Crime Rate Prediction using KNN

R Umamaheswaran[1], R. Radha[2], ALLA. Preethi[3]

[1,2,3]*Computer Science Engineering, SRM Institute of Science and Technology, Kattankulathur ,Tamil Nadu ,India.*

Abstract— *Crime is one of the biggest and dominating problem in our society and its prevention is an important. task. Daily there are huge numbers of crimes committed frequently. This require keeping track of all the crimes and maintaining a database for same which may be used for future reference. The current problem faced are maintaining of proper dataset of crime and analyzing this data to help in predicting and solving crimes in future. The objective of this project is to analyze dataset which consist of numerous crimes and predicting the type of crime which may happen in future depending upon various conditions. In this project, we will be using the technique of machine learning and data science. Before training of the model data preprocessing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The K-Nearest Neighbor (KNN) classification and various other algorithms will be tested for crime prediction and one with better accuracy will be used for training. Visualization of dataset will be done in terms of graphical representation of many cases for example at which time the criminal rates are high or at which month the criminal activities are high. The soul purpose of this project is to give a jest idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate.*

Keywords—preprocessing, KNN, Visualization ,detect, predict (*key words****)***

1. INTRODUCTION

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way.

The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster.

MACHINE LEARNING – OVERVIEW

Machine learning is a very hot topic for many key reasons, and because it provides the ability to automatically obtain deep insights, recognize unknown patterns, and create high performing predictive models from data, all without requiring explicit programming instructions.

This high level understanding is critical if ever involved in a decision-making process surrounding the usage of machine learning, how it can help achieve business and project goals, which machine learning techniques to use, potential pitfalls, and how to interpret the results.

WHAT IS MACHINE LEARNING?

Machine learning is a subfield of computer science, but is often also referred to as predictive analytics, or predictive modeling. Its goal and usage is to build new and/or leverage existing algorithms to learn from data, in order to build generalizable models that give accurate predictions, or to find patterns, particularly with new and unseen similar data.

magine a dataset as a table, where the rows are each observation (aka measurement, data point, etc), and the columns for each observation represent the features of that observation and their values.

At the outset of a machine learning project, a dataset is usually split into two or three subsets. The minimum subsets are the training and test datasets, and often an optional third validation dataset is created as well.

Once these data subsets are created from the primary dataset, a predictive model or classifier is trained using the training data, and then the model's predictive accuracy is determined using the test data.

As mentioned, machine learning leverages algorithms to automatically model and find patterns in data, usually with the goal of predicting some target output or response. These algorithms are heavily based on statistics and mathematical optimization.

Optimization is the process of finding the smallest or largest value (minima or maxima) of a function, often referred to as a loss, or cost function in the minimization case. One of the most popular optimization algorithms used in machine learning is called gradient descent, and another is known as the the normal equation.

In a nutshell, machine learning is all about automatically learning a highly accurate predictive or classifier model, or finding unknown patterns in data, by leveraging learning algorithms and optimization techniques.

2. LITERATURE SURVEY

[1]Sharmista Dutta et al proposed a paper , which mainly deals with identity crime related to credit card application, which nowadays is quite prevalent and costly even. A new data mining layer of defence has been proposed in this paper. This novel layer makes use of two algorithms-Communal Detection and Spike Detection for detecting frauds in applications. The existing non data-mining techniques for eliminating identity theft have some flaws and uses two algorithms-Communal Detection and Spike Detection for detecting frauds in applications. The proposed two data mining algorithm gives less accuracy and time consuming . The data set for this project is also not given because this projects deals with one's personal identity .

[2]Author Ying-Lung Lin and Liang-chih Yu proposes a data-driven method based on "broken windows" theory and spatial analysis to analyze crime data using machine mining algorithms and thus predict emerging crime hotspots for additional police attention in this paper. The Deep Learning algorithm provides better prediction results than other methods including Random Forest, and Naïve Bayes for potential crime hotspots. But still accuracy of this deep learning algorithm is less and this algorithm doesn't have high amount of processing power. The source of collected data sets has not disclosed due to sensitive contents.

[3]Nidhi Tomar and Amit Kumar Manjhvar has proposed Data mining that automates the finding predictive records procedure in big databases in this paper. Clustering is a most famous method in data mining and is an important methodology that is performed based on the similarity principle Clustering is a major field of data analysis and data mining application. It is a set of methodologies for producing high superiority clusters and high intra-cluster similarity and also low inter -class similarity. The types of data used for analysis of clustering are interval scatted variables binary, nominal, ordinal, ratio variables of mixed types. The segregation of a big database is a stimulating and task of time consuming. The source of the data sets has not provided in the paper due to security reasons.

[4]Olivera Kotevska et al conducted a study that presents a dynamic network model for improving service resilience to data loss. The network model identifies statistically significant shared temporal trends across multivariate spatiotemporal data streams and utilizes these trends to improve data prediction performance in the case of data loss. The ability to handle multivariate time-series/spatiotemporal data streams is essential for estimation accuracy. Future work on the system will investigate the use of time-varying coefficients in VAR to enhance dynamic performance. The impact of mixed frequency data on system performance can be improved. 116375 records were collected for crime events reported throughout Montgomery County, Maryland (MD) for the 1/1/2014 to 5/26/2016 period.

[5]Alexander Stec, and Diego Klabjan proposed a work in which the objective is to take advantage of deep neural networks in order to make next day crime count predictions in a fine-grain city partition. We make predictions using Chicago and Portland crime data, which is augmented with additional datasets covering weather, census data, and public transportation. The crime counts are broken into 10 bins and our model predicts the most likely bin for a each spatial region at a daily level. we are able to predict the correct bin for overall crime count with 75.6% and 65.3% accuracy for Chicago and Portland, respectively. Using deep neural networks the final accuracy % is near 65% - 75%, and KNN- algorithm gives much better accuracy as compared to this algorithm. For Chicago, the data sets are publicly available through the city of Chicago's data portal website, https://data.cityofchicago.org/. For Portland, the crime data was obtained from the National Institution of Justice Real-Time Crime Forecasting challenge.

[6]Ying-Lung Lin and Liang-Chih Yu proposed the geographic characteristics of the grid are discussed,

leaving prediction models unable to predict crime displacement. This study incorporates the concept of a criminal environment in grid-based crime prediction modeling, and establishes a range of spatial-temporal features based on 84 types of geographic information by applying the Google Places API to theft data for Taoyuan City, Taiwan. The best model was found to be Deep Neural Networks, which outperforms the popular Random Decision Forest, Support Vector Machine. After tuning, compared to our design's baseline 11-month moving average, the F1 score improves about 7% on 100-by-100 grids. Algorithms like random forest, SVM are time consuming process and gives less accuracy. Data sets collected from http://www.tyhp.gov.tw/newtyhp/upload/cht/article/file_extract/C0065.pdf.

[7]Julio Borges proposed a paper that analyzes characteristics of the urban environment in San-Francisco in the US and Natal in Brazil cities, deploying a machine learning model to detect categories and hotspots of criminal activities. Extensive evaluation on several years of crime records from both cities show how some features – such as the street network – carry important information about criminal activities. This paper proposed an extensive set of spatio-temporal & urban features which can significantly improve the accuracy of machine learning models for these tasks. Still the proposed work is a time consuming and yields less accuracy. The underlying dataset evaluated in this scenario contains incidents derived from San Francisco Police Department (SFPD) Crime Incident Reporting System. The data ranges from 1/1/2003 to 5/13/2015 with 1.762.311 crime records with 39 distinct crime categories, link: https://www.kaggle.com/c/sf-crime.

[8]Iqbal et al. conducted a study which compares two different classification algorithms, Naïve Bayesian and Decision Tree for predicting the crime category for different states in USA. The result shows that Decision Tree algorithm performed much better than Naïve Bayesian algorithm and achieved 83.9519% accuracy in prediction of the crime category for different states in USA.

[9]Chandra and Gupta proposed a paper about distance-based semi-supervised clustering algorithm functional link neural network (FLNN). The main motive of the work is to reduce disadvantages of the semi-supervised clustering techniques which have a base of the pair wise constraints. Mostly it fails to address the problem, which deals with attributes that have different weights. Based on the real-life applications, all attributes do not have similar importance and hence the equal weights can't be assigned for every attribute.

[10]Oatley and Ewart developed a project to help the West Midlands Police in the UK with high volume crime, burglary from dwelling houses [4]. They created software for utilizing mapping and visualization tools. To determine the causality in this domain they employed statistical methods which have data mining technology of neural network. The predictions of burglary have been predicted by calculated and by combining the total evidence into a Bayesian belief network that is embedded in the developed software system.

3. INFERENCE FROM THE SURVEY

The major problem faced in most of the projects is with the collected data sets in which redundant and improper values may present. These redundant and inaccurate values are solved in successive projects using different types of supervised and un-supervised machine learning techniques. From the above references on crime rate prediction, many algorithms and techniques like clustering, decision tree, naïve-bayes classifier are been used. Further, KNN algorithm has been implemented in this project to get better accuracy. KNN's decision boundary can take any form. KNN is good with correlated attributes, if the distinguishing characteristic of classification is not the marginal distributions but correlation. KNN Classifier can be updated online at low cost, where new instances with known classes are presented. Accuracy is better as compared to other algorithms and time consumption is also less.

4. PROPOSED WORK

In this proposed system the data sets are collected from https://archive.ics.uci.edu/ml/index.php (user generated content) in Internet . Here using classification or regression algorithm based on the data sets our requirement. In this process we are going to use K-Nearest Neighbours (KNN) and also some classification algorithm. Based on the accuracy of the algorithm, prediction rate can be found. Based on the accuracy the algorithm is been chosen.

There are totally 884263 crimes are recorded in the training data and test data for which the categories will be predicted are of the same number. The recorded fields are as follows:

1. Category – type of crime incident (train.csv). This is the target variable which is going to be predicted.
2. Dates – time stamp of the crime incident
3. DayOfWeek – day of week
4. Descript – detailed information of the crime incident
5. PdDistrict – name of police department district
6. Address – address of the crime incident spot
7. X - longitude
8. Y – latitude

5. SYSTEM REQUIREMNTS

The software requirements specification is produced at the culmination of the analysis task. The function and performance allocated to software as part of system engineering are refined by establishing a complete information description as functional representation of system behavior, an indication of performance requirements and design constraints, appropriate validation criteria.

HARDWARE REQUIREMENTS

System	: Pentium IV 2.4 GHz
Hard Disk	: 40 GB
Floppy Drive	: 1.44 Mb
Monitor	: 15 VGA Colour
Mouse	: Logitech
Ram	: 512 Mb

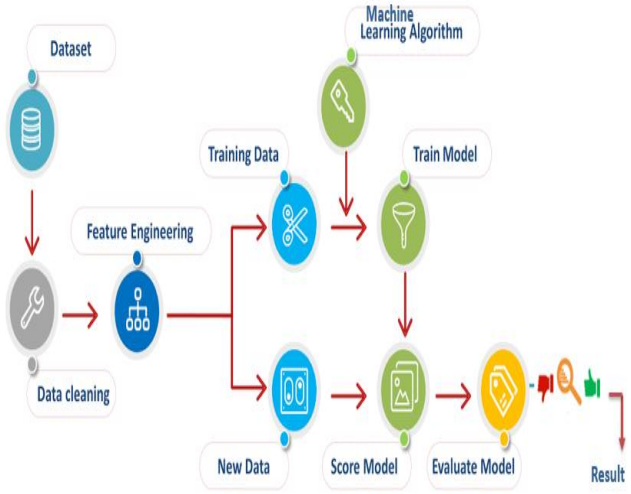
SOFTWARE REQUIREMENTS

Operating system	: Windows 10
IDE	: anaconda navigator
Coding Language	: python

6. SYSTEM ARCHITECTURE

Design is a multi- step that focuses on data structure software architecture, procedural details, procedure etc… and interface among modules. The design procedure also decode the requirements into presentation of software that can be accessed for excellence before coding begins. Computer software design change continuously as novel methods; improved analysis and border understanding evolved. Software proposal is at relatively primary stage in its revolution.

Therefore, software design methodology lacks the depth, flexibility and quantitative nature that are usually associated with more conventional engineering disciplines. However methods for software designs do exit, criteria for design qualities are existing and design notation can be applied.



7. SYSTEM IMPLEMENTATION

A. MODULES:

- Collecting Dataset
- Pre-processing
  - Data cleaning
  - Data transformation
  - Data selection
- Data input

B. ***MODULE DESCRIPTION:***

Collecting Dataset

Data Collection is one of the most important tasks in building a machine learning model. We collect the specific dataset based on requirements from internet. The dataset contains some unwanted data also. So first we need to pre-process the data and obtain perfect data set for algorithm.

Pre-processing

It is the gathering of task related information based on some targeted variables to analyse and produce some valuable outcome. However, some of the data may be noisy, i.e. may contain inaccurate values, incomplete values or incorrect values. Hence, it is must to process the data before analysing it and coming to the results. Data pre-processing can be done by data cleaning, data transformation, data selection.

Data cleaning includes Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.

Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data.

Data selection includes some methods or functions which allow us to select the useful data for our system.

Data input

Dataset values converted into array values which is going to given to the algorithm to find accuracy. Select the algorithm based on the accuracy and analyse the data by using the algorithm.

8. SYSTEM ENVIRONMENT

A. PYTHON TECHNOLOGY

Python is an interpreted , object-oriented programming language similar to PERL, that has gained popularity because of its clear syntax and readability. Python is said to be relatively easy to learn and portable, meaning its statements can be interpreted in a number of operating systems, including UNIX-based systems, Mac OS, MS-DOS, OS/2, and various versions of Microsoft Windows 98. Python was created by Guido van Rossum, a former resident of the Netherlands, whose favourite comedy group at the time was Monty Python's Flying Circus. The source code is freely available and open for modification and reuse. Python has a significant number of users. A notable feature of Python is its indenting of source statements to make the code easier to read. Python offers dynamic data type, ready-made class, and interfaces to many system calls and libraries. It can be extended, using the C or C++language. Python can be used as the script in Microsoft's Active Server Page (ASP) technology. The scoreboard system for the Melbourne (Australia) Cricket Ground is written in Python. Z Object Publishing Environment, a popular Web application server, is also written in the Python language

B. PYTHON PLATFORM

Apart from Windows, Linux and MacOS, CPython implementation runs on 21 different **platforms**. IronPython is a .NET framework based **Python** implementation and it is cabable of running in both Windows, Linux and in other environments where .NET framework is available.

C. PYTHON LIABRARY

Machine Learning, as the name suggests, is the science of programming a computer by which they are able to learn from different kinds of data. A more general definition given by Arthur Samuel is – "Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed." They are typically used to solve various types of life problems.

In the older days, people used to perform Machine Learning tasks by manually coding all the algorithms and mathematical and statistical formula. This made the process time consuming, tedious and inefficient. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reason is its vast collection of libraries. Python libraries that used in Machine Learning are:

- Numpy
- Scipy
- Scikit-learn
- Theano
- TensorFlow
- Keras
- PyTorch
- Pandas
- Matplotlib

9. ALGORITHM

Dataset values converted into array values which are going to be given to the algorithm to find accuracy .Select the algorithm based on the accuracy and analyse the data by using the algorithm.

What is K-nearest Neighbor rule?

One of the simplest decision procedures that can be used for classification is the k-nearest neighbor (NN) rule. It classifies a sample based on the category of its nearest neighbor. The nearest neighbor based classifiers use some or all the patterns available in the training set to classify a test pattern.

K nearest neighbor algorithm works based on minimum distance from the query instance to the training samples to determine the K-nearest neighbors. The data for KNN algorithm consist of several multivariate attributes name that will be used to classify Load the data. Initialize K to your chosen number of neighbors. For each example in the data calculate the distance between the query example and the current example from the data. Add the distance and the index of the example to an ordered collection. Sort the ordered collection of distances and

indices from smallest to largest (in ascending order) by the distances. Pick the first K entries from the sorted collection. Get the labels of the selected K entries. If regression, return the mean of the K labels. If classification, return the mode of the K labels.

ALGORITHM:

1. Load the data
2. Initialize K to your chosen number of neighbors
3. For each example in the data

   3.1 Calculate the distance between the query example and the current example from the data.

   3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels

10. RESULTS AND DISCUSSION

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

11. Conclusion

KNN algorithm is one the simplest classification algorithm. It can give highly competitive results, even with simple classification. KNN algorithm has been implemented in this project to get better accuracy. KNN's decision boundary can take any form. KNN is good with correlated attributes, if the distinguishing characteristic of classification is not the marginal distributions but correlation. KNN Classifier can be updated online at low cost, where new instances with known classes are presented. Accuracy is better as compared to other algorithms and time consumption is also less. A small value of K means that noise will have higher influence on the result and large value make it computationally expensive. In this project data sets will be pre-processed before compiling it. By using data cleaning, data transformation and data selection the numerical data sets will be ready for compilation process.

12. REFERENCES

[1] Amit Kumar Manjhvar (Asst.prof) and Nidhi Tomar (Research Scholar) "An Improved Optimized Clustering Technique For Crime Detection "2016 IEEE Symposium on Colossal Data Analysis and Networking (CDAN)

[2] Khushboo Sukhija,Shailendra Narayan Singh and Jitendra Kumar"Spatial Visualization Approach for Detecting Criminal Hotspots: An Analysis of Total Cognizable Crimes in the State of Haryana" 2017 2nd IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT), May 19-20, 2017, India

[3] Julio Borges, Daniel Ziehr, Michael Beigl,Nelio Cacho, Allan Martins and Simon Sudrich, Samuel Abt,"Feature Engineering for Crime Hotspot Detection" 2017 IEEE

[4] Julio Borges and Adelson Araujo Jr.,Cacho "Towards a Crime Hotspot Detection Framework for Patrol Planning "2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th Intl. Conference on Data Science and Systems

[5] Shoaib KhalidI3*, liechen Wangl2, Muhammd Shakeel3, Xia Nan I "Spatio-temporal Analysis of the Street Crime Hotspots in Faisalabad City of Pakista n " 2015 23rd International Conference on Geoinformatics

[6] Chung-Hsien Yu1 , Max W. Ward1 , Melissa Morabito2 , and Wei Ding"Crime Forecasting Using Data Mining Techniques "2011 11th IEEE International Conference on Data Mining Workshops

[7] Samina Kausar Xu Huahu, Iftikhar Hussain , Zhu Wenhao , And Misha Zahid "Integration of Data Mining Clustering Approach in the Personalized E-Learning System "Received October 26, 2018, accepted November 13, 2018, date of publication November 20, 2018, date of current version December 19, 2018.

[8] X.Alphonse Inbaraj and A. Seshagiri Rao " Hybrid Clustering Algorithms for Crime Pattern Analysis" Department of Computer Science and Engineering PACE Institute of Technology and Sciences Ongole 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India .

[9] Clifton Phua, Member, IEEE, Kate Smith-Miles, Senior Member, IEEE, Vincent Lee, and Ross Gayler, "Resilient Identity Crime Detection ", IEEE transactions on knowledge and data engineering,year 2012 .

[10] Peng Chen and Justin Kurlan"Time, Place, and Modus Operandi: A Simple Apriori Algorithm Experiment for

Crime Pattern Detection " New Zealand Institute for Security and Crime Science University of Waikato Hamilton

[11] Shayam Varan Nath,"Crime pattern detection using data mining ", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 09, pp. 41-44, 2010.

[12] R. Adderley and P. B. Musgrove." Data mining case study: Modelling the behaviour of offenders who commit serious sexual assaults." In Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01), pages 215.220, NEWYORK