CRIME DETECTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering Degree in computer science and Engineering

by

ANAGANI RAVI TEJA – 39110058

AMARENDRA KUMAR G - 39110049

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI – 600119

OCTOBER 2022

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

(Established under Section 3 of UGC Act, 1956)

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai– 600119

www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Professional Training Report is the bonafide work of **ANAGANI RAVI TEJA (39110058) and AMARENDRA KUMAR GOLLAPALLI (39110049**) who carried out project entitled "**CRIME DETECTION USING MACHINE LEARNING"** under my supervision from JUNE 2022 to SEPTEMBER 2022.

Internal Guide
Dr. Cruz Antony J, Ph.D.

Head of the Department
Dr. L. LAKSHAMANAN,M.E., Ph.D

Submitted for Viva voce Examination held on 01.02.2023

Internal Examiner **External Examiner**

DECLARATION

I, **ANAGANI RAVITEJA(39110058) and AMARENDRA KUMAR GOLLAPALLI (39110049**) hereby declare that the Project report that entitled **"CRIME DETECTION USING MACHINE LEARNING "** done by us under theguidance of **Dr. Cruz Antony J, Assistant Professor, Ph.D.** It is submitted in partial fulfilment of the requirements for the award of Bachelor of Engineering degree in **COMPUTER SCIENCE AND ENGINEERING**

DATE:01 FEB 2023

PLACE:CHENNAI

SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E., Ph.D., Dean, School of Computing** , **Dr. L. Lakshamanan, M.E., Ph.D., Head of the Department, Dept. of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. Cruz Antony J, Assistant Professor, Ph.D.,** for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of **Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

A crime is a deliberate act that can get 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.

In order to help 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, using efficient methods ie 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 is being used in multiple domains, an example to such cases are automated or self driving cars. By Machine learning algorithms, there is a way to predict certain results based upon 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.

CHAPTER 1
INTRODUCTION

Day after day crime data rate has been increasing as a result of modern technologies and hi-tech methods as it is helping criminals to perform illegal activities. Evidently to 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 are acquired from various blogs, news and websites. The huge 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 hotspots 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 neighbours.

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 behaviour. 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 dataset 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 behaviour that 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 forecasted using a variety of multi-linear regression algorithm.

1.3 PROBLEM STATEMENT

.

- The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of 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

Literature survey is the most important step in 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 lot of external support. This support can be obtained from senior programmers, from 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 survey all the required needs for developing the project. For every project Literature survey is the most important sector in software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, man power, economy, and company strength. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what are all the necessary softwareare needed to proceed with the next step such as developing the tools, and the associated operations

2.2 An Exploration of Crime Prediction Using Data Mining on Open Data

Ella Haig and Ginger saltos in 2017 has 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 behaviour 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 behaviour crimes. Three algorithms are used from different categories of approaches: instance based learning, regression and decision trees. The data are from the UK police and contain 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 behaviour frequency.

2.3 Crime Analysis and Prediction Using Data Mining

Shuji Sathyadevan and others (2021) has demonstrated on Crime analysis and prevention is a systematic approach for identifying and analysing patterns and trends in crime. This System can predict regions which have high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the 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 an 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.

2.4 Crime Detection Techniques Using Data Mining and K-Means

Khushabu A. Bokde and others in 2018 signified on how Crimes will somehow influence organizations and institutions when occurred frequently in a society. Thus, it seems necessary to study reasons, factors and relations between occurrence of different crimes and finding 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 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 within 1990 to 2011. Further weights are assigned to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm (GA) is used for optimizing of Outlier Detection operator parameters using Rapid Miner tool.

2.5 Survey on crime analysis and prediction using data mining techniques

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 had been performed. However, only few efforts have made in the criminology field. Many few have taken efforts for comparing 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 the criminal movements and criminal activity involvement in the 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 has been applied towards criminal identification. This project presents the survey on the Crime analysis and crime prediction using several Data Mining techniques.

2.6 Crime Pattern Analysis, Visualizations and Prediction Using Data Mining

Tushar Sonawanev and others in 2017 has demonstrated about how crime against women these days has become problem of every nation around the globe many countries are trying to curb this problem. Preventive are 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 crime.

This data can prove very useful in analysing and predicting crime and help us prevent the crime to some extent. Crime analysis is an area of vital importance in police department. Study of crime data can help us analyse crime pattern, interrelated clues & important hidden relations between the crimes. That is why data mining can be great aid to analyse, visualize and predict crime using crime data set. Classification and correlation of data set makes it easy to understand similarities & dissimilarities amongst the data objects. Data objects are grouped using clustering technique. Dataset is 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 administrationto plan strategies for prevention of crime, further using data mining technique data can be predicted and visualized in various form in order to provide better understanding of crime patterns.

2.7 Crime Analysis and prediction using Data Mining Techniques

Rajkumar .S and others (2019) has 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 high probability for crime occurrence and can visualize crime prone areas. With the increasing advent of computerized systems, crime data analysts can help the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, previously unknown, useful information from an unstructured data are 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.

2.8 Systematic Review of Crime Data Mining

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 the Law enforcement officers to speed up the process of solving crimes. Using the concept of data mining, previously unknown, useful information from an unstructured data can be analysed. Predictive policing means, using analytical and predictive techniques, to identify criminal and it has been found to be pretty much effective in doing the same. Because of the increased crime rate over the years, there is a huge amount of crime data stored in warehouses that has to be handled which would be very difficult to analyse manually, and also now-a-days, criminals are becoming technologically advance, so there is need to use advance technologies in order to keep police ahead of them. This project mainly focuses on the review of algorithms and techniques used for identify the criminals.

2.9 Survey paper on Crime Prediction using Ensemble Approach

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

CHAPTER 3

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, a Sting Ray can help us only when the suspect is using a phone, which should be switched on.

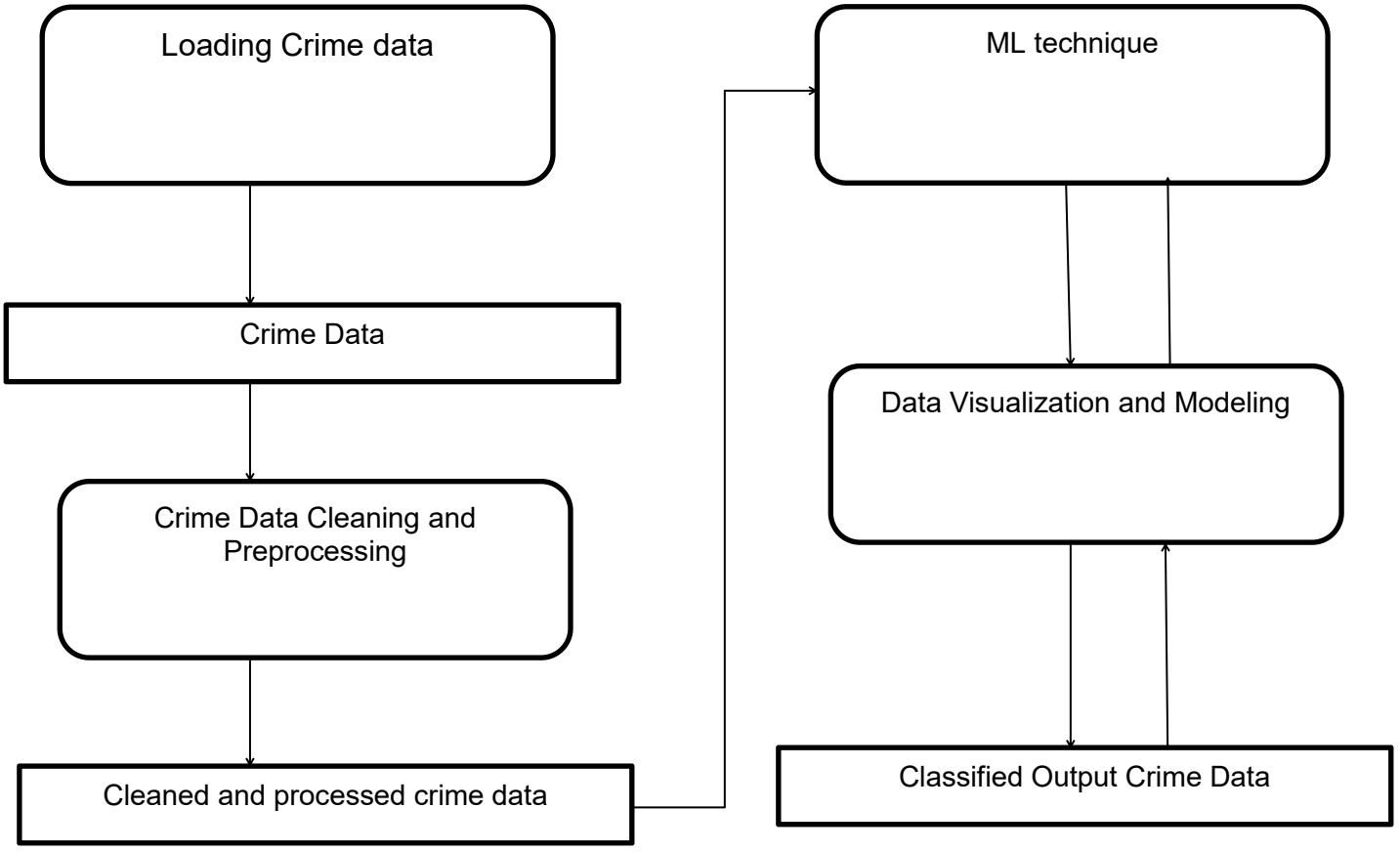
Thus, it is only useful when the information regarding the stake out location is correct. Based on this information, the ever-evolving technology has yet again produced a smart way to conduct surveillance. The introduction of deep learning, ML,and computer vision techniques has provided us with a new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

3.2 PROPOSED SYSTEM

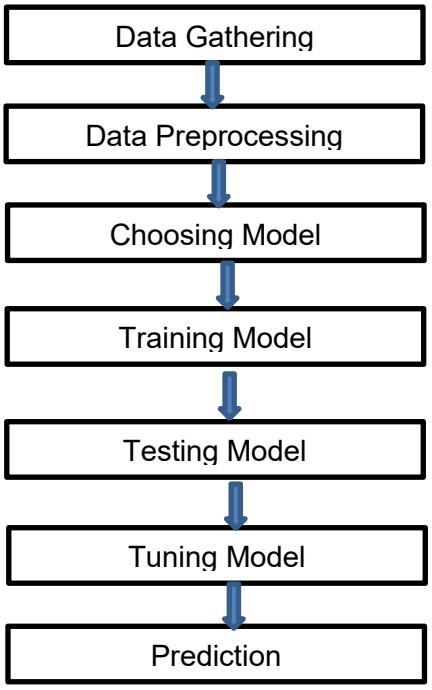
An approach uses SVM filtering methods to detect the crime-related post from the social media data set can be proposed. There are four main phases. In the first phase, social media text posts that relate to the crimes are extracted. In the second phase, the preprocessing techniques are applied to clean 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 random forest classifier is used for classification to categorize the data. The main steps include Formatting, cleaning and sampling. Cleaning process is usedfor removal or fixing of some missing data there may be data that are incomplete.

3.2.1 BLOCK DIAGRAM



3.2.2 FLOW DIAGRAM



CHAPTER 4

4.1 Machine learning

What are the 7 steps of machine learning?

7 Steps of Machine Learning

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

Introduction:

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 performance.

Python Libraries that would be need to achieve the task:

1.Numpy

2.Pandas

3.Sci-kit

4. Matplotlib

Understanding the machine learning workflow

Following can define the machine learning workflow in 3 stages.

1. Gathering data
2. Data pre-processing
3. Researching the model that will be best for the type of data
4. Training and testing the model
5. Evaluation

What is the machine learning Model?

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

1. 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 using 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 utiltized 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.

2. Data pre-processing

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

What is data pre-processing?

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to 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 the analysis.
Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

Why do we need it?

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 outliners, 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.

Three Types of Data

1. Numeric e.g. income, age

2. Categorical e.g. gender, nationality

3. Ordinal e.g. low/medium/high

How can data pre-processing be performed?

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

1. **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.

2. **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.

3. **Filling the missing values:** Whenever there is an encounter about missing data in the data set then the missing data can be filled manually, most commonly the mean, median or highest frequency value is used.

4. **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.

5. **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 = 800 Kg; due to mistyping of extra 0]

4.2 Supervised Learning:

In Supervised learning, an AI system is presented with data which is labelled, which means that each data tagged with the correct label.

The supervised learning is categorized into 2 other categories which are "**Classification**" and "**Regression**".

4.3 Classification:

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".

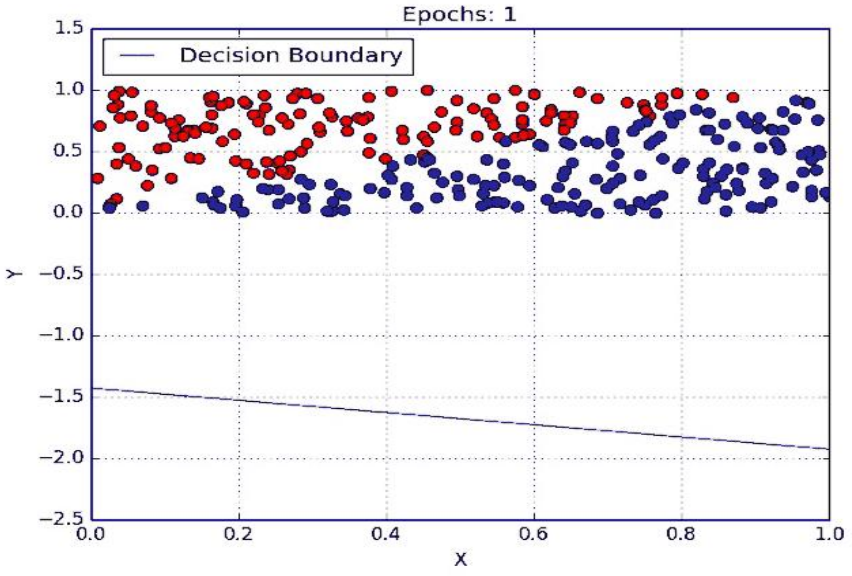


Fig no - 4.3.1 (Classification)

As shown in the above representation, we have 2 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 imaged, so the best fit line that separates both classes of flowers can be created.

These some most used classification algorithms.

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

4.4 Regression:

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

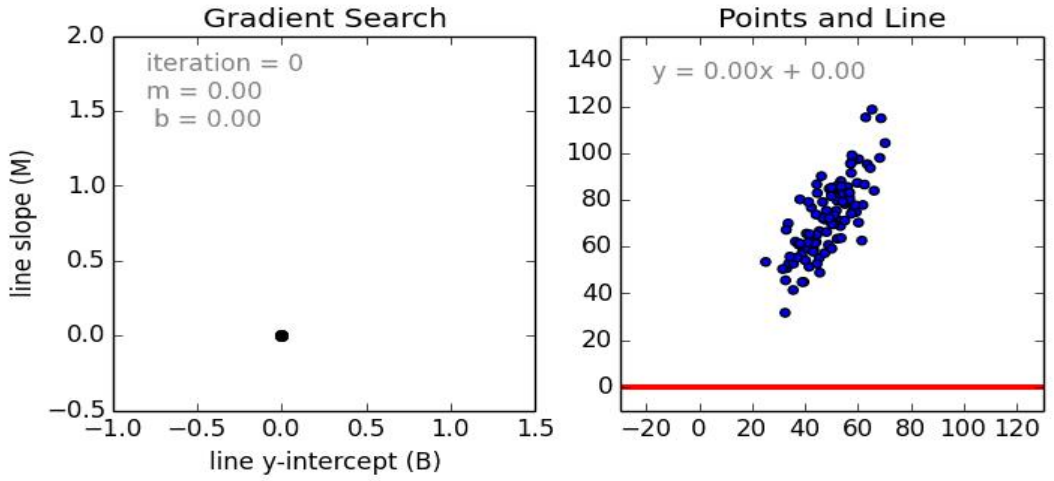


Fig no – 4.4.1 (Regression)

As shown in the above representation, the graph's X-axis is the 'Test scores' and the Y-axis represents 'IQ' can be imagined. So we try to 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 some most used regression algorithms.

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

4.5 Unsupervised Learning:

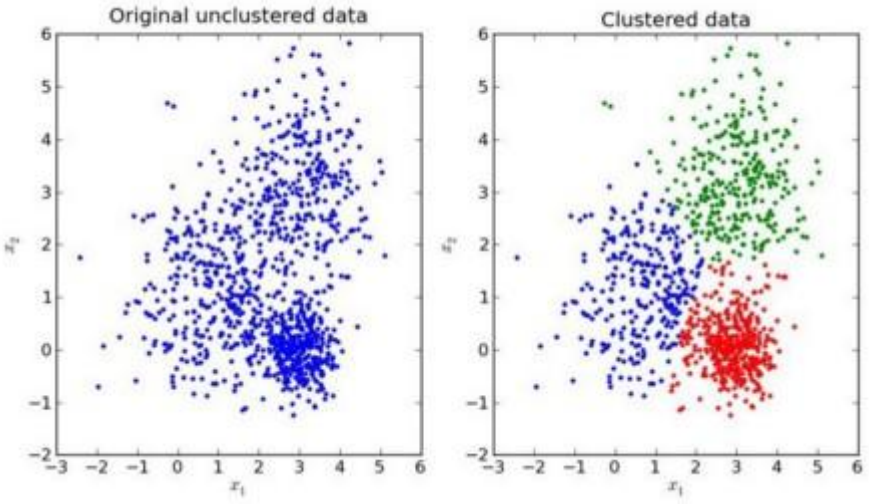


Fig no- 4.5.1 (Unsupervised)

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

4.6 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.

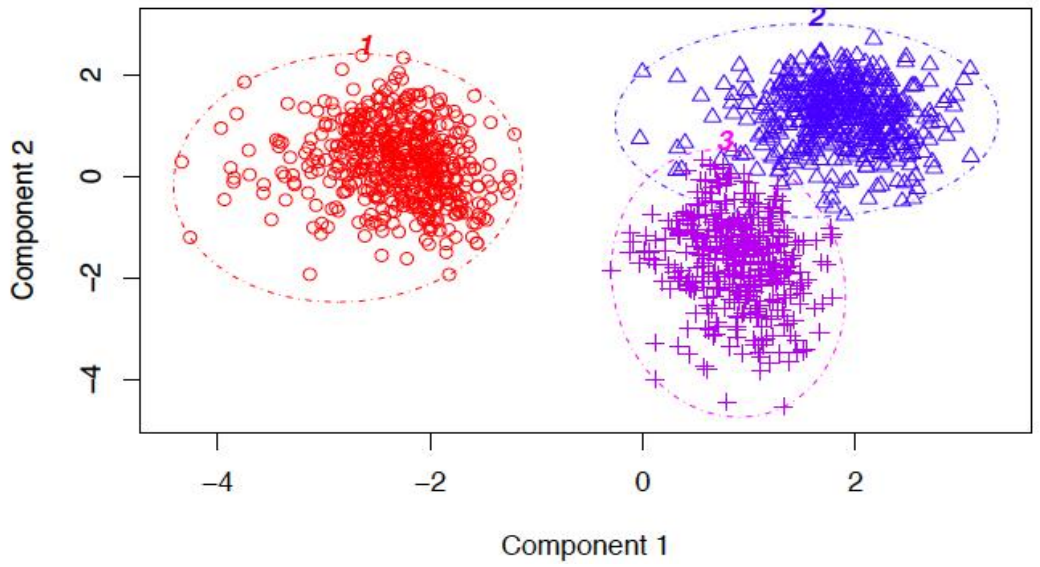


Fig no- 4.6.1(Clustering)

Methods used for clustering are:

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

CONCLUSION

Future enhancements of this research works 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 on the currently available research. Therefore, SVM was employed in the study.

REFERENCES

[1] Ginger Saltos and Mihaela Coacea, 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] Sarpreet kaur, Dr. Williamjeet Singh, Systematic review of crime data mining, International Journal of Advanced Research in computer science, 2015.

[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 minig techniques , International journal of Enginnering 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).