# Projet STT 3795

Olivier Déry-Prévost 20214548
et
Sidya Galakho 20207299
28 avril 2024
Prédiction des résultats de match de soccer

# 1. Introduction

Le soccer est l'un des sports les plus populaires au monde, avec des millions de fans qui suivent les matchs et les compétitions avec passion. Cependant, prédire le résultat d'un match de soccer est une tâche complexe qui dépend de nombreux facteurs, tels que la force des équipes, la stratégie des entraîneurs, les performances des joueurs et les conditions du terrain. Les équipes de soccer sont constamment à la recherche de moyens pour améliorer leurs chances de victoire, tandis que les bookmakers et les parieurs cherchent à anticiper les résultats pour prendre des décisions éclairées.

Malgré l'importance de la prédiction des résultats de matchs de soccer, c'est difficile à cause de la complexité du jeu et de la multitude de facteurs qui influent sur le résultat. Les modèles de prédiction traditionnels, basés sur des analyses statistiques et des règles de décision, ont des limitations importantes et ne peuvent pas prendre en compte la totalité des facteurs qui influent sur le résultat d'un match.

Dans ce contexte, l'apprentissage automatique et l'analyse de données offrent de nouvelles perspectives pour améliorer la prédiction des résultats de matchs de soccer. En utilisant des techniques pour analyser de grandes quantités de données, il est possible d'identifier les facteurs clés qui influencent  le résultat d'un match et de développer des modèles de prédiction plus précis.

# 2. Objectifs

Les objectifs de ce projet sont clairs et ambitieux. Nous visons à développer un modèle prédictif robuste et fiable qui puisse aider les équipes de soccer, les entraîneurs et les parieurs à prendre des décisions éclairées.

**Développement d'un modèle prédictif robuste :**

Créer un modèle capable de prédire les résultats des matchs de soccer avec une bonne précision, en utilisant des techniques d'analyse de données avancées telles que l'Analyse en Composantes Principales (PCA) et des algorithmes de classification.

**Identification de facteurs de performance clés :**

Identifier les facteurs qui influent sur le succès des équipes, tels que les attributs des joueurs, les performances passées et d'autres variables pertinentes. Nous cherchons à comprendre comment ces facteurs interagissent entre eux et comment ils influent sur le résultat d'un match.

Pour atteindre cet objectif, nous allons utiliser des techniques d'analyse de données telles que l'analyse de régression et l'analyse de variance. Nous allons utiliser des visualisations de données pour identifier les tendances et les patterns dans les données.Nous allons également utiliser des techniques de sélection de variables pour identifier les facteurs les plus importants qui influent sur le résultat d'un match.

**Évaluation de la fiabilité des prédictions :**

Évaluer la fiabilité de nos prédictions en utilisant des techniques de validation croisée et d'évaluation des performances, pour mesurer l'efficacité de notre modèle et identifier les domaines où des améliorations sont nécessaires.

Pour atteindre cet objectif, nous allons utiliser des métriques d'évaluation telles que la précision, la recall, la F1-score.Nous allons aussi utiliser des techniques de validation croisée pour évaluer la performance de notre modèle sur des données inconnues.

En résumé, les objectifs de ce projet sont de développer un modèle prédictif robuste et fiable, d'identifier les facteurs de performance clés et d'évaluer la fiabilité des prédictions. Nous sommes convaincus que ce projet peut apporter une valeur ajoutée significative à la communauté du soccer et aux équipes qui cherchent à améliorer leurs performances.

## 3. Description des données analysées

Les données utilisées pour ce projet proviennent de l'ensemble de données sur le soccer disponible sur Kaggle, intitulé "European Soccer Database". Ce jeu de données offre plusieurs informations sur plus de 25 000 matchs, plus de 10 000 joueurs, ainsi que des détails sur les équipes, les formations tactiques, les cotes de paris et les événements de matchs.

**Source des données**

Le dataset est composé de données provenant de plusieurs sources, notamment :

- http://football-data.mx-api.enetscores.com/ pour les scores, les line-ups, les formations d'équipe et les événements de match
- http://www.football-data.co.uk/ pour les cotes de paris
- http://sofifa.com/ pour les attributs des joueurs et des équipes provenant des jeux vidéo FIFA d'EA Sports

**Type de données**

Les données sont de type quantitatif et qualitatif, et comprennent des informations sur les matchs, les équipes, les joueurs, les formations tactiques, les cotes de paris et les événements de matchs.

**Quantité**

Le jeu de données comprend plus de 25 000 matchs, plus de 10 000 joueurs et 11 pays européens avec leur championnat principal.

**Attributs**

Nous avons décidé de seulement utiliser la table `Team _Attributes` pour avoir les attributs d'équipe comme variables explicatives dans notre jeu de donné, car ses attributs ont été conçues en prenant en considération les attributs des joueurs individuellement et en ajoutant le facteur de cohésion du jeu d'équipe. Cette décision accélèrera l'entraînement de nos algorithmes.

**Traitement des données manquantes**

Les colonnes qui contiennent beaucoup de données manquantes ont été supprimées.

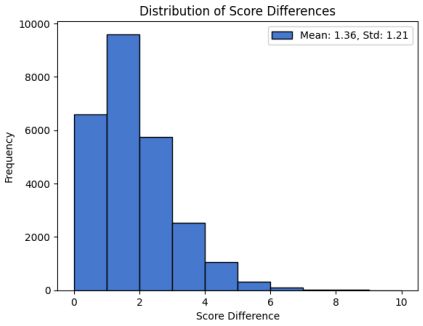
```
team_api_id                     0
date                            0
buildUpPlaySpeed                0
buildUpPlaySpeedClass           0
buildUpPlayDribbling          969
buildUpPlayDribblingClass       0
buildUpPlayPassing              0
buildUpPlayPassingClass         0
buildUpPlayPositioningClass     0
chanceCreationPassing           0
chanceCreationPassingClass      0
chanceCreationCrossing          0
chanceCreationCrossingClass     0
chanceCreationShooting          0
chanceCreationShootingClass     0
chanceCreationPositioningClass  0
defencePressure                 0
defencePressureClass            0
defenceAggression               0
defenceAggressionClass          0
defenceTeamWidth                0
defenceTeamWidthClass           0
defenceDefenderLineClass        0
dtype: int64

Number of null values: 969
```

Nous avons seulement effacé l'attribut **buildUpPlayDribbling**.

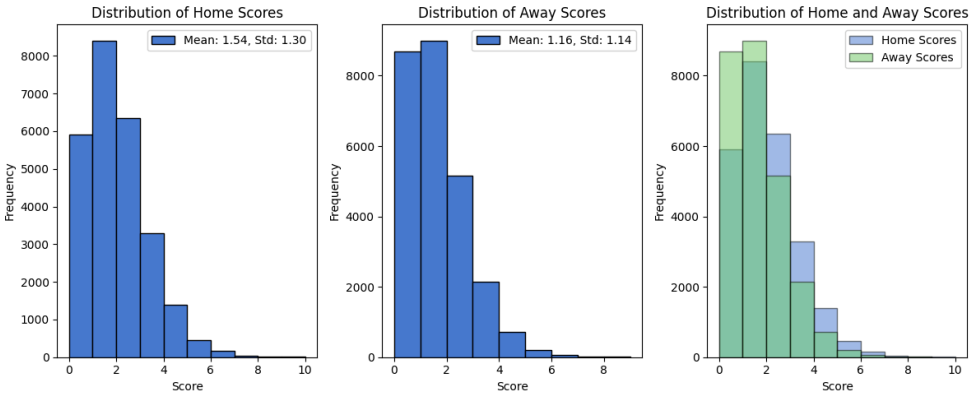
### Statistiques descriptives

Les statistiques descriptives du jeu de données comprennent :

- Moyenne et écart type des scores

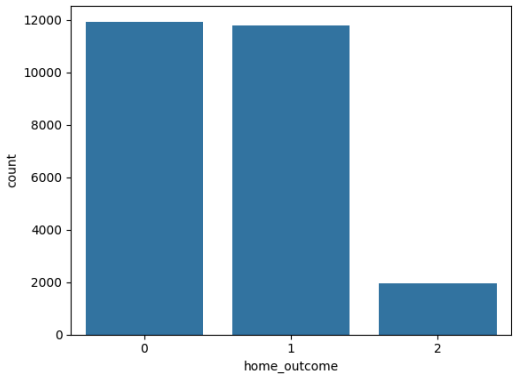



Nous voyons que les matchs ont tendance à ne pas finir partie nulle.



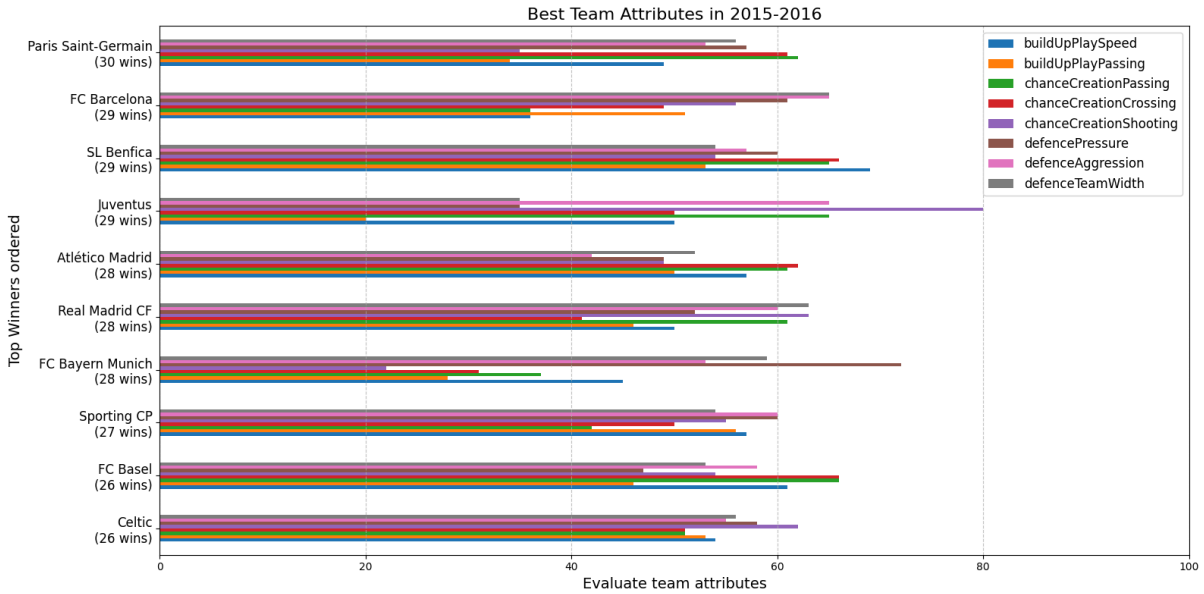
Nous constatons que les matchs joués à domicile ont tendance à être dominés par l'équipe à domicile.

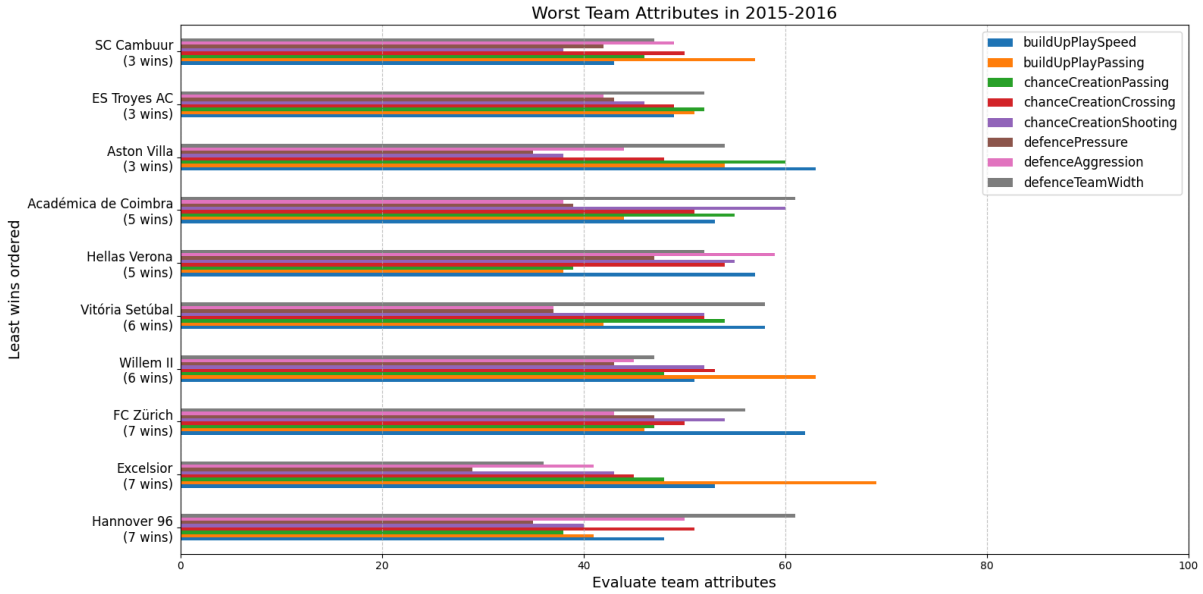
- Répartition des victoires, des défaites et des matchs nuls




Nous remarquons que la classe 2 correspondant aux matchs nuls est très minoritaire.
Par contre les classes 0 et 1 sont quasiment égales et largement majoritaires. Donc la classe 2 est largement non équilibrée par rapport aux autres classes.

- Corrélation entre les variables. Voir annexe 1)
- Distribution des attributs des 10 meilleures équipes de 2016






Nous constatons qu'il n'y a pas de tendance particulière entre les meilleures équipes et de même entre les équipes les moins performantes.Toutefois les mesures sont en moyenne plus élevées pour les meilleures équipes que pour les mauvaises équipes.

En résumé, les données utilisées pour ce projet sont fiables, exhaustives, non fortement corrélées et variées. Ce qui permettra d'obtenir des résultats précis et significatifs pour améliorer la compréhension du soccer et espérer prédire le résultat d'un match.

## 4. Méthodologie d'analyse

Etant donné que nos points de données sont étiquetés avec des classes et que notre objectif est de prédire les étiquettes des observations futures, les modèles de classifications sont de bons candidats pour résoudre notre problème.

### 1. Random Forest

Nous avons choisi d'entraîner notre modèle de prédiction avec Random Forest en premier lieu en raison de ses capacités à gérer efficacement les données complexes souvent rencontrées dans les résultats des matchs de soccer. En effet, ces données comportent généralement à la fois des variables qualitatives (comme le nom des équipes, le lieu du match, etc.) et des variables quantitatives (comme les statistiques des joueurs, les performances passées, etc.). Random Forest est bien adapté à ce

type de données mixtes, car il peut traiter à la fois les variables catégorielles et numériques, ce qui est crucial pour prédire avec précision les résultats des matchs.

De plus, Random Forest est connu pour sa capacité à résister au surapprentissage, un problème courant dans les modèles d'apprentissage automatique. Le surapprentissage se produit lorsque le modèle s'adapte trop étroitement aux données d'entraînement, ce qui peut entraîner une baisse des performances lorsqu'il est confronté à de nouvelles données. Random Forest utilise des techniques d'agrégation et de sélection aléatoire des caractéristiques pour construire plusieurs arbres de décision, ce qui réduit le risque de surapprentissage et permet au modèle de mieux généraliser aux nouvelles données.
De plus, Random Forest est robuste aux données aberrantes, c'est-à-dire aux valeurs extrêmes qui pourraient perturber la performance du modèle. Les arbres de décision individuels utilisés dans Random Forest sont moins sensibles aux données aberrantes que d'autres algorithmes, ce qui contribue à la stabilité et à la fiabilité des prédictions du modèle.

### 2. KNN

Nous avons choisi l'algorithme KNN pour prédire le résultat des matchs de soccer en fonction des attributs des équipes. KNN est approprié car il est simple à comprendre et à mettre en œuvre. Il fonctionne en recherchant les k échantillons d'entraînement les plus proches dans l'espace des attributs pour chaque point de test, puis attribue une étiquette basée sur la classe majoritaire parmi ces voisins. KNN est également choisi pour sa simplicité et sa flexibilité, ainsi que pour sa capacité à traiter des données non linéaires sans faire d'hypothèses sur leur distribution. De plus, des notions telles que la distance euclidienne, le choix judicieux de k via la validation croisée ont été utilisés.

### 3. SVM non-linéaire

Lorsque nos jeux de données ne sont pas linéairement séparables, c'est-à-dire que les frontières de décision entre les classes ne peuvent pas être représentées de manière linéaire, il est important d'utiliser des techniques adaptées pour modéliser ces relations complexes. Voir annexe 2)

Dans le contexte de la prédiction des résultats des matchs de soccer, où les interactions entre les différentes caractéristiques des équipes et les résultats des

matchs peuvent être non linéaires, l'utilisation de méthodes non linéaires comme les Support Vector Machines (SVM) avec des noyaux non linéaires peut être bénéfique.

Les SVM sont des algorithmes d'apprentissage supervisé utilisés pour la classification et la régression. Ils sont capables de trouver des frontières de décision complexes en transformant l'espace des caractéristiques d'entrée dans un espace de plus grande dimension, où les données peuvent être séparées de manière linéaire. Cela se fait en utilisant des fonctions noyau (kernel functions) qui calculent le produit scalaire entre les données dans cet espace de plus grande dimension, sans avoir à calculer explicitement les coordonnées de ces données dans cet espace.

Ainsi, en utilisant des SVM avec des noyaux non linéaires, nous pouvons capturer des relations complexes entre les caractéristiques des équipes et les résultats des matchs, même lorsque ces relations ne peuvent pas être modélisées de manière linéaire. Cela permet d'améliorer la capacité de notre modèle à faire des prédictions précises et à généraliser à de nouvelles données, ce qui est crucial pour la prédiction des résultats.

4. **Multimodal Naive Bayes**

Pour aborder notre problème de prédiction des résultats d'une partie de soccer avec des données non normalement distribuées (test statistique rejette l'hypothèse pour toutes les colonnes), nous avons choisi d'utiliser le Multimodal Naive Bayes (MNB). Étant donné sa capacité à modéliser les caractéristiques avec une distribution multinomiale. Contrairement au Gaussian Naive Bayes classique qui suppose une distribution normale des données, le MNB peut traiter efficacement les données non normalisées. Nous l'avons essayé car cet algorithme est très simple, rapide à entraîner et donne de relativement bon résultat malgré son assomption naïve, comme nous l'avons mentionné en classe.

5. **XGBoost**

Nous avons opté pour l'algorithme  GradientBoostingClassifier pour prédire les résultats des matchs de soccer en raison de sa capacité à fournir des performances élevées dans les problèmes de classification, notamment pour les ensembles de données de grande taille comme les nôtres d'après nos recherches sur le web. GradientBoostingClassifier a été choisi pour sa robustesse et sa capacité à gérer des données complexes, en exploitant des techniques d'optimisation de gradient pour améliorer la précision des prédictions. Son utilisation a été motivée par sa popularité dans le domaine de l'apprentissage automatique et son efficacité éprouvée dans une variété de tâches de classification.

### 6. Optimisation d’hyper paramètres

Pour optimiser nos différents modèles, nous avons utilisé des techniques d'optimisation d'hyper paramètres telles que GridSearchCV et Optuna. GridSearchCV est une méthode qui consiste à spécifier une liste de valeurs pour chaque hyperparamètre du modèle, puis à évaluer le modèle pour chaque combinaison possible de ces valeurs, en utilisant la validation croisée pour évaluer les performances. Cela permet d'identifier les meilleures valeurs d'hyper paramètres pour maximiser les performances du modèle.

D'autre part, Optuna est une bibliothèque d'optimisation d'hyper paramètres basée sur l'optimisation bayésienne. Elle explore de manière plus intelligente l'espace des hyperparamètres en utilisant des techniques probabilistes pour déterminer quelles configurations d'hyper paramètres sont les plus prometteuses à évaluer, ce qui permet une convergence plus rapide vers les meilleures valeurs d'hyper paramètres.

En utilisant ces techniques d'optimisation d'hyper paramètres, nous avons pu rechercher efficacement les meilleures configurations pour nos modèles, améliorant ainsi leurs performances et leur capacité à généraliser à de nouvelles données. Cela nous a permis d'obtenir des modèles plus performants pour la prédiction des résultats des matchs de soccer, ce qui est essentiel pour notre objectif de prédiction précise et fiable.

## 5. Résultats

Voir les graphiques des résultats (Cliquer sur les liens):
Précision Recall F1-Score Comparaison des modèles

**Performance globale :**

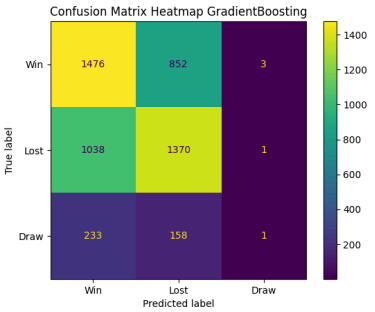
Nous pouvons voir comme baseline le modèle Random qui utilise une probabilité proportionnelle à la distribution des résultats de match pour classifier les parties.
La précision, le recall et le score F1 varient d'un modèle à l'autre. Certains modèles ont de meilleures performances que d'autres, mais il n'y a pas de modèle qui se démarque nettement dans tous les aspects. Nous voyons tout de même des performances strictement supérieures au modèle random.

**Sensibilité aux données transformées :**

Les performances des modèles basés sur PCA sont généralement légèrement inférieures à celles des modèles sans PCA.
Cela suggère que la réduction de dimensionnalité introduite par PCA peut entraîner une perte d'information pour certains modèles, affectant ainsi leurs performances.

**Meilleur modèle :**

Sur la base des résultats présentés, le modèle GradientBoostingClassifier semble être le meilleur selon tous les critères évalués avec une précision (weighted recall) de 55%




L'analyse des résultats met en lumière la diversité des performances des différents modèles dans la prédiction des résultats des parties de soccer. Pour une meilleure prédiction, il peut être nécessaire de combiner les forces de plusieurs modèles ou d'explorer davantage de techniques de prétraitement des données pour améliorer la qualité des prédictions.

## 6. Conclusions

En conclusion, malgré nos efforts pour contourner le déséquilibre des données, aucune des techniques utilisées, telles que SMOTE, les poids de classe, l'échantillonnage sous-représenté et l'échantillonnage sur-représenté, n'a fonctionné efficacement dans notre projet de prédiction des résultats des matchs de soccer. De plus, l'utilisation de PCA n'a pas apporté les bénéfices escomptés.

Cependant, certaines approches ont été fructueuses, notamment l'optimisation des hyperparamètres avec GridSearchCV et Optuna.

Pour l'avenir, il serait judicieux d'explorer des méthodes d'équilibrage des données plus avancées et spécifiques aux données de matchs de soccer. L'utilisation de modèles plus complexes ou de techniques d'apprentissage profond pourrait également être envisagée pour capturer les relations non linéaires entre les caractéristiques des équipes et les résultats des matchs. Enfin, l'intégration de données supplémentaires ou de caractéristiques plus spécifiques aux matchs de soccer pourrait aider à améliorer la qualité des prédictions.

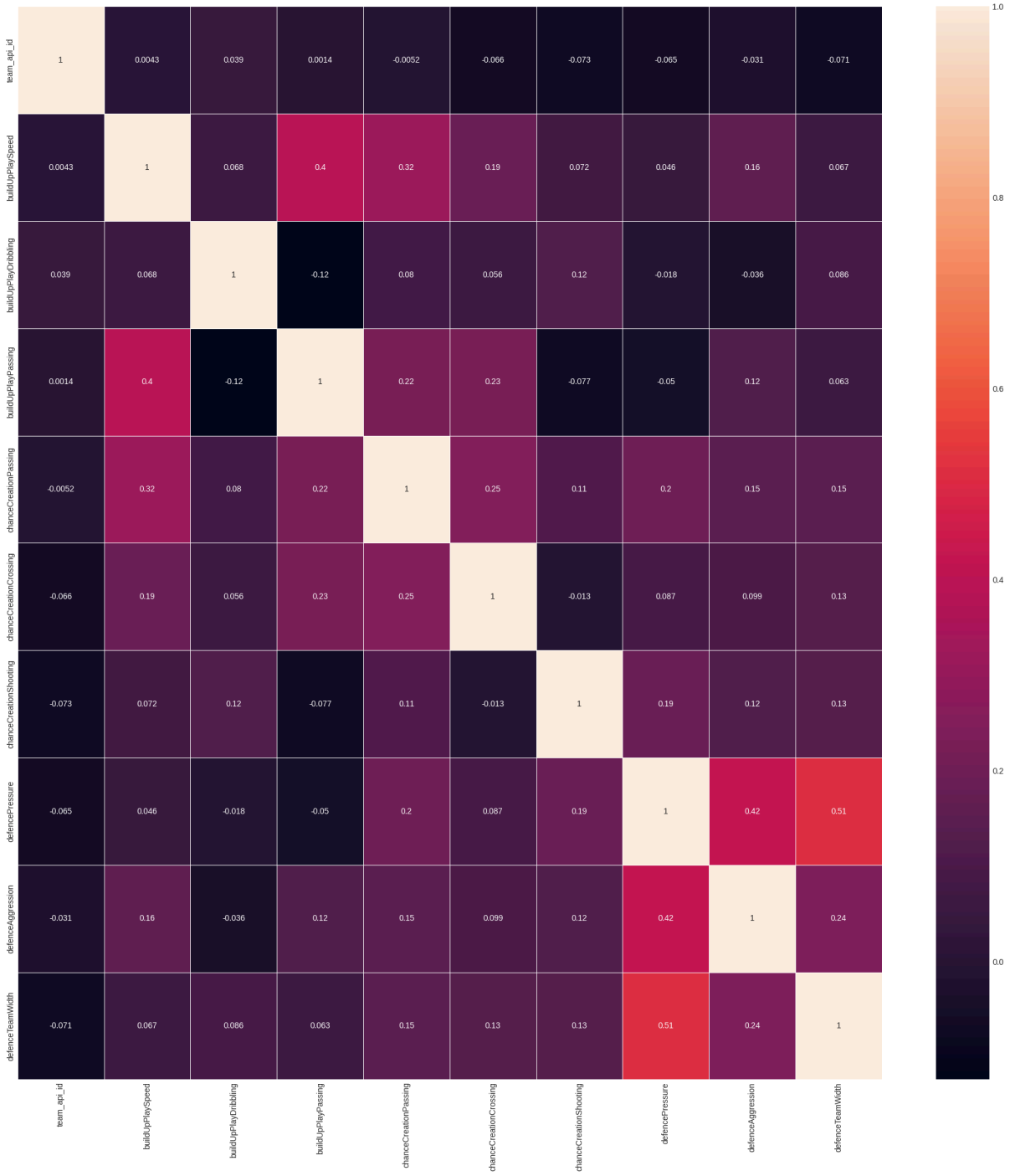
## 7. Contribution pour chaque membre

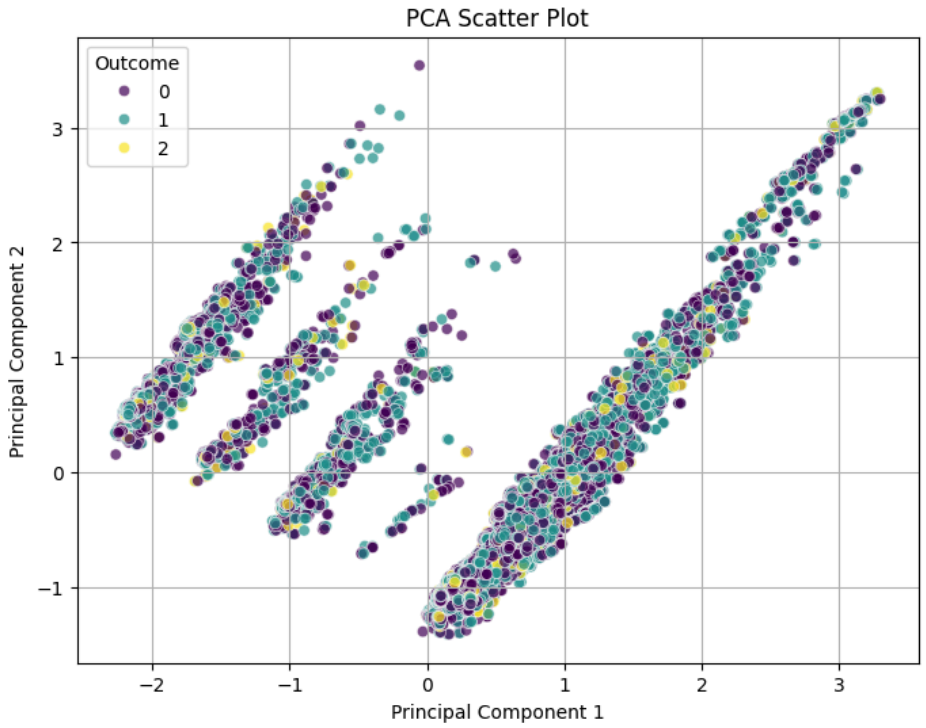
Dans le cadre de notre projet, nous avons travaillé en étroite collaboration, participant activement à toutes les étapes du processus. Nous avons débuté par la mise en place de l'environnement de développement nécessaire, puis nous avons collecté et prétraité les données relatives aux matchs de soccer. Ensemble, nous avons réalisé une analyse exploratoire approfondie des données, en utilisant des techniques de visualisation pour mieux comprendre les tendances et les patterns présents.

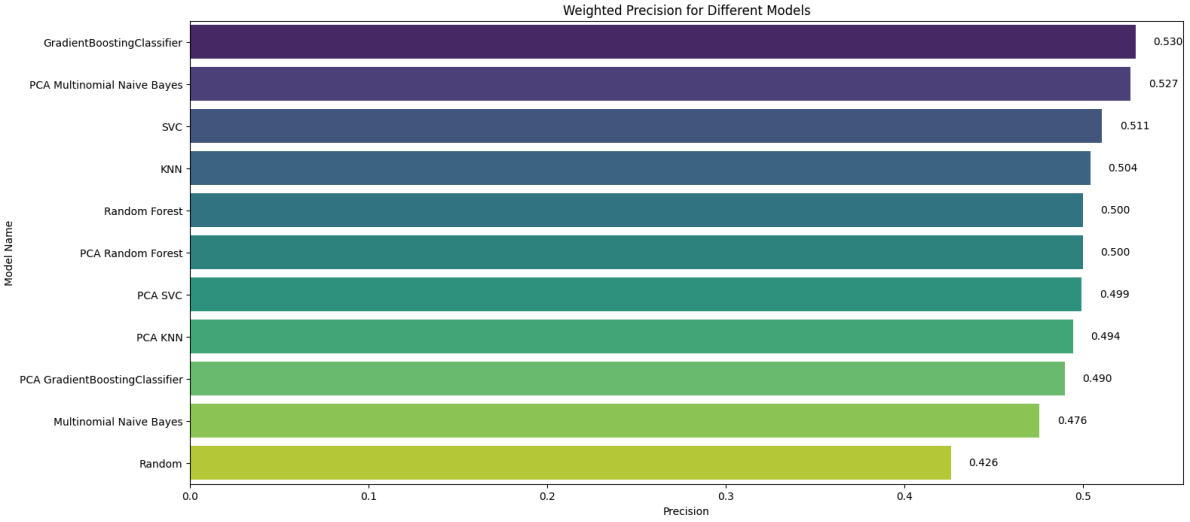
L'implémentation des modèles de prédiction ainsi que l'optimisation de leurs hyperparamètres ont également été des efforts communs. Nous avons également travaillé ensemble sur l'application des techniques d'équilibrage des données, telles que SMOTE et les poids de classe, pour résoudre le problème de déséquilibre des classes, bien que ces tentatives n'aient pas abouti à des résultats satisfaisants.

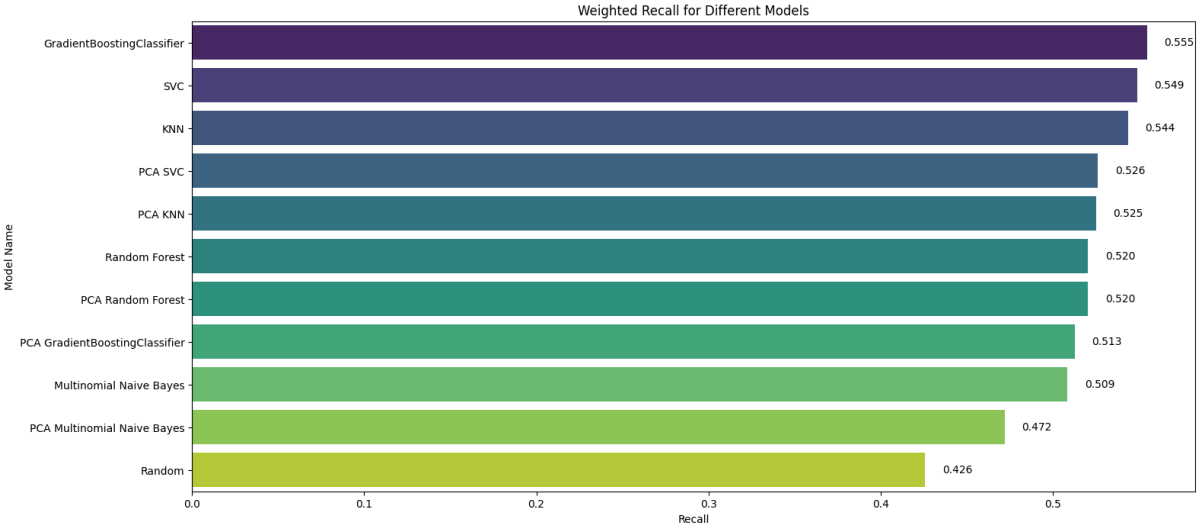
L'évaluation des performances des modèles, l'interprétation des résultats et la révision des approches ont été des tâches que nous avons menées conjointement. Enfin, la rédaction de la méthodologie du projet, de la partie introduction, de la revue de littérature, et des conclusions a été le fruit d'une collaboration étroite entre nous deux. Notre projet a ainsi bénéficié de notre complémentarité et de notre capacité à travailler ensemble de manière harmonieuse et efficace.
(Oui nous avons passé beaucoup de temps sur FaceTime)

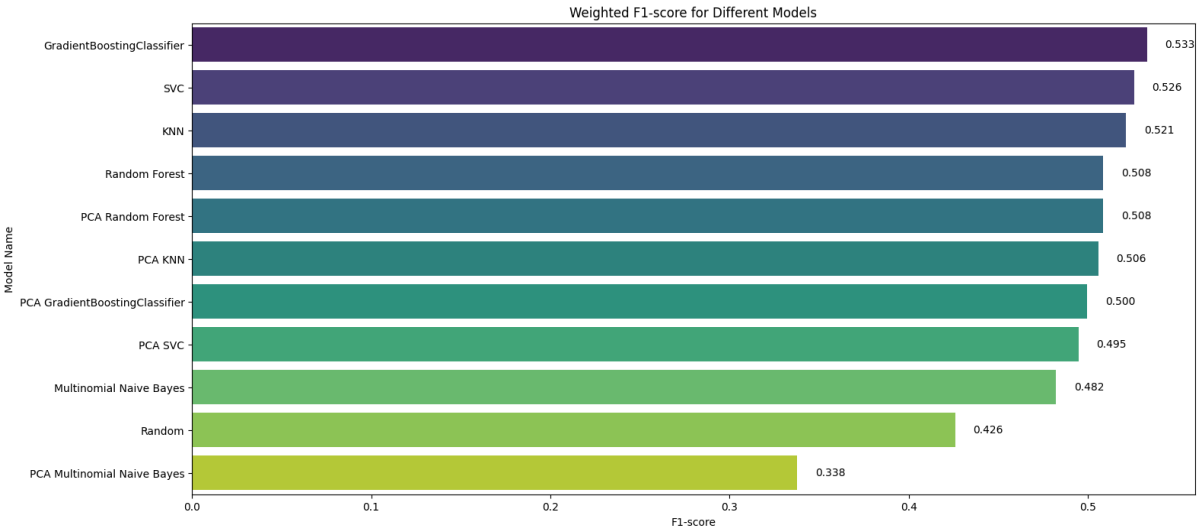
# Graphiques

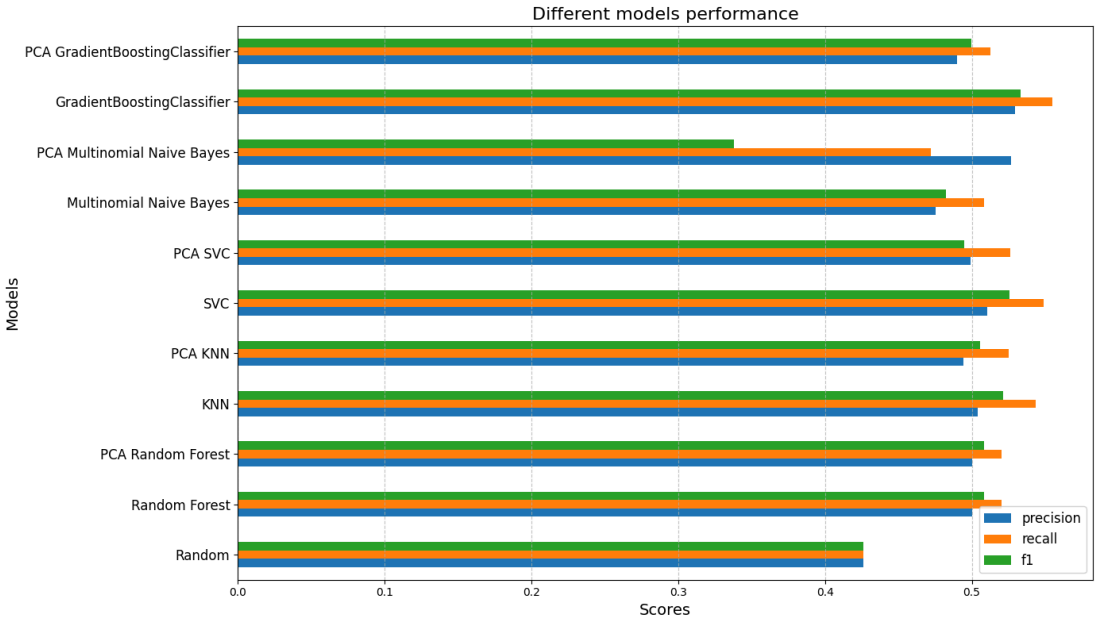

team_api_id
buildUpPlaySpeed
buildUpPlayDribbling
buildUpPlayPassing
chanceCreationPassing
chanceCreationCrossing
chanceCreationShooting
defencePressure
defenceAggression
defenceTeamWidth
1 0.0043 0.039 0.0014 -0.0052 -0.066 -0.073 -0.065 -0.031 -0.071
0.0043 1 0.068 0.4 0.32 0.19 0.072 0.046 0.16 0.067
0.039 0.068 1 -0.12 0.08 0.056 0.12 -0.018 -0.036 0.086
0.0014 0.4 -0.12 1 0.22 0.23 -0.077 -0.05 0.12 0.063
-0.0052 0.32 0.08 0.22 1 0.25 0.11 0.2 0.15 0.15
-0.066 0.19 0.056 0.23 0.25 1 -0.013 0.087 0.099 0.13
-0.073 0.072 0.12 -0.077 0.11 -0.013 1 0.19 0.12 0.13
-0.065 0.046 -0.018 -0.05 0.2 0.087 0.19 1 0.42 0.51
-0.031 0.16 -0.036 0.12 0.15 0.099 0.12 0.42 1 0.24
-0.071 0.067 0.086 0.063 0.15 0.13 0.13 0.51 0.24 1
1.0
0.8
0.6
0.4
0.2
0.0


PCA Scatter Plot
Outcome
0
1
2
Principal Component 2
Principal Component 1
3
2
1
0
−1
−2
−1
0
1
2
3


Weighted Precision for Different Models
GradientBoostingClassifier
PCA Multinomial Naive Bayes
SVC
KNN
Random Forest
PCA Random Forest
PCA SVC
PCA KNN
PCA GradientBoostingClassifier
Multinomial Naive Bayes
Random
0.530
0.527
0.511
0.504
0.500
0.500
0.499
0.494
0.490
0.476
0.426
Model Name
0.0
0.1
0.2
0.3
0.4
0.5
Precision


Weighted Recall for Different Models
GradientBoostingClassifier
SVC
KNN
PCA SVC
PCA KNN
Random Forest
PCA Random Forest
PCA GradientBoostingClassifier
Multinomial Naive Bayes
PCA Multinomial Naive Bayes
Random
0.555
0.549
0.544
0.526
0.525
0.520
0.520
0.513
0.509
0.472
0.426
Model Name
0.0
0.1
0.2
0.3
0.4
0.5
Recall


Weighted F1-score for Different Models
GradientBoostingClassifier
SVC
KNN
Random Forest
PCA Random Forest
PCA KNN
PCA GradientBoostingClassifier
PCA SVC
Multinomial Naive Bayes
Random
PCA Multinomial Naive Bayes
0.533
0.526
0.521
0.508
0.508
0.506
0.500
0.495
0.482
0.426
0.338
Model Name
0.0
0.1
0.2
0.3
0.4
0.5
F1-score


Different models performance
PCA GradientBoostingClassifier
GradientBoostingClassifier
PCA Multinomial Naive Bayes
Multinomial Naive Bayes
PCA SVC
SVC
PCA KNN
KNN
PCA Random Forest
Random Forest
Random
Models
0.0
0.1
0.2
0.3
0.4
0.5
Scores
precision
recall
f1