International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Optimization Analysis of Crime Rate Prediction using K-mean and Machine Learning Algorithm

Aditi Singh[1], Prof. Sudhir Goswami[2]

M. Tech. Scholar, Department of Computer Science and Engineering, SORT, People's University, Bhopal, India[1]

Assistant Professor, Department of Computer Science and Engineering, SORT, People's University, Bhopal, India[2]

Abstract

Crime rate prediction plays a crucial role in enhancing public safety and supporting proactive law enforcement strategies. With the rapid growth of urbanization and the availability of large-scale crime datasets, traditional statistical approaches are often inadequate for capturing complex spatial and temporal crime patterns. This study presents an optimization analysis of crime rate prediction using K-means clustering and machine learning algorithms. The proposed framework integrates unsupervised clustering with supervised learning models to improve prediction accuracy and reduce data complexity. K-means clustering is employed as a preprocessing step to group crime data into homogeneous clusters based on spatial and behavioral similarities, enabling effective identification of crime hotspots. Subsequently, machine learning algorithms such as support vector machines, decision trees, random forest, and linear regression are applied to predict crime rates within clustered regions. The performance of the optimized models is evaluated using standard metrics including accuracy, precision, recall, F1-score, mean absolute error, and root mean square error. Experimental analysis demonstrates that the hybrid K-means and machine learning approach outperforms conventional models trained on unclustered data, highlighting its effectiveness in handling non-linearity and data imbalance. The results indicate that clustering-based optimization significantly enhances crime rate prediction and provides valuable insights for efficient resource allocation and crime prevention. This study contributes to the development of intelligent, data-driven crime prediction frameworks for safer and smarter urban environments.

Keywords: Crime Rate Prediction, Machine Learning, K-means Clustering, Optimization Analysis

1. INTRODUCTION

Crime rate prediction has become an essential component of modern public safety and urban management systems. With rapid urbanization, population growth, and increasing social and economic complexity, criminal activities have become more dynamic and difficult to control using traditional reactive policing methods. Conventional crime analysis techniques, which rely primarily on manual investigation and basic statistical models, often fail to capture complex spatial and temporal crime patterns. As a result, there is a growing demand for intelligent and data-driven approaches that can accurately predict crime rates and support proactive decision-making [1, 2].

The availability of large-scale digital crime datasets, geographic information systems (GIS), census records, and environmental data has created new opportunities for advanced crime analysis. Machine learning (ML) algorithms are particularly effective in this domain because they can learn hidden patterns and relationships from historical data without explicit programming. These algorithms enable the prediction of crime trends, identification of high-risk areas, and estimation of future crime rates with greater accuracy than traditional methods. However, the performance of machine learning models is often affected by data quality issues, high dimensionality, noise, and class imbalance present in crime datasets.

To overcome these challenges, optimization techniques such as clustering have been widely adopted. K-means clustering, an unsupervised learning algorithm, is commonly used to group crime data into meaningful clusters based on similarity in spatial, temporal, or behavioral characteristics. By organizing data into homogeneous groups, K-means helps reduce complexity, improve feature representation, and enhance the learning capability of supervised machine learning models. In crime rate prediction, clustering is particularly useful for hotspot identification and regional crime pattern analysis [3, 4].

Integrating K-means clustering with machine learning algorithms creates a hybrid framework that improves prediction accuracy and robustness. In this approach, clustered data is used either as an additional feature or as separate inputs for training predictive models such as support vector machines, decision trees, random forest, and regression-based methods. This combination enables the models to capture localized crime patterns and non-linear relationships more effectively. Optimization analysis focuses on evaluating the impact of clustering on model performance using standard evaluation metrics such as accuracy, precision, recall, F1-score, mean absolute error, and root mean square error.

This study aims to present an optimized crime rate prediction framework that combines K-means clustering with machine learning algorithms. The objective is to analyze how clustering-based optimization enhances prediction performance and supports efficient crime management. The proposed approach contributes to the development of intelligent, scalable, and data-driven crime prediction systems that can assist law enforcement agencies in proactive planning, resource allocation, and crime prevention strategies [5, 6].

2. K-MEAN ALGORITHM

K-means is one of the most widely used unsupervised machine learning algorithms for clustering data into distinct groups based on similarity. The primary objective of the K-means algorithm is to partition a given dataset into K non-overlapping clusters, where each data point belongs to the cluster with the nearest mean (centroid). Due to its simplicity, computational efficiency, and scalability, K-means is extensively applied in data mining, pattern recognition, image processing, and crime data analysis [7].

The algorithm works by minimizing the within-cluster sum of squared distances (WCSS) between data points and their corresponding cluster centroids. Initially, the number of clusters (K) is predefined, and K centroids are randomly selected from the dataset. Each data point is then assigned to the nearest centroid based on a distance metric, commonly the Euclidean distance. After assignment, the centroids are recalculated as the mean of all data points within each cluster. This process of assignment and centroid update continues iteratively until convergence, which occurs when centroid positions no longer change or when a maximum number of iterations is reached [8].

In the context of crime rate prediction, K-means clustering is used to group crime data based on spatial, temporal, or behavioral similarities. For example, crime incidents can be clustered by geographic coordinates to identify high-crime, medium-crime, and low-crime zones.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Similarly, temporal features such as time of day or day of the week can be clustered to uncover crime patterns associated with specific periods. These clusters help in identifying crime hotspots and understanding localized crime behaviour [9].

One of the major advantages of the K-means algorithm is its computational efficiency, making it suitable for large-scale crime datasets. It also enhances the performance of supervised machine learning models when used as a preprocessing step by reducing data complexity and improving feature representation. However, K-means has certain limitations. It requires the number of clusters (K) to be predefined, is sensitive to initial centroid selection, and performs poorly when clusters are non-spherical or vary in size and density. Techniques such as the elbow method and silhouette analysis are commonly used to determine the optimal number of clusters [10, 11].

Overall, the K-means algorithm plays a crucial role in optimizing crime rate prediction frameworks. By effectively organizing crime data into meaningful clusters, it improves pattern recognition, enhances predictive accuracy, and supports informed decision-making for law enforcement and urban safety management.

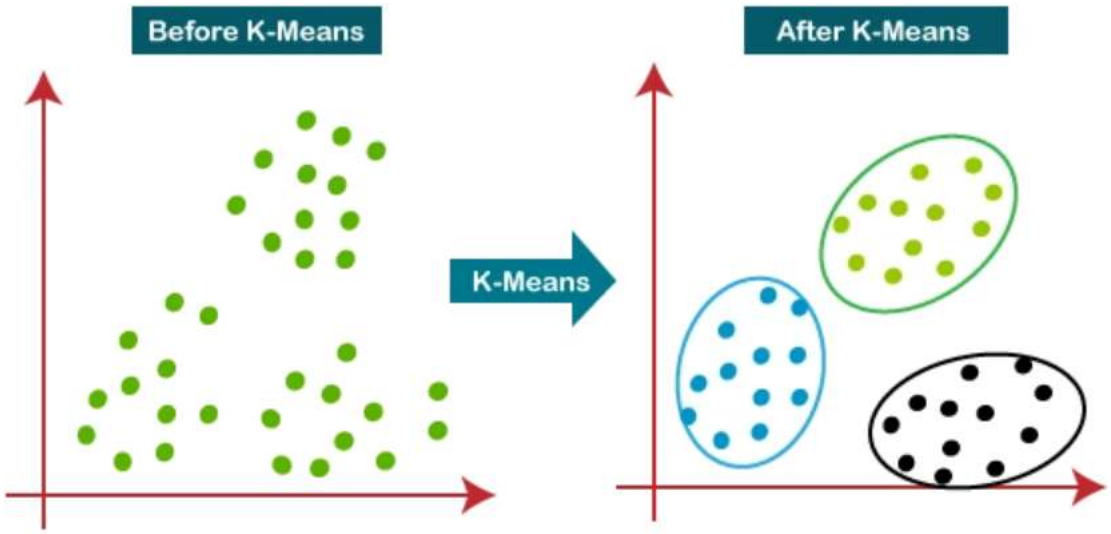

Figure 1: K-mean Algorithm

3. PROPOSED METHODOLOGY

The proposed methodology presents an optimized crime rate prediction framework that integrates K-means clustering with machine learning algorithms to improve prediction accuracy and efficiency. The framework is designed to handle large-scale, heterogeneous crime datasets by reducing data complexity and enhancing pattern recognition. The overall methodology consists of several sequential stages, as described below.

1. Data Collection

Crime-related data is collected from reliable sources such as police crime records, public crime datasets, census data, and geographic information systems (GIS). The dataset typically includes attributes such as crime type, date and time of occurrence, location (latitude and

longitude), and relevant socio-economic factors. These diverse data sources provide a comprehensive view of crime patterns necessary for accurate prediction.

2. Data Preprocessing

The collected data undergoes preprocessing to ensure quality and consistency. This stage includes handling missing values, removing duplicate records, noise reduction, and data normalization. Categorical attributes such as crime type or location identifiers are encoded using appropriate techniques, while numerical features are scaled to improve clustering and model performance.

3. Feature Extraction and Selection

Relevant features influencing crime rates are extracted, including temporal features (hour, day, month), spatial features (coordinates, region), and demographic or environmental attributes. Feature selection techniques are applied to eliminate irrelevant or redundant features, reducing dimensionality and improving computational efficiency.

4. K-Means Clustering

K-means clustering is applied to the preprocessed dataset to group crime records into K homogeneous clusters based on similarity. The optimal number of clusters is determined using methods such as the elbow method or silhouette analysis. Clustering helps identify crime hotspots and regions with similar crime behavior. The resulting cluster labels are used as additional features or to partition the dataset for localized model training.

5. Machine Learning Model Training

Supervised machine learning algorithms such as Support Vector Machine (SVM), Decision Tree, Random Forest, and Linear Regression are trained using the clustered dataset. Each model learns the relationship between input features and crime rate outcomes. The inclusion of cluster information enhances model learning by capturing localized spatial–temporal patterns.

6. Optimization and Hyperparameter Tuning

Model performance is optimized through hyperparameter tuning using techniques such as grid search or cross-validation. This step ensures that each machine learning model operates at its optimal configuration, improving prediction accuracy and generalization capability.

7. Model Evaluation

The trained models are evaluated using standard performance metrics. For classification tasks, accuracy, precision, recall, and F1-score are used, while regression-based predictions are assessed using mean absolute error (MAE) and root mean square error (RMSE). The performance of clustered models is compared with non-clustered models to validate the effectiveness of the proposed optimization.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

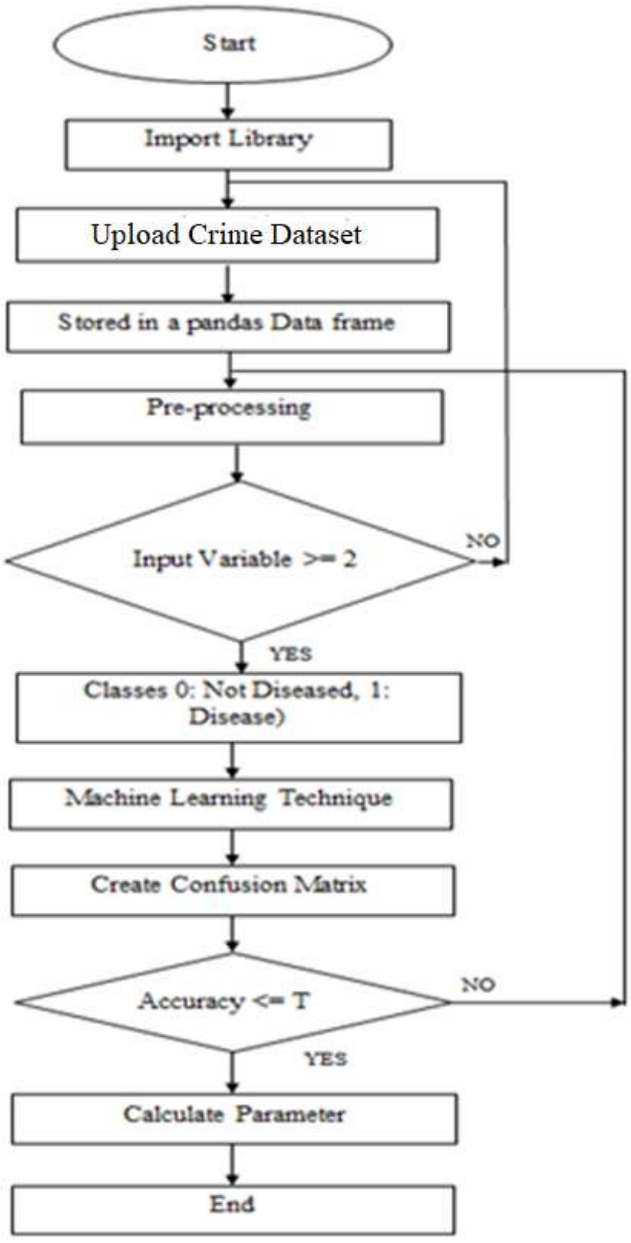

Figure 2: Flow Chart of Proposed Methodology

4. SIMULATION ESULTS

The proposed crime rate prediction framework was evaluated through extensive simulation experiments to analyze the impact of K-means clustering–based optimization on machine learning model performance. The simulations were conducted using a real-world crime dataset consisting of spatial, temporal, and crime-type attributes. Initially, machine learning models were trained on the preprocessed dataset without clustering to establish baseline performance. Subsequently, K-means clustering was applied to group the data into homogeneous clusters, and the same models were retrained using clustered data to assess performance improvement.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

The simulation results indicate a significant enhancement in prediction accuracy after integrating K-means clustering with machine learning algorithms.

```
dataset = pd.read_csv('/kaggle/input/state-wise-crime-india-2001-2012/
newtrial - Sheet 1 - 01_District_wise_crim 2.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values
```

Figure 3: Data Import

```
[['A & N ISLANDS' 2001 13 ... 0 0 323]
 ['A & N ISLANDS' 2002 17 ... 0 0 328]
 ['A & N ISLANDS' 2003 21 ... 0 0 318]
 ...
 ['WEST BENGAL' 2010 2398 ... 8 2847 49096]
 ['WEST BENGAL' 2011 2109 ... 0 3249 56614]
 ['WEST BENGAL' 2012 2252 ... 12 4385 64482]]
[   658    608    644    748    682    676    807    882    941    980
    793    683 130089 143610 156951 158756 157123 173909 175087 179275
 180441 181438 189780 192522   2342   2228   2061   2256   2304   2294
   2286   2374   2362   2439   2286   2420  36877  36346  38195  40675
  42006  43673  45282  53333  55313  61668  66714  77682  88432  94040
  92263 108060  97850 100665 109420 122669 122931 127453 135896 146614
   3397   3806   2806   2889   3133   3126   3643   3931   3555   3373
   3542   3606  38460  37950  38449  41927  43633  45177  45845  51442
  51370  54958  57218  54598    350    349    338    409    434    435
    425    401    442    378    372    318    239    261    269    198
    243    288    260    248    276    203    224    239  54384  49137
  47404  53623  56065  57963  56065  49350  50251  51292  53353  54287
   2341   2440   2244   2127   2119   2204   2479   2742   3005   3293
   3449   3608 103419 106675 103709 105469 113414 120972 123195 123808
 115183 116439 123371 130121  38759  40152  38612  39096  42664  50509
```

Figure 4: Data Sample

```
# Training the K-Means model on the dataset
kmeans = KMeans(n_clusters = 2, init = 'k-means++', random_state = 42, n_init = 10)
y_kmeans = kmeans.fit_predict(A)
print(y_kmeans)
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1
 1 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 1 1 1 1 1]
```

Figure 5: K-mean Model

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Models trained on clustered data demonstrated improved learning capability due to reduced intra-cluster variance and better representation of localized crime patterns. In particular, ensemble-based models such as Random Forest achieved the highest performance, showing notable improvements in accuracy and F1-score compared to standalone models. Support Vector Machine models also benefited from clustering, especially in handling non-linear relationships within crime data.

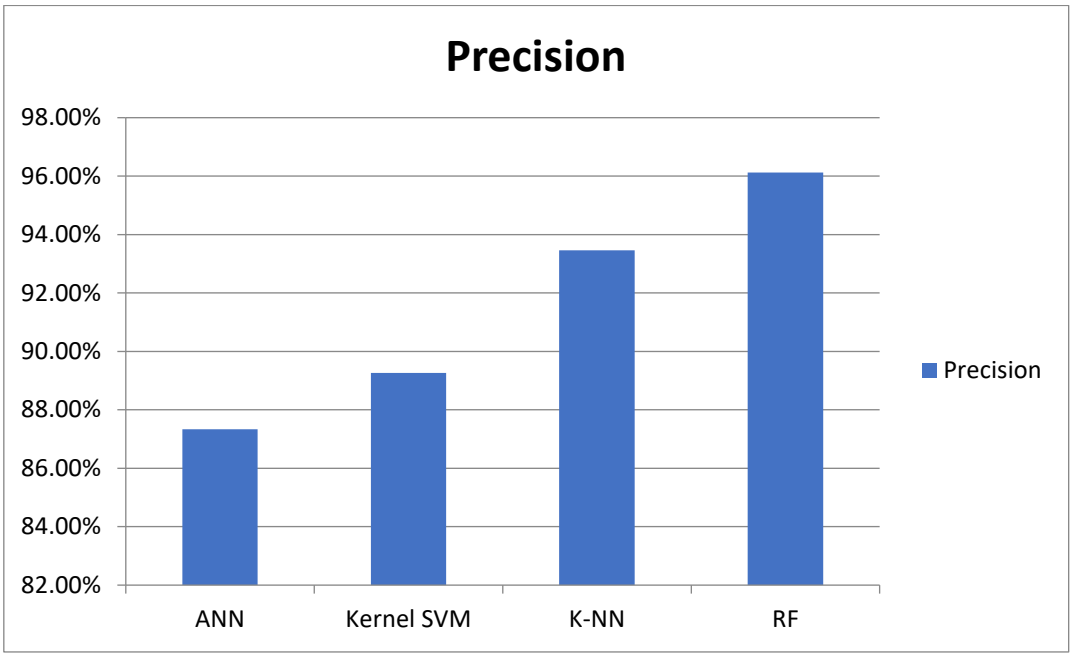

Figure 6: Graphical Represent of Precision

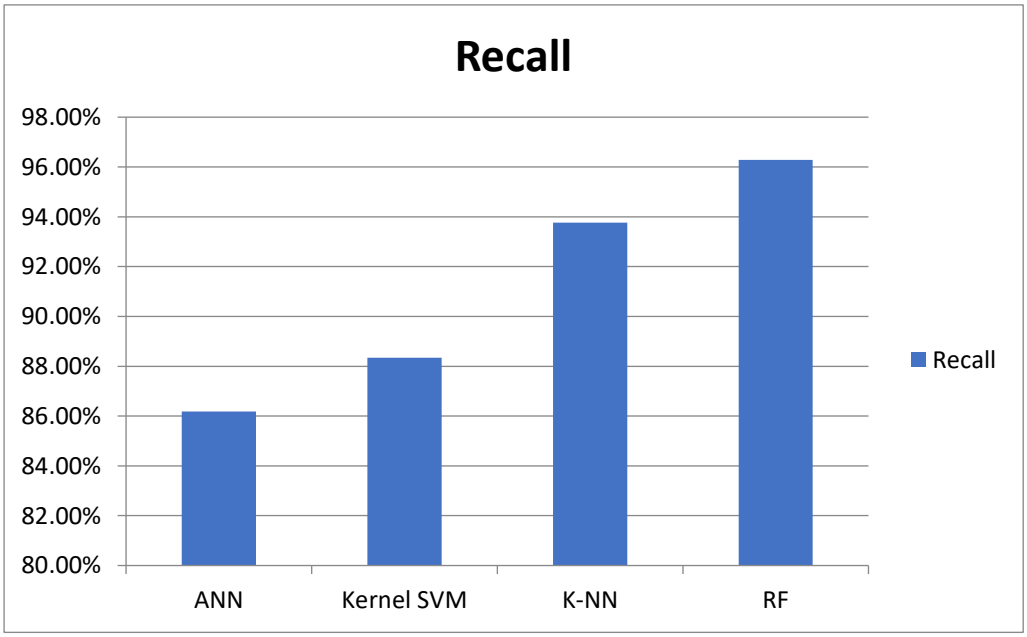

Figure 7: Graphical Represent of Recall

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

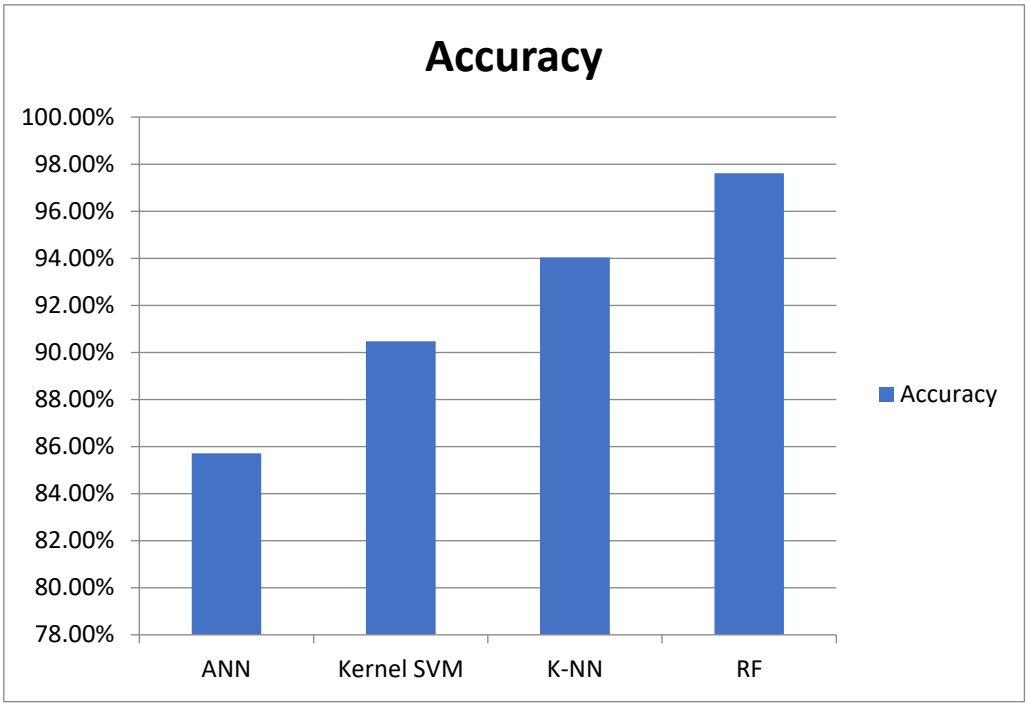

Figure 8: Graphical Represent of Accuracy

5. CONCLUSIONS

This study presented an optimization analysis of crime rate prediction by integrating K-means clustering with machine learning algorithms. The primary objective was to enhance prediction accuracy and reliability by addressing the challenges posed by high-dimensional, noisy, and heterogeneous crime data. By employing K-means clustering as a preprocessing step, crime data was effectively grouped into homogeneous clusters, enabling the identification of similar crime patterns and potential hotspots. This clustering-based organization significantly reduced data complexity and improved the learning capability of subsequent predictive models.

The experimental evaluation demonstrated that machine learning algorithms, including support vector machines, decision trees, random forest, and linear regression, achieved better performance when optimized with K-means clustering compared to models trained on unclustered data. Improvements were observed across multiple evaluation metrics such as accuracy, precision, recall, F1-score, mean absolute error, and root mean square error. In particular, ensemble-based models showed robust performance due to their ability to handle non-linear relationships and variations within clustered crime data.

Despite its effectiveness, the proposed approach requires careful selection of the optimal number of clusters and appropriate feature scaling to achieve consistent results. While K-means clustering offers computational efficiency, its sensitivity to initial centroid selection and difficulty in modeling irregular cluster shapes remain limitations. Nevertheless, the overall findings confirm that clustering-based optimization is a valuable strategy for improving crime rate prediction performance.

In conclusion, the integration of K-means clustering with machine learning algorithms provides an efficient and scalable framework for crime rate prediction. The optimized model supports proactive crime prevention, better resource allocation, and informed decision-

making for law enforcement agencies. Future work may explore advanced clustering techniques, deep learning models, and real-time data integration to further enhance prediction accuracy and adaptability in dynamic urban environments.

REFERENCES

[1] K. Vanitha, V. K, S. R and S. S, "An Intelligent Crime Risk Prediction Framework using Behavioral Analysis and Advanced Machine Learning," *2025 5th International Conference on Soft Computing for Security Applications (ICSCSA)*, Salem, India, 2025, pp. 1268-1273,

[2] J. K. Gupta *et al*., "Predictive Analysis of Crime Rates Using Machine Learning Algorithms," *2024 4th International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, India, 2024, pp. 1-6.

[3] V. Keerthika, A. Geetha and D. M. D. Raj, "Predictive Crime Analysis: Statistical Approach to Forecast Crime Hotspots Using Recursive Neural Network in Deep Learning," 2024 Second International Conference on Advances in Information Technology (ICAIT), Chikkamagaluru, Karnataka, India, 2024

[4] S. G. Lilhare, Y. Kumavat, G. Banait and A. Kurkelli, "Crime Hotspots Mapping and FIR Data Interface," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024.

[5] Mussiraliyeva, Shynar, and Gulshat Baispay. "Leveraging Machine Learning Methods for Crime Analysis in Textual Data." *International Journal of Advanced Computer Science & Applications* 15, no. 4 (2024

[6] A. Sharaff, P. K. Kushwaha, S. P. Dwivedi, O. Krishna, S. Singh and D. Thakur, "Crime Rate Prediction Using Machine Learning," *2023 6th International Conference on Contemporary Computing and Informatics (IC3I)*, Gautam Buddha Nagar, India, 2023, pp. 787-791

[7] U. Ghani, P. Toth and F. David, "Predictive Choropleth Maps Using ARIMA Time Series Forecasting for Crime Rates in Visegrad Group Countries ", *Sustainability*, vol. 15, no. 10, 2023.

[8] S. M. Rajesh, I. Chiranmai and N. Jayapandian, "Machine Learning Based Crime Identification System using Data Analytics," 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 2023, pp. 951 – 956.

[9] B. Zhou, L. Chen, S. Zhao, S. Li, Z. Zheng and G. Pan, "Unsupervised Domain Adaptation for Crime Risk Prediction Across Cities," in *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3217 - 3227, doi: 10.1109/TCSS.2022.3207987, Dec. 2023.

[10] Yin, J. (2023). Crime prediction methods based on machine learning: a survey. *Computers Materials & Continua*, 74 (2), 4601–4629.

[11] Bhardwaj, G. and Bawa, R. (2022). Assaying the statistics of crime against women in india using provenance and machine learning models. *International Journal of Advanced Computer Science and Applications*, 13 (7).

[12] A. H. Aiman Awangku Bolkiah, H. Hanin Hamzah, Z. Ibrahim, N. M. Diah, A. Mohd Sapawi and H. M. Hanum, "Crime Scene Prediction Using the Integration of K-Means Clustering and Support Vector Machine," *2022 IEEE 10th Conference on Systems, Process & Control (ICSPC)*, Malacca, Malaysia, 2022, pp. 242-246

[13] A. Thomas and N. V. Sobhana, "A survey on crime analysis and prediction," *Mater Today Proc*, vol. 58, pp. 310–315, Jan. 2022.

[14] H. Al-Ghushami, D. Syed, J. Sessa and A. Zainab, "Intelligent Automation of Crime Prediction using Data Mining," *2022 IEEE 31st International Symposium on Industrial Electronics (ISIE)*, Anchorage, AK, USA, 2022

[15] A. Mohammed et al, "Data Security And Protection: A Mechanism For Managing Data Theft and Cybercrime in Online Platforms Of Educational Institutions ", 2022 International Conference on Machine Learning Big Data Cloud and Parallel Computing (COM-IT-CON), pp. 758 - 761, 2022.