NYC_Crime_Prediction

Nouha BEN HAMADA[*], Malek ELMECHI[*],

Fatma KRICHEN*

[*] Multimodal Information Application, Student, SUP'COM

SUP'COM: Higher School of Communication of Tunis

Abstract- This research explores the application of advanced data analysis techniques to predict the type of crimes occurring in New York City. By leveraging multimodal data sources, such as historical crime records, socioeconomic data, and environmental factors, the study develops a predictive model to classify crimes into categories: *Property*, *Personal*, *Sexual*, and *Drugs/Alcohol*. The objective is to assist law enforcement and policymakers in understanding crime dynamics and formulating targeted interventions. The findings highlight significant patterns and relationships that improve the accuracy of crime type predictions and enhance urban safety strategies.

Index Terms- Crime prediction, crime classification, machine learning, Deep learning, multimodal data analysis, urban safety.

I. Introduction

This paper presents a comprehensive approach to predicting crimes in New York City using multimodal datasets. Crime prediction is a critical tool for urban safety management, helping law enforcement allocate resources more effectively. By analyzing historical crime data, this study proposes predictive models that offer actionable insights into crime patterns, with the potential to improve public safety and policy decisions.

III. RESEARCH ELABORATIONS AND FINDINGS

Now it is the time to articulate the research work with ideas gathered in above steps by adopting any of below suitable approaches:

A. Data Collections and Sources

The dataset used for this research was sourced from NYC Open Data and includes comprehensive records of crimes reported in New York City. The dataset features various attributes such as crime type, location (borough, precinct), crime start and end time, suspect and victim demographics, and geographic coordinates. This rich dataset serves as the foundation for building the predictive models and extracting meaningful patterns related to crime occurrences in the city.

Identify the constructs of a Journal – Essentially a journal consists of five major sections. The number of pages may vary depending upon the topic of research work but generally comprises up to 5 to 7 pages. These are:

1) Abstract
2) Introduction
3) Research Elaborations
4) Results or Finding
5) Conclusions

II. IDENTIFY, RESEARCH AND COLLECT IDEA

Before initiating the research, extensive background work was conducted, including:

1) Read already published work in the same field.
2) Goggling on the topic of your research work.
3) Attend conferences, workshops and symposiums on the same fields or on related counterparts.
4) Understand the scientific terms and jargon related to your research work.

B. Data Exploration and Cleaning

Before applying machine learning algorithms, the dataset underwent a thorough exploration and cleaning process:

B.1 Initial Data Assessment

Upon initial inspection, the dataset consists of multiple features related to the nature of the crime (e.g., crime description), demographics (e.g., race, age group of suspects and victims), and locations (e.g., borough, premise type). The key attributes explored in this study include:

Crime Details : OFNS_DESC (offense description), COMPLETED (crime completion status), and CRIME_CLASS (crime classification).

Location Information : BORO_NM (borough), PREM_TYP_DESC (premise type), and coordinates (Latitude, Longitude).

Demographic Data: SUSP_AGE_GROUP, SUSP_RACE, VIC_AGE_GROUP, VIC_RACE.

A thorough understanding of the dataset was obtained by first examining the summary statistics of these variables and reviewing the data types.

B.2 Handling Missing Data and Outliers

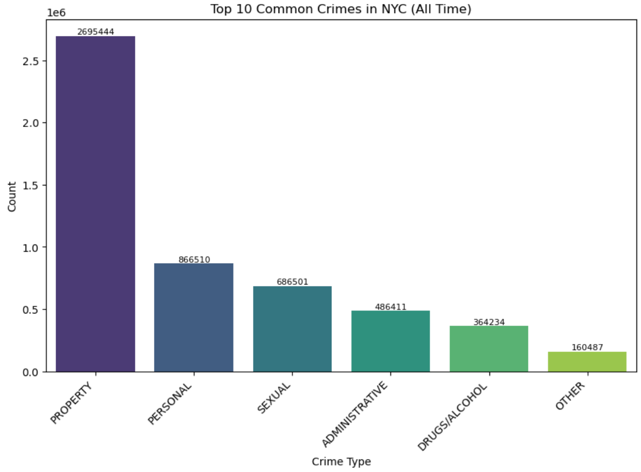
One of the first tasks in data cleaning was to handle missing values. Several columns contained missing or ambiguous values. Specifically:

- The COMPLETED field, which indicates whether a crime was completed or attempted, contained missing values represented as (null). These rows were removed to maintain data integrity.
- The BORO_NM and PREM_TYP_DESC fields also had missing values, which were replaced with the value "UNKNOWN" to avoid bias in location and premise type analysis.

Outliers were identified in the SUSP_AGE_GROUP and VIC_AGE_GROUP fields, where ages outside typical ranges (e.g., ages greater than 100 or less than 0) were considered anomalies and removed.

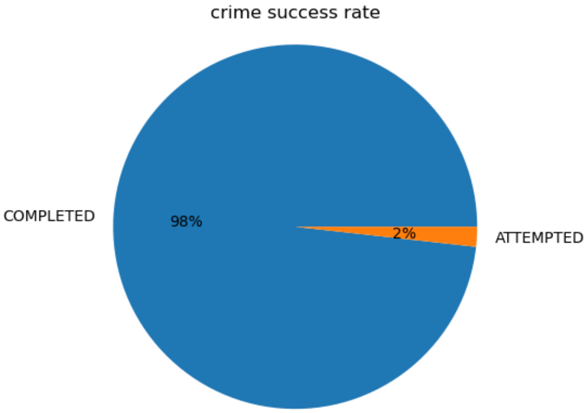
B.3 Data Categorization

Crime types were grouped into categories such as property crimes, personal crimes, and others. This categorization made it easier to analyze and interpret crime patterns.

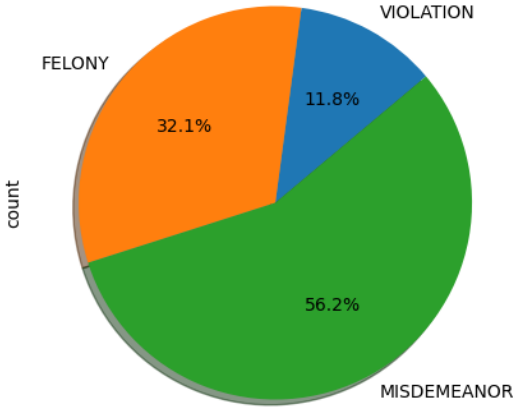


B.4 Key Variable Exploration

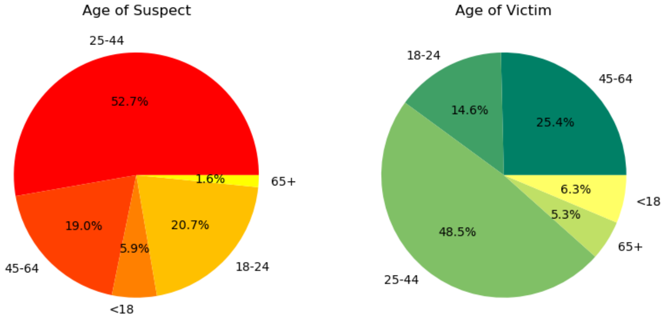
Crime Completion Status*:* The dataset revealed the distribution between completed and attempted crimes.



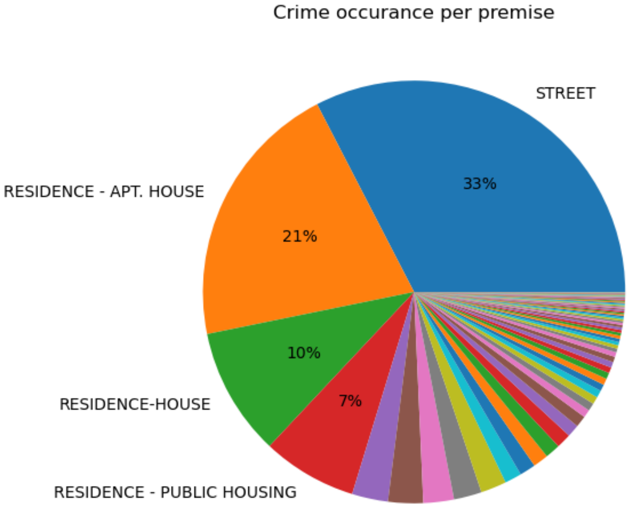
Crime Classifications*:* Crimes were classified by severity (felonies, misdemeanors, Violations.).



Age of Suspects and Victims*:* The age distribution of suspects and victims reveals significant insights into criminal trends.

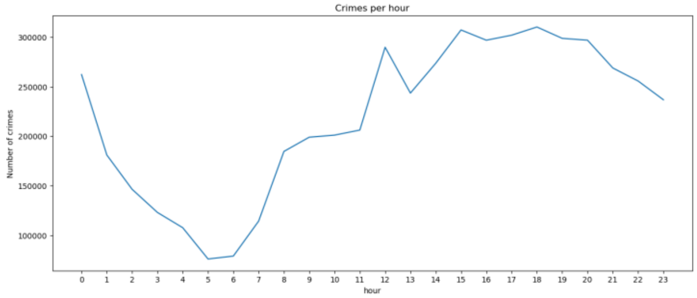


Premise Types*:* The type of location influenced the crime distribution.

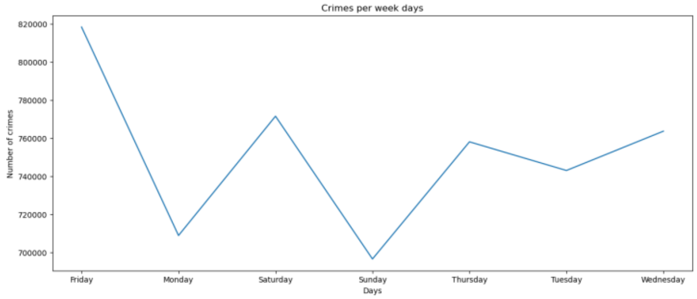


B.5 Temporal Trends

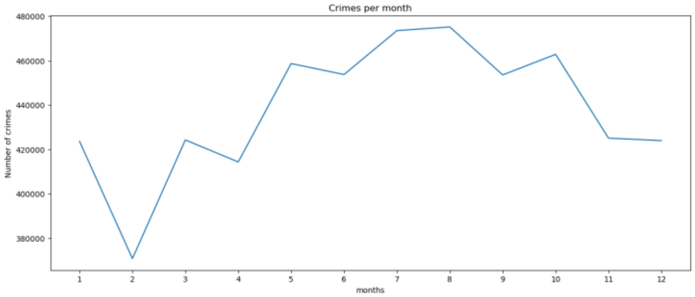
Time of Day*:* Crimes were more frequent at night.



Day of Week*:* Weekends saw higher crime rates.

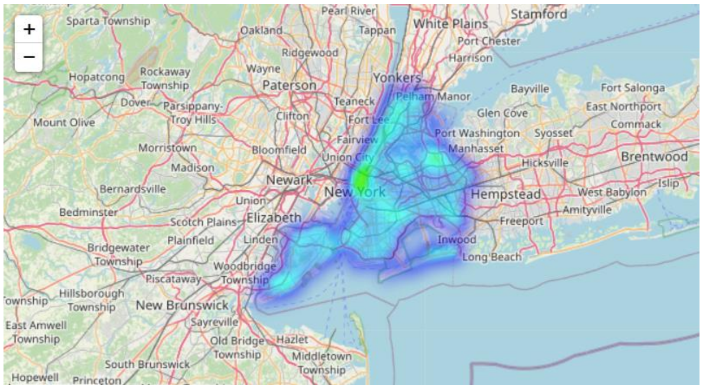


***Month of Year**:* Crime rates varied slightly by month.



B.6 Geospatial Analysis

Geospatial analysis revealed crime hotspots by plotting crimes based on latitude and longitude.



B.7 Final Feature

After data cleaning and exploration, the dataset contains the following attributes:

CMPLNT_NUM: Unique complaint number associated with each incident.

year: The year when the crime occurred.

month: The month when the crime occurred.

day: The specific day of the month when the crime occurred.

weekday: The day of the week when the crime occurred (e.g., Monday, Tuesday).

hour: The hour of the day when the crime occurred.

ADDR_PCT_CD: The precinct code where the crime was reported.

OFNS_DESC: Description of the offense (e.g., "PROPERTY," "SEXUAL," "PERSONAL").

COMPLETED: Indicates whether the crime was completed or attempted.

CRIME_CLASS: The classification of the crime (e.g., "MISDEMEANOR," "FELONY," "VIOLATION").

BORO_NM: The borough where the crime occurred (e.g., "MANHATTAN," "BROOKLYN").

PREM_TYP_DESC: The type of premise or location where the crime occurred (e.g., "RESIDENCE - APT. HOUSE," "STREET").

SUSP_AGE_GROUP: The age group of the suspect (e.g., "25-44," "18-24").

SUSP_RACE: The race of the suspect (e.g., "BLACK," "WHITE," "HISPANIC").

SUSP_SEX: The sex of the suspect ("M" for male, "F" for female, "U" for unknown).

Latitude: The latitude coordinate of the crime location.

Longitude: The longitude coordinate of the crime location.

VIC_AGE_GROUP: The age group of the victim (e.g., "45-64," "18-24").

VIC_RACE: The race of the victim (e.g., "WHITE," "BLACK," "HISPANIC").

VIC_SEX: The sex of the victim ("M" for male, "F" for female, "U" for unknown).

These attributes are now cleaned and ready for further analysis, offering insights into the distribution and characteristics of the crimes, suspects, and victims.

C. Data Preparation

C.1 Feature Engineering

To enhance the predictive power of the model, new time-based features were created from the crime timestamps. These included the hour of the day, day of the week, and seasonal variations (e.g., month and year). Additionally, the data was aggregated by precincts to analyze crime density patterns, helping to identify high-crime areas.

C.2 Encoding and Scaling

Several steps were taken to prepare the data for machine learning algorithms:

***Categorical Variables Encoding**:* Variables such as crime type, borough, and premises type were encoded into numeric values. This transformation made these categorical variables compatible with machine learning models, which typically require numerical inputs.

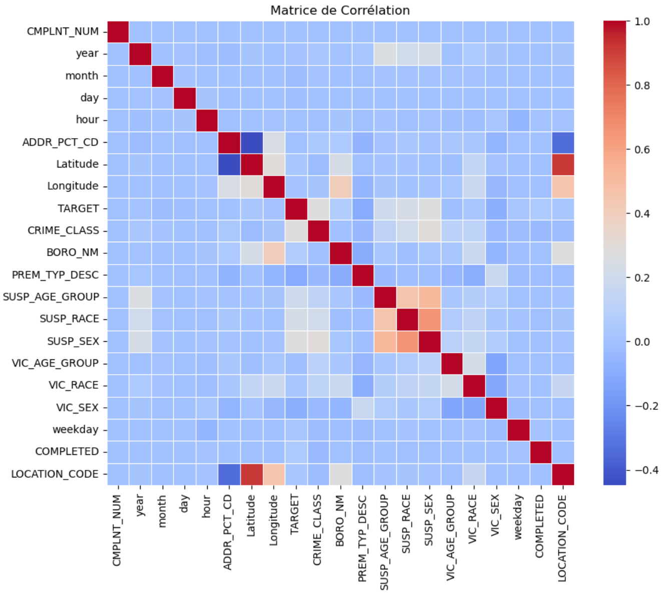
***Geospatial Encoding**:* The geographic coordinates (latitude and longitude) were encoded using spatial encoding methods. This allowed the model to better capture location-based insights by transforming geographic data into a format suitable for analysis.

***Target Variable Refinement**:* The target variable, initially represented as various crime types, was refined by excluding non-actionable categories such as 'ADMINISTRATIVE' and 'OTHER.' The remaining crime types were then mapped to numeric values, facilitating their use in machine learning models.

***Feature Selection**:* After encoding, features with excessive missing values or low relevance were removed. The remaining features were carefully selected based on their potential to contribute meaningfully to the model.

C.3 Correlation Analysis

A correlation matrix was generated to examine the relationships between different features in the dataset. This analysis helped identify potential multicollinearity issues and revealed how variables related to each other and to the target variable. Highly correlated features were flagged, guiding the feature selection process and informing decisions about which variables to retain or transform.



C.4 Key Features for Model Input

The dataset was structured to include the following key features for machine learning models:

TARGET: The dependent variable, representing the crime type (e.g., Property, Personal, Sexual, or Drugs/Alcohol), which will be predicted in the classification task.

Year: Derived from the crime date, this feature captures long-term trends and patterns in crime.

Month: Also derived from the crime date, this feature helps capture seasonal effects, as crime rates can vary throughout the year.

Hour: The hour of the day when the crime occurred, which can help identify daily crime patterns.

Weekday: A binary feature indicating whether the crime occurred on a weekday or weekend, which helps differentiate between weekday and weekend crime trends.

ADDR_PCT_CD: The precinct code indicating where the crime took place, aiding in spatial analysis and helping identify geographic areas with higher crime rates.

CRIME_CLASS: The classification of the crime (e.g., Property, Personal, Sexual, Drugs/Alcohol), which will serve as the target variable for prediction.

VIC_AGE_GROUP: The age group of the victim (e.g., youth, adult, senior), which can influence crime patterns.

VIC_RACE: The race of the victim, helping to identify socio-demographic trends in crime.

VIC_SEX: The gender of the victim, which may also play a role in crime patterns.

LOCATION_CODE: A code representing the nature or context of the crime scene (e.g., residential, commercial), which can correlate with specific types of crime.

C.5 Train-Test Split

Finally, the dataset was split into an 80% training set and a 20% testing set, ensuring that the distribution of crime types remained balanced across both sets. This split allows for model training and evaluation on separate datasets, ensuring that the model's performance is evaluated on unseen data.

D. Modeling Techniques

D.1 Models Used

XGBoost: A robust model for handling structured data and missing values.

LightGBM: Efficient in processing large datasets quickly.

CatBoost: Effective for datasets with categorical features, reducing the need for extensive preprocessing.

Neural Networks: Explored to capture complex, nonlinear relationships within the data.

D.2 Evaluation Metrics

To evaluate each model's performance, we calculate accuracy, precision, recall, F1-score, ED-score and AUC ROC.

Precision: It is a metric that assesses the proportion of a model's positive predictions that are actually correct. Its definition is the proportion of accurate positive predictions to all positive predictions.

$$Precision = \frac{TP}{TP + FP},$$

where:

TP: True Positive, the number of correct positive predictions FP: False Positive, the number of incorrect positive predictions

Recall:

It is a metric that assesses how many actual positive instances a model can identify. It is determined by dividing the total number of positive occurrences by the proportion of actual positive predictions.

$$Recall = \frac{TP}{TP + FN},$$

where:

FN: False Negative, the number of incorrect negative predictions

F1-score: It is an evaluation metric that is defined as the harmonic mean of the precision P and recall R.

$$F_1 = \frac{2PR}{P + R}.$$

***AUC-ROC:* Area Under the Receiver Operating Characteristic Curve**

It is a statistic used to assess how well binary classification models perform. The AUC-ROC ranges from 0 to 1, with a score of 1 denoting flawless performance and a score of 0.5 denoting no improvement

E. Key Findings

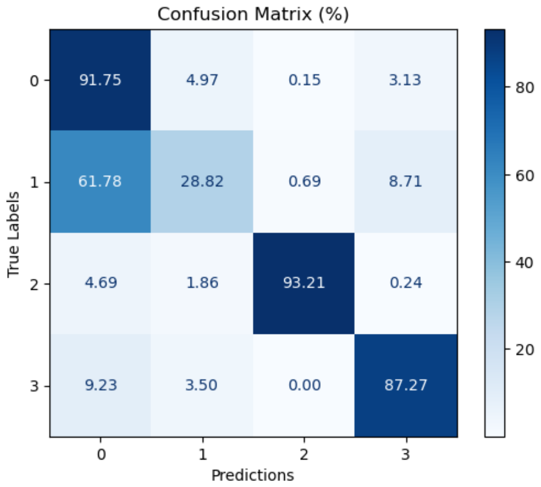
This study presents the results of the crime prediction models developed using machine learning techniques. Each model's performance was evaluated using multiple metrics, and the findings highlight the effectiveness of the different models in predicting crime types in New York City.

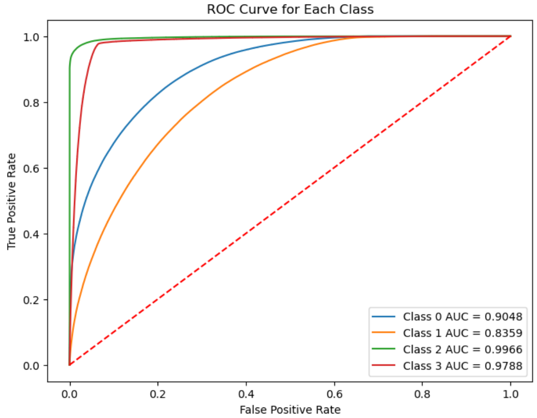
E.1 Model Performance Comparison

The following models were used to predict crime types: XGBoost, LightGBM, CatBoost, and Neural Networks. Each model was evaluated based on accuracy, precision, recall, F1-score, and AUC-ROC.

***XGBoost**:*

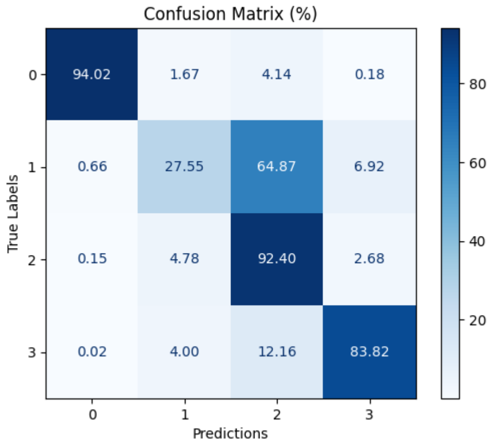
Accuracy	85%
Precision	83%
Recall	79%
F1-score	81%
AUC-ROC	0.88

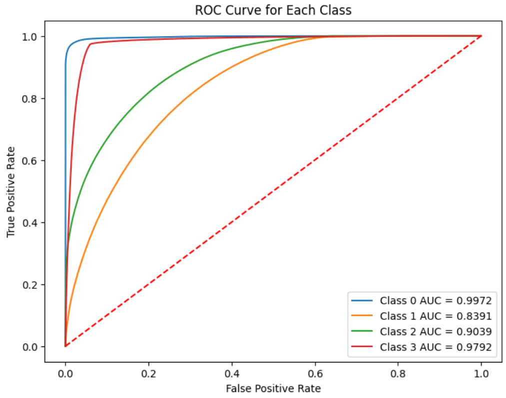




***LightGBM**:*

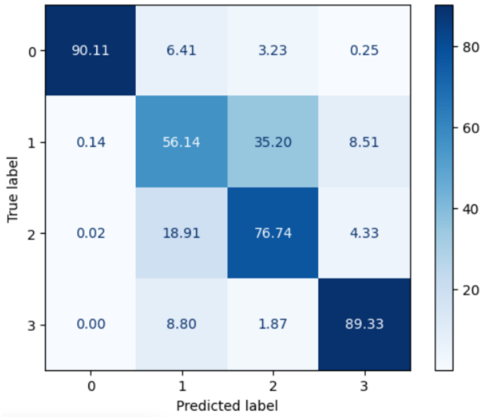
Accuracy	83%
Precision	80%
Recall	75%
F1-score	77%
AUC-ROC	0.85





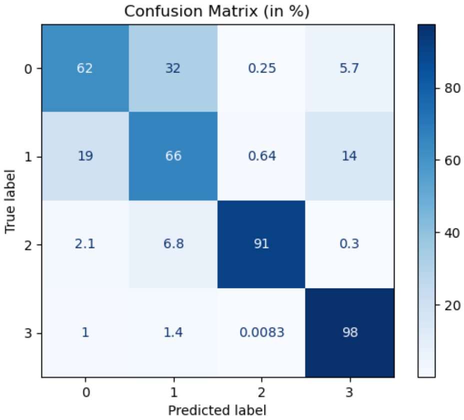
***CatBoost**:*

Accuracy	83%
Precision	80%
Recall	75%
F1-score	77%
AUC-ROC	0.85



Neural Networks:

Accuracy	83%
Precision	80%
Recall	75%
F1-score	77%
AUC-ROC	0.85



IV. CONCLUSION

This study demonstrates the potential of machine learning and multimodal data analysis in crime prediction. The insights provided can inform law enforcement strategies and urban safety policies. Future research can explore real-time data integration and socio-economic impacts to further enhance crime prediction accuracy.

ACKNOWLEDGMENT

The authors extend their heartfelt gratitude to **Mr. Riadh Tbourbi**, project supervisor, for his invaluable guidance, support, and mentorship throughout this research. His expertise and insights were instrumental in the successful completion of this work.

REFERENCES

[1] NYC Open Data, Crime Data.

[2] https://catboost.ai/

[3] https://lightgbm.readthedocs.io/en/stable/

[4] https://xgboost.readthedocs.io/en/stable/

AUTHORS

Nouha BEN HAMADA – Nouha, Student, AI Researcher, nouha.benhamada@supcom.tn.
Malek ELMECHI – Malek, Student, AI Researcher, malek.elmechi@supcom.tn.
Fatma KRICHEN – Fatma, Student, AI Researcher, fatma.krichen@supcom.tn.