JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Methodical way of Describing and Analysing Categorization of Criminal Incidents.

[1]Prof.Shivanand Konade, [2]Prof. Arun Gawande, [3]Prof. Shradha Haeshal Chaudhari, [4]Dr. Karuna Kirwale, [5]Prof.Utkarsha Pawar
1,2,3,5 Assistant Professor, Electrical Engineering Department, Smt. Indira Gandhi College of Engineering. 4Assistant Professor in Amity Institute of Information Technology, Amity University Mumbai

Abstract
Analyzing and predicting crime involves a methodical process of identifying criminal activities. System capable dentify areaswith a high likelihood of crime and create a visual representation of these crimeprone areas.By utilizing the techniqu e of data mining, we are able to uncover valuable information that was previously unknown.Data without structure. New information is projected to be obtained by analyzing the current datasets.Crime is a prevalent and deceitful social issue that is experienced globally. Quality of life is compromised by crimes.The wellbeing, financial development, and standing of a country. In order to protect society from criminal activities,There is a requirement for more sophisticated systems and innovative methods to enhance crime analytics. Keeping their community safe

Keywords- Crime Analysis, data mining, clustering, prediction.

I. INTRODUCTION

What is Machine Learning ?

Machine learning is a set of computer algorithms that have the ability to improve themselves using examples, without requiring e xplicit programming by a human. Intelligence that involves computers learning from data without being explicitly programmed.U tilizing statistical tools to analyze data and predict outcomes for practicalapplications represents intelligent use of information.acti onable findingsThe breakthrough occurs when realizing that a machine has the ability to learn independently from the data provid ed (i.e., examples).generate precise outcomes. Machine learning is strongly connected to datamining and Bayesian prediction. modeling. The machine takes in data as input and then employs an algorithm to generate responses.

One common task in machine learning is to offer recommendations. If you have a Netflix account, allMovie or series suggestions are determined by analyzing the user's viewing history. Tech firms are utilizingImproving user experience through unsupervised l earning for personalized recommendations.Machine learning is utilized for various tasks such as fraud detection, predictive maint enance, and portfolio management.Optimizing processes, automating tasks, and similar activities.

Machine Learning vs. Traditional Programming

Machine learning is very different from traditional programming. In traditional programming, a programmer consults with an expert in the field for which software is being produced before coding every rule. Every rule has a logical basis, and the machine will produce an output in accordance with the coherent assertion. More rules must be written as the system gets more complicated. It can easily become too expensive to maintain. Machine learning is very different from traditional programming. In traditional programming, a programmer consults with an expert in the field for which software is being produced before coding every rule. Every rule has a logical basis, and the machine will produce an output in accordance with the coherent assertion. More rules must be written as the system gets more complicated. It can easily become too expensive to maintain.

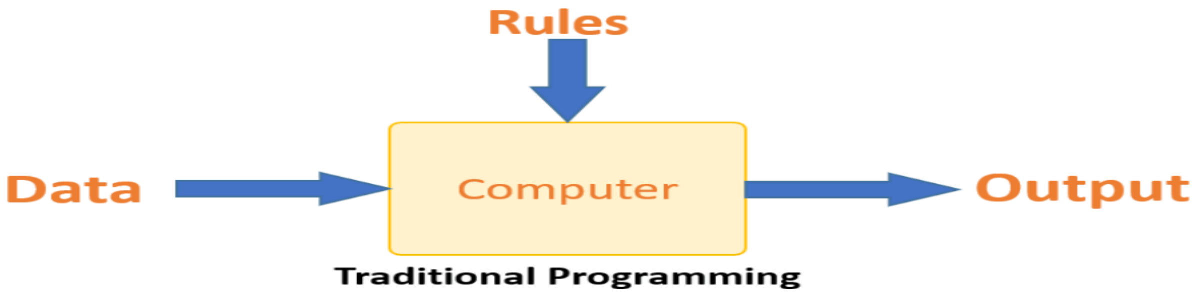

Figure.1 Traditional Programming

This problem is meant to be solved by machine learning. After determining the correlation between the input and output data, the machine creates a rule. Every time there is new data, the programmers don't have to write new rules. Over time, the algorithms' efficacy is enhanced by their ability to adjust to fresh data and experiences.

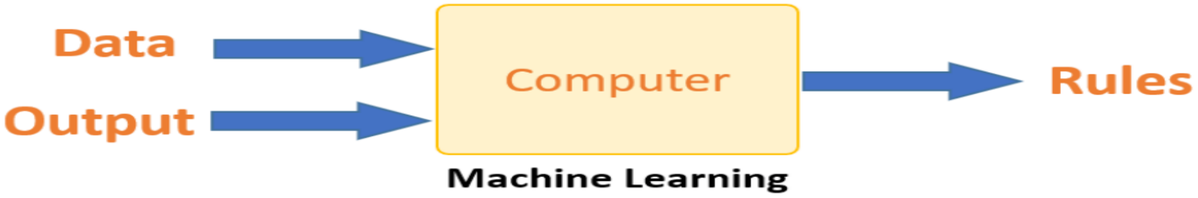

Figure.2 Machine Learning

How does Machine Learning Work?

The brain that does all the learning is machine learning. Machine learning is comparable to human learning. Experience teaches humans new things. Predicting becomes easier the more information we have. By analogy, our chances of success are lower in an unknown situation than in a recognized one. Computers receive the same training. The machine sees an example in order to produce an accurate prediction. The machine can determine the result when we offer it another example that is similar. But much like a human, the machine finds it difficult to forecast if it is fed an example that hasn't been seen before.Learning and inference are machine learning's main goals. Initially, the machine gains knowledge by identifying patterns. This finding is made The machine turns this discovery into a model by applying sophisticated algorithms to simplify reality. As a result, the data are described and condensed into a model during the learning step.

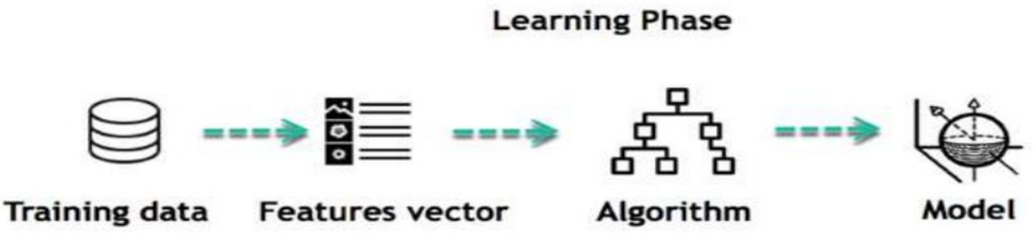

Figure.3 Learning Phase

For example, the machine is attempting to comprehend the correlation between an individual's income and the probability of dining at a fine dining establishment. It appears that the algorithm discovers a favourable correlation between earning a living and dining at upscale restaurants: The model is this one.

Inferring

Once the model is constructed, its effectiveness can be evaluated on previously unseen data. After being converted into a features vector, the new data are run through the model to produce a forecast. This is the most lovely aspect of it all. of artificial intelligence. Retraining the model or updating the rules are not necessary. The model that has been previously trained can be used to draw conclusions from fresh data.

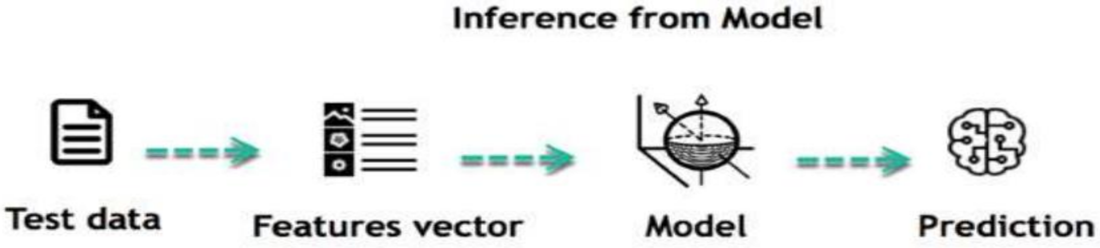

Figure.4 Inference from Model

The following points can be used to sum up the simple life of machine learning programs:
1. Describe the query.
2. Gather information
3. Display data visually
4. Algorithm training 5. Algorithm testing
6. Gather input 7. Iterate the algorithm
8. Repetition 4–7 until desired outcomes are obtained
9. Create a prediction using the model.
The algorithm applies its acquired skill at making the correct decisions to new sets of data as it becomes proficient at doing so.

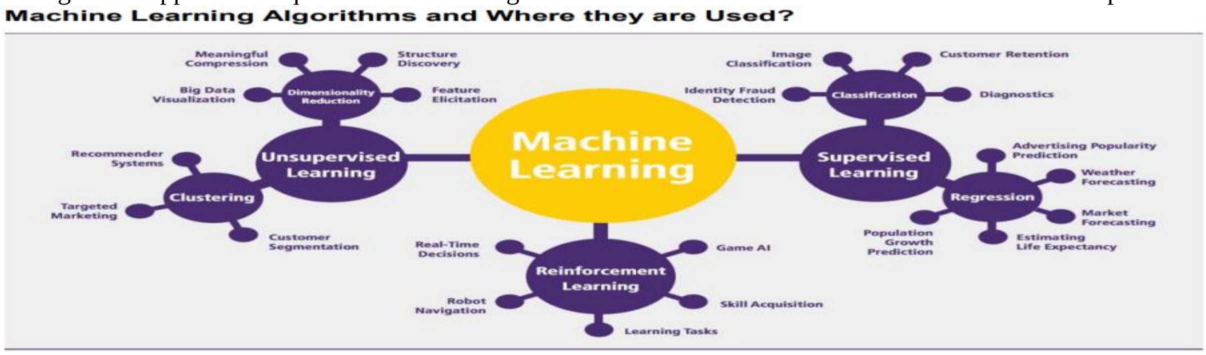

Figure.5 Machine learning Algorithms

Supervised and Unsupervised are the two main categories into which machine learning may be divided. Numerous alternative algorithms exist.

supervised instruction

An algorithm learns the link between given inputs and provided outputs by using human feedback and training data. For example, a practitioner can input weather forecast and marketing expense data to forecast the can sales. When the output data is known, supervised learning can be used. New data will be predicted by the program. Two types of supervised learning exist: Tasks for classification and regression

Grouping

Let's say you wish to determine a customer's gender for an ad. You will begin compiling information from your customer database on height, weight, occupation, pay, shopping basket, etc. You are aware of the gender of There can only be one gender per customer: male or female. Assigning a probability of being male or female (i.e., the label) based on the data (i.e., features you have gathered) is the classifier's goal. Once the program has figured out whether something is male or female, you may use fresh data to predict. For example, you would like to know if the new information you received from an unidentified consumer is male or female. If male is predicted by the classifier30% of the customers are female, while 70% are male. Two or more classes may be included in the label. While the machine learning example above only has two classes, each object in the world has thousands of classes (e.g., glass, table, shoes, etc.) that a classifier must predict. symbolizes a class.

II. Experimental or Methods and Materials: Algorithm used

Description of Modules:

Data Collection:

Gathering data is the first concrete step in creating a machine learning model. This is a crucial stage that will eventually affect how excellent the model is; the more and better data we collect, the more effective our model will be. Data can be gathered using a variety of methods, including physical interventions, web scraping, and more.

Dataset:

There are 520 distinct pieces of data in the dataset. The dataset consists of 23 columns, each of which is explained below.
1. ID: The record's unique identification.
2. Case Number: The incident-specific Records Division Number (RD Number) of the Chicago Police Department.
3. Date: The exact day the incident happened.
4. Block: the location of the incident
5. Illinois Uniform Crime Reporting Code (IUCR).
6. Primary Type: The IUCR code's principal description.
7. Description: A subcategory of the primary description, this is the secondary description of the IUCR code.
8. Location Description: Describe the area where the event took place.
9. Arrest: Says if someone has been taken into custody.
10. Domestic: This indicates whether the Illinois Domestic Violence Act's definition of "domestic" applies to the incident.
11. Beat: The beat on which the incident took place is indicated. The smallest police geographic region is called a beat. There is a police beat car assigned to each beat.
12. District: The incident's police district is indicated here.
13. Ward: The incident's ward (a district of the City Council).
14. Community Area: Identifies the locality in which the incident took place. There are 77 community areas in Chicago.
15. FBI Code: This is the National Incident-Based Reporting System (NIBRS) criminal classification as specified by the FBI.
16. X Coordinate: In the State Plane Illinois East NAD 1983 projection, the x coordinate of the incident's location.
17. Y Coordinate: In the State Plane Illinois East NAD 1983 projection, the y coordinate of the incident's location.
18. Year: The year the event took place.
19. Last Updated On: The record was last updated at this time.
20. Latitude: The latitude at which the incident took place. Although it is on the same block, this position has been moved from the actual location for partial redaction.
21. Longitude: The location of the incident, measured in longitude. This site has been moved from the real place for a temporary

Gather and organize data

In order to get it ready for training. Clean up anything that could need it (get rid of duplicates, fix mistakes, handle missing values, normalize, convert data types, etc.). Data should be randomized to eliminate the impact of the specific sequence in which we gathered and/or otherwise processed the data. Use data visualization to carry out additional exploratory analysis or to identify pertinent correlations between variables or class imbalances (bias alert!). Divided into sets for training and assessment

Model Selection:

We applied the Random Forest Classifier machine learning technique after finding that it produced an 80.7% accuracy on the test set.

The Algorithm of Random Forest

Let's take a more general look at the algorithm. Let's say you want to take a trip and you want to go somewhere you will enjoy yourself. What steps do you take then to locate a location you will enjoy? You have a few options: ask your friends, search online, and read reviews on travel blogs and portals. Assume for the moment that you have chosen to ask your friends, and that you have discussed with them their prior experiences traveling to different locations. Every friend will give you some recommendations. You now need to compile a list of those suggested locations. Next, you ask them to choose one location from your list of suggested locations to vote for or choose as the trip's best site.

Two steps make up the decision-making process described above. First, find out which destination your friends would recommend from a list of countries they have visited by asking them about their own travel experiences. Utilizing the decision tree algorithm is similar to this section. Here, each friend lists the locations they have already been to. The voting process to choose which recommendation on the list should be placed highest comes in the second section, which is the collection of all the recommendations. The random forests algorithm refers to the entire procedure of soliciting recommendations from friends and using a vote system to determine which location is best. Technically, it is an ensemble method of decision trees created on a randomly split dataset, based on the divide-and-conquer strategy. This assortment of decision trees

How is the algorithm implemented?

There are four steps to its operation:

Choose arbitrary samples from the provided dataset. Create a decision tree for every sample, then use each decision tree to derive a prediction. Cast a vote for each anticipated outcome. Select the prediction result with the most votes as the final prediction.

Advantages:

Because so many decision trees are involved in the process, random forests are thought to be a very resilient and reliable method. It is not affected by the overfitting issue. The primary reason is that the biases are eliminated by averaging all of the forecasts. The approach is applicable to issues involving both regression and classification. Missing values can also be handled using random forests. There are two approaches to deal with these: calculating the proximity-weighted average of missing data and substituting continuous variables with median values. Obtaining the relative feature importance facilitates the process of choosing the features that contribute the most to the classifier.

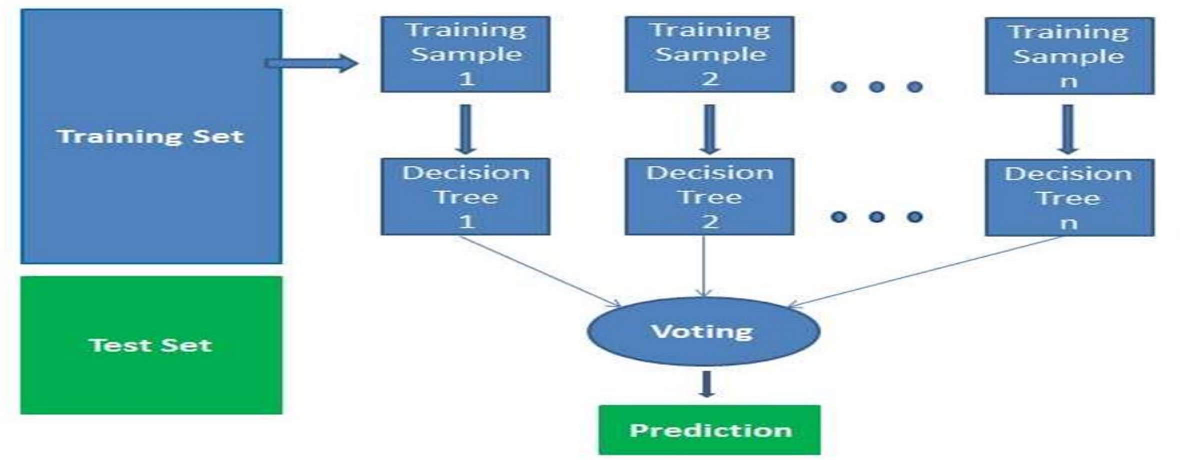

Figure.6 Machine learning prediction

Disadvantages:

1. Because random forests have several decision trees, they generate forecasts slowly. Every time it makes a forecast, every tree in the forest must predict the same thing, given the identical input, and then cast a vote on it.
2. The entire procedure takes a lot of time. In contrast to a decision tree, which makes decision-making simple by following the path within the tree, the model is more challenging to interpret.

Finding important features

Additionally, random forests provide a useful feature selection metric. An additional variable that Scikit-learn includes with the model indicates the relative significance or input of each feature to the prediction. The relevance score of every feature in the training phase is automatically calculated. Next, it climbs the decreasing importance such that all scores add up to one. When developing a model, this score will assist you in selecting the most crucial characteristics and eliminating the less crucial ones. The significance of each feature is determined using Random Forest using gini relevance or mean reduction in impurity (MDI). The entire decrease in node impurity is another name for Gini significance. This represents the reduction in accuracy or model fit that occurs when a variable is removed. The greater the

Random Forests vs Decision Trees

A group of several decision trees is called a random forest. Overfitting is a potential problem for deep decision trees, but random forests guard against it by building trees on arbitrary subsets. The computing speed of decision trees is higher. A decision tree is simply interpreted and may be turned into rules, whereas random forests are more difficult to understand.

Analyze and Prediction:

Only eight features were selected from the actual dataset:

1. Year: The year the incident took place.

2. Month: The month the incident happened.

3. Day: The day the incident took place.

4. Day of Week: The day on which the incident took place.

5. Minute: The exact moment the incident happened.

6. Second: the moment the incident happened.

7. Latitude: The latitude at where the incident took place. Although it is on the same block, this position has been moved from the actual location for partial redaction.

8 Longitude: This is the longitude of the incident's location. Although it is on the same block, this location is different from the one used for partial redaction.

SYSTEM ANALYSIS

EXISTING SYSTEM:

1. During pre-work, duplicate values and features are eliminated from the dataset that was gathered from the public domain.
2. Decision trees have been utilized to extract features from vast amounts of data and to identify patterns in crime. It offers the fundamental framework needed for additional classification procedures.
3. Deep Neural Network is used to extract features related to classified crime patterns. The performance is computed for both trained and test values based on the prediction. Crime prediction aids in anticipating future instances of criminal activity and enables law enforcement to take swift action to put an end to it.

DISADVANTAGES OF EXISTING SYSTEM:

1. Because the classifier employs categorical values, which lead to a biased result for the nominal qualities with higher values, the previous works explain the low accuracy.
2. The classification methods are not appropriate for areas with inaccurate data and genuine values.
3. Characteristics. As a result, an ideal value must be assigned since the classifier's value needs to be adjusted.

PROPOSED SYSTEM:

1. To remove repetitive and unnecessary data values, the acquired data is first pre-processed using a machine learning approach called a filter and wrapper. Additionally, it lowers the dimensionality, cleaning the data. After that, the data goes through one more dividing procedure. It is divided into test and trained dataset.
2. The training and testing datasets are used to train the model. After that comes mapping. To make classification easier, the crime type, year, month, time, date, and place are mapped to an integer. The Random Forest Classifier is first used to analyze the independent influence between the attributes.
3. The criminal aspects are labeled, making it possible to analyze the incidence of crime at a specific time and place. At last, the most common crimes are discovered, together with temporal and spatial details. The accuracy rate is used to determine the prediction model's performance. Python was the language used to create the prediction model, which is based on machine learning and data analysis.

ADVANTAGES OF PROPOSED SYSTEM:

1. Since the majority of the included properties rely on the time and location, the proposed method is highly suited for the detection of crime patterns.
2. It also solves the issue of the attributes' independent effect analysis.
3. Since the best value takes into consideration both nominal and actual values, as well as the region with inadequate information, there is no need to initialize it.
4. Comparing the accuracy to other machine learning prediction models, it has been comparatively high.

SYSTEM ARCHITECTURE

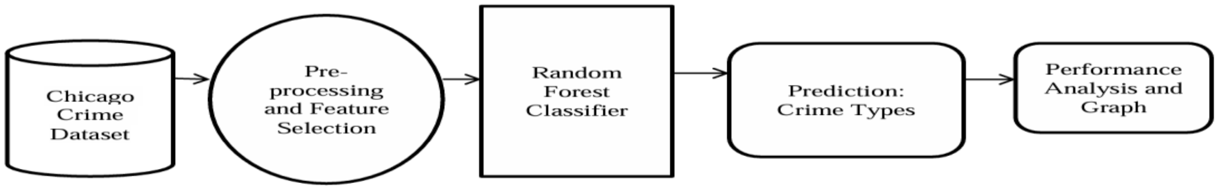

Figure.7 System Architecture

DATA FLOW DIAGRAM:

1. The bubble chart is another name for the DFD. It is a straightforward graphical formalism that may be used to depict a system in terms of the data that the system receives as input, the data that the system processes in different ways, and the data that the system generates as output.
2. One of the most crucial modeling tools is the data flow diagram (DFD). The components of the system are modeled using it. These elements consist of the system's procedure, the data it uses, an outside party that communicates with it, and the information flows it uses.
3. DFD illustrates the flow of information through the system and the various changes that alter it. It's a visual method for representing information flow.

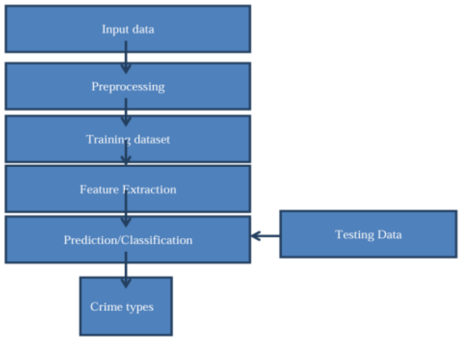

Figure.8 Flow Chart

UML Schemas

Unified Modeling Language is known as UML. An industry-standard general-purpose modeling language used in object-oriented software engineering is called UML. The Object Management Group both developed and oversees the standard. The intention is for UML to spread as a standard language for modeling object-oriented computer programs. The two main parts of UML as it exists now are a notation and a meta-model. In the future, UML may also include other processes or methods that are connected to it. A common language for business modeling and other non-software systems, as well as for defining, visualizing, building, and documenting software system artifacts, is called the Unified Modeling Language. A collection of the best engineering practices is represented by the UML.

GOALS:

The following are the main objectives for the UML design:
1. Give users access to an expressive, ready-to-use visual modeling language so they can create and trade insightful models.
2. Offer methods for specialization and extendibility to expand the fundamental ideas.
3. Be unaffected by specific development processes and programming languages.
4. Offer an official foundation for comprehending the modeling language.
5. Promote the market expansion for OO tools.
6. Encourage the use of higher level development ideas like components, frameworks, partnerships, and patterns.
7. Include industry best practices.

USE CASE DIAGRAM:

A use case diagram is a kind of behavioral diagram specified by and generated from a use-case analysis in the Unified Modeling Language (UML). Its objective is to provide a visual summary of the features. supplied by a system with respect to its actors, their objectives (depicted as use cases), and any interdependencies among those use cases. A use case diagram's primary goal is to display which actors are served by which system functionalities. It is possible to illustrate the roles of the system's actors.

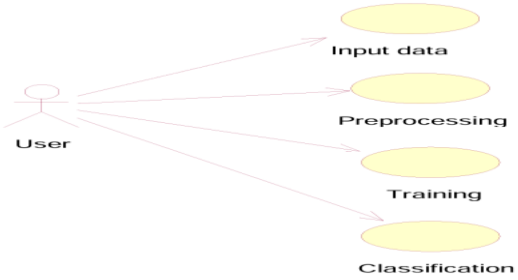

Figure.9 Case Diagram

CLASS DIAGRAM:

A class diagram in software engineering is a kind of static structural diagram that displays the classes, attributes, and other components of a system using the Unified Modeling Language (UML). procedures (or techniques), as well as the connections between the classes. It indicates which class has the data.

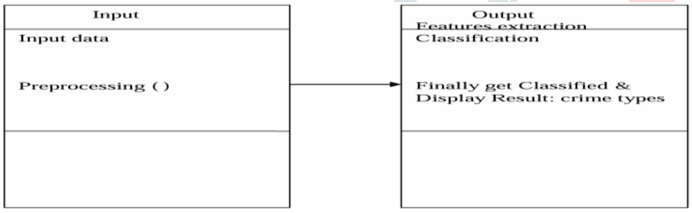

Figure.10 Class Diagram

Sequence diagram:

Extraction of features Grouping : Get Classified at Last & Show Results: Crime Types In the Unified Modeling Language (UML), a sequence diagram is a type of interaction diagram that illustrates the relationships and sequence in which processes operate with one another. It is a Message Sequence Chart construct. Event diagrams, event situations, and timing diagrams are other names for sequence diagrams.

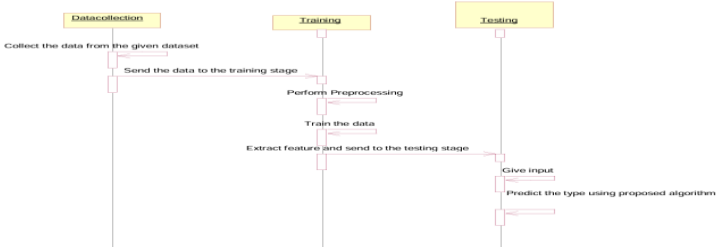
Datacollection
Training
Testing
Collect the data from the given dataset
Send the data to the training stage
Perform Preprocessing
Train the data
Extract feature and send to the testing stage
Give input
Predict the type using proposed algorithm

Figure.11 Sequence Diagram

ACTIVITY DIAGRAM:

With support for choice, iteration, and concurrency, activity diagrams are graphical depictions of workflows consisting of sequential activities and actions. Activity diagrams in the Unified Modeling Language can be utilized to elucidate the sequential business and operational processes of the system's constituent parts. An activity diagram displays the total control flow.

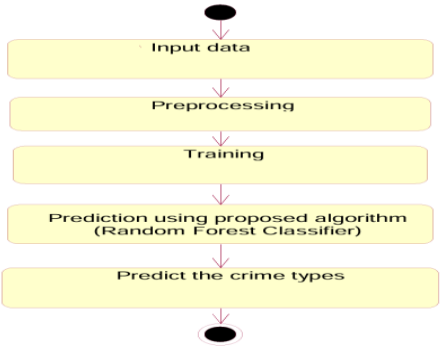
Input data
Preprocessing
Training
Prediction using proposed algorithm
(Random Forest Classifier)
Predict the crime types

Figure.12 Activity Diagram

Conclusion

In this work, two classifiers, such as Multinominal NB and Gaussian NB, are used to tackle the challenge of handling the nominal distribution and real valued characteristics. It is ideal for real-time forecasts and doesn't require a lot of training time. It also solves the issue of having to deal with continuous target set of variables that the previous research was unable to fit into. Thus, Random Forest Classification could be used to forecast and identify the crimes that occur most frequently. A few common metrics are also used to compute the algorithm's performance. The measures that are primarily used in algorithm evaluation include average precision, recall, F1 score, and accuracy. By using machine learning methods, the accuracy value might be improved significantly.

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).