

Beat the Bookmakers\*

Promotion : Data Analyst - septembre 2020.

Participants :

Sofiane

Anthony

Maxime Le Tutour

\*Battre les bookmakers

# Présentation et objectifs du projet

Dans le cadre de la formation de Data Analyst nous allons présenter notre projet de data science, une analyse détaillée d'un jeu de données permettant de maximiser un profit.

Les données qui nous allons analyser représentent les matchs de tennis professionnels ayant eu lieu entre 2000 et 2018 contenant certaines caractéristiques pour chaque match et les côtes de deux sites de paris sportifs (Pinnacle et Bet 365). Au total 44 708 matchs sont répertoriés.

Les côtes des paris sportifs pour un match sont définies et calculées par les bookmakers en fonction de leurs estimations des probabilités de survenance de l'évènement en se basant notamment sur des données statistiques, l'historique des rencontres et de manière générale toutes les informations disponibles pertinentes.

Notre objectif est d'obtenir un potentiel retour sur investissement maximum en élaborant une stratégie de pari optimale. Pour cela, nous allons utiliser ces données pour entrainer un algorithme qui calculera la probabilité du vainqueur de chaque match. Nous la comparerons à la probabilité des sites de bookmaker pour décider s'il est avantageux de parier ou non.

Au terme de notre analyse et de la création de l'algorithme nous verrons qu'il est potentiellement possible de gagner environ 1200€ net en pariant 1€ sur 10000 paris, soit un gain moyen de 0,12€ par pari.

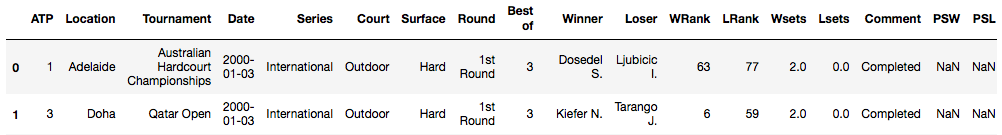
Les membres de l'équipe n'ont d'expertise du domaine que la formation de DataScientest du parcours Data Analyst. Pour ce projet il n'est pas nécessaire d'avoir des connaissances particulière en tennis, seulement de connaitre quelques vocabulaires.

# Étape 1 : Découverte du jeu de données

## Introduction

Le jeu de données est disponible librement sur le site *http://tennis-data.co.uk/data.php*.

Initialement le dataframe comprend 44708 matchs et 23 variables dont voici les premières lignes :



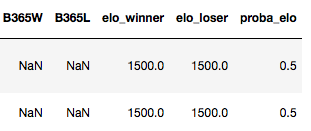


Image 1 : Jeu de donnée initial

Afin de comprendre ce que signifie les informations du jeu de donnée, voici quelques détails :

|  |  |
| --- | --- |
| Serie | La catégorie du tournois |
| Court | Match ayant lieu à l'intérieur ou l'extérieur |
| Surface | Le type de surface du terrain |
| Best of | Si le match est gagné est trois ou deux sets |
| WRank/Lrank | Le rang mondial du gagnant et du perdant |
| PSW/PSL | La côte du gagnant/perdant du site Pinnacle |
| B365W/B365L | La côte du gagnant/perdant du site Bet 365 |
| Elo | Le nombre de points du joueur au classement elo |
| Proba elo | Probabilité à priori que le gagnant remporte le match |
| Winner | Variable cible |

Tableau 1: Explication des variables

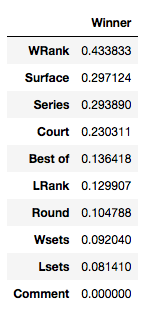
A présent vérifions la qualité du jeu de données et quelques caractéristiques des variables.

|  |  |
| --- | --- |
| Lignes | 44708 |
| Valeurs manquantes | 35626 |
| Doublon | 0 |
| Variables | 23 |
| Catégoriques | 8 |
| Numériques | 10 |
| Texte | 4 |
| Date | 1 |

Tableau 2 : Description du jeu de données

On s'aperçoit que de nombreuses valeurs sont manquantes. Nous verrons plus tard, lors de la transformation du jeu de données comment nous les traiterons.

## Corrélations entre les variables



Pour mesurer la corrélation des variables qualitatives avec la variable **winner,** nous utilisons le V de Cramer. Plus le résultat est proche de 1 plus la corrélation est robuste. Un indice supérieur à 0.25 est considéré comme significatif.

On peut s'apercevoir, sans surprise, que la variable la plus corrélée à la variable **winner** est **le rang mondial du vainqueur**. Suivi de **la surface du terrain** et de **la catégorie du tournois.**

Figure 1 : Corrélation des variables qualitatives avec la variable cible winner

#### Corrélation des variables quantitatives avec la variable cible **winner**

Dans le but de mesurer la corrélation des variables quantitatives avec la variable cible nous utiliserons l'analyse de la variance (**ANOVA**) à un facteur. Plus la p-value est faible et plus la statistique F élevée, plus la corrélation est forte.

|  |  |  |
| --- | --- | --- |
| Variables | Statistique F | P-value |
| PSW | 6.06 | 0.0 |
| PSL | 14.16 | 0.0 |
| B365W | 8.36 | 0.0 |
| B365L | 19.14 | 0.0 |
| elo-winner | 139.09 | 0.0 |
| elo-looser | 8.92 | 0.0 |
| proba elo | 25.69 | 0.0 |

Tableau 3 : Corrélation des variables quantitatives avec la variable cible

Toutes ces variables quantitatives sont corrélées à un certain degré à la variable cible (p-value = 0). Nous voyons clairement que le nombre de points elo du gagnant est la variable la plus corrélée à la variable cible.

#### Corrélation entre variables quantitatives

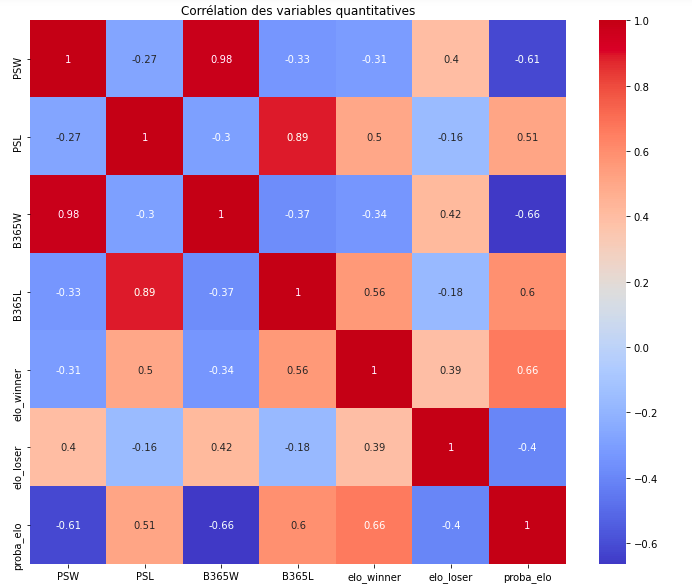
Pour comprendre les relations entre les variables quantitatives nous avons réalisé une heatmap utilisant le coefficient de Pearson. Les résultats sont compris entre -1 et 1. Une corrélation forte et négative se rapprochera de -1 et une corrélation forte et positive se rapprochera de 1.

Figure 2 : Heatmap corrélation des variables quantitatives

On s'aperçoit que de nombreuses variables sont corrélées entre elles. En effet, l'issu d'un match de tennis est souvent prédictibles grâce au rang du joueur et son nombre de points Elo. Les variables ont donc tendance à donner la même information.

## Data visualisation

#### Distribution de quelques variables

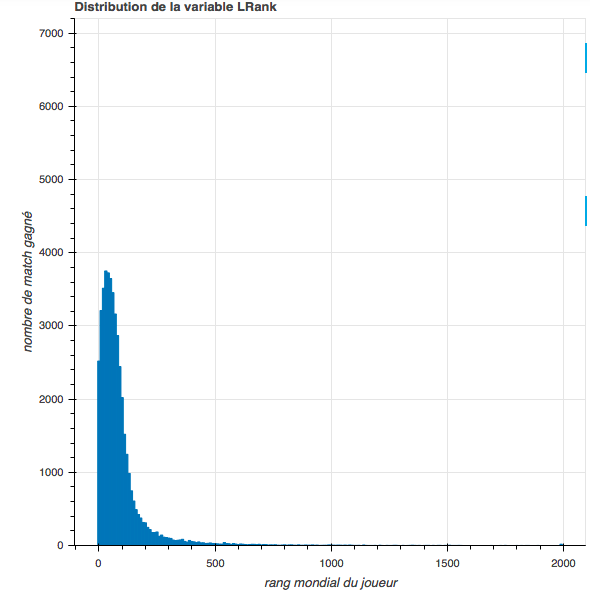
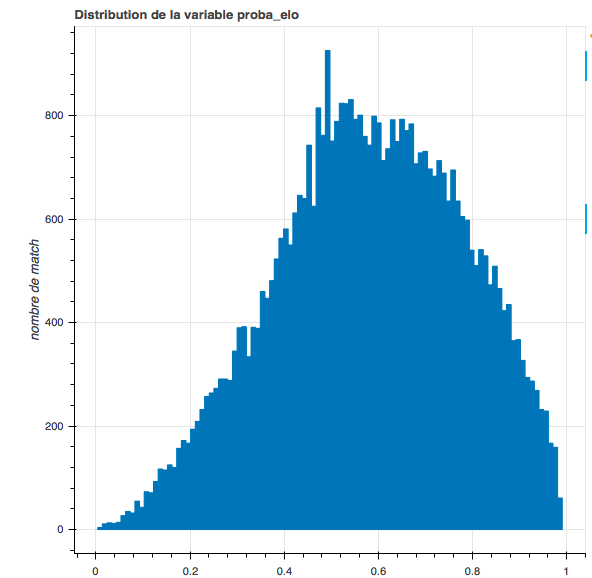


Figure 3 : Distribution du rang mondial des gagnant et des perdants

Le classement mondial des joueurs se situe majoritairement entre 1 et 300. On note, en toute logique, que les gagnants sont massivement situés dans les premiers classés.



**La probabilité elo** est la probabilité a priori que le gagnant gagne le match. Ce graphe indique le degré de précision de la prédiction. La courbe semble suivre une loi normale avec une proba elo se situant vers 0.6 en moyenne et une variance relativement élevée.

Figure 4 : Distribution de la proba elo

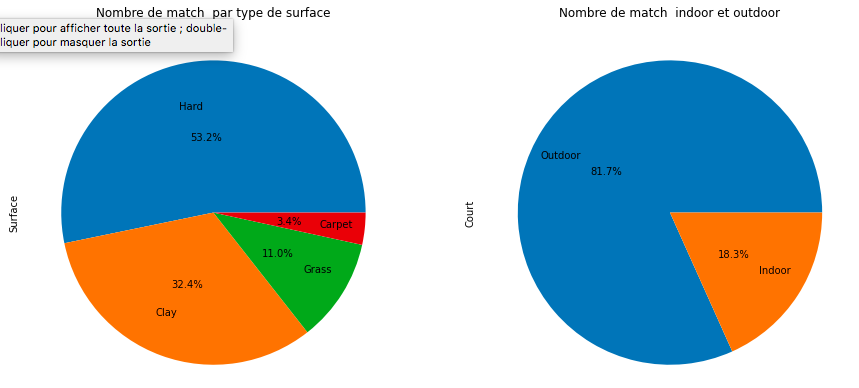


Figure 5 : distribution du type de surface et du court

Les matchs se jouent principalement à l'extérieur et sur une surface dur ou terre battue (clay).

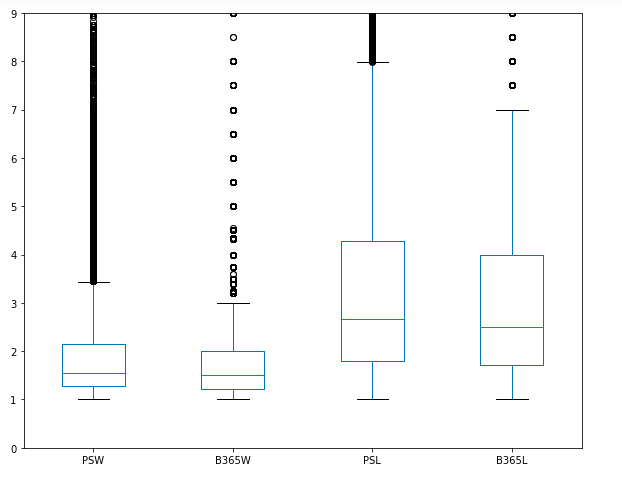


Figure 6: comparaison des cotes des perdants et des gagnants des deux sites de paris

Sur ces boites à moustache nous voyons clairement que les cotes des gagnants(PSW, B365W) sont nettement inférieures au côtes de perdants (PSL, B365L). Ce qui confirme que le résultat des matchs sont correctement prédits par les algorithmes des bookmakers. Le deuxième point intéressant est la différence notable des cotes des deux sites de paris. Le site Pinnacle propose aux parieurs des cotes plus intéressantes que son concurrent Bet 365. De plus les outliers des cotes des gagnants du site Pinnacle sont bien plus nombreux. Ils proposent une cote élevé pour le joueur qui gagnera le match. Ce sont donc de bonnes opportunités de pari. Il sera donc plus intéressant pour nous de parier sur Pinnacle que Bet 365.

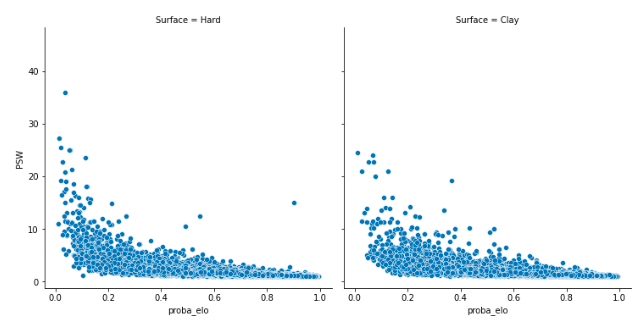


Figure 7 : proba elo en fonction de la cote du vainqueur

Il est remarquable que plus la proba elo est élevé, plus la cote du joueur est faible. Il semble assez clair que les bookmakers se fie à cette variable pour définir la cote du joueur.

#### Nos premiers paris

Avant de se lancer dans une analyse plus poussée, nous nous sommes posé la question :

*"Combien l'on gagnerait en pariant 1 euro uniquement sur les cotes les plus faibles ? Ou sur la proba elo la plus faible ? ou sur les mieux classés ?"*

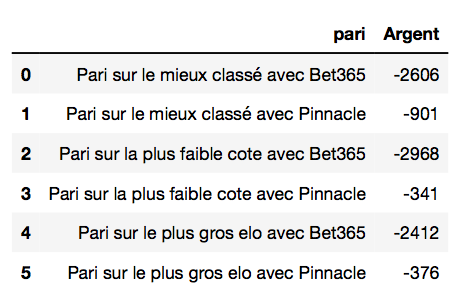
Voici les gains que l'on obtiendrait si on optait pour une de ces stratégies de pari :

Tableau 4 : Les gains potentiels avec des stratégies de pari basiques

Il va donc falloir travailler un peu plus si l'on ne veut pas rendre notre banquier nerveux...

# Étape 2 : Nettoyage du jeu de données et création de variables

## Nettoyage

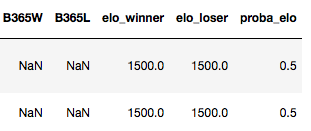
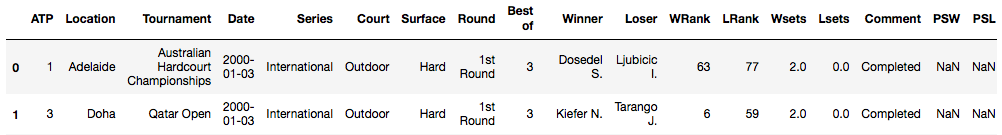


Figure 8: Jeu de données initial

Le nettoyage du jeu de données consiste a supprimer les informations non pertinentes pour atteindre notre objectif. Après avoir étudié chaque variable, nous décidons de supprimer les variables explicatives inutiles au déploiement de notre modèle, à savoir : "Tournament", "Round", "Location", "Date", "Series", "ATP".

Toutes les lignes avec des valeurs manquantes, principalement des cotes des sites de paris, sont supprimées. Le jeu de données nettoyé compte maintenant 28990 lignes.

## Création de nouvelles variables explicatives et d'une nouvelle variable cible

#### Les variables favori/challenger

Afin d'entrainer un algorithme d'apprentissage supervisé sur nos données nous devons créer une nouvelle variable cible. En effet de nombreuses variables de notre dataset donnent les informations du vainqueur et du perdant, c'est-à-dire les informations à posteriori du match. Notre objectif est d'abord de prédire le vainqueur, ces variables biaiseront donc considérablement la prédiction. Nous devons transformer le dataset pour supprimer ces biais. Pour cela nous n'utiliserons plus les variables **Winner** et **Looser**, qui sont des informations à posteriori, mais les variables **favori** et **challenger**, le favori étant systématiquement le joueur le mieux classé du match. Ainsi nous avons des informations à priori au match.

En conséquence les variables concernant le vainqueur et le perdant :

*Winner, Looser, WRank, LWrank, Wset, Lset, PSW, PSL, B365W, B365L, elo\_winner, elo\_looser* seront également supprimées pour être remplacées par celles concernant le favoris et le challenger : *Favoris, Challenger, Rank\_F, Rank\_C, set\_F, set\_C, PS\_F, PS\_C, B365\_F, B365\_C, elo\_favoris, elo\_challenger*

Nous créons ainsi la nouvelle variable cible **Vainqueur** qui prend deux modalités : Favoris et Challenger.

Il est important de comprendre à ce stade que nous ne remplaçons pas *winner* par *favori*, et *looser* par le *challenger*. Le favori et le challenger sont seulement déterminés par leur classement mondial avant le match. La variable cible **Vainqueur** est donc soit le mieux classé soit le moins bien classé du match.

#### Nouvelles variables explicatives

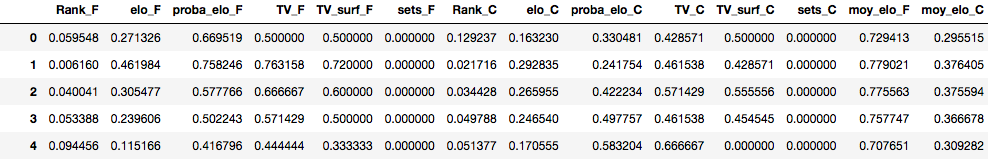
Nous créons ensuite de nouvelles variables explicatives en tirant des informations de notre jeu de données :

- le taux de victoire de leurs 150 derniers matchs afin d'évaluer les performances récentes des joueurs : **TV\_F** et **TV\_C** (taux de victoire favoris, et taux de victoire challenger).

- le taux de victoire sur la surface du match en question de leurs 150 derniers matchs pour estimer l'affinité du joueur avec cette surface : **TV\_surf\_F**, **TV\_surf\_C**

- le nombre de sets joués par les joueurs sur le tournois en question pour évaluer la potentielle fatigue du joueur : **sets\_F**, **sets\_C**

- La différence entre les points elo actuel et la moyenne des points elo des 150 derniers matchs afin de savoir si le joueur est plutôt dans une phase de progression ou de régression.



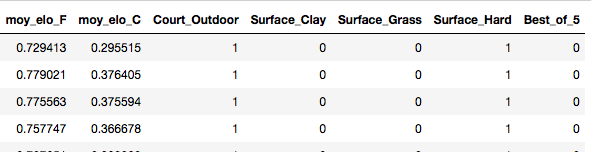


Figure 9 : Jeu de données nettoyé et amélioré prés pour l'apprentissage supervisé

# Étape 3 : Développement du modèle

## Prédiction du vainqueur

Nous allons maintenant prédire le vainqueur entre le favori et le challenger ainsi que la probabilité de victoire de chacun des joueurs. Ensuite nous chercherons à déterminer l'écart significatif entre les probabilités de victoire générées par l'algorithme d'apprentissage supervisé.

Tout d'abord nous allons tester plusieurs algorithmes avec un grid pour trouver la combinaison de paramètres la plus performante et mesurer trois indicateurs : la précision, le rappel et l'accuracy. Voici les résultats résumés dans ce tableau :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Nom algorithme | accuracy | | rappel | | précision | |
| Train | Test | Favori | Challenger | Favori | Challenger |
| Regression logistique | 0.68 | 0.68 | 0.91 | 0.20 | 0.70 | 0.53 |
| Forêt aléatoire | 0.91 | 0.67 | 0.88 | 0.25 | 0.70 | 0.51 |
| Arbre de décision | 0.68 | 0.67 | 0.88 | 0.25 | 0.70 | 0.51 |
| K plus proche voisins | 0.70 | 0.66 | 0.86 | 0.26 | 0.70 | 0.48 |
| XGBoost | 0.71 | 0.68 | 0.90 | 0.24 | 0.70 | 0.54 |

Tableau 5 : Résultats de prédictions des algorithmes de machine learning

L'accuracy du set d'entrainement (train) et du set test (test) nous permet de vérifier l'absence d'overfitting. Dans un premier temps, nous pouvons d'abord éliminer l'algorithme *Forêt aléatoire* qui présente un overfitting trop important pour être utilisé.

Dans un second temps, nous pouvons noter qu'aucun des algorithmes ne semblent se distinguer par leur performance. En outre, une tendance semble se dégager : il est difficile de prédire les challengers. La précision et le rappel du challenger de chaque algorithme est relativement faible. Or les challenger seront ceux qui créeront la surprise lorsqu'ils gagneront le match. Cela signifie une cote plus élevée, donc un potentiel de gain plus important. Partant de ce constat, nous choisissons de continuer le projet avec XGBoost qui présente un rappel et une précision légèrement supérieur aux autres algorithmes.

## Élaboration d'une stratégie de pari

Afin d'élaborer une stratégie de pari nous allons créer un nouveau dataframe.

Tout d'abord, comme nous l'avons vu dans la première partie, nous ne parierons que sur les côtes du site Pinnacle car celles-ci sont plus avantageuses.

Ensuite, nous allons intégrer dans le dataframe la probabilité de victoire de chaque joueur calculées par XGBoost ainsi que les probabilités de victoire calculées par les bookmakers en prenant l'inverse de leur cote.

Nous créerons deux variables supplémentaires : le gain théorique du favoris (gain\_theorique\_F) et du challenger (gain\_theorique\_C). Celles-ci sont le produit de la probabilité de victoire calculée par notre algorithme et la cote du site Pinnacle. Le gain théorique le plus élevé des deux joueurs sera notre indice de confiance que l'on placera dans la colonne "gain\_confiance", et qui nous indiquera sur quel joueur parier. La variable "gain\_confiance" est donc notre indicateur de pari. Plus il est élevé plus il est avantageux de parier. Nous ordonnons notre dataframe dans l'ordre décroissant de la variable *gain\_confiance.* Les paris les plus avantageux se retrouvent donc en premier.

Enfin, nous rajoutons la colonne "pari gagné" qui indique si le pari est gagné et la colonne "cote" qui est la cote du pari joué. Voici le dataframe résultant :

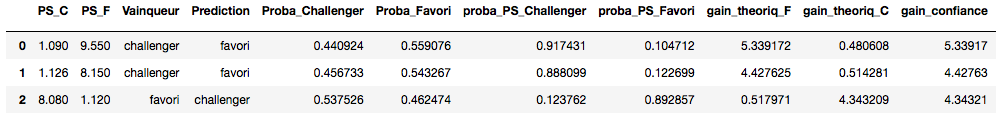
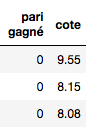


Tableau 6 : Dataframe utilisé pour les paris

## Pertinence

*Avez vous eu à nettoyer et à traiter les données ? Si oui, décrivez votre processus de traitement.*

*Quelles variables vous semblent les plus pertinentes au regard de vos objectifs ?*

*Quelle est la variable cible ?*

*Décrivez la distribution de ses valeurs ?*

*Avez vous identifié des relations entre différentes variables ? Entre variables explicatives ? et Entre vos variables explicatives et la/les cible(s) ?*

*Quelles particularités de votre jeu de données pouvez-vous mettre en avant ?*

# Projet

## Classification du problème

*À quel type de problème de machine learning votre projet s’apparente-t-il ? (Classification, régression, clustering ….)*

*À quelle tâche de machine learning votre projet s’apparente-t-il ? ( détection de fraude, reconnaissance faciale, analyse de sentiment …) ?*

*Quelle est la métrique de performance principale utilisée pour comparer vos modèles ?*

*Avez vous utilisé d’autres métriques de performances qualitative ou quantitative) ? Si oui, détaillez.*

## Choix du modèle & Optimisation

*Quels algorithmes avez vous essayé ?*

*Décrivez celui / ceux que vous avez retenu et pourquoi ?*

*Qu’est ce qui a engendré une amélioration significative de vos performances ?*

*Avez vous analysé les erreurs de votre modèle ?*

*Cela a-t-il contribué à son amélioration ? Si oui, décrivez.*

*Détaillez quelle a été votre contribution principale dans l’atteinte des objectifs du projet.*

# Description des travaux réalisés

## Répartition de l’effort sur la durée et dans l’équipe

*Morceler votre projet en un maximum de tâches unitaires. Produisez le diagramme de Gant a posteriori en spécifiant qui s’est occupé de quelle tâche et à quelle moment. (joindre le diagramme en annexe du rapport)*

## Bibliographie

*Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc... ) vous êtes vous appuyé pour réaliser votre projet ?*

## Difficultés rencontrées lors du projet

*Quel a été le principal verrou scientifique rencontré lors de ce projet ?*

*Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elle vous ont ralenti dans la mise en place de votre projet :*

*Prévisionnel : (tâche qui ont pris plus de temps que prévu etc ….)*

*Jeux de données : (Acquisition, volumétrie, traitement, aggregation etc….)*

*Compétences technique / théoriques : (Timing d’acquisition des compétences, compétence non proposée en formation etc…)*

*Pertinence : ( de l’approche, du modèle, des données etc ...)*

*IT : ( puissance de stockage, puissance computationnelle, etc…. )*

*Autres*

# Bilan & Suite du projet

*En quoi votre projet a-t-il contribué à un accroissement de connaissance scientifique ?*

*Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.*

*S’ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s’inscrire ? Détaillez.*

*Dans le cas contraire, quelles pistes d’amélioration suggérez-vous pour améliorer les performances de votre modèle ?*

# Annexes

## Diagramme de gant

## Description des fichiers de code