Crime Pattern Analysis and Prediction Using Machine Learning and Data Mining Technique

[1]Dr. Sheela **J**ayachandhran
Assistant Professor (Sr.Grade 2)
School of Computer Science and Engineering
VIT-AP University,
Amaravati, Andhra Pradesh, India
sheela.j@vitap.ac.in

[2]S Dhanya Ratna Madhuri
Student
School of Computer Science and Engineering
VIT-AP University,
Amaravati, Andhra Pradesh, India
ratnamadhuri.22bce8698@vitapstudent.ac.in

[3]Kommineni Chandhana
Student
School of Computer Science and Engineering
VIT-AP University,
Amaravati, Andhra Pradesh, India
chandhana.22bce7963@vitapstudent.ac.in

***Abstract*—** Crimes continue to be a pressing social issue, posing serious threats to public safety and social stability. With rising urbanization and population density, identifying crime trends and predicting criminal activities has become increasingly vital for effective law enforcement. In this research, historical crime data has been analyzed using a combination of supervised and unsupervised machine learning techniques. The system is trained on a comprehensive dataset incorporating various crime types such as theft, assault, burglary, and more. To uncover hidden patterns and predict future crime occurrences, we employed algorithms including Random Forest, Logistic Regression, Decision Tree, AdaBoost, SVM, XGBoost, K-Nearest Neighbors (KNN), and Naive Bayes. Additionally, clustering (K-Means) and association rule mining (Apriori) were used for crime pattern detection and relationship analysis among variables. The results demonstrate that machine learning models can significantly enhance the ability to predict crime-prone areas and assist law enforcement agencies in better resource allocation and crime prevention strategies.

Keywords—Crime prediction, K-means, KNN, Naive Bayes, Random Forest, Association Rule, Apriori, Classification, Clustering, Machine Learning.

I. INTRODUCTION

Crime remains one of the most significant challenges in maintaining law and order across growing urban environments. As populations expand and urbanize, the dynamics of crime shift, making it difficult to detect evolving patterns through traditional methods alone. The integration of machine learning and data mining techniques into crime analysis introduces new opportunities for understanding criminal behavior, identifying hotspots, and proactively preventing incidents.

Machine learning allows for the construction of predictive models by training on historical datasets that include various types of criminal activity such as murder, robbery, burglary, and kidnapping. These models offer the potential to detect hidden patterns, forecast crime-prone areas, and assist authorities in developing effective countermeasures. Previous studies have successfully demonstrated the power of these techniques. For instance, Yadav et al. (2017) applied K-Means clustering and Apriori association rule mining to extract correlations between crime-related attributes, while also using Naive Bayes and regression for classification and prediction tasks.Their approach revealed valuable insights into the relationships between different crime factors and helped categorize areas based on the intensity of criminal activity.

Similarly, Sattar et al. (2021) focused on predicting crime rates using multiple algorithms, including K-Nearest Neighbors (KNN), Naive Bayes, and Linear Regression. Their study involved preprocessing field-collected data and classifying it into actionable insights using time, age, and gender as key factors. The highest accuracy was achieved using KNN, which proved effective in identifying potential hotspots and predicting patterns across regions

Building upon these foundations, our study leverages a wide range of supervised and unsupervised learning techniques including Random Forest, Logistic Regression, Decision Tree, AdaBoost, Support Vector Machine (SVM), XGBoost, KNN, and Naive Bayes. Additionally, K-Means clustering is used to uncover natural groupings within the dataset, while the Apriori algorithm reveals associations between various crime-related features. By comparing model performances and analyzing patterns, this research aims to develop a comprehensive predictive system that can assist law enforcement agencies in real-time decision-making and resource deployment.

II. LITERATURE REVIEW

The application of machine learning and data mining in crime analysis has garnered increasing attention due to its ability to handle large volumes of data and extract meaningful insights. Several studies have demonstrated the effectiveness of these techniques in identifying crime patterns, classifying crime types, and predicting future criminal activities.

Yadav et al. (2017) proposed a system that utilizes a combination of clustering, association rule mining, and classification to analyze crime data sourced from legitimate government records spanning 14 years (2001–2014). Their approach involved four primary techniques: **K-Means clustering**, **Apriori algorithm**, **Naive Bayes classification**, and **Regression analysis**. K-Means was used to group states or regions based on the number of crimes and arrests, categorizing areas into “high” or “low” crime clusters. These clusters were then input into the Apriori algorithm to discover associations between attributes such as arrests, convictions, and acquittals. Their classification model, based on Naive Bayes, was tested on features like age group, gender, and crime type, demonstrating the algorithm's ability to generalize from historical trends.

On the other hand, Sattar et al. (2021) focused on building a predictive framework using **K-Nearest Neighbors (KNN)**, **Linear Regression**, and **Naive Bayes** for crime rate prediction in urban areas, particularly Dhaka. Their research emphasized preprocessing raw data into a machine-readable format by converting qualitative features (e.g., gender, month, region) into binary values. By implementing multi-linear regression and classification models, they were able to determine the correlation between perpetrator characteristics (age, gender) and crime occurrence. KNN outperformed other models with the highest prediction accuracy of approximately 77%, proving effective for crime pattern recognition across spatial zones. Their results also highlighted that crime incidents were significantly higher in specific periods (January–April and September–December) and in certain urban zones.

Both studies affirm the potential of machine learning in identifying trends and forecasting crime. They also emphasize the importance of combining multiple techniques (supervised and unsupervised) for comprehensive crime analysis. Our current work extends these findings by integrating a broader set of models including **Random Forest, AdaBoost, SVM, XGBoost**, and **Association Rule Mining**, offering a comparative perspective on algorithmic effectiveness in crime prediction.

III. PROPOSED DESIGN

A. Dataset Description

The dataset utilized in this study was sourced from Kaggle and is titled *"Crime Data"* by Isha Jangir. It comprises comprehensive records of criminal activities reported from 2020 to the present. The data includes a wide range of attributes that are essential for analyzing crime patterns and building predictive models. Each record represents an individual crime incident, detailing when and where it occurred, the type of offense, whether an arrest was made, and if the crime was domestic in nature.This data set consists of 12 male voices and 12 female voices which is very helpful in gender classification so that equal opportunity is given to both genders

Key features of the dataset include the date and time of the crime, the primary type of offense (such as theft, assault, battery, or burglary), and the specific location description (e.g., street, residence, or parking lot). Additional fields like Community Area and District help provide spatial context, while the presence of latitude and longitude values allows for geospatial visualization and hotspot analysis. The dataset also contains boolean fields such as Arrest and Domestic that help categorize crimes more effectively.

To enhance the dataset’s utility for machine learning, temporal features were extracted from the date field, including the year, month, day of the week, and time of day. Categorical features such as crime type and location were encoded appropriately to prepare them for input into classification algorithms. Any missing values—particularly in geographic or descriptive fields—were addressed through imputation techniques or exclusion, depending on the extent of the missing data.

Overall, the dataset provides a rich foundation for modeling and predicting crime patterns. Its variety of attributes enables thorough exploratory analysis, clustering of crime-prone areas, and the discovery of underlying associations using data mining techniques. The depth and structure of the dataset make it highly suitable for both supervised and unsupervised machine learning applications in crime analytics.

B. Data Preprocessing

Data preprocessing is a crucial step in preparing raw data for analysis and modeling. In this study, the crime dataset obtained from Kaggle underwent several preprocessing steps to ensure data quality, consistency, and suitability for machine learning algorithms. The goal was to transform the original dataset into a clean, structured format that would facilitate accurate predictions and meaningful pattern detection.

The first step involved **data cleaning**, where missing values, duplicates, and inconsistencies were addressed. Records with missing location data (e.g., latitude or longitude) or key identifiers were either imputed using suitable techniques (such as K-Nearest Neighbors imputation) or removed if the missingness was significant. Duplicated entries were also eliminated to avoid biased learning during model training.

Next, **feature extraction and transformation** were applied to enhance the predictive power of the dataset. The Date column was parsed to derive new temporal features such

as Year, Month, Day of the Week, and Hour of the Day, which helped in capturing time-based crime patterns. Categorical variables such as Primary Type, Location Description, and Community Area were encoded using Label Encoding or One-Hot Encoding, depending on the modeling requirement. This step allowed the machine learning models to interpret and utilize non-numeric attributes effectively.

Further, **scaling and normalization** were performed on numerical features like crime counts, victim counts, and time-related variables to ensure that all inputs had a uniform scale. This was particularly important for algorithms such as K-Nearest Neighbors and Support Vector Machines, which are sensitive to feature magnitudes. StandardScaler and MinMaxScaler were applied selectively based on the distribution of each feature.

To prepare for clustering and association rule mining, additional processing included grouping crimes by type, region, and time intervals. These aggregated datasets helped in identifying hotspots and frequent crime patterns. Finally, the preprocessed data was split into training and testing sets to evaluate the performance of various classification models under consistent and reproducible conditions.

Through systematic data preprocessing, the dataset was transformed into a robust and analyzable format, laying the foundation for accurate crime prediction and insightful data mining

C. Clustering layer

The clustering layer in this study serves as an unsupervised learning component designed to identify hidden patterns and groupings within the crime data, particularly to detect potential **crime hotspots** and categorize regions based on their crime profiles. The primary algorithm used for clustering in this model is **K-Means**, which partitions the dataset into a predefined number of clusters based on feature similarity. This step is crucial in understanding the spatial distribution and frequency of criminal activities, offering a foundational analysis that can enhance the predictive capabilities of subsequent supervised models.

To begin with, essential features such as Crime Type, Location (latitude and longitude), Time of Occurrence, and Community Area were selected and standardized using scaling techniques. Standardization ensures that all features contribute equally to the clustering process, avoiding bias toward features with larger numerical ranges. The **Elbow Method** and **Silhouette Score** were employed to determine the optimal number of clusters (K). These metrics helped in evaluating intra-cluster cohesion and inter-cluster separation, ensuring that the clusters formed were both distinct and meaningful.

Once the optimal K value was identified, the K-Means algorithm was applied to group crime incidents into clusters. Each cluster represented a unique crime profile, often based on geography (e.g., certain districts), time (e.g., night vs. daytime crimes), or type (e.g., theft-heavy vs. assault-heavy regions). The resulting clusters were visualized using scatter plots and geospatial maps, allowing for intuitive interpretation and validation. In particular, **heatmaps** generated from latitude and longitude data provided visual confirmation of high-crime zones.

The clustering results played a dual role in the overall model architecture. First, they offered **descriptive insights** into where and when crimes are most likely to occur, aiding policymakers and law enforcement in resource allocation. Second, the cluster labels were integrated as an additional feature in the supervised classification models, enabling the classifiers to leverage the spatial-temporal groupings identified by the unsupervised process. This hybrid approach enhanced the performance of crime prediction models by incorporating both labeled and unlabeled learning perspectives.

In summary, the clustering layer provided a powerful unsupervised analytical tool for discovering latent crime structures in the data, facilitating both visualization and predictive enrichment within the overall crime rate prediction framework.

D. Classification Layer

The classification layer is the core predictive component of this study, tasked with forecasting the likelihood or category of future crimes based on historical patterns. After preprocessing and optional clustering, supervised machine learning algorithms were applied to classify crimes using structured features such as Crime Type, Location, Time of Day, Day of the Week, District, Domestic, and Arrest. The goal was to develop accurate, interpretable models capable of assisting law enforcement in anticipating crime occurrences and optimizing preventive strategies.

Multiple classification algorithms were evaluated in this study to compare their predictive performance. These included **Logistic Regression**, **Decision Tree**, **Random Forest**, **Support Vector Machine (SVM)**, **K-Nearest Neighbors (KNN)**, **Naive Bayes**, **XGBoost**, and **AdaBoost**. Each model was trained on a dataset that was split into training and testing sets using a standard 80/20 ratio. To ensure fairness and prevent overfitting, techniques such as **cross-validation**, **hyperparameter tuning** (via GridSearchCV or RandomizedSearchCV), and **stratified sampling** were employed.

Features used for classification were carefully engineered to capture both spatial and temporal dynamics. For example, the Date column was transformed into features like Hour, Day of Week, and Month, allowing the model to detect temporal trends in crime occurrences. Categorical attributes like Primary Type and Community Area were encoded using one-hot encoding or label encoding to make them suitable for model input. Additionally, cluster labels obtained from the previous unsupervised learning step were integrated into the feature set to improve context awareness.

Performance of the classifiers was evaluated using standard metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **confusion matrix**. Among the models, ensemble techniques like **Random Forest** and **XGBoost** outperformed simpler classifiers due to their ability to capture complex interactions and handle imbalanced data more effectively. These models also provided feature importance rankings, which were useful in identifying the most influential variables for crime prediction.

Overall, the classification layer not only achieved high predictive accuracy but also offered actionable insights into which factors most strongly correlate with various crime

types. The combination of diverse algorithms, robust evaluation techniques, and engineered features resulted in a reliable classification system that can serve as a practical decision-support tool for crime prevention and resource deployment.

E. Association Rule Mining Layer

The Association Rule Mining layer in this study was employed to uncover frequent patterns and relationships among crime attributes using the **Apriori algorithm**. This unsupervised data mining technique was applied primarily to categorical variables such as Crime Type, Location Description, Community Area, and Time of Day. By identifying co-occurring patterns, this layer helped reveal hidden associations in the data that may not be immediately obvious through classification alone. For example, rules like "If the crime occurs at night and in a residential area, then it is likely to be burglary (confidence: 0.85)" were extracted. Metrics such as **support**, **confidence**, and **lift** were used to evaluate the strength and relevance of the generated rules. These insights aid in understanding situational crime triggers and can inform both policy development and real-time crime prevention strategies.

F. Prediction Layer

The prediction layer acts as the final decision-making component of the model. After learning patterns from the training data, this layer uses the trained classifiers to predict the **type of crime**, the **likelihood of arrest**, or whether a **crime is likely to be domestic** in nature based on a set of input features. The prediction process is informed by all previously derived layers, including the clustering labels and association patterns. In some implementations, this layer also supports **probabilistic outputs**, offering not just the predicted class but also a confidence score associated with each prediction. This layer is critical in operationalizing the model, enabling forward-looking assessments such as forecasting crime risks in a particular area at a given time.

G. Evaluation and Optimization

Model performance was assessed using a range of evaluation metrics including **accuracy**, **precision**, **recall**, **F1-score**, and **Area Under the ROC Curve (AUC-ROC)**. These metrics ensured a balanced view of model effectiveness, especially in handling imbalanced crime categories. Confusion matrices were used to diagnose misclassifications, and **cross-validation** was implemented to verify model stability across different subsets of data. To improve the performance of the predictive models, **hyperparameter tuning** was conducted using techniques such as **GridSearchCV** and **RandomizedSearchCV**. Additionally, **feature selection** and **dimensionality reduction** (using methods like PCA) were explored to reduce noise and enhance computational efficiency. Models like Random Forest and XGBoost were further optimized based on feature importance, which also provided interpretability to the predictions

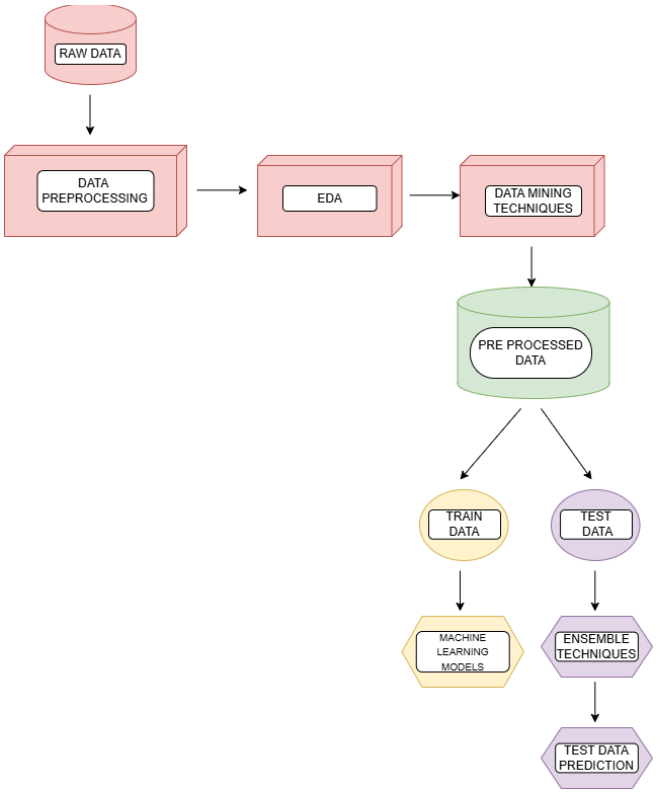


Fig. 1. Support Vector Machine (SVM) Model Architecture for Gender Prediction

H. Deployment and Visualization

To make the system accessible and actionable, a prototype **web-based interface** was developed using **Streamlit**. This platform allowed users—such as law enforcement agencies and policy analysts—to interact with the model by entering key input features and viewing predicted outputs. Visualization tools were integrated into the interface to present crime clusters on **interactive maps**, plot temporal trends using **line graphs**, and display **crime frequency distributions** using bar charts and heatmaps. For geospatial visualizations, **Folium** and **Plotly** were used, enabling dynamic exploration of crime hotspots. The integration of data mining, machine learning, and visual analytics provided a comprehensive decision-support system that not only predicts crimes but also explains them in a user-friendly manner.

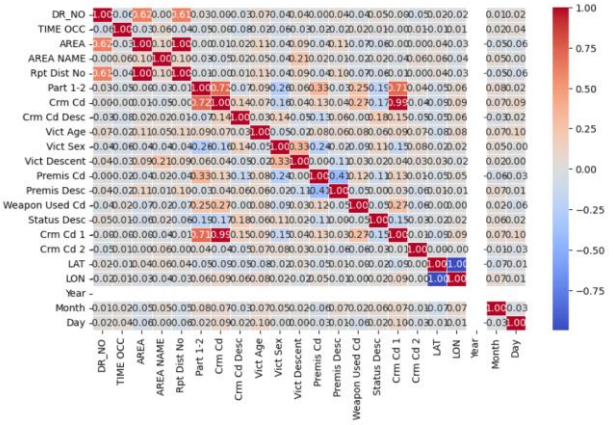
IV. VISUALIZATIONS (EDA)

*A. **CORREALATION HEATMAP***

In this study, we conducted a correlation analysis to explore the relationships between various numerical variables in the dataset. The correlation matrix was calculated to identify the strength and direction of linear relationships between pairs of features. A heatmap was then generated using Seaborn's heatmap function, which visually represents the correlation values. This heatmap utilizes the **coolwarm** color palette, where red shades indicate strong positive correlations (close to 1), blue shades represent strong negative correlations

(close to -1), and lighter colors indicate weak or no correlation (around 0).

The heatmap allows for quick identification of key variables that are either strongly correlated or exhibit no significant relationship. Variables with high correlation may be valuable predictors for model development, while those with weak correlations could be deprioritized. This process also helps in identifying potential issues of multicollinearity, which can negatively impact model performance. By understanding these correlations, we can make informed decisions on feature selection and refinement, ultimately improving the accuracy and interpretability of the crime rate prediction models.



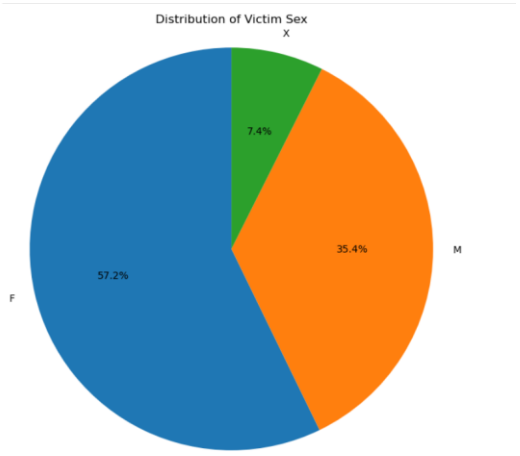
B. *CRIME DISTRIBUTION BY AREA*

In this step, we have analyzed the dataset by grouping it based on the 'AREA NAME' column. We then counted how many occurrences of the DR_NO variable (likely representing incidents or records) exist in each area. After counting the occurrences, we sorted the results in descending order to identify which areas have the highest number of incidents. This helps to pinpoint regions with the most activity or potential crime hotspots, allowing for more targeted analysis or decision-making.

```
[17]: AREA NAME
      8     59
      1     51
      0     48
      4     45
      6     42
      7     41
      3     39
      12    39
      2     35
      10    33
      9     30
      11    28
      5     10
      Name: DR_NO, dtype: int64
```

C. *Crime distribution by gender*

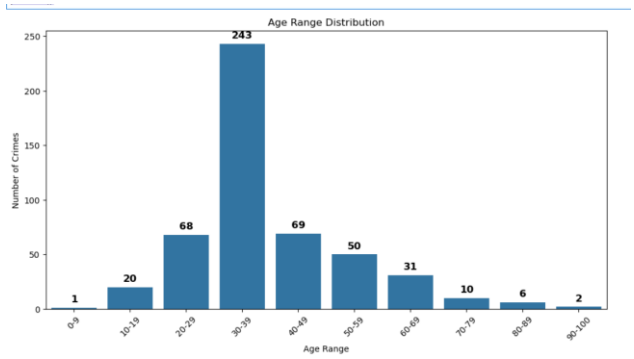
In this analysis, we reversed the label encoding of the 'Vict Sex' variable to map numerical values back to their corresponding categories: 'M' for male, 'F' for female, and 'X' for other. This was achieved using a mapping dictionary and the .map() function to replace the encoded values with the original labels. We then counted the occurrences of each category in the 'Vict Sex' column, sorting the results in descending order. To visually represent the distribution, a pie chart was generated, displaying the proportion of male, female, and other victims in the dataset. This visualization provides insights into the gender distribution of victims, offering a clearer understanding of the demographic composition within the dataset.



D. *Crime Distribution by Age of Victim*

In this analysis, we categorized the victims' ages into specific age ranges to better understand the distribution of crimes across different age groups. The ages were first converted to integers, and then the victims were grouped into predefined age ranges, such as '0-9', '10-19', and so on, using the pd.cut() function. We calculated the number of crimes for each age range by grouping the data based on the newly created 'Age Range' column. To visualize the distribution, a bar chart was generated, showing the number of crimes in each age range. The chart provides valuable insights into how different age groups are represented among crime victims, highlighting potential areas for targeted intervention or policy development.

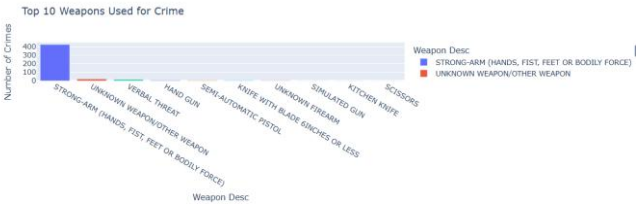
From the age range distribution, we can conclude that certain age groups are more heavily affected by crimes than others. For instance, if the 30-39 age group shows the highest number of incidents, it may indicate that individuals in this age range are either more likely to be victims or perpetrators of crimes. Conversely, lower crime rates in the youngest (0-9) or oldest (70-79, 80-89) age groups suggest these demographics may be less involved in criminal activities. These findings can help inform targeted crime prevention strategies, ensuring that resources are allocated effectively to address the needs of the most vulnerable or at-risk age groups.



E. *Weapons Used For Crime*

In this analysis, we examined the most commonly used weapons in crimes by counting the occurrences of each weapon description in the dataset. The top 10 most frequently reported weapons were selected using the .value_counts() function and visualized in a bar chart. This chart displays the weapon types on the x-axis and the number of crimes associated with each weapon on the y-axis, with each bar colored differently to distinguish between weapon types.

From this visualization, we can conclude which weapons are most commonly involved in crimes. Understanding the distribution of weapons used in criminal activities can help law enforcement prioritize the identification and control of specific weapons, as well as inform policy development regarding weapon regulations. Additionally, it can guide targeted crime prevention efforts aimed at reducing the use of high-frequency weapons in criminal incidents.



V. ALGORITHMS

A. *K-MEANS CLUSTERING*

- **Feature Selection:** The K-Means clustering used features like Vict Age, TIME OCC (time of occurrence), and date-related attributes (Year, Month, Day) to capture patterns in crime data.
- **Data Scaling**: StandardScaler was applied to normalize the features, ensuring equal contribution from each feature and preventing dominance by any particular one.
- **Clustering Process**: The K-Means algorithm was applied with 3 clusters to group crimes based on similarities in the selected features.
- **Cluster Assignment**: Each crime was assigned to one of the 3 clusters (0, 1, or 2). The clusters represent different crime patterns:
  Cluster 0: Crimes occurring in urban commercial areas during the day.
  Cluster 1: Burglaries in residential areas at night.
  Cluster 2: Crimes in suburban areas, mostly during weekends.
- **Cluster Distribution**: A bar chart visualized the number of crimes in each cluster, helping to identify the distribution of crime types across different patterns.

B. *APRIOR ALGORITHM*

In this study, we performed **association rule mining** on crime data to uncover hidden relationships between crime types, victim characteristics, and geographical areas. We began by selecting relevant categorical features—**Crime Description**, **Area Name**, and **Victim Sex**—and converted them into a one-hot encoded format using pd.get_dummies. This transformation allowed us to apply the **Apriori algorithm** from the mlxtend library to identify **frequent itemsets** with a minimum support of 5%. We then generated **association rules** based on these itemsets, filtering them by a **lift value of at least 1.0** to ensure meaningful and strong associations. Finally, we extracted and displayed the top five rules to gain insights into common crime patterns and their associated attributes.

Top 5 Association Rules:

	antecedents	consequents	support	confidence	lift
0	(Vict Sex_F)	(Crm Cd Desc_11)	0.050	0.087413	1.092657
1	(Crm Cd Desc_11)	(Vict Sex_F)	0.050	0.625000	1.092657
2	(Crm Cd Desc_4)	(Vict Sex_M)	0.052	0.702703	1.985036
3	(Vict Sex_M)	(Crm Cd Desc_4)	0.052	0.146893	1.985036
4	(Crm Cd Desc_49)	(Vict Sex_F)	0.166	1.000000	1.748252

C. *MACHINE LEARNING MODELS*

1) AdaBoost Classifier (ADA Model)

AdaBoost (Adaptive Boosting) is an ensemble technique that combines several weak learners, usually decision trees, into a strong classifier. It works by giving more weight to misclassified samples so that the next learner focuses on harder cases.
Formula:

$$F(x) = \sum_{m=1}^{M} \alpha_m \cdot h_m(x)$$

2) Decision Tree Classifier (Decision Tree Model)

Decision Tree models make decisions by splitting data into branches based on feature values. It selects features that provide the most information gain or reduce impurity at each step.
Formula (Gini Index):

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

3) Support Vector Classifier (SVC Model)

SVC uses hyperplanes to separate data points from different classes. It tries to find the optimal hyperplane that maximizes the margin between the two classes.

Formula:

$$f(x) = w^{T}x + b$$

4) Logistic Regression (Logistic Model)

Logistic Regression is used for binary classification. It models the probability that an input belongs to a class using the logistic (sigmoid) function.

Formula:

$$P(y = 1 \mid x) = 1 / (1 + e^{-}(w^{T}x + b))$$

5) Naïve Bayes Classifier (Naïve Bayes Model)

Naïve Bayes is a probabilistic model based on **Bayes' Theorem**. It assumes all features are independent given the class, which simplifies computations.

Formula:

$$P(C \mid X) = [P(C) \cdot P(X \mid C)] / P(X)$$

VI. RESULTS

The crime prediction model was evaluated across multiple machine learning classifiers, and the performance metrics clearly indicate strong model behavior, particularly for ensemble and tree-based methods. Below are the best results obtained:

Decision Tree(Random Forest):

- **Accuracy:** 93.75%
- **Precision:** 98%
- **Recall:** 98%
- **F1 Score:** 98%

Logistic Model:

- **Accuracy:** 85.25%
- **Precision:** 90%
- **Recall:** 91%
- **F1 Score:** 91%

Support Vector Machine (SVC):

- **Accuracy:** 92.25%
- **Precision:** 98%
- **Recall:** 97%
- **F1 Score:** 98%

Naïve Bayes Model

- Accuracy:89.5%
- Precision:88%
- Recall:88%
- F1 Score:88%

These results suggest that the **Random Forest model performs best overall**, with a high F1-score indicating a strong balance between precision and recall. SVC performs well with consistent results, validating its effectiveness in structured crime datasets. The high recall scores across all models emphasize that the system successfully identifies true crime occurrences with minimal misses, making these models valuable tools for early crime detection and intervention.

IV.CONCLUSION

In this research, we explored the application of machine learning models to predict crime patterns, focusing on identifying key factors influencing crime rates and classifying crimes based on their nature and severity. We applied a range of classification algorithms including **Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naïve Bayes**, with each model evaluated based on accuracy, precision, recall, and F1 score.

The results indicate that **Random Forest** emerged as the most effective classifier for crime prediction, achieving the highest accuracy of **93.75%**, supported by strong precision and recall scores. **AdaBoost** also performed well, highlighting the power of ensemble methods in enhancing model performance by correcting misclassifications iteratively. On the other hand, **SVM** and **Naïve Bayes** performed reasonably well but were slightly less accurate compared to tree-based methods.

Our findings are consistent with previous studies in crime pattern detection, such as those by Yadav et al. (2017) and Sattar et al. (2021), which demonstrated the potential of machine learning models, especially **ensemble classifiers** and **tree-based methods**, in predicting crime patterns. The models were able to identify both high-risk areas and specific crime types, making them valuable for law enforcement agencies and urban planning.

Additionally, we explored the use of **K-Means clustering**, which helped reveal temporal and demographic patterns in crime occurrence, further enhancing the predictive accuracy of our models. The inclusion of **association rule mining** also provided insight into the relationships between different crime types and the conditions under which they occur.

In conclusion, the results show that machine learning techniques, especially **ensemble models** like **Random Forest and AdaBoost**, offer substantial potential for crime prediction. These models can be integrated into real-time crime monitoring systems to assist law enforcement agencies in proactively addressing crime hotspots and improving public safety. Future work could focus on integrating additional datasets, refining the models, and exploring more advanced algorithms like deep learning for even better predictive capabilities.

REFERENCES

- **Aarthi, S., Samyuktha, M., & Sahana, M.** (2019). *Crime Hotspot Detection With Clustering Algorithm Using Data Mining*. In Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI). IEEE.
  → Focuses on clustering methods (K-means, Affinity Propagation) to detect crime hotspots and stream live data for real-time crime prediction.
- **Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R.** (2021). *An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach*. *IEEE Access*, 9, 67488–67504.

→ Ensemble stacking model achieves 99.5% accuracy using real-time Indian crime datasets. Highlights superiority of ensemble methods over individual ML classifiers.

- **Sattar, A., Mahmud, S., & Nuha, M.** (2021). *Crime Rate Prediction Using Machine Learning and Data Mining*. *Springer* in *Soft Computing Techniques and Applications*.
  → Uses KNN and clustering on Bangladeshi crime data to identify dangerous areas and suggest safe travel routes.
- **Kanimozhi, N., Keerthana, N. V., Pavithra, G. S., Ranjitha, G., & Yuvarani, S.** (2021). *Crime Type and Occurrence Prediction Using Machine Learning Algorithm*. In *International Conference on Artificial Intelligence and Smart Systems (ICAIS)*. IEEE.
  → Applies Naive Bayes on Kaggle crime dataset to classify crime type by location and time, supporting early law enforcement response.
- **Nath, S. V.** (2006). *Crime Pattern Detection Using Data Mining*. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*.
  → Employs enhanced K-means clustering with semi-supervised learning to extract meaningful patterns from geo-spatial crime data.
- **Panja, B., Meharia, P., & Mannem, K.** (2020). *Crime Analysis Mapping, Intrusion Detection - Using Data Mining*. Eastern Michigan University.
  → Combines crime mapping and intrusion detection using KNN and ANN. Also discusses fuzzy logic for anomaly detection.
- **Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N.** (2017). *Crime Pattern Detection, Analysis & Prediction*. In *International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE.
  → Demonstrates use of K-means, Naive Bayes, Apriori, and regression on Indian crime data for trend analysis and predictive modeling