FINAL SUBMISSION-ML PROJECT REPORT

INTRODUCTION: -

In today's society, safety and security are paramount considerations for individuals and families when selecting a place to reside. Understanding the crime rates of different neighbourhoods can greatly influence decisions related to housing and community selection. As such, our project focuses on addressing this critical need by developing a robust Crime Rate Prediction system.

The aim of our project is to assist real estate companies, home builders, and the general public in making informed decisions about choosing neighbourhoods with low crime rates or no crime. Leveraging advanced machine learning algorithms, we analysed historical crime data from various districts and states to predict future crime rates accurately. By providing insights into crime trends and patterns at a granular level, our solution empowers stakeholders to prioritize safety and security when considering residential locations.

Our Crime Rate Prediction system offers a user-friendly interface that allows users to input specific parameters such as location, district, and year of interest. The system then utilizes sophisticated predictive models to generate accurate forecasts of crime rates for the specified criteria. By harnessing the power of data-driven insights, our project aims to contribute to creating safer and more secure communities for everyone.

PROBLAME STATEMENT: -

CRIME RATE PREDICTION

You are working for an organization which is supporting real estate companies, home builders and general public to choose the right place/neighbour which has no crime or lowest crime rate.

The problem at hand involves developing a Crime Rate Prediction system to aid real estate companies, home builders, and the general public in selecting neighbourhoods with minimal or no crime. This system aims to provide accurate forecasts of crime rates in different areas, allowing stakeholders to make informed decisions when choosing residential locations. By leveraging historical crime data and advanced machine learning algorithms, our goal is to empower individuals and organizations to prioritize safety and security when selecting places to live or build homes.

ALGORITHM DEFINITION: -

As per problem statement understanding we are going to use below models

- **Support Vector Regressor (SVR):**

  - SVR is a supervised machine learning algorithm used for regression tasks.
  - It works by mapping input data into a high-dimensional feature space and finding the hyperplane that best separates the output variable (crime rate) into different classes.
  - SVR aims to minimize the error between the predicted and actual crime rates while maximizing the margin of separation between different crime rate categories.
  - In the context of crime rate prediction, SVR can be used to model the relationship between various features such as location, district, and year, and predict the corresponding crime rates.

- **Neural Network (NN):**

  - Neural Networks are a class of algorithms inspired by the structure and functioning of the human brain.
  - In the context of crime rate prediction, a neural network can be trained to learn complex patterns and relationships between input features (such as location, district, and year) and the corresponding crime rates.
  - NNs consist of multiple layers of interconnected neurons (nodes) that process input data through a series of mathematical operations to generate output predictions.
  - They are known for their ability to capture nonlinear relationships in data and can be trained using historical crime data to make accurate predictions of future crime rates.

- **K-Nearest Neighbours (KNN):**

- KNN is a simple yet effective supervised learning algorithm used for classification and regression tasks.
- It works by finding the 'k' nearest data points to a given query point based on a distance metric (e.g., Euclidean distance) and averaging their target values to make predictions
- In the context of crime rate prediction, KNN can be used to predict the crime rate of a specific location or district based on the crime rates of its nearest neighbours.

➢ **Random Forest (RF):**

- Random Forest is an ensemble learning technique that combines multiple decision trees to make predictions.
- Each decision tree is trained on a random subset of the training data and a random subset of features, leading to a diverse set of trees.
- In the context of crime rate prediction, Random Forest can be used to model the relationship between input features (e.g., location, district, and year) and crime rates by learning from historical crime data.
- RF is known for its robustness to overfitting, scalability to large datasets, and ability to handle nonlinear relationships between features and target variables.

➢ **Linear Regression (LR):**

- Linear Regression is a classical statistical method used for modelling the relationship between one or more independent variables (features) and a dependent variable (target).
- In the context of crime rate prediction, Linear Regression can be used to establish a linear relationship between input features (e.g., location, district, and year) and crime rates.
- LR works by fitting a linear equation to the observed data points, aiming to minimize the difference between the predicted and actual crime rates.

EXPERIMENTAL EVALUATION: -

Methodology: -

➢ **Data Collection:**

Gathered historical crime data from Kaggle. Datasets containing relevant features such as demographics, socioeconomic indicators, geographic information, etc district.

- Source: The data was sourced from government databases, law enforcement agencies, and crime reporting platforms.
- Time Period: The dataset covers a time period ranging from 2001 to 2014, with crime rate observations recorded at regular intervals (e.g., monthly, annually).
- Scope: The dataset includes crime rate statistics for different categories of crimes, such as violent crimes, property crimes, social offenses, and crimes against women.
- Granularity: Crime rate data is available at the district and state levels, allowing for both localized and regional analyses.

➢ **Features:**

1. *State/UT Classification*: Indicates whether the location is classified as a state or a union territory.
2. *State/UT Name*: Name of the state or union territory.
3. *District*: Name of the district within the state/union territory.
4. *Year*: The year in which the crime rate observation was recorded.
5. *Murder*: Intentional killing of another person, excluding cases of manslaughter.
6. *Attempted Murder*: Failed attempts to unlawfully cause the death of another individual.
7. *Culpable Homicide Not Amounting to Murder*: Cases of non-intentional killings that do not qualify as murder.
8. *Rape*: Forcible sexual intercourse without the consent of the victim.
9. *Custodial Rape*: Instances of rape that occur while the victim is in the custody of law enforcement or legal authorities.
10. *Kidnapping & Abduction*: Unlawful taking away or confinement of an individual against their will.
11. *Kidnapping and Abduction of Women and Girls*: Specific instances of kidnapping and abduction involving females.
12. *Kidnapping and Abduction of Others*: Similar incidents involving individuals other than women and girls.
13. *Other Rape*: Additional cases of rape not falling under custodial rape.

14. *Assault on Women with Intent to Outrage Her Modesty*: Assaults or attacks on women with the intention to violate their modesty.
15. *Insult to Modesty of Women*: Offenses that involve offending or humiliating the modesty of women.
16. *Cruelty by Husband or His Relatives*: Acts of cruelty or harassment against married women by their husbands or in-laws.
17. *Importation of Girls from Foreign Countries*: Illegal trafficking or importing of girls from foreign nations.
18. *Dacoity*: Armed robbery involving a group of people.
19. *Preparation and Assembly for Dacoity*: Acts of planning or organizing dacoity.
20. *Robbery*: Theft or unlawful taking of property or money from an individual or institution.
21. *Burglary*: Illegal entry into a building with the intent to commit theft or any other crime.
22. *Theft*: Unlawful taking of someone else's property with the intention of permanently depriving the owner.
23. *Auto Theft*: Theft or stealing of motor vehicles.
24. *Other Theft*: Additional instances of theft not classified under specific categories.
25. *Riots*: Public disturbances involving violence, disorder, or tumultuous behaviour by a group of people.
26. *Criminal Breach of Trust*: Breach of trust or confidence by a person entrusted with property or power.
27. *Cheating*: Deceptive practices or fraudulent representations to deceive another person.
28. *Counterfeiting*: Production of fake or unauthorized copies of items such as currency, documents, or products.
29. *Arson*: Intentional or malicious setting of fire to property or buildings.
30. *Hurt/Grievous Hurt*: Infliction of bodily injury or harm, ranging from minor hurt to severe injuries.
31. *Dowry Deaths*: Deaths of married women caused by harassment or cruelty related to dowry demands.
32. *Causing Death by Negligence*: Unintentional causing of death due to negligence or carelessness.

- **Data Preprocessing:**

This data generated my local agencies so lot of inconsistency. Clean the collected data by handling missing values, removing duplicates, and correcting any inconsistencies or errors. Perform feature engineering to extract meaningful insights from the data, such as creating new features or aggregating existing ones.
For example-:

'D&N HAVELI':'DADRA AND NAGAR HAVELI',
'D & N HAVELI':'DADRA AND NAGAR HAVELI',
'A&N ISLANDS':'ANDAMAN AND NICOBAR ISLANDS',
'A & N ISLANDS':'ANDAMAN AND NICOBAR ISLANDS'

- **Feature Selection:**

Aggregate all same crimes based on similarly, category, and problem statement. We only interested in crimes measure by police stations of particular area. So, we can neglect the crime happens with families, crimes in railways, crimes come under other agencies like ROW, CID, ED

1. *Violent Crimes:*
   Murder: Intentional killing of another person, excluding cases of manslaughter.
   Attempted Murder: Failed attempts to unlawfully cause the death of another individual.
   Culpable Homicide Not Amounting to Murder: Cases of non-intentional killings that do not qualify as murder.
   Kidnapping & Abduction: Unlawful taking away or confinement of an individual against their will.
   Kidnapping and Abduction of Women and Girls: (aggregating in Kidnapping & Abduction column)
   Kidnapping and Abduction of Others: (aggregating in Kidnapping & Abduction column)

2. *Crime against Women:*
   Rape: Forcible sexual intercourse without the consent of the victim.
   Custodial Rape: Instances of rape that occur while the victim is in the custody of law enforcement or legal authorities.
   Other Rape: (added in Rape columns)
   Assault on Women with Intent to Outrage Her Modesty: Assaults or attacks on women with the intention to violate their modesty.
   Insult to Modesty of Women: Offenses that involve offending or humiliating the modesty of women.
   Cruelty by Husband or His Relatives: Acts of cruelty or harassment against married women by their husbands or in-laws.
   Importation of Girls from Foreign Countries: Illegal trafficking or importing of girls from foreign nations.

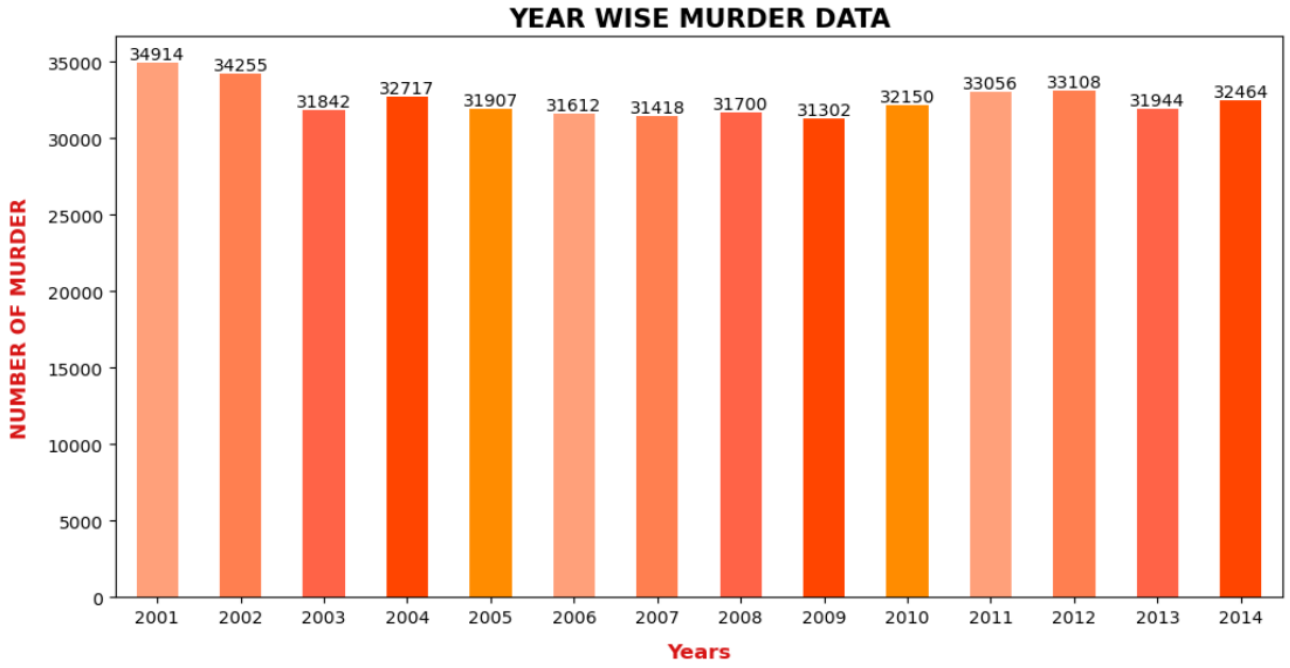
3. *Property Crimes:*
   Dacoity: Armed robbery involving a group of people.
   Preparation and Assembly for Dacoity: Acts of planning or organizing dacoity.
   Robbery: Theft or unlawful taking of property or money from an individual or institution.
   Burglary: Illegal entry into a building with the intent to commit theft or any other crime.
   Theft: Unlawful taking of someone else's property with the intention of permanently depriving the owner.
   Auto Theft: Theft or stealing of motor vehicles.
   Other Theft: Additional instances of theft not classified under specific categories.

4. *Other Offenses:*
   Riots: Public disturbances involving violence, disorder, or tumultuous behaviour by a group of people.
   Criminal Breach of Trust: Breach of trust or confidence by a person entrusted with property or power.
   Cheating: Deceptive practices or fraudulent representations to deceive another person.
   Counterfeiting: Production of fake or unauthorized copies of items such as currency, documents, or products.
   Arson: Intentional or malicious setting of fire to property or buildings.
   Hurt/Grievous Hurt: Infliction of bodily injury or harm, ranging from minor hurt to severe injuries.
   Dowry Deaths: Deaths of married women caused by harassment or cruelty related to dowry demands.
   Causing Death by Negligence: Unintentional causing of death due to negligence or carelessness.

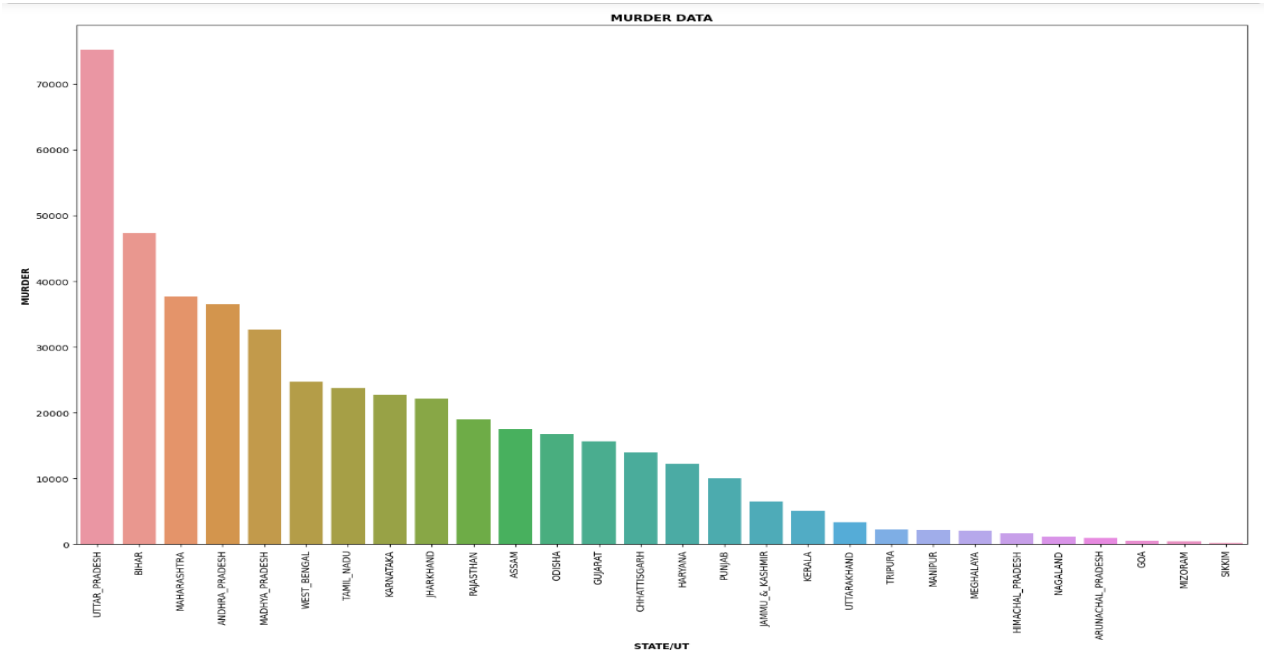
➢ **Exploratory Data Analysis (EDA):**

Conduct a thorough exploratory analysis of the dataset to understand the distribution of crime rates, identify trends, and uncover patterns. Visualize the data using histograms, box plots, heatmaps, and other techniques to gain insights into the relationships between variables.

- *Bivariant analysis of crime over the years (we have all crimes plots in jupyter notebook): -*



- *Bivariant analysis of crime over the states/UT (we have all crime plots in jupyter notebook): -*



- *Bivariant analysis of Violent crimes over the years (we have all plots in jupyter notebook): -*

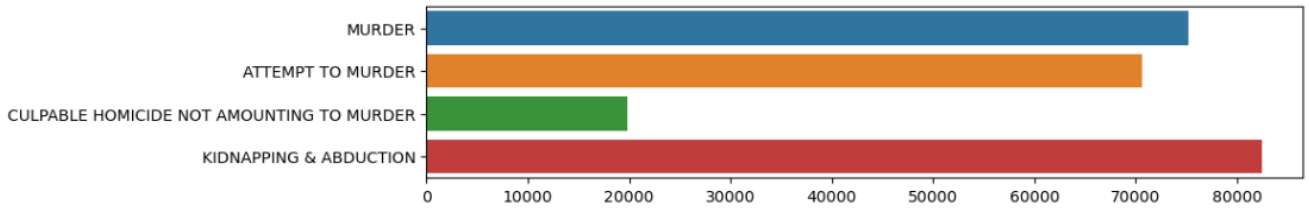
-------VIOLENT CRIME------

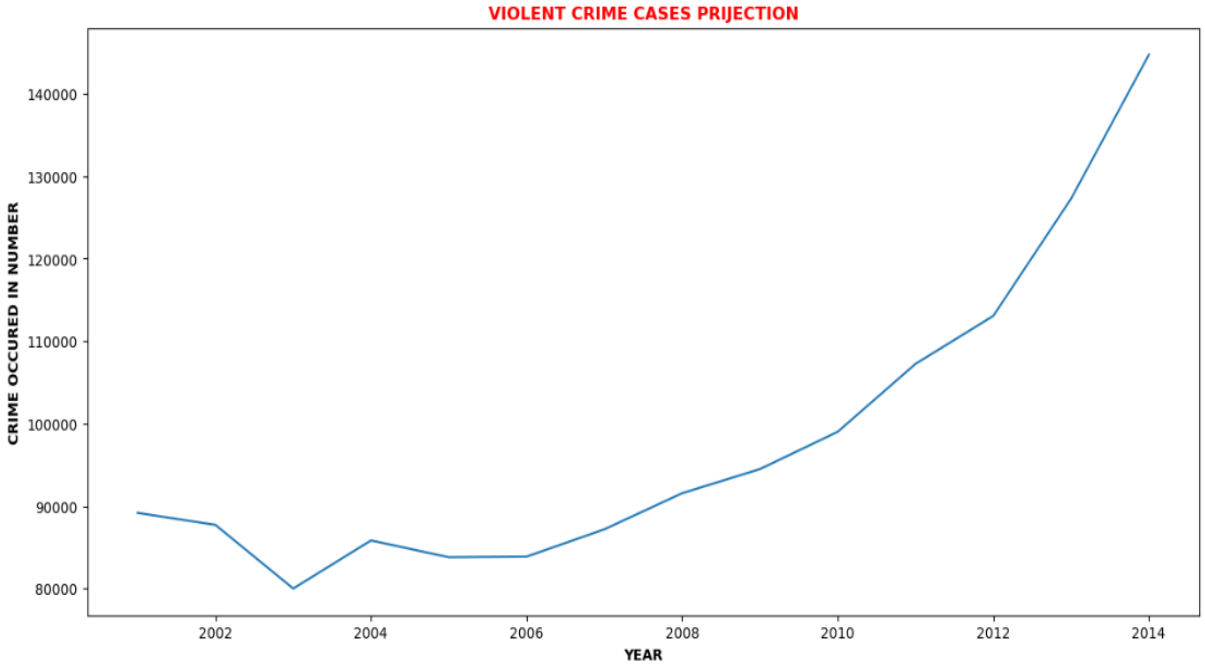
MURDER
TOTAL MURDER Cases 14 YEARS:- **454389**
Maximum cases UTTAR_PRADESH: 75170
Minimum cases SIKKIM: 207

ATTEMPT TO MURDER
TOTAL ATTEMPT TO MURDER Cases 14 YEARS:- **417142**
Maximum cases UTTAR_PRADESH: 70599
Minimum cases SIKKIM: 136

CULPABLE HOMICIDE NOT AMOUNTING TO MURDER
TOTAL CULPABLE HOMICIDE NOT AMOUNTING TO MURDER Cases 14 YEARS:- **49384**
Maximum cases UTTAR_PRADESH: 19843
Minimum cases TRIPURA: 24

KIDNAPPING & ABDUCTION
TOTAL KIDNAPPING & ABDUCTION Cases 14 YEARS:- **454033**
Maximum cases UTTAR_PRADESH: 82415
Minimum cases MIZORAM: 110





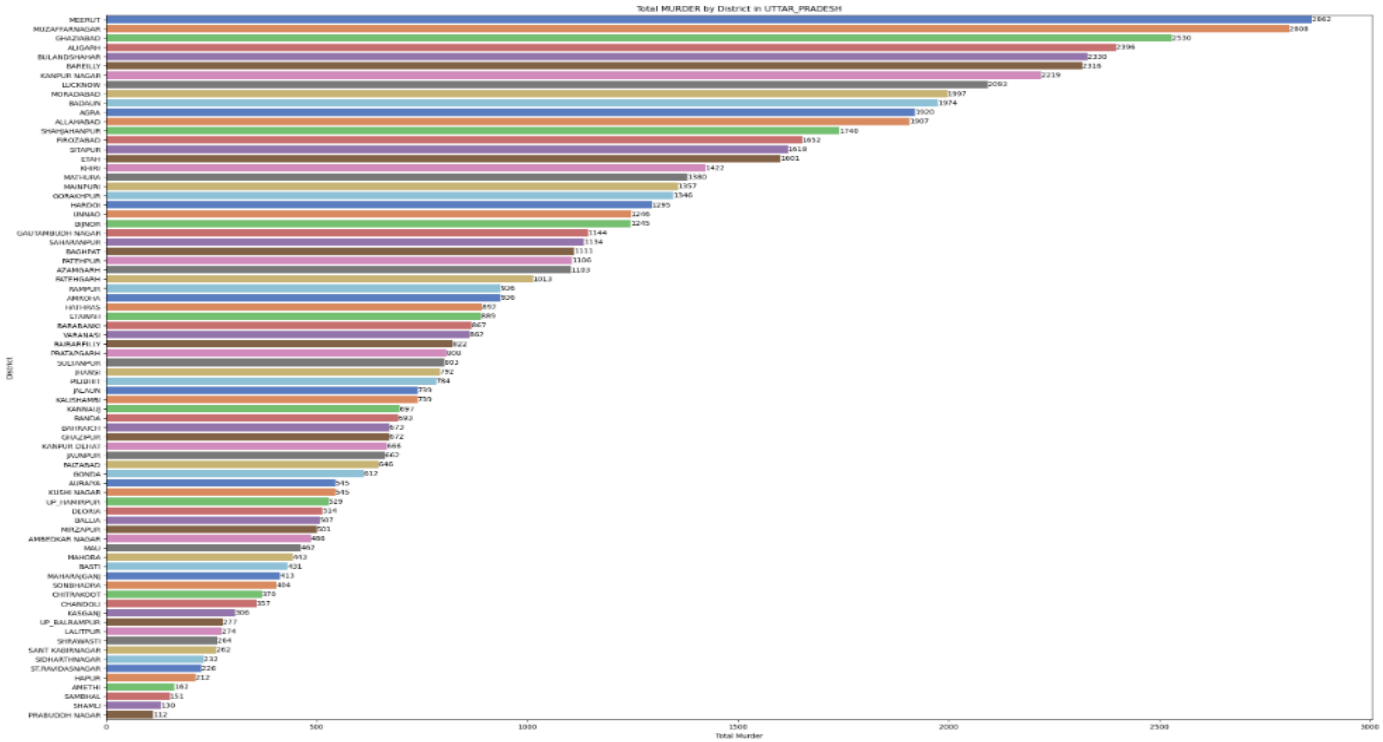
- *Bivariant analysis of Violent crimes against Districts of states has Maximum number of crimes (all plots in jupyter notebook): -*

VIOLENT CRIME

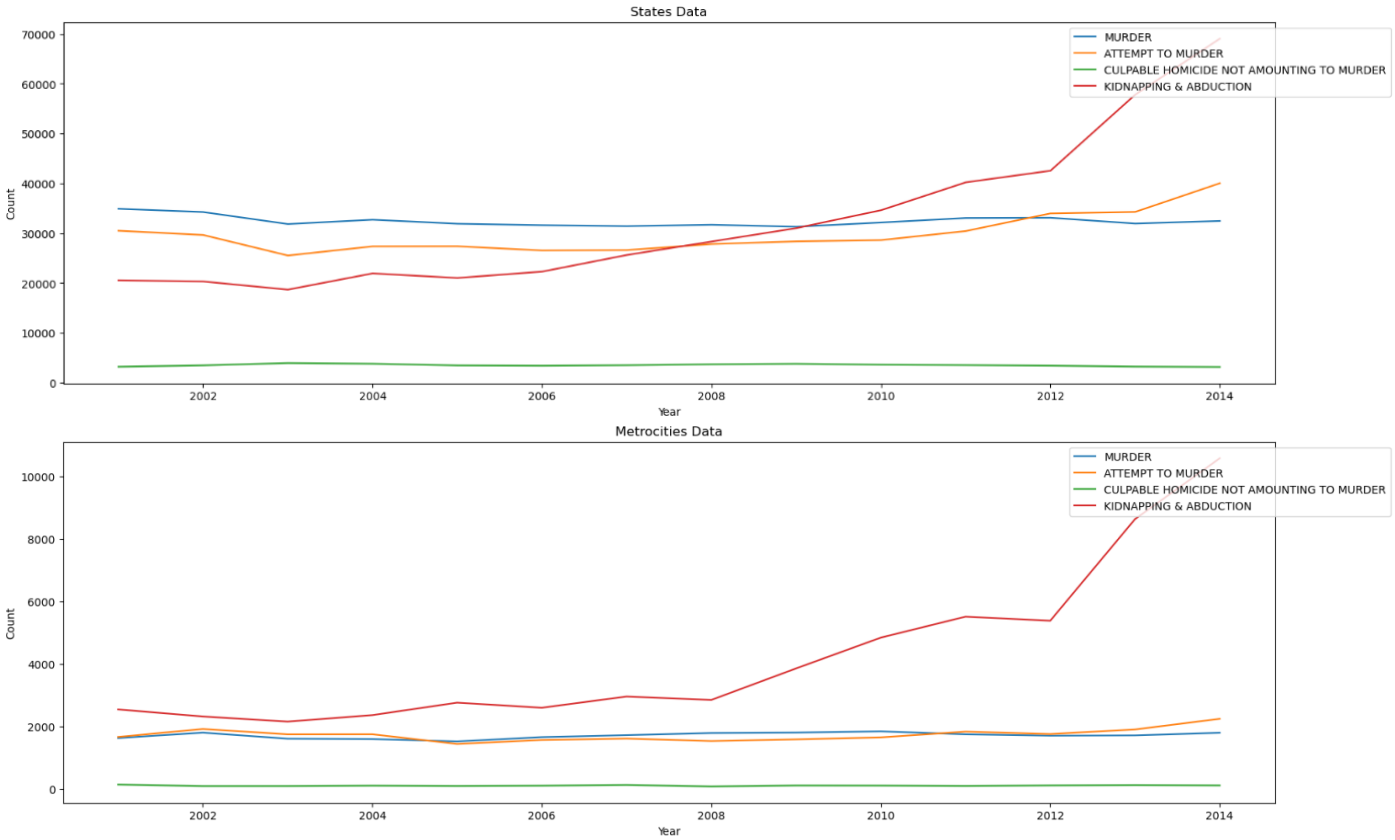
MURDER

TOTAL MURDER RATE IN 14 YEARS:- 462599

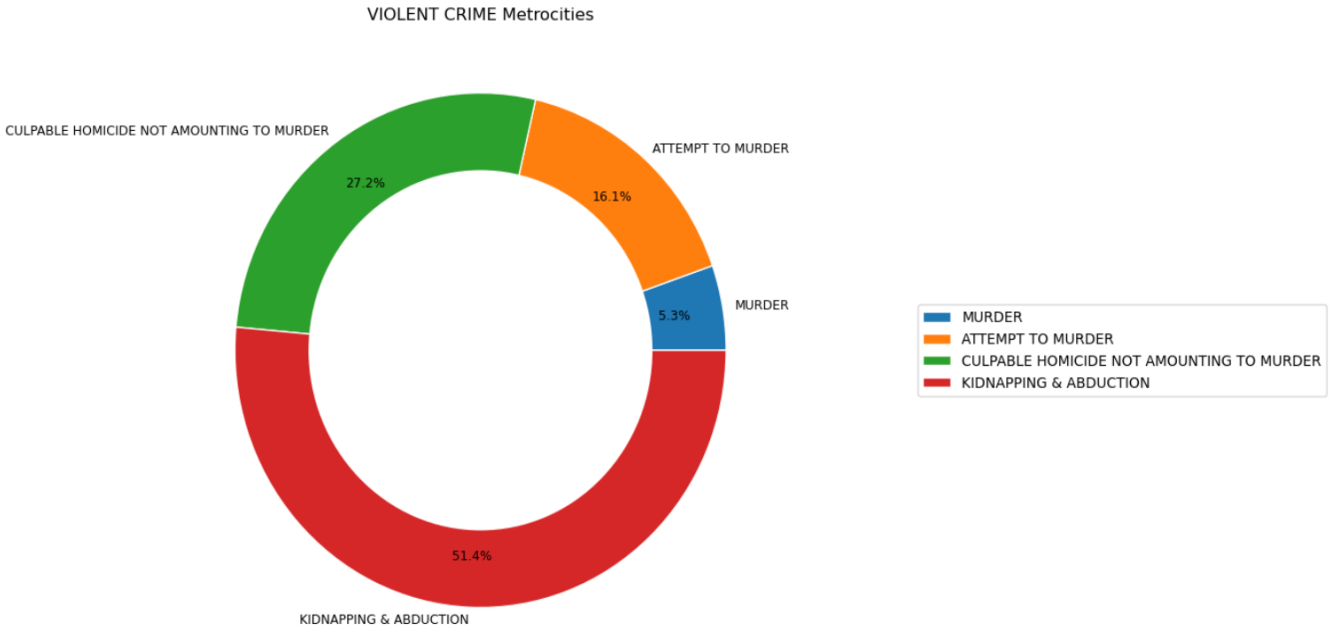
UTTAR_PRADESH: 75170 has maximun total Cases of MURDER

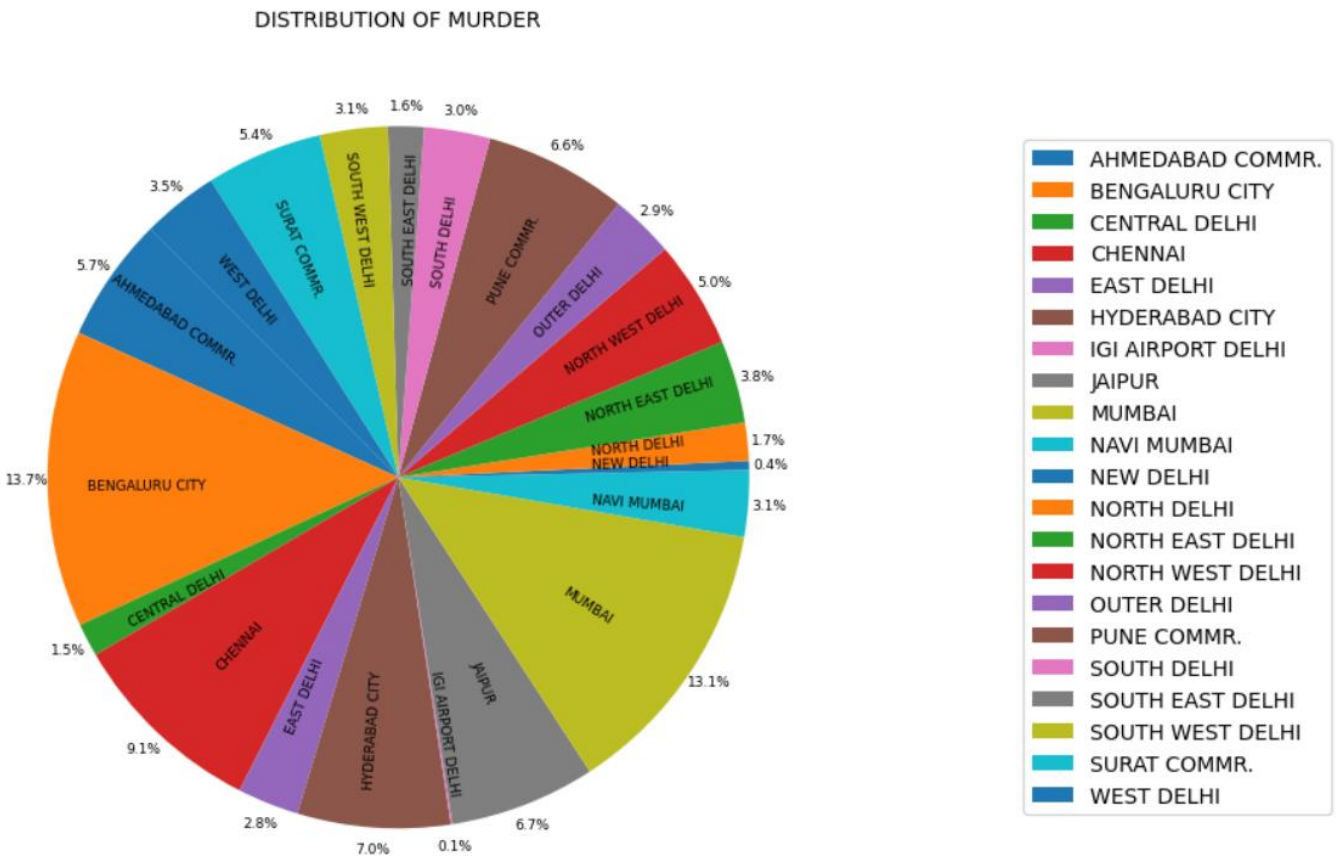


- *Multiveriant Projection Analysis of Violent crimes over the years of State data/Metricities data (all plots in jupyter notebook): -*



- *Multiveriant analysis Donut Plot Violent Crime distribution over the Metricities (all plots in jupyter notebook): -*





- **Model Selection:**

  - In our project, we conducted a comprehensive evaluation and comparison of multiple regression algorithms, namely Support Vector Regressor (SVR), Neural Network (NN), K-Nearest Neighbours (KNN), Random Forest (RF), and Linear Regression (LR). Our objective was to determine the most suitable algorithm for predicting crime rates with high accuracy.

  - To assess the performance of each algorithm, we employed various evaluation metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2) score. These metrics provided valuable insights into the predictive capabilities of each model and helped us gauge their effectiveness in capturing the underlying patterns in the data.

  - To optimize the performance of the chosen models further, we utilized techniques such as fine-tuning of hyperparameters. This involved employing grid search or randomized search methodologies to identify the optimal set of hyperparameters for each model. By fine-tuning the hyperparameters, we aimed to enhance the models' predictive accuracy and generalization capabilities.

  - To ensure the robustness of our models, we adopted a rigorous approach to dataset splitting. We partitioned the dataset into training and testing sets in an 80:20 ratio, allowing us to train the models on a sufficient amount of data while reserving a portion for evaluation purposes. Additionally, we implemented techniques like k-fold cross-validation during the training phase. This approach helped validate the models' performance across multiple subsets of the data, mitigating the risk of overfitting and ensuring their reliability in real-world scenarios.

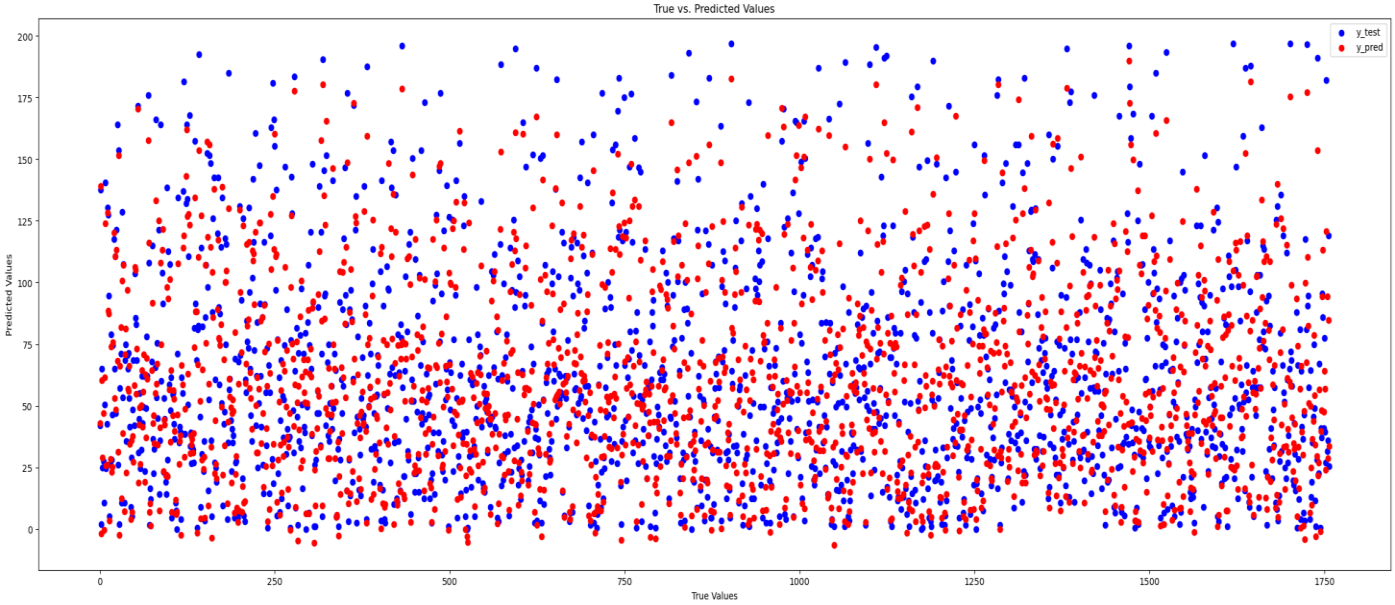
  - Finally, we trained the selected regression models on the training data using the optimized hyperparameters and state-of-the-art optimization techniques. By leveraging advanced machine learning methodologies and rigorous evaluation procedures, our project aimed to deliver accurate and robust predictions of crime rates, thereby assisting stakeholders in making informed decisions related to community safety and security.

➢ **Model Evaluation:**

1. Support Vector Regressor (SVR)

Mean Absolute Error (MAE): 11.76
Mean Squared Error (MSE): 296.90
Root Mean Squared Error (RMSE): 17.23
R-squared (R2) score: 0.86

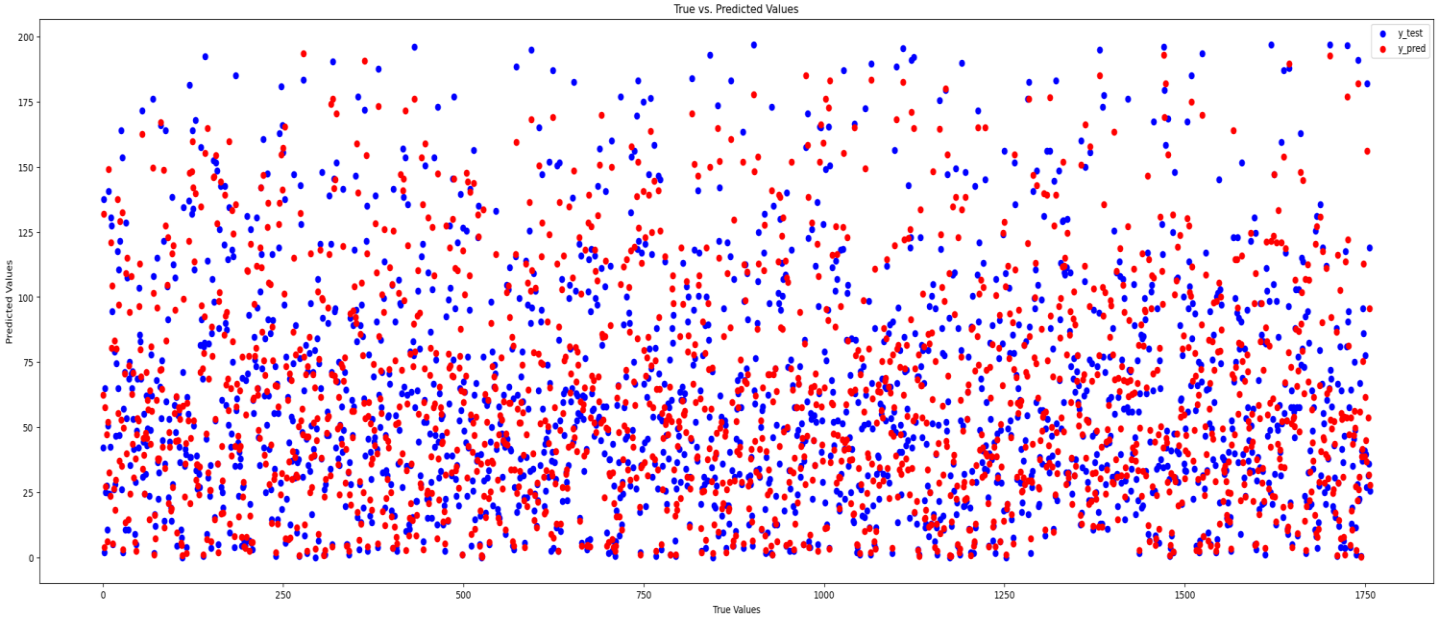
True Vs Predicted Values: -



2. K-Nearest Neighbours regressor (KNN)

Mean Absolute Error (MAE): 10.67
Mean Squared Error (MSE): 253.60
Root Mean Squared Error (RMSE): 15.92
R-squared (R2) score: 0.88

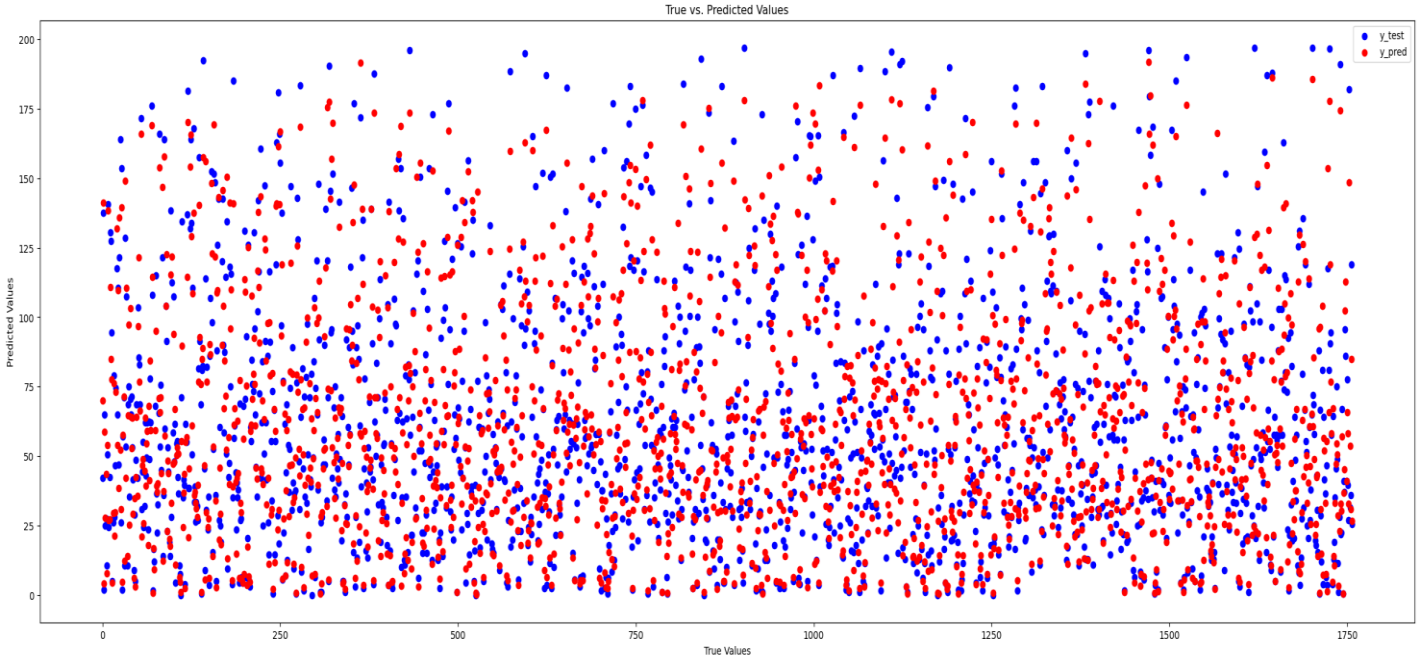
True Vs Predicted Values: -



3. Random Forest regressor (RF)

Mean Absolute Error (MAE): 9.76
Mean Squared Error (MSE): 215.53
Root Mean Squared Error (RMSE): 14.68
R-squared (R2) score: 0.90

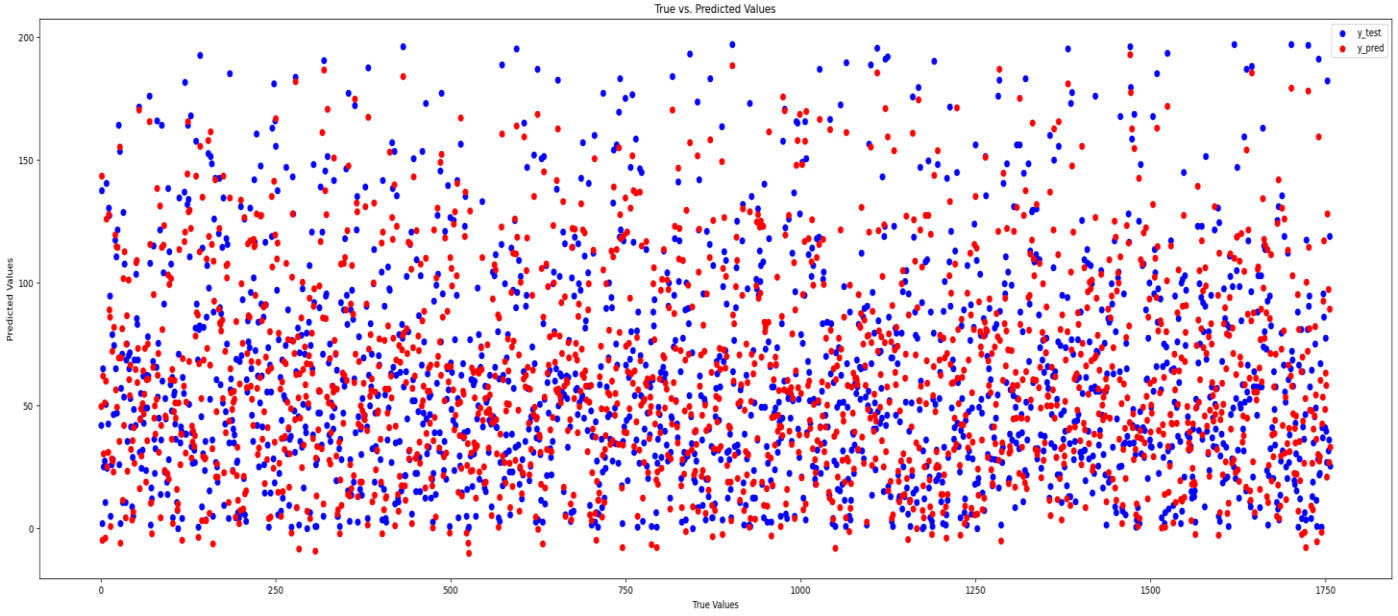
True Vs Predicted Values: -



4. Linear Regression: -

Mean Absolute Error (MAE): 11.98
Mean Squared Error (MSE): 285.41
Root Mean Squared Error (RMSE): 16.89
R-squared (R2) score: 0.86

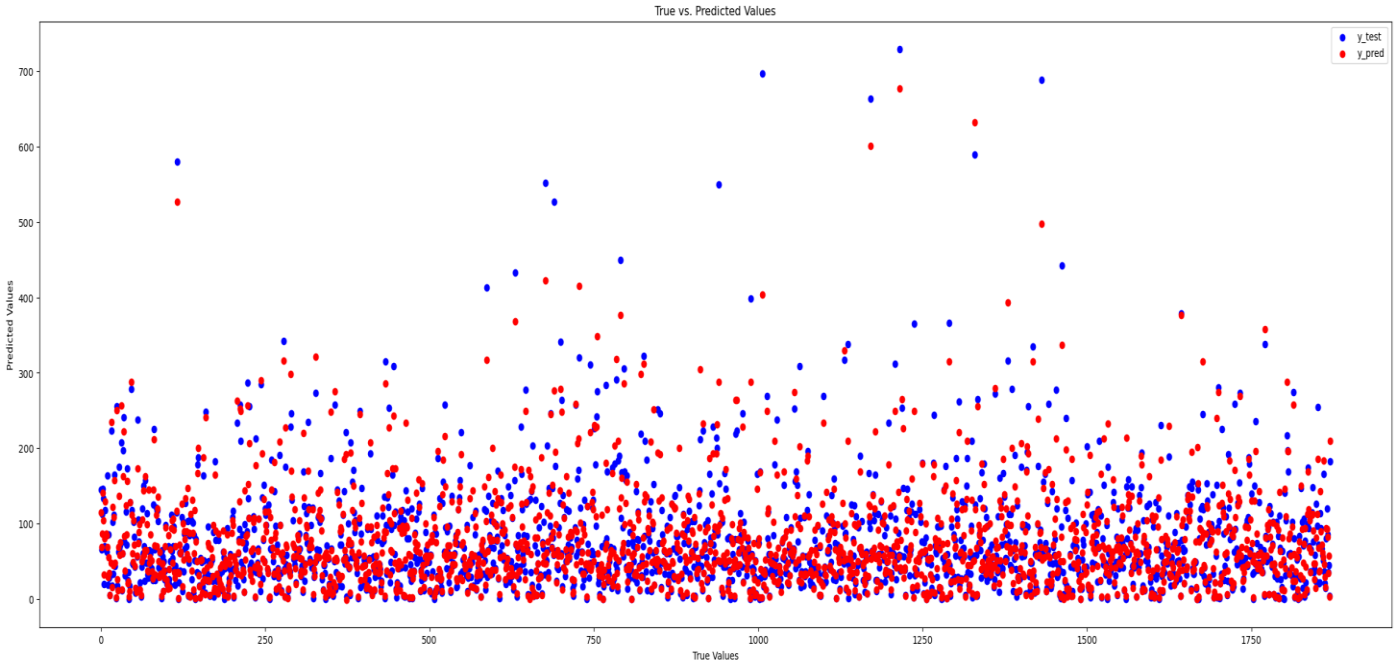
True Vs Predicted Values: -



5. Neural Network (NN)

Mean Absolute Error (MAE): 10.591938432793432
Mean Squared Error (MSE): 235.8002860487319
Root Mean Squared Error (RMSE): 15.36
R-squared (R2) score: 0.89

True Vs Predicted Values: -



CONCLUSION: -

- Among the regression algorithms evaluated, Random Forest (RF) and Neural Network (NN) emerged as the top performers based on key evaluation metrics.
- Random Forest exhibited slightly lower Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) compared to Neural Network, indicating its superior predictive accuracy.
- However, Neural Network demonstrated a slightly higher R-squared (R2) value, indicating its ability to explain a greater proportion of the variance in the target variable. In summary, while Random Forest excels in prediction accuracy, Neural Network offers a better balance between prediction accuracy and explanatory power.
- Thus, for our crime rate prediction task, Neural Network stands out as the preferred model due to its comprehensive performance across all metrics.

RECOMMENDATION: -

Based on our findings, we recommend the adoption of Neural Network for crime rate prediction tasks due to its comprehensive performance across various evaluation metrics. Neural Network not only achieves competitive prediction accuracy but also offers a deeper understanding of the underlying factors influencing crime rates. Moreover, its flexibility and adaptability make it well-suited for handling complex and dynamic datasets, ensuring reliable predictions in real-world scenarios.