

PREDICTION OF CRIME OCCURRENCE FROM MULTI-MODEL DATA USING MACHINE LEARNING ALGORITHM

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Engineering Degree in Electronics and Communication Engineering

by

KANDULA SRUTHI (39130205)

KAKARLA SAI MANI SRI (39130196)



DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

SCHOOL OF ELECTRICAL AND ELECTRONICS

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119

APRIL - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC
Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119
www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICAITON ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **KANDULA SRUTHI (39130205)** and **KAKARLA SAI MANI SRI (39130196)** who carried out a project entitled "**Prediction of Crime Occurrence from Multi-Model Data Using Machine Learning Algorithm**" under our supervision from November 2022 to April 2023.

Internal Guide

Dr. A. SAHAYA ANSELIN NISHA, M.E., Ph.D.

Head of the Department

Dr. T. RAVI, M.E., Ph.D.

Submitted for Viva voce Examination held on 20/04/2023

Internal Examiner

External Examiner



DECLARATION

We, **KANDULA SRUTHI** and **KAKARLA SAI MANI SRI** hereby declare that the Project report entitled “**Prediction of Crime Occurrence from Multi- Model Using Machine Learning Algorithm**” done by us under the guidance of **Dr. A. SAHAYA ANSELIN NISHA, M.E., Ph.D.** It is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in Electronics and Communication Engineering.

DATE: 20/04/2023

PLACE: Chennai

SIGNATURE OF THE CANDIDATES

1. 
2. 

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to the **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. We are grateful to them.

We convey our thanks to **Dr. N. M. NANDHITHA, M.E., Ph.D., Professor & Dean, School of Electrical and Electronics** and **Dr. T. RAVI, M.E., Ph.D., Professor & Head, Dept. of Electronics and Communication Engineering** for provide us with necessary support and details at the right time during the progressive reviews.

We would like to express my sincere and deep sense of gratitude to our Project Guide, **Dr. A. SAHAYA ANSELIN NISHA, M.E., Ph.D.**, her valuable guidance, suggestions and constant encouragement paved the way for the successful completion of our project work.

We wish to express our thanks to all teaching and non-teaching staff members of the Department of **ELECTRONICS AND COMMUNICATION ENGINEERING** who were helpful in many ways for the completion of the project.

ABSTRACT

A crime is a deliberate act that can cause physical or emotional distress, as well as property damage or loss, and can lead to discipline by a state or other authority according to the inflexibility of the crime. The number and forms of felonious conditioning are adding at an enhancing rate, forcing agencies to develop effective styles to take preventative measures. In the present situations, traditional crime-working ways are unfit to deliver results, being slow-paced and less effective.

To help the police, information can be derived from text sources by using social media text analytics. Social media platforms are among the text-based datasets that can benefit from text analysis. Today's society is dealing with a lot of crime issues, and social media is no exception. Crime has had a negative impact on both economic growth and life quality. By looking for and analyzing previous data, criminal patterns are spotted and make predictions about future crimes. But because there isn't enough evidence, some crimes go unreported and unsolved. As a result, finding criminals is still a difficult task. Social media can be used to monitor criminal activity. Because people who utilize social media occasionally post statements about their surroundings on those platforms.

To achieve this, use efficient methods i.e including machine learning (ML) and computer vision algorithms and techniques. Machine learning is the branch of science where computers decide without human intervention. In recent times, machine learning has been used in multiple domains. An example of such cases is automated or self-driving cars. With Machine learning algorithms, there is a way to predict certain results based on the inputs and provide a solution to solve crime cases. In this project, the results of certain cases where such approaches were used, are described.

TABLE OF CONTENTS

CHAPTER	TITLE	PAGE NO.
	ABSTRACT	v
1	INTRODUCTION	1
	1.1 Scope of the Project	2
	1.2 Objective Of The Project	2
	1.3 Problem Statement	2
2	LITERATURE REVIEW	3
3	SYSTEM ANALYSIS	8
	3.1 Existing System	8
	3.2 Proposed System	8
4	MACHINE LEARNING	10
	4.1 Introduction	10
	4.2 Three Types of Data	12
	4.3 Training and testing the Model on Data	17
5	MODULES	20
	5.1 Machine Learning Modules	20
	5.2 Support Vector Machine(SVM) Algorithm	25
	5.3 Random Forest Algorithm	30
6	RESULTS AND DISCUSSION	37
	6.1 Analyse crime	37
	6.2 Detect Crime	38
7	CONCLUSION & FUTURE SCOPE	39
	7.1 Conclusion	39
	7.2 Future Scope	39
	REFERENCES	40

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
3.1	Block Diagram	9
3.2	Flow Diagram	9
4.1	Classification	13
4.2	Regression	14
4.3	Unsupervised Learning	15
4.4	Clustering	15
4.5	Machine Learning Models	16
4.6	Training and Testing	17
4.7	Cross Validation	18
4.8	Matrix	18
5.1	Support Vector Machine	26
5.2	Scenario-1	27
5.3	Scenario-2	27
5.4	Scenario-3	28
5.5	Scenario-4	28
5.6	Scenario-5	29
5.7	Real Life Analogy	31
5.8	Bagging	32
6.1	Analyse Crime	37
6.2	Detect Crime	38

CHAPTER 1

INTRODUCTION

Day after day, the crime data rate has been increasing as a result of modern technologies and hi-tech methods as it helps criminals to perform illegal activities. Evidently, according to the Crime Record Bureau, crimes like burglary, arson, and so on have been increasing, while crimes like murder, sex abuse, gang rape, etc. have been increasing. Crime data is acquired from various blogs, news, and websites. A huge amount of data is utilized as a record for creating a crime report database. The knowledge which is acquired from the data mining techniques will help in reducing crimes as it helps in finding the culprits faster. A vital component of existence is security. Our most basic needs cannot be satisfied unless all are safe. Therefore, having a sense of security is essential to achieving our objectives, whether they will be shared or personal. Criminal activity is a social issue that has a significant negative impact on our society. Both local authorities and residents are getting more concerned with the ability to spot crime and pinpoint the most recent crimes in a certain area. On the other hand, when residing in a bustling environment, people are constantly interested in enhancing safety and developing trustworthy connections with neighbors.

One of the biggest problems facing societies worldwide, especially those in urban areas, is the incidence of crime. While social crimes have been the subject of increasing research, social media has only been used in a limited number of studies involving crime and criminal behavior. As a result, the study attempts to propose a prediction model (algorithm) by using the machine-learning technique, which is intended to hold a high capability to forecast crimes by aspects of social media datasets using the Data Mining idea. Social media is the primary source of our data. The primary objective is to locate every hidden data source and forecast outcomes.

1.1 SCOPE OF THE PROJECT

Much of the current work is focused in two major directions:

- Predicting surges and hotspots of crime.
- Understanding patterns of criminal behavior could help in solving criminal investigations.

1.2 OBJECTIVE OF THE PROJECT

- The project's primary goal is to forecast crime rates and examine those that will actually occur in the future. The authorities can take responsibility and attempt to lower the crime rate based on this information.
- To forecast the graph between the types of Crimes (Independent Variable) and the Year, Multi Linear Regression is employed (Dependent Variable)
- To assist detectives in solving crimes more quickly, the system will examine how to turn criminal information into a regression problem.
- Crime analysis to identify trends in crime based on information already available. Based on the territorial distribution of the available data and crime recognition, the frequency of occurring crimes can be forecast using a variety of multi-linear regression algorithms.

1.3 PROBLEM STATEMENT

- The project is to make crime predictions using the features present in the dataset. The dataset is extracted from the official sites. With the help of a machine learning algorithm, using python as core, the type of crimes which will occur in a particular area are predicted.

CHAPTER 2

LITERATURE REVIEW

2.1 LITERATURE SURVEY

A literature survey is the most important step in the software development process. Before developing the tool it is necessary to determine the time factor, economy, and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool, the programmers need a lot of external support. This support can be obtained from senior programmers, from a book, or from websites. Before building the system, the above consideration is taken into account for developing the proposed system. The major part of the project development sector considers and fully surveys all the required needs for developing the project. For every project, the Literature survey is the most important sector in the software development process. Before developing the tools and the associated designing, it is necessary to determine and survey the time factor, resource requirements, manpower, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine the software specifications in the respective system, such as what type of operating system the project would require, and what all necessary software is needed to proceed with the next step, such as developing the tools, and the associated operations

Ella Haig and Ginger Saltos in 2017 have presented that the increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to better understand criminal behavior and ultimately prevent future crimes. While many of these approaches make use of clustering and association rule mining techniques, there are fewer approaches focusing on predictive models of crime. In this project, various models are explored for predicting the frequency of several types of crimes by LSOA code (Lower Layer Super Output Areas — an administrative system of areas used by the UK police) and the frequency of anti-social behavior crimes.

Three algorithms are used from different categories of approaches: instance-based learning, regression, and decision trees. The data is from the UK police and contains over 600,000 records before pre-processing. The results, looking at predictive performance as well as processing time, indicate that decision trees (M5P algorithm) can be used to reliably predict crime frequency in general as well as anti-social behavior frequency. In this project, various models are explored for predicting the frequency of several types of crimes by LSOA code (Lower Layer Super Output Areas — an administrative system of areas used by the UK police) and the frequency of anti-social behavior crimes. Three algorithms are used from different categories of approaches: instance-based learning, regression, and decision trees. The data is from the UK police and contains over 600,000 records before pre-processing. The results, looking at predictive performance as well as processing time, indicate that decision trees (M5P algorithm) can be used to reliably predict crime frequency in general as well as anti-social behavior frequency[1].

Shuji Sathyadevan and others (2021) have demonstrated crime analysis and prevention is a systematic approach for identifying and analyzing patterns and trends in crime. This system can predict regions that have a high probability of crime occurrence and can visualize crime-prone areas. With the increasing advent of computerized systems, crime data analysts can help law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, previously unknown, useful information can be extracted from unstructured data. Here there is an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc[2].

Khushabu A. Bokde and others in 2018 signified how crimes will somehow influence organizations and institutions when they occur frequently in a society. Thus, it seems necessary to study reasons, factors and relations between the occurrence of different crimes and find the most appropriate ways to control and avoid more crimes. The main objective of this project is to classify clustered crimes based on occurrence frequency during different years.

Data mining is used extensively in terms of analysis, investigation and discovery of patterns for the occurrence of different crimes. A theoretical model based can be applied on data mining techniques such as clustering and classification to real crime dataset recorded by police in England and Wales from 1990 to 2011. Further weights are assigned to the features in order to improve the quality of the model and remove the low value of them. The Genetic Algorithm (GA) is used for optimizing of Outlier Detection operator parameters using Rapid Miner tool [3].

Benjamin Fredrick David. H and A. Suruliandi (2017) has demonstrated that Data Mining is the procedure which includes evaluating and examining large pre-existing databases in order to generate new information which may be essential to the organization. The extraction of new information is predicted using the existing datasets. Many approaches for analysis and prediction in data mining have been performed. However, only a few efforts have been made in the criminology field. Many few have taken the effort to compare the information all these approaches produce. The police stations and other similar criminal justice agencies hold many large databases of information which can be used to predict or analyse criminal movements and criminal activity involvement in society. The criminals can also be predicted based on the crime data. The main aim of this work is to perform a survey on the supervised learning and unsupervised learning techniques that have been applied towards criminal identification. This project presents a survey on crime analysis and crime prediction using several Data Mining techniques [4].

Tushar Sonawanev and others in 2017 have demonstrated about how crime against women these days has become a problem in every nation around the globe. Many countries are trying to curb this problem. Preventive is taken to reduce the increasing number of cases of crime against women. A huge amount of data set is generated every year on the basis of reporting of crimes. This data can prove very useful in analysing and predicting crime and help us prevent crime to some extent. Crime analysis is an area of vital importance in the police department. Study of crime data can help us analyse crime patterns, interrelated clues & important hidden relations between crimes.

That is why data mining can be a great aid to analyse, visualize and predict crime using the crime data set. Classification and correlation of the data set makes it easy to understand similarities & dissimilarities amongst the data objects. Data objects are grouped using clustering technique. Datasets are classified on the basis of some predefined condition. Here grouping is performed according to various types of crimes against women taking place in different states and cities of India. Crime mapping will help the administration plan strategies for prevention of crime. Further, using data mining techniques, data can be predicted and visualized in various forms in order to provide better understanding of crime patterns[5].

Rajkumar .S and others (2019) have also presented that crime analysis and prevention is a systematic approach for identifying and analysing patterns and trends in crime. Thus, the system can predict regions which have a high probability of crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, previously unknown, useful information from unstructured data is extracted. Here there is an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Instead of focusing on causes of crime occurrence like criminal background of offender, political enmity etc, here the main focus is on crime factors of each day[6].

Chhaya chauhan and Smriti sehgal (2017) has presented that Crime analysis is a methodical approach for identifying and analysing patterns and trends in crime. With the increasing origin of computerized systems, crime data analysts can help law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, previously unknown, useful information from unstructured data can be analysed. Predictive policing means, using analytical and predictive techniques, to identify criminals and it has been found to be pretty much effective in doing the same[7].

Kalyani Kadam and Ayishu Almaw in 2018 signifies that crime is a foremost problem where the top priority has been concerned by individual, community and government. This project investigates a number of data mining algorithms and ensemble learning which are applied to crime data mining. This survey paper describes a summary of the methods and techniques which are implemented in crimedata analysis and prediction. Crime forecasting is a way of trying to mining out and decrease the upcoming crimes by forecasting the future crime that will occur. Crimedata prediction uses historical data and, after examining data, predicts the upcoming crime with respect to location, time, day, season and year. Present crime cases rapidly increasing so it is an inspiring task to foresee upcoming crimes closely with better accuracy. Data mining methods are too important to resolving crime problems with investigating hidden crime patterns. So, the objective of this study could be to analyse and discuss various methods which are applied to crime prediction and analysis. This paper delivers reasonable investigation of data mining techniques and ensemble classification techniques for discovery and prediction of upcoming crimes[8].

CHAPTER 3

SYSTEM ANALYSIS

3.1 EXISTING SYSTEM

After finding and understanding various distinct methods used by the police for surveillance purposes, the importance of each method has been demonstrated. Each surveillance method can perform well on its own and produce satisfactory results, although for only one specific characteristic, that is, Sting Ray can help us only when the suspect is using a phone, which should be switched on.

Thus, it is only useful when the information regarding the stakeout location is correct. Based on this information, the ever-evolving technology has yet again produced a smart way to conduct surveillance. The introduction of deep learning, ML, and computer vision techniques has provided us with a new perspective on ways to conduct surveillance.

3.2 PROPOSED SYSTEM

An approach using SVM filtering methods to detect crime-related posts from the social media data set can be proposed. There are four main phases. In the first phase, social media text posts that relate to the crimes are extracted. In the second phase, the reprocessing techniques are applied to clean the data set.

Then, the TF-IDF values are calculated for each pre-processed post in the third phase. Finally, SVM-Based Filter is applied to remove non-related data. Then a random forest classifier is used for classification to categorize the data. The main steps include formatting, cleaning and sampling. The cleaning process is used for removal or fixing of some missing data. There may be data that is incomplete.

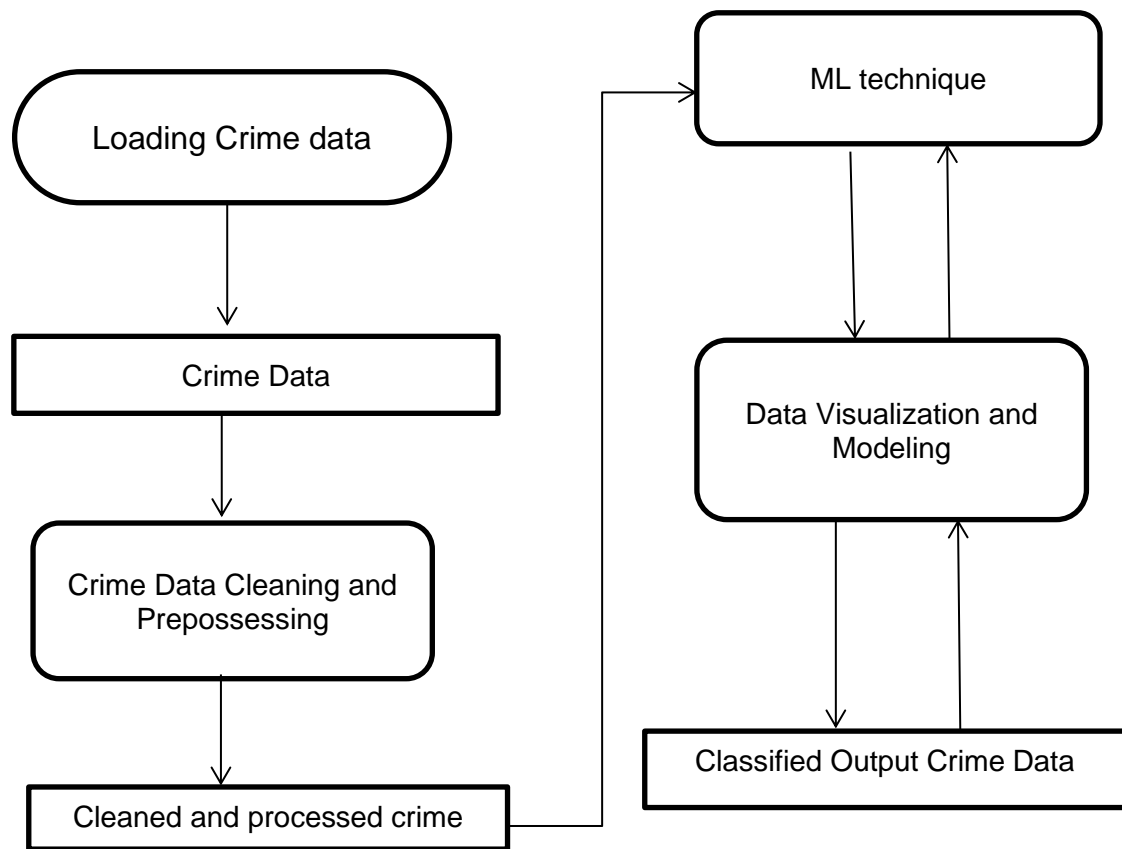


Fig 3.1: Block Diagram

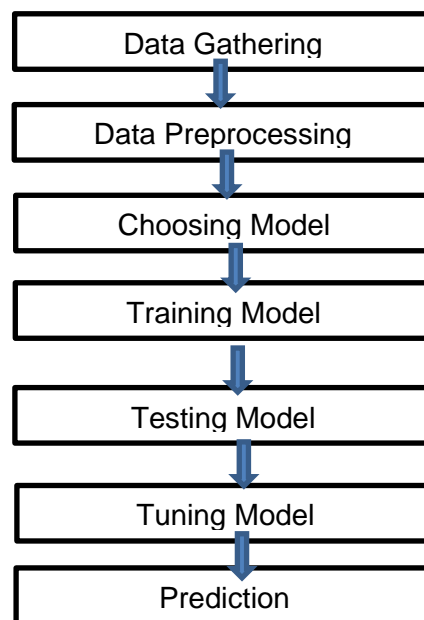


Fig 3.2: Flow Diagram

CHAPTER 4

MACHINE LEARNING

4.1 INTRODUCTION

There are totally 7 steps of Machine Learning. Those are:

- Step 1: Gathering Data. ...
- Step 2: Preparing for Data. ...
- Step 3: Choosing a Model. ...
- Step 4: Training. ...
- Step 5: Evaluation. ...
- Step 6: Hyper parameter Tuning. ...
- Step 7: Prediction.

In this blog, the workflow of a machine learning project includes all the steps required to build the proper machine learning project from scratch. And also data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. A couple of pre-modeling steps that can help to improve the model's performance.

Python Libraries that would be needed to achieve the task:

1. Numpy
2. Pandas
3. Sci-kit
4. Matplotlib

Understanding The Machine Learning Workflow:

The following can define the machine learning workflow into 3 stages.

- i. Gathering data
- ii. Data pre-processing
- iii. Researching the model that will be best for the type of data
- iv. Training and testing the model
- v. Evaluation

The machine learning model is nothing but a piece of code; an engineer or data scientist makes it smart through training with data. So, if a garbage is given to the model, a garbage is returned as an output, i.e. the trained model will provide false or wrong predictions.

Gathering Data:

The process of gathering data depends on the type of project that is desired to make, an ML project that uses real-time data, an IoT system that uses different sensors data can be built. The data set can be collected from various sources such as a file, database, sensor and many other such sources, but the collected data cannot be used directly for performing the analysis process as there might be a lot of missing data, extremely large values, unorganized text data or noisy data. Therefore, to solve this problem, data preparation is done. Free data sets can be utilized which are present on the internet. Kaggle and UCI Machine learning Repository are the repositories that are used the most for making machine learning models. Kaggle is one of the most visited websites that is used for practicing machine learning algorithms. They also host competitions in which people can participate and get to test their knowledge of machine learning.

Data pre-processing:

Data pre-processing is one of the most important steps in machine learning. It is the most important step that helps to build machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time on data pre-processing and 20% time on actually performing the analysis.

Data pre-processing is a process of cleaning the raw data i.e The data is collected in the real world and is converted into a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn't feasible for analysis. Therefore, certain steps are taken to convert the data into a small clean data set. This part of the process is called data pre-processing.

As it is clear that data pre-processing is a process of cleaning the raw data into clean data, so that can be used to train the model. So, it requires data pre-processing to achieve good results from the applied model in machine learning and deep learning projects.

Most of the real-world data is messy. Some of these types of data are:

1. **Missing data:** Missing data can be found when it is not continuously created or due to technical issues in the application (IOT system).
2. **Noisy data:** This type of data is also called outliers. This can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.
3. **Inconsistent data:** This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

4.2 THREE TYPES OF DATA

- Numeric e.g. income, age
- Categorical e.g. gender, nationality
- Ordinal e.g. low/medium/high

These are some of the basic pre-processing techniques that can be used to convert raw data.

Conversion of data: As it is clear that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

Ignoring the missing values: Whenever missing data in the data set is encountered then the row or column of data can be removed. This method is known to be efficient, but it shouldn't be performed if there are a lot of missing values in the dataset. Most commonly, the mean, median or highest frequency value is used.

Machine learning: If there is some missing data then what data shall be present at the empty position can be predicted by using the existing data.

Outliers detection: There are some error data that might be present in our data set that deviates drastically from other observations in a data set. [Example: human weight of = 800 Kg; due to mistyping of extra 0]

Supervised learning:

In supervised learning, an AI system is presented with data which is labelled, which means that each data is tagged with the correct label. The supervised learning is categorized into 2 other categories, which are “**Classification**” and “**Regression**”.

Classification:

The **classification** problem is when the target variable is **categorical** (i.e. the output could be classified into classes — it belongs to either Class A or B or something else). A classification problem is when the output variable is a category, such as “red” or “blue”, “disease” or “no disease” or “spam” or “not spam”.

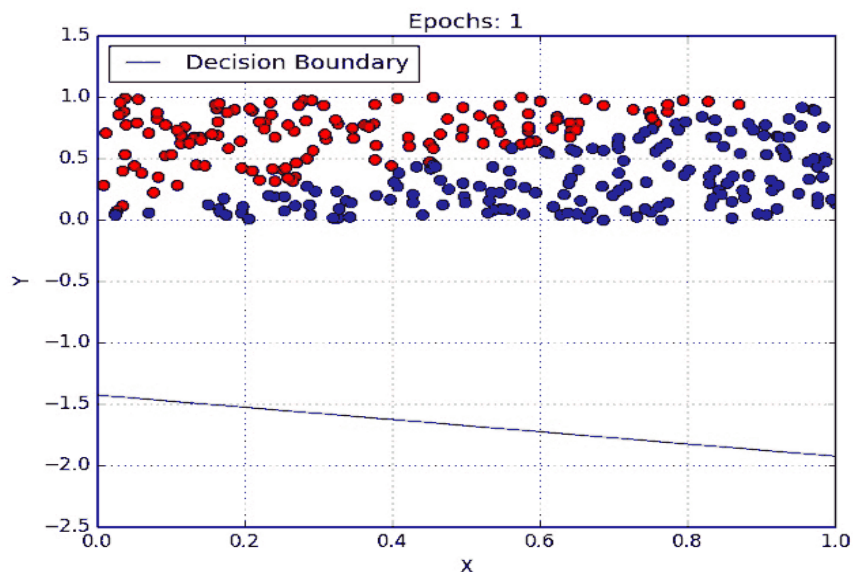


Fig 4.1: Classification

As shown in the above representation, The classes which are plotted on the graph i.e. red and blue which can be represented as 'setosa flower' and 'versicolor flower', the X-axis as the 'Sepal Width' and the Y-axis as the 'Sepal Length' can be imagined, so the best fit line that separates both classes of flowers can be created. These are some of the most used classification algorithms.

- **K-Nearest Neighbour**
- **Naive Bayes**
- **Decision Trees/Random Forest**
- **Support Vector Machine**
- **Logistic Regression**

Regression:

While a regression problem is when the target variable is **continuous** (i.e. The output is numeric).

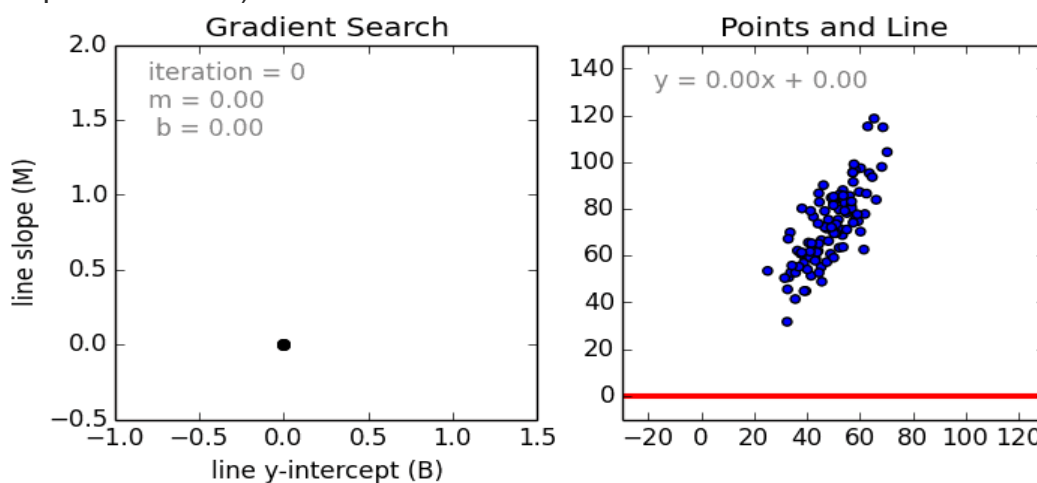


Fig 4.2: Regression

As shown in the above representation, the graph's X-axis is the 'test scores' and the Y-axis represents 'the IQ' can be imagined. So create the best fit line in the given graph so that that line can be used to predict any approximate IQ that isn't present in the given data.

These are some of the most used regression algorithms.

- **Linear Regression**
- **Support Vector Regression**
- **Decision Tress/Random Forest**
- **Gaussian Progresses Regression**
- **Ensemble Methods**

Unsupervised learning:

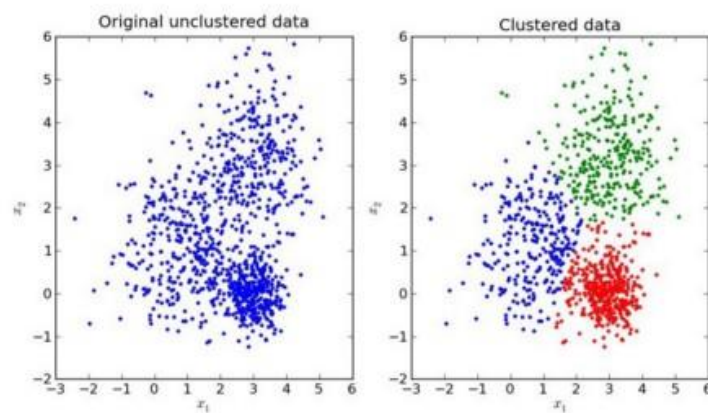


Fig 4.3: Unsupervised Learning

The unsupervised learning is categorized into 2 other categories which are “**Clustering**” and “**Association**”.

Clustering:

A set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

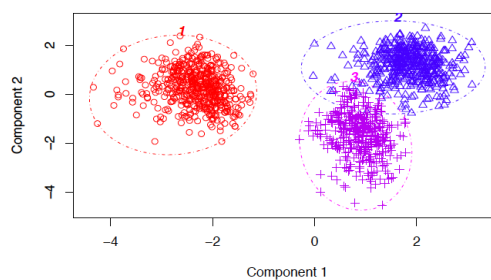


Fig 4.4: Clustering

Methods used for clustering are:

- **Gaussian mixtures**
- **K-Means Clustering**
- **Boosting**
- **Hierarchical Clustering**
- **K-Means Clustering**
- **Spectral Clustering**

Overview of models under categories:

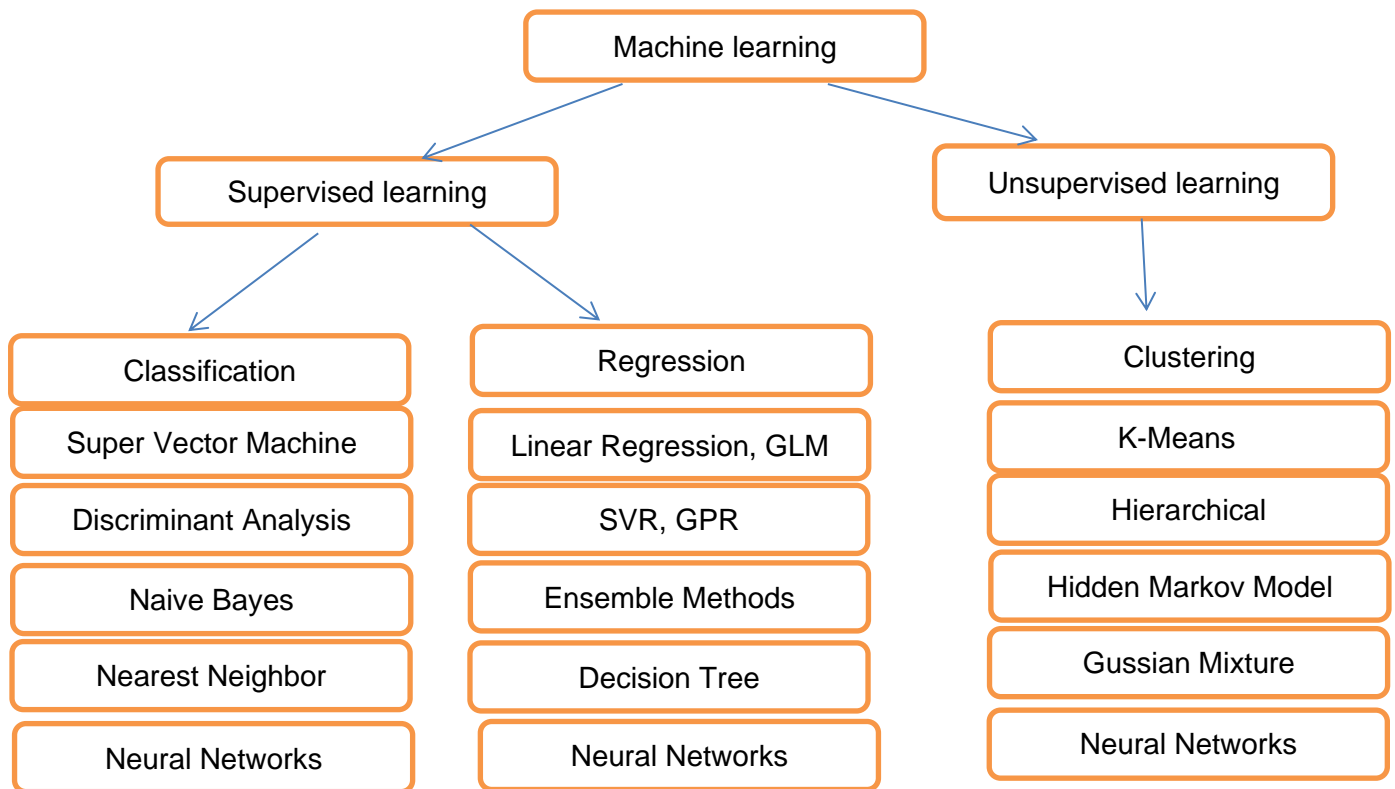


Fig 4.5: Machine Learning Models

4.3 Training and testing the model on data

For training a model, initially split the model into 3 three sections, which are 'Training data', 'Validation data' and 'Testing data' and train the classifier using 'training data set', tune the parameters using 'validation set', and then test the performance of the classifier on unseen 'test data set'. An important point to note is that during training the classifier only the training and/or validation set is available. The test data set must not be used during the training classifier. The test set will only be available during testing the classifier.

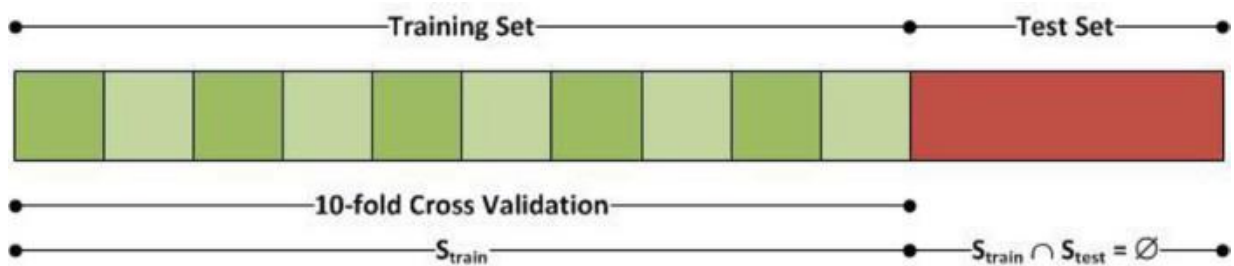


Fig 4.6: Training and Testing

Training set:

The training set is the material through which the computer learns how to process information. Machine learning uses algorithms to perform the training part. A set of data is used for learning, that is to fit the parameters of the classifier.

Validation set:

Cross-validation is primarily used in applied machine learning to estimate the skill of machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.

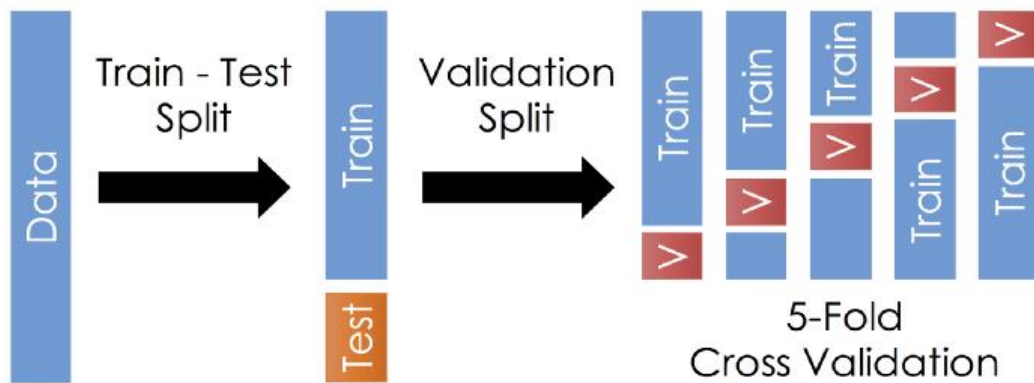


Fig 4.7: Cross Validation

Once the data is divided into the 3 given segments, the training process will started. In a data set, a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually, a data set is divided into a training set, a validation set (some people use ‘test set’ instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration.

The model uses any one of the models that has been chosen in step 3/ point 3. Once the model is trained, the same trained model is used to predict testing data i.e. the unseen data. Once this is done, it can develop a confusion matrix. This tells us how well the model is trained. A confusion matrix has 4 parameters, which are ‘**True positives**’, ‘**True Negatives**’, ‘**False Positives**’ and ‘**False Negative**’. S prefer to get more values in the true negatives and true positives to get a more accurate model. The size of the Confusion matrix completely depends upon the number of classes.

n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Fig 4.8: Matrix

- **True positives:** These are cases in which are predicted TRUE and the predicted output is correct.
- **True negatives:** predicted FALSE and our predicted output is correct.
- **False positives:** predicted TRUE, but the actual predicted output is FALSE.
- **False negatives:** predicted FALSE, but the actual predicted output is TRUE.

So find out the accuracy of the model using the confusion matrix.

$$\text{Accuracy} = (\text{True Positives} + \text{True Negatives}) / (\text{Total number of classes})$$

i.e. for the above example:

$$\text{Accuracy} = (100 + 50) / 165 = 0.9090 \text{ (90.9\% accuracy)}$$

Evaluation:

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. To improve the model it might tune the hyper-parameters of the model and try to improve the accuracy and also look at the confusion matrix to try to increase the number of true positives and true negatives.

CHAPTER 5

MODULES

5.1 MACHINE LEARNING MODULES

Data Collection:

- Social media posts are collected through the Search API.
- The search of the social media posts must be based on a set of keywords that can be used to classify the crime situations.
- Thus, in the first filter, use the main crime-related keywords according to crime categories.

Data Cleaning:

- Data cleaning is a critically important step in any machine learning project.
- In this module, data cleaning is done to prepare the data for analysis by removing or modifying the data that may be incorrect, incomplete, duplicated or improperly formatted. In tabular data, there are many different statistical analysis and data visualization techniques, so it is used to explore the data in order to identify data cleaning operations.

Data Pre-processing:

- As the next step, it is very important to apply the pre-processing techniques to the extracted data set. Because, there may be typos, unwanted content like URLs, and stop words in the social media post.
- Thus, data which is obtained from social media is highly unstructured and noisy.
- Pre-processing techniques will generate clean tweet data that will be used for the next process.
- First, remove the stop words such as is, then, which, have, etc. The words do not convey any positive or negative meaning. So, we can easily remove the stop word without affecting the meaning of the message.
- Then, remove the URLs, hashtags, symbols, usernames, expressions, quotes, etc. Next, combine words are split by applying tokenizing techniques.
- Finally, apply a stemming algorithm to reduce a word to its word stem that affixes to suffixes and prefixes or the roots of words.

Feature Extraction:

- This is done to reduce the number of attributes in the dataset, hence providing advantages like speeding up the training and accuracy improvements.
- In machine learning, pattern recognition, and image processing, feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases, leading to better human interpretations. Feature extraction is related to dimensionality reduction
- When the input data to an algorithm is too large to be processed and it is suspected to be redundant (e.g. the same measurement in both feet and meter, or the repetitiveness of images presented as pixels), then it can be transformed into a reduced set of features (also named a feature vector).
- Determining a subset of the initial features is called feature selection. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

Data Preparation:

- After completing the pre-processing, social media posts are transformed into vectors to generate the feature vectors. The vectors are used in the learning phase for machine learning algorithms.
- This research, used the term frequency-inverse document frequency (TF-IDF) values to create the vectors.
- TF-IDF value reflects the importance of a term in a document to the collection of documents.

Second Stage Filtering:

- Despite a large amount of data collected via keyword filtering, few of them are concerning crimes. Most information is considered as noise.
- According to the above examples, the searched keyword is 'shoot'. The expected gun shooting data relates to crime scenes. However, the above post is related to a film shoot and photo shoot. So, it can consider it as a noisy post.
- To eliminate the noise and increase the accuracy, implemented a machine learning-based filter at this stage. Used Support Vector Machine (SVM) as the machine learning algorithm.
- The SVM is a state-of-the-art classification method that uses a learning algorithm based on structural risk minimization.
- The classifier can be used in many disciplines because of its high accuracy, ability to deal with high dimensions, and flexibility in modeling diverse sources of data.

Model Training:

- A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output. The training model is used to run the input data through the algorithm to correlate the processed output against the sample output. The result from this correlation is used to modify the model.
- This iterative process is called "model fitting". The accuracy of the training dataset or the validation dataset is critical for the precision of the model.
- Model training in machine language is the process of feeding an ML algorithm with data to help identify and learn good values for all attributes involved.
- There are several types of machine learning models, of which the most common ones are supervised and unsupervised learning.
- In this module, use supervised classification algorithms like linear regression to train the model on the cleaned dataset after dimensionality reduction.

Testing Model:

- In this module, testing the trained machine learning model using the test dataset
- Quality assurance is required to make sure that the software system works according to the requirements. Were all the features implemented as agreed? Does the program behave as expected? All the parameters that were tested in the program are against and it should be stated in the technical specification document.
- Moreover, software testing has the power to point out all the defects and flaws during development and do not want clients to encounter bugs after the software is released and come to waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

Performance Evaluation:

- In this module, evaluating the performance of trained machine learning model using performance evaluation criteria such as F1 score, accuracy and classification error.
- In case the model performs poorly, optimize the machine learning algorithms to improve the performance.
- performance Evaluation is defined as a formal and productive procedure to measure an employee's work and results based on their job responsibilities. It is used to gauge the amount of value added by an employee in terms of increased business revenue, in comparison to industry standards and overall employee return on investment (ROI).
- All organizations that have learned the art of "winning from within" by focusing inward towards their employees, rely on a systematic performance evaluation process to measure and evaluate employee performance regularly.
- Ideally, employees are graded annually on their work anniversaries based on which they are either promoted or are given a suitable distribution of salary raises
- . Performance evaluation also plays a direct role in providing periodic feedback to employees, such that they are more self-aware in terms of their performance metrics.

Prediction:

- "Prediction" refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days.
- The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.
- The word "prediction" can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when they use machine learning to determine the next best action in a marketing campaign.
- Other times, though, the "prediction" has to do with, for example, whether or not a transaction that had already occurred was fraudulent.
- In that case, the transaction has already happened, but they are making an educated guess about whether or not it was legitimate, allowing them to take the appropriate action.
- In this module, using a trained and optimized machine learning model to predict whether the crime is accruing on social media.

Nltk:

- NLTK is a toolkit built for working with NLP in Python.
- It provides us with various text-processing libraries with a lot of test datasets.
- A variety of tasks can be performed using NLTK such as tokenizing, parse tree visualization, etc.
- Tokenization
- Lower case conversion
- Stop Words removal
- Stemming
- Lemmatization
- Parse tree or Syntax Tree generation
- POS Tagging

Tensorflow:

- TensorFlow is an open source framework developed by Google researchers to run machine learning, deep learning and other statistical and predictive analytics workloads. The TensorFlow software handles data sets that are arrayed as computational nodes in a graph.
- It is an open source artificial intelligence library, uses data flow graphs to build models. It allows developers to create large-scale neural networks with many layers. TensorFlow is mainly used for: Classification, Perception, Understanding, Discovering, Prediction and Creation

Keras:

- Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library.
- Designed to enable fast experimentation with deep neural networks, it focuses on being user-friendly, modular, and extensible.
- Keras contains numerous implementations of commonly used neural-network building blocks such as layers, objectives, activation functions, optimizers, and a host of tools to make working with image and text data easier to simplify the coding necessary for writing deep neural network code.

5.2 SUPPORT VECTOR MACHINE(SVM) ALGORITHM

Introduction

Mastering machine learning algorithms isn't a myth at all. Most beginners start by learning regression. It is simple to learn and use[9]. Think of machine learning algorithms as an armory packed with axes, sword, blades, bow, dagger, etc. Having various tools, but they are ought to learn and to use them at the right time. As an analogy, think of 'Regression' as a sword capable of slicing and dicing data efficiently, but incapable of dealing with highly complex data.

On the contrary, 'Support Vector Machines' is like a sharp knife – it works on smaller datasets, but on the complex ones, it can be much stronger and powerful in building machine learning models. By now, I hope they have now mastered Random Forest, Naive Bayes Algorithm and Ensemble Modeling. If not, I'd suggest taking out a few minutes and read about them as well. In this project , they guide through the basics advanced knowledge of a crucial machine learning algorithm, support vector machines[10].

Support Vector Machine:

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.

In the SVM algorithm, plot each data item as a point in n-dimensional space (where n is the number of features you have) with the value of each feature being the value of a particular coordinate. Then, perform classification by finding the hyper-plane that differentiates the two classes very well (look at the below snapshot).

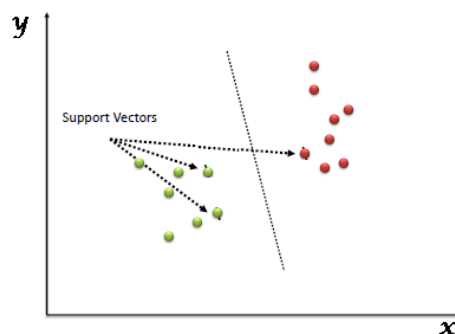


Fig 5.1: Support Vector Machine

Support Vectors are simply the co-ordinates of individual observation. The SVM classifier is a frontier which best segregates the two classes (hyper-plane/ line). So looking at support vector machines and a few examples of its working are here. Above, got accustomed to the process of segregating the two classes with a hyper-plane.

- **Identify the right hyper-plane (Scenario-1):** Here, it shows three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify stars and circles.

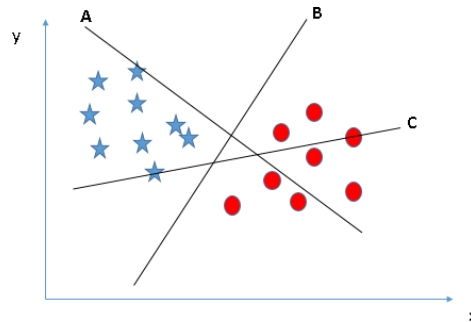


Fig 5.2: Scenario-1

So, need to remember a thumb rule to identify the right hyper-plane: “Select the hyper-plane which segregates the two classes better”. In this scenario, hyper-plane “B” has excellently performed this job.

- **Identify the right hyper-plane (Scenario-2):** Here, the three hyper-planes (A, B and C) and all segregate the classes well. Here, maximizing the distances between the nearest data point (either class) and the hyper-plane will help to decide the right hyper-plane. This distance is called the Margin. Let’s look at the below snapshot:

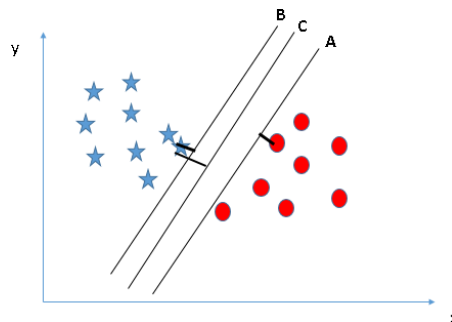


Fig 5.3: Scenario-2

Above, figure shows that the margin for hyper-plane C is high as compared to both A and B. Hence, name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If selected a hyper-plane has a low margin, then there is a high chance of misclassification[11].

- **Identify the right hyper-plane (Scenario-3):** Hint: Use the rules as discussed in the previous section to identify the right hyper-plane.

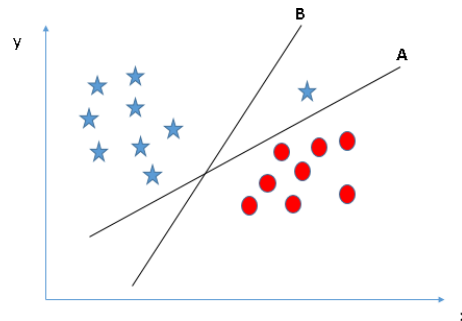


Fig 5.4: Scenario-3

Some of them may have selected the hyper-plane **B** as it has higher margin compared to **A**. But, here is the catch: SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has been classified all correctly. Therefore, the right hyper-plane is **A**.

- **Can we classify two classes (Scenario-4)?** : Below, unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.

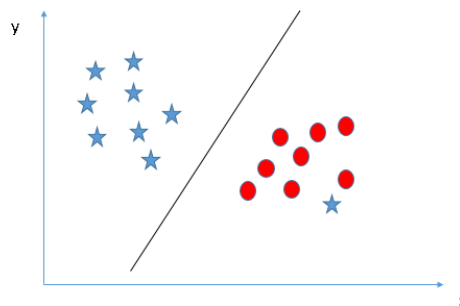


Fig 5.5: Scenario-4

As already mentioned, one star at the other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, SVM classification is robust to outliers.

- **Find the hyper-plane to segregate into classes (Scenario-5):** In the scenario below, linear hyper-plane does not have between the two classes. SVM can solve this problem. Easily! It solves this problem by introducing additional features. Here, it will add a new feature $z=x^2+y^2$. Now, let's plot the data points on axis x and z:

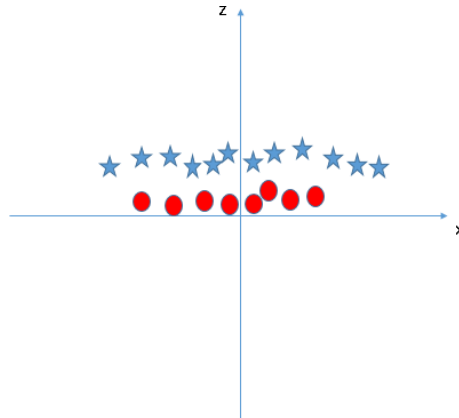


Fig 5.6: Scenario-5

In the above plot, points to consider are:

- All values for z would always be positive because z is the squared sum of both x and y .
- In the original plot, red circles appear close to the origin of the x and y axes, leading to the lower value of z and star relatively away from the origin, resulting in a higher value of z .

In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But, another burning question which arises is, whether we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the kernel trick. The SVM kernel is a function that takes lower dimensional input space and transforms it to a higher dimensional space i.e. It converts a separable problem into a separable problem. It is mostly useful in non-linear separation problems[12]. Simply put, it does some extremely complex data transformations, then finds out the process of separating the data based on the labels or outputs that are defined.

Pros and Cons associated with SVM

Pros:

- It works really well with a clear margin of separation. It is effective in high dimensional spaces.
- It is effective in cases where the number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

Cons:

- It doesn't perform well when we have large a data set because the required training time is higher
- It also doesn't perform very well, when the data set has more noise i.e. target class.
- SVM doesn't directly provide probability estimates, these are calculated using an expensive five- fold cross-validation. It is included in the related SVC method of Python scikit-learn library.

5.3 Random Forest Algorithm

Introduction

Random forest is a ***Supervised Machine Learning Algorithm*** that is ***used widely in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression[13].

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing ***continuous variables*** as in the case of regression and ***categorical variables*** as in the case of classification. It performs better results for classification problems.

Real Life Analogy

Let's dive into a real-life analogy to understand this concept further. A student named X wants to choose a course after his 10+2, and he is confused about the choice of course based on his skill set. So he decides to consult various people, like his cousins, teachers, parents, degree students, and working people. He asks them varied questions like why he should choose, job opportunities with that course, course fee, etc. Finally, after consulting various people about the course, he decides to take the course suggested by most of the people.

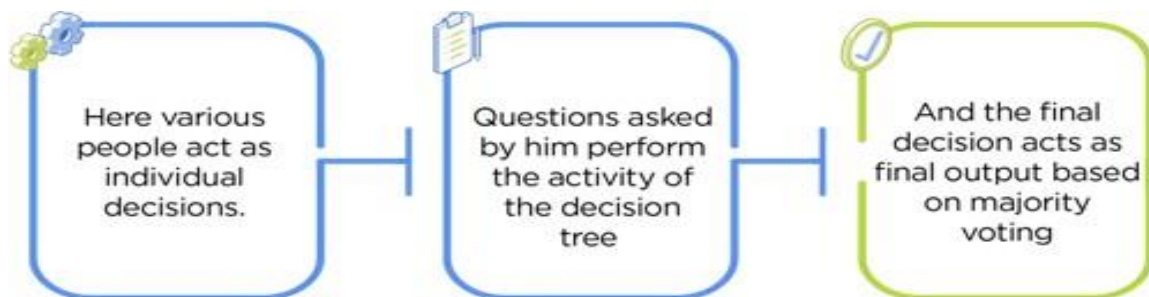


Fig 5.7: Real Life Analogy

Working of Random Forest Algorithm

Before understanding the working of the random forest, just look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model.

Ensemble uses two types of methods:

1. Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest.

2. Boosting– It combines weak learners with strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

Bagging

Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forests. Bagging chooses a random sample from the data set. Hence, each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now, each model is trained independently, which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation[14].

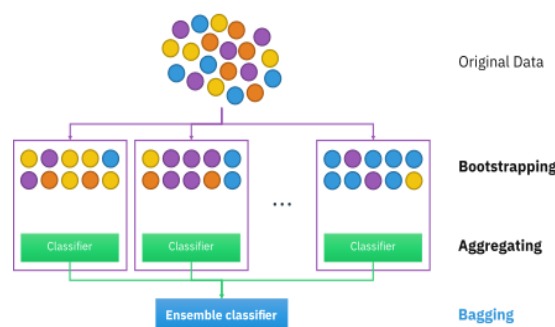


Fig 5.8: Bagging

Now let's look at an example by breaking it down with the help of the following figure. Here the bootstrap sample is taken from actual data (Bootstrap sample 01, Bootstrap sample 02, and Bootstrap sample 03) with a replacement which means there is a high possibility that each sample won't contain unique data.

Now the models (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently[15]. Each model generates results as shown. Now happy emojis have a majority when compared to sad emoji. Thus, based on majority voting, final output is obtained as Happy emoji

Steps involved in random forest algorithm:

Step 1: In Random forest n number of random records are taken from the data set of k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

For example: consider the fruit basket as the data as shown in the figure below. Now n number of samples are taken from the fruit basket and an individual decision tree is constructed for each sample[16]. Each decision tree will generate an output as shown in the figure. The final output is considered based on majority voting. In the below figure can see that the majority decision tree gives output as an apple when compared to a banana, so the final output is taken as an apple.

Important Features of Random Forest

1. Diversity - Not all attributes/ variables/ features are considered while making an individual tree, each tree is different.
2. Immune to the curse of dimensionality - Since each tree does not consider all the features, the feature space is reduced.
3. Parallelization - Each tree is created independently out of different data and attributes. This means that can make full use of the CPU to build random forests.
4. Train -Test split- In a random forest you don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
5. Stability - Stability arises because the result is based on majority voting/ averaging.

Important Hyperparameters:

Hyperparameters are used in random forests to either enhance the performance and predictive power of models or to make the models faster.

Following hyperparameters increases the predictive power:

1. **n_estimators**– number of trees the algorithm builds before averaging the predictions.
2. **max_features**– maximum number of features random forest considers splitting a node.
3. **mini_sample_leaf**– determines the minimum number of leaves required to split an internal node.

Following hyperparameters increases the speed:

1. **n_jobs**– it tells the engine how many processors it is allowed to use. If the value is 1, it can use only one processor, but if the value is -1 there is no limit.
2. **random_state**– controls randomness of the sample. The model will always produce the same results if it has a definite value of random state and if it has been given the same hyperparameters and the same training data.
3. **oob_score**– OOB means out of the bag. It is a random forest cross-validation method. In this, one-third of the sample is not used to train the data, instead used to evaluate its performance. These samples are called out of bag samples.

Advantages and Disadvantages of Random Forest Algorithm:

Advantages:

1. It can be used in classification and regression problems.
2. It solves the problem of overfitting as output is based on majority voting or averaging.
3. It performs well even if the data contains null/missing values.
4. Each decision tree created is independent of the other, thus it shows the property of parallelization.
5. It is highly stable as the average answers given by a large number of trees are taken.
6. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
7. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
8. It doesn't have to segregate data into trains and tests as there will always be 30% of the data which is not seen by the decision tree made out of bootstraps.
9. It can be used in classification and regression problems.
10. It solves the problem of overfitting as output is based on majority voting or averaging.
11. It performs well even if the data contains null/missing values.
12. Each decision tree created is independent of the other, thus it shows the pr property of parallelization.
13. It is highly stable as the average answers given by a large number of them are taken.
14. It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
15. It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.
16. It doesn't have to segregate data into trains and tests as there will always be 30% of the data which is not seen by the decision tree made out of bootstraps.

Disadvantages:

1. Random forest is highly complex when compared to decision trees where decisions can be made by following the path of the tree.
2. Training time is longer compared to other models due to its complexity. Whenever it has to make a prediction each decision tree has to generate output for the given input data.

CHAPTER 6

RESULT AND DISCUSSION

6.1 ANALYSE CRIME:

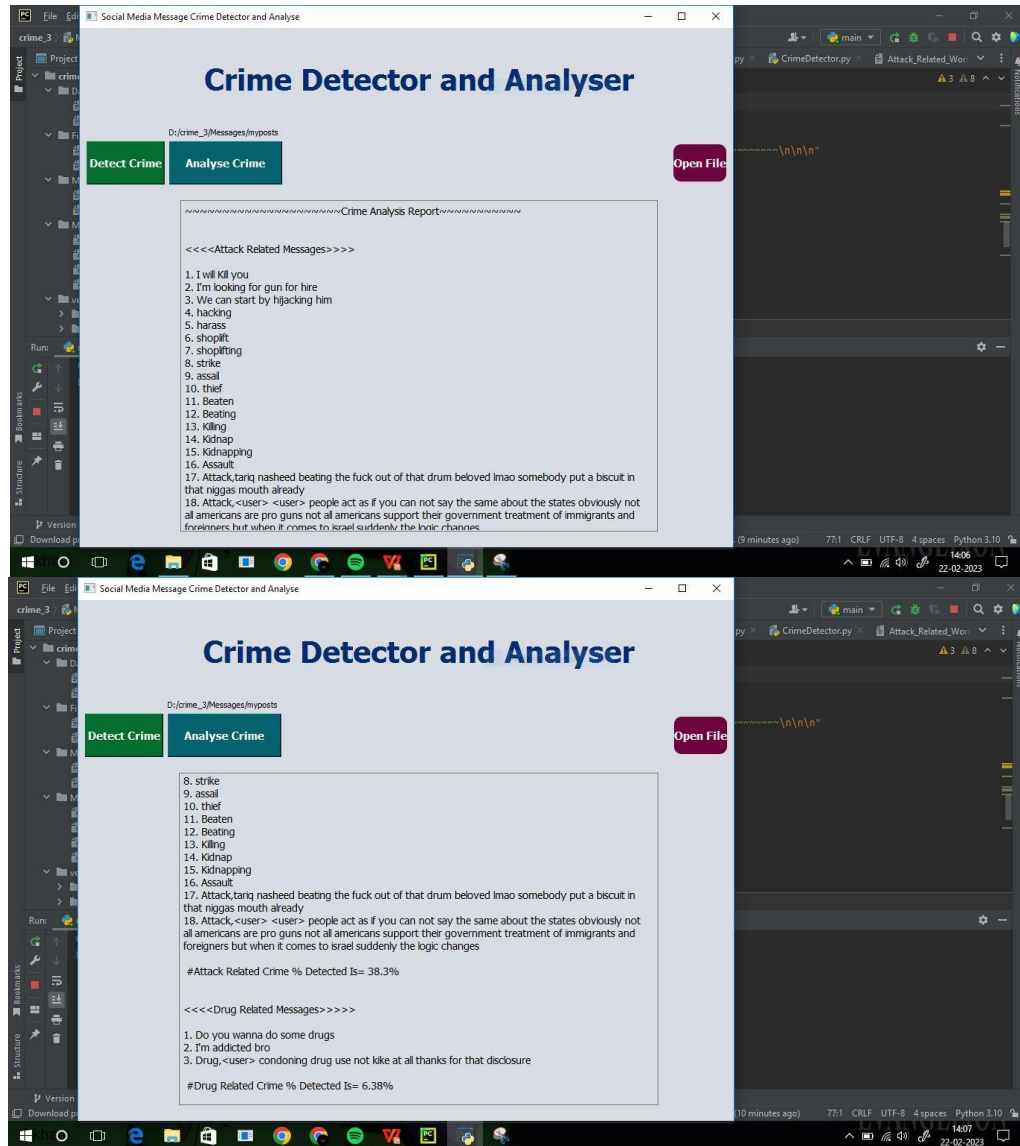


Fig 6.1: Analyse Crime

A method that uses SVM filtering techniques can be used to find postings from the social media data collections that are related to crimes. By testing and training the data we will analyze the type of the crime percentage.

6.2 DETECT CRIME:

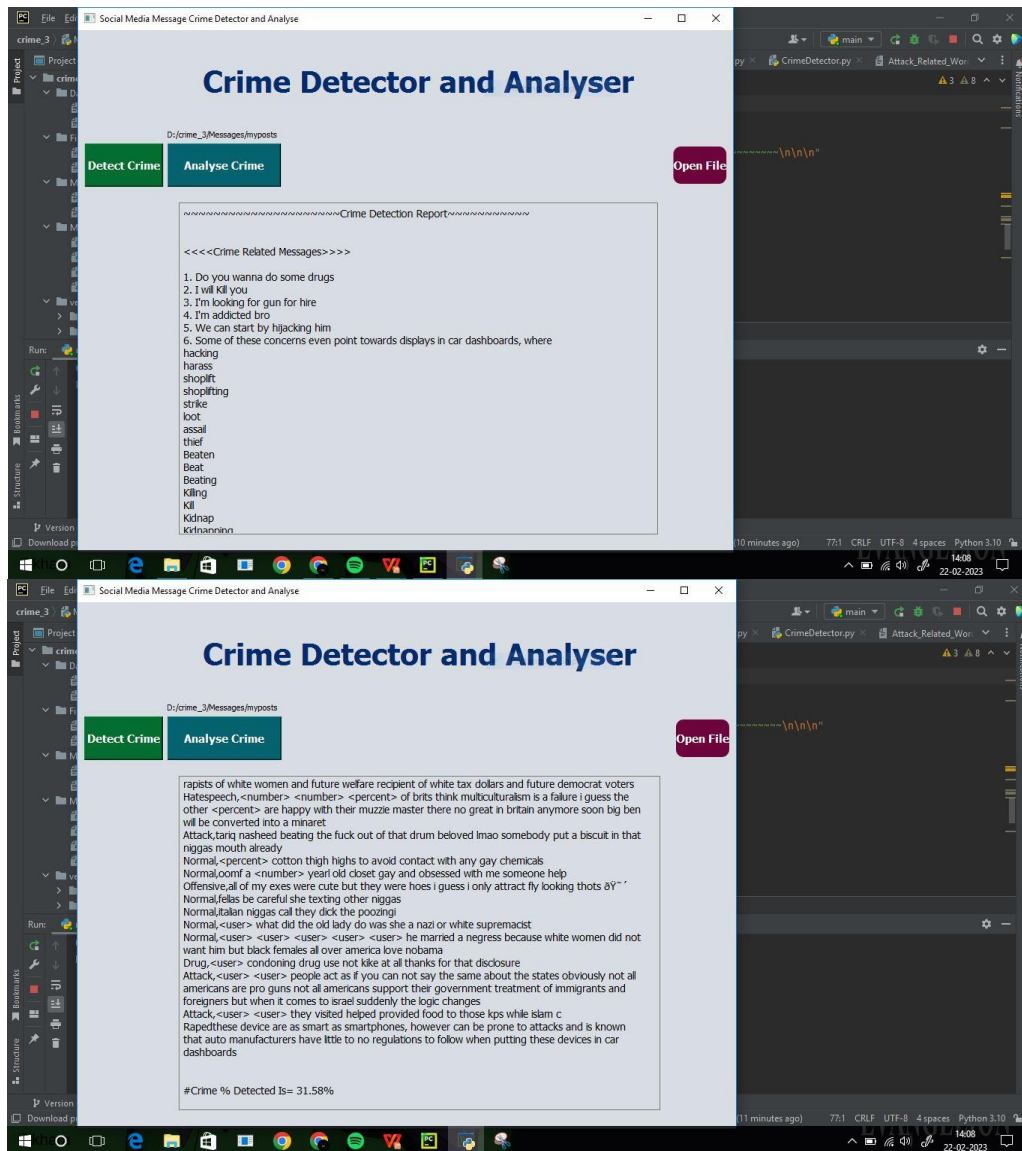


Fig 6.2: Detect Crime

By observing crime related messages through social media using SVM filter and random forest we conclude that type of the crime and percentage of the crime detected.

CHAPTER 7

CONCLUSION & FUTURE SCOPE

7.1 CONCLUSION

The project improves the accuracy of the detection of crime related posts from social media text messages. Here, First we applied a keyword based filter, and then to remove the noise, we applied the SVM based filter and random forest classification. According to the existing works, SVM has the best accuracy among the classifiers. So, The SVM is used for the research.

In conclusion, utilizing multi-model data and employing various machine learning algorithms can enhance the accuracy and effectiveness of crime prediction models. However, continuous refinement and evaluation of the models are necessary to ensure their reliability and validity in real-world applications.

7.2 FUTURE SCOPE

Future enhancements of this research work on training bots to predict the crime prone areas by using machine learning techniques. method for enhancing the detection of posts about crimes in social media text messages. Here, a keyword-based filter is initially used, and then an SVM-based filter and a random forest classification are utilized to get rid of the noise. SVM has the highest accuracy of all the classifiers, according to the currently available research. Therefore, was employed in the studies.

REFERENCES

- [1] Ginger Saltos and Ella Haig, An Exploration of Crime prediction Using Data Mining on Open Data, International journal of Information technology & Decision Making, 2017.
- [2] Shiju Sathyadevan, Devan M.S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing (IEEE) 2014.
- [3] Khushabu A. Bokde, Tisksha P. Kakade, Dnyaneshwari S. Tumasare, Chetan G. Wadhai B.E Student, Crime Detection Techniques Using Data Mining and K-Means, International Journal of Engineering Research & technology (IJERT) ,2018.
- [4] H. Benjamin Fredrick David and A. Suruliandi, Survey on crime analysis and prediction using data mining techniques, ICTACT Journal on Soft computing, 2017.
- [5] Tushar Sonawanev, Shirin Shaikh, rahul Shinde, Asif Sayyad, Crime Pattern Analysis, Visualization and prediction Using Data Mining, Indian Journal of Computer Science and Engineering (IJCSE), 2015.
- [6] RajKumar. S, Sakkarai Pandi.M, Crime Analysis and prediction using data mining techniques, International Journal of recent trends in engineering & research, 2019.
- [7] Chhaya Chauhan, Smriti Sehgal, Crime analysis using data mining techniques and algorithms, International Conference on Computing, Communication and Automation, 2017.
- [8] Ayisheshim Almaw, Kalyani Kadam, Survey Paper on Crime Prediction using Ensemble Approach, International journal of Pure and Applied Mathematics, 2018.
- [9] Dr .M.Sreedevi, A.Harha Vardhan Reddy, ch.Venkata Sai Krishna Reddy, Review on crime Analysis and prediction Using Data Mining Techniques, International Journal of Innovative Research in Science Engineering and technology, 2018.
- [10] K.S.N .Murthy, A.V.S.Pavan kumar, Gangu Dharmaraju, international journal of engineering, Science and mathematics, 2017.
- [11] Deepiika k.K, Smitha Vinod, Crime analysis in india using data mining techniques, International journal of Engineering and technology, 2018.

- [12] Hitesh Kumar Reddy ToppyiReddy, Bhavana Saini, Ginika mahajan, Crime Prediction & Monitoring Framework Based on Spatial Analysis, International Conference on Computational Intelligence Data Science (ICCIDS 2018).
- [13] Johnson, S.D., Bowers, K.J., Birks, D.J. et al. Predictive Mapping of Crime by Proactive Police Deployment. *J Quant Criminol* 24, 237–256 (2008). DOI: 10.1007/s10940-008-9051-2
- [14] Mohler, G., Short, M., Brantingham, P. et al. Randomized Controlled Field Trials of Predictive Policing. *J Exp Criminol* 9, 247–274 (2013). DOI: 10.1007/s11292-013-9184-4
- [15] Ratcliffe, J.H., Taniguchi, T., Groff, E.R. The Philadelphia Foot Patrol Experiment: A Randomized Controlled Trial of Police Patrol Effectiveness in Violent Crime Hotspots. *Criminol Public Policy* 16, 589–618 (2017). DOI: 10.1111/1745-9133.12289
- [16] Gerber, M.S., Johnson, B.D., Buerger, M.E. Predicting Gun Violence in Chicago. In: Gerber M.S., Johnson B.D. (eds) *Policing and Crime Control*. Springer, Cham (2016). DOI: 10.1007/978-3-319-28676-4_4.

