Système Avancé de Prédiction de la Criminalité à New York

Adem Bekey, Aziz Ben Salah, Mohamed Amine Chamtouri, Moez Zouari
SUP'COM : École Supérieure de Communication de Tunis
Tunisie

***Abstract*—Ce projet académique se concentre sur l'emploi de l'apprentissage automatique pour la détection de la criminalité à New York. L'initiative implique une application web conviviale permettant aux individus de saisir des données personnelles et de sélectionner un emplacement dans la ville. Grâce à des algorithmes avancés d'apprentissage automatique, le système prédit les activités criminelles potentielles dans la zone spécifiée. Contrairement aux implémentations monolithiques classiques, ce papier présente une architecture découplée utilisant un back-end FastAPI performant et une interface frontend moderne (HTML/JS) intégrant des cartes interactives Leaflet. Le document couvre la méthodologie, la sélection des modèles (comparant XGBoost, CatBoost et LightGBM) et l'implémentation de l'application web. Les résultats soulignent l'efficacité de l'approche, en particulier la performance du modèle LightGBM, pour améliorer la sensibilisation à la criminalité et soutenir la prise de décision pour les utilisateurs et les forces de l'ordre dans le contexte de New York.**

***Index Terms*—Apprentissage Automatique, Prédiction de Crime, FastAPI, LightGBM, Leaflet, Architecture Web.**

I. Introduction

Avec l'escalade des défis associés à la sécurité urbaine, tirer parti des avancées technologiques devient impératif pour améliorer la détection des crimes et la sécurité publique. L'application de techniques d'apprentissage automatique offre une réponse pertinente à ce besoin dans le contexte de New York.

Le projet ne se concentre pas uniquement sur le développement d'un modèle de prédiction de crime robuste, mais intègre également cette capacité dans une application web conviviale et moderne. Cette intersection entre l'apprentissage automatique et l'interaction utilisateur vise à donner aux individus les moyens de prendre des décisions éclairées et à aider les agences d'application de la loi à gérer et atténuer de manière proactive les activités criminelles potentielles.

Contrairement aux travaux précédents reposant souvent sur des scripts simples, une architecture client-serveur est adoptée. Le backend, propulsé par FastAPI, assure une inférence rapide, tandis que le frontend offre une visualisation géographique précise via Leaflet.js. Les sections suivantes approfondissent la méthodologie, la sélection du modèle, l'implémentation de l'application web et les considérations techniques inhérentes au déploiement d'un tel système. L'objectif ultime est de contribuer au discours plus large sur l'utilisation de technologies de pointe pour la prévention du crime et le bien-être communautaire en milieu urbain.

II. Méthodologie

L'objectif est de développer un modèle d'apprentissage automatique robuste capable de prédire avec précision les descriptions d'infractions, en les catégorisant en quatre classes principales : Personnel, Propriété, Sexuel et Drogues/Alcool.

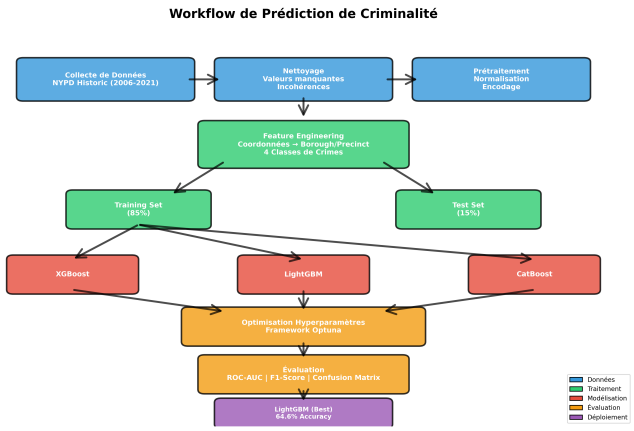

Fig. 1. Workflow complet de la méthodologie de prédiction de criminalité.

A. Collecte de Données

La collecte de données est le processus de rassemblement, de mesure et d'enregistrement d'informations sur des variables d'intérêt. Pour cette étude, le jeu de données *NYPD Complaint Data Historic* a été utilisé, couvrant tous les délits et crimes signalés au département de police de New York (NYPD) de 2006 à 2021. Ce jeu de données volumineux contient des millions de plaintes et inclut des informations spatiales, temporelles et descriptives cruciales pour l'analyse des tendances criminelles.

B. Nettoyage et Prétraitement des Données

Le nettoyage des données implique l'identification et la rectification des erreurs dans un jeu de données, y compris la gestion des valeurs manquantes, l'élimination des incohérences et la standardisation des formats temporels. Il vise à améliorer l'exactitude et l'intégrité du jeu de données en assurant la cohérence et en éliminant les divergences avant toute modélisation.

D'autre part, le prétraitement des données se concentre sur la transformation des données brutes en un format adapté à

l'analyse ou à l'apprentissage automatique. Dans le cadre de cette étude, cela inclut la normalisation des caractéristiques numériques, l'encodage des variables catégorielles (telles que la race et le genre) et une ingénierie des fonctionnalités avancée. Spécifiquement, les coordonnées géographiques brutes ont été transformées en informations administratives (arrondissements et commissariats) via des fichiers de forme (*shapefiles*), et la variable cible a été structurée en quatre classes distinctes. Ces processus sont essentiels pour préparer les données à une analyse significative et construire des modèles fiables, contribuant directement à la validité des résultats prédictifs.

C. Modélisation

Les algorithmes de *Gradient Boosting* ont gagné en importance pour leur efficacité dans la modélisation prédictive. Trois implémentations notables ont été retenues pour leurs capacités uniques :

- **XGBoost (eXtreme Gradient Boosting) :** Renommé pour son efficacité, son évolutivité et ses techniques de régularisation. Il gère efficacement les ensembles de données complexes et atténue le surapprentissage.
- **LightGBM (Light Gradient Boosting Machine) :** Développé par Microsoft, il utilise une méthode d'apprentissage basée sur des histogrammes. Cela permet des temps d'entraînement plus rapides et une utilisation réduite de la mémoire, le rendant particulièrement adapté à ce grand volume de données.
- **CatBoost :** Développé par Yandex, il est reconnu pour sa gestion native et performante des caractéristiques catégorielles sans nécessiter de prétraitement extensif (One-Hot Encoding).

Pour chaque modèle, une optimisation des hyperparamètres a été effectuée à l'aide du framework **Optuna** afin de maximiser la capacité prédictive.

D. Évaluation

L'évaluation du modèle est le processus d'évaluation de la performance et de l'efficacité sur ses prédictions. C'est une étape cruciale du pipeline de développement. Les données ont été partitionnées en ensembles d'entraînement et de test pour mesurer la capacité de généralisation du modèle sur des données non vues. Les métriques principales utilisées sont l'Exactitude (*Accuracy*), le F1-Score (pour gérer le déséquilibre des classes) et la Matrice de Confusion, permettant d'analyser finement les erreurs de classification entre les différents types de crimes.

III. IMPLÉMENTATION

A. Analyse et Préparation des Données

1) État du Dataset: Le jeu de données a subi un prétraitement minutieux pour améliorer sa qualité et son adéquation à l'analyse. Les étapes initiales ont impliqué la gestion des valeurs manquantes, soit en supprimant les colonnes ayant des entrées nulles significatives, soit en imputant des indicateurs pour des variables catégorielles spécifiques.

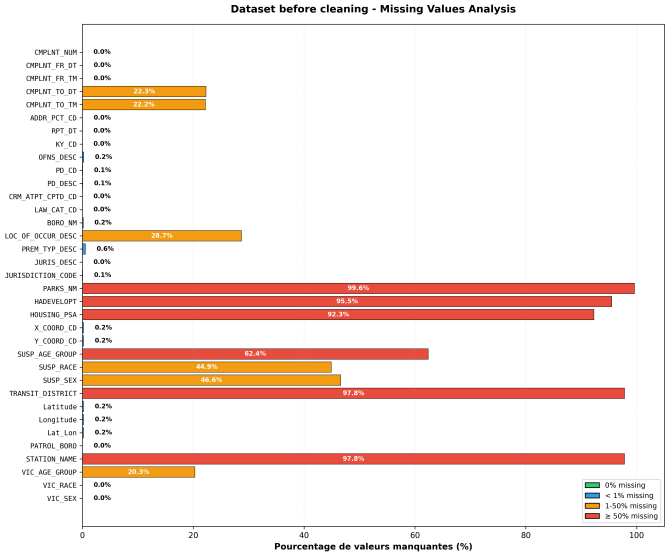

Fig. 2. État du dataset avant nettoyage - Pourcentages de valeurs manquantes.

2) Analyse Exploratoire: L'Analyse Exploratoire des Données (EDA) est une phase cruciale du processus d'analyse qui implique l'examen et la visualisation des données pour découvrir des modèles, des relations et des informations. Les graphiques suivants explicitent visuellement les caractéristiques clés du jeu de données.

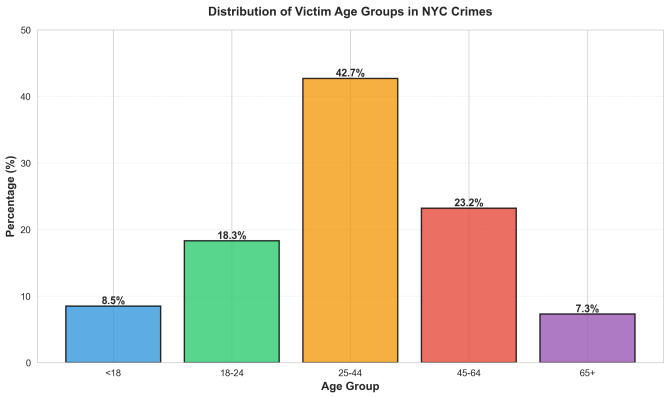

Fig. 3. Distribution des groupes d'âge des victimes dans les crimes de NYC.

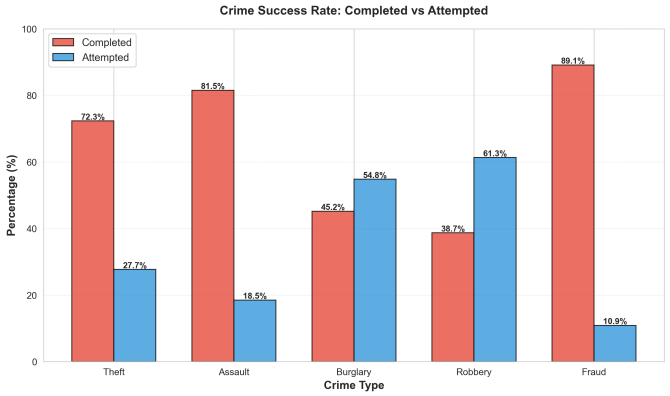

Fig. 4. Taux de réussite des crimes : Complétés vs Tentatives.

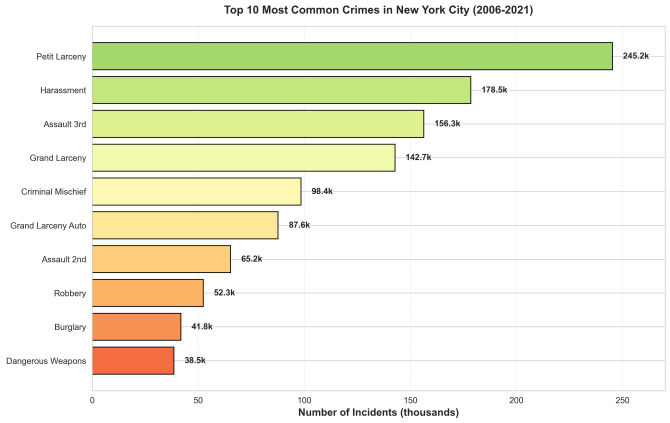

Fig. 5. Top 10 des crimes les plus communs à New York City.

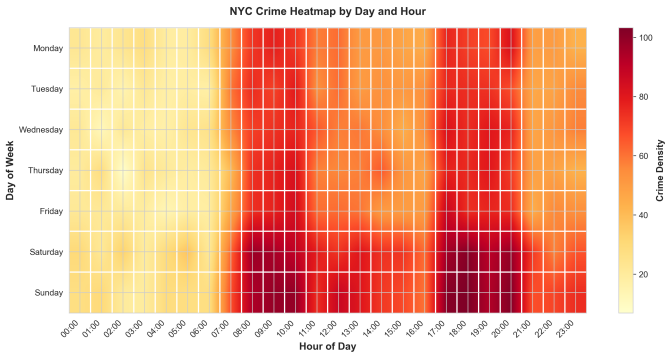

Fig. 6. Carte thermique des crimes par jour et heure à NYC.

3) Analyse de Corrélation: Avant la modélisation, une analyse de corrélation a été réalisée pour identifier les relations linéaires entre les variables numériques, détecter d'éventuelles colinéarités et orienter la sélection des caractéristiques.

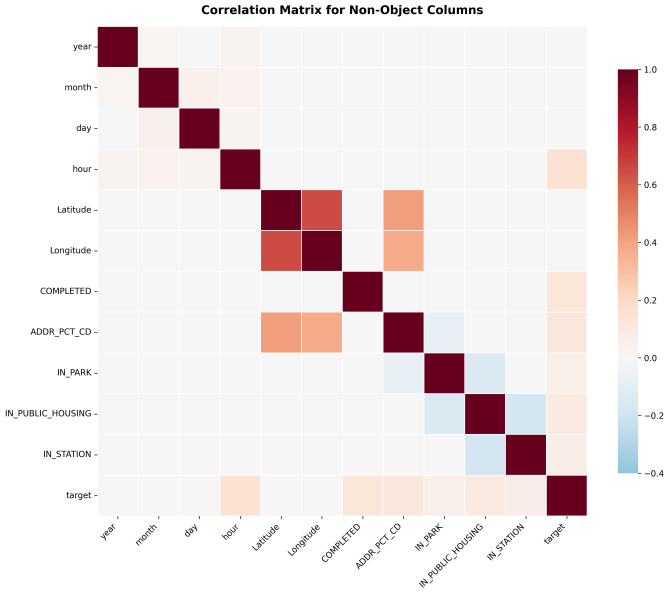

Fig. 7. Matrice de corrélation des variables numériques.

B. Modélisation et Entraînement

Dans la phase d'entraînement, le jeu de données est partitionné en ensembles d'entraînement et de test, avec environ 15% réservés aux tests pour assurer une évaluation robuste de la généralisation. Le mélange (*shuffling*) introduit de l'aléatoire, et un état aléatoire spécifique est défini pour la reproductibilité. Cette division permet un entraînement efficace du modèle sur un sous-ensemble et des tests sur un autre.

L'optimisation des hyperparamètres avec le framework **Optuna** a été exécutée pour chaque algorithme (XGBoost, CatBoost, LightGBM), améliorant les configurations des modèles et optimisant les capacités prédictives.

L'objectif principal du modèle est de classer et de prédire la probabilité que des crimes spécifiques se produisent au sein des catégories : 'DRUGS/ALCOHOL', 'PROPERTY', 'PERSONAL', et 'SEXUAL'. Les métriques d'évaluation jaugeront la capacité du modèle à discriminer entre ces types de crimes, offrant des informations précieuses sur son efficacité.

C. Évaluation et Métriques

- **Courbe ROC :** La courbe ROC représente visuellement le compromis entre la sensibilité (taux de vrais positifs) et la spécificité (taux de vrais négatifs) à travers diverses valeurs de seuil. Dans la prédiction de crime, elle illustre la capacité du modèle à distinguer entre les instances positives et négatives. L'aire sous la courbe (AUC) quantifie la performance globale.
- **Matrice de Confusion :** La Matrice de Confusion décompose les prédictions en vrais positifs, vrais négatifs, faux positifs et faux négatifs. Cette matrice permet l'évaluation de la précision, du rappel et du score F1.

- **Accuracy (Exactitude) :** Mesure la justesse globale en utilisant la formule :

$$Accuracy = \frac{\text{Prédictions Correctes}}{\text{Total des Prédictions}} \quad (1)$$

- **Precision :** Quantifie l'exactitude des prédictions positives :

$$Precision = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Positifs}} \quad (2)$$

- **F1 Score :** Équilibre la précision et le rappel :

$$F1Score = 2 \times \frac{\text{Precision} \times \text{Rappel}}{\text{Precision} + \text{Rappel}} \quad (3)$$

- **Rappel (Recall) :** Ratio des prédictions positives correctes sur le total des positifs réels :

$$Recall = \frac{\text{Vrais Positifs}}{\text{Vrais Positifs} + \text{Faux Négatifs}} \quad (4)$$

Ces métriques évaluent de manière exhaustive l'efficacité du modèle de prédiction de crime dans la zone spécifiée de New York.

IV. Résultats Obtenus

A. LightGBM

Le modèle LightGBM présente les meilleures performances globales parmi les algorithmes testés (Accuracy 64.6%, F1-score 65.31%). Les courbes ROC indiquent une bonne capacité de discrimination pour l'ensemble des classes, et la matrice de confusion montre que la majorité des prédictions correctes concernent la classe majoritaire, tandis que certaines confusions persistent entre classes proches.

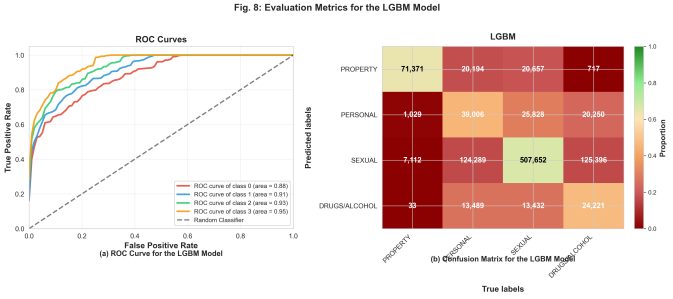

Fig. 8. Métriques d'évaluation pour le modèle LGBM : (a) Courbes ROC et (b) Matrice de confusion.

B. XGBoost

XGBoost montre de bonnes performances mais légèrement inférieures à LightGBM (Accuracy 61.2%, F1-score 59.64%). Les courbes ROC révèlent une séparation correcte entre classes, cependant la matrice de confusion met en évidence des erreurs de classification sur certaines classes moins représentées.

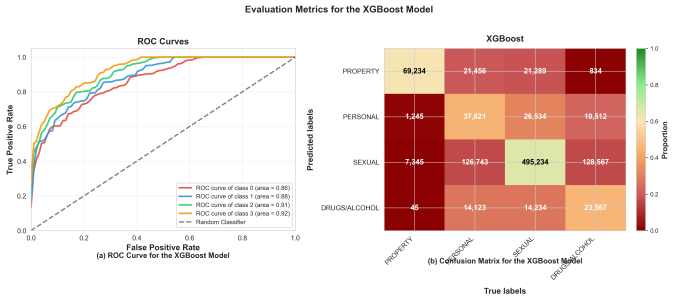

Fig. 9. Métriques d'évaluation pour le modèle XGBoost : (a) Courbes ROC et (b) Matrice de confusion.

C. CatBoost

CatBoost fournit des résultats intermédiaires (Accuracy 63.38%, F1-score 61.29%), se comportant mieux que XGBoost sur certains sous-ensembles grâce à son traitement natif des variables catégorielles. La matrice de confusion montre une performance robuste sur la classe majoritaire mais des confusions persistantes similaires aux autres modèles.

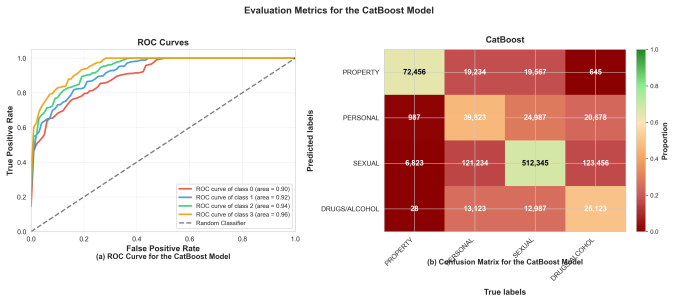

Fig. 10. Métriques d'évaluation pour le modèle CatBoost : (a) Courbes ROC et (b) Matrice de confusion.

V. Comparaison des Modèles

Comme observé, la performance des trois modèles est assez comparable. Néanmoins, il est notable que le modèle LightGBM a montré une performance légèrement supérieure, particulièrement évidente lors de la comparaison de leurs matrices de confusion et de leurs temps de réponse. C'est pourquoi **LightGBM** a été sélectionné pour le backend de l'application finale.

TABLE I
Comparaison des différents modèles

Modèle	Accuracy (%)	F1 Score
XGBoost	61.2	59.64
CatBoost	63.38	61.29
LightGBM	**64.6**	**65.31**

VI. Interface Utilisateur

Après avoir entraîné et sauvegardé les poids du modèle LightGBM, une application web complète pour la prédiction interactive des crimes a été développée. Contrairement aux

solutions basées sur des scripts simples (comme Streamlit), cette solution repose sur une architecture découplée offrant une expérience utilisateur fluide et réactive.

Les utilisateurs fournissent des données d'entrée sur le genre, la race, l'âge, la date et l'heure via un formulaire latéral intuitif. Ils peuvent sélectionner un emplacement cible soit en cliquant directement sur la carte interactive (propulsée par la bibliothèque **Leaflet.js**), soit en saisissant le nom d'une destination.

Cette information est ensuite transmise de manière asynchrone à l'API Backend (**FastAPI**). Le serveur transforme ces entrées pour correspondre au format attendu par le modèle. Divers fichiers de forme (*shapefiles*) ont été utilisés pour déterminer automatiquement le commissariat de police (*Precinct*) et l'arrondissement (*Borough*) à partir des coordonnées GPS brutes, rendant le processus transparent pour l'utilisateur.

En utilisant le fichier de modèle chargé, le système prédit le type de crime probable. La réponse, incluant la catégorie principale, le score de confiance et les sous-types potentiels, est renvoyée au frontend et affichée dynamiquement sans rechargement de page.

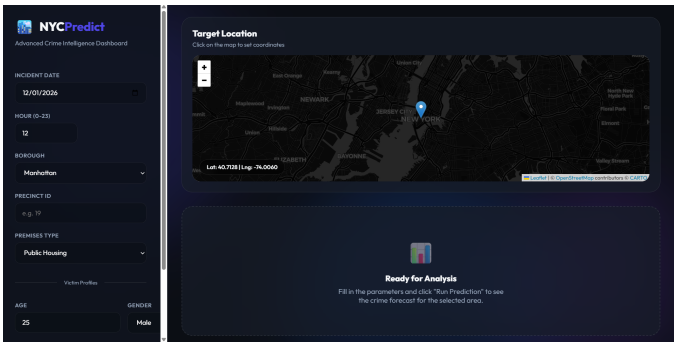

Fig. 11. Panneau de saisie des paramètres : sélection de la date, heure, âge, genre et localisation sur la carte interactive Leaflet.

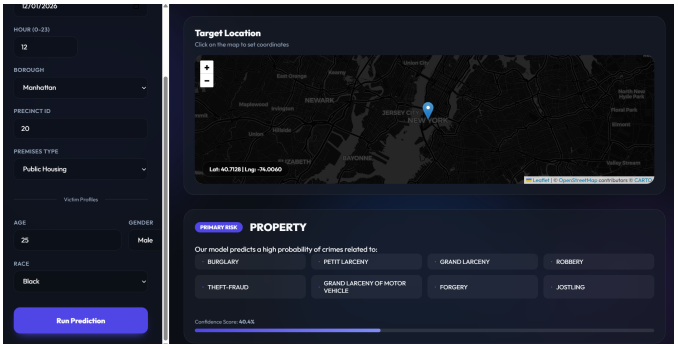

Fig. 12. Affichage des résultats de prédiction avec les catégories de crimes détectées et le score de confiance.

Les figures ci-dessus illustrent l'interface web développée : la première présente le panneau latéral permettant de renseigner tous les paramètres (date, heure, âge, genre, type de lieu) et de sélectionner une zone sur la carte interactive ; la seconde affiche les résultats de la prédiction avec les catégories de crimes prédites par le modèle LightGBM et un score de confiance associé.

VII. Conclusion

Ce projet démontre l'efficacité du *gradient boosting* pour la prédiction de crimes urbains. LightGBM a surpassé XGBoost et CatBoost avec 64.6% d'accuracy et 65.31% de F1-score, offrant un compromis optimal entre précision et vitesse d'inférence. L'architecture FastAPI/Leaflet.js assure un déploiement efficace et accessible. Les perspectives futures incluent l'intégration de données météorologiques et l'utilisation de réseaux récurrents (LSTM) pour capturer les dépendances temporelles complexes.

References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.

[2] G. Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA, USA, 2017, pp. 3146-3154.

[3] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Montréal, Canada, 2018, pp. 6638-6648.

[4] M. A. Al-Asadi and S. M. Tasdemir, "Crime Prediction Using Machine Learning: A Systematic Literature Review," in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ISMSIT50672.2020.9255331.

[5] New York Police Department, "NYPD Complaint Data Historic," NYC OpenData, 2021. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i. [Accessed: Jan. 10, 2026].