International Journal of Inventive Engineering and Sciences (IJIES)
ISSN: 2319-9598 (Online), Volume-10 Issue-3, March 2023

Crime Analytics using Machine Learning

Suman Acharya

Abstract: ***Crime is one of the most significant and pervasive problems in our society, and preventing it is a crucial duty. A large number of crimes are perpetrated each day. Maintaining and analyzing crime data to forecast and solve crimes is the current issue. This project explores a large dataset of crimes and predicts future crimes based on conditions. This project utilises data science and machine learning to predict crime data in India. Thus, Decision Tree, Logistic Regression, Multi-Regression, k-NN, Lasso & Ridge, and Random Forest are all involved in the supervised classification problem. Predicting crimes and classifying effective pattern detection and visualisation equipment. Utilising crime data trends from the past allows us to correlate aspects that may help us comprehend the breadth of crimes in the future. This study employs visualisation and machine learning methods to predict future crime rates. First, raw datasets were processed and displayed.***

Keywords: Predictive Analytics, Crime Analytics, Machine Learning, Performance Enhancement, Pattern Detection, Decision Learning.

I. INTRODUCTION

Crime poses the greatest threat to humanity. Numerous crimes occur at regular intervals. Perhaps it is growing and spreading rapidly and widely. From small villages and towns to major metropolitan areas, criminal activity is prevalent. There are various types of crimes, including robbery, murder, rape, assault, imprisonment, and kidnapping, etc. [1][2]. Crime prediction enables analysts to visualise criminal networks, mitigate risks, and enhance productivity [3][4]. A good prediction technique helps predict crime rates, evolve crime data sets faster, and track crime analysis resources. Crime analysis can be done using machine learning techniques. Machine learning approaches utilise computers and mathematics to program systems to operate autonomously. These methods help prevent and identify crime. Crime analysis includes pattern extraction, prediction, and detection. With such a wealth of crime data, the police force struggles to predict crime accurately. Technology is required for faster case resolution. Train a prediction model. The test dataset will validate the training dataset. Based on accuracy, more intelligent algorithms will build the model [5][6]. This work enables Indian law enforcement organisations to predict and detect crimes more accurately, thereby reducing crime rates. Machine learning algorithms have significantly enhanced crime prediction using prior data.

Manuscript received on 07 January 2023 | Revised Manuscript received on 14 March 2023 | Manuscript Accepted on 15 March 2023 | Manuscript published on 30 March 2023.
*Correspondence Author(s)
Prof. Suman Acharya*, Assistant Professor, University of Engineering and Management, Jaipur (Rajasthan), India. E-mail id: Suman.acharya@uem.edu.in, ORCID ID: https://orcid.org/0009-0002-2039-7239

The goal of this project is to utilise machine learning models to analyse and predict crime trends over the next few months. It creates a model to estimate monthly crime by type. In this research, several machine learning models, including regression techniques, K-NN, and boosted decision trees, will be utilised to predict criminal behaviour. A month-wise crime analysis can be performed to comprehend the crime pattern [7]. Using various visualization tools and graphs can aid law enforcement organizations in detecting and predicting crimes with greater precision. This will indirectly help reduce crime rates and enhance security in these essential areas. We utilised a dataset of crime reports for India from 2022, which was made available on the official website for datasets. The data, which originated from the Indian Police Department, had over 5473 records, or data points [7]. Each data point consists of 14 attributes representing diverse information on the major head, minor head, crimes committed in the past years and months, and the severity of the reported offence [7].

II. LITERATURE SURVEY

Machine learning models forecast crime using India's crime data collection. This paper compares KNN, SVM, and regression models. Dataset and feature selection affect prediction. KNN predicts 80%, Linear Regression 91%, SVC 86%, Lasso and Ridge Regression 85%, Logistic Regression 87%, and SVC 84% [8][9][10]. Crime pattern detection is a big issue. Understanding datasets is also crucial. To avoid wasting resources on erroneous or missing numbers, we took the mean. Additionally, employing a strategy for categorising the crime rate as high, medium, or low, no one has identified the types of crimes that can occur or their likelihood of occurrence. Crime analysis and prediction are crucial activities that can be optimised through the use of various approaches and processes. Numerous researchers conduct extensive studies in this field. Existing work is confined to identifying the number of crimes committed in the current month using datasets [11][12] [13].

III. PROPOSED SYSTEM

The proposed system employs the Lasso-KNN combined technique to get greater precision and outcomes than other superior algorithms such as random forest, SVR, KNN, and decision tree classifiers. Lasso employed two-thirds of the Crime dataset for training. Then we retrained Lasso's projected value using KNN to achieve more accurate results—a collection of dataset-record decision trees. Lasso and KNN have a precision of 85%, random forest 84%, and SVR 90%. Our primary goal is to increase the accuracy of the KNN regressor method from 80% to 85%, predict accurate results for our Crime dataset using Lasso-KNN, and avoid overfitting and underfitting.

Our proposed approach employs Lasso and KNN, which are more accurate than previous regression algorithms by 85%. The next step uses Lasso and SVR, which are 91% more accurate than the previous regression algorithms and Lasso-KNN. Numerous graphs, such as bar graphs, may be applied in this representation. Using the matplotlib package from Sklearn [14]. The crime dataset is analyzed through the use of graphs

IV. METHODOLOGY

- **Data Collection Methods**:

Using important data sources, data collection was done for crime prediction. These are:

- Social media analysis
- Records of crime from the website of the Indian government
- Social media analysis
- newspapers;
- CRDs (Call Data Records)

The Indian Police Department's "Crime Data 2022" is the most extensive dataset maintained by a law enforcement agency. A sizable number of crime prediction systems are being tried in India. Social media is another important method of gathering information for crime prediction [7].

To accurately estimate the number of victims for the current month, we applied specific machine learning techniques. Using a combination of KNN and Lasso also improves the accuracy of KNN. SVR and the Random Forest Regressor are used to make accurate predictions. We estimate the appropriate number of victims for August and December of the current year using databases of all crimes committed in those months.

- **Approach**:

Lasso-KNN:

Here, we employ a combination approach of Lasso and KNN to enhance the accuracy of KNN from 80% to 85%. Specifically, we first train the model using Lasso and then retrain it with KNN to increase the accuracy to 85%.

```
: import matplotlib.pyplot as plt
  plt.scatter(x_test['Major head'],y)
  plt.show()
```

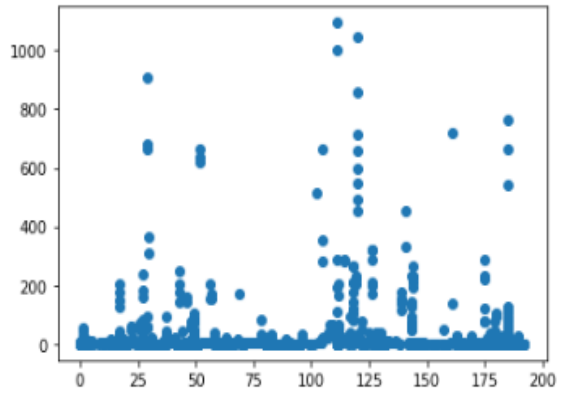

Fig 4.1: Scatter plot of Lasso-KNN of Crime Dataset Between major Head and Y Predict.

Lasso-SVR:

Here, we utilise a combination of Lasso and SVR to enhance the accuracy of the previously proposed method, increasing it from 90% to 91%. To be more specific, we train the model with Lasso first and then retrain it with SVR to achieve an accuracy of 91%.

```
In [48]: import matplotlib.pyplot as plt
         plt.scatter(x_test['Major head'],y)
         plt.show()
```

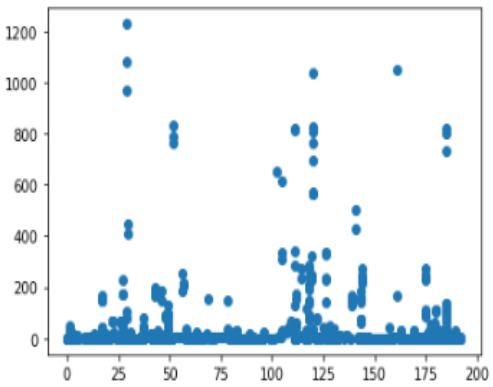

Fig 4.2: Scatter Plot of Lasso-SVR of Crime Dataset Between Major Head and Y Predict.

In this project, we examined the various plots and algorithms available for estimating future criminal activity.

V. RESULTS

Score Table:

Algorithm	Score
Linear Regression	91%
Logistic Regression	87%
K-Nearest Neighbor	80%
Random Forest Regressor	84%
Random Forest Classifier	91%
Lasso Regression	91%
Decision Tree Classifier	90%
Lasso-KNN	85%
Support Vector Regressor	90%
Lasso-SVR	91%

Given that the output of Lasso-SVR is superior to that of competing algorithms, we adopt a Lasso-KNN strategy to boost the precision of KNN.

Dataset:

The dataset was obtained from the government dataset website by month, but it must be organised and preprocessed before usage.

It has columns such as "Month of the previous year," "Major Head," "Minor Head," "End of month," "prior month," "current month," and "Name of month." We heavily rely on the previous year's month, the last month, the end of the month, the major head, and the name of the month to forecast the value of the current month.

Retrieval Number: 100.1/ijies.F74440311623
DOI: 10.35940/ijies.F7444.0310323
Journal Website: www.ijies.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

	Act	Major head	Minor head	End Month	Month Previous	Previous month	Current month	Month
0	A - IPC Crime	Murder (Sec.302/303 IPC)	For gain	6.0	4.0	3.0	6.0	1
1	A - IPC Crime	Murder (Sec.302/303 IPC)	Over Property Dispute	4.0	2.0	5.0	4.0	1
2	A - IPC Crime	Murder (Sec.302/303 IPC)	Due to Personal Vendetta or enemity	2.0	1.0	2.0	2.0	1
3	A - IPC Crime	Murder (Sec.302/303 IPC)	Due to Sexual jealousy	3.0	2.0	2.0	3.0	1
4	A - IPC Crime	Murder (Sec.302/303 IPC)	For dowry by burning	0.0	1.0	0.0	0.0	1
...	...	...	...	...	...	...	...	...
5468	D.CRIME AGAINST CHILDREN	Probation of Offenders Act	NaN	3.0	0.0	0.0	0.0	9
5469	E. CRIME AGAINST SCHEDULED CASTES /TRIBES BY ...	Murder	NaN	69.0	7.0	3.0	6.0	9
5470	E. CRIME AGAINST SCHEDULED CASTES /TRIBES BY ...	Rape	NaN	161.0	15.0	25.0	13.0	9
5471	E. CRIME AGAINST SCHEDULED CASTES /TRIBES BY ...	Kidnapping	NaN	16.0	0.0	6.0	2.0	9
5472	E. CRIME AGAINST SCHEDULED CASTES /TRIBES BY ...	Offences under the Protection of Civil Rights ...	NaN	4.0	0.0	0.0	2.0	9

5473 rows × 8 columns

Figure 5.1: Crime dataset

KNN Score:

Here, we utilised the KNN method to determine the predicted output and assess its accuracy. Here, we first train our model using the fit method in Python, and then we use it to predict the test data, which accounts for 30% of our dataset. Seventy per cent of our dataset is used for training purposes. Here, KNN achieves an accuracy of approximately 80%. Here, we must enhance the precision of KNN by combining the Lasso and KNN approaches.

```
In [117]: from sklearn.neighbors import KNeighborsRegressor

In [118]: kn=KNeighborsRegressor(8)

In [119]: kn.fit(x_train,y_train)
Out[119]: KNeighborsRegressor(n_neighbors=8)

In [120]: l=kn.predict(x_test)

In [123]: kn.score(x_test,y_test)
          0.803787230493467
```

Figure 5.2: Model Score after applying the KNN approach

Lasso-KNN Score:

Here, we first use the Lasso method for training, which predicts all training outputs and stores all predicted results in the g_train variable. Then, we use the KNN approach for retraining, which includes x_train and g_train (the predicted Lasso results). Then, we construct a model that accurately predicts the x-test results. This strategy increases the KNN accuracy by 5%.

```
In [38]: from sklearn.neighbors import KNeighborsRegressor
         from sklearn.linear_model import Ridge,Lasso
         b=Lasso(alpha=2.0)
         b.fit(x_train,y_train)
         g_test=b.predict(x_train)

In [39]: kn=KNeighborsRegressor(8)

In [40]: kn.fit(x_train,g_test)
Out[40]: KNeighborsRegressor(n_neighbors=8)

In [41]: kn.predict(x_test)
Out[41]: array([-0.09811477,  0.24767629,  0.05188954, ...,  6.90322526,
                0.48076657,  0.52521337])

In [43]: kn.score(x_test,y_test)
Out[43]: 0.8461155447033872
```

Figure 5.3: Model Score after applying the Lasso-KNN approach

After employing the lasso-KNN method, we successfully improved the KNN accuracy.

Lasso-SVR Score:

Here, we begin by training with the Lasso method, which predicts all training output and stores all predicted results in the g_train variable. Then, we retrain using the SVR method, which includes x_train and g_train (predicted Lasso results). Then, we construct a model that accurately predicts the x-test results. This strategy improves the accuracy of the Lasso by 6%.

```
In [26]: from sklearn.svm  import SVR

In [27]: a=SVR(kernel='linear')

In [29]: from sklearn.linear_model import Ridge,Lasso
         b=Lasso(alpha=2.0)
         b.fit(x_train,y_train)
         g_test=b.predict(x_train)

In [30]: a.fit(x_train,g_test)
Out[30]: SVR(kernel='linear')

In [31]: a.predict(x_test)
Out[31]: array([-0.04806038,  0.28591292,  0.1020071 , ...,  9.8274206 ,
                0.56392885,  0.52531227])

In [33]: a.score(x_test,y_test)
Out[33]: 0.9141598344895
```

Figure 5.4: Model Score after applying Lasso-SVR approach

Predicted output:

The number of victims is expected to be announced in August. We are utilizing appropriate supervised learning techniques to make accurate forecasts. When the month is set to 8, the major head is set to 14, the previous year's month is set to 5, the prior year's month is set to 9, and the current year up to the end of the month under review is set to 2, the output is 3.47 for Lasso-KNN and 8.22 for SVR. Our proposed algorithm delivers a better result than SVR. As our proposed algorithm, Lasso-KNN, yields an output of 3.47, which is somewhat larger than 2 for the current year up to the end of the month under review. However, it also yields 8.22, which is significantly greater than 2.

Therefore, there is a high probability of an overfit error occurring. However, we require more accurate results than our proposed method, Lasso-KNN, can provide, so we are using our proposed method, Lasso-SVR, instead. Our proposed method, Lasso-SVR, yields a result of 6.57, which outperforms the results of all other algorithms. We are still using the two proposed methods, but Lasso-SVR is giving better results than Lasso-KNN. However, by using the Lasso-KNN approach, we were able to improve the accuracy of KNN from 80% to 85%. We utilise Python, a widely used programming language for statistical analysis, to make predictions about the frequency of various types of crime in this project.

```
In [213]: a.predict([[8,14,5,9,2]])
Out[213]: array([8.22962592])
```

Figure 5.5: Predicted value of the SVR algorithm

```
In [62]: kn.predict([[8,14,5,9,2]])
Out[62]: array([3.4742516])
```

Figure 5.6: Predicted value of the Lasso-KNN approach

```
In [168]: kn.predict([[8,14,5,6,2]])
Out[168]: array([[0.5]])
```

Figure 5.7: Predicted value of the KNN algorithm

```
In [259]: a.predict([[8,14,5,9,2]])
Out[259]: array([6.57459174])
```

Figure 5.8: Predicted value of Lasso-SVR algorithm

VI. DISCUSSION

This paper discusses crime prediction systems that utilise machine learning techniques and examines how these systems operate. In machine learning, algorithms such as K-Means, SVR, Lasso-KNN, linear regression, logistic regression, Random Forest, and decision trees are used to predict events. From these used data analysis algorithms, "K-Means" yields a lower score, so we need to combine it with "Lasso" to increase the score and prevent "overfitting," which occurred with the "SVR" algorithm. This happened because the criminal data was reliable and relevant. However, K-Means won't work well with noisy data, so we need to preprocess it to handle missing and NaN values. The algorithm also won't give a better average when there are more clusters. Because of this, K-Means works best when the data is not too noisy and there aren't too many clusters. However, when we combined Lasso and KNN, we achieved a prediction accuracy of 85% of the time. On the other hand, we require more accurate results, so we need to utilise the Lasso-SVR approach to achieve the most precise output of 6.57, which is significantly better than the Lasso-KNN approach. In this project, we employ two proposed methods: first, we increase the score of KNN to 85%, and then we utilise the best approach to achieve a score of 91%. These are two possible ways to do something. The area being discussed has grown because crime pattern detection systems now also utilise image processing.

VII. CONCLUSION

With the use of machine learning technologies, it has become easier to identify relationships and patterns within diverse data sets. This study primarily focuses on predicting the types and numbers of crimes that may occur. Using the concept of machine learning, we constructed a model with training datasets that underwent data cleansing and modification. When we use Lasso-KNN, the model can predict the type of crime with 0.846% accuracy. When we use Lasso-SVR, it achieves the same result with 0.914% accuracy. Data visualization facilitates data set analysis. The graphs consist of bar, line, and scatter graphs, each with its distinct qualities. We developed numerous graphs and discovered intriguing statistics that helped us comprehend Indian crime datasets that can aid in identifying elements that contribute to a safer society. Our program provides a framework for analysing the number of crimes and their patterns using a variety of machine learning algorithms. Utilising a range of interactive visualisations, the initiative supports criminal analysts in analysing these crimes. The interactive and visual feature applications will aid in the reporting and identification of criminal patterns. Law enforcement agencies can greatly benefit from utilising machine learning algorithms to combat crime and save lives.

ACKNOWLEDGEMENT

This paper's author is grateful to Mr. Somen Nayak and Dr. Yogesh Kumar Jakhar, Assistant Professor, for their assistance and support during this independent research endeavour.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval or consent to participate, as it presents evidence that is not subject to interpretation.
Availability of Data and Material/ Data Access Statement	The data has been extracted from the website of the Indian government (https://data.gov.in/catalog/crime-review-report-2022) and is being pre-processed to improve its format.
Authors Contributions	I am the sole author of the article.

REFERENCES

1. Lakshman Narayana Vejendla and A Peda Gopi, (2019),” Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology”, Revue d'Intelligence Artificielle, Vol. 33, No. 1, [CrossRef]
2. Gopi, A.P., Jyothi, R.N.S., Narayana, V.L.et al. (2020), “Classification of tweets data based on polarity using improved RBF kernel of SVM”. Int. j. inf. tecnol. (2020). [CrossRef]
3. A Peda Gopi and Lakshman Narayana Vejendla, (2019),” Certified Node Frequency in Social Network Using Parallel Diffusion Methods”, Ingénierie des Systèmes d' Information, Vol. 24, No. 1, 2019, pp 113-117. [CrossRef]
4. Lakshman Narayana Vejendla and Bharathi C R,(2018), “Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs”, Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. [CrossRef]
5. Lakshman Narayana Vejendla and Bharathi C R, (2018), “Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS”, Modelling, Measurement and Control A, Vol. 91, Issue 2, pp.73-76. [CrossRef]
6. Lakshman Narayana Vejendla, A Peda Gopi and N.Ashok Kumar,(2018),“ Different techniques for hiding the text information using text steganography techniques: A survey”, Ingénierie des Systèmes d'Information, Vol . 23, Issue 6, pp. 115-125. [CrossRef]
7. Crime review report of India, Open government data platform, NIC, Ministry of Electronics & Information Technology, https://data.gov.in/catalog/crime-review-report-2022.
8. Patibandla, R.S.M.L., Veeranjaneyulu, N. (2018), “Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria”, Arab J Sci Eng, Vol . 43, pp.4379–4390. [CrossRef]
9. R S M Lakshmi Patibandla, Santhi Sri Kurra and N.Veeranjaneyulu, (2015), “A Study on Real-Time Business Intelligence and Big Data”, Information Engineering, Vol . 4, pp.1-6. [CrossRef]
10. K. Santhisri and P.R.S.M. Lakshmi,(2015), “Comparative Study on Various Security Algorithms in Cloud Computing”, Recent Trends in Programming Languages, Vol . 2, No.1, pp.1-6.
11. Patibandla, R. S. M. Lakshmi et al., (2016), “Significance of Embedded Systems to IoT.”, International Journal of Computer Science and Business Informatics, Vol. 16, No.2, pp. 15-23.
12. Anveshini Dumala and S. PallamSetty. (2020), “LANMAR routing protocol to support real-time communications in MANETs using Soft computing technique”, 3rd International Conference on Data Engineering and Communication Technology (ICDECT-2019), Springer, Vol. 1079, pp. 231-243. [CrossRef]
13. Anveshini Dumala and S. PallamSetty.(2019), “Investigating the Impact of Network Size on LANMAR Routing Protocol in a Multi-Hop Ad hoc Network”, i-manager’s Journal on Wireless Communication Networks (JWCN), Volume 7, No. 4, pp.19-26. [CrossRef]
14. Paul E. Barrett, J.Hunter, J.T. Miller, J.-C.Hsu,”matplotlib, A portable Python plotting package”, Conference, Astronomical data analysis software and systems XIV, Volume:347.

AUTHORS PROFILE

Prof. Suman Acharya is an assistant professor at the University of Engineering and Management in Jaipur, Rajasthan. He received his BCA in Computer Applications from the Institute of Engineering and Management, Kolkata, West Bengal, and his MCA in Computer Applications from the University of Engineering and Management, Kolkata, West Bengal. He also participated in several workshops and faculty development programs. His research focuses on machine learning, data mining, and decision learning, all of which combine to form the discipline of artificial intelligence. Currently, he must shift his research focus to quantum computing to apply his expertise in artificial intelligence to quantum artificial intelligence.