Crime Analysis and Prediction of Kanpur City

Anjali Singh Chauhan[1], Abhishek Pandey[2], Pragati Khare[3], Chesta Jain[4], Kshitiz Pathak[5]

[1-4]Computer Science Department, Inderprastha Engineering College, Ghaziabad, India

Abstract - *In this paper, we will explore the crime analysis and prediction of crime in the Kanpur city problem. We have used the datasets given from iit kanpur during a seminar , which provides us with the information of the types of crime occurring in particular areas within kanpur city. The difficulty of this task arises from the irregularities in the data sets as the crime in an area varies by huge extent and it was very difficult for us to use any of the machine learning models available. This paper presents the task to predict which category of crime is most likely to occur given a time and place in Kanpur. The use of AI and machine learning to detect crime via sound or cameras currently exists, is proven to work, and expected to continue to expand. To be better prepared to respond to criminal activity, it is important to understand patterns in crime.In our project, we analyse crime data from the city of Kanpur, scraped from publicly available website of Kanpur Police. We also attempt to make our classification task more meaningful by merging multiple classes into larger classes. Finally, we report and reflect on our results with different classifiers, and dwell on avenues for future work.this supervised learning algorithm try to model relationships and dependencies between the target prediction output and the input features such that we can predict the output values for new data based on those relationships which it learned from the previous data sets.*

Key Words: *Classification; Clustering; Learning; MLP; SOM; Supervised learning; Unsupervised learning;*

1. INTRODUCTION

Crime-Scanner means predicting crimes, many important questions in public safety and protection relate to crime, and a better understanding of crime is beneficial in multiple ways: it can lead to targeted and sensitive practices by law enforcement authorities to mitigate crime, and 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:

Rash Driving

Murder

Rape

Forgery

Kidnapping

Suicide

The output is the class of crime that is likely to have occurred. We try out multiple classification algorithms, such as KNN (K-Nearest Neighbours), Decision Trees, and Random Forests. We also perform multiple classification tasks – we first try to predict which crimes are likely to have occurred.

1.1 MACHINE LEARNING

The term machine learning refers to the automated detection of meaningful patterns in data. In the past couple of decades it has become a common tool in almost any task that requires information extraction from large data sets. We are surrounded by a machine learning based technology: search engines learn how to bring us the best results, anti-spam software learns to filter our email messages, and credit card transactions are secured by a software that learns how to detect frauds. Digital cameras learn to detect faces and intelligent personal assistance applications on smart-phones learn to recognize voice commands. Cars are equipped with accident prevention systems that are built using machine learning algorithms.

Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide an explicit, fine detailed specification of how such tasks should be executed. Taking example from intelligent beings, many of our skills are acquired or refined through learning from our experience (rather than following explicit instructions given to us).

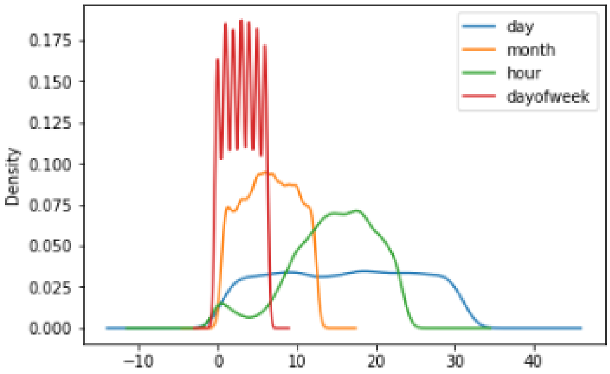

Fig 1. dataset

1.2 KNN (K-Nearest neighbours)

A powerful classification algorithm used in pattern recognition K nearest neighbours stores all available cases and classifies new cases based on similar measures(e.g. distance function). It is one of the top data algorithms used today. A non-parametric lazy learning algorithm(An instance based learning method).

KNN: Classification Approach

- An object(a new instance) is classified by a majority of votes for its neighbour classes.
- The object is assigned to the most common class amongst its K nearest neighbour(measured by distance function).

Nearest-neighbour classifiers are built on training by metaphor, i.e. a testing tuple is compared with a given training tuple that is identical to it. The training tuples are described by n attributes. Every tuple is represented by a point in an n-dimensional space. Thus, every training tuple is stored in an n-dimensional pattern space. When an unknown tuple is given, the K-NN classifier searches for the pattern space for K training tuples that are near to the unknown tuple. These k training tuples are the k “nearest neighbours” of the unknown tuple. When a large training set are given K- Nearest Neighbour is labour intensive. It is widely used in the field of pattern recognition. In addition to this it also has many applications in intrusion detection and data mining. Let n be the training data samples and p be the unknown point. we store n training samples in an array a []. Each element of the array represents a training tuple (x, y), – Now calculate Euclidean distance for every training data sample. Euclidean distance id d (a [], p).

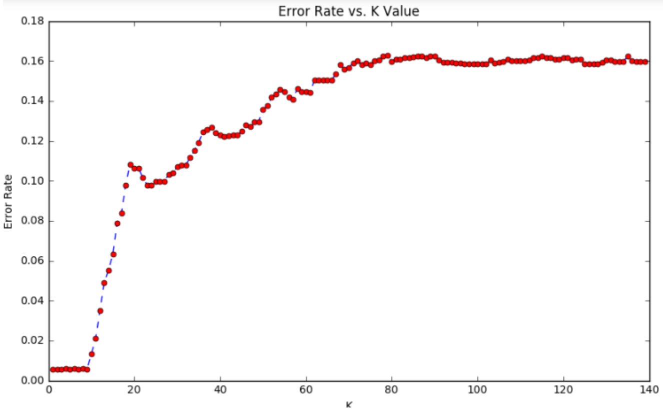

Fig 2 :knn model

1.3 Decision Tree

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.A decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods. Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike linear models, they map nonlinear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression). Decision Tree algorithms are referred to as **CART (Classification and Regression Trees).**

As the name says all about it, it is a tree which helps us by assisting us in decision-making. Used for both classification and regression, it is a very basic and important predictive learning algorithm.

 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: Decision tree considers the most important variable using some fancy criterion and splits the dataset based on it. It is done to reach a stage where we have **homogenous subsets** that are giving predictions with utmost surety.

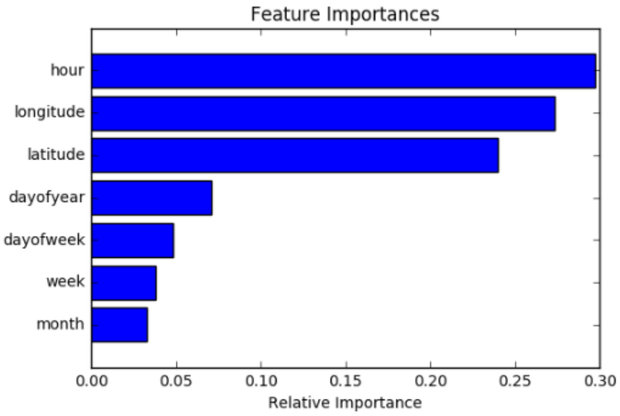

Fig 3. Decision tree model

1.4 Random forest

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks). In this post we'll learn how the random forest algorithm works, how it differs from other algorithms and how to use it. Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

Random Forests is a very popular ensemble learning method which builds a number of classifiers on the training data and combines all their outputs to make the best predictions on the test data. Thus, the Random Forests algorithm is a variance minimizing algorithm that uses randomness when making split decisions to help avoid overfitting on the training data.

A random forests classifier is an ensemble classifier, which aggregates a family of classifiers h(x|θ1),h(x|θ2),..h(x|θk). Each member of the family, h(x|θ), is a classification tree and k is the number of trees chosen 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.

.

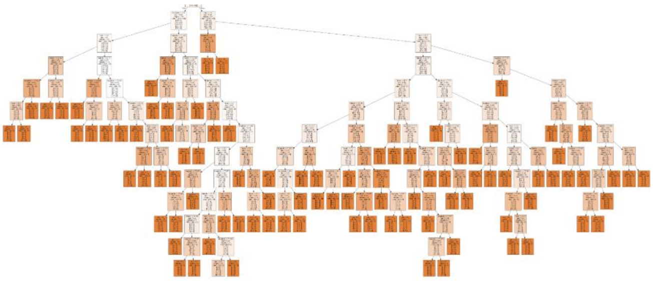

Fig 4. Random forest

1.5 CONCLUSIONS

Finally, we train each of our machine learning models and check the cross-validation results. Here, we have used only our train_data. We have made plots between different factors to find relationships between them.

We have seen which model gives more accurate results and found the most suitable model that we will be using in our project for further work. As a result, we have analysed the most frequent crimes in particular areas, time of the day in which crimes are more likely to occur.

We have calculated accuracy of all three models, that is the decision tree model, k-nn model and random forest model.

Thus, by analysing all three models we have found that the random forest model has highest accuracy. Random forest model has accuracy around 99%.So, the conclusion is drawn that random forest is the best suited model to use in our project.

REFERENCES

1. Alkesh Bharati1, Dr. Sarvana guru RA.K2 (2018, Sep). “Crime Prediction and Analysis Using Machine Learning”. In International Research Journal of Engineering and Technology (IRJET) , (Vol. 05).

2. Sushant Bharti, Ashutosh Mishra. “Prediction of Future Possible offenders’s Network and Role of Offenders”. In 2015 Fifth International Conference on Advances in Computing and Communications

3. Xiangyu Zhao, Jiliang Tang. “Exploring Transfer Learning for Crime Prediction”. 2017 IEEE International Conference on Data Mining Workshops.

4. Jerome H. Friedman, Ron Kohavi, Yeogirl Yun. “Lazy Decision Trees”. To appear in AAAI-96.

5. Tony C.Smith. “Introducing Machine Learning concepts using WEKA”. In Methods of Molecular Biology.