# CRIME DATA ANALYSIS AND VISUALIZATION USING MACHINE LEARNING

# NAGA RAJITHA BHOGADI TRIBHUVAN REDDY KOTHAPALLY

[**Project repository**](https://github.com/sncmedikonduru/Plant_Leaf_Disease_Prediction)

# ABSTRACT

Crime outcomes not only impact the safety of the community but also the economy of the concerned locality, hence they should be predicted and taken care of. Law enforcement agencies and policy makers would find great help in dealing with the status quo by utilizing the application of this project. Economically, as per the project’s title, “Crime Data Analysis and Prediction Using Machine Learning”, this covers advanced big data frameworks and historical crime data analysis to predict based on the regions of interest. For the automated pipeline project to be successful there is need to find Azure Databricks which makes it very easy to join data discovery and cloud computing as well as data analysis and visualization together with ML integration. As far as specific examples of usage of this analysis goes, crime prediction models have been developed utilizing Random Forest and K-Means Clustering and Exponential Smoothing techniques. Areas for improvement and attention are made clear with the help of interactive Power BI dashboards, which also support adequate resource allocation and intervention measures. This initiative amply demonstrates the success of a case study that seeks to counter measure crime rate increases and enhance safety of a specific community while using a data driven approach that is not limited to a single community but can be replicated.

# INTRODUCTION

Crime has made it impossible for society to skip its negative effects on security, economy and general liveability. An increase in population and urbanization, along with inequality in a society, has proven to increase the likelihood of crime, necessitating the introduction of effective prevention measures. In this context, information oriented methods have successfully proven their worth in exploring, predicting and visualizing the enforcement agency in the context of crime combat. [1]

To address these challenges, the project, “Crime Data Analysis and Visualization using Machine Learning,” aims to benefit from modern technologies. The overall goal of the project is to develop forecast models and visualization techniques regarding crime evolution through several years and regions as provided by the Federal Bureau of Investigation (FBI) database incorporating over a variety of years. This study is one of the most broad-scaled assessments of this kind ever performed, focusing on the integration of data from all 50 states as well as more than 7000 city authorities across the country, in contrast to frequent strategies that focus on small focused areas. [8]

One of the main goals of this initiative is to provide a dual analysis and visualization solution to fight crime. Analytical tools predict the probability of different types of crimes. from the previous model And visualization helps stakeholders identify high-risk locations. Also known as crime hotspots. With these insights, law enforcement organizations can allocate resources strategically and take focused action. The project results will also help people make informed decisions about their personal safety. and helping lawmakers create safer urban environments.   
  
More than that, the project houses some latest technologies like power BI for visualization, Azure Databricks data processing, and machine learning models for prediction. Random forest is among effective machine learning techniques for pattern analysis and risk classification. Such clustering and K-means clustering establish a solid foundation for prevention, implementation by guaranteeing prediction accuracy [4][6][7].

The major objective of this project is to develop a system which is sound, scalable, and easy to use. The project uses data driven strategies for crime-related problems. This project strengthens law enforcement agencies and fusion of technology with societal needs To promote a safer and better-by-information society, it has given birth to the evidence of technology's greatest revolution as a problem solver of complex social problems.

# OBJECTIVES

The goal of this project titled “Crime Data Analysis and Visualization Using Machine Learning” is crime prevention through the use of data-centered strategies. This project focuses on utilizing advanced analytics and visualization techniques in order to achieve the following high-level goals:

* Data Integration and Processing: Research and pre-process a variety of crime data from valid sources (for example, FBI and other law enforcement agencies) so that the output data might be analyzed properly in terms of accuracy and consistency.
* Advanced Crime Trend Prediction: There is a potential use of Predict Crime Trends advanced. This will be achieved by implementing Random forests, and Ensemble Learning and developing different types of ML models to predict crimes, crime rates, and geographic hot-spots for crimes.
* Improved Visualization Tools: Employ Power BI, and other similar applications to design user-friendly dashboards and visual representations. These will enable the stakeholders to comprehend complex data and even identify trends and hotspots with ease.
* Law Enforcement Resource Optimization: Provide information that would enable targeted actions and proper deployment of resources so that law enforcement resources are optimally employed.
* Public Safety Awareness: Educate the public on crime trends and areas to avoid in order to enhance their decision making and trust in the use of data centric approaches.

By achieving these goals, the project creates a scalable framework for tackling crime with technology and evidence-based strategies, making communities safer.

# LITERATURE REVIEW

Mass surveillance and real-time crime predictions can significantly reduce crime rates. Numerous approaches and procedures can be used to enhance the crucial activity of crime analysis and prediction. Crimes are pervasive social problems that have an effect on a nation's reputation, economic development, and standard of living. They developed a framework for visualizing crime networks and assessing them using a variety of machine learning techniques using Google Maps and many R packages. Nevertheless, there is little interaction in the application, and a variety of crime types were not examined. We evaluated a number of crime categories across different locations after analyzing all available data.[9]

Criminal justice agencies used machine learning (ML), data mining, and deep learning to help combat crime by utilizing historical crime data to find and identify crime hotspots and patterns, predict future crimes, and apprehend suspects and offenders. Two authentic crime datasets from the US cities of San Francisco and Chicago underwent cutting-edge massive crime data processing and visualization, claim Mokhtar and Xia [10].

Big data analytics may help police departments and other law enforcement organizations better understand criminal concerns and gain insights that will aid in activity monitoring, event prediction, and decision-making. According to Mingchen Feng's 2019 study, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," big data analytics, or BDA, has become a popular technique for processing data and learning. This study used a variety of state-of-the-art big data analytics and visualization techniques to analyze massive amounts of crime data from three US locations. Time series models will be useful in forecasting future criminal behavior on these datasets.[11]

Spencer Chainey and his team looked into the effectiveness of hotspot mapping in locating criminal hotspots. In order to address the issue of crime, crime scene analysis is crucial. Hotspot mapping is used by many police and crime-reduction specialists to identify spatial patterns of crime. This indicates that the efficiency of hotspot maps in tracking the investigators and predicting geographic patterns of crime was assessed. Thus, it serves as a foundational method for predicting possible crime hotspots, predicated on the notion that historical crime patterns may be useful predictors of future patterns. This study compelled us to tailor our investigative methods based on the types and locations of crimes.[12]

# TOOLS

1. **Azure Cloud Services:** 
   1. Azure Blob Storage: To store the raw and processed datasets
   2. Azure Databricks: For extract, transform, load of the dataset and to train and implement ML models

1. **Hardware Specifications:** 
   1. Processor: standard\_ds3\_v2 CPU (Intel Neon)
   2. RAM: 14GB, 4 core
   3. Storage: It can store upto 500 GB
2. **Programming Environment**
   1. Apache Spark : 16.0 ML (includes Apache Spark 3.5.0, Scala 2.12)
   2. Python 3.12
   3. Databricks Notebook connected to compute cluster

1. **Data Processing & Machine Learning:** 
   1. Pandas
   2. NumPy
   3. Scikit-learn

1. **Data Visualization Tools:** 
   1. Power BI : To develop interactive crime dashboards.

# PROPOSED APPROACH

To predict and visualize the crime data on a map, the following approach is used. It involves four major steps and are explained in the following sub-sections:

The subsections are

1. Process Workflow
2. Data Processing
3. ML Model Implementation
4. Data Visualization

# 1.PROCESS WORKFLOW

This project involves many datasets, To handle data efficiently spark is used. The whole investigation is carried out by using big data concept Spark in Azure. The lifecycle of a process flow starts with a problem and then the data required is collected using FBI API and stored in azure blob storage. The data is transformed in the spark cluster running on Azure Databricks and after the ETL operations are performed the data is stored in Blob storage. We will access the master data for Databricks ML Notebooks. Then data predictions are consumed in Power Bi to create Visualizations. [4] [7]

Figure 1: Process Workflow

**2.DATA PROCESSING**

There are three pronounced widely accepted criminal justice reporting systems and one among those is most commonly used by the professionals of America for crime monitoring and crime-data assessment and public safety measures evaluation. The first three would be National Crime Victimization Survey (NCVS), Uniform Crime Report (UCR), and National Incident-Based Reporting System (NIBRS). This study depended on UCR and NIBRS databases, which are available through the FBI API. The dataset is truly genuine and authentic, and it has been collected from the official website of the FBI. It covers a few states from 2020 to 2023 and a few types of crime.[8]

So as to ensure that raw data is converted into a clean, organized, and useable format, data preparation is a vital activity involved in a data-driven project. Automated, scalable, and efficient big data pretreatment was carried out by Azure Databricks in the project "Crime Data Analysis and Visualization Using Machine Learning". This stage was the preparation of crime data for machine learning models and visualization tasks by bringing together all the work to resolve inconsistencies, cleanse data, create valuable features and automate the entire process.

Azure Databricks, a single platform that combines Apache Spark and offers scalable computing resources, was used to carry out the preparation activities. A Standard\_DS3\_v2 cluster with four cores and 14 GB of memory was set up in the environment. Set to 16.0.x-cpu-ml-scala2.12, the runtime environment was optimized for tasks including machine learning. The data processing workflow was streamlined by the development of pipelines to automate repetitive processes made possible by Databricks' flexibility. To ensure a smooth transition from data import to the deployment of machine learning models, a modular pipeline structure including two primary tasks—Data\_Preprocessing and ML\_Tasks—was employed.

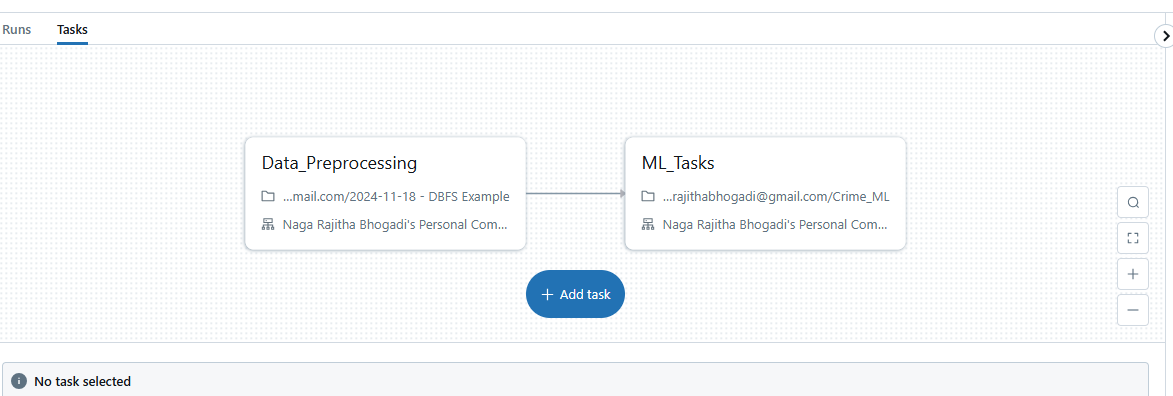


Figure 2: Automated Pipeline in Azure Databricks

Data ingestion was the first step of the preprocessing stage. Importing the raw crime data was managed through reliable sources such as the FBI’s uniform report for crimes (UCR) or the national incident-driven reporting system (NIBRS) storage using Azure blob storage. These datasets contained information on various categories of crime, their occurrence sites and the periods within which these crimes occurred. Azure Databricks acted as the processing layer where the raw data was extracted and created for deep exploration. The design of the pipeline made the entire intake system automated thus ensuring efficiency and scalability.

The once the raw data was ingested the first step was rigorous cleansing to account for missing values, outliers and duplicates. Important variables like crime type or location with missing data were adapted using statistical imputation of the data which were then removed if they were considered not important for analysis. Instances of dispersed data files were deleted to prevent data being recorded more than once and affecting data. These cleaning steps enhanced reliability of the files and the dataset and its usability.

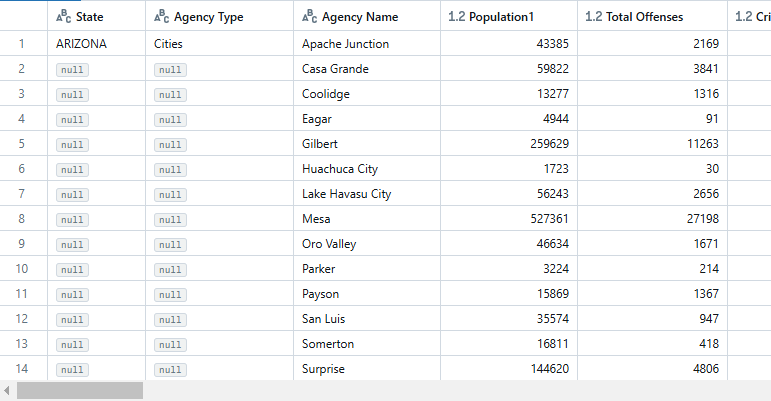


Figure 3: Before Data Pre-processing

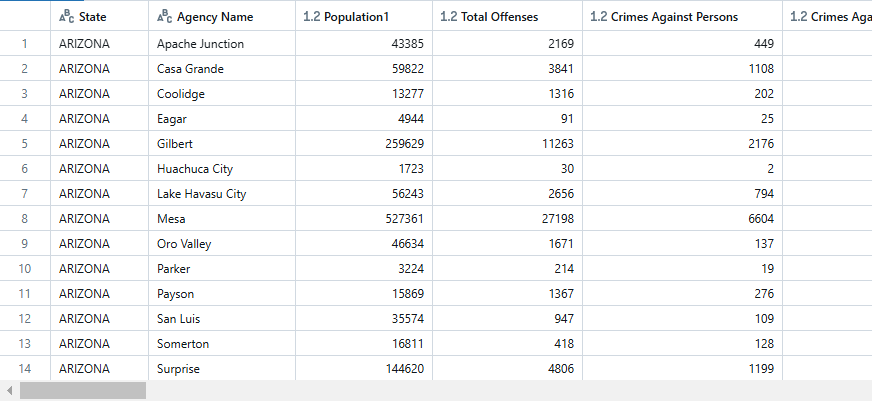


Figure 4: After Data Pre-processing

Figure 3. represents the dataset before preprocessing, where several columns, such as Agency Type, contain null or missing values, and the data structure appears inconsistent. The challenges related to the understanding of the unlabelled data set raw data indicate that cleaning is entirely necessary and normalization and cleaning techniques have to be applied. On the other hand, Figure 4. shows the dataset after preprocessing, where null values have been removed or imputed, columns have been streamlined, and the data is more structured and ready for analysis. Important variables such as State, Population, Total Offenses, and Crimes Against Persons and Affections Against Property are all cleaned up allowing for further better analysis and better machine learning later.

The dataset was therefore later adjusted so that it meets the requirements for machine learning applications, particularly the requirements for the trained and structured dataset. For instance, sensitive algorithms to features magnitude were compensated for by normalizing the population size, as well as the number of crimes reported. Next, against Stated categorical variables, State or Crime Type which were incorporated in the models were generated by means of label encoding or one hot encoding. Through such alterations, the data set was made uniform and corresponded to excelled performance even in predictive modelling tasks.

At the early stages of development pixel normalization had to be automated. The process that included feature engineering as well as data cleaning was done completely automatically on Azure Databricks the degree of human involvement was reduced and the uniformity of datasets was achieved. Each new dataset had no impact on the scalability and flexibility structure of the pipeline since it was prepared in accordance with the same principles. The modularity of the pipeline made it possible to perform tasks with sufficient looseness and thus allowed the team to respond to the changing environment rather quickly. The incorporation of preprocessing activities with machine learning tasks within the pipeline made the pre-training and test processes efficient with lesser processing time. After the pre-processed data in the form of a dataset had been divided into different subsets, 20% for testing purposes and the remaining 80% to train the machine learning models. By balancing the data this way the chances of the model overfitting were the least since the models would perform well on unseen data as well. The last stage of pre-processing the dataset before using it was that it was saved in the Azure Blob Storage which enabled an easy retrieval of the dataset for visualization as well as further analysis and research.

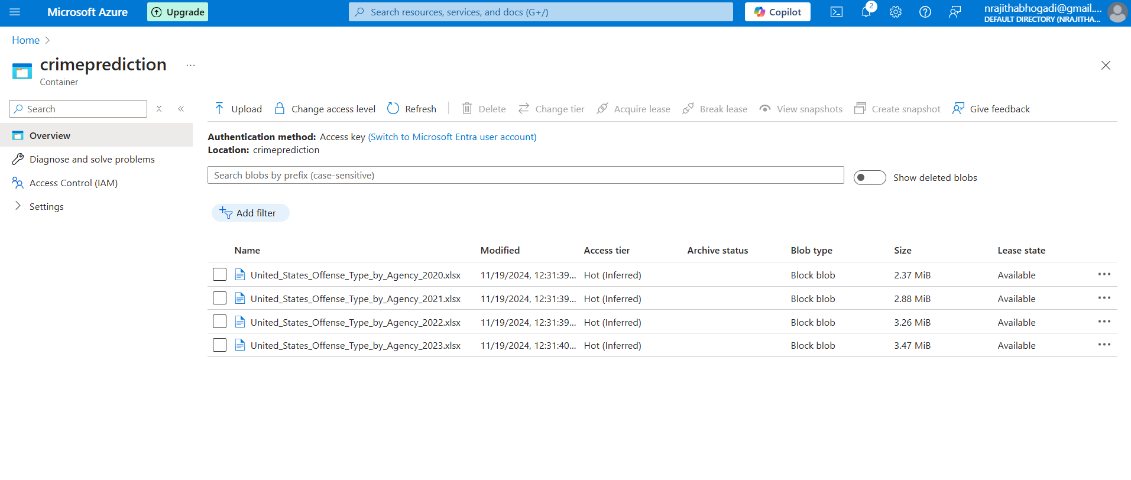


Figure 5: Azure Blob Storage container - Input & Output

Following preprocessing, the dataset was divided into subsets for testing and training, with 20% put aside for testing and the remaining 80% utilized to train machine learning models. This divide decreased the possibility of overfitting by ensuring that the models generalized well to unseen data. Azure Blob Storage was used to store the final pre-processed dataset, making it easily accessible for display and additional research.

Using Azure Databricks for preprocessing saw a number of advantages. This virtually seamless and efficient preprocessing is made possible by the elastic compute resources of the service automatically scaling to cope with massive workloads. You can work around the application and develop debugging tools in real-time during the writing stage by using PySpark scripts in combination with the Databricks notebooks. The automation of preprocessing activities reduced human error and saved time, in addition. Moreover, since all preparation tasks were done in common notebooks, it was much easier to cooperate and collaborate in the environment supported by Databricks.

Besides, the system was able to cope with the changes in the crime data due to the event driven structure so it supported batch as well as real-time data processing. Such flexibility made it possible for the preprocessing pipeline to handle both new incoming data streams and old datasets remarkably well. Also, because Power BI would be used as the visualization tool, storing the pre-processed data in Azure Blob Storage was the very reason why the pre-processed data would be useful.

As a conclusion, the pretreatment stage of the unprocessed control crime data in this project was very extensive owing to the fact that it had to prepare the data into a consistent usable format. The project also went on to build a preprocessing pipeline that is cost effective and very easy to build using the many powerful features of Azure Databricks. This solid example of the problem has pointed out the importance of data preprocessing in the context of large-scale analysis projects by getting ready the required data set for machine learning.

# 3.ML MODEL IMPLEMENTATION

Once the data has gone through preprocessing, it will be modelled on Azure Databricks for machine learning (ML) using the cleaned final dataset. This project employs a host of machine learning techniques for the training and development of the prediction models. These algorithms will be able to predict the high-risk areas, crime trends, and most common types of crimes that are likely to happen in the respective areas. Predictive samples like that of crime estimations in 2024 and 2025 will be used to develop new datasets that will be used for further visualization.

The models are geared towards a number of important predictive tasks. The first of these projects estimates from demographic and historical data the expected crime figures for 2024 and 2025. Each likely crime is considered individually according to the probabilistic distribution by cities described using the statistics of geography. The models are concerned about all of the different crimes within each subdivision of the country while enforcing how they would focus their interest on those items and demonstrate trends in crime within particular regions. Lastly, the models classify cities or neighbourhoods into different bands, such as high risk, medium risk, and low risk, combined with regression outputs and clustering methods such as K-Means. This provides increased efficiency for law enforcement allocation of resources.

It displays the various forecasting approaches as they exist in combination as components of the forecasting system. One can then deduce anything valuable from processed data. Supervised maturities like Random Forest Regressor as well as Classifier have to be used for a class-proportional sub-categorization and forecasting related to diverse crimes against the overall counts of crimes, involving characteristics such as historical trends, population density, etc. The K-means clustering algorithm is an example of an unsupervised algorithm that helps to discriminate areas into high-, medium-, and low-zone criminal activities on the basis of comparable crime intensity. Seasonal and future crime trends will also be forecasted for 2024 and 2025 by the Exponential Smoothing method. These models provide good prediction capabilities as well as good knowledge.

# 1.Exponential Smoothing

The Exponential Smoothing Model has widely been used in popular time-series forecasting for estimating trends and seasonal movements from historical data. In this approach, older observations tend to get lower exponential weights than more recent observations. The Exponential Smoothing Model predicts the total crimes that might occur in 2024 as well as 2025 in this project. It is a very practical option to study crime patterns over long durations because of its simplicity and efficient handling of seasonality. The model makes forecasts based on derived crime data from previous years, giving relatively real insights for proactive planning and decision making.

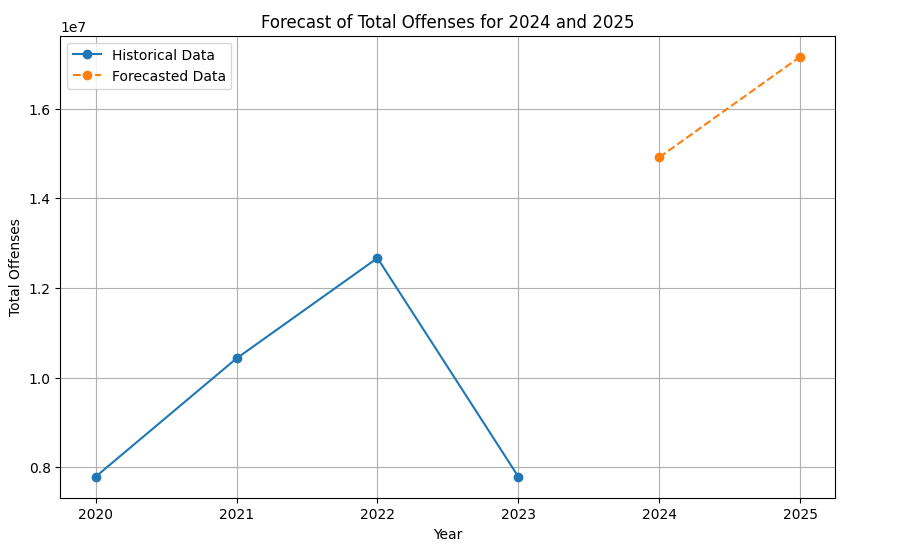


Figure 6: Forecast of Total Crime across USA in 2024 & 2025

The time-series prediction concerning crime from the year 2020 to the year 2025 is set up with the Exponential Smoothing Model and has been shown in Figure 6. Actual historical data extract from 2020 to 2023 is represented by the blue solid line, which peaks total offenses in 2022, after which it significantly drops in 2023. An orange dashed line indicates the forecasted data for 2024 and 2025 to show that this is when bulged offenses are expected at this time.

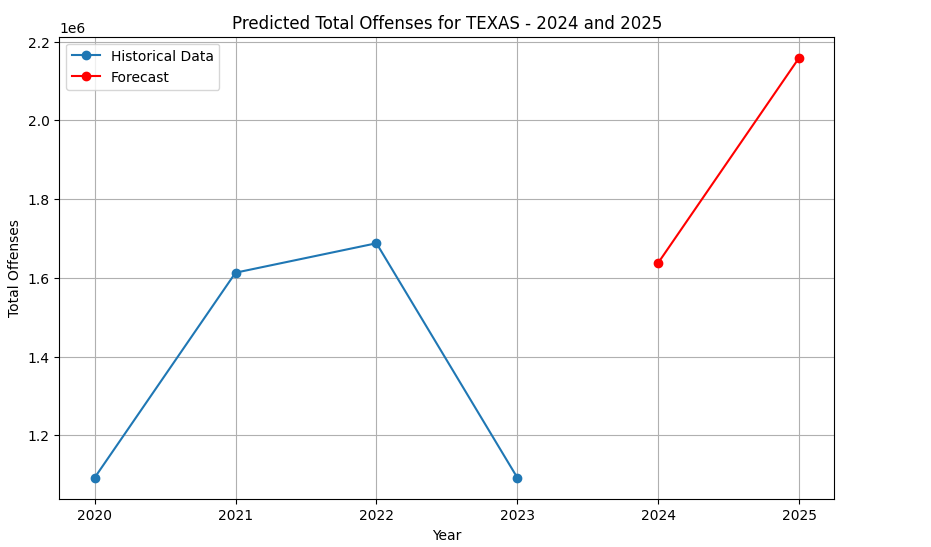


Figure 7: Forecast of Total Crime across State Texas in 2024 & 2025

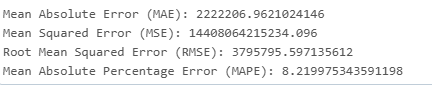


Figure 8: Exponential Smoothing Metrics

The performance of the exponential smoothing model was measured using several metrics. The average absolute difference between predicted and actual numbers is 2,222,206 or mean absolute error (MAE). A high MAE values implies significant prediction errors and can be problematic in case of modelling highly valuable data. Some of the variability in predictions are highlighted by strong aberrations or high as well as mean square error (MSE) with a value that equate to 1.44 × 1010 which penalises even greater errors. These observed deviations are confirmed to exist further more by root means square error (RMSE), it measures size of the prediction error and it equals 3,795,795. The average model prediction deviates from an actual value by 3,795,795.8 which is equal to average based on percentage from a real anticipated value being equal to 8.22% according to mean absolute percentage error (MAPE), this makes satisfaction for obtained broad outline of crime types prognosis accuracy and efficiency.

# 2.Random Forest Regressor

The Random Forest Regressor is a supervised machine learning tool that helps predict a number value using decision trees during training and finding the average of these output values for accuracy and reliability. This work forecasts the total crime offenses per state for the years 2024 and 2025 using the concept of the Random Forest Regressor. Indeed, it is well-equipped with capabilities of dealing with complicated data sets in the form of interrelated features with non-linear correlations; hence, this method will prove a stronghold for all the crime data processing information. The method becomes a potential way to identify high-crime-geography by giving insights into the geospatial distribution of crimes which will help the interested parties to organize interventions and distribute resources more effectively.

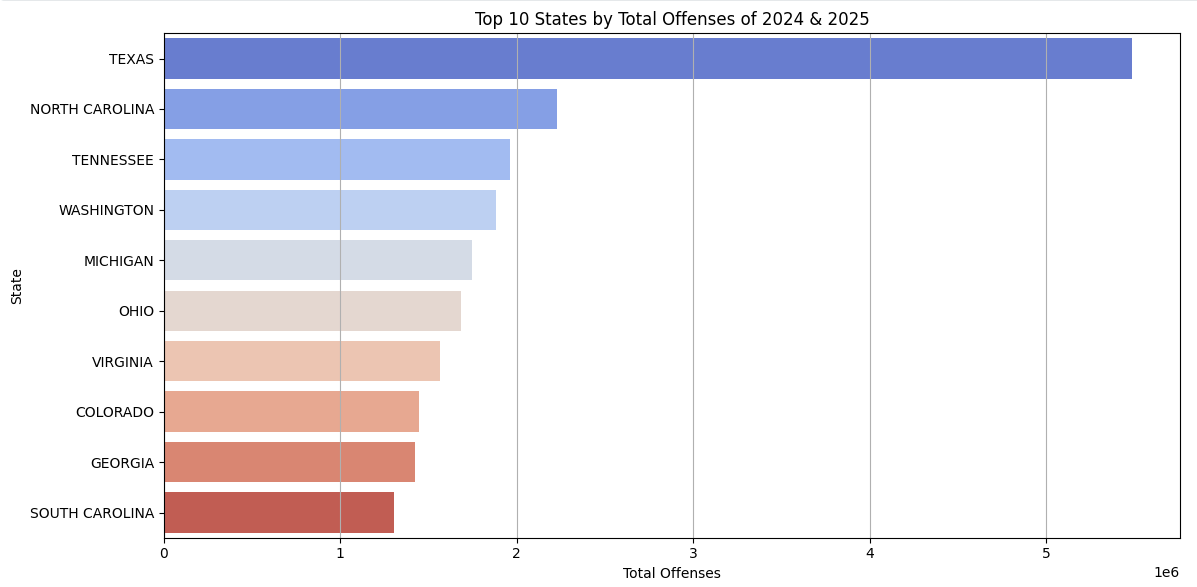


Figure 9: Geographical Distribution of Crimes of 2024 & 2025

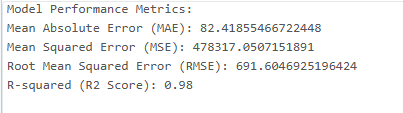


Figure 10: Random Forest Regressor Metrics

The performance of Random Forest Regressor is evaluated using some important metrics. Mean Absolute Error (MAE) is 82.42, which means the average difference between predicted and actual crime counts is low, so this shows the precision of the model in predicting. Mean Squared Error (MSE) (478317) penalizes the largest errors, while it is low so it depicts that this model predictions are consistent. Root Mean Squared Error (RMSE) (691.60), provides an easy to understand way of amplifying the errors’ magnitude together with accurate estimation P-values by interpreting its value more one can conclude about our assertion regarding this model accuracy. R-squared 0.98, means that 98% of variance in crime data can be explained by our model.

Random Forest Regressor performed very well as indicated high P-value, It also generated output for geographical distribution prediction of total crime offense and has been concluded that states like Texas/North-Carolina/Tennessee having highest predicted/No-of-crime count from year 2024–2025.

# 3.Ensemble Model:

The Ensemble Model, which fuses the prediction of three powerful machine learning algorithms i.e, Random Forest Classifier, Gradient Boosting Classifier, and XGBoost Classifier was used to predict the most common crimes type in a particular region. Ensembled method takes advantage of all base classifiers using soft vote for final prediction to achieve higher accuracy. Hence this way of modelling is robust as it dedicates equal probabilities from all models and choose best final predictions to make final decisions.

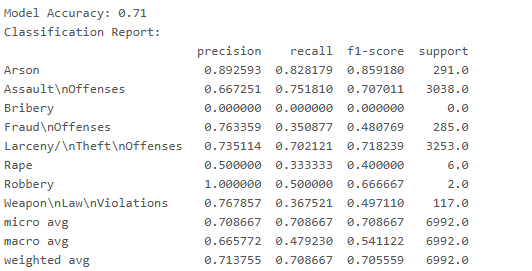


Figure 11: Ensemble Method Classification Report

The model gave an overall Accuracy of 71% which clearly indicates that the model learnt well among features in dataset. Diverse Performance parameter gives us different information like where model does well and where not. In our case Arson has highest Precision i.e, 89.25%, Recall i.e, 82.18% And F1 Score i.e, 85.91% So our model can be referred as accurate classification due High(Values) precision-recall scenario .Likewise one often repeated crime category Larceny/Theft Offenses has balanced result hence can refer as average category because our data Neither High Nor low preciseness neither high recall value but average precision ad recall hence balance result. A category Assault Offenses provide moderate performance parameters values Precision=67.25%, recall =75.18% at minimizing False positive there quite challenging also increase true values classification. Bottom line is more over these parameters this business problem required to concentrate on other output parameter. Bribery performance measure could not know because we didn’t have example for that in our dataset. The macro avg F1 score being only 48.07%, the lowest performing class “Bribery” clearly brings down the avg ,but if we consider weighted F1 Score then it is around somewhat good performing 71%.

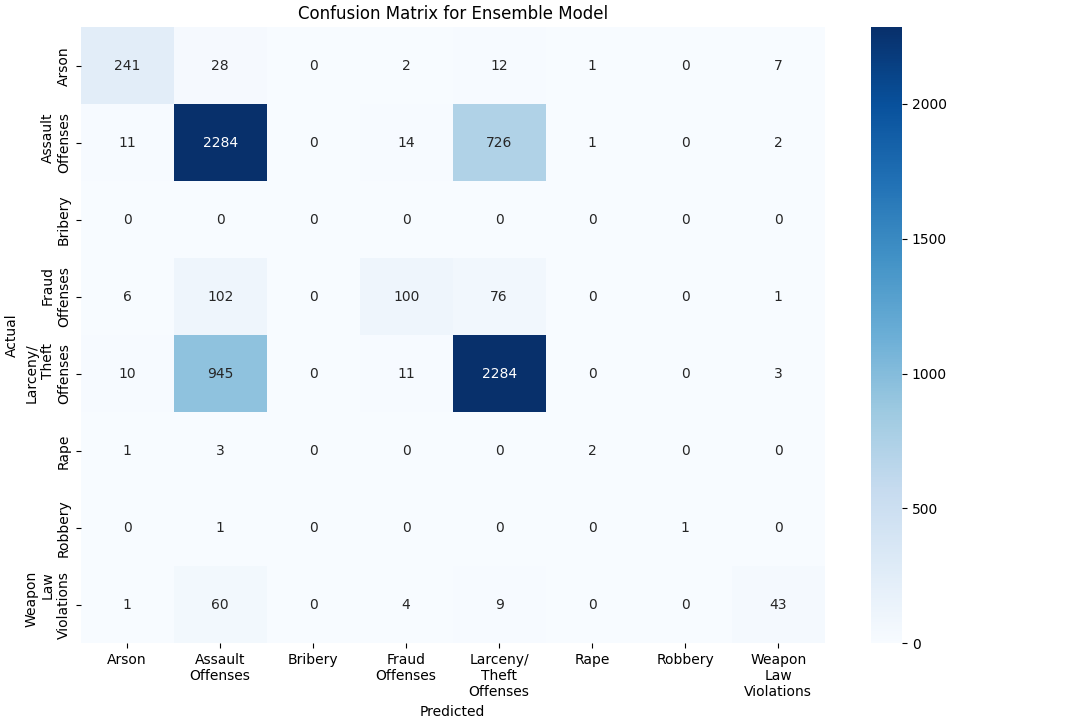


Figure 12: Confusion Matrix of Ensemble Model

The confusion matrix provide idea about which class the model may be predict correct. For example, Assault Offenses classes have a very high rate of instance which predicted correctly (2,284) and other overlapping class, like Larceny/Theft Offenses and Fraud Offenses, are more considerable instances will be misclassified.

Ensemble Model is well learning these imbalanced data (as possible as it can). But again general performance seem to be significantly impacted by dataset nature since they are real-word ones crime.

# 4.Random Forest Classifier:

The Random Forest Classifier is a supervised machine learning algorithm that will be used in this particular project to come up with probable types of crime occurrence by making use of historical crime data and its features. Its operations involve creating several decision trees and aggregating each one's decision to produce the final prediction. This makes it a robust classifier when it comes to accuracy. Features such as population, state, city, total offenses, and other demographics will be modelled to predict common crimes such as arson offenses, assault offenses, bribery, fraud offenses, larceny/theft offenses, rape, robbery, and weapons law violations. Random forests are, therefore, the best choice for modelling crime classification; they can handle extremely imbalanced datasets at very high dimensions.

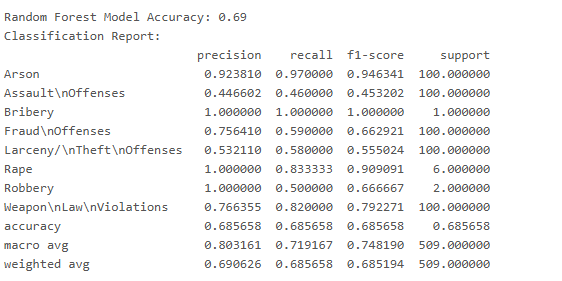


Figure 13: Random Forest Classifier Classification Report

The model achieved an accuracy rate of 69%, which means that approximately 69% of the time, correct category classifications were predicted. The detailed scores help us get more clarity about where our model excelled along with areas where we can improve our results/performance further. For example, the precision value for arson is coming out as 92.32%, recall comes out as 97%, and the F1 score is coming out as 94.63%. This indicates that there will be very minimal false assigned values (may be actual innocent people being considered under this category) when he/she actually falls into the arson criminal activity range or vice versa if there exists any backlog that might not have been cleared leading to incorrect assignment by enforcement agencies). Thus we can see here how much confidence our model has shown us in its capability of making right predictions even without looking at some other aspects such as heatmaps, etc. Similar results are obtained for criminal activities like Bribery (100% precision, recall & F1 score), assault offenses (45.32%), larceny/theft offenses (55.52%) Macro avg F1-Score: 74.81%Weighted avg F1-Score: 68.15.

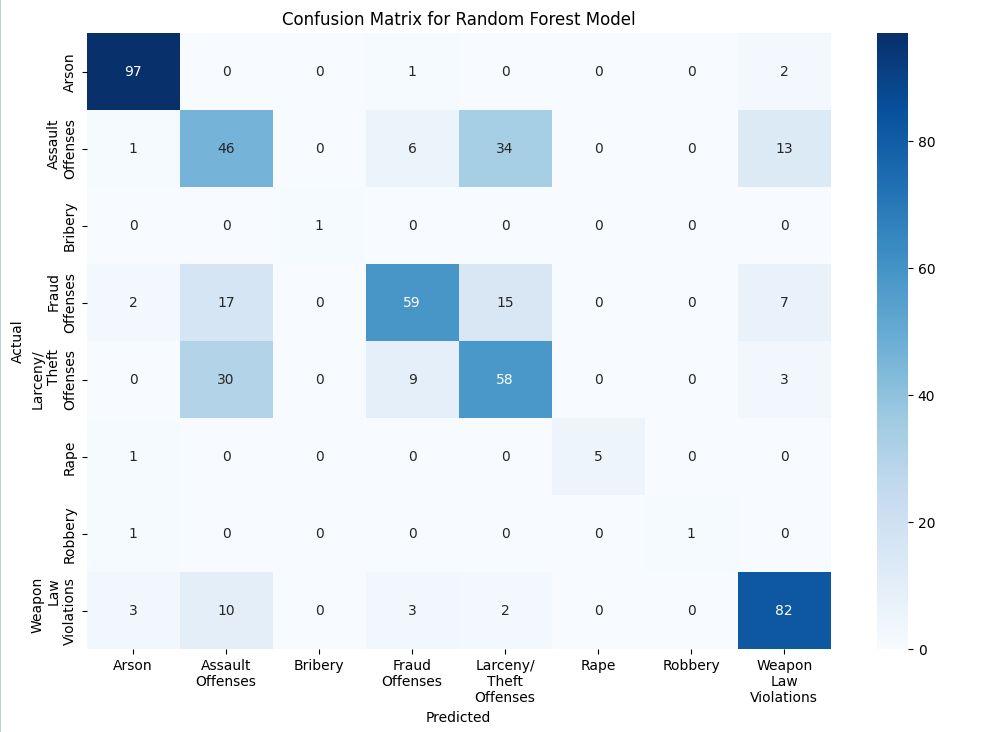


Figure 14: Confusion Matrix for Random Forest Classifier Model

The predictions of the model can be visualized with a confusion matrix. For example, Arson is predicted correctly 97 times and has only a few misclassifications. Larceny/Theft Offenses and Assault Offenses have more misclassifications and suggest parts where the model could do better. On the flip side, Larceny/Theft Offense and Assault Offense more misclassifications can be seen as areas where the model can be improved. This so high number of correct predictions Within the Weapon Law Violations (82 correct cases), indicates the model's potential in being confident for showing performance even with predominant classes that are well represented.

Thus Random Forest Classifier suits quite well to the data, but the overall accuracy and performance metrics are influenced by the dataset themselves as it is real-world crime data.

# 5.K-Means Clustering:

The efficient identification of crime hotspots will be possible through k-means clustering algorithm for that it classifies an area having varying crime statistics into different levels of risk. This is a tool of unsupervised learning which simply means it clusters the data points-of-similarity by minimizing the intra-cluster variance while maximizing inter-cluster difference. K-means have been successfully applied in this project in classifying the regions into geographic region high; medium; low risk classes and the spaces are open for strategic decisions in resource allocation for law enforcement agencies.

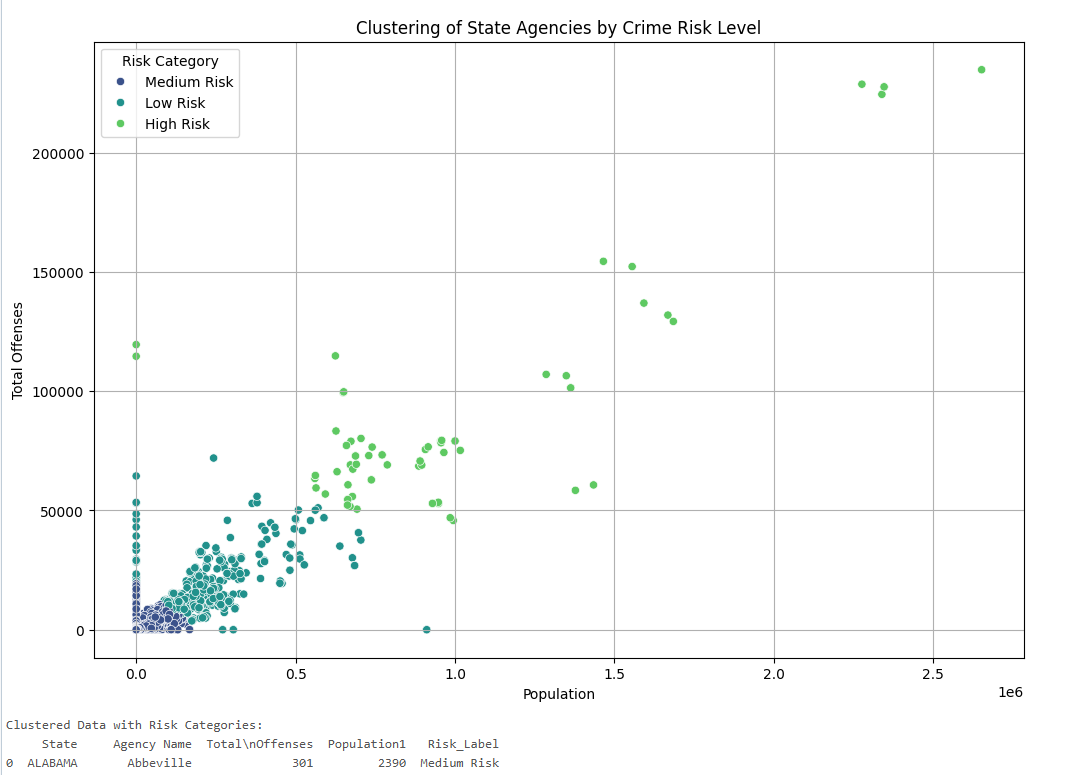


Figure 15: Categorizing each city into zones

Figure 15. demonstrates the categorization of data points, regions that can be defined by total offenses and population density whose values are also critical to crime risk. Regions that can be characterized as having a high population and high offense totals are defined as being "High Risk," moderate-low offense-total regions and moderate-high population were labelled "Medium Risk"; while regions with low population density and the least number of offenses were classified under "Low Risk”. This clear distinction and isolation among clusters serve as one of the signs of how much effective the model is in differentiating one stratum from another with respect to crime scenarios.

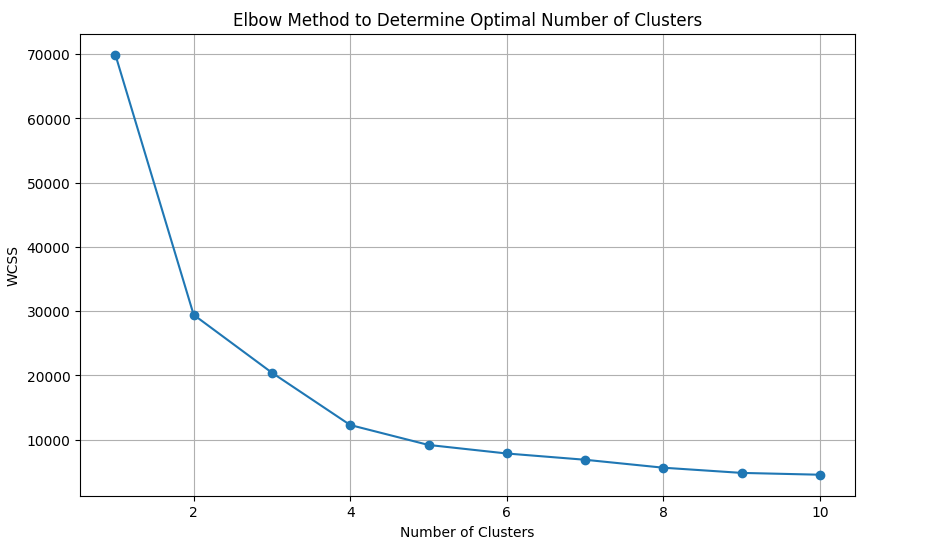


Figure 16: Elbow Method to determine Clusters

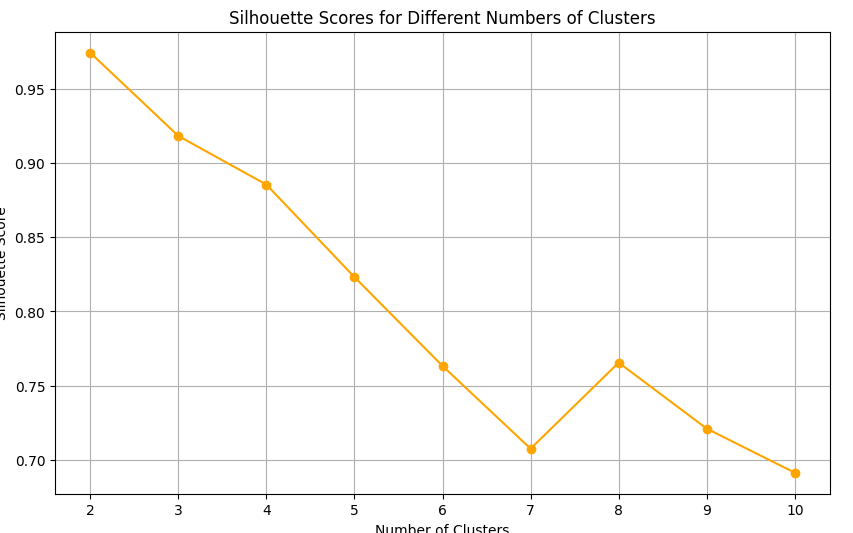


Figure 17: Silhouette Score to determine Clusters

The optimum number of clusters was determined using the Elbow method and silhouette scores, two of the metrics mostly used to evaluate performance of clustering techniques. The elbow technique is to plot the WCSS with the increasing number of clusters, and seek the elbow point where the increment in clusters stops yielding significant improvements. Three clusters appeared optimum according to the elbow method; this was visible by the significant decline of WCSS up to that point, then flattening. Similarly, the silhouette scores were computed for different numbers of clusters; the higher the score, the better defined the clusters. As for the Silhouette Score, that was at a maximum for 3 clusters, thus supporting this choice.

# 6.Predicted Outputs

The results produced from the implemented machine learning models are stored as structured datasets in a secured Azure Blob Container. This makes the data storage scalable and accessible while integrating it with visualization applications such as Power BI. The datasets hold some of the following critical results: future crime predictions for the years 2024 and 2025 might predict types of crimes occurring the most by city and by state and categorized risk levels of areas depending on their crime intensity. Machine learning techniques have been able to derive outputs organized into structured datasets that have been stored in an Azure Blob Container. This data becomes scalable and accessible while also integrating with visualization applications such as Power BI. The datasets hold some of these key results as follows: Future crimes predictions for 2024 and 2025, types of crimes occurring the most by city and by state, and categorized risk levels of areas according to their crime intensity.

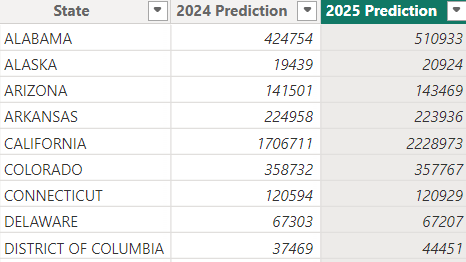


Figure 18: 2024 & 2025 Total Offenses

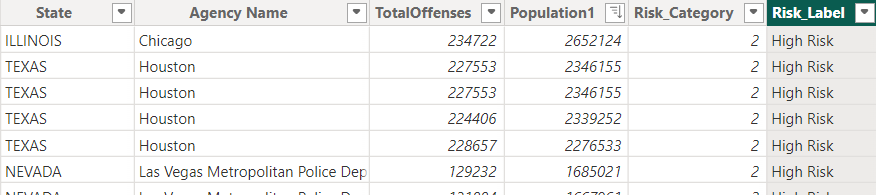


Figure 19: Categorization based on Risk

The machine learning models have been integrated into an automated pipeline, and this includes the ensemble models, clustering algorithms, and predictive regressors; the entire process is automated without manual interference and delivers continuously updated results upon the entry of new data. The real-time insights, predictions for future happenings, and other in-the-moment actionable insights, such as crime trends with high-risk areas, will then be available to stakeholders. Therefore, it is an automated pipeline covering the entire process - from data preprocessing, model execution, to storing the outcome and thereby diminishing manual overheads while simultaneously improving system efficiency.

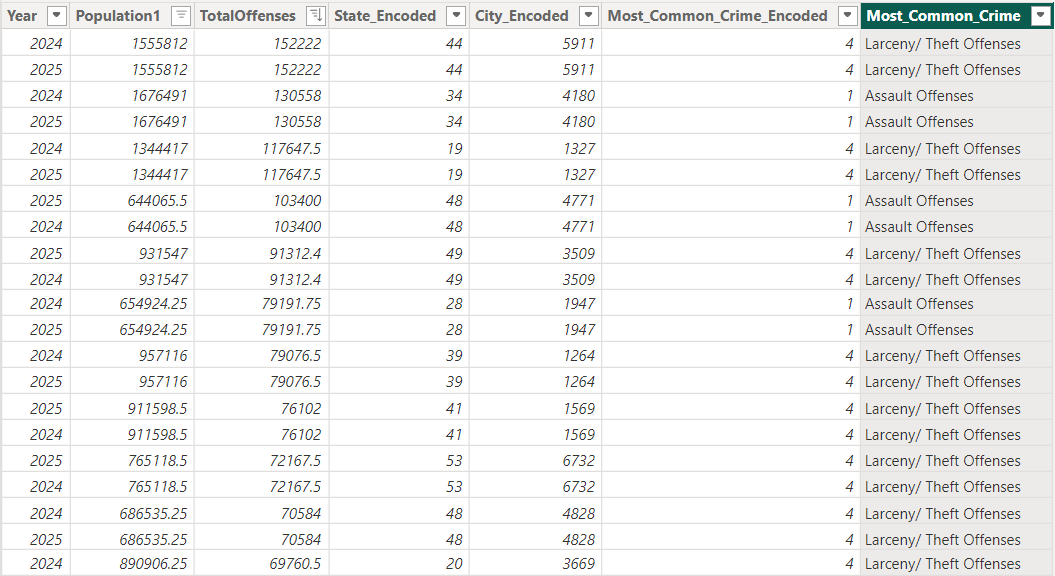


Figure 20: 2024 & 2025 Top predicted Crimes

Thus, this automated, scalable approach assures that the machine learning framework embodies a dynamic continuity of relevance, adapting to changes in the data and maintaining the accuracy of the insights over time.

# 4. DATA VISUALIZATION

The end of this project was marked by the visualization of the results obtained from the machine learning models using Power BI dashboards, merging dynamic and interactive views with stakeholders. The next step was to integrate Power BI and hence seamlessly import the structured datasets, stored in Azure Blob Storage, into the project. This step is transforming raw outputs from machine learning into intuitive visuals to make it very handy for decision-making in crime analysis and detection.

It is both art and science combined together. It has a form of visual/pictorial communication and involves research and development. It also helps improve our ability to judge and interpret figures and facts. Producing crime density maps aimed at criminals would help criminal analysts study and understand trends in crimes. The trend of criminal activities gives knowledge to law enforcement and intelligence agencies about investigating and preventing criminal activities. Understanding is considerably facilitated with the help of crime data with intervals on the map relative to location. The research presents a new way of mapping and predicting historical data on crimes and their development over time into new directions.

The Microsoft platform’s interactive and visual tool Power BI aids in the visualisation of the dataset across states. The police and law enforcement investigators can study the local crime kinds with the use of crime hotspot maps. With the help of this tool, users will be able to visually filter the dataset so they can make judgments.

# 1. Crime Hotspot Maps

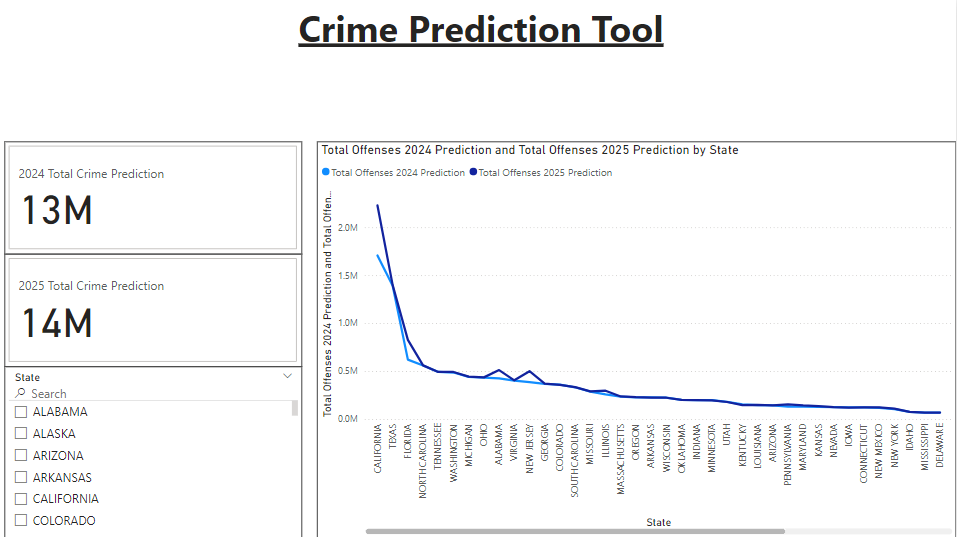


Figure 21: 2024 & 2025 Crime offenses by State Line Graph

# 2.Crime Hotspot Maps

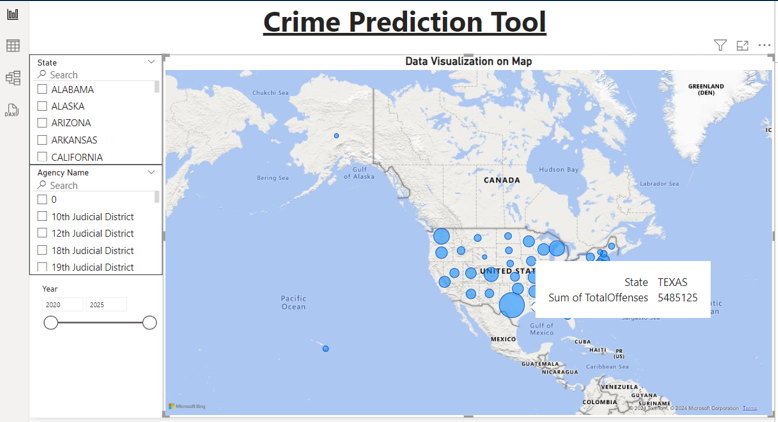


Figure 22:Crime Hotspot Maps By offenses across US

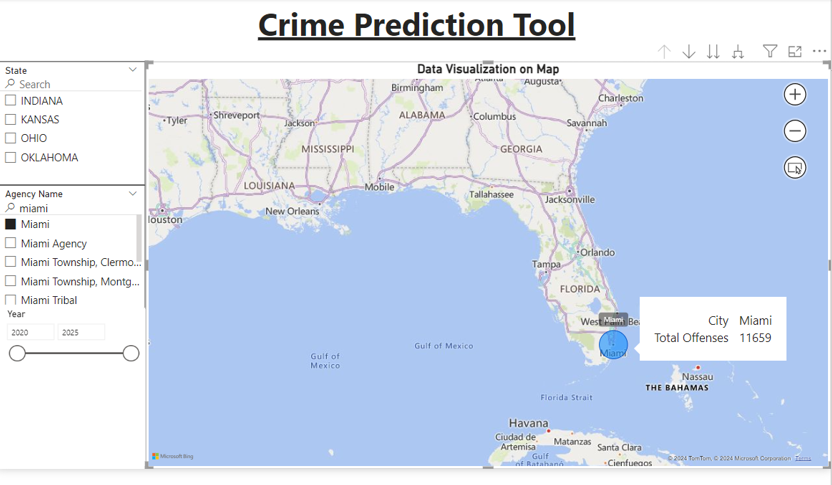


Figure 23: Crime Hotspot Maps by Offenses across Cities

# 3.Crime Distribution Visualization

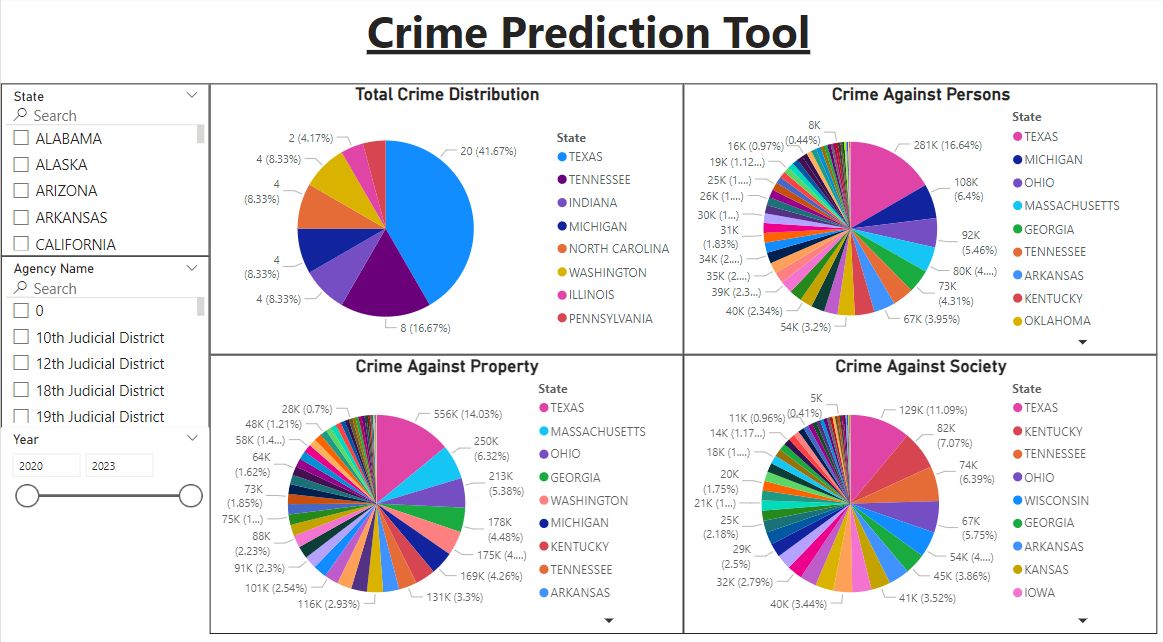


Figure 24: Crime Distribution across US by crime types

# 4.Crime Risk Areas

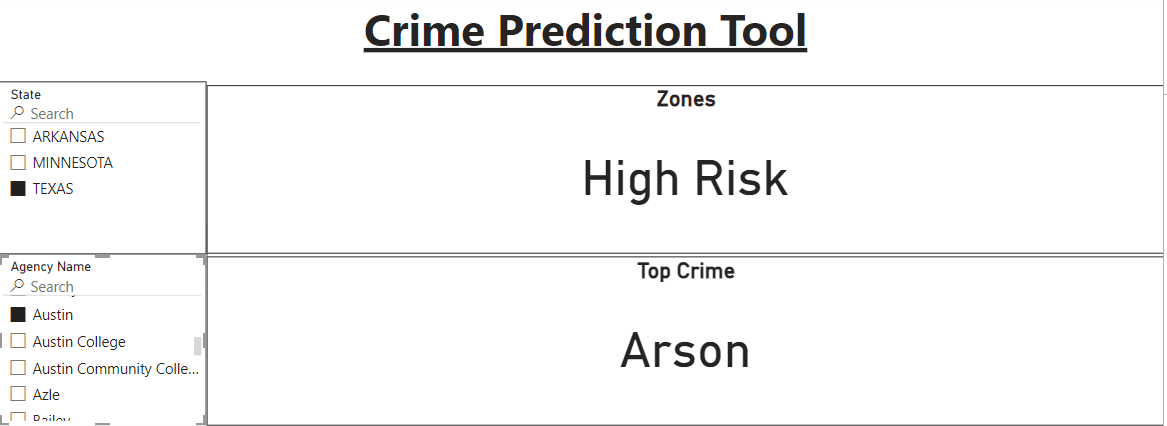


Figure 25: Risk Zone Prediction by City and State

Power BI dashboard encompasses several informational visuals for the efficient display and interpretation of crime prediction data. Among them is a line graph which reflects the magnitudes of states such as California and Texas that would have high expected crime rates. In addition to a direct comparison of total offenses from 2024 and those in 2025 across states, the graph reveals that which succinctly shows crime trends over the years, enabling high-end stakeholders to see where real trends exist and where they must focus further efforts on high-risk areas.

Prominently displayed through their KPIs, total crime projection for 2024 was 13 million while that for 2025 was 14 million. Such KPIs give decision makers a rather succinct view of the breath of infractions expected making it quick for one to discern the larger trends.

# CONCLUSION

The project 'Crime Data Analysis and Visualization Using Machine Learning' deals with existing machine learning technology to efficiently guide big data in interactive visualization around various issues of crime. The future of crime prediction has been made possible through high-precision current futuristic models-the Random Forest and Ensemble Methods-and accordingly, high crime-risk areas have been confirmed and crime levels classified. K-means clustering is deployed. Moreover, the automated pipeline involves a full feeding of datasets, model executions, and intake of fresh real-time snippets for effective scaling.

Power BI dashboards turn analytical output into actionable insights that allow stakeholders to dynamically visualize crime trends. By using crime distribution, risk category, and prediction for 2024-2025, law enforcement agencies can better allocate resources and plan proactive interventions. Even though this project has fulfilled all its objectives, some future enhancements, such as addressing data imbalance and incorporating socioeconomic factors, can have a greater impact on it. This is an example of how data-driven solutions can create opportunities and interactions for communities and better decision-making overall.

# FUTURE WORK

In enhancing accuracy forecasting, the inclusion of socioeconomic indicators like income level, unemployment rate, and education information should also come into the study. Including an array of local and international crime reports to the dataset provided avail and generalizability of the developed model. Neural networks and some other sophisticated techniques in deep learning could be helpful to detect those advanced patterns and improve prediction and classification of crime. These attributes will make the system more robust, scalable, and adaptable to any scenario.

# ACKNOWLEDGEMENT

We would like to thank our professor Mentor: Dr. Samira Zad, Instructor: Dr. Ananda Mondal who gave us this opportunity to work on this project. We got to learn a lot from this project about Data Science and Data Visualization. We’d like to express our gratitude to Azure for providing free services and FBI for providing access to data.

# REFERENCES

[1] Chauhan, Tirthraj & Aluvalu, Rajanikanth. (2016). Using Big Data Analytics for developing Crime Predictive Model.

<https://www.researchgate.net/publication/302026832_Using_Big_Data_Analytics_for_developing_Crime_Predictive_Model>

[2] Mokhtar Mansour Salah and Kewen Xia. 2022. Big Crime Data Analytics and Visualization. In Proceedings of the 2022 6th International Conference on Compute and Data Analysis (ICCDA '22). Association for Computing Machinery, New York, NY, USA, 24–28. <https://doi.org/10.1145/3523089.3523094>

[3] Anjana Ravi, Praseetha V.M.2021. CRIME PREDICTION AND ANALYSIS USING BIG DATA

<https://www.jetir.org/papers/JETIR2107201.pdf>

[4] Scikit-learn developers. (n.d.). Ensemble methods. Retrieved from

<https://scikit-learn.org/stable/modules/ensemble.html#forest>

[5] GeeksforGeeks. (n.d.). K-means clustering introduction. Retrieved from <https://www.geeksforgeeks.org/k-means-clustering-introduction/>

[6] Microsoft. (n.d.). Power BI: Business analytics tools. Retrieved from <https://www.microsoft.com/en-us/power-platform/products/power-bi>

[7] Apache Software Foundation. (n.d.). Apache Spark™: Unified analytics engine for big data. Retrieved from s<https://spark.apache.org/>

[8] U.S. Department of Justice. (n.d.). Crime data explorer (CDE). Retrieved from <https://cde.ucr.cjis.gov/>

[9] H. ToppiReddy, B. Saini, G. Mahajan. “Crime Prediction Monitoring Framework Based on Spatial Analysis”. Procedia Computer Science 132 (Jan. 2018), pp. 696–705.

[10] Mokhtar Mansour Salah and Kewen Xia. 2022. Big Crime Data Analytics and Visualization. In 2022 The 6th International Conference on Compute and Data Analysis (ICCDA 2022). Association for Computing Machinery, New York, NY, USA, 24–28. <https://doi.org/10.1145/3523089.3523094>

[11] M. Feng et al., ”Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data,” in IEEE Access, vol. 7, pp. 106111-106123, 2019, doi: 10.1109/ACCESS.2019.2930410.

[12] Chainey, Spencer Tompson, Lisa Uhlig, Sebastian. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. Security Journal. 21. 4-28. 10.1057/palgrave.sj.8350066.