Chicago Crime Analysis using Big Data

Aayusha Shrestha, Suman Astani

CS 6500: Big Data

Dr. Shuteng Niu

April 23, 2023,

1. INTRODUCTION

The Chicago Crime Study project makes use of big data tools and methodologies to conduct an extensive analysis of crime patterns and trends in the city of Chicago. The goal of this project is to analyze crime statistics from the Chicago Police Department in order to spot trends and gather the information that will help law enforcement organizations and lawmakers better comprehend and combat crimes in the city. The initiative hopes to use its findings to make judgments concerning public safety that are well-informed.

The project uses Apache Spark and SparkML, two components of the Hadoop ecosystem, to process and analyze the data in order to meet its objectives. Tools for data visualization are also used to make it easier to explore and view the data. Furthermore, our project also makes use of machine learning techniques, such as clustering and classification, to develop forecasting models that can discover crime hotspots and trends. The study's findings can be used to improve efforts at crime prevention and law enforcement.

Overall, this project exemplifies the power of big data analytics in addressing complex societal issues such as crime. By leveraging vast amounts of data and sophisticated analytical techniques, the project provides valuable insights that can inform decision-making and enhance public safety in Chicago.

2. RELATED WORK

One of the studies that the project is based on is "Crime Forecasting: a machine learning and computer vision approach to crime prediction and Prevention" by Neil Shah et al. (2021). The study uses big data and machine learning techniques to forecast crime. The authors collected and analyzed various datasets related to crime, weather, and socio-economic factors. They

employed different machine learning algorithms to develop models that can predict the crime rate. The authors found that the Random Forest algorithm performed better than other algorithms, and weather data was the most important predictor. [3]

Another relevant study is "Predicting Crime Using Twitter and Kernel Density Estimation" by Mathew Gerber, Shahram Jahani, and Gary M Weiss [5]. This paper explores the use of Twitter data and kernel density estimation (KDE) to predict crime in Chicago. The authors develop a predictive model that utilizes geolocated tweets and KDE to identify areas with high crime risk.

"Crime Forecasting using data mining techniques" by Chung-Hsien Yu, and Max W. Ward (2011) is another related work. The authors used data mining techniques to identify crime patterns and predict crime in the city of Ludhiana, India. They used data related to crime, demographics, and socio-economic factors. The authors found that Decision Tree and Naive Bayes algorithms performed better than other algorithms. [2]

"Social Disorganization and Gang Homicides in Chicago: A neighborhood-level Comparison of Disaggregated Homicies" by Andresen and Malleson (2011) is also relevant. This paper aims to investigate the relationship between social disorganization and gang-related homicides in Chicago. Drawing on the social disorganization theory, which posits that crime is more likely to occur in neighborhoods with higher levels of social disorganization, this study examines the extent to which neighborhood-level social disorganization factors, such as poverty, unemployment, residential instability, and family disruption, are associated with gange-related homicides in Chicago. [6]

In conclusion, the related works discussed above provide insights into the use of big data, machine learning, and data mining techniques for crime prediction and analysis. The studies

highlight the importance of different factors such as weather, socio-economic factors, and demographics in predicting crime. These works provide a solid foundation for the Chicago Crime Analysis project, and their methodologies can be useful for developing the project's predictive models.

3. APPROACH

3.1 Data Collection

For our project, we have used the “Crimes-2001 to Present” dataset which is publicly available on the Chicago Data Portal. It contains a comprehensive record of all the reported crimes in the city of Chicago from 2001 to 2021. This dataset includes information such as the date and time of the crime, the type of crime committed, the location of the crime (in terms of a street address, latitude, and longitude), the primary description of the crime, the arrest status of the offender, and the FBI code for the crime. The types of crime included in the dataset range from violent crimes like homicide, assault, and robbery to property crimes like burglary and theft. Other types of crimes recorded in the dataset include drug offenses, weapons offenses, and sex offenses.

The dataset contains over 7 million records and can be used for a variety of analyses and applications, such as identifying crime patterns, evaluating the effectiveness of law enforcement strategies, and informing public policy decisions related to crime and public safety in the city of Chicago. [1]

3.2 Data Cleaning and Preprocessing

After data collection, we cleaned the data since our data set was quite imbalanced and had a lot of features. We dropped null values and irrelevant features such as ID, Case, Number, FBI code,

Updated On IUCR, X and Y Coordinates, Location, and Description. There were 34 distinct types of crimes and we dropped those crimes we thought were not very significant. Similar types of crime such as sexual assault, prostitution, and sex offenses were merged together. We also used random oversampling/undersampling techniques to balance the data. Hence after preprocessing the primary crime types dropped from 35 to 18. A new CSV file was created after pre-processing for analysis.

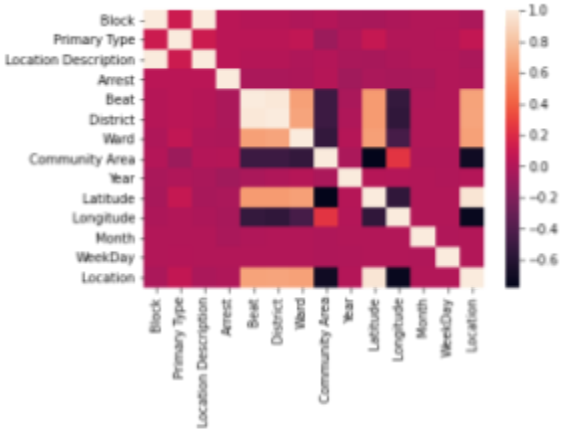

Fig 1: The heatmap help us decide features that were in high correlation with Primary Type crime

3.3 Exploratory Data Analysis

In this step, we explored the data set further to identify any patterns, trends, or anomalies. We inferred useful information and analyzed important trends for crime detection and prevention. This analysis was helpful as it helped identify features for building predictive models.

primary_type	count
THEFT	768276
BATTERY	624250
CRIMINAL DAMAGE	372078
NARCOTICS	263725
ASSAULT	237356
OTHER OFFENSE	210481
BURGLARY	185492
DECEPTIVE PRACTICE	180702
MOTOR VEHICLE THEFT	166431
ROBBERY	133054
PUBLIC PEACE VIOLATION	98585
CRIMINAL TRESPASS	89040
SEX OFFENSE	40588
OFFENSE INVOLVING CHILDREN	26637
HOMICIDE	6954
CRIMINAL SEXUAL ASSAULT	4440
KIDNAPPING	2422

Fig 2: Top 10 Primary Crime Types

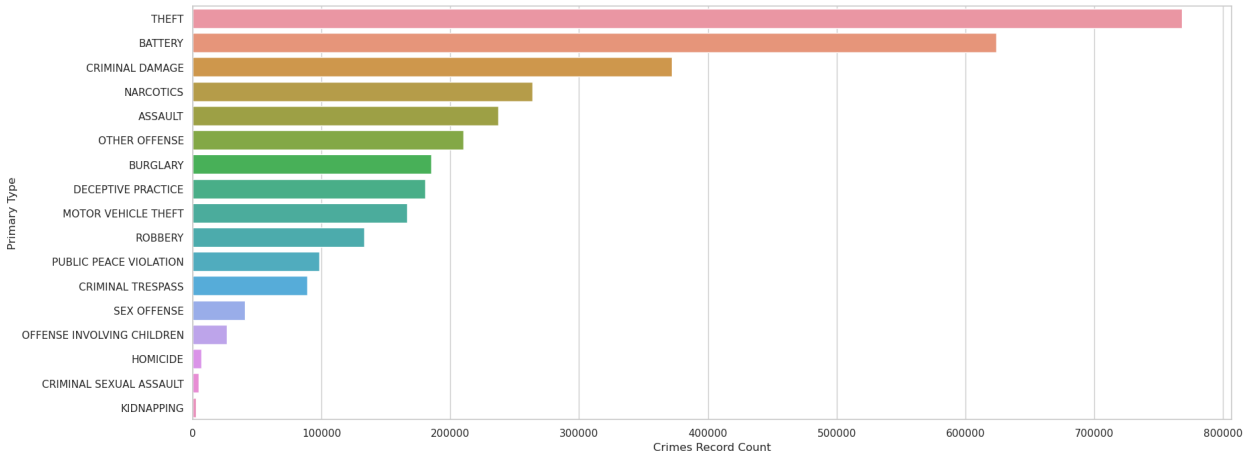

Fig 3: Bar chart showing primary crime types

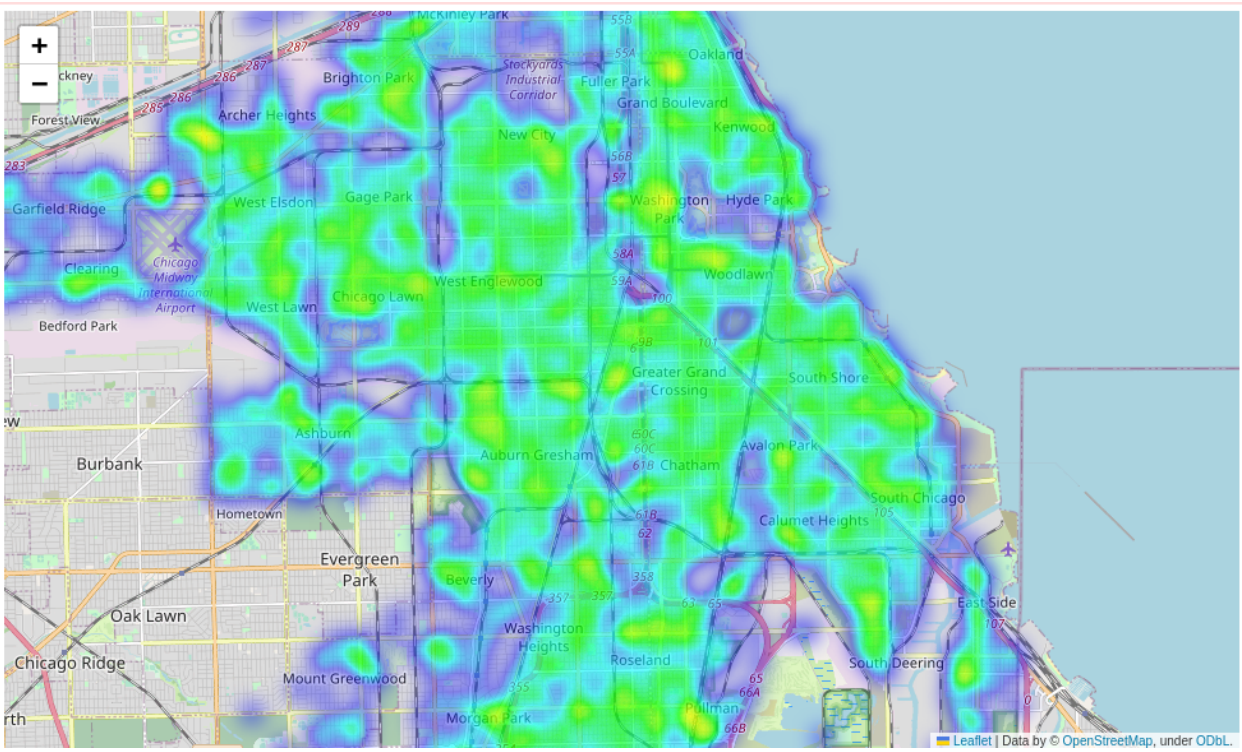

Fig 4: Heatmap of places where crime is at its peak

3.4 Crime Trend Analysis

We then finally analyzed the crime trends in Chicago from 2001 to 2020. The analysis included the overall crime rate, the crime rate by year, month, and day of the week, and the top crimes committed in Chicago.

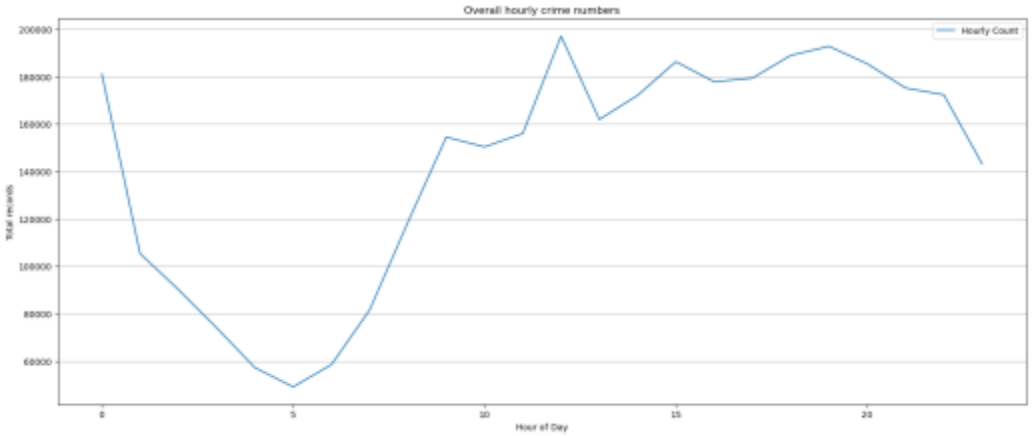

Fig 5 : Time of the day when crimes are at peak

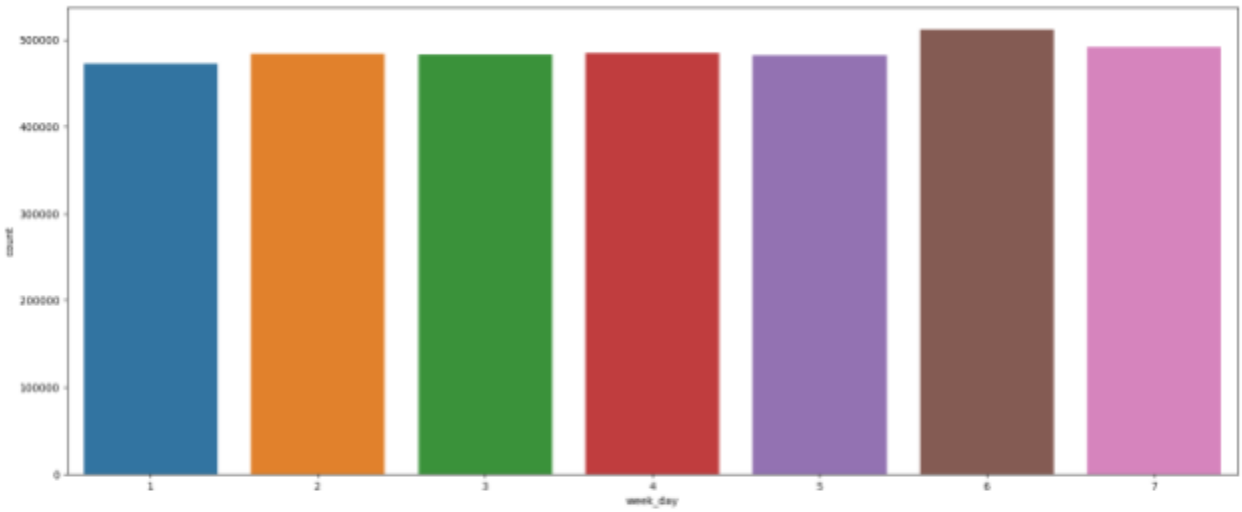

Fig 6 : Crime incident count distribution for different days of week

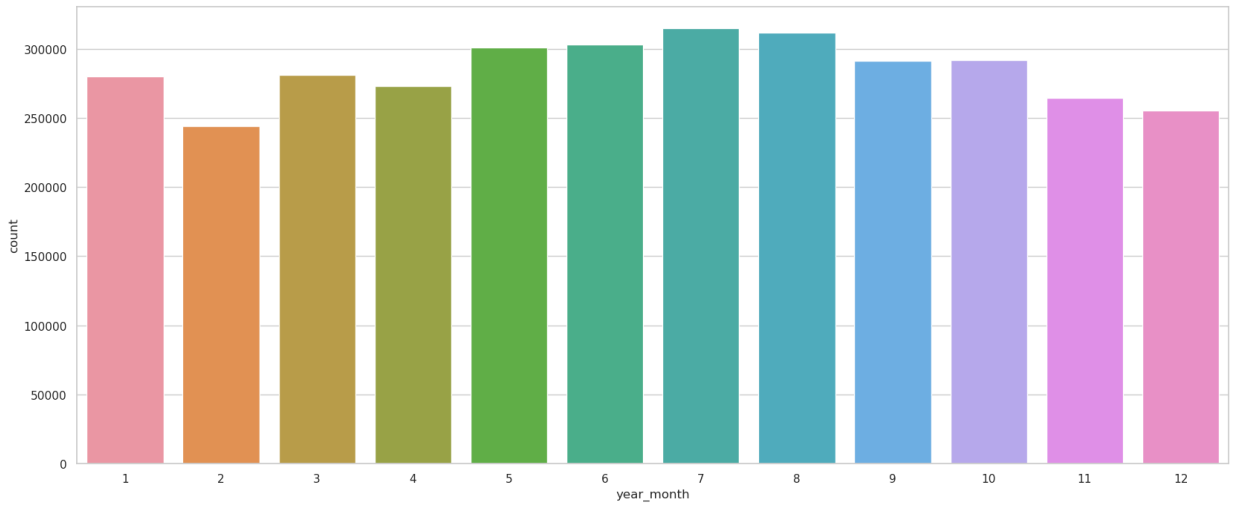

Fig 7: Crime incident count distribution for different months of the year

From the above figures, it is clear that most crimes happen either at midnight or on Fridays. Also, it is evident that the peak crime months are from May to August. Weather conditions, Peak tourism, Economic factor [i,e summer month] people earn and travel a lot.

3.5 Crime Prediction Analysis

In the last step, we employed a robust evaluation strategy by utilizing a random split and k-Fold Cross-Validation technique during the training phase. Additionally, we incorporated additional features such as Location Description, Arrest, etc. to improve the accuracy of our model. To handle categorical data, we transformed it into binary vectors using either One Hot Vector or Label Encoding, depending on the context. We leveraged ExtraTreesClassifier, Correlation Matrix/HeatMap, and Principal Component Analysis (PCA) as feature selection techniques to identify the most relevant features for our model.

To optimize our model's performance, we fine-tuned the hyperparameters, such as the number of neighbors in KNN and the number of trees in Random Forest, through rigorous

experimentation. We also employed an ensemble approach by combining multiple classification models and utilizing soft voting for generating the final output. This allowed us to leverage the strengths of different models and enhance the overall accuracy and robustness of our predictions.

Finally weak was used s a feature to predict the crime based on time. [4]

4. EVALUATION AND RESULTS

After training and testing of the model, a base performance was recorded, that is, performance based on random train split samples and analysis made on the results for the model. The metrics used to measure the performance were accuracy and an F1 score. The experiment was performed using the train test split method where we divided the dataset into 70% training set and 30% testing set. We used the MulticlassClassificationEvaluator from PySpark to calculate the accuracy.

MODEL	**ACCURACY**	**F1 SCORE**
KNN	46.45%	25.11%
Random Forest	21.67%	15.9%

5. CONCLUSION

The machine learning model used in our analysis has shown promise, but its performance can be impacted by the quality of the dataset. Upon examining our dataset, we have observed that the correlation between features is crucial for accurate predictions. However, our correlation matrix reveals that some features have a low correlation with a target variable.

To improve our predictions, we conducted experiments with different features, such as utilizing only the week of the crime, and location, and incorporating additional variables such as location and arrest rate. While the results have shown improvement with the inclusion of additional features, the improvement was not statistically significant.

5.1 Limitations

- Data Quality: The dataset employed in this study was found to have various issues, such as missing values, inconsistent data, and duplicate values. These limitations could have potentially affected the quality and reliability of the findings derived from the analysis.
- Data imbalance: In our dataset, certain crime types were grouped separately, even though they exhibited similarities. This resulted in redundant and overlapping categories, which could have negatively impacted the accuracy and consistency of our models. Additionally, this may have led to the loss of information and introduced noise in the modeling process, potentially affecting the reliability of our results.

5.2 Future Work

- Incorporating Additional Data**:** Considering other relevant data types, such as economic indicators, demographic information, and weather data, to improve the accuracy and comprehensiveness of the crime prediction model. These additional data can provide valuable insights into the social, economic, and environmental factors that may influence crime patterns and can help enhance the predictive capabilities of the model.
- Exploring different machine learning models: Apart from the existing models, different machine learning algorithms can be explored that are specifically designed for handling imbalanced data, such as Random Forest with SMOTE or other ensemble methods.

- Focusing on Specific Crime Types: Narrowing down the focus to specific crime types, such as burglary, theft or assault can provide better insights and intuition into the patterns and trends associated with those specific crimes. This can enable the development of targeted crime prevention strategies and interventions that are tailored to address the unique characteristics and dynamics of each crime type, resulting in more accurate predictions and more effective crime prevention measures.
- Exploring Sampling Techniques: We can consider experimenting with a combination of oversampling and undersampling techniques to address the class imbalance issue commonly found in crime datasets. Techniques such as SMOTE, and ADASYN can help generate synthetic data or remove redundant data points to balance the distribution of minority and majority classes. This can potentially improve the model's ability to capture patterns in minority-class crimes, leading to better predictions.

6. REFERENCES

[1]Chicago Data Portal, ”Crimes-2001 to Present”

[2] Chung-Hsien Yu, Max W. Ward, Melissa morabito, Wei Ding “Crime Forecasting Using Data Mining Techniques,” 2011.

[3] Neil Shah, Nandish Bhagat, Manan Shah, “Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention, ” 2021.

[4] Vaibhav3M, “Chicago-crime-analysis”

[5] Matthew S. Gerber, “Predicting crime using Twitter and kernel density estimation”, 2013.

[6] Dennis Mares, “Social Disorganization and Gang Homicides in Chicago: A Neighborhood Level Comparison of Disaggregated Homicides”, 2010.

7. APPENDIX

7.1 Aayusha Shrestha

Some of the things I learned after completing this project are:

- The main takeaway from this project is that I learned to handle large datasets. It has helped me understand the challenges and techniques involved in handling big data. I learned how to efficiently process and manipulate large datasets using scalable tools and techniques like Apache Spark to handle the volume, velocity, and veracity of big data.
- I also learned to use different SparkML libraries and write code for different models.
- I learned to integrate and transform diverse data sources for big data analysis. I learned how to handle data in different formats, and how to clean, preprocess and transform the data into a suitable format for analysis.
- I also learned to use Apache Spark to perform exploratory data analysis on big data including data visualization, summary statistics, and data profiling. This has helped me develop skills in scalable EDA techniques which are crucial for big data.

My key contributions include

- Data collection: I was responsible for collecting the relevant data for the Chicago crime analysis project. This involved identifying and obtaining the necessary datasets from reliable sources, such as public datasets or APIs, and ensuring data quality and integrity.
- Data preprocessing: I performed data preprocessing tasks to clean, transform, and prepare the data for analysis. This included tasks such as data cleaning, handling missing values, data normalization, and feature engineering using Spark MLib.

- Data visualization: I created visualizations to explore and analyze the data using various visualization tools and libraries. This involved generating plots, charts, and other visual representations of the data to gain insights and communicate findings effectively.
- Research and Report Writing: I conducted research on the Chicago crime dataset along with my team member to analyze crime patterns, trends, and correlations. I performed the statistical analysis, and data interpretation, and drew meaningful conclusions from the data. I then documented the findings in a comprehensive report, including data visualizations, analysis results, and recommendations for further action or insights.

Overall, my contributions to the project involved a wide range of activities including data collection, data preprocessing, data visualization, and report writing. These activities helped to uncover insights and generate meaningful conclusions from the data which contributed to the overall understanding and analysis of the Chicago Crime Analysis project.

7.2 Suman Astani

Summary of key contribution

- Literature Review: Conducted an extensive and comprehensive literature review on crime analysis, encompassing previous research studies, academic papers, reports, and other relevant resources. Reviewed blogs, articles on Medium, and publicly available Github repositories to gather additional references and insights.
- Model Fitting: A major part of my responsibility was to develop Random Forest and KNN for model prediction. I trained the Random Forest model and KNN using the training data, tuned the hyperparameters for optimal performances and evaluated the

model's accuracy and F1 score. I iteratively refined the model by adjusting the hyperparameters and feature selection techniques to achieve the best possible results.

- Report Writing: I also contributed to the reporting and documentation of the project. This involved using Google Docs as a collaborative tool to create comprehensive reports that summarized the findings and results of the project.
- Updates: After receiving feedback from the instructor, following the project presentation I made several updates to the project to further improve its accuracy and effectiveness. Some of which are incorporating the KNN algorithm to the project, addressed the issue of null or missing values in the dataset such as imputation of null values.

From this project, I learned several valuable lessons:

- The importance of good data: While having a large dataset with millions of observations is beneficial, it's crucial to ensure that the data quality is high and that the features used for prediction are relevant and informative. In this project, I realized that the quality of the dataset and the relevance of the features had a significant impact on the performance of the prediction models. It highlighted the importance of data preprocessing, feature engineering, and data quality assessment in ensuring accurate and reliable predictions.
- Collaborative remote work with GitLab: Working collaboratively with team members remotely using GitLab was a valuable experience. It demonstrated the effectiveness of version control and collaborative coding in a team setting. It allowed for seamless integration of code changes, easy tracking of modifications, and efficient collaboration among team members, even when working remotely.
- Utilizing Google Colab for ML projects: While encountering memory and connection error issues while working with PySpark on local machines, I learned the benefits of

using cloud-based platforms like Google Colab for machine learning projects. It provided a scalable and easily accessible environment for running large-scale ML models, and I also discovered the flexibility of customizing server configurations to suit project requirements.

- Real-world application of curriculum guidance: The experience of working with a large dataset and training machine learning models to predict crime types based on other variables allowed me to implement the guidance from the curriculum in a practical setting. It provided insights into real-world challenges and opportunities in applying machine learning techniques to real-life projects, reinforcing the importance of the practical application of learned concepts.