Formalisation du problème :

Prédire à Chicago si un crime se finit par une arrestation ou non? (apprentissage supervisé; classification binaire)

Les données :

Supprimer les données où il manque une feature, réduire la taille du dataset:

- sed '/,,/d' Crimes2001.csv | sed 1000000q > Crimes1M.csv (garde 1M de données au lieu de 7M, en enlevant les features vides)

Transformer la feature date en:

- matin/journée/soir/nuit
- jours de la semaine
- week-end/semaine
- mois

Ajouter la feature suivante:

- associer chaque crime à un cluster (avec k-means)

Supprimer les features suivantes:

- Case Number
- ID
- Updated on
- Longitude
- Latitude
- Location
- X Coordinate
- Y Coordinate

(Peut-être rééquilibrer les classes car il y a plus de crimes sans arrestation que de crimes avec arrestation)

Formalisation des données :

Entrée :

-Les crimes répertoriés à Chicago selon X features (ou une partie des features pour savoir lesquelles influencent) ; label = {0;1} qui correspond à la feature arrestation

Sorties :

-Vecteur binaire si oui ou non il y a eu une arrestation.

-Quelles features influent le plus sur l'arrestation (comprendre pourquoi il n'y a pas eu arrestation: nature du crime, quartier, moment de la journée...)

Features à comparer (qui influencent le mieux):
-Wards; Community Area; K-Means; District → Lieu
-Matin/journée/soir/nuit ; jours de la semaine; week-end/semaine; mois → Moment

Question :

-Certaines de nos features sont des strings (exemple : Primary type ou description). Faut-il les transformer en int pour la classification?