Analyse et prédiction des crimes dans la ville de New-York en utilisant une approche d'apprentissage automatique et des données spatiotemporelles

Taieb Hadjkacem et Mohamed Chaaben

***Abstract*— Pour les New-Yorkais, l'idée d'un retour aux années 1990, période de forte criminalité, fait froid dans le dos. Or le spectre de cette ère violente a resurgi ces dernières années dans la presse locale. Le nombre de fusillades a augmenté de 97% et celui de meurtres de près de 45%, dont 15 semaines de violences armées consécutives, en 2020 dans la ville de New York, selon les dernières statistiques de la police de New York (NYPD). Les analystes civils et les fonctionnaires du ministère de la police de New York utilisent un outil de calcul unique pour repérer les tendances dans les données sur la criminalité. Une collection de modèles d'apprentissage de la machine, nommée Patternizr, a été créée en 2016, mais le département n'a révélé le système qu'en 2019. L'analyse prédictive vise donc à optimiser l'utilisation de ces données pour anticiper les événements criminels. Dans ce travail, nous utilisons des modèles d'apprentissage automatique pour la prédiction spatio-temporelle des différents types de crimes à New York, à partir d'ensembles de données sur la criminalité de NYPD. Nos résultats vont être intégrés dans une solution web pour lutter et prévenir contre la criminalité en aidant les utilisateurs à prédire les crimes selon les données fournies.**

I. INTRODUCTION

Beaucoup considèrent la criminalité comme un problème social, un problème défini par la société, comme le sans-abri, la toxicomanie, etc. D'autres diraient que la criminalité est un problème sociologique que les sociologues définissent comme un problème et qui devrait être traité en conséquence par les sociologues.

La criminalité fait partie intégrante de toute société, et ses coûts et ses conséquences touchent presque tout le monde dans une certaine mesure. Les types de coûts et d'impacts sont très différents. En outre, certains coûts sont à court terme, tandis que d'autres survivent tout au long de la vie. D'autres coûts pour les victimes peuvent inclure les frais médicaux, les pertes d'actifs et les pertes de revenus. Les pertes subies par les victimes peuvent également prendre la forme de dépenses accrues en matière de sécurité, telles que le renforcement des serrures, l'éclairage supplémentaire, le stationnement dans des propriétés plus chères et sûres, les avertissements de sécurité pour les maisons et les voitures et l'entretien des chiens de garde. Des sommes considérables sont dépensées pour éviter les victimes. D'autres types de dépenses peuvent concerner une victime ou une personne, la peur du crime, le déménagement dans un nouveau quartier, les frais funéraires, les frais d'avocat et la perte des jours d'école.

La technologie peut servir à réagir ou même à prévenir des crimes. Les systèmes d'information géographique, la télédétection et les applications de cartographie des données permettent aux organismes d'application de la loi de faire d'énormes progrès en matière de renseignement criminel, d'analyse de la criminalité, d'intervention d'urgence et de maintien de l'ordre.

Les technologies de cartographie utilisent de l'information en temps réel pour évaluer où se produisent les crimes, selon le type de crimes qu'ils sont. Ils peuvent repérer les points chauds et les tendances d'activité, ce qui leur permet de cibler les enquêtes et d'améliorer la prévention du crime.

De même, l'analyse moderne des données utilise l'apprentissage automatisé pour identifier les modèles et, dans un laps de temps nettement plus court, pour établir des profils criminels plus précis qui pourront ensuite être utilisés pour détecter les criminels et aussi prédire les crimes selon des données spécifiques.

L'inclusion de données spatiales et temporelles dans les ensembles de données de criminalité à l'aide du SIG a révolutionné les systèmes de prévision de la criminalité. L'information spatio-temporelle aide les chercheurs à présenter des systèmes de prévision de la criminalité plus crédibles et plus précis qui peuvent être fiables en matière de prévention de la criminalité.

II. TRAVAUX CONNEXES

A. Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention [1]

Le but de cette étude est de déterminer comment une combinaison de l'apprentissage automatique et de vision par ordinateur peut être utilisée par les autorités répressives ou les autorités pour détecter, prévenir et résoudre les crimes, à un rythme beaucoup plus rapide et plus précis. Elle prouve que les techniques de l'apprentissage automatique et de vision par ordinateur peuvent faire évoluer les cabinets d'avocats.

B. Crime Prediction Model using Deep Neural Networks [2]

Ce projet étudie la faisabilité d'utiliser des techniques d'apprentissage de la machine, en particulier des réseaux de neurones, pour prédire les comportements criminels sur la base de l'histoire des arrestations. L'expérience doit gérer des fréquences déséquilibrées. Il s'est concentré sur la manière dont les réseaux neuronaux peuvent être avantageux dans la classification de la prédiction de la criminalité.

III. Méthodologie

Dans cette section, nous décrivons notre méthodologie pour construire des modèles d'apprentissage automatique et les valider sur les données de la criminalité de New York.

Notre objective est de prédire les types de crimes qui peuvent subir une personne selon des données bien spécifiques comme son âge, son sexe, sa position etc...

Nous allons présenter maintenant la démarche de notre solution dans la figure 1.

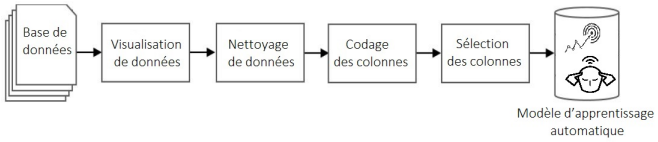

Fig. 1. Pipline du projet

A. La collection de données

Nous allons utiliser une base de données fournis par le département de police de New York (NYPD)[3]. Cette base contient tous les crimes, délits et blessures signalés au (NYPD) entre 2006 et la fin de l'année 2019 sous format csv.

Cette base de données est publiée depuis 2016 et elle est mise à jour chaque année. Elle contient 35 colonnes où nous trouvons 55 types de délits, 7 million de lignes où chacune présente une plainte et un fichier de documentation explique la signification de chaque colonne fournie. Par exemple, elle comporte la date et le temps d'occurrence du crime et même les données spatiales du lieu d'occurrence, des données sur la victime et le suspect comme l'âge, la race, et le sexe... Elle comporte des données numériques, catégoriques et time series.

B. La visualisation de données

Un grand nombre d'informations présentées sous forme graphique ou sous forme de statistique sont plus faciles à comprendre et à analyser. Nous commençons par comprendre la signification de chaque attribut puis visualiser les différentes valeurs pour chaque colonne.

Ensuite, nous déterminons le pourcentage de lignes vides pour chaque colonne.

La figure 2 illustre les victimes des agressions sexuelles selon l'âge.

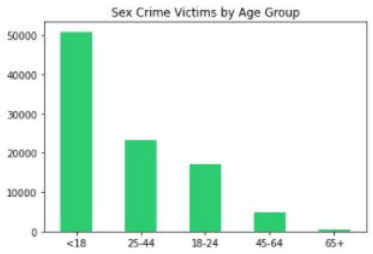

Fig. 2. Les victimes des agressions sexuelles selon l'âge

C. Le nettoyage de données

Cette procédure permet d'éliminer le bruit et de corriger les incohérences dans les données par le remplissage des données manquantes en utilisant des techniques d'imputation.

Comme première étape, nous avons supprimé les colonnes avec un pourcentage supérieur à 60% de lignes vides (PARKS_NM, STATION_NAME, TRANSIT_DISTRICT, HADEVELOPT, HOUSING_PSA). Aussi, nous avons constaté que les données liées à la personne suspecte ne sont pas à la disposition de l'utilisateur qui veut savoir s'il va être une victime de crime alors il faut les supprimer.

Pareillement, nous avons décidé de supprimer les lignes avec des valeurs NaN pour des colonnes et de remplacer ces valeurs NaN par le terme UNKNOWN dans d'autres selon le pourcentage de vide.

Enfin, nous avons choisi de remplir les autres valeurs manquantes en fonction de la distribution des valeurs dans l'ensemble de données temporelles. Quant aux valeurs horodatées, nous avons remplacé toutes les valeurs nulles par la valeur médiane dans chaque colonne.

D. Le codage des colonnes

Un modèle d'apprentissage automatique exige que toutes les valeurs soient numériques. Nous devons convertir toutes les valeurs en chiffres utilisant les techniques de One hot encoding et Label Encoder. Pour le One hot encoding, c'est transformez toutes les valeurs uniques en listes de 0 et 1, la valeur cible étant 1 et le reste 0. Et pour le Label Encoder, c'est transformez les étiquettes en valeurs numériques uniques.

Nous avons construit des valeurs dérivées à partir de nos données initiales qui sont plus informatives et non redondantes. Nous avons commencé par générer des colonnes d'année, de mois et de jour basées sur CMPLNT-FR-DT.

Ensuite, à partir de CMPLNT_FR_TM, nous avons classé les différents jours en quatre classes : morning, afternoon, evening et night en donnant à chaque classe un numéro de 1 à 4. De la même façon, avec 55 types de crimes, nous les avons regroupés en seulement dix classes les plus existantes comme illustré dans la figure 3 et nous avons regroupé la colonne âge en 6 classes.

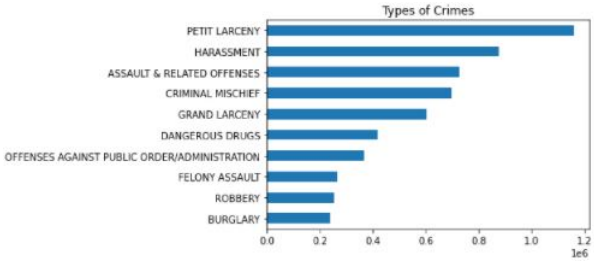

Fig. 3. Les types de crimes

Les 5 colonnes VIC_RACE, VIC_SEX, BORO_NM, PATROL_BORO et LOC_OF_OCCUR_DESC ont été encodés

en utilisant la technique One hot encoding car ces variables ne présentent aucun ordre naturel à prendre en considération.

E. La sélection des colonnes

Il s'agit de sélectionner les colonnes les plus pertinents de notre ensemble de données. Le but est la réduction possible de la surcharge et du temps d'entraînement (moins de données globales et moins de données redondantes pour l'entraînement) et amélioration de la précision.

Nous avons gardé seulement les colonnes les plus pertinentes pour les utiliser dans notre modèle. Nous avons seulement préservé les colonnes de longitude, latitude et le quartier, l'emplacement par rapport à un local, les informations de la victime comme le sexe, la race et l'âge, les données temporelles. En total, nous disposons 42 colonnes.

F. Modèle d'apprentissage automatique

Dans cette étape, nous avons mis en œuvre plusieurs méthodes de classification puis nous avons les évalué avec la matrice de confusion et différentes autres métriques de performances.

- Random forest : est une méthode qui permet d'obtenir des modèles prédictifs pour la classification et la régression. L'idée générale derrière la méthode est la suivante : au lieu d'essayer d'obtenir une méthode optimisée en une fois, on génère plusieurs prédicteurs avant de mettre en commun leurs différentes prédictions.
- La régression logistique : est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives Xi et une variable qualitative Y. Il s'agit d'un modèle linéaire généralisé utilisant une fonction logistique comme fonction de lien. Un modèle de régression logistique permet aussi de prédire la probabilité qu'un événement arrive (valeur de 1) ou non (valeur de 0) à partir de l'optimisation des coefficients de régression.
- Decision Tree : est un classificateur arborescent, où les nœuds internes représentent les caractéristiques d'un ensemble de données, les branches représentent les règles de décision et chaque feuille représente le résultat.

IV. RÉSULTATS

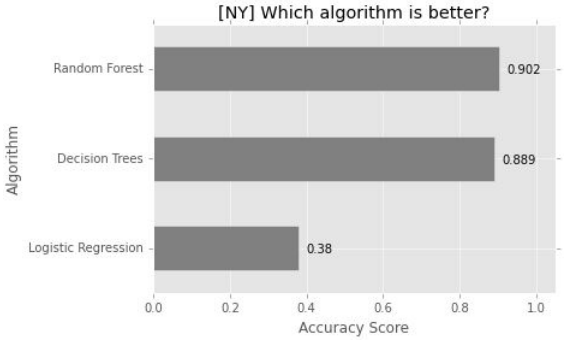

Fig. 4. Les accuracy des algorithmes

Pour cette étude, nous avons entrainé et testé tous les algorithmes mentionnés ci-dessus. À partir des résultats mentionnés dans la figure 4, nous observons que l'accuracy du modèle Random forest est la plus élévée. Aussi ce modèle a atteint une valeur de Recall égale à 0.9, une valeur de Precision égale à 0.9, et une valeur de F1 Score égale à 0.9.

Quant à la visualisation des résultats finaux de notre recherche, nous avons créé une interface utilisateur graphique sous la forme d'une application web utilisant le framework Flask ainsi que des techniques de cartographie web. Dans cette interface, nous avons tracé la carte de New York et nous avons donné à l'utilisateur la possibilité de choisir un endroit dans la carte pour laquelle il a l'intention de visiter, entrer ses données telles que son âge, son sexe, sa race et l'heure à laquelle il se rendra sur le site, puis comme résultat, nous avons affiché le type de crime le plus probable qui sera commis contre lui. Dans le côté serveur de notre application, nous avons intégré notre modèle d'apprentissage automatique prêt à l'emploi sous format h5 et nous avons créé une API qui prend les données fournies par l'utilisateur, appliquer les transformations nécessaires sur elle, prédit le type de crime à l'aide de notre modèle, puis renvoie le résultat au côté client de notre application.

V. CONCLUSIONS

Dans ce travail, on a utilisé la base de données fournie par NYPD. Tout d'abord, nous avons compris la signification des données puis nous avons fait le nettoyage des données ainsi que le codage pour qu'elles soient prêtes à l'utiliser avec des modèles d'apprentissage automatique. Ensuite, on choisit seulement les colonnes les plus pertinentes comme entrée de nos modèles. Puis, nous avons appliqué plusieurs modèles d'apprentissage automatique dans le but de prédire les types de crimes qui peuvent subir une personne selon des données fournies par l'utilisateur. Enfin, nous avons créé une application web qui aide nos utilisateurs à entrer des données particulières pour afficher comme résultat le type de crime le plus probable qui sera commis contre lui.

Dans les prochains travaux, on vise à schématiser les densités des crimes dans une carte dans notre application et aussi améliorer notre modèle en ajoutant quelques données qui peuvent être significatives.

REFERENCES

[1] Shah, N., Bhagat, N. and Shah, M. Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention. Vis. Comput. Ind. Biomed. Art 4, 9 (2021)
[2] https://dl.acm.org/doi/fullHtml/10.1145/3325112.3328221
[3] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i