Proceedings of the International Conference on Sustainable Computing and Data Communication Systems (ICSCDS-2023)
DVD Part Number: CFP23AZ5-DVD; ISBN: 978-1-6654-5579-4

PREDICTION OF CRIME DATA USING MACHINE LEARNING TECHNIQUES

Dr.P.Varaprasada Rao
Department of CSE
Gokaraju Rangaraju Institute of Engineering & Technology
Hyderabad, India
prasadp.griet@gmail.com

Sukesh Sunkari
Department of CSE
Gokaraju Rangaraju Institute of Engineering & Technology.
Hyderabad, India
sunkarisukesh@gmail.com

Tejeshwararao Raghumandala
Department of CSE
Gokaraju Rangaraju Institute of Engineering & Technology.
Hyderabad, India
tejeshwararao21@gmail.com

Ganesh Koka
Department of CSE
Gokaraju Rangaraju Institute of Engineering & Technology.
Hyderabad, India
ganesh.koka159@gmail.com

Deepak Chowdary Rayankula
Department of CSE
Gokaraju Rangaraju Institute of Engineering & Technology.
Hyderabad, India
deepakchowdaryrdc@gmail.com

Abstract— **Crime has a devastating effect on society and can take many forms, ranging from violent acts such as murder and assault to less serious but still damaging offenses like burglary and vandalism. Crime records are records of various criminal activities in a given area. These records typically include information such as the date, time, and location of reported crime, the nature of the crime, and the name of the suspect. However, there are many problems associated with keeping accurate and up-to-date records of criminal activity. The most common issue is a lack of uniformity in the way different states and municipalities track and record criminal activity. Additionally, many states and municipalities lack the resources to properly maintain accurate records, resulting in incomplete or inaccurate information. Finally, maintaining crime records is time consuming and expensive, leading to a backlog in the processing of reported crimes. This can complicate investigations and create a backlog of cases that need to be addressed.**

This model's objective is to make criminal investigation systems more effective. The two main facets of this endeavour are crime analysis and prediction. Based on these aspects the datasets which contain all the data is pre processed using different pre processing techniques. Further the dataset is processed using machine learning algorithms like SVM, Random forest classifier, Decision tree, K-Means. In conclusion, crime data analysis and prediction can provide valuable insights into criminal activity and aid in the development of effective crime prevention strategies.

Keywords—SVM, Random Forest Classifier, Decision Tree.

I. INTRODUCTION

Crime data analysis is the study of criminal behavior by examining data collected from various sources such as police reports, court records, and surveys. This data is used to examine patterns of criminal activity, identify possible causes of the crime, and develop strategies for preventing future criminal activity. By analyzing crime data, researchers can better understand the root causes of crime and develop evidence-based solutions to reduce crime in communities. Additionally, crime data analysis can help law enforcement officials identify trends in criminal activity and target resources to areas where they are most needed.

Information separation is a key component of AI. Prescient research, also known as quantifiable learning, is the study area at the intersection of insights, artificial intelligence, and software engineering. In recent years, the use of AI technology has permeated every aspect of daily life. Numerous cutting-edge websites and devices contain AI calculations at its core, from programmed suggestions of which movies to view, to what meals to request or which products to buy, to personalised web-based radio and recognising your friends in your photos. When you look at a confusing website like Facebook, Amazon, or Netflix, every part of the site certainly uses a separate AI model. Beyond applications for businesses, The way information-driven research is conducted nowadays has been impacted by AI. The tools described in this book have been used for a variety of logical tasks, including grasping stars, locating distant planets, discovering new particles, evaluating DNA groupings, and administering specialised malignant growth medications. To benefit from AI, your application does not need to have the same breadth or potential to change the world as these guides. In this section, we'll explain why artificial intelligence (AI) has gained such notoriety and look at the kinds of problems it can solve. Then, we'll explain how to put together your most memorable AI model while outlining key concepts.

II. OBJECTIVE

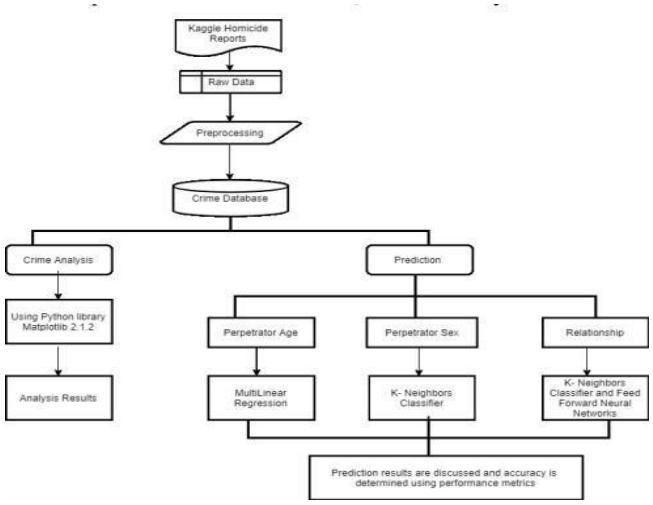
A. Outcome

Our research's main goal is to categorize algorithms according to how accurate they are. From the evidence present at the crime scene, one can determine whether the crime is real and even foresee its character. The investigation department will be able to resolve cases more quickly as a result of employing this methodology.

B. Block Diagram

A block diagram is a picture that shows the parts of a system and how they relate to one another, such as in a machine learning project. Project managers, engineers, and data scientists, among other stakeholders, can utilise it to explain the system's architecture and organisational structure. The blocks in the diagram stand in for several system parts, including model selection, feature engineering, data pre-processing, and evaluation. The direction of information or data flow between the blocks is indicated by arrows. The block diagram gives a summary of the system's architecture and can be used to spot any possible problems or process bottlenecks.

Block diagrams are distinctive visual representations that help explain the structure and design of a system by decomposing it into parts, such as data pre-processing, model selection, and evaluation. They are particularly useful in machine learning projects. They give a broad perspective of the architecture of the system and can be used to spot possible problems and bottlenecks. Project managers, engineers, and data scientists can use them to comprehend the system's architecture and make sure the project is successful.



III. PROPOSED SYSTEM

A. *Pre Processing*

To achieve the pre processing technique we are just following the basic pre processing techniques like once we have the dataset, the dataset is sent to the data cleaning, data slicing, feature selection and after data has been processed then the algorithms are applied to the dataset.

Preprocessing is a crucial phase in the drafting of reports, especially when the reports are based on data analysis or contain statistical analysis. Data must be cleaned and prepared for preprocessing in order to be in an analysis-ready format. This phase is important because data inaccuracies or inconsistencies can have a big influence on how accurate the results are.

Some common preprocessing techniques include removing duplicates, handling missing values, and transforming the data to meet the assumptions of the statistical methods being used.

Data Cleaning

```
[ ] df=df.rename(columns = {'Êcommunityname':'Community Name'})
    df = df.replace('?', '0')
    df.head()
```

Data Slicing

```
[ ] df1 = df.iloc[:200]
    df1.head(200)
```

Feature Selection for Clustering Algorithms

```
features = ['householdsize', 'racepctblack']
X = df1[features].values
y = df1['violent_crime_occurence'].astype(float).values
```

B. *Algorithms*

- **Support Vector Machine-** The popular machine learning method known as support vector machine, or SVM for short, is used for applications such as classification and regression. It is a reliable and flexible approach that may be used to a wide range of problems, including text classification and image classification. SVM is a reliable and well-liked machine learning technique that may be used to many different problems. It is praised for handling highly dimensional data and having a high rate of generalisation to fresh input.
- **Decision Tree Classifier-** A decision tree is a machine learning algorithm used for classification and prediction tasks. It works by creating a tree-like structure that breaks down a dataset into smaller subsets based on the values of different features. At each node of the tree, a decision is made about which feature to split on based on its ability to separate the data into different classes. This process continues recursively until each leaf node represents a single class label. Decision trees are easy to interpret
- **Random Forest Classifier-** Consider yourself attempting to make a significant decision and you don't want to depend just on one person's viewpoint. As an alternative, you opt to get a group of people's opinions before selecting a choice based on the poll's findings.

  The Random Forest Classifier operates in a similar manner. To create a more accurate model, several decision trees are integrated using the ensemble learning method. Each decision tree in the forest is trained on a random subset of data and attributes, similar to polling a group of individuals from diverse walks of life for their thoughts.

  By combining the predictions of all the decision trees in the forest, Random Forest is able to generate forecasts that are more accurate than any individual tree in the forest. It's similar to accepting the choice made by the majority of the people you surveyed.

C. Performance Metrics

- **r2_score-** An efficiency metric used to evaluate the quality of fit of a linear regression model is the coefficient of determination, often known as the R2 score. It determines the proportion of the dependent variable's variance that the model's independent variable or variables can explain. The R2 score is a value between 0 and 1, with 1 signifying the best possible match and 0 signifying no fit. When R2 is negative, the model does worse than the simple mean of the dependent variable.

  The independent variable(s) in the model may be responsible for 50% of the variation in the dependent variable if R2 is equal to 0. When the independent variable or variables in the model have an R2 value of 0.8, they are said to be responsible for 80 percent of the variation in the dependent variable.

  To completely assess the model's performance, it is always important to incorporate additional performance metrics and visualisations, such as the residual plots.

- **Accuracy score-**The most common performance metric used to assess the effectiveness of an ANN model and a multilane classification model is the accuracy score function. The subset accuracy is returned by this metric. The effectiveness of a classification model is assessed using its accuracy as a performance metric in machine learning. A proportion of all predictions is used to determine the model's accuracy.

  Although accuracy is a performance indicator that machine learning systems frequently use, it may not necessarily be the best one, particularly if the classes are unequal. For instance, a model that simply forecasts negative for all observations would have great accuracy but be useless if the positive class was far less than the negative class. Other performance metrics, such as recall, accuracy, F1 score, or AUC-ROC, may be more appropriate in some circumstances

D. Analysis

The dataset we used includes 2000 items for crimes that occurred between 1980 and 2014. In the Analysis step, x is analysed and determined. Unsolved crimes as a percentage x Tools of crime in unsolved cases The month with the most unresolved crimes is x

E. Literature review

The details were presented as an instrument used to describe and analyse questionable exercises is wrongdoing research. If the research conducted thus far should appear to be more explicitly valuable, it is typically because it shows which suspect types are useful in reducing wrongdoing, so for the most part they would be places where severe violations are reduced. Given that each area can be broken out by methodology and that data is acquired for each cycle to be examined, it is an excellent tool for assessing the crime rate. Due to the rapid development of data technology, fraud examiners will need to continue updating the examinations and assisting them with interpreting the evidence on the example bunching and pre-processing to obtain unstructured evidence, and then search for breaches within it. In this way, individuals who have been previously investigated and subsequently apprehended or identified as having engaged in a comparable suspicious behaviour may frequently be examined for examples like suspect history or occurrence reports rather than merely offences themselves. This is generally supposed to direct law enforcement to potential hotspots for criminal activity without giving it any attention as to who is actually responsible. Bayesian classifiers were used as the ongoing plan was being used set up in the ongoing approach.

F. Merits and demerits

- **Merit-** Identifies high-risk individuals or areas, allowing law enforcement to focus resources where they are most needed.

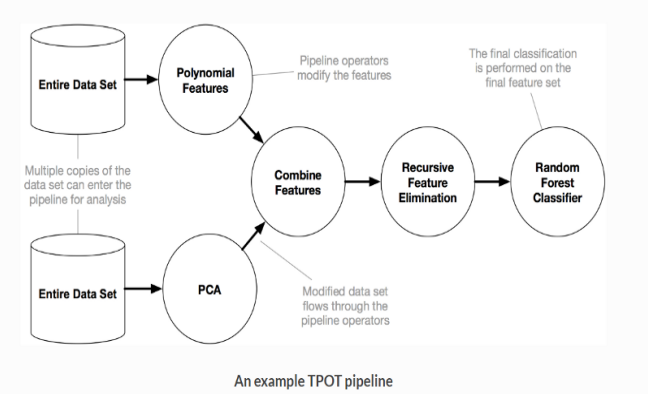
  Improving the effectiveness of investigations by giving officers access to a broader array of integrated, actionable intelligence needed to solve crimes.

  Technology can enhance the efficiency of incident investigations and help increase clearance rates.

- **Demerit-** The logic required to make a decision is specific to a single domain and task. Changing the task even slightly might require a rewrite of the whole system.

IV. Methadology

A. Flowchart



An example TPOT pipeline

B. Algorithm

Step 1: BEGIN

Step 2: Analyze the crime scene

Step 3: Collect the evidences from the crime scene and

gives them as input to the model to predict

Step 4: Perform multiple algorithms to check over the evidences relation from the old records

Step 5: Using the evidences the model generates the graphical representation of the Histograms

Step 6: Predicting the crime scene according to the inputs given to the model

Step 7: Exit.

C. *Graphical representation*

a) Positioning Figures and Tables: Figures and tables should go at the top and bottom of each column. Keep them away from the centre of columns. Tables and figures with a lot of data may fill both columns. Table heads should appear above the tables, and figure captions should be below the figures. After they are referenced in the text, include the figures and tables. Even at the beginning of a phrase.

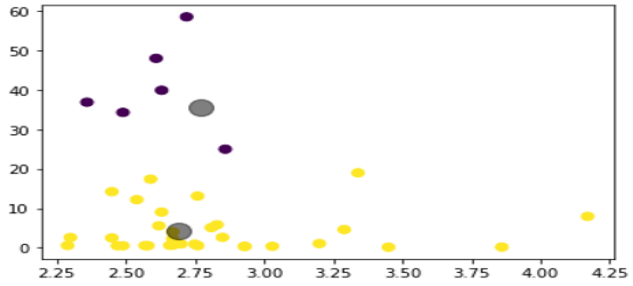


Fig. K-means

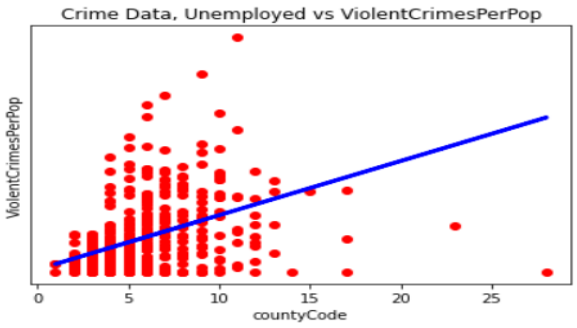


Fig. Crime data Unemployed vs Violent crimes

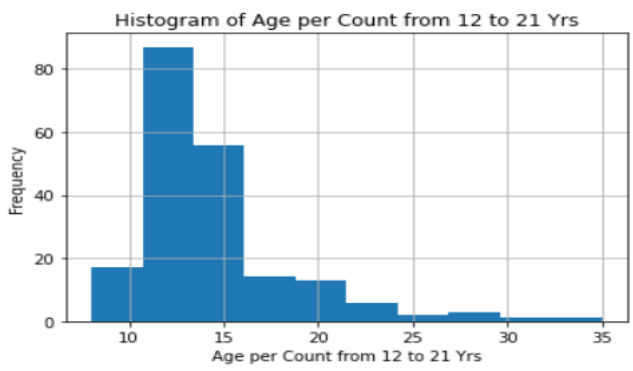


Fig. Histogram of age

D. *Dataset*

The Crime dataset is a collection of data from various sources that provide information on criminal activity. It includes details such as the type of crime, the location of the crime, the date and time of the incident, the identity of the perpetrator, and the outcome of the incident. It is often used by law enforcement agencies, researchers, and policy makers to better understand the prevalence and trends of criminal activity. The dataset can also be used to inform strategies for reducing crime and improving public safety.

Together with records on the police department and the officers participating in investigations and responding to crimes, these databases could also contain information on charges, convictions, and punishments.

Law enforcement organizations, decision-makers, and scholars frequently utilize crime databases to comprehend patterns and trends in criminality, pinpoint crime-prevention hotspots, and create public safety policies. Data analysts and machine learning experts may potentially utilize these resources to create forecasting model for information extraction and mitigation.

	crimeOccurence	Écommunityname	state	countyCode	communityCode	\
0	1	BerkeleyHeightstownship	NJ	39	5320	
1	1	Marpletownship	PA	45	47616	
2	1	Tigardcity	OR	?	?	
3	1	Gloversvillecity	NY	35	29443	
4	1	Bemidjicity	MN	7	5068	
...	...	...	...	...	...	
2210	1	Mercedcity	CA	?	?	
2211	1	Pinevillecity	LA	?	?	
2212	1	Yucaipacity	CA	?	?	
2213	1	Beevillecity	TX	?	?	
2214	1	WestSacramentocity	CA	?	?	

E. *Principal Component Analysis(PCA)*

PCA (Principal Component Analysis) is a technique used in machine learning to reduce the dimensionality of the dataset without losing much information. In other words, it helps in identifying the most important features or variables that contribute the most to the dataset's variance. By reducing the dataset's dimensions, it reduces the computational complexity of the algorithms and helps in avoiding the curse of dimensionality. PCA works by transforming the original features into a new set of uncorrelated features called principal components. These components are ranked based on their importance, and the least important ones can be discarded. This helps in reducing the dataset's dimensionality and can improve the performance of machine learning algorithms. Yet it's important to keep in mind that not all machine learning projects require or profit from PCA. It ought to be utilized rarely and only after carefully assessing the project's goals and the dataset's characteristics.

PCA

```
[ ] from sklearn.model_selection import cross_val_score
```

```
df
```

	population	householdsize	medIncome	PctUnemployed	PolicReqPerOffic	murders	rapes	burglaries	robberies	violent_crime_occurence
0	11980	3.10	75122	2.70	0.0	0	0	14	1	0
1	23123	2.82	47917	2.43	0.0	0	1	57	5	0
2	29344	2.43	35669	4.01	0.0	3	6	274	56	0
3	16656	2.40	20580	9.86	0.0	0	10	225	10	0
4	11245	2.76	17390	9.08	0.0	0	0	91	4	0
...	...	...	...	...	...	...	...	...	...	...
2210	56216	3.07	24727	9.99	0.0	10	30	1376	121	1
2211	12251	2.68	20321	7.90	0.0	0	4	104	1	0
2212	32824	2.46	27182	5.18	0.0	5	5	628	24	0
2213	13547	2.89	19899	12.12	0.0	0	2	192	7	1
2214	28898	2.61	23287	9.27	0.0	5	19	791	102	1

2215 rows × 10 columns

RESULT AND CONCLUSION

The predictions made by various classification algorithms show the occurrence possibility of a crime whether a crime will occur or not, if a crime occurs, will it be a violent or a non-violent crime or if a crime occurs, is the cause of the crime murder or not. These predictions might help the local police departments as well as the FBI solve many cases with efficiency and accuracy.

It was necessary to combine many datasets that display different patterns in order to identify distinct features that may help in the prevention or elimination of crime. Even though the majority of the research for this project has been devoted to estimating crimes that have already occurred, recent advancements in the use of artificial intelligence and big data analysis have made it possible to more easily and clearly highlight dynamic linkages and relationships based on big data. We created a model utilising the data, which had undergone thorough data cleaning and analysis using techniques based on machine learning.

REFERENCES

[1] John Braithwaite et al. Crime, shame and reintegration. Cambridge University Press,1989.

[2] Marcelo Beckmann, Nelson FF Ebecken, Beatriz SL Pires de Lima, et al. A knnundersampling approach for data balancing. Journal of Intelligent Learning SystemsandApplications,7(04):104, 2015.

[3] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, andAlex Pentland. Once upon a crime: towards crime prediction from demographics andmobiledata.InProceedingsofthe16thinternationalconferenceonmultimodalinteraction,pages 427–434,2014.

[4] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer.Smote: synthetic minority over-sampling technique. Journal of artificial intelligenceresearch,16:321–357,2002.

[5] SoonAeChun,VenkataAvinashPaturu,ShengchengYuan,RohitPathak, Vijayalakshmi Atluri, and Nabil R. Adam. Crime prediction model using deep neuralnetworks. In Proceedings of the 20th Annual International Conference on DigitalGovernmentResearch,pages 512–514,2019.

[6] IsaacEhrlich.Ontherelationbetweeneducationandcrime.InEducation,income,andhumanbehavior,pages 313–338.NBER,1975.

[7] Richard B Freeman. The economics of crime. Handbook of labor economics 3 (c).editedbyo.ashenfelterandd.card,1999.

[8] Rizwan Iqbal, Masrah Azrifah Azmi Murad, Aida Mustapha, Payam Hassany ShariatPanahy,andNasim Khan AhmadLiravi.An experimental study of classificationalgorithmsforcrimeprediction.IndianJournalofScienceandTechnology,6(3):4219–4225,2013.

[9] Hyeon-Woo Kang and Hang-Bong Kang. Prediction of crime occurrence from multi-modaldatausingdeeplearning.PloSone,12(4),2017.

[10] Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri. Crime analysisthroughmachinelearning.In2018IEEE9thAnnualInformationTechnology,ElectronicsandMobileCommunicationConference(IEMCON),pages415–420.IEEE

[11] J Kiran and K Kaishveen. Prediction analysis of crime in India using a hybrid clustering approach. In 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile,Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on, pages 520– 523. IEEE, 2018.

[12] Shen Ting Ang, Weichen Wang, and Silvia Chyou. San francisco crime classification.UniversityofCaliforniaSanDiego,2015.