Détection de bout en bout des crimes de la ville de New York grâce à l'apprentissage automatique

Khiyari Oussema, *Etudiant* Jendoubi Firas, *Etudiant*
Teborbi Riadh, *Professeur*

***Abstract*—La criminalité est un problème social précaire et courant dans le monde entier. Les crimes affectent la qualité de vie, la croissance économique et la réputation d'une nation. Au cours des dernières années, le taux de criminalité a énormément augmenté. En réponse à cette augmentation, il existe un besoin de systèmes avancés et de nouvelles approches pour améliorer l'analyse de la criminalité afin de protéger les communautés. Bien que la prédiction précise de la criminalité en temps réel contribue à réduire les taux de criminalité, elle reste un problème difficile pour la communauté scientifique car les occurrences de la criminalité dépendent de nombreux facteurs complexes. Dans ce travail, diverses techniques de visualisation et des algorithmes d'apprentissage automatique sont adoptés pour prédire la distribution des crimes dans la ville de New York. Dans un premier temps, un ensemble de données brutes a été traité, et de multiples techniques de visualisation ont été adoptées pour mieux comprendre les données et les relations entre les différentes variables. Ensuite, plusieurs algorithmes d'apprentissage automatique ont été utilisés pour prédire les types de délits en fonction des données fournies par les utilisateurs et en tenant compte de leur situation géographique. La dernière étape consiste à développer une interface utilisateur en utilisant Streamlit pour rendre l'interaction avec l'utilisateur plus simple.**

I. Introduction

Les crimes sont des problèmes sociaux courants qui affectent les essais.Les crimes sont des facteurs importants qui affectent diverses décisions vitales de la vie d'un individu, comme déménager dans un nouvel endroit, se déplacer au bon moment, éviter les zones à risque, etc. Les délits affectent et diffament l'image d'une communauté. Les crimes affectent et diffament l'image d'une communauté. Ils affectent également l'économie d'une nation en imposant une charge financière au gouvernement en raison du besoin de forces de police supplémentaires, de tribunaux, etc.

Comme les crimes augmentent de manière drastique, nous sommes alarmés de les réduire à un rythme encore plus rapide. Globalement, l'indice de criminalité de la ville de New York a augmenté de 1,3 pourcentage en 2021 par rapport à 2020. Seuls les cambriolages ont connu une baisse de 13,7 pourcentage par rapport à 2020, mais les vols qualifiés ont augmenté de 15,8 pour cent et les agressions criminelles de 13,8 pourcentage . Nous pouvons réduire ces chiffres si nous pouvons analyser et prévoir les scènes et les lieux de crime et prendre des mesures préventives à l'avance.

Les taux de criminalité peuvent être réduits de manière significative grâce à la prévision de la criminalité en temps réel et à la surveillance de masse, qui permettent de sauver les vies les plus précieuses. Une analyse appropriée des données sur la criminalité antérieure permet de prévoir les crimes et contribue ainsi à réduire le taux de criminalité. Le processus d'analyse consiste à examiner les rapports de criminalité et à identifier le plus rapidement possible de nouveaux modèles, séries et tendances.

Cette analyse permet de préparer des statistiques, des requêtes et des cartes à la demande. Le type de crime peut être prédit car les criminels sont actifs et opèrent dans leurs zones de confort, et ils sont susceptibles de reproduire le même crime s'ils réussissent le premier. Les criminels trouvent généralement un lieu et une heure similaires pour tenter le prochain crime.

Bien que cela ne puisse pas être exact dans tous les cas, la possibilité de répétitions est élevée, selon les études, ce qui rend les crimes prévisibles. Cet article propose une application web et une interface visuelle d'outil de prédiction de crime construite avec python en utilisant diverses bibliothèques telles que Streamlit pour l'interface utilisateur, Pandas pour le traitement des données, Folium pour la visualisation des données géographiques, etc. Le cadre proposé utilise différentes techniques de visualisation pour montrer la tendance des crimes et différentes façons de prédire les crimes en utilisant des algorithmes d'apprentissage automatique.

Le cadre proposé utilise différentes techniques de visualisation pour montrer la tendance des crimes et les différentes manières de prédire les crimes à l'aide d'algorithmes d'apprentissage automatique. En bref, la phase de prétraitement consiste à nettoyer et à transformer les données. La phase de visualisation génère divers rapports et cartes pour le processus de diagnostic et d'analyse, et enfin, dans la phase de construction du modèle, différents algorithmes d'apprentissage automatique sont utilisés pour la classification des crimes qui peuvent se produire dans un endroit particulier.

II. Dataset

Ce travail s'appuie sur le jeu de données historique NYPD Complaint Data . Cet ensemble de données comprend tous les crimes, délits et violations valides signalés au département de police de la ville de New York (NYPD) de 2006 à 2019. L'ensemble de données contient 6901167 plaintes et 35 colonnes comprenant des informations spatiales et temporelles

sur les occurrences de crimes ainsi que leur description et leur classification pénale.

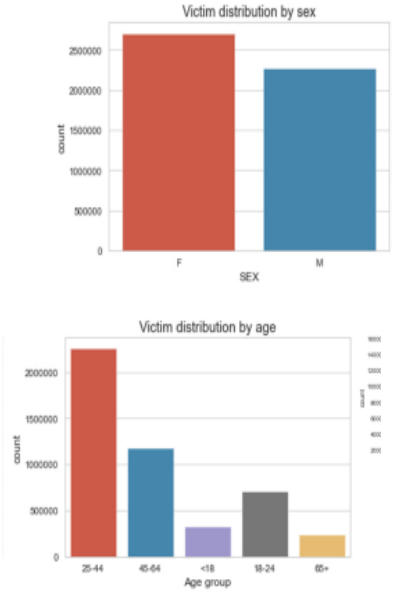
A. Analyse exploratoire des données

Afin de mieux comprendre les données en main et d'analyser les les différentes distributions et relations entre les caractéristiques, nous procédons à une analyse exploratoire des données (AED) afin de répondre aux questions sur la nature, le lieu et le moment des crimes.

B. Classification des crimes

Les types de crimes sont classés selon le code pénal en 3 types : les crimes, les délits et les violations, classés du plus au moins grave. On montre que les crimes les plus courants sont les délits. On constate également que l'ensemble de données n'est pas équilibré. Pour résoudre ce problème lors de la formation du modèle, nous prenons un sous-ensemble de 800 000 plaintes pour chaque classe.

Il est également important de savoir quels crimes spécifiques sont les plus fréquents . En outre, nous analysons le profil des victimes pour savoir quelles catégories sont les plus susceptibles d'être victimes de crimes. Nous pouvons constater que les femmes sont plus touchées par les crimes. De plus, les personnes âgées de 25 à 44 ans sont plus susceptibles d'être victimes que tout autre groupe d'âge. Ces caractéristiques peuvent être utiles pour prédire la probabilité d'un crime une fois combinées à d'autres caractéristiques spatio-temporelles.



Répartition des victimes d'actes criminels selon le sexe et l'âge

C. Répartition géographique et temporelle

Pour étudier la répartition géographique des crimes dans la ville de New York et leurs caractéristiques, l'ensemble de données fournit les coordonnées géographiques (latitude et longitude) de la scène du crime. Nous utilisons ces coordonnées pour tracer une carte thermique et découvrir les points chauds. La figure 44 illustre un exemple de ces cartes de chaleur pour la journée du 02/04/2018. Le résultat présenté va bien avec la répartition des crimes le long des cinq arrondissements de New York . Il montre des niveaux de criminalité élevés à Manhattan, Brooklyn et dans le Bronx Nous étudions la distribution temporelle de l'occurrence des crimes à travers le mois, les jours et les heures de la journée. L'analyse prouve que la plupart des crimes se produisent pendant la saison estivale, les jours de travail et le soir, comme le montre .

D. Nettoyage des données

Afin de préparer l'ensemble de données pour la phase de modélisation, nous avons procédé à un nettoyage des données. Tout d'abord, nous avons supprimé manuellement les caractéristiques redondantes telles que les codes de criminalité, les dates de rapport et les dates d'expiration. Les coordonnées géographiques dans d'autres systèmes. Ensuite, les caractéristiques temporelles ont été transformées pour extraire les années, les mois, les jours et les heures, chacun séparément.Ensuite, nous avons traité les données manquantes soit en supprimant les lignes dans le cas où seules quelques valeurs sont manquantes pour la colonne. valeurs manquantes pour la colonne, soit en considérant les valeurs comme des booléens et les valeurs manquantes sont fausses (parc, logement, caractéristiques de la station). La dernière étape consiste à ne conserver que les valeurs valides dans les colonnes,en éliminant les valeurs non pertinentes, par exemple dans la caractéristique groupe d'âge.Ces procédures permettent d'obtenir un jeu de données propre comportant 6882147 plaintes et 23 colonnes.

III. Modélisation

Après le pré-traitement des données, afin de classifier trois types de crimes, en fonction de leur gravité, plusieurs algorithmes d'apprentissage algorithmes d'apprentissage automatique ont été appliqués afin de comparer leurs résultats, notamment les classificateurs Random Forest (RF) et XGBoost . La classification est principalement utilisée pour reconnaître les classes étiquetées en connaissant leurs attributs (caractéristiques) dans l'ensemble de données. l'étiquette de classe pour les instances dont les caractéristiques sont connues. Ainsi,l'utilisation des classificateurs dans la prédiction de la criminalité permet de construire un modèle orienté vers l'avenir pour identifier le type de criminel dans un délai spécifique.. Dans cette section, une description de tous les classificateurs utilisés est présentée dans le tableau.

IV. Résultats expérimentaux

La tâche de classification a été réalisée à l'aide de trois classificateurs. Pour les tâches de classification, la matrice de confusion et la courbe ROC sont des mesures appropriées pour évaluer la performance du modèle.

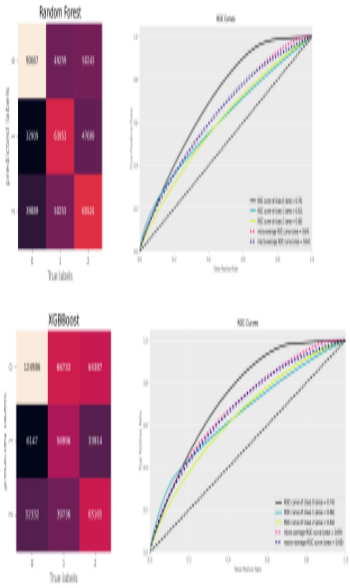
A. Matrice de confusion

Il s'agit d'une mesure de performance pour les problèmes de classification par apprentissage automatique où les résultats peuvent être deux ou plusieurs classes (3 classes dans notre cas). Il contient quatre mesures importantes

B. Courbe ROC

Il s'agit d'un graphique montrant les performances d'un modèle de classification à tous les seuils de classification. Cette courbe représente le taux de vrais positifs et le taux de faux positifs.

Ainsi, le score f1 a été mesuré en calculant les valeurs de préci- sion et de rappel. En outre, la comparaison a été faite entre les trois modèles, où la précision a également été mesurée pour chacun des modèles. Les formules de la matrice de con- fusion et de la courbe ROC sont présentées ci-dessous :



Les matrices de confusion et les courbes ROC

Par conséquent, comme le montre le tableau ci dessous, pour les scores des classifieurs en utilisant la matrice de confusion. Cependant, les modèles RandomForest et XG-Boost ont tendance à avoir des scores de validation très proches qui sont respectivement de 0,52 et 0,6. Cependant, le modèle XGBoost tend à être le meilleur modèle de classification pour prédire correctement les classes de crimes avec des scores de précision de 0,60. En outre, à partir des cartes thermiques de la matrice de confusion, nous avons une tâche de classification multi-classes avec 3 classes de 0 à 2 où il représente les la- bels de classe qui ont remplacé les noms de classe réels. D'après ce que nous pouvons voir dans la matrice de confusion et les courbes ROC, nous pouvons constater que le modèle XGBoost surpasse les autres classificateurs avec les prédictions les plus élevées dans chaque classe. Pourtant, ils ont tous tendance à avoir un faible nombre de vraies classifications pour chaque classe.

Classifier	Précision	Rappell	F1 score	Efficacité
Random Forest	0.52	0.51	0.49	0.52
XGBoost	0.62	0.60	0.59	**0.60**

Méthodes de classification et leur descriptions

V. Interface utilisateur

Après avoir entraîné le modèle et sauvegardé le fichier des poids, nous avons construit une application web en utilisant Streamlit et Folium pour permettre à l'utilisateur d'interagir avec la carte et de pré-dicter le type de crime qui pourrait se produire. L'utilisateur peut entrer son sexe, sa race, son âge, la date et l'heure à laquelle il veut prédire le type de crime, l'emplacement sur la carte et enfin, le lieu (dans un parc, dans un logement social ou dans une gare). Ces informations sont ensuite transformées pour s'adapter à l'entrée du modèle,et ensuite, en utilisant le fichier de poids du modèle chargé, nous prédisons le type de délit et le renvoyer à l'utilisateur avec les sous-types potentiels de ce délit.

VI. Conclusion

La criminalité est un problème important dans de nombreuses villes. Par conséquent, de nombreux chercheurs ont essayé de le résoudre et de prédire les points chauds les plus criminels afin d'améliorer la compréhension des endroits dangereux à certains moments. Dans cet article, nous avons analysé les données de la ville de New York afin de reconnaître les modèles spatio-temporels des incidents criminels. Ainsi, à partir de l'analyse des données, nous pouvons distinguer trois grands types de crimes qui se sont produits de 2006 à 2019. De plus, à partir de l'analyse de la visualisation temporelle sur l'échelle des jours, le taux le plus élevé d'activité et de récurrence des incidents criminels se situe les samedis et les dimanches, plus précisément de 12h à 6h du matin.

De plus, nous avons proposé une méthodologie pour classer et prédire le type de crimes en classant les données spatio-temporelles.

à l'aide de deux algorithmes d'apprentissage automatique : Random Forest et XGBoost.Les classificateurs Random Forest et XGboost ont obtenu une précision de prédiction très proche de 0,52 et 0,6, respectivement. Cependant, ils ne parviennent pas à obtenir des prédictions décentes en général. Enfin, nous avons créé une interface utilisateur pour permettre aux utilisateurs d'entrer leurs informations et d'obtenir la classe du crime qui peut se produire dans un endroit particulier à un moment précis.

Dans le cadre de travaux futurs, pour une meilleure classification, il semble que l'apprentissage profond, comme les réseaux neuronaux artificiels profonds ou les encodeurs automatiques profonds, pourrait être utilisé. Ces dernières années, l'apprentissage profond a pris le pas sur les modèles ML traditionnels. Il a donc un bon potentiel pour améliorer les performances de classification en prédisant les types de criminalité et les points chauds.