Rapport projet Apprentissage automatique

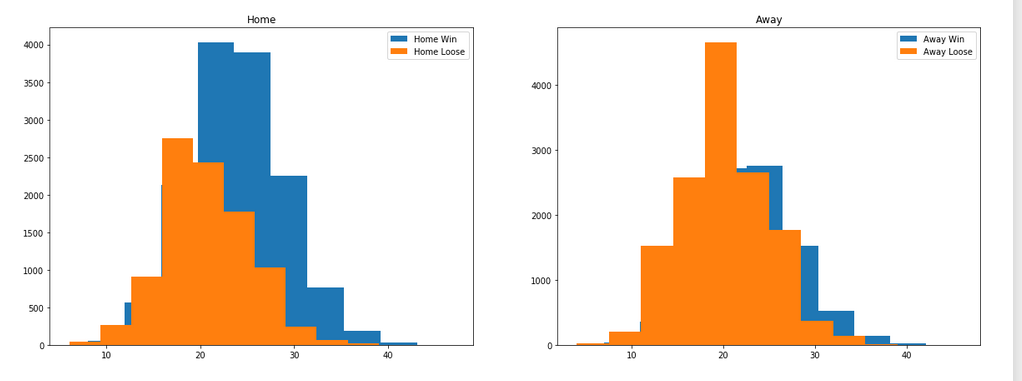
# Exploration des données

Pour ce projet nous avons choisi un dataset sur kaggel sur le thème des matchs NBA. Nous avions plusieurs choix de fichier csv possibles et les prendre tous nous auraient fait un trop grand nombre de données à traiter. Nous avons donc choisi de prendre celui qui répertoriait les statistiques des matchs (<https://www.kaggle.com/nathanlauga/nba-games>).

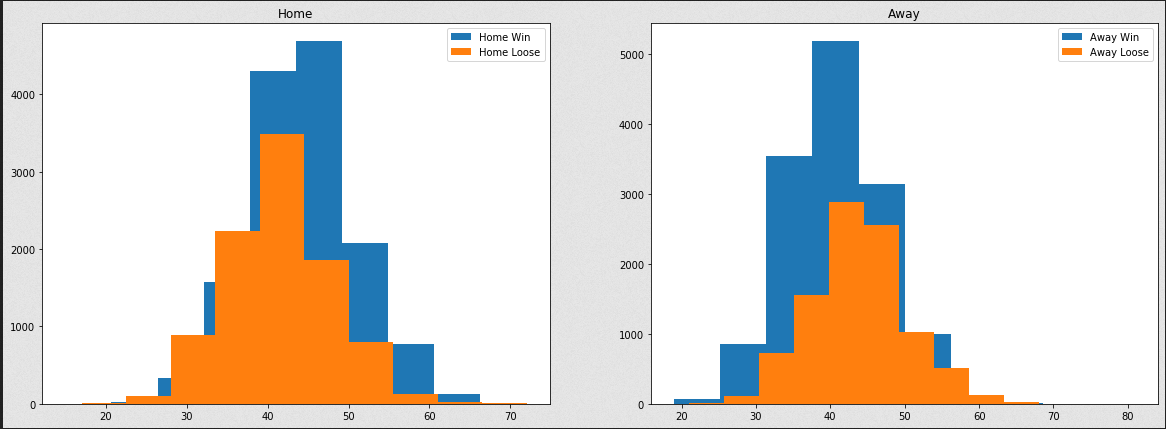
Le dataset original possède les informations suivantes : GAME\_DATE\_EST, GAME\_ID, GAME\_STATUS\_TEXT, HOME\_TEAM\_ID, VISITOR\_TEAM\_ID, SEASON, TEAM\_ID\_home, PTS\_home, FG\_PCT\_home, FT\_PCT\_home, AST\_home, REB\_home, TEAM\_ID\_away, PTS\_away, FG\_PCT\_away, FT\_PCT\_away, FG3\_PCT\_away, AST\_away, REB\_away, HOME\_TEAM\_WINS.

Dans un premier temps, nous avons donc regardé les features disponibles et fait un premier tri en enlevant celles qui ne nous paraissaient pas pertinentes c’est-à dire celles qui n’influent pas directement sur le résultat du match. Nous avons donc retiré les ID des équipes, la saison, la date du match, le statut du match ainsi que l’id du match. On obtient ainsi un dataset de 23520 lignes pour 12 features.

A partir de ce nouveau dataset, nous avons comparé les valeurs de différentes features en fonction de si le match est gagné ou perdu par les joueurs à domicile afin de déterminer la pertinence de celles-ci. Nous avons fait de même pour les visiteurs. Nous obtenons les graphes ci-après.

* Graphiques associés au nombre de passes décisives :

Observations : On remarque que lorsque home gagne les matchs, elle a fait plus de passes décisives. Quand les visiteurs (away) perdent, leur nombre moyen de passes décisives par match est plus bas que lorsqu’ils gagnent.

* Graphiques associés au nombre de rebonds par match :

Observations : On remarque que lorsque les joueurs à domicile (home) font des rebonds ils gagnent plus souvent et plus le nombre de rebond est élevé plus ils gagnent. En revanche, pour les visiteurs, l'équipe gagne plus quand le nombre de rebond est moins élevé.

Ces graphiques nous montent tout de même que la proportion de matchs gagnés par les joueurs à domicile et plus élevée que celle des matchs gagnés par les visiteurs. Le dataset n’est pas forcément super équilibré ?

# Définition de la tâche

A partir des observations que nous avons faites, nous avons pu déterminer l’objectif de ce projet : prédire si les joueurs à domicile gagnent le match ou pas. Nous allons donc faire une classification binaire supervisée car nous avons déjà les résultats des matchs.

# Pre-processing

Avant de réellement commencer à traiter notre dataset nous nous sommes assurés qu’il soit complet et correctement utilisable. Nous avons remarqué qu’une bonne quantité de données étaient Nan, pour palier cela et sans pour autant trop ‘fausser’ les données nous avons choisi de les remplacer par les valeurs moyennes de chacune des features concernées.

De plus, nous avons choisi de ne pas prendre en compte les données sur les points marqués par chacune des équipes car cela nous a semblé trop faciliter la prédiction du résultat.

Choix du validation/train test etc...

# Architecture

SVM 🡪 expliquer son fonctionnement + pk on l’a choisi

Bayesien 🡪 expliquer son fonctionnement

# Ouverture/Conclusion

* Features qui dit qd tu joues à domicile t’as tant de % de gagner