CRIME DATA ANALYTICS

Divya K
PES1UG20CS135
Dept. of CSE, PES University
divya110702@gmail.com

Dhanalakshmi K M
PES1UG20CS124
Dept. of CSE, PES University
kmdhanalakshmi25@gmail.com

Syed Azfar Rayan
PES1UG20CS453
Dept. of CSE, PES University
munzernouman@gmail.com

Abstract— **A crime is any unlawful action punishable by a state, and often creates a huge social impact. Our project aims on using data analytic tools for crime prediction and uncovering trends and patterns in crimes. By identifying crime hotspots and early warnings, police services can be increased in those areas to effectively prevent occurrences of further crimes. Various machine learning techniques will be used for trend identification, prediction, and visualization.**

I. Introduction

Safety of its citizens is a matter of prime importance for governments and the police. Crimes have a detrimental social and economic impact on society. Hence, crime forecasting can prove to be a useful tool to keep communities better-informed of crime patterns and warn people living in areas vulnerable to a particular crime. When attempting to stop crimes, law enforcement agencies are faced with several difficulties. To help law enforcement agencies undertake descriptive, predictive, and prescriptive analysis on crime data, we suggest a Crime Data Analytic Platform. Keeping this in mind, our project aims to address questions such as:
-Which are the major crime indicator categories?
-What types of crimes are most frequently committed and what are the trends in the crimes?
-Which hours of the day do these crimes occur and is there a pattern?
-Which months and which parts of the month are crimes likely to occur?
-Which are the crime hotspots in the city and the safest neighborhoods in the city?
-Which days of the week are crimes most prevalent?
-Are there different trends observed for different categories of crime?
-Can we predict the likely crimes before it happens?
Our proposed approach is the supervised prediction technique of classification for building a predictive model that can predict the category of crimes. Algorithms such as Decision tree, KNN Classifier, Naïve Bayes, and Random Forest will be tested in order to identify the best performing model for crime prediction. We also use clustering to

II. DATASET

Toronto Crime dataset on major crime indicators has been used for the year of 2021. This dataset includes all Major Crime Indicators (MCI) occurrences by reported date and related offenses. The MCI categories include Assault, Break and Enter, Auto Theft, Robbery and Theft Over. It is published on the Toronto police public safety data portal. *[1]*

Major Crime Indicators (MCI) - Data Field Descriptions

Field	Field Name	Description
1	Index	Unique Identifier
2	event_unique_id	Offence Number
3	Division	Police Division where Offence Occurred
4	occurrence_date	Date of Offence
5	reporteddate	Date Offence was Reported
6	location_type	Location Type of Offence
7	premises_type	Premises Type of Offence
8	ucr_code	UCR Code for Offence
9	ucr_ext	UCR Extension for Offence
10	Offence	Title of Offence
11	reportedyear	Year Offence was Reported
12	reportedmonth	Month Offence was Reported
13	reportedday	Day of the Month Offence was Reported
14	reporteddayofyear	Day of the Year Offence was Reported
15	reporteddayofweek	Day of the Week Offence was Reported
16	reportedhour	Hour Offence was Reported
17	occurrenceyear	Year Offence Occurred
18	occurrencemonth	Month Offence Occurred
19	occurrenceday	Day of the Month Offence Occurred
20	occurrencedayofyear	Day of the Year Offence Occurred
21	occurrencedayofweek	Day of the Week Offence Occurred
22	occurrencehour	Hour Offence Occurred
23	MCI	MCI Category of Occurrence
24	Hood_ID	Identifier of Neighbourhood
25	Neighbourhood	Name of Neighbourhood
26	Long	Longitude Coordinates (Offset to nearest intersection)
27	Lat	Latitude Coordinates (Offset to nearest intersection)

Table 1

III. LITERATURE REVIEW

A. Crime Analyses Using Data Analytics [2]

- Developed a supervised learning model to predict arrest status based on Chicago crime data for three consecutive years (2017-2019)

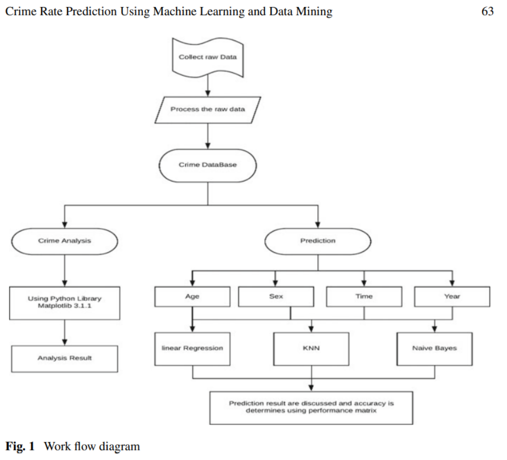
- Employed three classification algorithms: probabilistic-Naïve Bayes, Rule Induction-Repeated Incremental Pruning to Produce Error Reduction (RIPPER), and Decision Tree- C4.5 on features identified using feature selection methods.

-Synthetic Minority Oversampling Technique (SMOTE) applied before the training phase to balance the class in the dataset.

- The results obtained after data sampling by the machine learning algorithms were surprisingly low with C4.5 being superior with classification accuracy of 63.5%.

B. *Crime Rate Prediction Using Machine Learning and Data Mining [3]*

In this paper, the major focus is on analyzing crime indicators and identifying the hotspots in the city which eventually tells the safe routes to travel.



Crime Rate Prediction Using Machine Learning and Data Mining 63

Fig. 1 Work flow diagram

Three major algorithms have been compared in this research paper

- Linear Regression:

Multi-linear regression is a sort of mathematical approach to finding a relation between the dependent variables (Victim age) and a set of independent variables whose input values gathered from the crime spot. This methodology predicts the Era of the victims age values based on the input characteristics indicated in the metadata column.

- KNN:

K-nearest neighbors is used when the target variable must be classified in more than two classes.

In this paper the dataset has three classes of target variable perpetrator sex: male, female, and unknown and similarly, three categories of young, old and kid are defined in age, hence KNN is used.

- Naïve Bayes:

Naive Bayes algorithms are mostly used to determine emotions, delete spam, suggestions, etc. which are fast and easy to implement, but their biggest downside is the need for autonomous predictors. In most real-life situations, the predictors are dependent, hampering the output of the classifier.

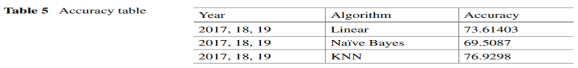


Table 5 Accuracy table

Year	Algorithm	Accuracy
2017, 18, 19	Linear	73.61403
2017, 18, 19	Naïve Bayes	69.5087
2017, 18, 19	KNN	76.9298

C. *Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques [4]*

-The major focus in this study applied different machine learning algorithms, namely, the logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), and time series analysis by long-short term memory (LSTM) and autoregressive integrated moving average (ARIMA) model to better fit the crime data.

- The performance of LSTM for time series analysis was reasonably adequate in order of magnitude of root mean square error (RMSE) and mean absolute error (MAE)

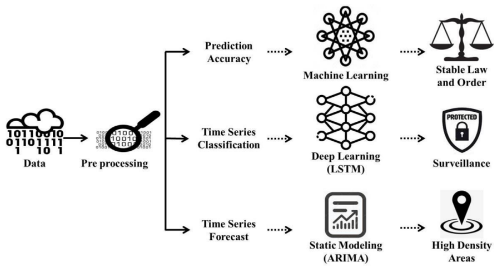


FIGURE 1. Proposed methodology and study framework.

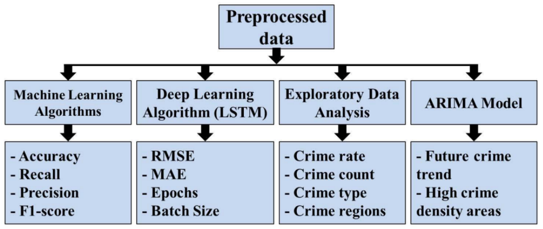
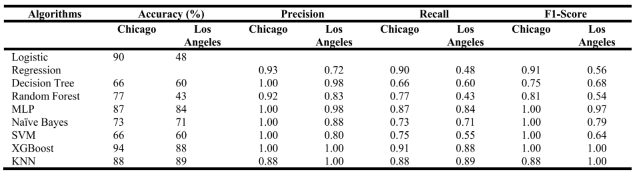


FIGURE 2. Experimental flow chart.

- Predictive accuracy:

TABLE 1. Performance parameters for Chicago and Los Angeles datasets.



Algorithms	Accuracy (%)		Precision		Recall		F1-Score	
	Chicago	Los Angeles	Chicago	Los Angeles	Chicago	Los Angeles	Chicago	Los Angeles
Logistic Regression	90	48	0.93	0.72	0.90	0.48	0.91	0.56
Decision Tree	66	60	1.00	0.98	0.66	0.60	0.75	0.68
Random Forest	77	43	0.92	0.83	0.77	0.43	0.81	0.54
MLP	87	84	1.00	0.98	0.87	0.84	1.00	0.97
Naïve Bayes	73	71	1.00	0.88	0.73	0.71	1.00	0.79
SVM	66	60	1.00	0.80	0.75	0.55	1.00	0.64
XGBoost	94	88	1.00	1.00	0.91	0.88	1.00	1.00
KNN	88	89	0.88	1.00	0.88	0.89	0.88	1.00

Among the different algorithms, XGBoost achieves the maximum accuracy on Chicago datasets and KNN achieves the maximum accuracy on Los Angeles.

- Time Series Analysis through LSTM:

LSTM is an elegant variation in the RNN architecture, which is an approach that can be applied to model sequential data. The structure of LSTM makes it an effective solution to combat the vanishing gradient problem of RNNs.

- Forecasting with an ARIMA model:

The objective of an ARIMA analysis is to determine the best predictive performance for the data of interest.

D. *Crime Analysis and Prediction Using Data Mining [5]*

Different steps are involved in doing the crime analysis:

- Classification

For classification we are using an algorithm called Naïve Bayes which is a supervised learning method as well as a statistical method for classification. The advantage of using Naive Bayes Classifier is that it is simple, and converges quicker than logistic regression. Compared to other algorithms like SVM (Support Vector Machine) which takes a lot of memory, the ease of implementation and high performance makes it different from other algorithms.

- Pattern Identification

Pattern identification phase where we have to identify trends and patterns in crime. For finding crime patterns that occur frequently we are using Apriori algorithm. So, crime occurs only if particular patterns occur on a day.

- Prediction:

For prediction, a decision tree concept. The main advantage of using a decision tree is that it is simple to understand and interpret. The other advantages include its robust nature and also it works well with large data sets. This feature helps the algorithms to make better decisions about variables.

Classification is done based on the Bayes theorem which showed more than 90% accuracy.

The problem is that the paper is not predicting the time in which the crime is happening. Since time is an important factor in crime, we have to predict not only the crime prone regions but also the proper time.

E. Paper on Different Approaches for Crime Prediction system [6]

The paper constitutes of the following techniques for Crime Prediction: a) Data mining technique b) Crime cast technique c) Deep learning technique

- Data mining technique works by analyzing trends & patterns from crime databases that are already stored. The authors have focused on the Association Rule Mining, Classification, Clustering techniques to carry out the analysis by Data Mining.

- Crime Cast is a simulation technique that finds the crime rate, type of crime, & predicts for the future using the previous database. In this technique, a region with a higher crime rate is to be found and is called Hotspots. The region with the lower crime rate is called Coldspots. Crime casts can be introduced in the hotspots by simulating probabilistic model implementation and ANN.

- *Deep Learning* method was also used. This comprises multiple layers which includes nonlinear operations. Deep Learning uses Algorithms to convert the raw info to higher representations. The Graph Loader component loads the graph data store for analyzing events which predicts the time & space of the crime events. The deep learning model gives very inaccurate results when a small dataset is used.

F. CRIME RATE PREDICTION [7]

-The objective was to train a model for prediction. That was done using a training data set which will be validated using the test dataset. The Multi Linear Regression (MLR) was used for crime prediction.

-The concept of MLR was implemented by the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable).

- Linear Regression and Logistic Regression models were also tested, but the MLR produced the minimal error.

- Logistic Regression & KNN was used to implement MLR on the used dataset. Logistic regression classified the data into 2 broad categories as 0 or 1. In KNN, the output is a class membership. It is classified by the plurality vote of k nearest neighbors of the object. MLR is an extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable.

IV. PROPOSED SOLUTIONS

EDA and Visualizations

Irrelevant columns such as X,Y,Object_ID have been removed as they do not provide any information. There are no empty fields in the data. Dates have been converted to datetime and extra spaces have been stripped from the data. Various diagrams have been plotted to come up with findings.

Data Preprocessing

Feature selection is a data preprocessing technique for selecting a subset of the best variables prior to constructing a model. It helps to remove irrelevant variables i.e., variables that do not share a strong relationship with the target variable. The dataset is split into 70% training set and 30% test set. Feature selection is performed on the training set.

SMOTE is an oversampling technique that generates synthetic samples from the minority class. It is used to obtain a synthetically class-balanced or nearly class-balanced training set, which is then used to train the classifier.

Modeling

Clustering is the grouping of a set of data in such a way that data in the same group(cluster) are very similar to one another than the data that are in the other clusters. It is a very popular way of grouping variables. This allows us to understand & organize data in a better fashion. In addition, clustering is an integral part of data visualization. Here, we have employed K - Means Clustering.

K - means clustering

It partitions the data into k numbers of clusters in which each data observed is assigned to the nearest centroid. The algorithm does this iteratively till the location of the cluster in two successive iterations doesn't change. The number of clusters can be specified and the algorithm partitions the data into the required number of clusters.. This process will be repeated until all the data is assigned to a cluster. Heuristic-based clustering approach was taken to define police district boundaries in a way that the identified districts have equitable distribution of criminal cases with compact shape.

Classification is a supervised prediction technique. Different classification algorithms are considered. More specifically, algorithms namely Decision tree, KNN Classifier, Naïve Bayes, Adaboost and Random forest ensemble model are tested, compared and evaluated in order to identify the best performing model for crime prediction.

Decision tree classification model forms a tree structure from a dataset. Decision tree is built by dividing a dataset into smaller pieces. At each step in the algorithm, a decision tree node is splitted into two or more branches until it reaches leaf nodes. Leaf nodes indicate the class labels or results. At each step, the decision tree chooses a feature that best splits the data with the help of two functions: Gini Impurity and Information Gain. Gini Impurity measures the probability of classifying a random sample incorrectly if the label is picked randomly according to the distribution in a branch.

Gaussian Naive Bayes is a supervised classifier that uses the naive assumption that there is no dependency between two features. This classifier is implemented by applying Bayesian Theorem.

Logistic regression uses linear boundaries to classify data into different categories. Logistic regression can work on both binary and multiclass problems. For a multiclass dataset, one vs the rest scheme is used. In this method, logistic regression trains separate binary classifiers for each class. Meaning, each class is classified against all other classes, by assuming that all other classes are one category.

Ensemble learning is a method of combining multiple learning algorithms together to achieve better performance over a single algorithm. In a random forest ensemble model several decision trees are built using samples drawn with replacement from the training set. The splitting of each node of a tree is not based on the best split of all features, rather the best split among a random set of features.

Adaboost or Adaptive Boosting is a boosting algorithm. Adaboost combines several weak learners to produce a stronger model. The final output is obtained from the weighted sum of the weak models. As it is a sequential process, in each step a weak learner is changed in favor of misclassified data points in previous classifiers.

V. RESULTS AND CONCLUSIONS

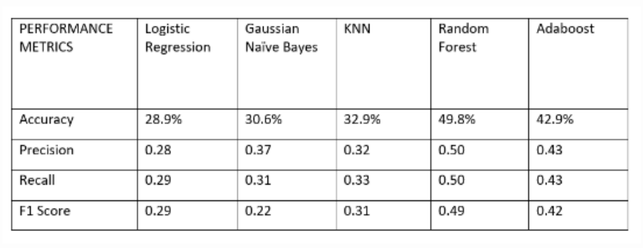
Our goal of the project was to compare various classification as well as clustering models to check which model would work better.

Metrics used include confusion matrix which returns a table layout that helps to visualize the performance of an algorithm rather than producing a numerical value that indicates the goodness of the algorithm. Precision and recall is found from this. Accuracy measures how many predictions are matched exactly with the actual or true label of the testing dataset and returns the percentage of correct results.Log loss is used to measure performance of classifiers by penalizing false classifications.

Key findings from EDA and Visualizations:

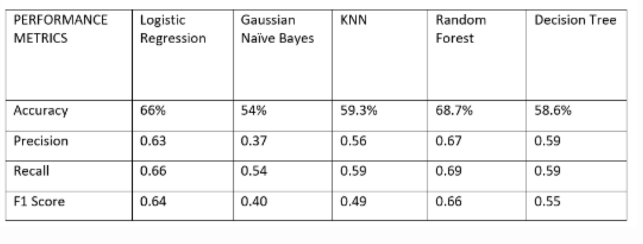
- In Toronto, the average number of crimes per month is 2856, and the average number of crimes per day is 93.
- Monday and Tuesday see lower crime rates, while Friday and weekends see higher crime rates.
- First day of the month sees a peak in the crime rates.
- July-October, being the season of summer and fall in Toronto, sees the highest crime rate.
- Crime rates show a peak at noon and increase through the evening and night, seeing a maximum at midnight hours.
- Most crimes happen outside followed by apartments and commercial establishments.
- Assault is the most prevalent crime followed by auto theft and break and enter.
- Assault is a prevalent crime especially in apartments, Auto thefts are more prevalent outside.
- WaterfrontCommunities-The Island, Church-Yonge Corridor are neighborhoods with the highest crime rate.
- Woodbine-Lumsden,LambtonBabyPoint, Guildwood are the safest neighborhoods.

Modeling Results Summary



PERFORMANCE METRICS	Logistic Regression	Gaussian Naïve Bayes	KNN	Random Forest	Adaboost
Accuracy	28.9%	30.6%	32.9%	49.8%	42.9%
Precision	0.28	0.37	0.32	0.50	0.43
Recall	0.29	0.31	0.33	0.50	0.43
F1 Score	0.29	0.22	0.31	0.49	0.42

With one-hot encoding



PERFORMANCE METRICS	Logistic Regression	Gaussian Naïve Bayes	KNN	Random Forest	Decision Tree
Accuracy	66%	54%	59.3%	68.7%	58.6%
Precision	0.63	0.37	0.56	0.67	0.59
Recall	0.66	0.54	0.59	0.69	0.59
F1 Score	0.64	0.40	0.49	0.66	0.55

We can draw important inferences through CLUSTERING and EDA of the Crime Dataset. In the future socio-economic factors such as income, education, electrification etc can be also used to get deeper insights.

REFERENCES

[1] Public Safety Data Portal. [Online]. Available: https://data.torontopolice.on.ca/.

[2] T. Dayara, F. Thabtah, H. Abdel-Jaber, and S. Zeidan, “Crime analyses using data analytics,” International Journal of Data Warehousing and Mining, vol. 18, no. 1, pp. 1–15, 2022.

[3] Mahmud, S., Nuha, M., &amp; Sattar, A. (2020). Crime rate prediction using machine learning and Data Mining. Advances in Intelligent Systems and Computing, 59–69. https://doi.org/10.1007/978-981-15-7394-1_5

[4] Safat, W., Asghar, S., &amp; Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and Deep Learning Techniques. IEEE Access, 9, 70080–70094. https://doi.org/10.1109/access.2021.3078117

[5] Sathyadevan, S., Devan, M. S., &amp; Gangadharan, S. S. (2014). Crime analysis and prediction using data mining. 2014 First International Conference on Networks &amp; Soft Computing (ICNSC2014). https://doi.org/10.1109/cnsc.2014.6906719

[6] Paper on Different Approaches for Crime Prediction system, International Journal of Engineering Research & Technology (IJERT) 2017.

[7] CRIME RATE PREDICTION, Journal of Engineering Sciences (JES), 2020.