Crime Analysis

Nandika Arora(2K19/MC/084) , Shashank(2K19/EE/229)

Methods in Data Analytics
Delhi Technological University

***Abstract*— It is critical to recognise crime patterns in order to be better prepared to respond to criminal behaviour. In this study, we evaluate crime data from the city of Indore, which we scraped from the Indore Police's publicly accessible website.**
The goal is to estimate which type of crime is most likely to occur at a given time and location in Indore.
The use of AI and machine learning to identify crime using sound or video is now in use, has been demonstrated to function, and is likely to grow.
The use of AI/ML to forecast crimes or a person's chance of committing a crime has potential, but it is still a work in progress. The most difficult task will most likely be "proving" to legislators that it works. It's tough to establish the negative when a system is meant to prevent something from happening. A positive feedback loop would certainly benefit companies who are directly involved in providing governments with AI capabilities to monitor areas or predict crime. Improvements in crime prevention technology will almost certainly lead to an increase in overall spending on this technology.
We also try to make our categorization work more relevant by grouping many classes together into larger groups. Finally, we present and discuss our findings using several classifiers, as well as discuss future research directions.

I. Introduction

This is a template document. The conference website has an electronic copy available for download. Please contact the conference publications committee as mentioned on the conference website if you have any issues about the paper guidelines. The conference website has information on how to submit your final work.
Many important questions in public safety and protection are related to crime, and a better understanding of crime can lead to more targeted and sensitive law enforcement practises to reduce crime, as well as more concerted efforts by citizens and authorities to create healthy neighbourhood environments. With the advent of the Big Data era and the availability of fast, efficient algorithms for data analysis, understanding patterns in crime from data is an active and growing field of research.

The inputs to our algorithms are time (hour, day, month, and year), place (latitude and longitude), and class of crime:

Act 379 - Robbery
Act 13 - Gambling
Act 279 - Accident
Act 323 - Violence
Act 302 - Murder
Act 363 - Kidnapping

The type of crime that is most likely to have occurred is the output. We use KNN (K-Nearest Neighbors), Decision Trees, and Random Forests to test a variety of categorization algorithms.
We also do numerous classification tasks, first attempting to forecast which of six crime classifications is most likely to have occurred, and then attempting to distinguish between violent and non-violent crimes.

II. RATIONALE

Indore, the commercial centre of Madhya Pradesh, topped the country's crime record in 2008, followed by Bhopal and Jaipur. According to the National Crime Record Bureau's (NCRB) publication "Crime in India 2008," the crime rate in Indore was 941.4, the highest in the country.

With the increased urbanisation and expansion of large cities and towns, crime rates are also rising. This unprecedented increase in crime and offences in cities is a source of tremendous concern and fear for all of us.

Robberies, murders, rapes, and other crimes are common. Thefts, burglaries, robberies, murders, killings, rapes, shoplifting, pick pocketing, drug misuse, illegal trafficking, smuggling, and vehicle thefts, among other things, have caused average citizens to experience troubled nights and days.

In the presence of anti-social and bad elements, they feel exceedingly insecure and vulnerable. The perpetrators have been acting in a well-organized manner, with national and international ties and links in some cases.

III. GOAL

Much of the present research is focused on two main areas:
• predicting crime surges and hotspots, and
• understanding criminal behaviour patterns that could aid in solving criminal investigations.

IV. OBJECTIVE

The objective of our work is to:

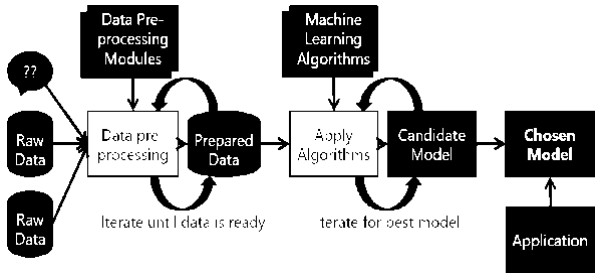
- Predicting crime before it takes place.
- Predicting hotspots of crime.
- Understanding crime pattern.
- Classify crime based on location.
- Analysis of crime in Indore.

V. METHODOLOGY

A. Machine Learning

The automated recognition of meaningful patterns in data is referred to as machine learning. It has become a common technique in practically every endeavour that demands information extraction from massive data sets over the last few decades. Search engines learn how to provide us the best results (while placing pro table advertisements), anti-spam software learns to filter our email communications, and credit card transactions are safeguarded by software that learns how to spot frauds. Intelligent personal assistance software on smartphones learn to understand voice commands and digital cameras learn to distinguish faces.Cars are equipped with accident prevention systems that are built using machine learning algorithms.

In addition to bioinformatics, medicine, and astronomy, machine learning is frequently employed in other scientific fields. A common aspect of all of these applications is that, unlike more traditional uses of computers, due to the complexity of the patterns that must be identified, a human programmer cannot provide an explicit, fine-tuned specification of how such activities should be carried out in these circumstances. Many of our skills are acquired or re ned by learning from our experiences, as shown by intelligent



beings (rather than following explicit instructions given to us). The goal of machine learning technologies is to provide programmes the ability to learn and adapt.

Machine learning process

The inputs to our algorithms are time (hour, day, month, year), place (latitude and longitude), class of crime

- Act 379-Robbery
- Act 13-Gambling
- Act 279-Accident
- Act 323-Violence
- Act 302-Murder
- Act 363-Kidnapping

The type of crime that is most likely to have occurred is the output. We use KNN (K-Nearest Neighbors), Decision Trees, and Random Forests to test a variety of categorization algorithms.

B. Dataset

The data we use is scraped on a daily basis from the Indore Police Department's publicly accessible website.

C. Preprocessing

Before implementing machine learning algorithms on our data, we went through a series of preprocessing steps with our classification task in mind. These included:

- Dropping features such police station, station number, Complainant name & address
  Accused name & address
- Dropping features such as Resolution, Description and Address: The resolution and description of a crime are only known after the crime has occurred, and thus are of little use in a practical, real-world scenario where one is attempting to forecast what type of crime has occurred, hence they were left out. We eliminated the address because we already knew the latitude and longitude, and the address didn't provide much value in that context.
- The timestamp contained the year, date and time of occurrence of each crime. This was decomposed into five features: Year (2018), Month (1-12), Date (1-31), Hour (0- 23) and Minute (0-59).

As part of our early exploratory efforts, we ran some out-of-the-box learning algorithms after these preprocessing processes. Our new feature set included nine features, all of which were now numeric.

timestamp	act379	act13	act279	act323	act363	act302	latitude	longitude
28-02-2018 21:00	1	0	0	0	0	0	22.73726	75.87599
28-02-2018 21:15	1	0	0	0	0	0	22.72099	75.87608
28-02-2018 10:15	0	0	1	0	0	0	22.73668	75.88317
28-02-2018 10:15	0	0	1	0	0	0	22.74653	75.88714

Dataset after Preprocessing

D. After Preprocessing

Following the preprocessing, we were left with three different classification problems to address, which we approached using a variety of classification algorithms. The algorithms that we employed are as follows:

- KNN(K- Nearest neighbors)
- Decision Tree
- Random Forests

VI. IMPLEMETATION

The idea was implemented in Indore first, in order to minimise the area covered by the prediction and make it less complicated. The data was sorted and translated into a new format of timestamp, longitude, and latitude, which would be used by the machine to estimate the crime rate in a specific place or city.

The entries were made solely to teach the machine what it needed to know about the input and what the desired outcome was. After the machine had learned the algorithms and the process, the accuracy of various algorithms was tested, and the algorithm with the highest accuracy, Random forest, was chosen as the prediction kernel.OF PROPER IMPLEMENTATION AND FUNCTIONING SEVERAL ALGORITHMS AND TECHNIQUES WERE USED. FOLLOWING ARE THE ALGORITHMS USED:

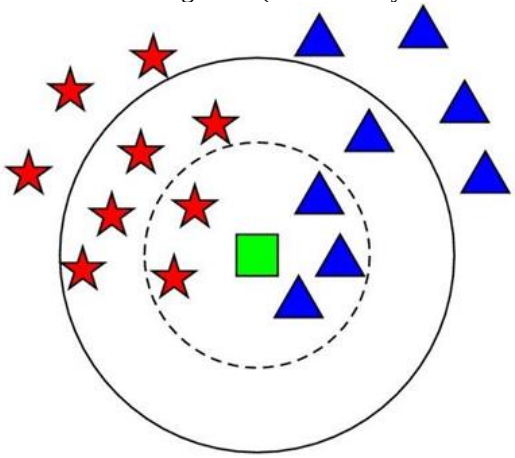
KNN (K-Nearest neighbors):

A pattern recognition algorithm that uses a sophisticated classification mechanism. K nearest neighbours keeps track of all available examples and categorises new ones using a similarity metric (e.g. distance function). One of the most widely used data mining methods today. A lazy learning algorithm that is non-parametric (An Instance based Learning method).

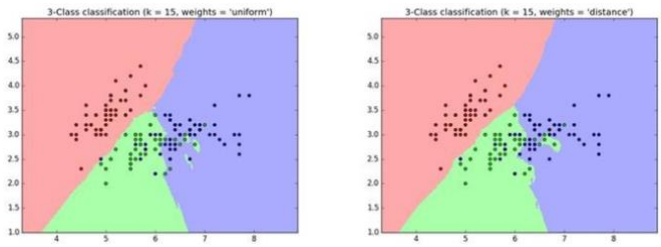
KNN: Classification Approach

An object (a new instance) is classified by a majority votes for its neighbor classes.

The object is assigned to the most common class amongst its K nearest neighbors.(measured by distance function)



Principle diagram of KNN



graphical representation of KNN

Decision Trees: It is, as the name implies, a tree that aids us in decision-making. It is a very fundamental and important predictive learning technique that may be used for both classification and regression.

- It is different from others because it works intuitively i.e., taking decisions one-by-one.
- Non-parametric: Fast and efficient.

It consists of nodes which have parent-child relationships:

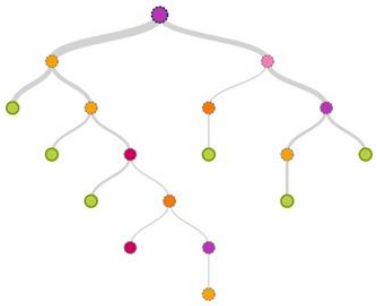
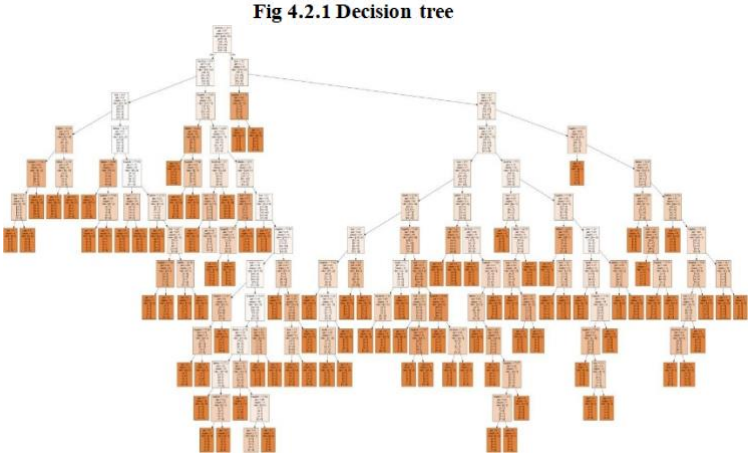


Fig 4.2.1 Decision tree



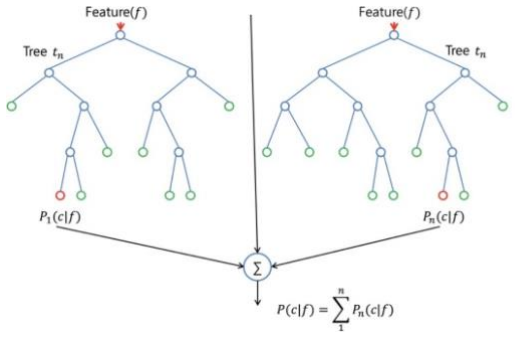
Desision tree of our analysis

Random forest:

Random Forests is a common ensemble learning method that uses training data to create a number of classifiers and then combines their outputs to generate the best predictions on test data.

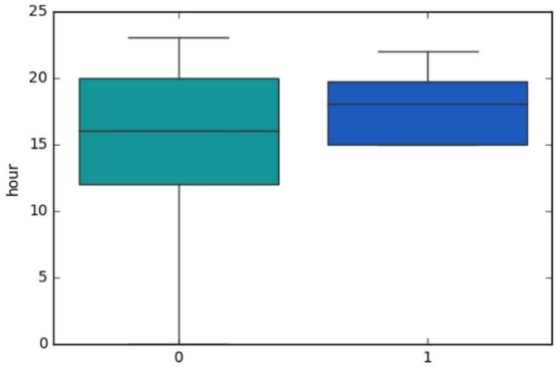
As a result, the Random Forests algorithm is a variance-minimizing algorithm that makes split decisions using randomness to minimise overfitting on the training data.
A random forests classifier is an ensemble classifier that combines h(x|1), h(x|2),...h(x|k) classifiers. h(x|) is a classification tree for each member of the family, and k is the number of trees generated from a model random vector. Also, each θk is a randomly chosen parameter vector. If D(x,y) denotes the training dataset, each classification tree in the ensemble is built using a different subset Dθk(x,y) ⊂ D(x,y) of the training dataset.
Thus, h(x|θk) is the kth classification tree which uses a subset of features xθk ⊂ x to build a classification model. Each tree then works like regular decision trees: it partitions the data based on the value of a particular feature (which is selected randomly from the subset), until the data is fully partitioned, or the maximum allowed depth is reached. The final output y is obtained by aggregating the results thus:

$$y = \text{argmax}_{p \in \{h(x_1)..h(x_k)\}} \{ \sum_{j=1}^{K} (I(h(x|\theta_j) = p)) \}$$

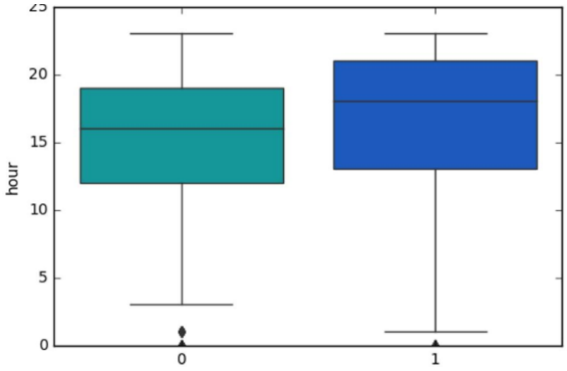


Random Forest Example

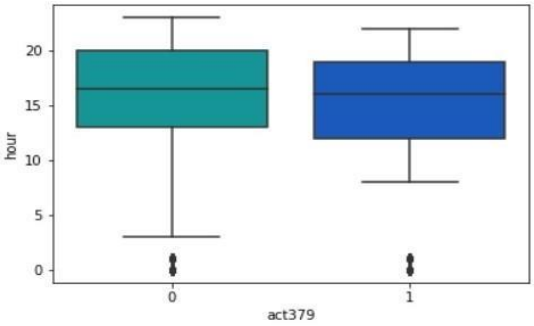
VII. DATA VISUALIZATIONS



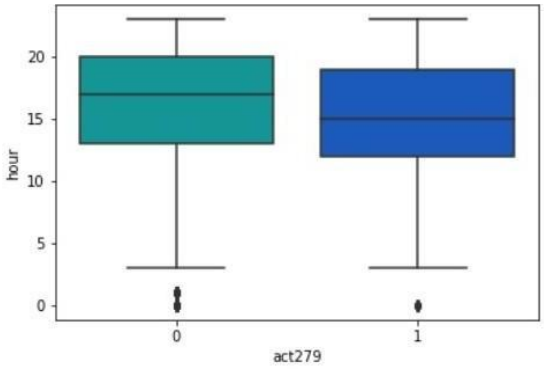
Act13(Gambling vs Hour)



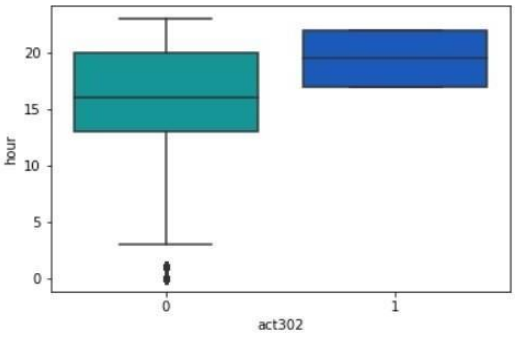
Act323(Violence vs Hour)



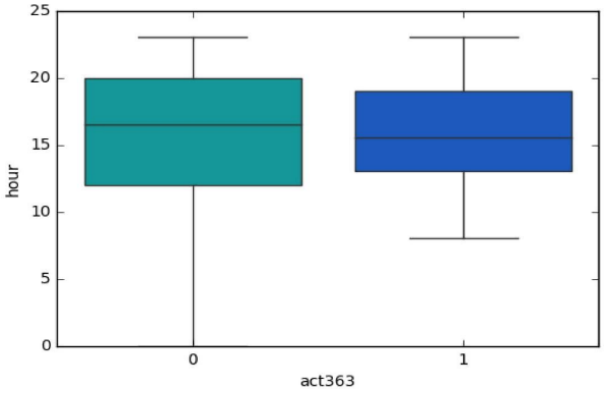
Act379(Robbery vs Hour)



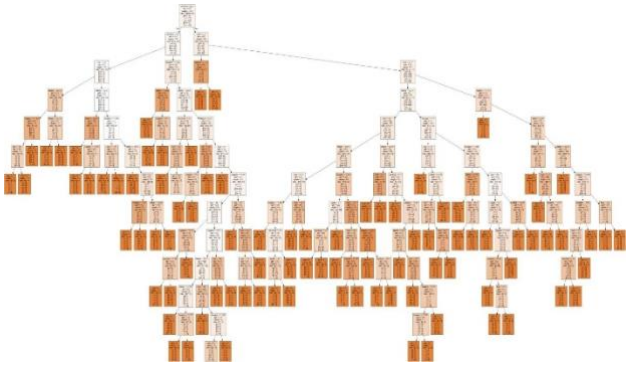
Act279(Accident vs Hour)



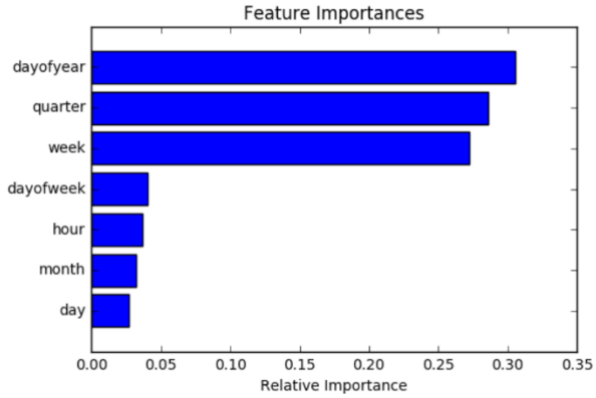
Act302(Murder vs Hour)



Act363(Kidnapping vs Hour)

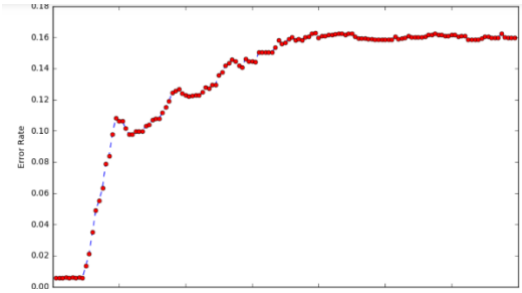


Tree Visualization(Decision Tree Analysis)

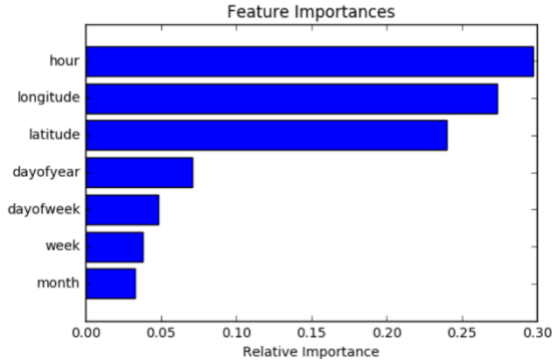


A plot on Feature importances after applying Random Tree Model

VIII. OTHER OUTPUTS



Error rate vs k value plot after using elbow method for getting optimum value for k



A plot on Feature importances after applying Decision Tree Model

IX. TECH STACK

Python (3.6.5)

Packages Used:

- Flask (0.12.2)
- Pandas (0.22.1)
- Numpy (1.14.2)
- Sklearn (0.19.1)
- Geopy (1.13.0)

Jupyter Notebook

X. CONCLUSION

The first task of classifying six separate crime categories was a difficult multi-class classification problem, and our initial data set lacked enough predictability to achieve high accuracy. In order to find organisation in the data, we discovered that collapsing the criminal categories into fewer, larger groups was a more useful method. On Prediction, we had a high level of accuracy and precision. However, using the same classifiers, the Violent/Non-violent crime classification did

not produce impressive results – this was a far more difficult classification task. As a result, collapsing crime categories is a difficult operation that necessitates careful selection and study.

Time-series modelling of the data to identify temporal relationships in it, which may subsequently be used to forecast surges in different types of crime, is one way to extend this work. It would also be interesting to investigate links between surges in other types of crimes — for example, it's possible that two or more types of crimes spike and SINK at the same time, which would be a fascinating relationship to investigate. Implementing a more accurate multi-class classifier and investigating better ways to show our results are two other areas to work on.

XI. FUTURE SCOPE

The goal of any society shouldn't be to just catch criminals but to prevent crimes from happening in the first place

- Predicting Future Crime Spots: We can anticipate where future crimes will occur by using historical data and monitoring where current crimes occurred. A spike in burglaries in one region, for example, could signal an increase in burglaries in other locations in the near future. On a map, the system shows potential hotspots where police should consider increasing patrols.
- Predicting Who Will Commit a Crime: Face Recognition is being used to predict whether or not a person will commit a crime before it occurs. If there are any suspicious changes in their behaviour or unexpected movements, the system will alert them. For example, if a person appears to be walking back and forth in a specific area repeatedly, this could indicate that they are a pickpocket or are scoping the area for a future crime. Individuals will be tracked throughout time as well.
- Pretrial Release and Parole: Most people are released after being charged with a crime until their case goes to trial. Judges used to use their best judgement when selecting who should be released pretrial or what an individual's bail should be set at. Judges had to decide in a matter of minutes whether someone was a flight risk, a significant threat to society, or a witness who would be harmed if freed. It's a flawed system prone to bias. According to the media organization's investigation, the system may have a severe racial bias. "Black defendants who did not recidivate over a two-year period were nearly twice as likely as their white counterparts to be misclassified as greater risk (45 percent vs. 23 percent)," they discovered. The paper raises the question of whether stronger AI/ML will lead to more accurate predictions in the future or if it will exacerbate present issues. Any system will be based on real-world data, but if that data is generated by biassed cops, the AI/ML will be biassed as well.

XII. REFERENCES

[1] Geeks For Geeks
[2] Stack Overflow
[3] Towards Data Science
[4] Machine Learning Mastery
[5] Kaggle
[6] Analytics Vindhya