



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY



Accredited with "A" grade by NAAC
(DEEMED TO BE UNIVERSITY)

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 Project Report is the bonafide work of DEVARAKONDA KARTHEEK(RegNo:38110125), ANAGANI SYAM PRAKASH(RegNo:38110031)

who carried out the project entitled "**CRIME DETECTION USING MACHINE LEARNING**"
under my supervision from AUG 2021 TO MARCH 2022

Internal Guide
Dr. K. Mohana Prasad M.E., Ph.D

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

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

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. Lakshmanan M.E., Ph.D. ,** and **Dr.S.Vigneshwari M.E., Ph.D. Heads** of the Department of Computer Science and Engineering for providing 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./Mr./Ms for his 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.

CRIME DETCTION USING MACHINE LEARNING

ABSTARCT:

Social media text analytics is the process of deriving information from text sources. Text analysis can be applied to any text-based dataset, including social media. Crime is a major problem faced today by society even we are finding in social media. Crimes have affected the quality of life and economic growth badly . We can identify the crime patterns and predict the crimes by detecting and analyzing the historical data. However, some crimes are unregistered and unsolved due to a lack of evidence. Thus, detecting crimes is a still challenging task. We can use social media to detect crimes related activities. Because social media users sometimes convey messages related to his or her surrounding environment via social media message. In this paper, we proposed a machine learning approach to detect the crimes and analyses it on its type. As the first step, we fetch the text messages using predefined keywords relating to the crimes. Then, after the preprocessing, we applied a support vector machine- based filtering approach to eliminate the noise. And then Random forest is used for classification . Then in the final stage, we analyses and categorize the crime type.

Chapter 1

INTRODUCTION

- Security is a very necessary aspect of life. Unless we are safe, our most important needs cannot be met. Security is therefore a requirement in human life that helps us to achieve our goals collectively or individually. Crimes are a social problem, which costs our society deeply in many aspects. The ability to identify unsafe areas for crime and identify the most recent crime in a particular location has become a growing concern for both local authorities and residents. On the other hand, people are always interested in improving safety and make reliable relationships with neighbors when living in a busy society. The prevalence of crime is one of the greatest challenges for societies around the world, particularly in metropolitan areas. There are more researches regarding social crimes in the world, but using social media, there are few types of research about crimes and their behavior. Therefore, the paper aims in presenting a prediction model (algorithm) by using the machine-learning technique, which is meant to possess a strong capability to predict crimes by factors of social media dataset using the Data Mining concept. Our main data source is social media. The main goal is to identify each hidden data source and predict results

OBJECTIVES OF THE PROJECT:

- The main objective of the project is to predict the crime rate and analyze the crime rate to be happened in future. Based on this Information the officials can take charge and try to reduce the crime rate.
- The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable)
- The system will look at how to convert crime information into a regression problem, so that it will help detectives in solving crimes faster.
- Crime analysis based on available information to extract crime patterns. Using various multi linear regression techniques, frequency of occurring crime can be predicted based on territorial distribution of existing data and Crime recognition.

Existing system:

- After finding and understanding various distinct methods used by the police for surveillance purposes, we determined the importance of each method.
- Each surveillance method can perform well on its own and produce satisfactory results, although for only one specific characteristic, that is, if we use a Sting Ray, it 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, we can see how 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.

PROBLEM STATEMENT

- Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities
- 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 we can predict the type of crime which will occur in a particular area.

Chapter 2

LITERATURE REVIEW

2.1 CRIME RATE PREDICTION

Abstract

Crime is one of the biggest and dominating problem in our society. Daily there are huge number of crimes committed frequently. Here the dataset consists of the date and the crime rate that has taken place in the corresponding years. In this project the crime rate is only based on the robbery. We use linear regression algorithm to predict the percentage of the crime rate in the future by using the previous data information. The date is given as an input to the algorithm and the output is the percentage of the crime rate in that particular year. Keywords: Crime rate, number of crimes, regression algorithm, Machine learning.

1. INTRODUCTION

Crimes are the significant threat to the humankind. There are many crimes that happen in regular intervals of time. Perhaps it is increasing and spreading at a fast and vast rate. Crimes happen from small village, town to big cities. Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which the case solving could be faster. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster. 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 we can predict the type of crime which will occur in a particular area with crime perception. The objective would be to train a model for prediction. The training would be done using Training data set which will be validated using the test dataset. The Multi Linear Regression (MLR) will be used for crime prediction. Visualization of dataset is done to analyze the crimes which

may have occurred in a particular year and based on population and number of crimes. This work helps the law enforcement agencies to predict and detect the crime_per capita in an area and thus reduces the crime rate.

METHODOLOGY:

Initially, the data set is prepared manually. After identifying the relationships and visualizing the data, we create a regression model for forecasting the per capita. For this model, we have used Multi Linear regression model. Other models such as the Linear Regression and Logistic Regression models were also tested, but the Multi Linear regression produced the minimal error while training the model. This regression model predicts the per capita of Crime rate that is going to be happen in future by taking different parameters. Logistic Regression: In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". KNN: In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a nonparametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression. In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors. Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y. In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression that involves more than one explanatory variable

RESULTS

The Support vector Regression model results in classifying the test data. It predicts the stock crime rate value that is going to be happening in future. It uses train set to train and fit the model. The test data is transformed and predicts the accurate result. By Using Django and Aws we prepared the UserInterface. The below figure describes that home page takes the year, murder, attempts, cheating etc., values as input in text boxes.

2.2 Crime forecasting: a machine learning and computer vision approach to crime prediction and prevention

Abstract

A crime is a deliberate act that can cause physical or psychological harm, as well as property damage or loss, and can lead to punishment by a state or other authority according to the severity of the crime. The number and forms of criminal activities are increasing at an alarming rate, forcing agencies to develop efficient methods to take preventive measures. In the current scenario of rapidly increasing crime, traditional crime-solving techniques are unable to deliver results, being slow paced and less efficient. Thus, if we can come up with ways to predict crime, in detail, before it occurs, or come up with a “machine” that can assist police officers, it would lift the burden of police and help in preventing crimes. To achieve this, we suggest including machine learning (ML) and computer vision algorithms and techniques. In this paper, we describe the results of certain cases where such approaches were used, and which motivated us to pursue further research in this field. The main reason for the change in crime detection and prevention lies in the before and after statistical observations of the authorities using such techniques. The sole purpose of this study is to determine how a combination of ML and computer vision can be used by law agencies or authorities to detect, prevent, and solve crimes at a much more accurate and faster rate. In summary, ML and computer vision techniques can bring about an evolution in law agencies.

Introduction

Computer vision is a branch of artificial intelligence that trains the computer to understand and comprehend the visual world, and by doing so, creates a sense of understanding of a machine’s surroundings. It mainly analyzes data of the surroundings from a camera, and thus its applications are significant. It can be used for face recognition, number plate recognition, augmented and mixed realities, location determination, and identifying objects. Research is currently being conducted on the formation of mathematical techniques to recover and make it possible for computers to comprehend 3D images. Obtaining the 3D visuals of an object helps us with object detection, pedestrian detection, face recognition, Eigenfaces active appearance and 3D shape models, personal photo collections, instance recognition, geometric alignment, large databases, location recognition, category recognition, bag of words, part-based models, recognition with segmentation, intelligent photo editing,

context and scene understanding, and large image collection and learning, image searches, recognition databases, and test sets. These are only basic applications, and each category mentioned above can be further explored. In, VLFeat is introduced, which is a library of computer vision algorithms that can be used to conduct fast prototyping in computer vision research, thus enabling a tool to obtain computer vision results much faster than anticipated. Considering face detection/human recognition, human posture can also be recognized. Thus, computer vision is extremely attractive for visualizing the world around us.

Machine learning (ML) is an application that provides a system with the ability to learn and improve automatically from past experiences without being explicitly programmed. After viewing the data, an exact pattern or information cannot always be determined. In such cases, ML is applied to interpret the exact pattern and information. ML pushes forward the idea that, by providing a machine with access to the right data, the machine can learn and solve both complex mathematical problems and some specific problems. In general, ML is categorized into two parts: (1) supervised ML and (2) unsupervised ML. In supervised learning, the machine is trained on the basis of a predefined set of training examples, which facilitates its capability to obtain precise and accurate conclusions when new data are given. In unsupervised learning, the machine is given a set of data, and it must find some common patterns and relationships between the data its own. Neural networks, which are important tools used in supervised learning, have been studied since the 1980s. the author suggested that different aspects are needed to obtain an exit from nondeterministic polynomial (NP)-completeness, and architectural constraints are insufficient. However, it was proved that NP-completeness problems can be extended to neural networks using sigmoid functions. Although such research has attempted to demonstrate the various aspects of new ML approaches, how accurate are the results?

Although various crimes and their underlying nature seem to be unpredictable, how unforeseeable are they? the authors pointed out that as society and the economy results in new types of crimes, the need for a prediction system has grown. crime trends and prediction technology called Mahanolobis and a dynamic time wrapping technique are given, delivering the possibility of predicting crime and apprehending the actual culprit. As described in 1998, the United States National Institute of Justice granted five grants for crime forecasting as an extension to crime mapping. Applications of crime forecasting are currently being used by law enforcement in the United States, the United Kingdom, the Netherlands, Germany, and Switzerland. Nowadays, criminal intellect with the help of advances in technology is improving with each passing year. Consequently, it has become necessary for us to provide the police department and the government with the means of a new and powerful machine (a set of programs) that can help them in their process of solving crimes. The main aim of crime forecasting is to predict crimes before they occur, and thus, the importance of using crime forecasting methods is extremely clear. Furthermore, the prediction of crimes can sometimes be crucial because it may potentially save the life of a victim, prevent lifelong trauma, and avoid damage to private property. It may even be used to predict possible terrorist crimes and activities. Finally, if we implement predictive policing with a considerable level of accuracy, governments can apply other primary resources such as police manpower, detectives, and funds in other fields of crime solving, thereby curbing the problem of crime with double the power.

In this paper, we aim to make an impact by using both ML algorithms and computer vision methods to predict both the nature of a crime and possibly pinpoint a culprit.

Beforehand, we questioned whether the nature of the crime was predictable. Although it might seem impossible from the outside, categorizing every aspect of a crime is quite possible. We have all heard that every criminal has a motive. That is, if we use motive as a judgment for the nature of a crime, we may be able to achieve a list of ways in which crimes can be categorized. Herein, we discuss a theory where we combine ML algorithms to act as a database for all recorded crimes in terms of category, along with providing visual knowledge of the surroundings through computer vision techniques, and using the knowledge of such data, we may predict a crime before it occurs.

Present technological used in crime detection and prediction

Crime forecasting refers to the basic process of predicting crimes before they occur. Tools are needed to predict a crime before it occurs. Currently, there are tools used by police to assist in specific tasks such as listening in on a suspect's phone call or using a body cam to record some unusual illegal activity. Below we list some such tools to better understand where they might stand with additional technological assistance.

One good way of tracking phones is through the use of a stingray, which is a new frontier in police surveillance and can be used to pinpoint a cellphone location by mimicking cellphone towers and broadcasting the signals to trick cellphones within the vicinity to transmit their location and other information. An argument against the usage of stingrays in the United States is that it violates the fourth amendment. This technology is used in 23 states and in the district of Columbia. the authors provide insight on how this is more than just a surveillance system, raising concerns about privacy violations. In addition, the Federal Communications Commission became involved and ultimately urged the manufacturer to meet two conditions in exchange for a grant: (1) "The marketing and sale of these devices shall be limited to federal, state, local public safety and law enforcement officials only" and (2) "State and local law enforcement agencies must advance coordinate with the FBI the acquisition and use of the equipment authorized under this authorization." Although its use is worthwhile, its implementation remains extremely controversial.

A very popular method that has been in practice since the inception of surveillance is "the stakeout". A stakeout is the most frequently practiced surveillance technique among police officers and is used to gather information on all types of suspects. the authors discuss the importance of a stakeout by stating that police officers witness an extensive range of events about which they are required to write a report. Such criminal acts are observed during stakeouts or patrols; observations of weapons, drugs, and other evidence during house searches; and descriptions of their own behavior and that of the suspect during arrest. Stakeouts are extremely useful, and are considered 100% reliable, with the police themselves observing the notable proceedings. However, are they actually 100% accurate? All officers are humans, and all humans are subject to fatigue. The major objective of a stakeout is to observe wrongful activities. Is there a tool that can substitute its use? We will discuss this point herein.

Another way to conduct surveillance is by using drones, which help in various fields such as mapping cities, chasing suspects, investigating crime scenes and accidents, traffic management and flow, and search and rescue after a disaster. legal issues regarding the use of drones and airspace distribution problems are described. Legal issues include the privacy concerns raised by the public, with the police gaining

increasing power and authority. Airspace distribution raises concerns about how high a drone is allowed to go.

Other surveillance methods include face recognition, license plate recognition, and body cams. the authors indicated that facial recognition can be used to obtain the profile of suspects and analyze it from different databases to obtain more information. Similarly, a license plate reader can be used to access data about a car possibly involved in a crime. They may even use body cams to see more than what the human eye can see, meaning that the reader observes everything a police officer sees and records it. Normally, when we see an object, we cannot recollect the complete image of it. the impact of body cams was studied in terms of officer misconduct and domestic violence when the police are making an arrest. Body cams are thus being worn by patrol officers. the authors also mentioned how protection against wrongful police practices is provided. However, the use of body cams does not stop here, as other primary reasons for having a body camera on at all times is to record the happenings in front of the wearer in hopes of record useful events during daily activities or during important operations.

Although each of these methods is effective, one point they share in common is that they all work individually, and while the police can use any of these approaches individually or concurrently, having a machine that is able to incorporate the positive aspects of all of these technologies would be highly beneficial.

ML techniques used in crime prediction

a comparative study was carried out between violent crime patterns from the Communities and Crime Unnormalized Dataset versus actual crime statistical data using the open source data mining software Waikato Environment for Knowledge Analysis (WEKA). Three algorithms, namely, linear regression, additive regression, and decision stump, were implemented using the same finite set of features on communities and actual crime datasets. Test samples were randomly selected. The linear regression algorithm could handle randomness to a certain extent in the test samples and thus proved to be the best among all three selected algorithms. The scope of the project was to prove the efficiency and accuracy of ML algorithms in predicting violent crime patterns and other applications, such as determining criminal hotspots, creating criminal profiles, and learning criminal trends.

When considering WEKA the integration of a new graphical interface called Knowledge Flow is possible, which can be used as a substitute for Internet Explorer. IT provides a more concentrated view of data mining in association with the process orientation, in which individual learning components (represented by java beans) are used graphically to show a certain flow of information. The authors then describe another graphical interface called an experimenter, which as the name suggests, is designed to compare the performance of multiple learning schemes on multiple data sets.

the potential of applying a predictive analysis of crime forecasting in an urban context is studied. Three types of crime, namely, home burglary, street robbery, and battery, were aggregated into grids of 200 m × 250 m and retrospectively analyzed. Based on the crime data of the previous 3 years, an ensemble model was applied to synthesize

the results of logistic regression and neural network models in order to obtain fortnightly and monthly predictions for the year 2014. The predictions were evaluated based on the direct hit rate, precision, and prediction index. The results of the fortnightly predictions indicate that by applying a predictive analysis methodology to the data, it is possible to obtain accurate predictions. They concluded that the results can be improved remarkably by comparing the fortnightly predictions with the monthly predictions with a separation between day and night.

crime predictions were investigated based on ML. Crime data of the last 15 years in Vancouver (Canada) were analyzed for prediction. This machine-learning-based crime analysis involves the collection of data, data classification, identification of patterns, prediction, and visualization. K-nearest neighbor (KNN) and boosted decision tree algorithms were also implemented to analyze the crime dataset. In their study, a total of 560,000 crime datasets between 2003 and 2018 were analyzed, and crime prediction with an accuracy of between 39% and 44% was obtained by predicting the crime using ML algorithms. The accuracy was low as a prediction model, but the authors concluded that the accuracy can be increased or improved by tuning both the algorithms and crime data for specific applications.

a ML approach is presented for the prediction of crime-related statistics in Philadelphia, United States. The problem was divided into three parts: determining whether the crime occurs, occurrence of crime and most likely crime. Algorithms such as logistic regression, KNN, ordinal regression, and tree methods were used to train the datasets to obtain detailed quantitative crime predictions with greater significance. They also presented a map for crime prediction with different crime categories in different areas of Philadelphia for a particular time period with different colors indicating each type of crime. Different types of crimes ranging from assaults to cyber fraud were included to match the general pattern of crime in Philadelphia for a particular interval of time. Their algorithm was able to predict whether a crime will occur with an astonishing 69% accuracy, as well as the number of crimes ranging from 1 to 32 with 47% accuracy.

the authors analyzed a dataset consisting of several crimes and predicted the type of crime that may occur in the near future depending on various conditions. ML and data science techniques were used for crime prediction in a crime dataset from Chicago, United States. The crime dataset consists of information such as the crime location description, type of crime, date, time, and precise location coordinates. Different combinations of models, such as KNN classification, logistic regression, decision trees, random forest, a support vector machine (SVM), and Bayesian methods were tested, and the most accurate model was used for training. The KNN classification proved to be the best with an accuracy of approximately 0.787. They also used different graphs that helped in understanding the various characteristics of the crime dataset of Chicago. The main purpose of this paper is to provide an idea of how ML can be used by law enforcement agencies to predict, detect, and solve crime at a much better rate, which results in a reduction in crime.

a graphical user interface-based prediction of crime rates using a ML approach is presented. The main focus of this study was to investigate machine-learning-based techniques with the best accuracy in predicting crime rates and explore its applicability with particular importance to the dataset. Supervised ML techniques were used to analyze the dataset to carry out data validation, data cleaning, and data

visualization on the given dataset. The results of the different supervised ML algorithms were compared to predict the results. The proposed system consists of data collection, data preprocessing, construction of a predictive model, dataset training, dataset testing, and a comparison of algorithms, as shown in Fig. 1. The aim of this study is to prove the effectiveness and accuracy of a ML algorithm for predicting violent crimes.

Result:

After finding and understanding various distinct methods used by the police for surveillance purposes, we determined the importance of each method. Each surveillance method can perform well on its own and produce satisfactory results, although for only one specific characteristic, that is, if we use a Sting Ray, it 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, we can see how 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. This is an intelligent approach to surveillance because it tries to mimic a human approach, but it does so 24 h a day, 365 days a year, and once it has been taught how to do things it does them in the same manner repeatedly.

Although we have discussed the aspects that ML and computer vision can achieve, but what are these aspects essentially? This brings us to the main point of our paper discussion, i.e., our proposed idea, which is to combine the point aspects of Sting Ray, body cams, facial recognition, number plate recognition, and stakeouts. New features include core analytics, neural networks, heuristic engines, recursion processors, Bayesian networks, data acquisition, cryptographic algorithms, document processors, computational linguistics, voiceprint identification, natural language processing, gait analysis, biometric recognition, pattern mining, intel interpretation, threat detection, threat classification. The new features are completely computer dependent and hence require human interaction for development; however, once developed, it functions without human interaction and frees humans for other tasks. Let us understand the use of each function.

Chapter 3

EXISTING SYSTEM:

- finding and understanding various distinct methods used by the police for surveillance purposes, we determined the importance of each method.
- Each surveillance method can perform well on its own and produce satisfactory results, although for only one specific characteristic, that is, if we use a Sting Ray, it 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, we can see how 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.

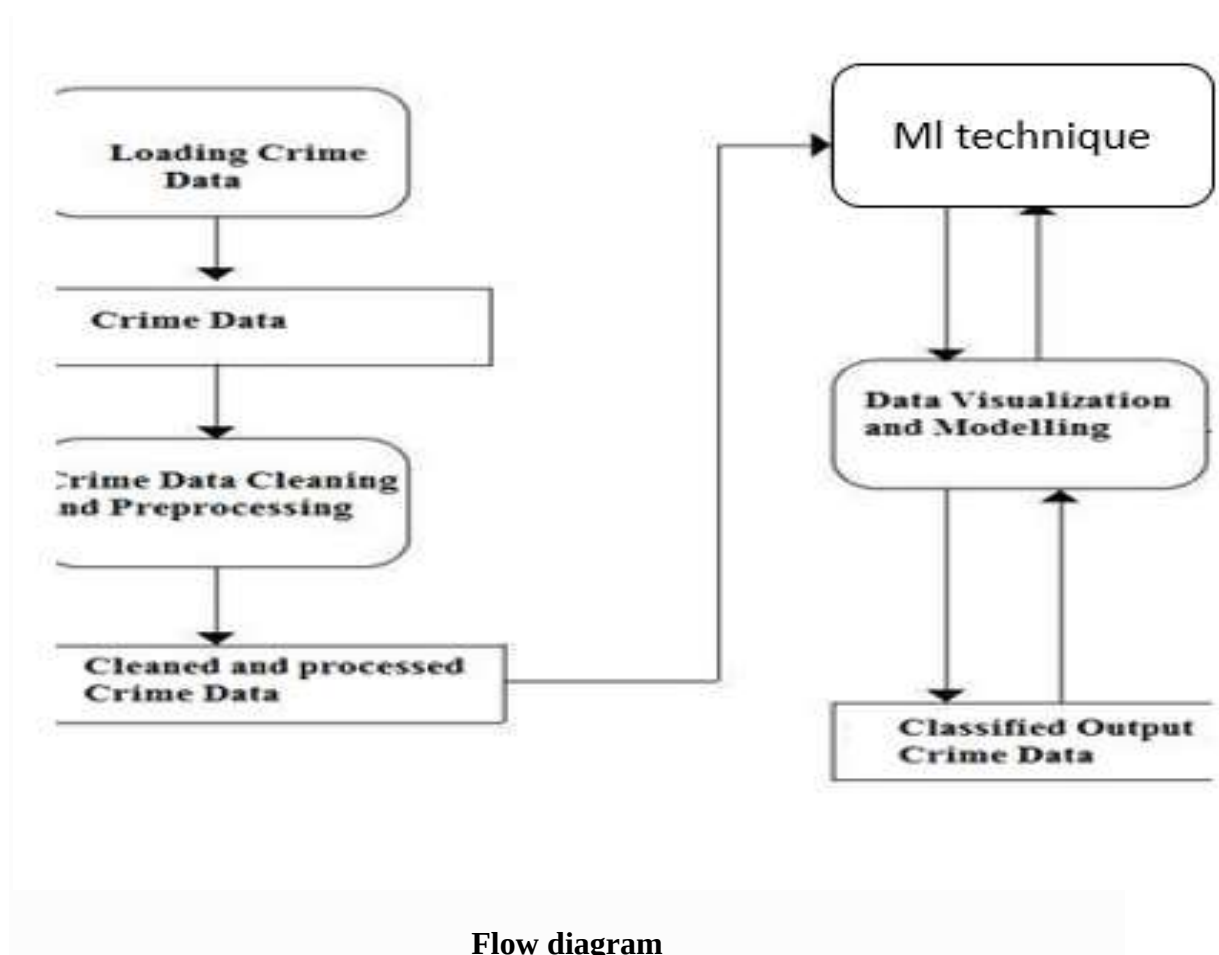
PROBLEM STATEMENT

- Crimes are of different type – robbery, murder, rape, assault, battery, false imprisonment, kidnapping, homicide. Since crimes are increasing there is a need to solve the cases in a much faster way. The crime activities have been increased at a faster rate and it is the responsibility of police department to control and reduce the crime activities
- 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 we can predict the type of crime which will occur in a particular area.

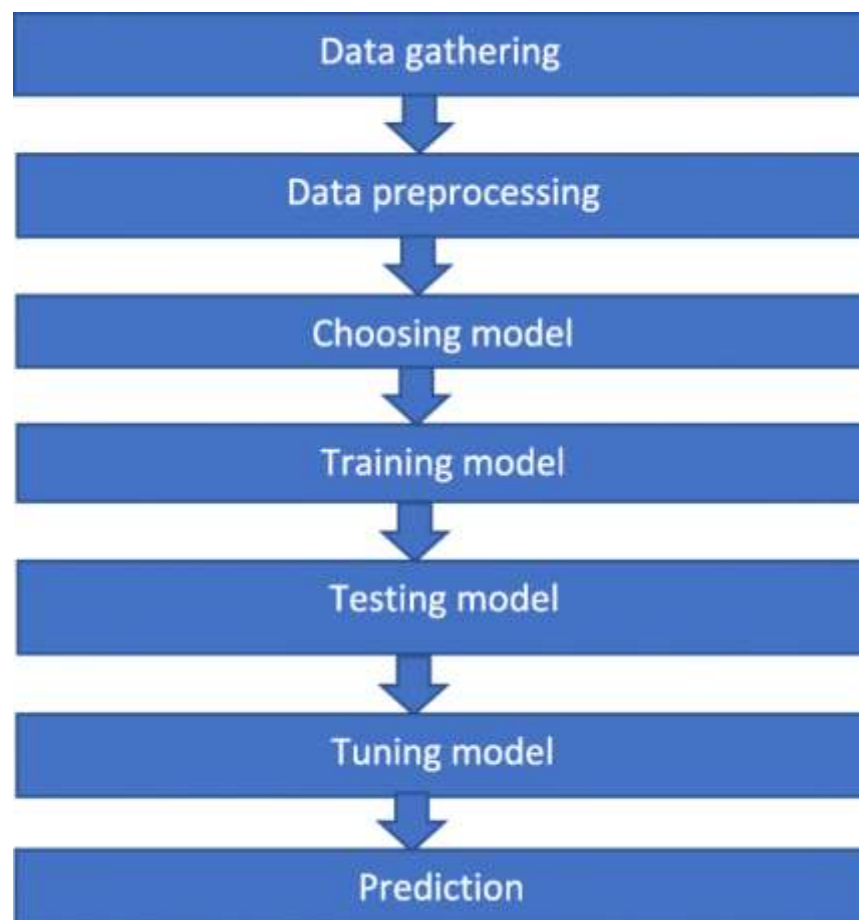
Propose system:

- Our proposed approach uses svm filtering methods to detect the crime-related post from the social media data set.
- There are four main phases. In the first phase, social media text posts that relate to the crimes are extracted. In the second phase, we applied the preprocessing techniques to clean the data set.
- Then, we calculated the TF- IDF values 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.

Block diagram



Flow diagram



Chapter 5

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, we will discuss the workflow of a Machine learning project this includes all the steps required to build the proper machine learning project from scratch.

We will also go over data pre-processing, data cleaning, feature exploration and feature engineering and show the impact that it has on Machine Learning Model Performance. We will also cover a couple of the pre-modelling 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 Learn
4. Matplotlib

Understanding the machine learning workflow

We 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

Okay but first let's start from the basics

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 you give garbage to the model, you will get garbage in return, 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 we desire to make, if we want to make an ML project that uses real-time data, then we can build an IoT system that using different sensors data. 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.

We can also use some free data sets 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 pre-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 we know 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, we definitely need 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.

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 we know 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 we encounter missing data in the data set then we can remove the row or column of data depending on our need. 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 we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4. **Machine learning:** If we have some missing data then we can predict what data shall be present at the empty position 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]

Researching the model that will be best for the type of data

Our main goal is to train the best performing model possible, using the pre-processed data.

Supervised Learning:

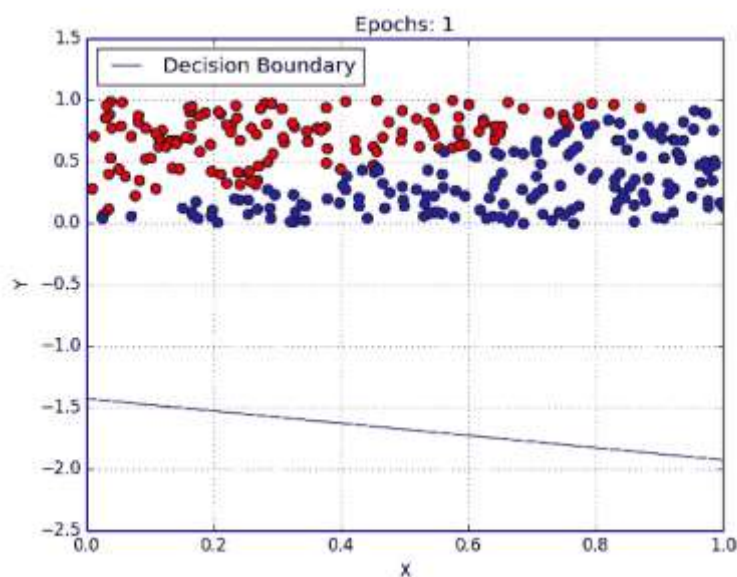
In Supervised learning, an AI system is presented with data which is labelled, which means that each data is tagged with the correct label.

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

Classification:

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”.



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’, we can image the X-axis as their ‘Sepal Width’ and the Y-axis as the ‘Sepal Length’, so we try to create the [best fit line](#) that separates both classes of flowers.

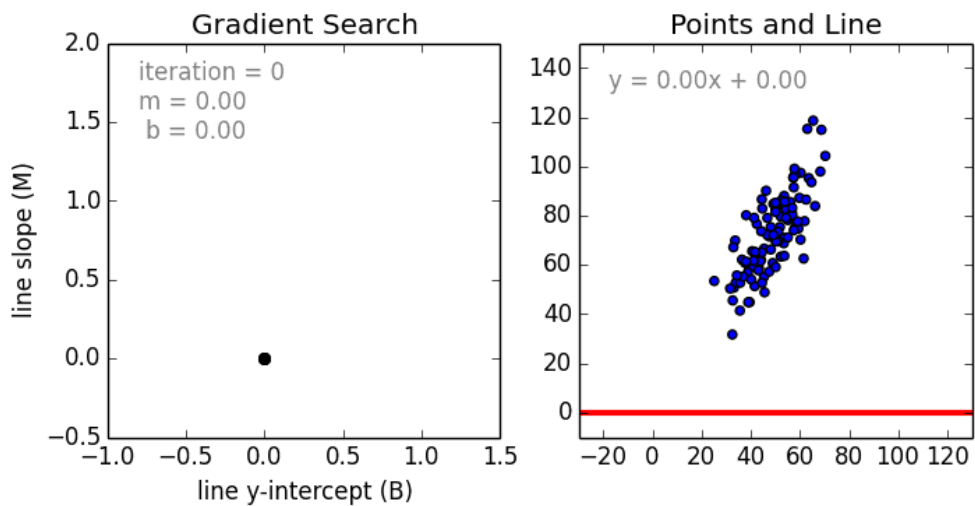
These some most used classification algorithms.

- **K-Nearest Neighbor**
- **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).

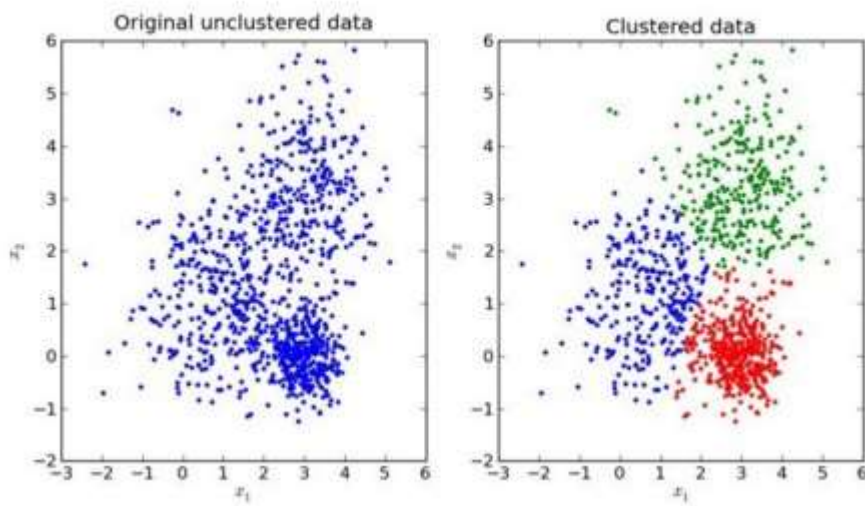


As shown in the above representation, we can imagine that the graph’s X-axis is the ‘Test scores’ and the Y-axis represents ‘IQ’. So we try to create the best fit line in the given graph so that we can use that line to predict any approximate IQ that isn’t present in the given data.

These some most used regression algorithms.

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

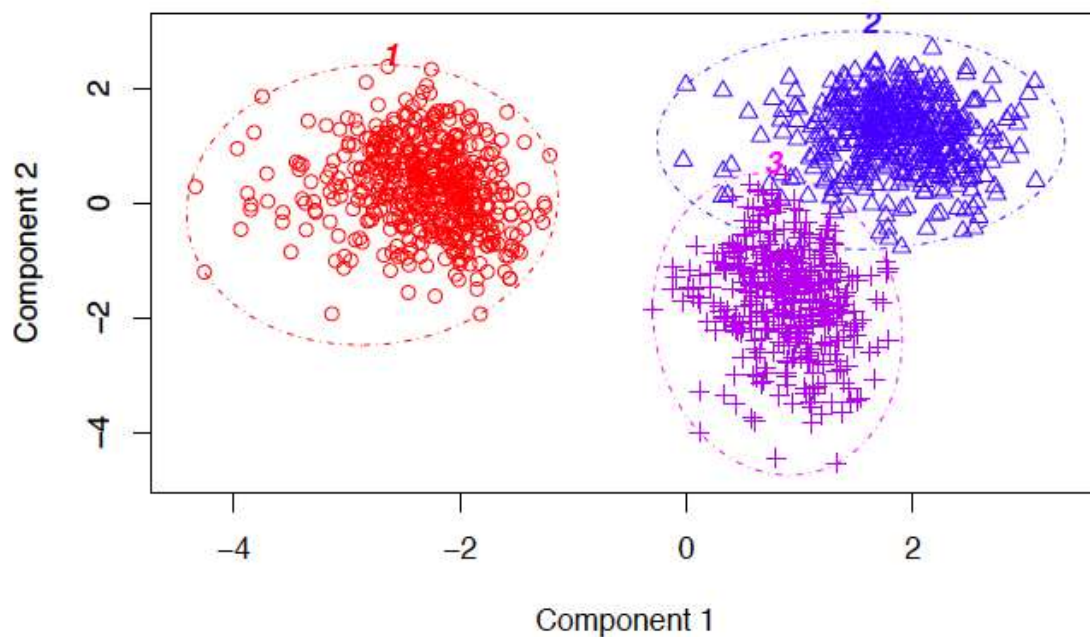
Unsupervised Learning:



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.

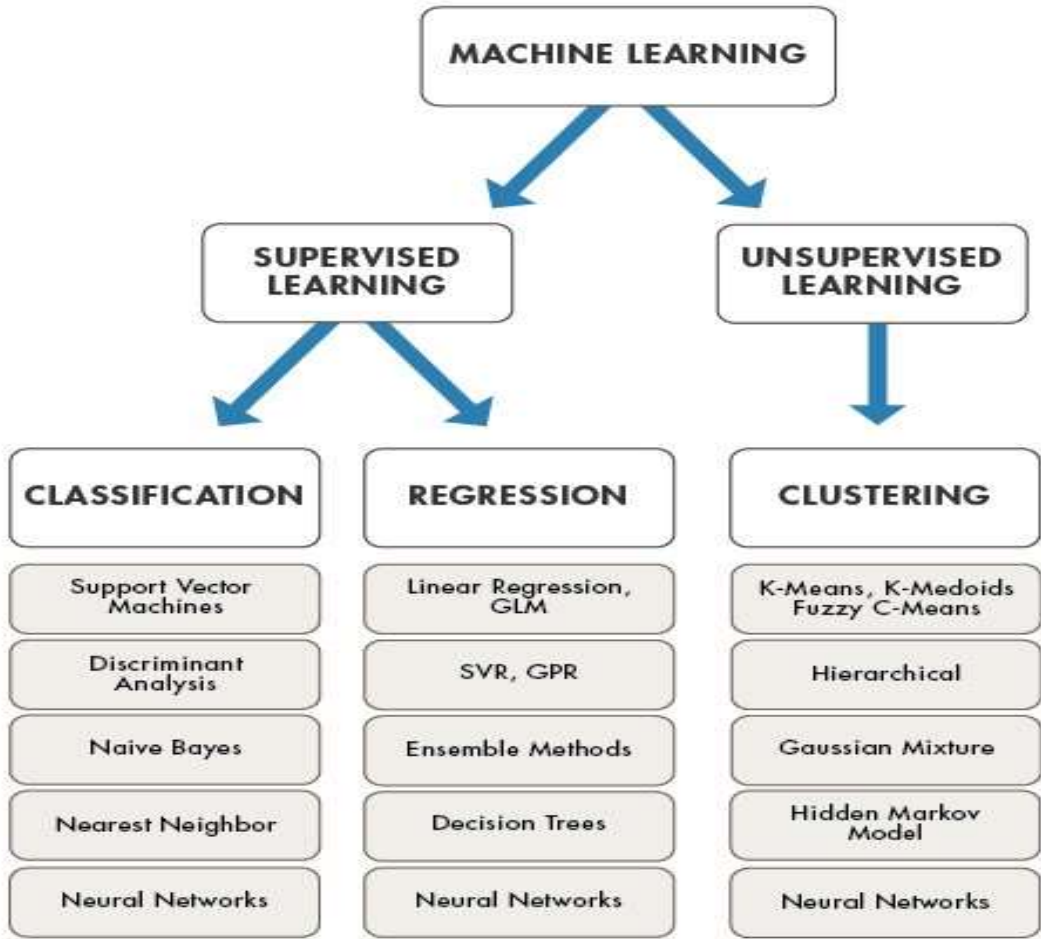


Methods used for clustering are:

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

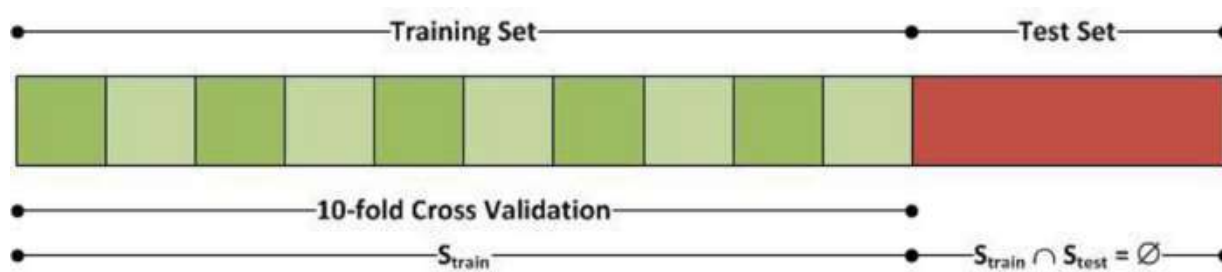
- **K-Means Clustering**
- **Spectral Clustering**

Overview of models under categories:



4. Training and testing the model on data

For training a model we initially split the model into 3 three sections which are ‘**Training data**’ , ‘**Validation data**’ and ‘**Testing data**’. You train the classifier using ‘**training data set**’, tune the parameters using ‘**validation set**’ and then test the performance of your 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 training the classifier. The test set will only be available during testing the classifier.

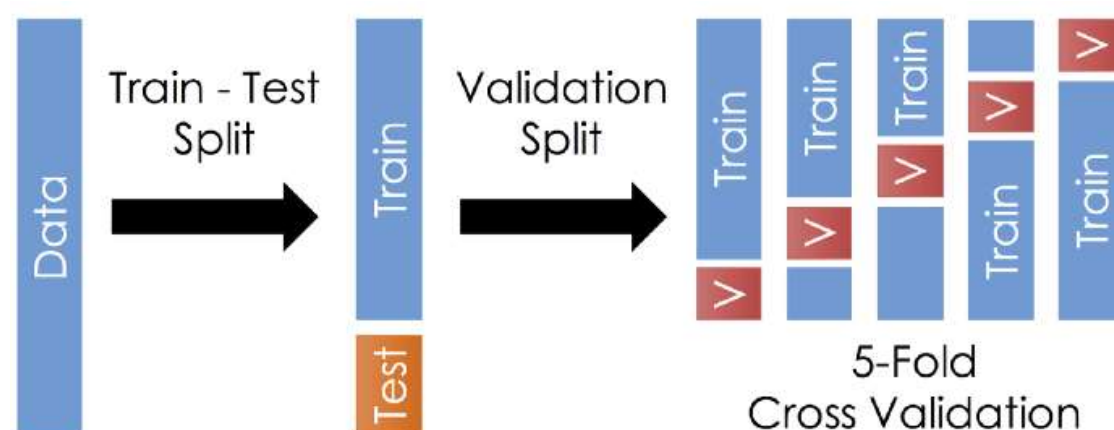


Training set:

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 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 a machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.



Once the data is divided into the 3 given segments we can start the training process.

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 we had chosen in step 3/ point 3. Once the model is trained we can use the same trained model to predict using the testing data i.e. the unseen data. Once this is done we can develop a

confusion matrix, this tells us how well our model is trained. A confusion matrix has 4 parameters, which are ‘**True positives**’, ‘**True Negatives**’, ‘**False Positives**’ and ‘**False Negative**’. We prefer that we 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

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

We can also 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:

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

5. 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 we might tune the hyper-parameters of the model and try to improve the accuracy and also looking at the confusion matrix to try to increase the number of true positives and true negatives.

Conclusion

In this blog, we have discussed the workflow a Machine learning project and gives us a basic idea of how a should the problem be tackled.

Chapter5

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, we used 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 you can use to explore your data in order to identify data cleaning operations you may want to perform

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, we removed the stop words such as is, the, 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, we removed URLs, hashtags, symbols, usernames, expressions, quotes, etc. Next, combine words are split by applying tokenizing techniques.
- Finally, we applied 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 meters, 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.
- In this research, we 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'. We have expected gun shoot data relates to crime scenes. However, the above post is related to a film shoot and photo shoot. So, we can consider it as a noisy post.
- To eliminate the noise and increase the accuracy, we implemented a machine learning-based filter in this stage. We have used the 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 we use supervised classification algorithms like linear regression to train the model on the cleaned dataset after dimensionality reduction.

Testing model:

- In this module we test 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 you test the program against should be stated in the technical specification document.
- Moreover, software testing has the power to point out all the defects and flaws during development. You don't want your clients to encounter bugs after the software is released and come to you waving their fists. Different kinds of testing allow us to catch bugs that are visible only during runtime.

Performance Evaluation

- In this module, we evaluate 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, we 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 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 you're using 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 already occurred was fraudulent.
- In that case, the transaction already happened, but you're making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

- In this module we use trained and optimized machine learning model to predict whether the crime accrue in social meadia

NLTK

- NLTK is a toolkit build for working with NLP in Python.
- It provides us 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 graph.
- It is an open source artificial intelligence library, using 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.

Support Vector Machine(SVM) algorithm

Introduction

Mastering machine learning algorithms isn't a myth at all. Most of the beginners start by learning regression. It is simple to learn and use, but does that solve our purpose? Of course not! Because you can do so much more than just Regression!

Think of machine learning algorithms as an armoury packed with axes, sword, blades, bow, dagger, etc. You have various tools, but you ought to learn 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 you've now mastered Random Forest, Naive [Bayes Algorithm](#) and Ensemble Modeling. If not, I'd suggest you take out a few minutes and read about them as well. In this article, I shall guide you through the basics to advanced knowledge of a crucial machine learning algorithm, support vector machines.

You can learn about Support Vector Machines in course format here (it's free!):

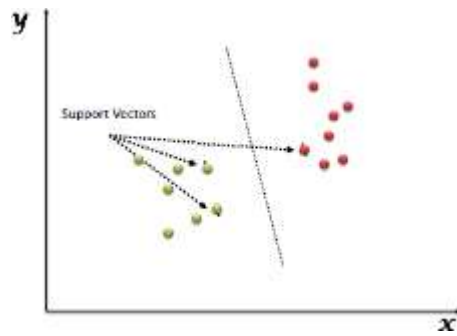
- Support Vector Machines (SVM) in Python and R

If you're a beginner looking to start your data science journey, you've come to the right place! Check out the below comprehensive courses, curated by industry experts, that we have created just for you:

- Introduction to Data Science
- Certified Program: Data Science for Beginners (with Interviews)

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



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).

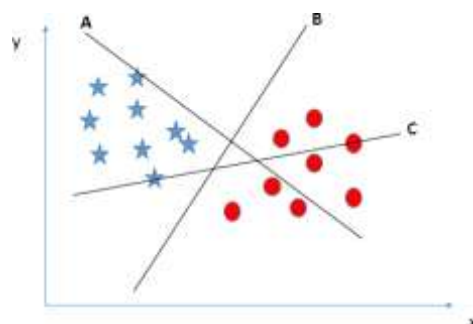
You can look at support vector machines and a few examples of its working here.

How does it work?

Above, we got accustomed to the process of segregating the two classes with a hyper-plane. Now the burning question is “How can we identify the right hyper-plane?”. Don’t worry, it’s not as hard as you think!

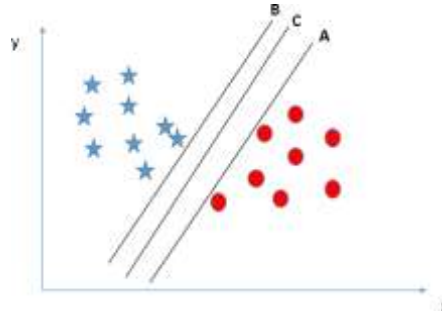
Let’s understand:

- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B and C). Now, identify the right hyper-plane to classify star and circle.

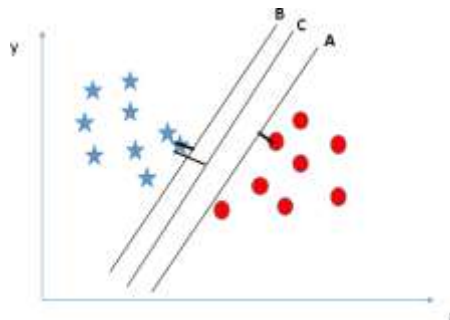


You need to remember a thumb rule to identify the right hyper-plane: “Select the hyper-plane which segregates the two classes better”. In this scenario, hyper-plane “B” has excellently performed this job.

- **Identify the right hyper-plane (Scenario-2):** Here, we have three hyper-planes (A, B and C) and all are segregating the classes well. Now, How can we identify the right hyper-plane?

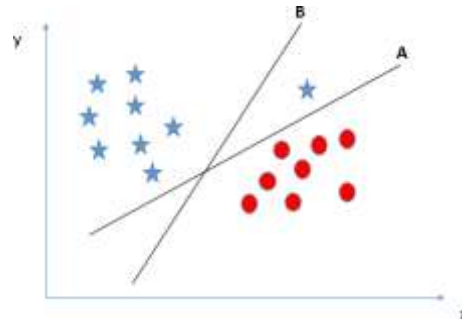


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look at the below snapshot:



Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss- classification.

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



Some of you 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 classified all correctly. Therefore, the right hyper-plane is **A**.

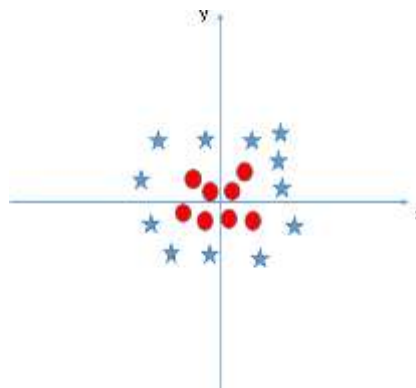
- **Can we classify two classes (Scenario-4)?:** Below, I am unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.



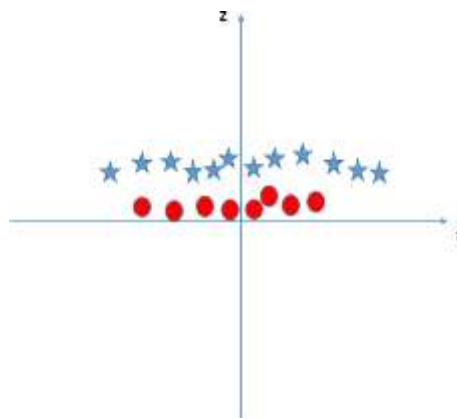
As I have already mentioned, one star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.



- **Find the hyper-plane to segregate to classes (Scenario-5):** In the scenario below, we can't have linear hyper-plane between the two classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyper-plane.



SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature $z = x^2 + y^2$. Now, let's plot the data points on axis x and z:

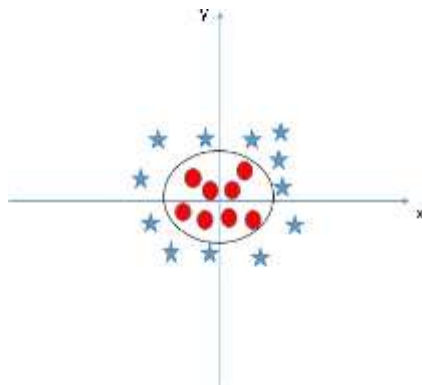


In above plot, points to consider are:

- All values for z would be positive always because z is the squared sum of both x and y
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to 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, should 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 low dimensional input space and transforms it to a higher dimensional space i.e. it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined.

When we look at the hyper-plane in original input space it looks like a circle:



Now, let's look at the methods to apply SVM classifier algorithm in a data science challenge.

Problem Statement

Dream Housing Finance company deals in all home loans. They have a presence across all urban, semi-urban and rural areas. A customer first applies for a home loan, after that the company validates the customer's eligibility for a loan.

Company wants to automate the loan eligibility process (real-time) based on customer details provided while filling an online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers' segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial data set.

Use the coding window below to predict the loan eligibility on the test set. **Try changing the hyperparameters for the linear SVM to improve the accuracy.**

Pros and Cons associated with SVM

• Pros:

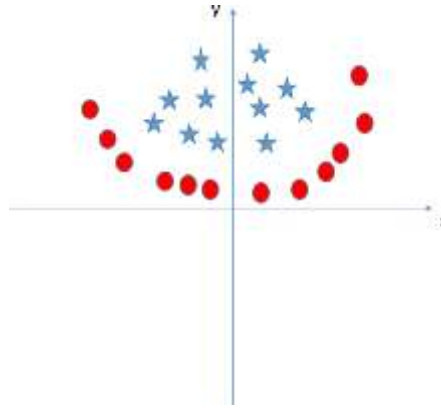
- It works really well with a clear margin of separation. It is effective in high dimensional spaces.
- It is effective in cases where the number of dimensions is greater than the number of samples.
- It uses a subset of training points in the decision function (called support vectors), so it is also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

• Cons:

- It doesn't perform well when we have large 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 classes are overlapping
- 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.

Practice Problem

Find right additional feature to have a hyper-plane for segregating the classes in below snapshot:



Answer the variable name in the comments section below. I'll shall then reveal the answer.

End Notes

In this article, we looked at the machine learning algorithm, Support Vector Machine in detail. I discussed its concept of working, process of implementation in python, the tricks to make the model efficient by tuning its parameters, Pros and Cons, and finally a problem to solve. I would suggest you to use SVM and analyse the power of this model by tuning the parameters. I also want to hear your experience with SVM, how have you tuned parameters to avoid over-fitting and reduce the training time

Random Forest Algorithm

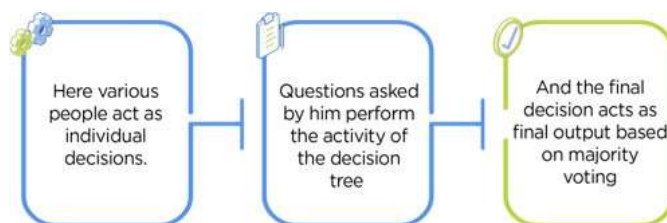
Introduction

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.

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.



Working of Random Forest Algorithm

Before understanding the working of the random forest we must look into the ensemble technique.

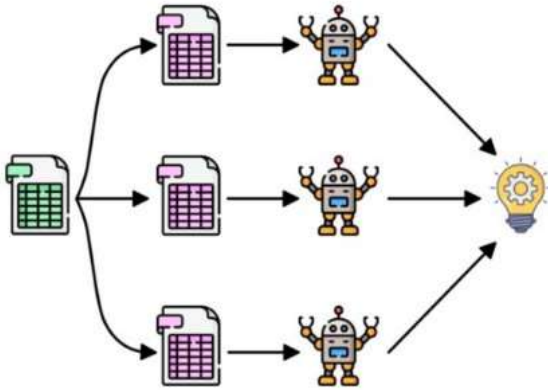
Ensemble simply means combining multiple models. Thus a collection of models is used to make predictions rather than an individual model.

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.

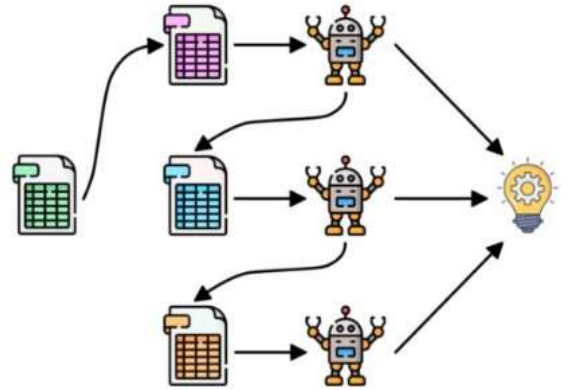
2. **Boosting**– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST

Bagging



Parallel

Boosting



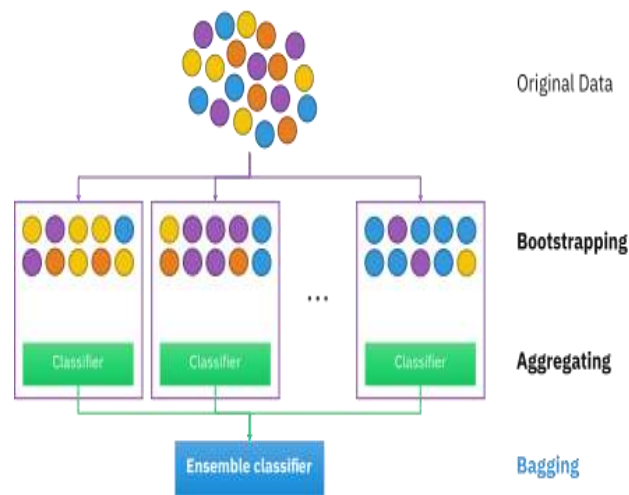
Sequential

As mentioned earlier, Random forest works on the Bagging principle. Now let's dive in and understand bagging in detail.

Bagging

Bagging, also known as **Bootstrap Aggregation** is the ensemble technique used by random forest. 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**.



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 model (Model 01, Model 02, and Model 03) obtained from this bootstrap sample is trained independently. Each model generates results as shown. Now Happy emoji is having 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 having 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. 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 you 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 we can make full use of the CPU to build random forests.
- 4. **Train-Test split**- In a random forest we 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.

Difference Between Decision Tree & Random Forest

Random forest is a collection of decision trees; still, there are a lot of differences in their behavior.

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower.
3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Thus random forests are much more successful than decision trees only if the trees are diverse and acceptable.

Important Hyperparameters

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

Use Cases

This algorithm is widely used in E-commerce, banking, medicine, the stock market, etc.

For example: In the Banking industry it can be used to find which customer will default on the loan.

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. We don't have to segregate data into train and test as there will always be 30% of the data which is not seen by the decision tree made out of bootstrap.

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 more 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.

Summary

Now, we can conclude that Random Forest is one of the best techniques with high performance which is widely used in various industries for its efficiency. It can handle binary, continuous, and categorical data.

Random forest is a great choice if anyone wants to build the model fast and efficiently as one of the best things about the random forest is it can handle missing values.

Overall, random forest is a fast, simple, flexible, and robust model with some limitations.

Chapter 6

Limitation & Future scope

- future work, we planned to evaluate other machine learning algorithms, such as neural networks and decision trees. We also plan to implement the crime prediction approach using SVM and deep neural network approach.
- Also, we intend to expand our analysis to include spatial and temporal analysis to find out when and where crime has spread in the past, and when and where it is most likely to spread in the future

CONCLUSION:

we proposed an approach to improving 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, we used SVM for the research.

