CRIME DATA ANALYSIS AND VISUALIZATION USING MACHINE LEARNING

NAGA RAJITHA BHOGADI TRIBHUVAN REDDY KOTHAPALLY

Project repository

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]

The idea of "Crime Data Analysis and Visualization using Machine Learning" is to employ current technologies in a transformation. The project's general objective will be aimed at forecasting models and visualization techniques concerning crime evolution in several years and geographical areas, drawing on the Federal Bureau of Investigation (FBI) database that spans a number of years. Unlike conventional methods that target specific small areas, this study becomes the widest ever done by at least number of states and number of city authorities with data from all 50 states and over 7000 city authorities across the entire nation. [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. In our project, we have extended these hotspot techniques to cover all states and cities across the United States, providing a comprehensive nationwide analysis.[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. By considering Big data frameworks like spark, we intend to build the large scale crime data prediction tool.[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

Azure Cloud Services

- Azure Blob Storage: Storing raw data and clean datasets in a secure way.
- Azure Databricks - Core to the project in receiving ETL of the data set and the modelling and deployment of machine learning models.

Hardware Specifications

- This project consumes the Intel Neon standard_ds3_v2 CPU that has 4 cores and 14 GB of RAM.
- Storage: The strong 500 GB capacity is dedicated to reliable data management and model storage.

Programming Environment

- The data analysis and modelling exercise is done in Python 3.12.
- Another distributed computation frame in Databricks Notebook that connects to a computing cluster will be of benefit: Apache Spark 3.5.0 with Scala 2.12.

Data Processing & Machine Learning

- Pandas and NumPy: for data manipulation and numerical computations.
- Scikit-learn: for building and evaluating the machine learning model.

Data Visualization Tools

- Power BI: an essential component of designing interactive and dynamic dashboards that assist in visualizing crime and trends, alongside predictions.

DATASET

This project gets its information from the Uniform Crime Reporting Database from the Federal Bureau of Investigation (FBI). The period covered in research is from 2020 to 2023 with predictive analysis and visualization for data from all fifty states and more than 7,000 city agencies. It is broken into 40,000 data points split on 80-20 for train-test purposes. For the

most part, continuous features in the dataset include population, total offenses, and crimes against persons, such as assaults and homicides, crimes against property, such as arsons and robberies, and society crimes, which include weapon and drug violations. Categorical features include state names, an id of either a city or an agency, and key variables such as crime types, crime counts, and years of reporting.

The raw data reside in Azure Blob Storage in the Azure Cloud environment where preprocessing is done using Azure Databricks and Apache Spark.

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]

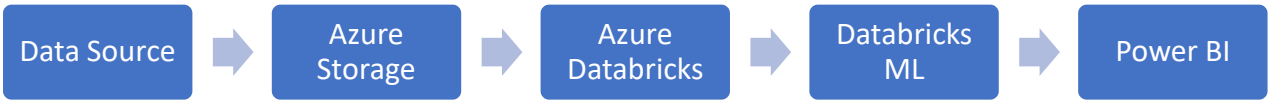


Figure 1: Process Workflow

2.DATA PROCESSING

In the United States, there are three major broadly known criminal justice reporting systems, one of which is most relied upon by the professionals in America in the areas of crime monitoring and crime-data assessment and public safety measures evaluation. The three include the National Crime Victimization Survey (NCVS), the Uniform Crime Report (UCR), and the National Incident-Based Reporting System (NIBRS). This study has relied on UCR and NIBRS databases which are all accessible through the FBI API, indicating that raw data is very genuine and authentic as it has been collected from the official website of FBI, and covers a few states from the years 2020-2023 and a few types of crime.[8]

Preparation of data is, therefore, one of the important steps in projects that take data as their drivers, ensuring that crude data makes it to a clean, organized, and usable form. Much data preprocessing work can thus be automated, scaled, and efficiently done in Azure Databricks as in the case of this project named: "Analysis and Visualization of Crime Data Using Machine Learning". It was an activity phase that prepared crime datasets for model and visualization activities involving collective efforts to address inconsistencies, cleansing data, creating useful features, and automating the entire process.

Azure Databricks, a platform that combines Apache Spark and scales computing resources to support it, executed activities at preparation. A cluster Standard_DS3_v2 with four cores and 14 GB was set in the environment. Set at 16.0. x-cpu-ml-scala2.12, the runtime environment was optimized for tasks such as machine learning. Pipelines were developed to automate repetitive processes so that an easy workflow can be established. The pipelines were modularized consisting of 2 main tasks-Dating, and ML_Tasks: Data Preprocessing to ensure seamless transition starting from the data import phase up to the deployment of machine learning models.

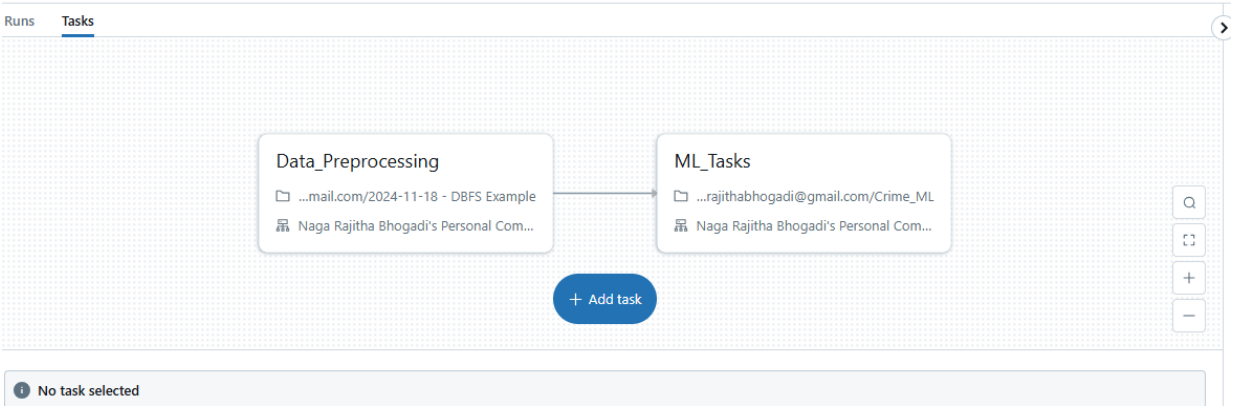
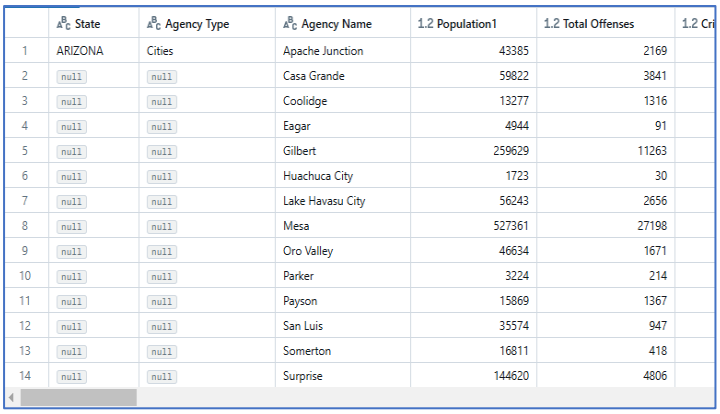


Figure 2: Automated Pipeline in Azure Databricks

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

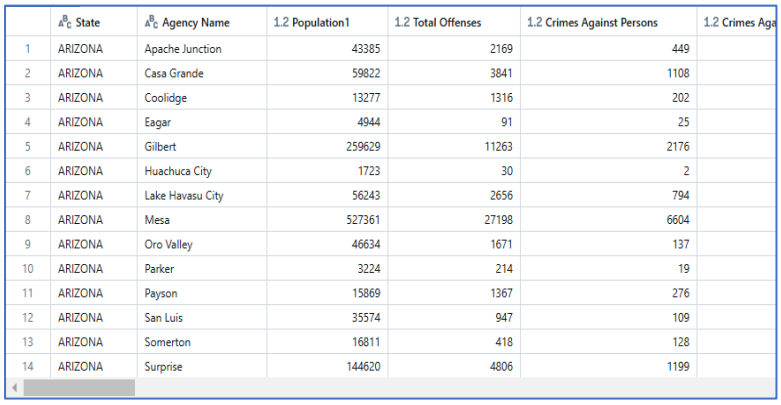
datasets contained information on various categories of crime, their occurrence sites and the periods within which these crimes occurred. Azure Databricks acted as the processing layer where the raw data was extracted and created for deep exploration. The design of the pipeline made the entire intake 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.



	State	Agency Type	Agency Name	1.2 Population1	1.2 Total Offenses	1.2 Cri
1	ARIZONA	Cities	Apache Junction	43385	2169	
2	null	null	Casa Grande	59822	3841	
3	null	null	Coolidge	13277	1316	
4	null	null	Eagar	4944	91	
5	null	null	Gilbert	259629	11263	
6	null	null	Huachuca City	1723	30	
7	null	null	Lake Havasu City	56243	2656	
8	null	null	Mesa	527361	27198	
9	null	null	Oro Valley	46634	1671	
10	null	null	Parker	3224	214	
11	null	null	Payson	15869	1367	
12	null	null	San Luis	35574	947	
13	null	null	Somerton	16811	418	
14	null	null	Surprise	144620	4806	

Figure 3: Before Data Pre-processing



	State	Agency Name	1.2 Population1	1.2 Total Offenses	1.2 Crimes Against Persons	1.2 Crimes Aga
1	ARIZONA	Apache Junction	43385	2169	449	
2	ARIZONA	Casa Grande	59822	3841	1108	
3	ARIZONA	Coolidge	13277	1316	202	
4	ARIZONA	Eagar	4944	91	25	
5	ARIZONA	Gilbert	259629	11263	2176	
6	ARIZONA	Huachuca City	1723	30	2	
7	ARIZONA	Lake Havasu City	56243	2656	794	
8	ARIZONA	Mesa	527361	27198	6604	
9	ARIZONA	Oro Valley	46634	1671	137	
10	ARIZONA	Parker	3224	214	19	
11	ARIZONA	Payson	15869	1367	276	
12	ARIZONA	San Luis	35574	947	109	
13	ARIZONA	Somerton	16811	418	128	
14	ARIZONA	Surprise	144620	4806	1199	

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.

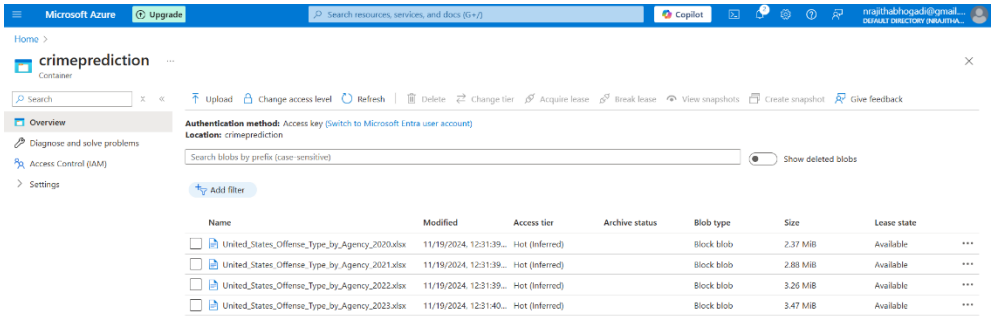


Figure 5: Azure Blob Storage container - Input & Output

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 gained great recognition among the famous and widespread techniques used in time-series forecasting to identify trends and seasonal movements from historical data. The older observations get lower exponential weights when compared with recent observations. This project involved the Exponential Smoothing Model, working towards predicting total crimes that might occur in 2024 as well as 2025. It is very simple in approach and can efficiently handle seasonality that allows studying patterns of crimes

across longer durations, forecasting based on predictions derived from crime data in earlier years to give relatively real inputs for advanced planning and decision making because it is such a suitable option for studying crime patterns.[4]

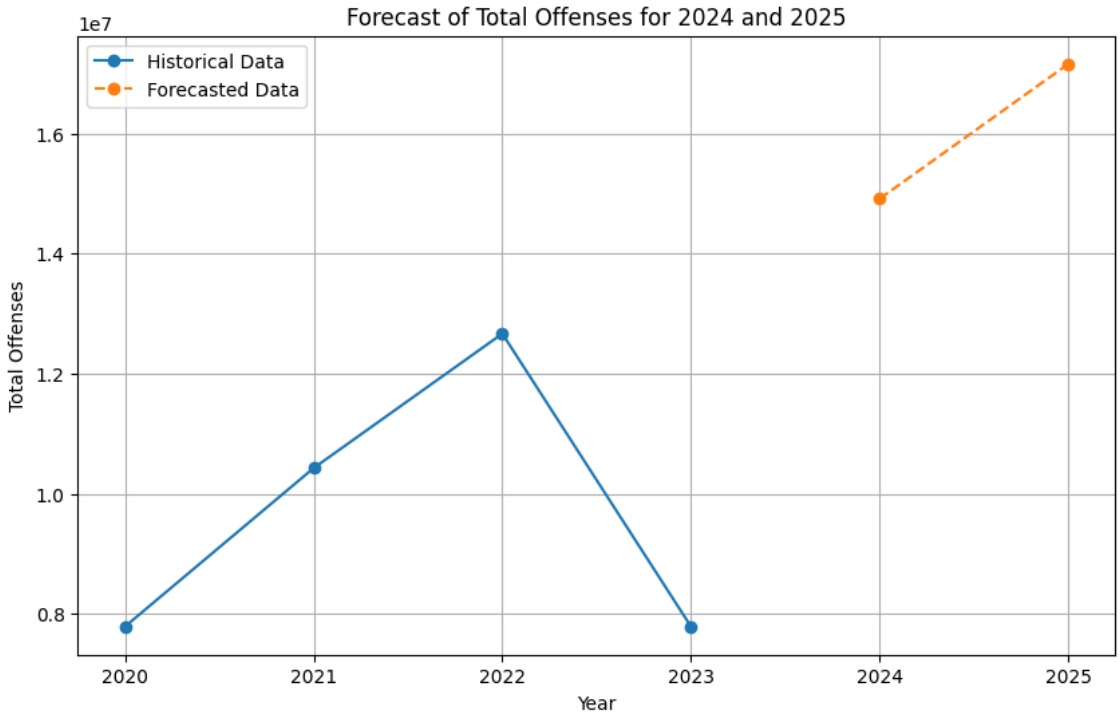


Figure 6: Forecast of Total Crime across USA in 2024 & 2025

The time-series forecast for offenses from the year 2020 to 2025 is captured in an Exponential Smoothing Model and shown in Figure 6. An actual historical data extract from 2020-2023 is represented by the blue solid line, which peaks total offenses in 2022, significantly drops afterward 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.

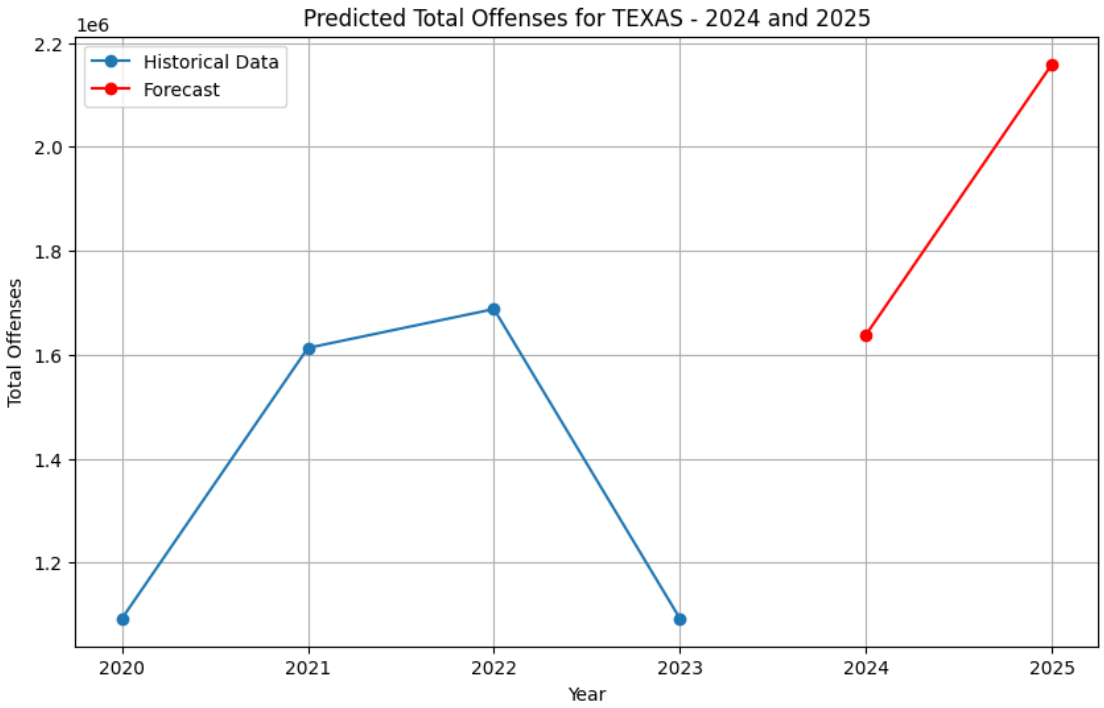


Figure 7: Forecast of Total Crime across State Texas in 2024 & 2025

```
Mean Absolute Error (MAE): 2222206.9621024146
Mean Squared Error (MSE): 14408064215234.096
Root Mean Squared Error (RMSE): 3795795.597135612
Mean Absolute Percentage Error (MAPE): 8.219975343591198
```

Figure 8: Exponential Smoothing Metrics

A number of indicators were used to draw performance of the exponential smoothing model. Mean absolute error (MAE) specifies a mean absolute difference between actual and forecast values besides the indicator of 2222206 in the modeling of really valuable data, which is troublesome with a high MAE number, as it shows large prediction errors. Strong differences or high mean square error (MSE) of 1.44 × 1010, punishing even larger mistakes, are among the diversities in forecasts. 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

A Random Forest Regressor is a supervised machine learning algorithm that predicts a number value using multiple decision trees during training and then averages their output values for accuracy and reliability. The work presented here concerns forecasting total crime offenses per state during the years 2024 and 2025 by employing concepts such as using Random Forest Regressor. It indeed holds an arsenal of weaponry to deal with the 'robustly' complicated data sets comprising of highly interrelated features with non-linear correlations; thus, this approach will serve as a bulwark for processing all the crime data information. This thus brings great potential in identifying rich crimes geography as it gives an understanding of geo-spatial distribution of crimes that would help interested parties organize interventions and distribute resources more efficiently.[4]

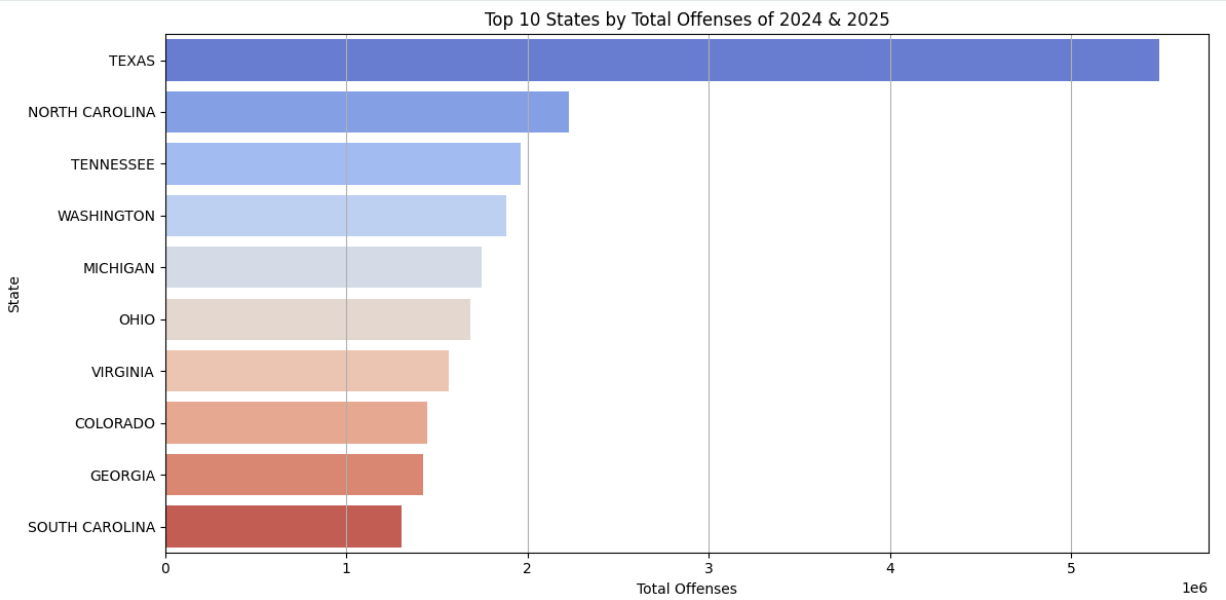


Figure 9: Geographical Distribution of Crimes of 2024 & 2025

```
Model Performance Metrics:
Mean Absolute Error (MAE): 82.41855466722448
Mean Squared Error (MSE): 478317.0507151891
Root Mean Squared Error (RMSE): 691.6046925196424
R-squared (R2 Score): 0.98
```

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.[4]

Model Accuracy: 0.71
Classification Report:

	precision	recall	f1-score	support
Arson	0.892593	0.828179	0.859180	291.0
Assault\nOffenses	0.667251	0.751810	0.707011	3038.0
Bribery	0.000000	0.000000	0.000000	0.0
Fraud\nOffenses	0.763359	0.350877	0.480769	285.0
Larceny/\nTheft\nOffenses	0.735114	0.702121	0.718239	3253.0
Rape	0.500000	0.333333	0.400000	6.0
Robbery	1.000000	0.500000	0.666667	2.0
Weapon\nLaw\nViolations	0.767857	0.367521	0.497110	117.0
micro avg	0.708667	0.708667	0.708667	6992.0
macro avg	0.665772	0.479230	0.541122	6992.0
weighted avg	0.713755	0.708667	0.705559	6992.0

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%.

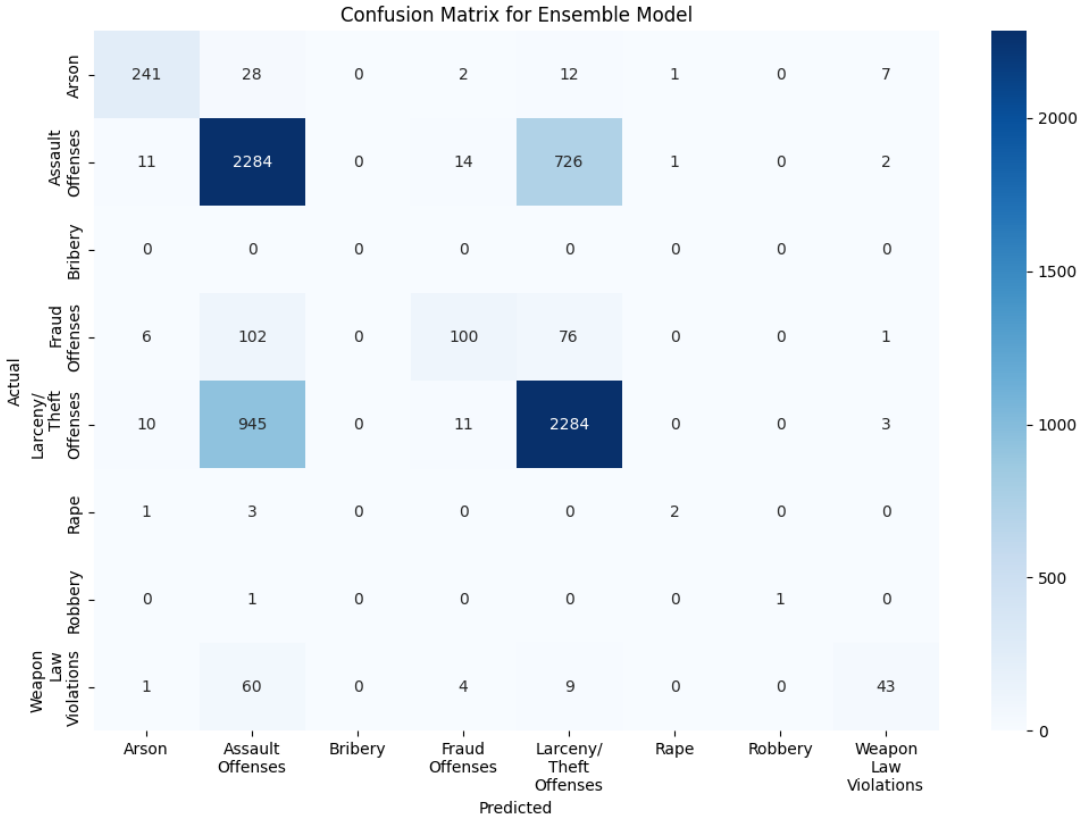


Figure 12: Confusion Matrix of Ensemble Model

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.[4]

Random Forest Model Accuracy: 0.69
Classification Report:

	precision	recall	f1-score	support
Arson	0.923810	0.970000	0.946341	100.000000
Assault\nOffenses	0.446602	0.460000	0.453202	100.000000
Bribery	1.000000	1.000000	1.000000	1.000000
Fraud\nOffenses	0.756410	0.590000	0.662921	100.000000
Larceny/\nTheft\nOffenses	0.532110	0.580000	0.555024	100.000000
Rape	1.000000	0.833333	0.909091	6.000000
Robbery	1.000000	0.500000	0.666667	2.000000
Weapon\nLaw\nViolations	0.766355	0.820000	0.792271	100.000000
accuracy	0.685658	0.685658	0.685658	0.685658
macro avg	0.803161	0.719167	0.748190	509.000000
weighted avg	0.690626	0.685658	0.685194	509.000000

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.

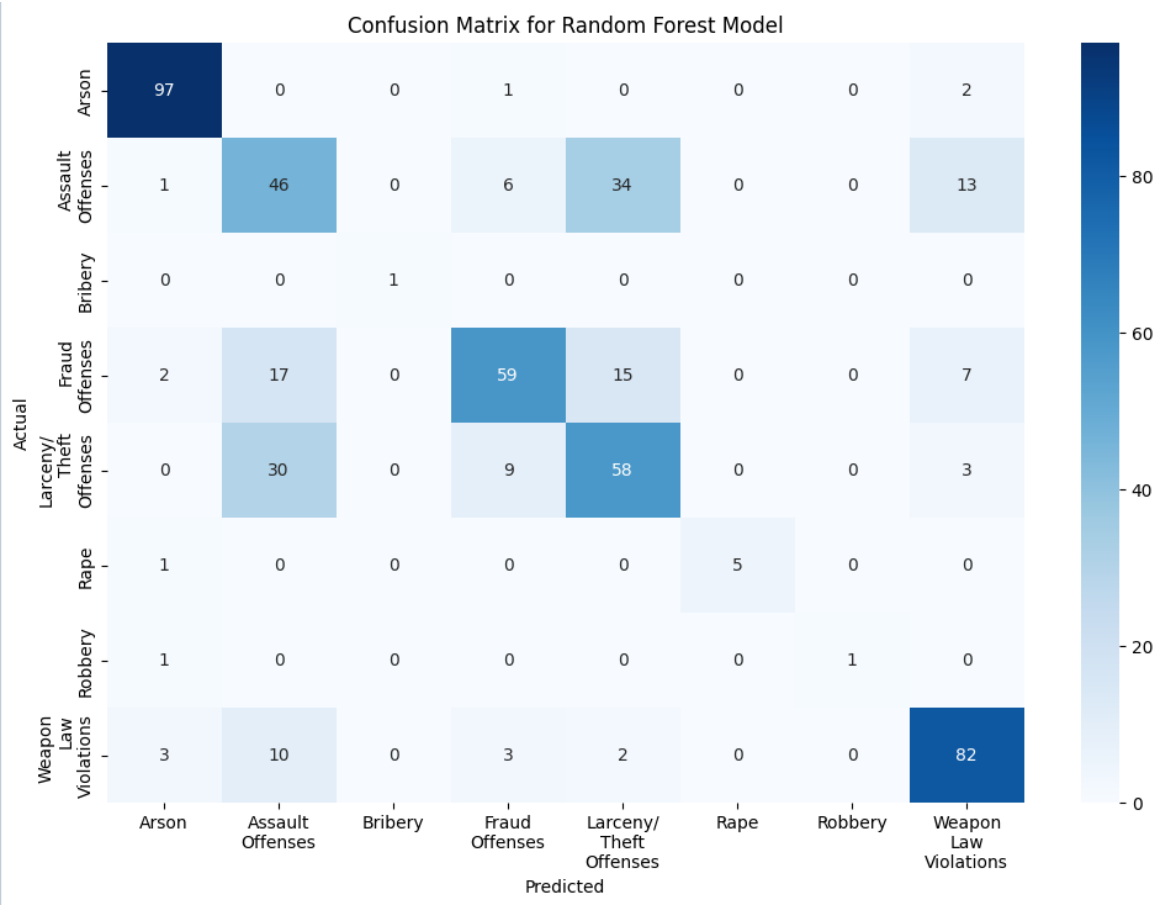


Figure 14: Confusion Matrix for Random Forest Classifier Model

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.[5]

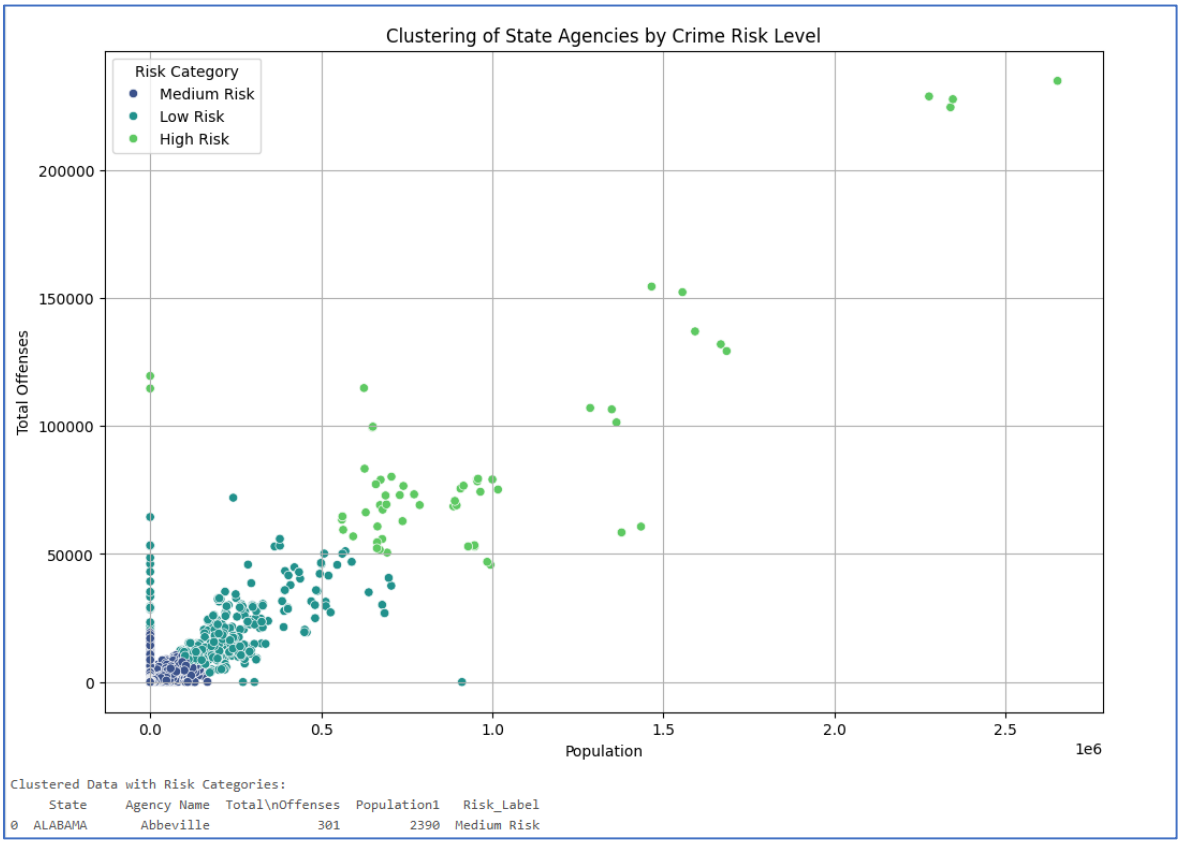


Figure 15: Categorizing each city into zones

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.

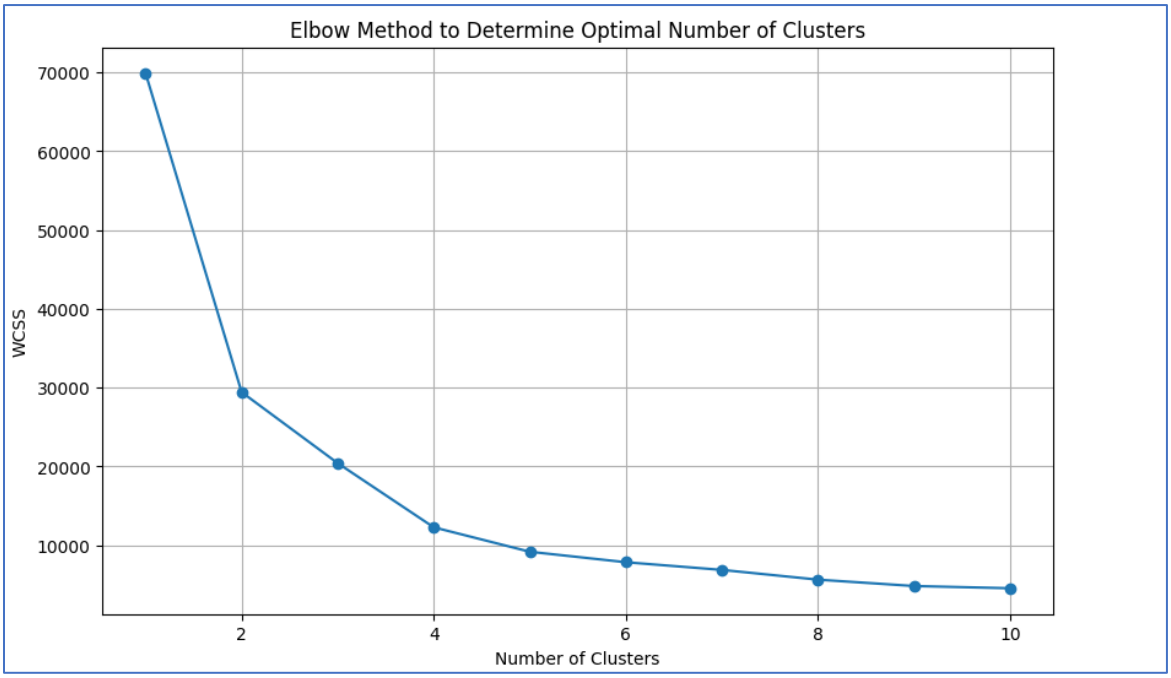


Figure 16: Elbow Method to determine Clusters

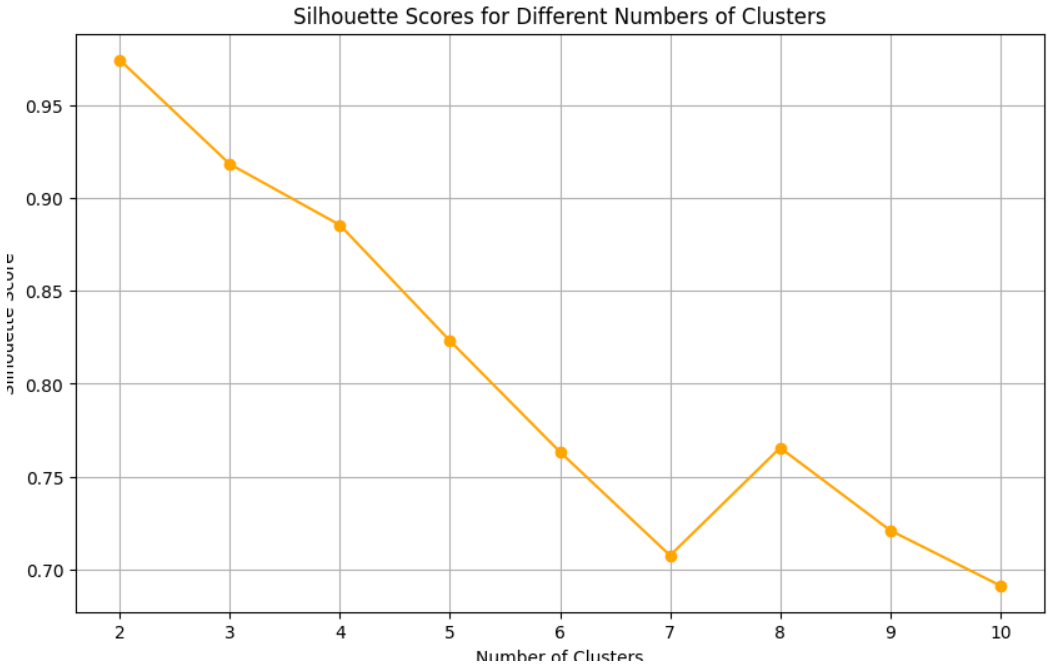


Figure 17: Silhouette Score to determine Clusters

The Elbow method and silhouette scores are two metrics generally used to measure the result of clustering techniques. According to the elbow method, the ideal number of clusters is determined by plotting WCSS against an increased number of clusters and identifying an elbow point where increase in number of clusters no longer adds appreciably to WCSS. Clearly, three clusters would seem to give the best solution, as drawn from the sharp decrease in WCSS until that point, then flattening. Similarly, silhouette scores are determined on multiple

counts of cluster number, the higher the count, the more defined the clusters. This was further ratified as Silhouette scores peak at three clusters.

6.Predicted Outputs:

The outputs of the machine learning models, which have been implemented, are saved in structured data sets within a secured Azure Blob Container. Thus, the scalable accessible data can be integrated into various visualization applications like Power BI. Some of the significant outputs of the data sets are: Future crime predictions, for example, 2024 and 2025, predict about which types will occur mostly by city and by state and categorized risk levels of areas depending on their crime intensity. Machine Learning techniques have led the outputs organized into structured datasets that are now stored at an Azure Blob Container. This data scales and becomes accessible, integrating as well visualization applications such as Power BI. The datasets carry some of the following key results: Future crime 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.

State	2024 Prediction	2025 Prediction
ALABAMA	424754	510933
ALASKA	19439	20924
ARIZONA	141501	143469
ARKANSAS	224958	223936
CALIFORNIA	1706711	2228973
COLORADO	358732	357767
CONNECTICUT	120594	120929
DELAWARE	67303	67207
DISTRICT OF COLUMBIA	37469	44451

Figure 18: 2024 & 2025 Total Offenses

State	Agency Name	TotalOffenses	Population1	Risk_Category	Risk_Label
ILLINOIS	Chicago	234722	2652124	2	High Risk
TEXAS	Houston	227553	2346155	2	High Risk
TEXAS	Houston	227553	2346155	2	High Risk
TEXAS	Houston	224406	2339252	2	High Risk
TEXAS	Houston	228657	2276533	2	High Risk
NEVADA	Las Vegas Metropolitan Police Dep	129232	1685021	2	High Risk

Figure 19: Categorization based on Risk

An automated pipeline constitutes a common approach for the integration of machine learning models, which consists of including ensemble models, clustering algorithms, and predictive regressors. This entire sequence has been thoroughly automated, and no manual interference is needed for the delivery of continuously updated results with the entry of new data. Hence, stakeholders can then obtain the real-time insights along with predictions to

reflect future happenings and other in-the-moment actionable insights, such as crime trends with high-risk areas. Thus, it is an automated pipeline that covers the entire process, from data preprocessing to model execution and storing the outcome, thereby diminishing manual overheads while simultaneously improving system efficiency.

Year	Population1	TotalOffenses	State_Encoded	City_Encoded	Most_Common_Crime_Encoded	Most_Common_Crime
2024	1555812	152222	44	5911	4	Larceny/ Theft Offenses
2025	1555812	152222	44	5911	4	Larceny/ Theft Offenses
2024	1676491	130558	34	4180	1	Assault Offenses
2025	1676491	130558	34	4180	1	Assault Offenses
2024	1344417	117647.5	19	1327	4	Larceny/ Theft Offenses
2025	1344417	117647.5	19	1327	4	Larceny/ Theft Offenses
2025	644065.5	103400	48	4771	1	Assault Offenses
2024	644065.5	103400	48	4771	1	Assault Offenses
2025	931547	91312.4	49	3509	4	Larceny/ Theft Offenses
2024	931547	91312.4	49	3509	4	Larceny/ Theft Offenses
2024	654924.25	79191.75	28	1947	1	Assault Offenses
2025	654924.25	79191.75	28	1947	1	Assault Offenses
2024	957116	79076.5	39	1264	4	Larceny/ Theft Offenses
2025	957116	79076.5	39	1264	4	Larceny/ Theft Offenses
2025	911598.5	76102	41	1569	4	Larceny/ Theft Offenses
2024	911598.5	76102	41	1569	4	Larceny/ Theft Offenses
2025	765118.5	72167.5	53	6732	4	Larceny/ Theft Offenses
2024	765118.5	72167.5	53	6732	4	Larceny/ Theft Offenses
2024	686535.25	70584	48	4828	4	Larceny/ Theft Offenses
2025	686535.25	70584	48	4828	4	Larceny/ Theft Offenses
2024	890906.25	69760.5	20	3669	4	Larceny/ Theft Offenses

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.

7.Experimentation and Findings:

In the course of developing the project, numerous machine-learning algorithms have been tested in order to determine the superior machine-learning models available for performing forecasting, classification, and clustering tasks. The algorithms selected were chosen as they outperformed and proved to be far more reliable than other tested models.

Exponential Smoothing has been found to be the best model for forecasting, time series whose precision score is 91%. It beats its alternatives such as SARIMAX and ARIMA. A typical example is SARIMAX, which was thought of by the researchers, but it failed terribly as it registered a score of about 60% and could not understand the variations in the irregularities as well as the seasonal anomalies in the data on crime. Almost all required huge parameter tuning, and it could not deal with long-term seasonality, compared to Exponential Smoothing.

Random Forest Regressor proved to be the best at predicting total crime counts with R^2 value of 0.98, which means that the model explained 98% of all variance in data. This new technique was a huge improvement over previous models like Gradient Boosting Regressor, which were

accurate but allowed for much greater overfitting along with a lot more work on hyperparameter tuning to get the models to behave well. Thus, it is a much more rounded application because random forest can handle high-dimensional and very complicated relationships.

The accuracy of the ensemble model comprises a random forest classifier, gradient boosting classifier, and the XGBoost classifier for classification tasks was 71%, leading all straightforward models capable of giving their predictions in the validation phase, such as logistic regression which could not deal with imbalanced datasets, and support vector machines (SVM), which were overly costly on computational resources and not scalable for high dimensions.

K-Means Clustering has successfully categorized the regions into high, moderate, and low-risk areas in performing spatial analyses. Although the hierarchical clustering method has been explored, it turned out to be very intensive in computation and not very interpretable; hence K-Means has been adopted as the preferred method for finding crime hotspots.

The comparisons justify the careful, evidence-based procedure followed in selecting models appropriate for the task, prioritizing accuracy, scalability, and computational efficiency. Hence, it ensured that its selected models were predictive and actionable by outperforming alternatives in performance metrics and practical utility.

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.[6]

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

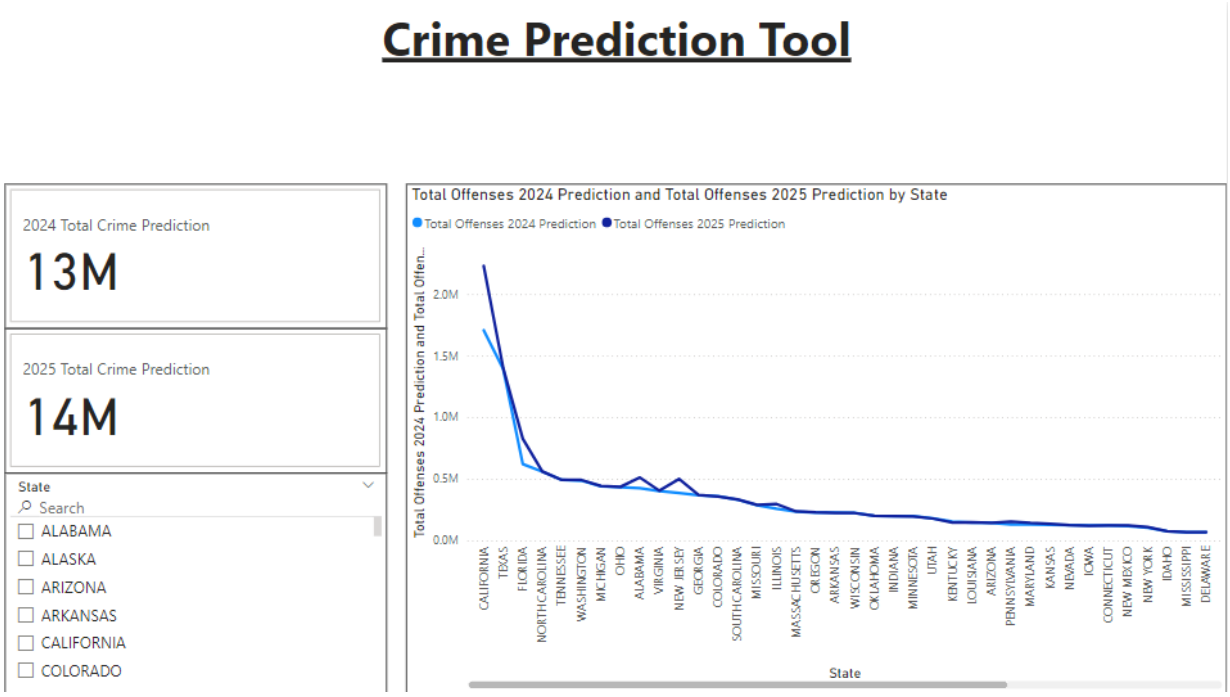


Figure 21: 2024 & 2025 Crime offenses by State Line Graph

The graphical way depicts the future trend of the crime rates across different states and years. This graph helps a lot in identifying trends across them. Using filter we can look at different states, cities and years data.

2.Crime Hotspot Maps

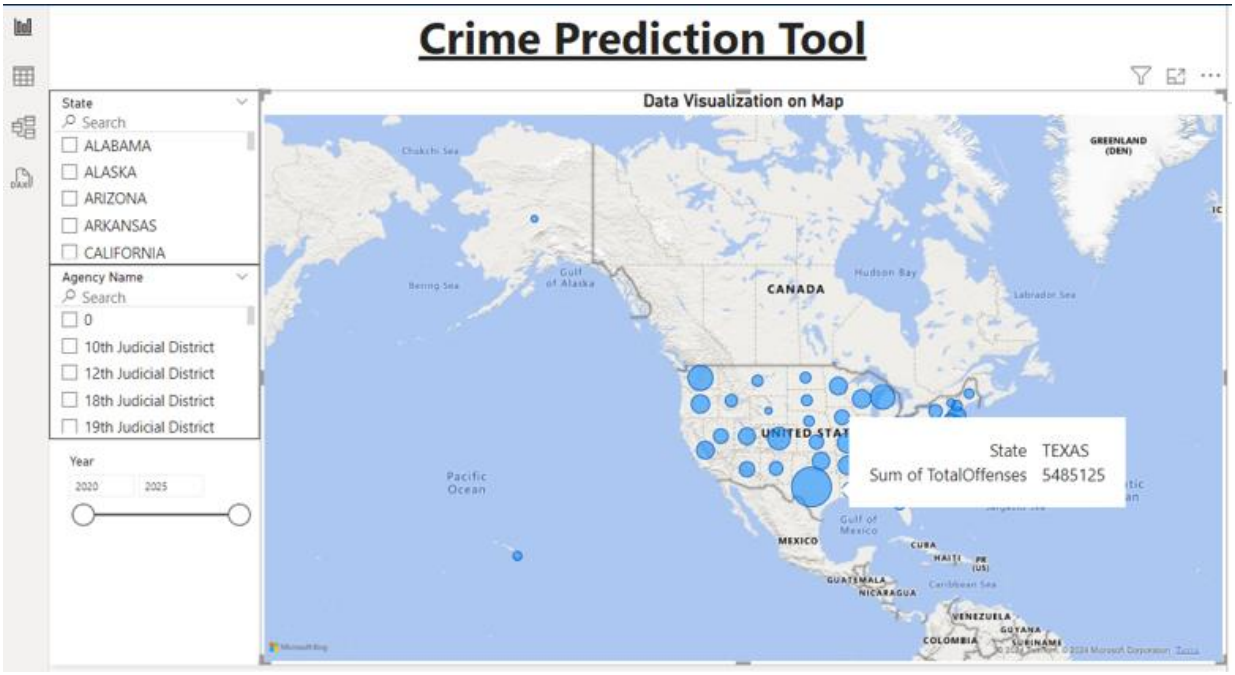


Figure 22:Crime Hotspot Maps By offenses across US

The crime data is derived from the dataset and is based on the state, city and total crimes that occurred. The display is based on the hotspot mapping approach. The crime is depicted across the US in Fig. 4 using point maps. The amount of offences will vary depending on the size and intricacy of the colour. It also gives a summary of the information. The data can also be represented based on the year and the data frame can also be adjusted based on the requirement and identify the crime trends across states over the years.

Figure 23: Crime Hotspot Maps by Offenses across Cities

3.Crime Distribution Visualization

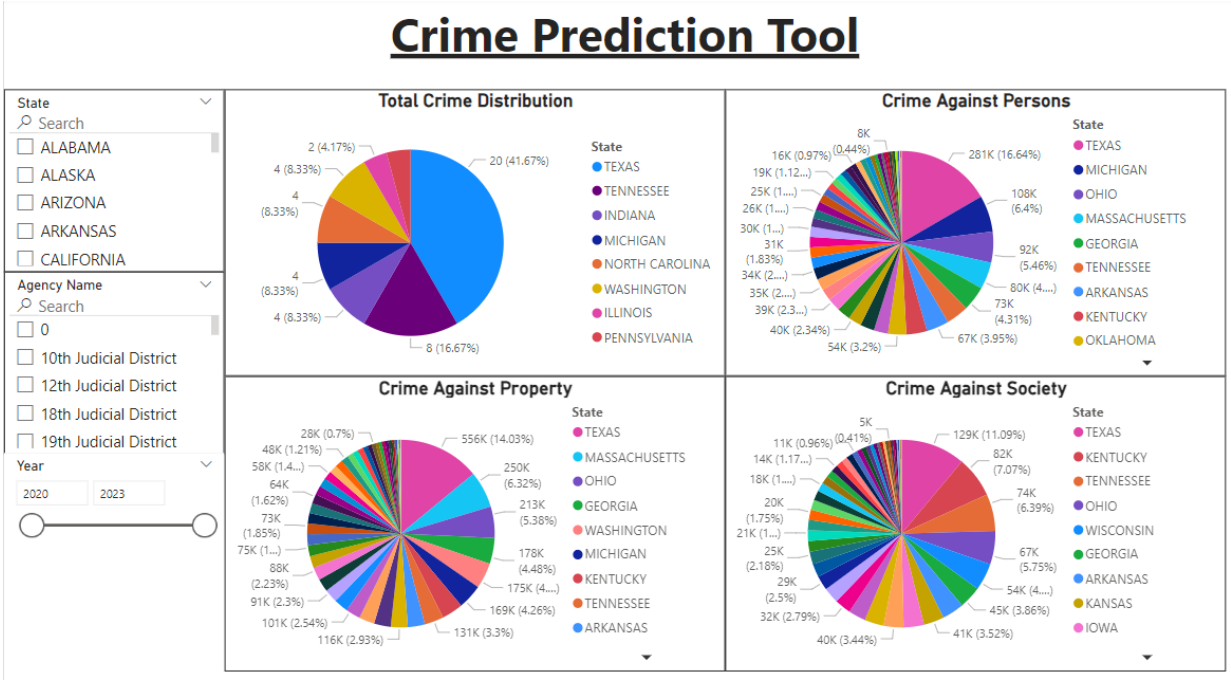


Figure 24: Crime Distribution across US by crime types

Based on the crime kinds, state name, and year trends, the dataset is used to produce the data needed for the graphics. The analysis of the sorts of crimes that commonly occur in a given location and the improvement of security measures based on those crimes are both aided by this type of depiction of crimes based on crime categories across different states.

4.Crime Risk Areas

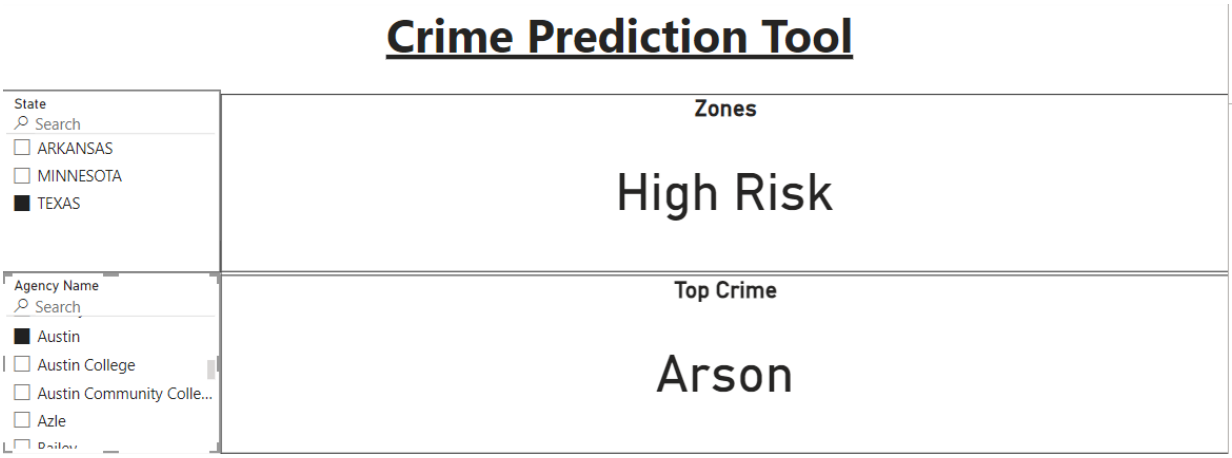


Figure 25: Risk Zone Prediction by City and State

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 professors Mentor: Dr. Samira Zad, Instructor: Dr. Ananda Mondal, Panel Member: Dr. Niemah Osman 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.

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 shttps://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.