Crimalytics

Neelkumar Patel
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
patel9t5@uwindsor.ca

Boond Marwaha
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
marwahab@uwindsor.ca

Param Patel
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
patel4q6@uwindsor.ca

Vraj Shah
Master of Applied Computing
University of Windsor
Windsor, ON, Canada
shah7g1@uwindsor.ca

***Abstract*—'Crimalytics' is a crime analysis and forecasting system designed for law enforcement in Toronto, a major Canadian city. By utilizing Ensemble learning with Random Forest & KNN, the system achieves an overall accuracy of 54.69% in making accurate predictions on crime occurrences, focusing on date-time and location/neighborhood data. The proposed system empowers law enforcement and the public with visualization capabilities to prevent crime in high-risk regions. Utilizing critical data features such as crime type, timestamp, and location, our data-driven approach offers actionable insights through a Power BI report, providing interactive visualizations of the crime data. Committed to excellence, 'Crimalytics' leads in data-centric predictive policing, contributing to safer communities. Our goal is to equip law enforcement with effective tools and foster public collaboration to combat crime proactively.**

Keywords—Crime analysis, Crime forecasting, Law enforcement, Ensemble learning, Random Forest classification, KNN, Prediction, Data-driven approach, Toronto, Crime probabilities, Power BI, Visualization capabilities, High-risk regions, Empirical method, Machine learning, Crime prevention, Supervised learning models, Criminal activity prediction, Data features, Safer communities, public safety, Crime Statistics, Crime prevention tools.

I. INTRODUCTION & MOTIVATION

The proposed solution aims to develop a crime analysis and prediction system that takes a data-driven approach to enable predictive policing. By leveraging historical crime data, this system anticipates criminal activity and empowers law enforcement agencies to proactively safeguard communities. Through comprehensive data collection, preprocessing, and advanced analytics, the system aims to provide accurate crime predictions and enhance understanding of crime hotspots. By continuously updating and evaluating the predictive models, the system ensures its relevance and effectiveness in contributing to the field of predictive policing.[1] This research aims to foster safer communities and equip law enforcement with valuable insights and tools to combat crime efficiently.

II. PROBLEM STATEMENT

Ensuring crime prevention and public safety are paramount for both communities and law enforcement agencies. However, the absence of comprehensive crime data analysis to identify trends and patterns can result in unreliable crime rate predictions, impacting public safety.[2] Thus, a critical necessity arises for a robust crime analysis and prediction system that can offer accurate forecasts based on historical crime data, pinpoint crime hotspots, and contribute to current law enforcement strategies effectively.[1] By addressing this need, the proposed system seeks to bolster public safety, empower law enforcement with precise insights, and foster a proactive approach to crime prevention.

III. RELATED WORK

Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention:

Authors recently investigated ML-based crime predictions in a study. In Vancouver, Canada, crime statistics from the previous 15 years were examined to make predictions. This crime analysis is based on machine learning and encompasses data gathering, categorization, pattern recognition, forecasting, and visualisation. The crime dataset was further examined using boosted decision tree and K-nearest neighbour (KNN) algorithms. The authors examined a dataset made up of multiple crimes and made predictions about the kinds of crimes that would happen soon considering certain circumstances. The accuracy of the crime prediction made by ML algorithms ranged from 39% to 44%. [2]

A study on predicting crime rates through machine learning and data mining using text:

This is the study containing a comprehensive description of studies on crime forecasting and prediction that used machine learning and data mining methods. Different algorithms were used to forecast crime trends in various cities, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Deep Neural Network (DNN), and others. In Columbus and the southern US states, SVM and LR were used to forecast crime locations. SVM proved successful when using a clustering strategy, however LR had trouble handling huge geo-areas. In the southern US states, RF with the SmoteR algorithm was used to forecast crime, but its accuracy remained subpar at 59.8%.[3]

Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India:

This study examines different crime types, their occurrences in various settings, and the dominating characteristic of murder in order to detect crimes using machine learning and data mining. On train and test data, the Scaled algorithm outperformed the Bayesian and Levenberg algorithms, which were compared. According to statistical study, the crime rate has decreased by 78%, with an accuracy of 0.78. The study plans to compare different machine learning techniques, such as deep learning and genetic algorithms, for improving crime detection on larger datasets in their future work.[4]

IV. PROPOSED MODEL

The crime prediction system uses a multi-model ensemble learning technique to increase the accuracy and dependability of crime forecasts. The system collects information from numerous sources, separates out pertinent properties, and goes through a rigorous model training and optimization procedure. Here is a basic outline of the steps we'll take to build a system that can forecast the likelihood that a crime will occur depending on input values.

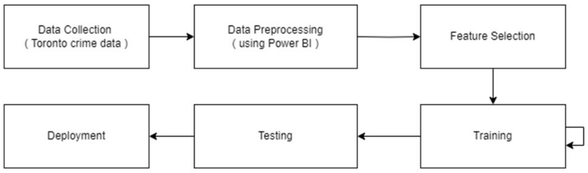


Fig 1. Proposed solution flow diagram

a) Data Collection: Continuing using the Toronto crime data as the primary dataset. Additionally, considering integrating other relevant data sources, such as socioeconomic data, weather data, demographics, and any other data that might influence crime patterns.[6]
b) Data Preprocessing: Handling missing values, duplicates, and inconsistencies in the data using Power BI. Applying feature engineering techniques to create new relevant features. For time series forecasting, ensuring that the data is organized into chronological order.
c) Feature Selection: Utilizing statistical testing, correlation analysis, and domain knowledge to select the most relevant features for each individual model. Taking into account the features suitable for time series forecasting and crime prediction using Random Forest and KNN.
d) Model Training: Training the three models separately: Random Forest, KNN, and the time series forecasting model using ARIMA. Each model will learn to predict crime occurrences based on different aspects of the data.
e) Ensemble Learning: Combining the predictions from the three individual models using an ensemble learning technique. Popular ensemble methods include Voting (e.g., Majority Voting). Deciding on the most suitable ensemble method for your specific use case.
f) Testing:
   I. Model Evaluation: We'll evaluate the Arima models using appropriate metrics, like mean absolute error (MAE) or R-squared. These metrics will indicate the correctness of the model and the degree to which it fits the training dataset. Similar metrics for classification tasks including recall, accuracy, and F1 score will be utilized to evaluate the KNN, Random Forest model. With the help of this evaluation, we can see how effectively the model reproduces the underlying patterns in the dataset.
   II. Model optimization: We will do model optimization in addition to model evaluation to improve the performance of the models. Adjusting hyperparameters particular to each model, such as the regularization parameter, may be required during this process. The optimization method will probably

involve iterative testing; to ensure continuous progress, we will carefully examine the outcomes at each iteration and compare them to the evaluation metrics.

g) Deployment: In the deployment phase, the trained models will be implemented in a production environment. This involves integrating the models into the system using Django framework and where we have a UI with which user can predict.

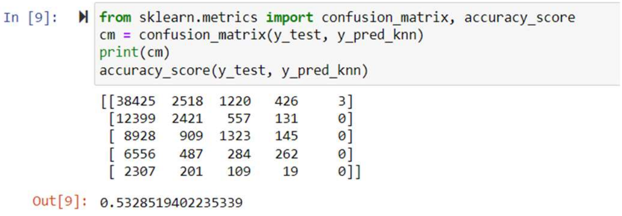
V. RESULT

1. Prediction:

   In this section, we present the results of our ensemble learning approach, combining the Random Forest and K-Nearest Neighbors (KNN) models. The ensemble model was trained on a dataset consisting of crime date-time neighborhood and to predict the type of crime occurred on any future data.

1.1 Ensemble learning Accuracy:

The performance of the Random Forest classification model was evaluated using various metrics, and the accuracy achieved was 54.81%. The KNN model has an accuracy of 53.28%. These two model's results are then used with the ensemble learning technique to get better representing the proportion of correctly predicted instances over the total number of instances in the test dataset. This accuracy indicates the effectiveness of the model in capturing the underlying patterns and making accurate predictions. This model is made to predict future data unlike other current models.[5]

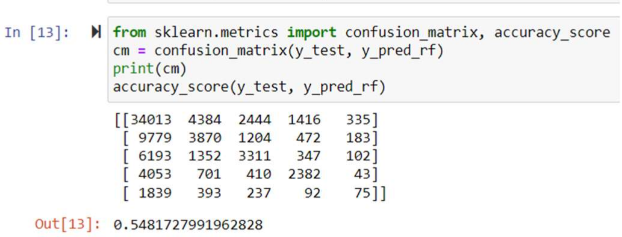


```
In [9]:  from sklearn.metrics import confusion_matrix, accuracy_score
         cm = confusion_matrix(y_test, y_pred_knn)
         print(cm)
         accuracy_score(y_test, y_pred_knn)

         [[38425  2518  1220   426     3]
          [12399  2421   557   131     0]
          [ 8928   909  1323   145     0]
          [ 6556   487   284   262     0]
          [ 2307   201   109    19     0]]

Out[9]:  0.5328519402235339
```

Fig 2. Accuracy of KNN



```
In [13]:  from sklearn.metrics import confusion_matrix, accuracy_score
          cm = confusion_matrix(y_test, y_pred_rf)
          print(cm)
          accuracy_score(y_test, y_pred_rf)

          [[34013  4384  2444  1416   335]
           [ 9779  3870  1204   472   183]
           [ 6193  1352  3311   347   102]
           [ 4053   701   410  2382    43]
           [ 1839   393   237    92    75]]

Out[13]:  0.5481727991962828
```

Fig 3. Accuracy of Random Forest

1.2 Time Series Forecasting:

In this section, we present the results of our ARIMA time series forecasting model. The model was applied to the same dataset to predict next month's future crime counts. To analyze the model, we have calculated the Mean Absolute error (MAE) which is 2.15. The results are shown in the below graph.

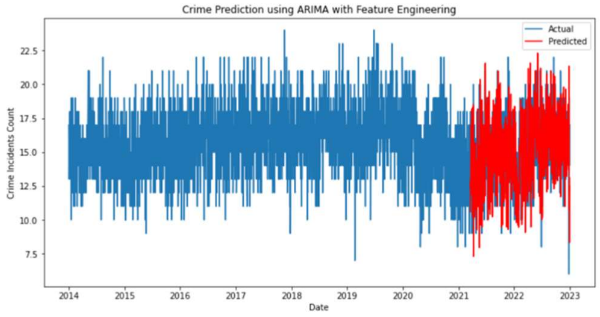
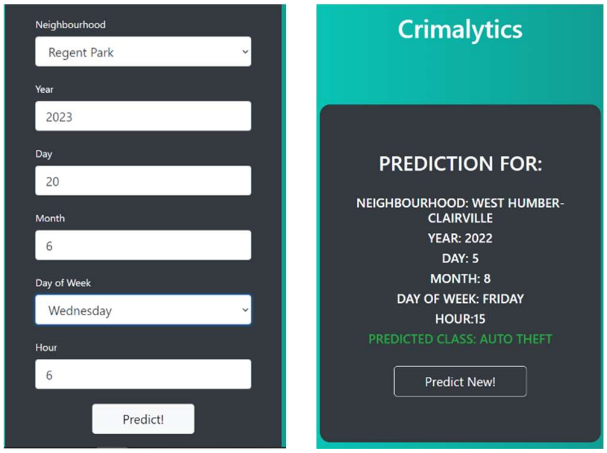


Fig 4. Time series Forecasting

We have also created a web-application using Django framework which gives user to give input like day, month, year and neighborhood in Toronto city with this information, model in the backend will use this data to predict type of crime and we also have added the option to view the analytical report of previous crime data created using Power BI.



Neighbourhood
Regent Park

Year
2023

Day
20

Month
6

Day of Week
Wednesday

Hour
6

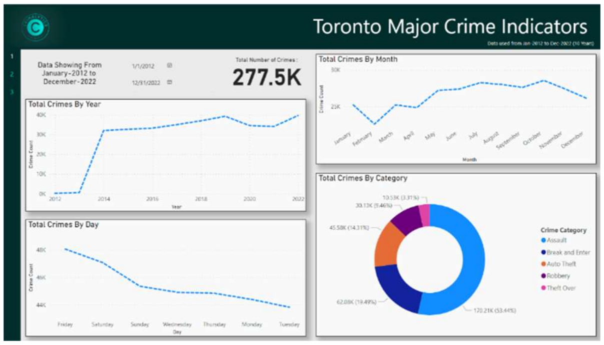
PREDICTION FOR:

NEIGHBOURHOOD: WEST HUMBER-CLAIRVILLE
YEAR: 2022
DAY: 5
MONTH: 8
DAY OF WEEK: FRIDAY
HOUR:15
PREDICTED CLASS: AUTO THEFT

Fig 5. Client Interface

2. Data Analysis:

   An analytical report is created on the Toronto Crime Dataset using Microsoft Power BI. The report created on the Toronto Crime Dataset likely included visualizations of the data, such as charts and graphs, to help users better understand the patterns and trends in the data. The report may have also included information on the types of crimes committed, the frequency of crimes in different areas of the city, and other relevant data points. Overall, the Power BI analytical report on the Toronto Crime Dataset likely provided valuable insights into crime patterns and trends in the city, which could be used to inform law enforcement and public policy decisions.

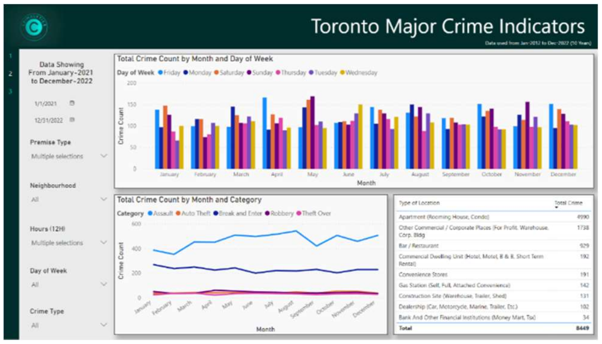


Data Showing From January-2012 to December-2022

1/1/2012

12/31/2022

Fig 6. Analytics report page - 1


Toronto Major Crime Indicators
Data used from Jan-2012 to Dec-2022 (10 Years)
1
2
3
Total Crime Count by Month and Day of Week
Day of Week ● Friday ● Monday ● Saturday ● Sunday ● Thursday ● Tuesday ● Wednesday
200
150
100
50
0
Crime Count
January
February
March
April
May
June
July
August
September
October
November
December
Month
Total Crime Count by Month and Category
Category ● Assault ● Auto Theft ● Break and Enter ● Robbery ● Theft Over
600
400
200
0
Crime Count
Month

Data Showing From January-2021 to December-2022

1/1/2021

12/31/2022

Premise Type: Multiple selections

Neighbourhood: All

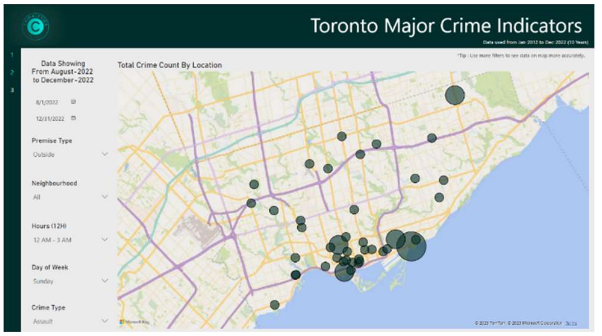
Hours (12H): Multiple selections

Day of Week: All

Crime Type: All

Type of Location	Total Crime
Apartment (Rooming House, Condo)	4990
Other Commercial / Corporate Places (For Profit, Warehouse, Corp. Bldg	1738
Bar / Restaurant	929
Commercial Dwelling Unit (Hotel, Motel, B & B, Short Term Rental)	192
Convenience Stores	191
Gas Station (Self, Full, Attached Convenience)	142
Construction Site (Warehouse, Trailer, Shed)	131
Dealership (Car, Motorcycle, Marine, Trailer, Etc.)	102
Bank And Other Financial Institutions (Money Mart, Tsx)	34
Total	**8449**

Fig 7. Analytics report page - 2


Toronto Major Crime Indicators
Data used from Jan-2012 to Dec-2022 (10 Years)
1
2
3
*Tip : Use more filters to see data on map more accurately.
Total Crime Count By Location
Microsoft Bing
© 2023 TomTom, © 2023 Microsoft Corporation Terms

Data Showing From August-2022 to December-2022

8/1/2022

12/31/2022

Premise Type: Outside

Neighbourhood: All

Hours (12H): 12 AM - 3 AM

Day of Week: Sunday

Crime Type: Assault

Fig 8. Analytics report page - 3

VI. LIMITATIONS OR CHALLENGES

In our research, we conducted crime prediction using ensemble learning, combining the Random Forest and K-Nearest Neighbors (KNN) models, along with time series forecasting using the ARIMA model. However, we encountered certain limitations that could have influenced our findings. The dataset's imbalance, containing only positive instances of crime occurrences, introduced bias in our model towards predicting positive occurrences. As well as 50-60% of total data is of only one class category which is creating difficulties in training the model and model is just getting trained to predict just one value that is resulting in low accuracy. Additionally, the quality of the crime data, model assumptions, lack of external validation, and interpretability challenges were important constraints. Despite these limitations, our research provided valuable insights, suggesting the need for addressing imbalanced datasets and exploring alternative methodologies for more robust crime prediction and forecasting applications.

In our experiment, we initially expected the ensemble learning approach, combining Random Forest and K-Nearest Neighbors (KNN) models, to significantly outperform the individual models in crime prediction. However, the results showed no difference due to the biased dataset, only positive values of crimes, might have influenced the ensemble's performance, leading to a difference between our anticipated and actual results.[2]

Similarly, for time series forecasting using the ARIMA model, we anticipated highly accurate predictions capturing underlying crime data patterns. Yet, the forecasting results displayed good enough accuracy. Despite these disparities from our initial expectations, the research outcomes still provide valuable insights into the complexities of crime prediction and time series forecasting. The research underscores the potential for further exploration and refinement of ensemble learning and time series forecasting techniques in the context of crime prediction.

VII. CONCLUSION

The crime prediction model using ensemble learning with Random Forest and KNN, along with the analysis using Power BI report, has provided valuable insights into crime patterns and trends. By leveraging the power of ensemble learning, which combines multiple algorithms, including Random Forest and KNN, the model was able to make accurate predictions on crime occurrences with overall accuracy of 54.69%. The Power BI report complements the crime prediction model by providing interactive visualizations of the crime data. Ensemble learning allows for the consideration of different perspectives and approaches, resulting in more robust and reliable predictions. Overall, the combination of ensemble learning with Random Forest and KNN, along with the analysis using Power BI report, has provided valuable insights into crime patterns and trends. This information can be used to inform law enforcement strategies, allocate resources effectively, and implement preventive measures to ensure public safety. The integration of advanced analytics techniques and interactive reporting tools like ensemble learning and Power BI demonstrates the power of data-driven approaches in addressing complex problems such as crime prediction.

VIII. FUTURE WORK

In future work, there are several opportunities to enhance the research presented in this report. Firstly, addressing the dataset's imbalance through advanced data augmentation techniques and improving feature engineering could lead to more accurate crime prediction and time series forecasting. Additionally, exploring advanced ensemble learning methods, such as stacking or weighted voting, and integrating temporal and geospatial analysis could further leverage the strengths of different models and improve predictive capabilities.

Furthermore, investigating real-time crime prediction and deploying the model in real-world settings could provide valuable insights for law enforcement agencies. Additionally, exploring long-term forecasting and considering temporal feature engineering may enable capturing crime trends over extended periods. By pursuing these future research avenues, researchers can contribute to the continuous improvement and application of crime prediction and time series forecasting techniques in enhancing public safety and crime prevention efforts.

IX. REFERENCES

1. Crime analysis and prediction using data mining," Crime analysis and prediction using data mining | IEEE Conference Publication | IEEE Xplore. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6906719

2. N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention - Visual Computing for Industry, Biomedicine, and Art," SpringerOpen, Apr. 29, 2021. [Online]. Available: https://vciba.springeropen.com/articles/10.1186/s42492-021-00075-z
3. R. M. Saeed and H. A. Abdulmohsin, "A study on predicting crime rates through machine learning and data mining using text," *De Gruyter*, Jan. 01, 2023. [Online]. Available: https://www.degruyter.com/document/doi/10.1515/jisys-2022-0223/html?lang=en
4. "Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India," *Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India - ScienceDirect*, Jun. 16, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920313417
5. "Random Forest&reg;: A Criminal Tutorial - KDnuggets," *KDnuggets*. [Online]. Available: https://www.kdnuggets.com/random-forest-a-criminal-tutorial.html
6. "Toronto Police Service Public Safety Data Portal," *Toronto Police Service Public Safety Data Portal*. [Online]. Available: https://data.torontopolice.on.ca/search?q=crime