionnels (ou forêts aléatoires de l'anglais random forest classifier) font partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de bagging. L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.



3.1 Cas d'utilisation

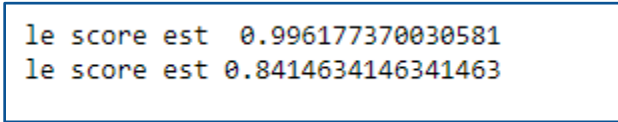
L'algorithme de forêt aléatoire est utilisé dans de nombreux domaines différents, comme la banque, la bourse, la médecine et le commerce électronique.

En finance, par exemple, il est utilisé pour détecter les clients plus susceptibles de rembourser leur dette à temps, ou d'utiliser plus fréquemment les services d'une banque. Dans ce domaine, il est également utilisé pour détecter les fraudeurs qui cherchent à arnaquer la banque. En trading, l'algorithme peut être utilisé pour déterminer le comportement futur d'une action.

Dans le domaine de la santé, il est utilisé pour identifier la bonne combinaison de composants en médecine et pour analyser les antécédents médicaux d'un patient afin d'identifier les maladies.

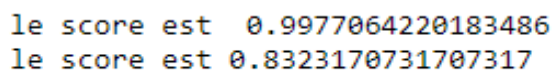
La forêt aléatoire est utilisée dans le commerce électronique pour déterminer si un client aimera réellement le produit ou non.

En effet on a appliqué l'algorithme **Random Forest Classifier** pour prédire **la classification_Contre_Partie et la Viabilité** ensuite on a entraîné nos deux modèles et finalement on a calculé les scores et on a obtenu **des scores d'entraînement égale à respectivement 0.99 , 0.99** et **des scores de test égale à respectivement 0.84, 0.83** .



```
le score est 0.996177370030581
le score est 0.8414634146341463
```

image 8 : Score du modèle classification_contre_partie

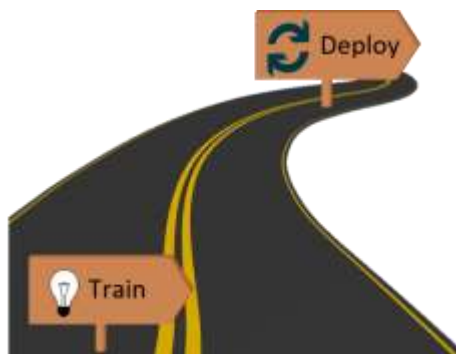


```
le score est 0.9977064220183486
le score est 0.8323170731707317
```

image 9 : Score du modèle viabilité

VI. DEPLOIEMENT

Il s'agit de l'étape finale du processus. Elle consiste en une mise en production pour les utilisateurs finaux des modèles obtenus. Son objectif : mettre la connaissance obtenue par la modélisation, dans une forme adaptée, et l'intégrer au processus de prise de décision. Le déploiement peut ainsi aller, selon les objectifs, de la simple génération d'un rapport décrivant les connaissances obtenues jusqu'à la mise en place d'une application, permettant l'utilisation du modèle obtenu, pour la prédiction de valeurs inconnues d'un élément d'intérêt.



1- Logiciel de déploiement

Django est un cadre de développement web open source en Python. Il a pour but de rendre le développement web 2.0 simple et rapide. Pour cette raison, le projet a pour slogan « Le framework pour les perfectionnistes avec des deadlines. ». Développé en 2003 pour le journal local de Lawrence (État du Kansas, aux États-Unis), Django a été publié sous licence BSD à partir de juillet 2005, Django est un cadre de développement qui s'inspire du principe MVC ou MTV (la vue est gérée par un gabarit) .



2-Réalisation de l'application

Tout d'abord , on a décidé de développer une application afin d'aider les banques à prédire le scoring des clients plus précisément les notes de projet , contrepartie , viabilité ainsi la classification projet , contrepartie et la viabilité des clients .

Notre application est développée avec le framework django ou on a déployé nos modèles .

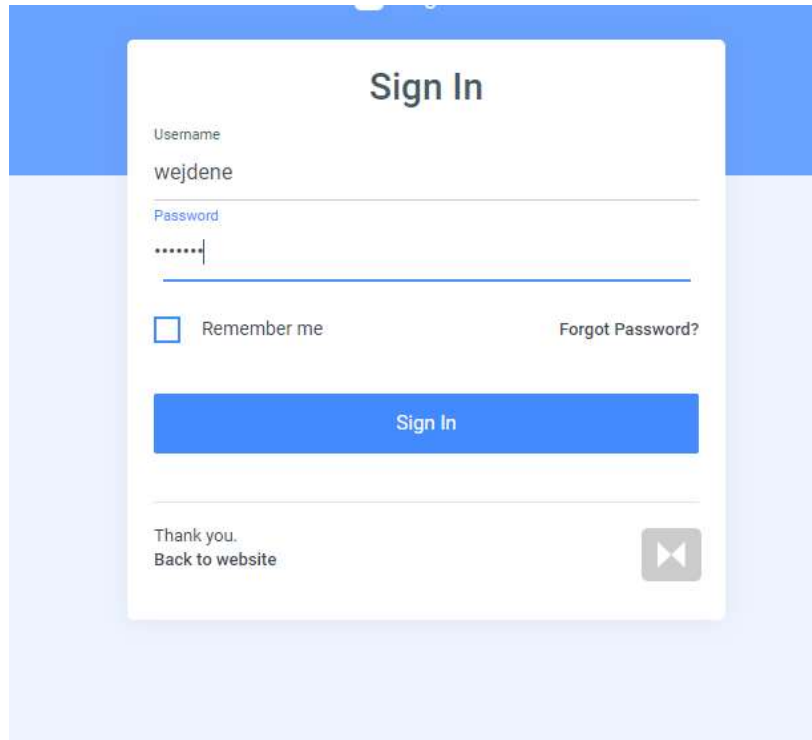
A screenshot of a web application's login page. The page has a blue header and a light blue background. A white card in the center contains the 'Sign In' form. The form has two input fields: 'Username' with the text 'wejdene' and 'Password' with masked characters '.....'. Below the password field is a checkbox labeled 'Remember me' and a link 'Forgot Password?'. A blue 'Sign In' button is at the bottom of the form. Below the button, it says 'Thank you.' and 'Back to website' with a small icon.

image 10 : Interface Login

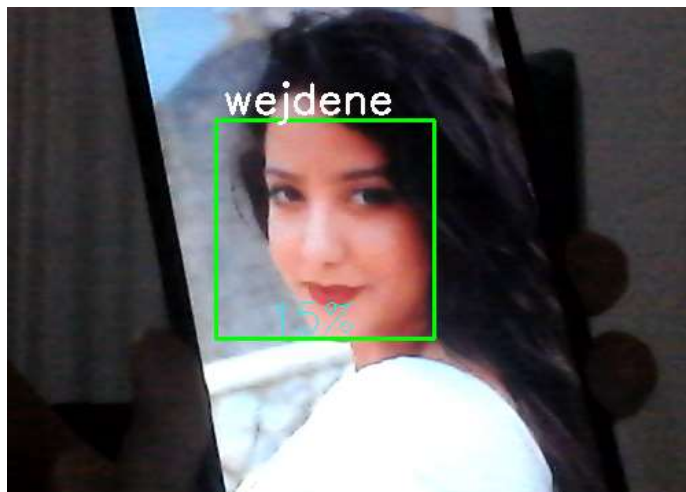


image 11 : Interface reconnaissance Faciale

La première étape , l'administrateur doit se connecter en ajoutant son login et son mot de passe, ainsi il doit faire la reconnaissance faciale afin de mieux sécuriser notre application.

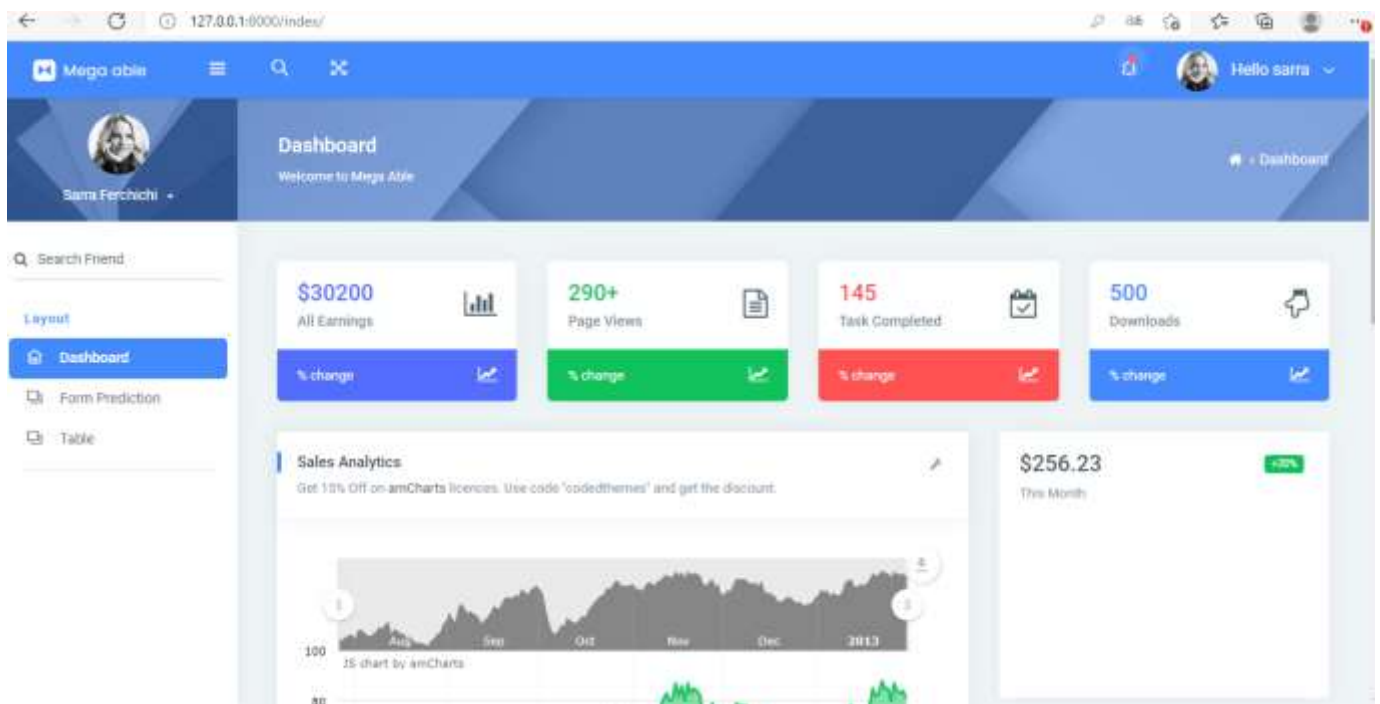


image 12 : Interface home

The screenshot shows the 'Form Prediction' interface. It features a blue header bar with the Mega Able logo and navigation icons. A sidebar on the left provides navigation options: 'Search Friend', 'Layout', 'Dashboard', 'Form Prediction' (selected), and 'Table'. The main content area is titled 'Enter all expected inputs' and contains two columns of input fields. The left column has a text input field with the value '4061998' and several numeric input fields with values: 24.25, 24.00, 18.00, 14.00, 16.00, 3, 3, 2, and 3. The right column has a series of numeric input fields with values: 4, 3, 2, 2, 3, 3, 2, 2, 2, and 2.

image 12 : Interface formulaire prédiction

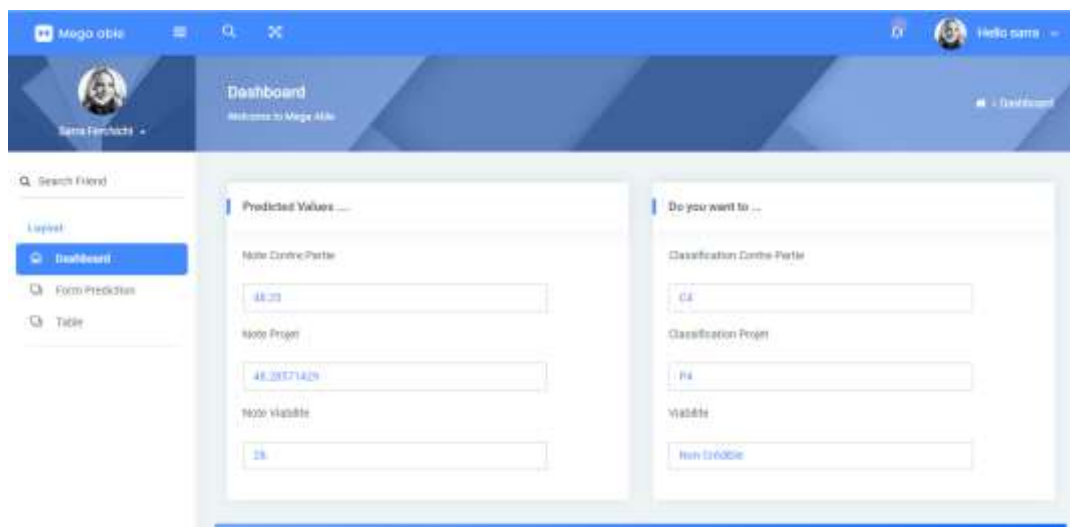


image 13 : Interface resultat (Notation)

Sara Ferchichi

Dashboard

Welcome to Mega Able

Search Friend

Layout

Dashboard

Form Prediction

Table

Contextual Classes

CODE DEMANDE	NOTE CONTRE PARTIE	CLASSIFICATION CONTRE PARTIE	NOTE PROJET	CLASSIFICATION PROJET	NOTE VIABILITE	VIABILITE
123	48.25	C4	48.28571429	P4	28.0	Non Credible
201	48.58	C4	51.0	P4	28.0	Non Credible
305	48.58	C4	51.0	P4	28.0	Non Credible
321	48.25	C4	48.28571429	P4	28.0	Non Credible
555	48.58	C4	51.0	P4	28.0	Non Credible

image 14 : Interface Tableau (historique des résultats)

26

CONCLUSION AND PERSPECTIVES

La notation de crédit est la base des institutions financières pour prendre des décisions d'octroi de crédit . Un bon modèle de notation de crédit sera en mesure de regrouper efficacement les clients dans groupe par défaut ou non par défaut. Plus il est efficace, plus le coût peut être économisé pour une institution financière.

La complexité du processus de prêt est l'un des cas les plus discutables dans le secteur bancaire

secteur et le processus d'automatisation est la solution pour moins de stress et plus de productivité. Notre produit aide les banques afin de prendre des décisions auprès des clients .

Pour améliorer notre produit, nous pouvons intégrer un système de chatbot qui aidera la banque à prendre la bonne décision.

Ce projet nous a permis de travailler de manière approfondie non seulement dans le domaine de la science des données, mais également dans celui des banques .