

# **MSPR BLOC.3**

## **Big Data & Business Intelligence**

**Thème : Analyses et Prédictions pour les  
Prochaines Élections Présidentielles**

**Membres :**

SEHAKI Sofiane  
GHERSBRAHAM Anis  
MERBAH Yanis  
DIANTOUADI-FRAY Kevine

**Encadré par :**

M. Rakib



## Table des matières :

Partie 1 : Introduction .....	3
1. Présentation de l'entreprise et de son activité.....	3
2. Objectif .....	3
Partie 2 : Plan de travaille.....	4
1) Choix de la zone géographique .....	4
2) Choix des critères.....	5
3) La démarche suivie et les méthodes employées .....	7
4) Les modèles testés.....	10
5) Conclusion .....	14

## Liste de figures :

FIGURE 1 RESULTAT DES ELECTIONS PRESIDENTIELS EN VAL D'OISE EN FRANCE.....	4
FIGURE 2 ARCHITECTURE DE DONNEES.....	6
FIGURE 3 CAPTURE D'ECRAN DE L'OUTIL TRELLO AU DEBUT DU PROJET.....	9
FIGURE 4 CAPTURE D'ECRAN DE L'OUTIL TRELLO AU MILIEU DU PROJET.....	9
FIGURE 5 CAPTURE D'ECRAN DE L'OUTIL TRELLO A LA FIN DU PROJET.....	10
FIGURE 6 LES RESULTATS DES MODELES 'RANDOM FOREST' .....	11
FIGURE 7 LES RESULTATS DU MODELE 'REGRESSION LOGISTIQUE' .....	12
FIGURE 8 LES RESULTATS DU MODELE 'MLP' .....	13

## Partie 1 : Introduction

### 1. Présentation de l'entreprise et de son activité

Jean-Edouard de la Motte Rouge a créé une start-up spécialisée dans le conseil sur la thématique des campagnes électorales.

La start-up comprend un expert en analyse politique, un business développeur, et un assistant.

Il souhaite pouvoir prédire, grâce à l'intelligence artificielle, les tendances des élections à venir, en se basant sur un certain nombre d'indicateurs, comme la sécurité, l'emploi, la vie associative, la population, la vie économique (nombre d'entreprises), la pauvreté...etc.

### 2. Objectif

L'entreprise nous a mandatés pour réaliser une preuve de concept (POC) visant à démontrer la possibilité d'utiliser l'intelligence artificielle pour prédire les tendances électorales futures dans une zone géographique donnée, en s'appuyant sur divers indicateurs économiques, sociaux et démographiques.

Cette POC a pour objectif de valider la faisabilité d'un modèle prédictif qui permettrait aux candidats, partis politiques et autres acteurs de mieux comprendre les attentes et préférences des électeurs locaux. Un tel outil offrirait la possibilité de cibler plus efficacement les campagnes électorales en les adaptant aux réalités socio-économiques propres à chaque région.

Au-delà des acteurs politiques, cette solution prédictive présenterait également un intérêt pour les organisations œuvrant dans les domaines de la démocratie et de la gouvernance. Elle leur permettrait d'anticiper les résultats électoraux et d'identifier les enjeux socio-économiques susceptibles d'influencer les comportements de vote.

## Partie 2 : Plan de travail

### 1) Choix de la zone géographique :

Pour notre projet, nous avons décidé de maintenir le département du Val d'Oise comme zone géographique d'étude. Ce choix s'appuie sur la représentativité socio-économique et démographique de ce territoire au regard du profil national.

Situé en Île-de-France, le Val d'Oise bénéficie d'une diversité démographique remarquable, avec une population cosmopolite aux origines culturelles et ethniques variées. Cette pluralité permettra d'analyser les attentes de différents groupes et d'assurer une meilleure représentativité de notre modèle prédictif.

De plus, le Val d'Oise combine à la fois des zones urbaines denses, telles que Cergy ou Argenteuil, et des territoires ruraux et semi-ruraux. Cette coexistence d'environnements urbains et ruraux reflète les réalités contrastées de la population française en matière de niveaux de vie et de développement socio-économique local.

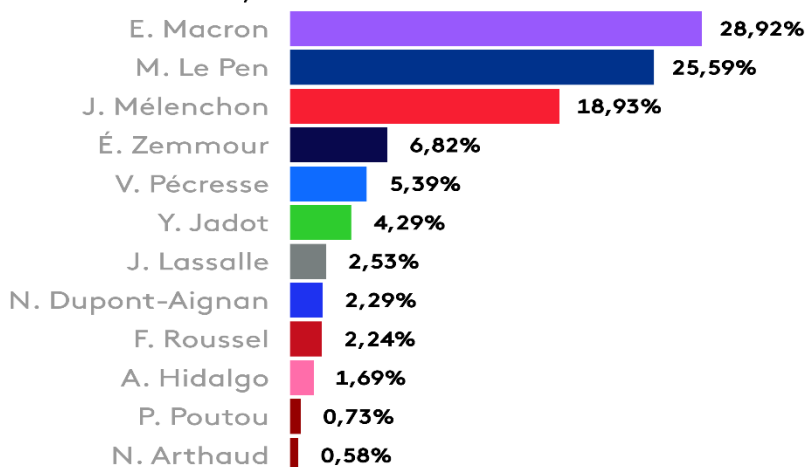
Sur le plan politique, le Val d'Oise joue un rôle clé dans les enjeux nationaux, du fait de sa proximité avec la capitale et des importants bassins de population qu'il abrite. Les tendances de vote dans ce département peuvent souvent présager des résultats à l'échelle nationale.

Ainsi, le choix du Val d'Oise comme terrain d'étude apporte une représentativité géographique, démographique, socio-économique et politique adaptée pour mener à bien cette preuve de concept et en assurer la pertinence au niveau national.

La figure ci-dessous illustre les résultats du 1<sup>er</sup> tour des élections présidentielles 2022 en Val d'Oise en France. Les couleurs représentent les différents candidats, et la taille des barres correspond à la proportion de voix obtenues par chaque candidat

## Loiret : résultats du 1er tour

Abstention : 25,02%



Source : Ministère de l'Intérieur. Crédits : franceinfo

*Figure 1 Résultat des élections présidentiels en Val d'Oise en France*

## 2) Choix des critères

Dans ce projet, nous désirons exploiter l'intelligence artificielle pour anticiper les tendances des élections futures, en nous appuyant sur divers indicateurs socio-économiques et politiques clés :

**L'éducation** : Un enjeu majeur qui façonne les orientations des électeurs. Les propositions sur l'investissement dans l'éducation publique, l'accès équitable à l'éducation et la réforme du système éducatif peuvent avoir un impact déterminant. Les partis mettant en avant des politiques éducatives progressistes, garantes de l'égalité des chances, séduiront un électorat soucieux de l'amélioration de l'éducation.

**La santé** : La qualité et l'accessibilité des services de santé sont des préoccupations centrales pour de nombreux citoyens. Les électeurs seront attentifs aux mesures proposées par les partis concernant l'accès aux soins, le financement de la protection sociale, la gestion des crises sanitaires, etc. Une offre de soins de qualité, abordable et inclusive, sera un atout électoral.

**Le taux d'activité** : Indicateur du dynamisme économique, les fluctuations du taux d'activité (part des 15-64 ans sur le marché du travail)

peuvent orienter les choix électoraux. Une hausse tendrait à favoriser les partis de gauche, généralement perçus comme défenseurs des travailleurs et de la protection sociale.

**Le pouvoir d'achat :** Préoccupation récurrente pour les ménages, les politiques économiques impactant directement le niveau de vie (salaires, coût de la vie, croissance, etc.) seront décisives. Les formations apportant des réponses concrètes pour préserver/augmenter le pouvoir d'achat séduiront cette frange de l'électorat.

**L'emploi :** Enjeu social majeur, les propositions sur l'emploi (création d'emplois, réduction du chômage, formation, etc.) mobiliseront les suffrages. Les partis avec un programme économique crédible et favorable à l'emploi pourront capter les attentes de ces franges de la population.

**Le logement :** Dans les zones tendues, l'accès au logement est une problématique sensible qui orientera les votes. Les mesures pour développer une offre de logements abordables et réguler le marché immobilier seront scrutées de près par cet électorat.

**La sécurité :** Sujet parfois clivant, les politiques de sécurité (maintien de l'ordre, lutte anti-terroriste et anti-délinquance) influenceront certains segments de l'électorat en quête de fermeté en la matière.

**La politique sociale :** Les attentes en termes de protection sociale, de droits sociaux, de lutte contre les inégalités seront déterminantes pour une partie de l'électorat. Les partis porteurs de politiques sociales progressistes et égalitaires pourront capter ces suffrages.

**L'immigration :** Thématique récurrente des campagnes, les propositions sur l'immigration (flux migratoires, intégration, contrôles, etc.) façonneront les comportements d'une frange de l'électorat, plus ou moins ouverte sur cette question.

### 3) La démarche suivie et les méthodes employées :

#### 1. Architecture de données :

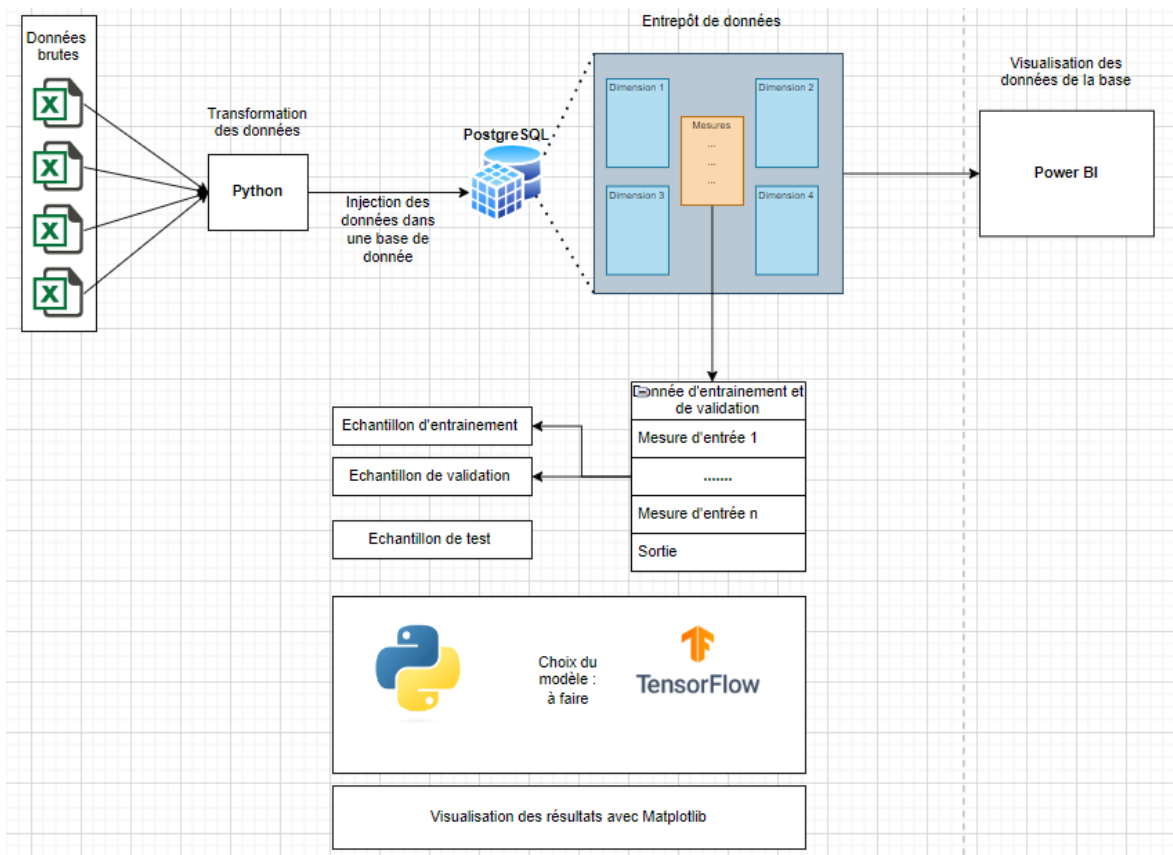


Figure 2 Architecture de données

#### Recherche et collecte des données :

- Identifier les sources de données pertinentes, telles que les résultats électoraux passés, les données démographiques, les données socio-économiques, etc.
- Explorer les bases de données publiques disponibles, telles que celles fournies par les institutions gouvernementales ou les organismes de sondage.

(<https://www.data.gouv.fr>),

(<https://www.insee.fr/fr/statistiques/6673769?sommaire=2500477> ).

### **Visualisation avant transformation :**

- Utiliser des techniques de visualisation de données, telles que les graphiques, les diagrammes ou les cartes, pour explorer les données collectées.
- Identifier les schémas, les tendances ou les anomalies potentielles dans les données afin de guider les prochaines étapes.

### **Nettoyage et transformation des données :**

- Traiter les valeurs manquantes, les doublons ou les erreurs dans les données collectées.
- Normaliser les variables pour les rendre comparables et cohérentes.
- Effectuer des opérations de manipulation de données, telles que le regroupement, le filtrage ou la création de nouvelles variables, pour préparer les données à l'étape suivante.

### **Stockage des données :**

- Utiliser une base de données ou un système de gestion de données adapté pour stocker les données nettoyées.
- Assurer la sécurité et la confidentialité des données conformément aux réglementations en vigueur.

### **Visualisation avant modélisation :**

- Effectuer une nouvelle visualisation des données préparées pour mieux comprendre les relations et les corrélations entre les variables.
- Identifier les caractéristiques les plus influentes ou significatives qui pourraient être utilisées dans la modélisation.

### **Choix du modèle IA adapté :**



- Explorer différentes techniques d'apprentissage automatique adaptées à votre sujet, telles que les modèles de régression, les modèles de classification ou les modèles de séries temporelles.
- Considérer des modèles d'IA avancés tels que les réseaux de neurones profonds (Deep Learning) ou les méthodes d'apprentissage ensembliste si nécessaire.

#### **Modélisation :**

- Diviser les données en ensembles d'entraînement et de test pour évaluer les performances du modèle.
- Entraîner le modèle d'IA en utilisant les données d'entraînement et ajuster ses paramètres pour maximiser sa précision et sa généralisation.
- Effectuer une validation croisée pour évaluer les performances du modèle sur des données non utilisées lors de l'entraînement.

#### **Prédiction :**

- Utiliser le modèle d'IA entraîné pour faire des prédictions sur de nouvelles données.
- Évaluer les performances du modèle en comparant les prédictions avec les résultats réels des élections présidentielles.
- Analyser les résultats obtenus et itérer si nécessaire pour améliorer la précision du modèle

## **2. Organisation d'équipe :**

Nous avons utilisé l'outil visuel « Trello » qui a permis à notre équipe de gérer le projet.

Après avoir listées toutes les tâches à faire, nous avons assignée à chaque membre du groupe, selon leurs points forts une tâche spécifique à faire.



Figure 3 Capture d'écran de l'outil Trello au début du projet



Figure 4 Capture d'écran de l'outil Trello au milieu du projet

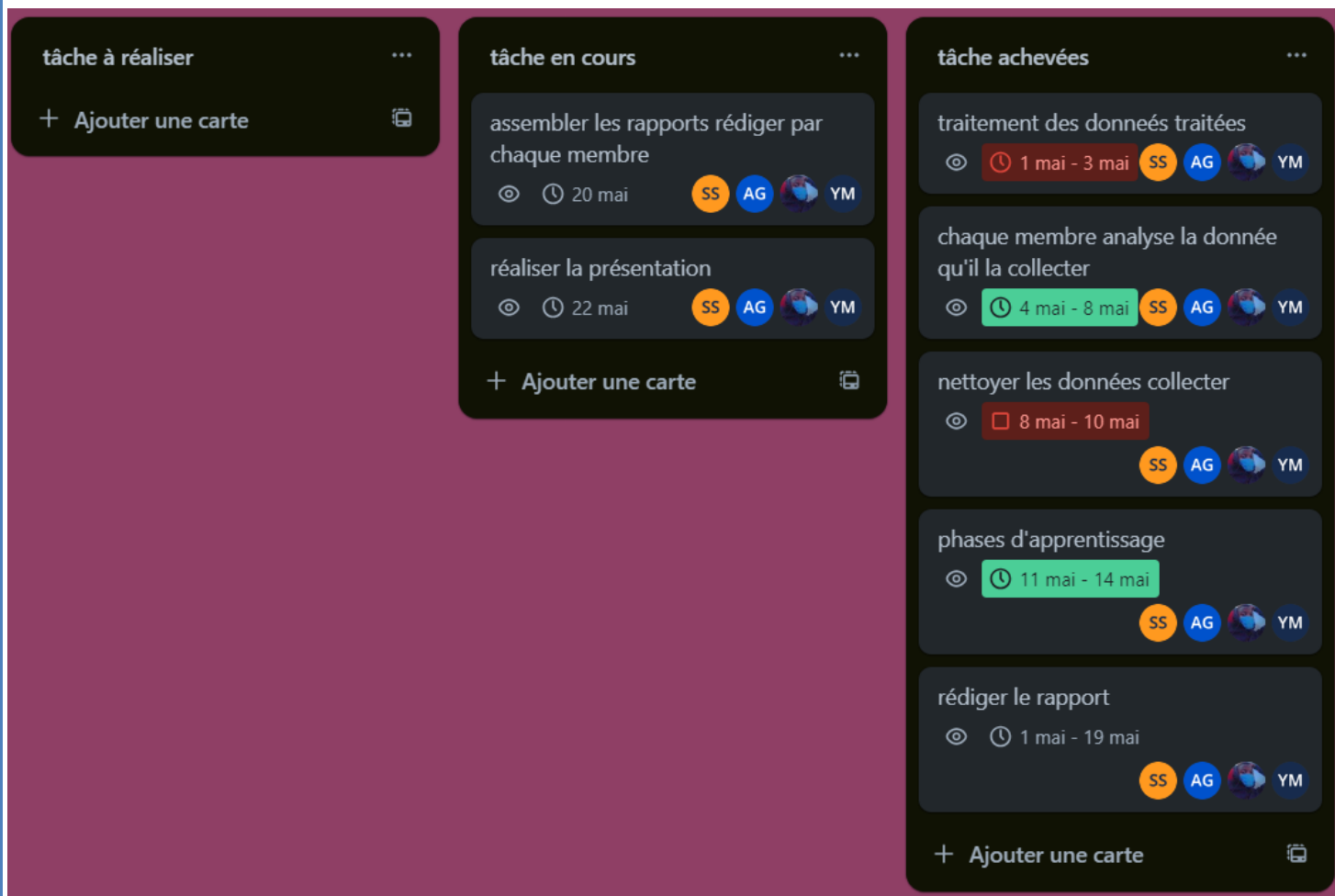


Figure 5 Capture d'écran de l'outil Trello à la fin du projet

### 3. Collaboration et communication :

Pour s'assurer d'avoir un bon suivi sur notre travail, nous avons organisé plusieurs réunions qu'elle soit en ligne (groupe Discord, appel téléphoniques et des discussions WhatsApp) ou en présentiels pour faire le point sur l'avancement de chacun de nos travaux.

### 4) Les modèles testés :

Il existe de nombreux modèles et techniques qui sont largement utilisés dans le domaine de la prédiction et qui sont adaptés à des problèmes spécifiques. Le choix de l'algorithme dépendra du type de données, de la nature du problème et des performances requises.

## 1. Définition des modèles :

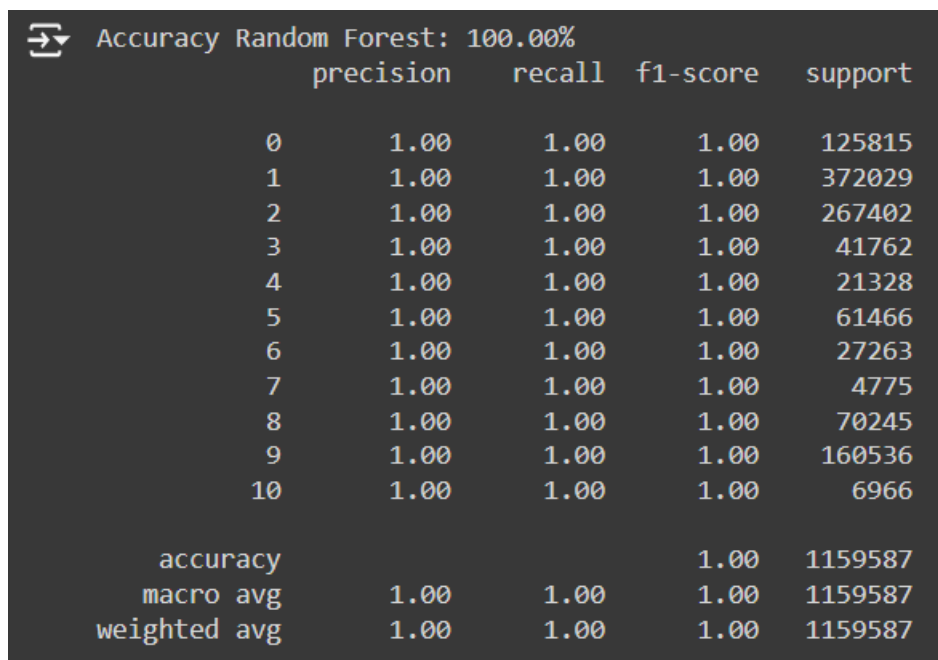
Voici quelques modèles utilisés pour la prédiction dans notre projet :

**Random Forest** : Une méthode d'ensemble qui combine plusieurs arbres de décision pour la prédiction. Chaque arbre est construit sur un sous-ensemble aléatoire des données et utilise une combinaison aléatoire des variables.

**Deep Learning** : Ce modèle utilise des réseaux de neurones profonds pour apprendre des structures complexes et effectuer des prédictions. Nous l'avons utilisé pour capturer des relations plus complexes entre les critères et les taux de vote.

**Régression logistique multi-classe** : Utilisée pour la prédiction de variables binaires, elle modélise la probabilité de succès en fonction des variables d'entrée à l'aide d'une fonction logistique

## 2. Les résultats obtenus :



Accuracy Random Forest: 100.00%				
	precision	recall	f1-score	support
0	1.00	1.00	1.00	125815
1	1.00	1.00	1.00	372029
2	1.00	1.00	1.00	267402
3	1.00	1.00	1.00	41762
4	1.00	1.00	1.00	21328
5	1.00	1.00	1.00	61466
6	1.00	1.00	1.00	27263
7	1.00	1.00	1.00	4775
8	1.00	1.00	1.00	70245
9	1.00	1.00	1.00	160536
10	1.00	1.00	1.00	6966
accuracy			1.00	1159587
macro avg	1.00	1.00	1.00	1159587
weighted avg	1.00	1.00	1.00	1159587

Figure 6 Les résultats des modèles 'Random Forest'

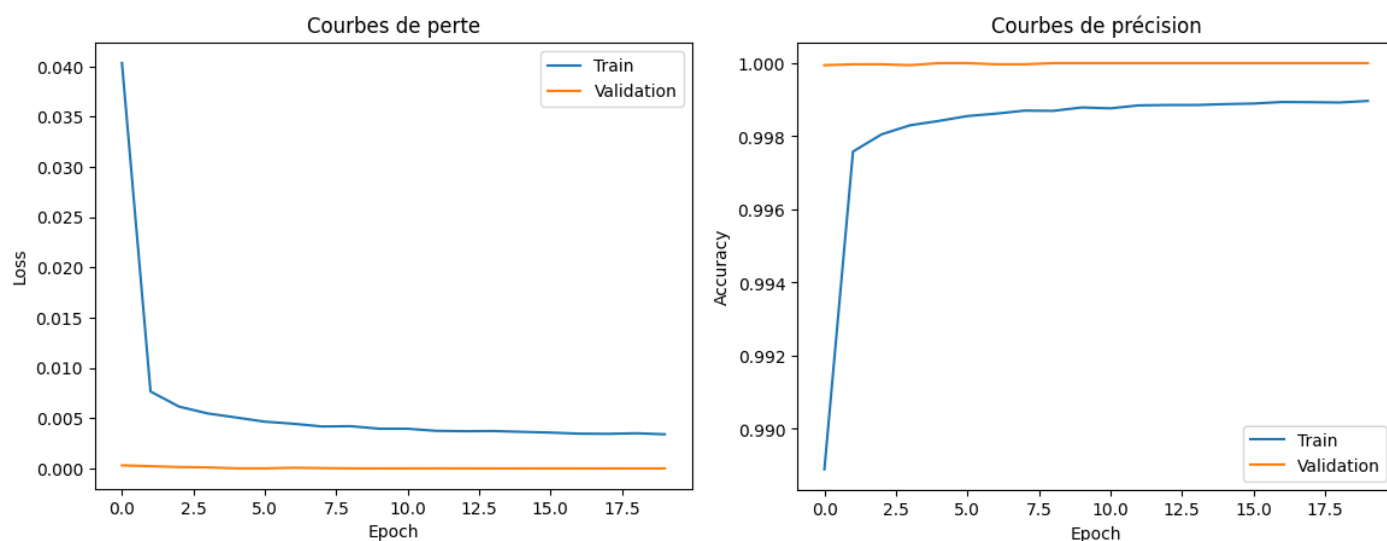
	precision	recall	f1-score	support
0	0.93	0.91	0.92	125815
1	0.92	0.96	0.94	372029
2	0.94	0.97	0.96	267402
3	0.91	0.61	0.73	41762
4	0.95	0.96	0.96	21328
5	0.81	0.69	0.75	61466
6	1.00	0.99	0.99	27263
7	0.00	0.00	0.00	4775
8	0.83	0.88	0.86	70245
9	0.83	0.86	0.84	160536
10	0.00	0.00	0.00	6966
accuracy			0.90	1159587
macro avg	0.74	0.71	0.72	1159587
weighted avg	0.89	0.90	0.90	1159587

*Figure 7 Les résultats du modèle 'Régression logistique'*

```

Epoch 12/20
36238/36238 [=====] - 129s 4ms/step - loss: 0.0037 - accuracy: 0.9988 -
Epoch 13/20
36238/36238 [=====] - 124s 3ms/step - loss: 0.0037 - accuracy: 0.9989 -
Epoch 14/20
36238/36238 [=====] - 124s 3ms/step - loss: 0.0037 - accuracy: 0.9989 -
Epoch 15/20
36238/36238 [=====] - 130s 4ms/step - loss: 0.0036 - accuracy: 0.9989 -
Epoch 16/20
36238/36238 [=====] - 137s 4ms/step - loss: 0.0036 - accuracy: 0.9989 -
Epoch 17/20
36238/36238 [=====] - 131s 4ms/step - loss: 0.0035 - accuracy: 0.9989 -
Epoch 18/20
36238/36238 [=====] - 120s 3ms/step - loss: 0.0034 - accuracy: 0.9989 -
Epoch 19/20
36238/36238 [=====] - 120s 3ms/step - loss: 0.0035 - accuracy: 0.9989 -
Epoch 20/20
36238/36238 [=====] - 119s 3ms/step - loss: 0.0034 - accuracy: 0.9990 -
36238/36238 [=====] - 60s 2ms/step - loss: 7.9347e-07 - accuracy: 1.0000
Accuracy MLP: 100.00%

```



*Figure 8 Les résultats du modèle 'MLP'*

### 3. Comparaison des modèles par rapport aux résultats :

Critère	Random Forest	Multilayer Perceptron (MLP)	Logistic Regression
Accuracy	100.00%	99.90% - 100.00%	90.00%
Precision	Très élevée (1.00 pour toutes les classes)	Très élevée (indicative des faibles pertes)	Varie par classe, jusqu'à 95% pour certaines classes
Recall	Très élevée (1.00 pour toutes les classes)	Très élevée (indicative des faibles pertes)	Varie par classe, certaines classes avec faibles valeurs
F1-Score	Très élevée (1.00 pour toutes les classes)	Très élevée (indicative des faibles pertes)	Varie par classe, certaines classes avec faibles valeurs
Loss	Non applicable directement, mais des faiblesses pour certaines classes	Très faible, indiquant un bon ajustement	Non applicable directement, mais des faiblesses pour certaines classes
Robustesse (Validation Croisée)	Très élevée (stratifiée sur 10 plis)	Très élevée (indicative des faibles pertes)	Moyenne (74% à 89%)
Overfitting	Moins probable	Possibilité d'overfitting (à vérifier)	Moins probable
Temps d'entraînement	Relativement rapide	Plus long en raison des nombreuses époques	Rapide
Utilisation des Données	Bonne performance sur des ensembles diversifiés	Performance excellente, mais nécessite validation croisée pour robustesse	Bonne performance sur des ensembles diversifiés
Complexité du Modèle	Moyenne (ensemble d'arbres décisionnels)	Élevée (réseaux de neurones profonds)	Faible (modèle linéaire)
Scénario Idéal	Vérification et robustesse sur des données diverses	Précision élevée et ajustement fin sur des données spécifiques	Scénarios simples, avec moins de classes

Dans notre comparaison des modèles d'apprentissage automatique, nous avons évalué plusieurs critères essentiels tels que l'accuracy, la précision, le recall, le F1-Score, la loss, la robustesse (validée par la validation croisée), la tendance à l'overfitting, le temps d'entraînement, l'utilisation des données, la complexité du modèle et le scénario idéal. Chaque modèle a ses propres forces et faiblesses, ce qui nous a permis de prendre des décisions éclairées sur le choix du modèle le plus approprié pour notre problème de prédiction des taux de vote dans le Val d'Oise.

## 5) Conclusion :

Suite à l'entraînement de différents modèles d'apprentissage automatique tels que la régression logistique multi-classe, les forêts aléatoires (Random Forest) et l'apprentissage profond (Deep Learning), nous avons identifié que les modèles de forêts aléatoires et d'apprentissage profond offraient les meilleures performances prédictives concernant les taux de vote dans le Val d'Oise. En particulier, le modèle Random Forest a montré une précision, un rappel et un F1-score parfaits, tandis que le modèle d'apprentissage profond a également atteint des performances très élevées avec une faible perte et une grande précision. Cependant, il convient de souligner que ces modèles plus complexes nécessitent des volumes de données d'entraînement conséquents pour exploiter pleinement leurs capacités et peuvent être sujets au sur ajustement. La régression logistique, bien que moins complexe, offre une solution rapide et efficace pour des scénarios plus simples avec des ensembles de données moins vastes. Pour des prédictions précises et robustes des taux de vote, les modèles Random Forest et d'apprentissage profond sont les meilleurs choix.

Le lien GitHub vers le projet MSPR : <https://github.com/sof-prog/MSPR-1.git>