Prédiction et Classification de Crimes de New York City

Abid Oumayma
SUP'COM
oumaima.abid@supcom.tn

Ghorbel Zeineb
SUP'COM
zeineb.ghorbel@supcom.tn

Kanzari Mohamed Yacine
SUP'COM
mohamedyacine.kanzari@supcom.tn

Selmi Rami
SUP'COM
rami.selmi@supcom.tn

RÉSUMÉ

La prédiction de la criminalité basée sur des données sociodémographiques traditionnelles a une valeur limitée car il ne parvient pas à saisir la complexité et la dynamicité de l'activité humaine dans les villes. Avec les nouvelles technologies omniprésentes, il est possible d'améliorer les modèles de prédiction de la criminalité avec des données de crowdsourcing qui produisent de meilleures procurations de l'activité humaine.

Dans ce papier, nous proposons l'utilisation de l'historique des données de plaintes à New York. Le dataset comprend tous les crimes, délits et infractions valides signalés au service de police de New York. Certains crimes sont terminés avec succès tandis que d'autres sont « tentés » mais ont échoué ou ont été interrompus prématurément. Nous cherchons à comprendre le schéma des activités criminelles réussies en créant des modèles de classification pour prédire le type du crime commis.

MOTS CLÉS

Prédiction, Crime, Classification, New York City

1 INTRODUCTION

La prédiction des crimes est intrinsèquement difficile. La criminalité est un phénomène social complexe produit par trois forces :

- La motivation du délinquant.
- La vulnérabilité de la victime.
- L'absence d'une tutelle compétente ou l'environnement (considéré comme le moment et le lieu de la victimisation) où le délinquant et la victime viennent ensemble.

Cela donne un système dynamique et complexe, et les chercheurs étudient encore diverses caractéristiques des trois forces du pouvoir prédictif. Traditionnellement, les études criminologiques se sont concentrées uniquement sur les attributs socio-démographiques en tant que facteurs en corrélation avec victimisation et ont remarqué que des groupes spécifiques de personnes étaient confrontés à un risque de victimisation plus élevé que autres groupes. Mais les données de recensement ont une limite intrinsèque, en ce qu'elles n'offrent qu'une image statique et parfois obsolète image de la ville, sans capturer la dynamique des gens dans le temps et dans l'espace. Il y a maintenant la possibilité de facteurs non conventionnels à intégrer dans la prévision de la criminalité modèles en exploitant de nouvelles sources de données qui reflètent la structure et la dynamique de nos villes. Avec l'émergence de nouvelles technologies de l'intelligence artificielle , une panoplie de sources de données peut désormais offrir de meilleurs proxys pour l'activité humaine.

Dans ce papier, nous évaluons le pouvoir prédictif des facteurs de prédiction de la criminalité dérivés d'informations provenant de l'historique des complaintes de New York City pour classifier ensuite ces crimes selon le type. Il s'agit d'une étape initiale mais essentielle vers l'élaboration de modèles de prévision de la criminalité robustes à partir de données sources décrivant la dynamique humaine en milieu urbain.

On commencera par étudier l'existant en donnant un bref aperçu de sur les modèles de prédiction de la criminalité existants dans la communauté d'exploration de données. Puis on explique en détail l'ensemble de données, l'approche proposée et les outils utilisés. Ensuite on procède à une évaluation des performances de notre approche et on finit par conclure et proposer nos perspectives.

2 ÉTAT DE L'ART

2.1 PredPol

PredPol [1] est un logiciel commercial proposé par une société américaine, PredPol Inc., située à Los Angeles où il a été expérimenté en premier lieu par le LAPD19. L'outil a pour objectif de prédire, avec précision et en temps réel, le lieu et le moment où les crimes ont le plus de risques de survenir. Autrement dit, cet outil identifie les zones à risque, suivant le modèle statistique utilisé en sismologie.

Ce service a convaincu plusieurs dizaines de villes aux États-Unis. Outre Los Angeles, il a par exemple été utilisé dans le Comté du Kent au Royaume-Uni ou encore à Chicago. Les données entrantes sont : les archives de la police d'une ville ou d'un territoire spécifique (procès-verbaux, suivis d'arrestations, appels au secours), afin d'en déduire les endroits où les crimes sont les plus fréquents pour « prédire »les lieux à surveiller en priorité. La cible visée porte sur des lieux et non sur des personnes.Les types d'infractions concernées sont les cambriolages, vols de voitures et vols dans les lieux publics.

Analyse critique de PredPol. Les études techniques de l'outil PredPol sont rares en raison d'une politique très limitée. Encadrement des risques techniques et juridiques des activités de police prédictive et secours prédictifs d'ouverture des données utilisées et a fortiori du code développé par la start-up, protégé par le secret commercial. Quelques études techniques ont néanmoins pu être menées, en se basant sur quelques bases de données mis en accès libre par des villes comme Chicago et sur des modèles semblables à PredPol. Il est difficile d'évaluer la valeur ajoutée de cette prévision par rapport aux cartes historiques hotspots ou par estimation

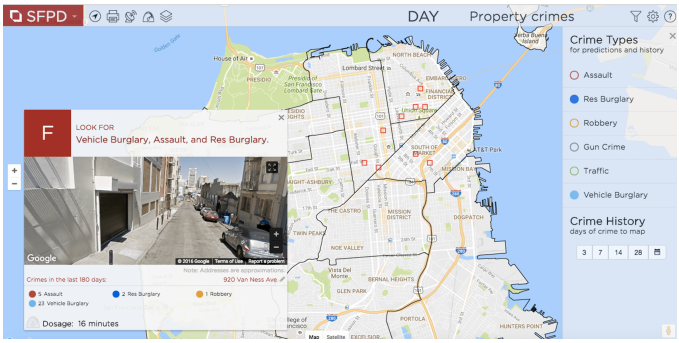

Fig. 1 : La police prédictive est construite autour d'algorithmes qui identifient les points chauds potentiels de la criminalité.

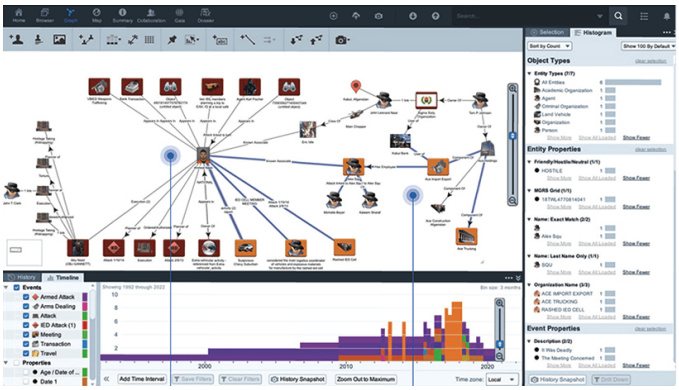

Fig. 2 : Prédiction des potentiels crimes.

de la densité par noyau. En effet, les rares travaux publiés d'évaluation de cette approche ne concerne pas la qualité de prévision mais les statistiques de criminalité qui sont sans doute plus sensibles à la stratégie de gestion des patrouilles qu'à une amélioration de la prévision par rapport à un historique classique[4].

2.2 Palantir Crime Risk Forecasting

Palantir Crime Risk Forecasting [2] est le brevet déposé et détenu par la société Palantir Technologies Inc. Ce dispositif a été déployé par exemple à Los Angeles, New York ou à la Nouvelle-Orléans qui a souscrit un contrat en 2012 afin d'aider le département de police à cibler les chauffeurs violents. Ce contrat n'a cependant pas été renouvelé en raison des controverses suscitées par son objet ciblant la population. Les plaques d'immatriculation sont en effet des données personnelles. Mais le plus souvent, les contrats conclus avec les villes sont secrets, si bien qu'il est difficile de les contester.

L'intention affichée de Palantir est de mettre en œuvre une technologie pour prévenir les risques criminels, afin d'aider la police à savoir où et quand la criminalité va survenir dans le futur. La prévision du risque de criminalité est associée à une cellule géographique et temporelle de base, par exemple de 250m carré, sur une période de 8h correspondant à la durée d'une patrouille de police. Ces valeurs, comme beaucoup d'autres du logiciel, peuvent être paramétrées. Des conventions graphiques : transparence, couleur... permettent de visualiser par cellule de la carte interactive l'intensité des risques, selon une échelle à définir, en fonction des types de criminalité choisis.

Analyse critique de Palantir.

- Les possibilités de combinaison, agrégation de modèles et algorithmes, associées à celles des très nombreuses variables explicatives qui peuvent être prises en compte, engendrent une très forte complexité et donc un nombre considérable de paramètres à estimer et d'hyper paramètres à optimiser. Le brevet ne dit rien sur la façon dont sont optimisés ces paramètres ni sur la qualité attendue des prévisions. Le brevet liste plutôt une immense combinatoire de combinaisons ou perspectives de combinaisons possibles sans rien dire sur la stratégie de choix à adopter.
- Il nous semble difficile qu'un service de police s'empare concrètement de cet outil sans une aide constante de spécialistes de la société Palantir.
- Tout algorithme ou combinaison produisant une prévision binaire ou plutôt une probabilité d'occurrence d'un crime ou délit peut être implanté sans rien changer aux implications juridiques de l'outil.

3 DATASET : NYPD COMPLAINT DATA HISTORIC

Ce dataset (Table 1) comprend 5 580 035 lignes et 24 colonnes, tous les crimes, délits et infractions valides signalés au département de police de New York (NYPD) de 2006 à la fin de l'année dernière (2019) [3].

4 MÉTHODOLOGIE

Une recherche approfondie, qui utilise divers types de ressources, a permis de définir certaines normes et concepts de base. Par conséquent, cette partie est dédiée à présenter les étapes qui sont liés à la réalisation de notre projet. En fait, nous expliquons d'abord chaque étape, ses prinicpes et les résultats de son application pour mettre ce projet dans un cadre clair. Notre approche vise à classifier les crimes selon le code KY_CD.

Étapes :

(1) Exploration des données.
(2) Nettoyage des données.
(3) Elaboration du modèle.
(4) Apprentissage et test.
(5) Evaluation.
(6) Résultat et Visualisation.

5 EXPLORATION DES DONNÉES

L'exploration de données, connue aussi sous l'expression de fouille de données, forage de données, prospection de données, data mining, ou encore extraction de connaissances à partir de données, a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données, par des méthodes automatiques ou semi-automatiques.

Colonne	Description
CMPLNT_NUM	ID persistant généré aléatoirement
CMPLNT_FR_DT	Date exacte d'occurrence de l'événement signalé
CMPLNT_FR_TM	Heure exacte de l'événement signalé
CMPLNT_TO_DT	Date de fin de l'événement signalé
CMPLNT_TO_TM	Fin de l'heure de l'événement pour l'événement signalé
ADDR_PCT_CD	L'enceinte dans laquelle l'incident s'est produit
RPT_DT	L'événement de date a été signalé à la police
KY_CD	Code de classification des infractions à trois chiffres
OFNS_DESC	Description de l'infraction correspondant au code clé
PD_CD	Code de classification interne à trois chiffres (plus granulaire que KY_CD)
PD_DESC	Description de la classification interne correspondant au code
CRM_ATPT_CPTD_CD	Indicateur du succès de la criminalité
LAW_CAT_CD	Niveau d'infraction : felony, misdemeanor, violation
BORO_NM	Le nom de l'arrondissement dans lequel l'incident s'est produit
LOC_OF_OCCUR_DESC	Lieu précis de l'événement dans ou autour des lieux
PREM_TYP_DESC	Description spécifique des lieux
JURIS_DESC	Description of the jurisdiction code
JURISDICTION_CODE	Juridiction responsable de l'incident.
PARKS_NM	Nom du parc ou des espaces verts de New York
HADEVELOPT	Nom du développement immobilier NYCHA de l'occurrence
HOUSING_PSA	Code de niveau de développement
X_COORD_CD	Coordonnée X, Plan de l'État de New York (FIPS 3104)
Y_COORD_CD	Coordonnée Y, Plan de l'État de New York (FIPS 3104)
SUSP_AGE_GROUP	Groupe d'âge du suspect
SUSP_RACE	Description de la race du suspect
SUSP_SEX	Description sexuelle du suspect
TRANSIT_DISTRICT	district de transit dans lequel l'infraction s'est produite.
Latitude	Latitude (EPSG 4326)
Longitude	Longitude (EPSG 4326)
Lat_Lon	(Latitude, Longitude)
PATROL_BORO	Le nom de l'arrondissement de patrouille dans lequel l'incident s'est produit
STATION_NAME	Nom de la station de transport en commun
VIC_AGE_GROUP	Groupe d'âge de la victime
VIC_RACE	Description de la race de la victime
VIC_SEX	Description du sexe de la victime

Tab. 1 : Description des colonnes du Datasset.

Dans un premier temps, nous essayons de comprendre quel type de données nous devons traiter. En effet, la compréhension des données constitue l'étape initiale pour savoir comment aborder un problème d'analyse de données et d'apprentissage automatique.

On a exploré les niveau des infractions classés par ans sur 3 niveaux : Crime, Délit, Violation sur les (Figure 5, 6, 7).

On a peut aussi voir Niveau des infractions par arrondissement (Figure 8).

On a aussi exploré la corrélation entre la victime et le suspect (Figure 9).

6 NETTOYAGE DES DONNÉES

Le nettoyage des données est un processus qui vise à identifier et corriger les données altérées, inexactes ou non pertinentes. Cette étape fondamentale du traitement des données améliore la cohérence, fiabilité et valeur des données.

Les causes les plus courantes d'inexactitude dans les données sont les valeurs manquantes, les entrées qui n'apparaissent pas dans l'emplacement adéquat et les fautes de frappe. Le nettoyage est basé sur la corrélation des données à travers la matrice de corrélation dans la (Figure 10)

7 ELABORATION DU MODÈLE

Le choix du modèle constitue l'étape la plus cruciale, un modèle performant qui renvoie des résultats envisageables et précis.

L'objectif principal est de former le modèle le plus performant possible en utilisant les données pré-traitées. Pour cela, il est nécessaire d'avoir une image claire des données. Il est donc important de :

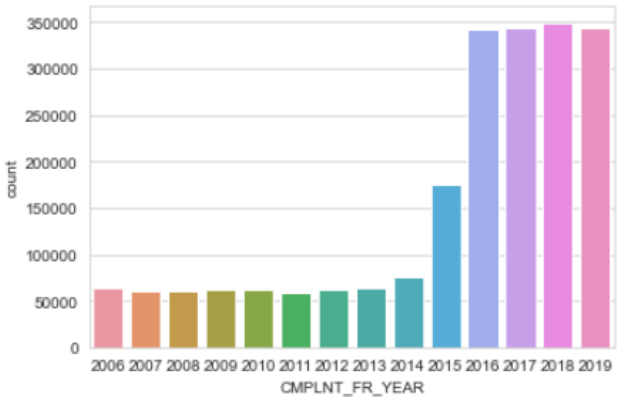

Fig. 3 : Nombre d'infractions par ans.

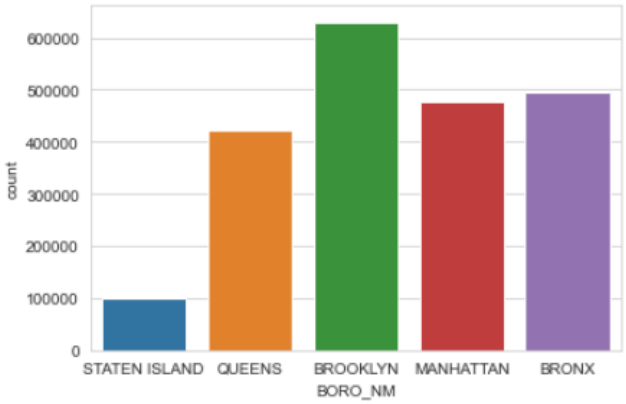

Fig. 4 : Nombre d'infractions par arrondissement.

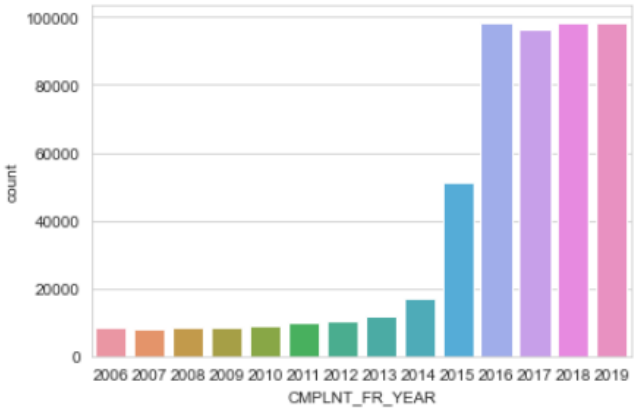

Fig. 5 : Nombre de Crimes par ans.

- Comprendre les données ; le type de données.

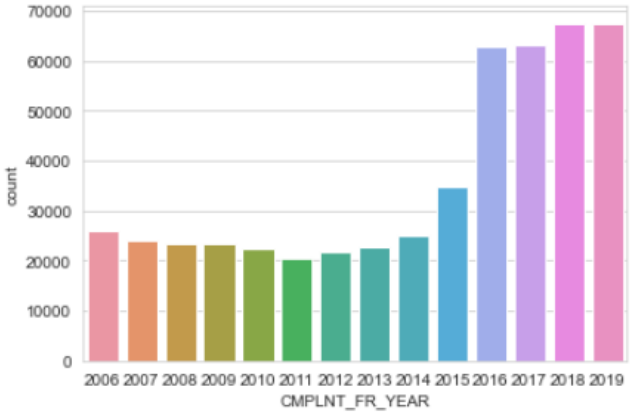

Fig. 6 : Nombre de Violations par ans.

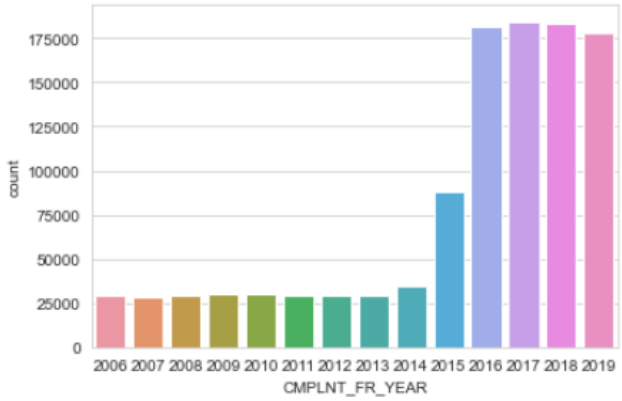

Fig. 7 : Nombre de Délits par ans.

- Catégoriser le problème ; les données sont étiquetées ou non et la sortie est un nombre, une classe ou un ensemble de groupes d'entrée.
- Comprendre les contraintes ; la capacité de stockage des données et la prédiction et la vitesse d'apprentissage souhaitées.

Nous avons utilisé un modèle séquentiel qui se compose de 5 couches dense dont 2 couches consistent l'input et l'output.

La couche d'entrée se compose d'un vecteur d'entrées qui constitue le nombre de features ou colonnes correspondant à 20 features. La profondeur des couches sont 256, 128, 64 et 32. La fonction d'activation utilisée est Relu.

La couche de sortie se compose de 6 classes identifiant le type de crime basé sur KY_CD. La fonction d'activation Softmax est utilisée pour classifier les crimes. On a choisi ADAM comme optimiser et 256 comme la taille du lot avec la fonction de perte categorical_crossentropy suite à la nature de notre sortie (6 classes) (Figure 11).

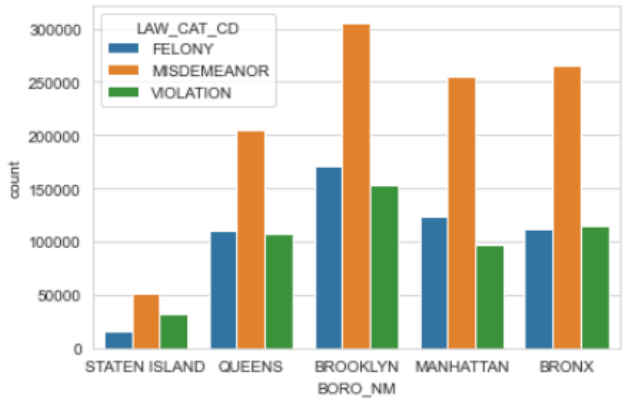

Fig. 8 : Nombre des différentes niveaux des infractions par arrondissement.

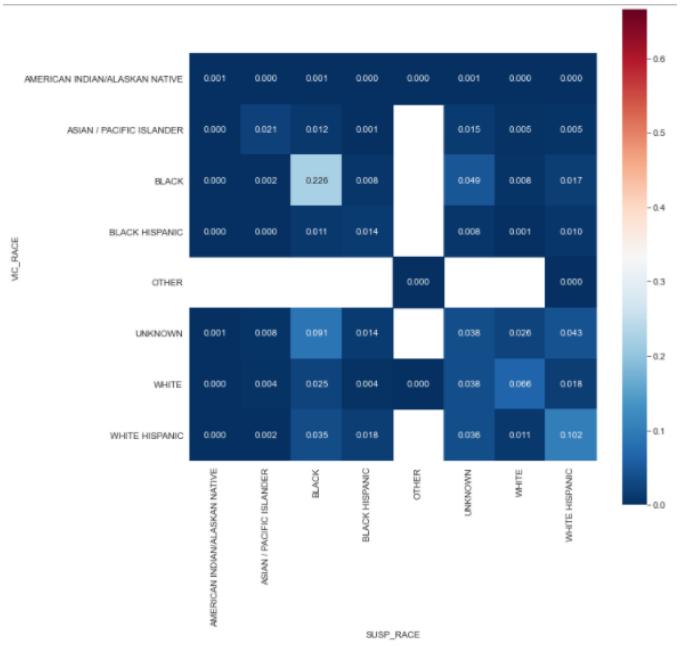

Fig. 9 : Corrélation entre la victime et suspect.

8 APPRENTISSAGE ET TEST

L'algorithme se construit une "représentation interne" afin de pouvoir effectuer la tâche de prédiction (Figure ??). Cette phase consiste à sélectionner les bonnes données test (Figure ??), choisir et entraîner le bon algorithme, en vérifiant grâce à l'analyse d'erreurs que le modèle devient de plus en plus performant et robuste et que les performances s'améliorent lorsqu'on lui fournit les données d'entraînement.

Décomposer le jeu de données en ensembles de formation et de tests pour la validation. Dans notre travail , on a divisé notre dataset en 30% pour le test et 70% pour le train.

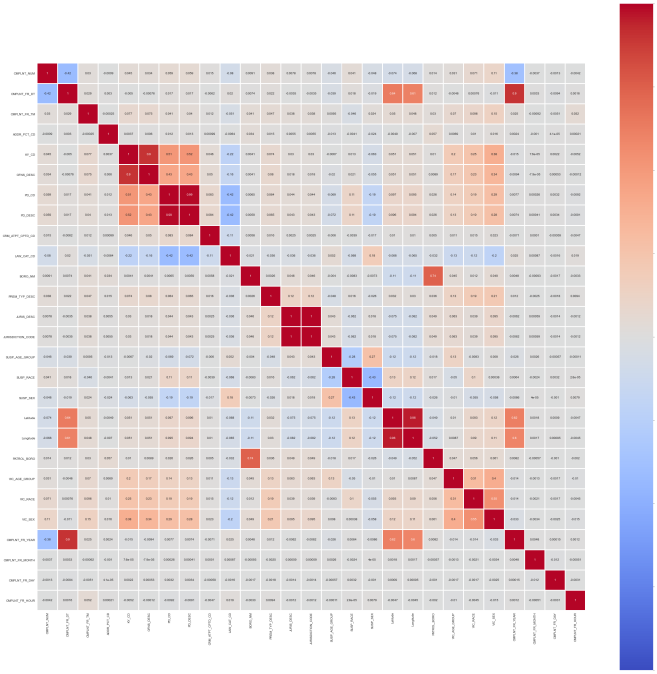

Fig. 10 : Corrélation des données.

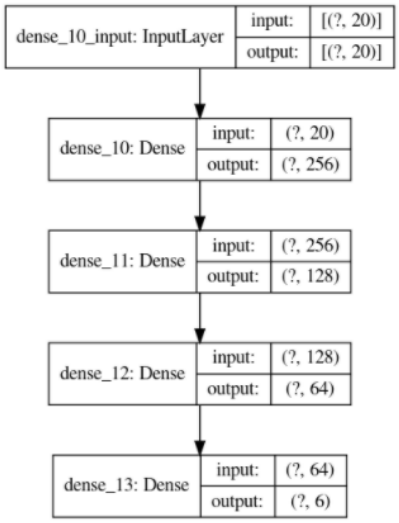

Fig. 11 : Modèle.

9 EVALUATION

Une évaluation rigoureuse des performances d'un modèle est une étape indispensable à son déploiement.

La matrice de confusion est un tableau à double entrée vérifiant la fréquence des prédictions correctes par rapport à la réalité (Figure 12).

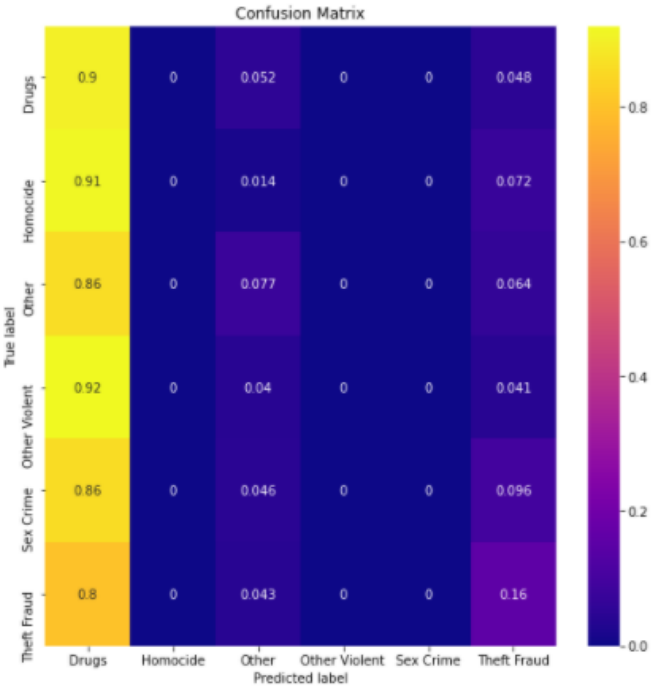

Fig. 12 : Matrice de Confusion.

10 RÉSULTAT ET VISUALISATION

Pour cette partie, React est utilisé comme front-end accompagné avec la bibliothèque ArcGIS et Flask servant comme back-end pour la prédiction.

Suite au taille énorme du dataset, la Visualisation est realisé sur l'année 2019 avec la possibilité de changer l'intervalle de temps et animer la visualisation selon intervalle du temps.

Pour la Visualisation 2D, on a eu recours aux couches : Métro NYC, Arrondissements NYC et Densité des Crimes alors que pour la Visualisation 3D les couches sont : Métro NYC, Arrondissements NYC, Bâtiments et les 10 meurtres infâmes.

La Visualisation 3D nous permet de voir les Bâtiments de New York pour chaque arrondissement ainsi que les 10 meurtres infâmes (Figure 13).

La Visualisation 2D nous permet de voir chaque arrondissement de New York et la densité des crimes sous forme de Heatmap (Figure 14) et lors du zoom une visualisation plus détaillée est révélée des crimes avec leurs types et niveau d'infraction (Figure 15).

On peut filter les crimes par Type Crime, Niveau d'infraction, Journée de la semaine et Heure.

Pour la prediction, l'utilisateur doit se localiser, sélectionner son sexe et éthinicité et sélectionner date et heure et lors de la prediction la résultat s'affiche dans un popup sur sa position (Figure 16).

11 CONCLUSION

De plus en plus de services de police cherchent à utiliser des approches fondées sur des preuves pour prévenir et répondre à la criminalité, et les cartes prédictives sont un outil utile pour confirmer les domaines de crimes. Les cartes montrant les points chauds peuvent être particulièrement utile pour aider à éduquer les nouveaux agents sur leurs domaines.

Dans notre approche, on a élaboré un modèle de prédiction et de classification des crimes en se basant sur la localisation, le sexe et la date pour prévenir un éventuel crime. Cette prédiction est ensuite affichée sur le map pour identifier le type du crime commis.

Comme perspectives, nous envisageons améliorer notre modèle et introduire de nouvelles fonctionnalités en temps réel.

BIBLIOGRAPHIE

[1] PredPol. Consulté à l'adresse https://www.predpol.com/

[2] Palantir. Consulté à l'adresse https://www.palantir.com/

[3] New York City Police Department (NYPD). 2020. NYPD Complaint Data Historic. Consulté à l'adresse https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i

[4] Fei Yang. 2019. Predictive Policing, Oxford Research Encyclopedia, Criminology and Criminal Justice, Oxford University Press. (septembre 2019). DOI :https ://doi.org/10.1093/acrefore/9780190264079.013.508

Fig. 13 : Visualisation 3D.

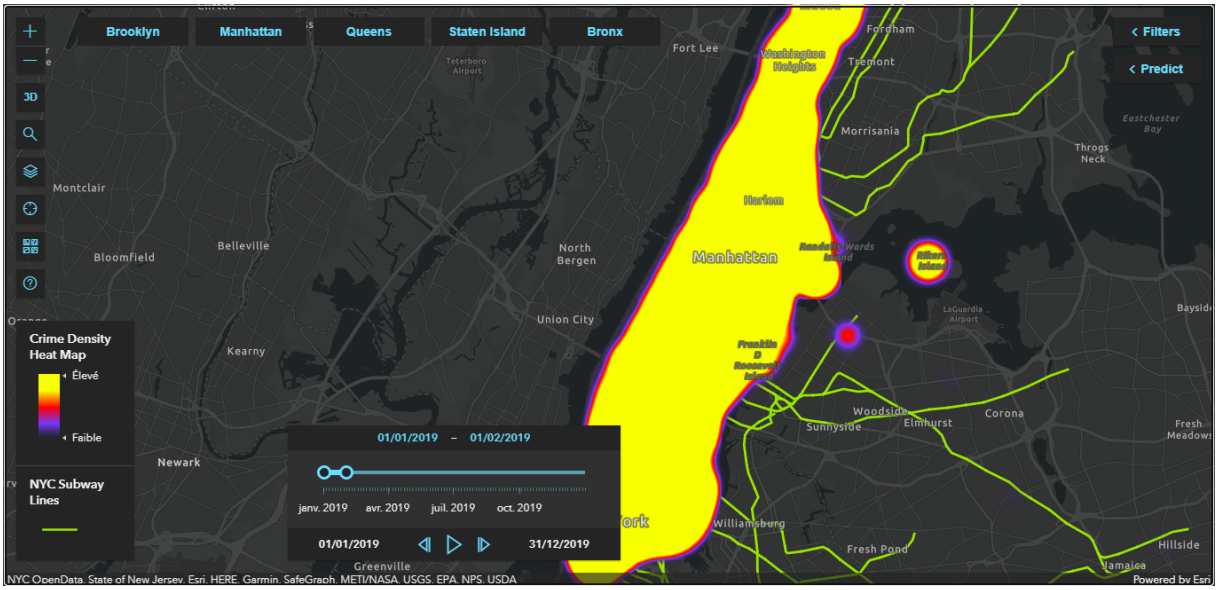

Fig. 14 : Visualisation 2D Globale.

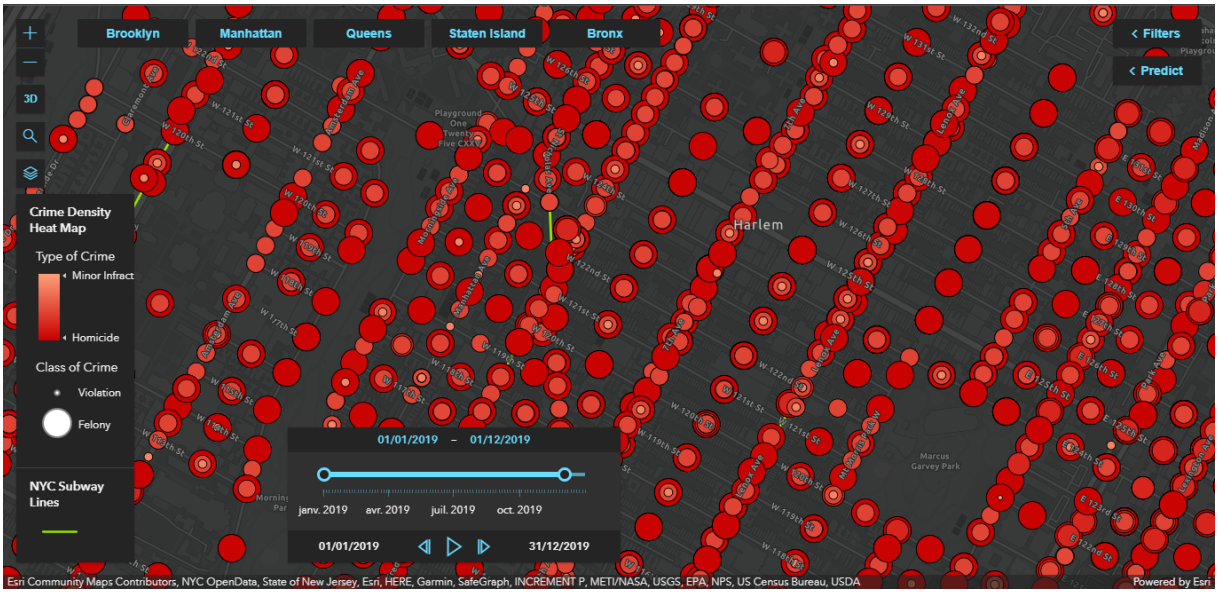

Fig. 15 : Visualisation 2D Détaillé (Arrondissement Manhattan).

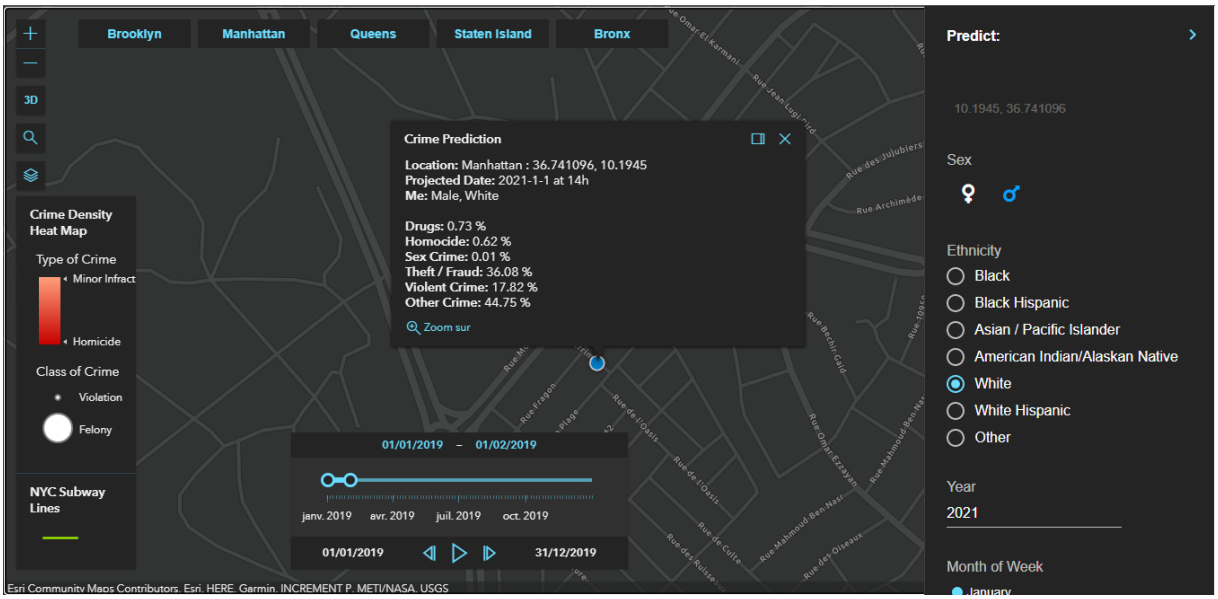

Fig. 16 : Résultat de la Prédiction.