CRIMINAL RECIDIVISM PREDICTION USING MACHINE LEARNING

INTRODUCTION: -

Machine learning is a branch of computer science closely related to the study of algorithms and statistical models that lie at the intersection of computer science, engineering, statistics and often appears in other disciplines. Machine Learning is a study used by computer systems to perform a certain task without being explicitly programmed. The machine learns itself based on experience and relies on patterns and inferences instead. [1]

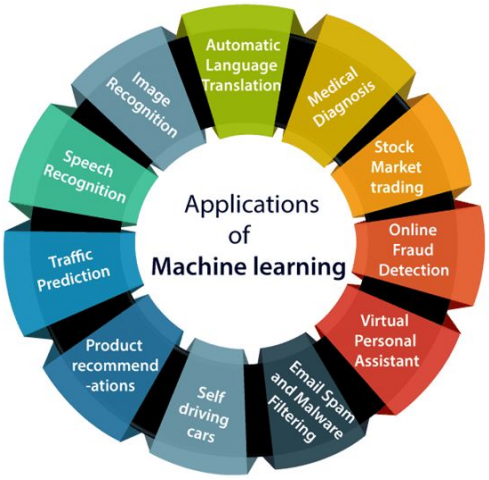
The use of Machine learning (ML) enables the computer systems to grasp, analyse and interpret information, without the need of programming it explicitly. Through machine learning and deep learning, the computers are made acquainted to new datasets, through which they can identify and analyse various trends. These enable the computer systems, also called models, to efficiently learn and adjust the parameters, such that the trends can be incorporated to optimise the performance of any algorithm to be implemented, like classification, regression, clustering or any other such application. To carry out the above, we need to perform various preprocessing steps to ensure the compatibility of the data with the code, using data mining, cleaning and preprocessing techniques. Machine and Deep learning models do not require to be explicitly programmed to understand and interpret trends in the dataset, they automatically enable the computer system to do it, once trained on a particular set of data (called the training sample). These trends can be used to predict and compare the performance of machine learning models, with actual data, in the testing phase (using the testing sample.) [2]

Leading industries today have realised the lucrativeness of incorporating Machine Learning into their existing systems. The adeptness of ML of gauging hidden trends in the data, is what drives numerous industries into adopting machine learning solutions for various problems faced, hence burgeoning their business. Traits of machine learning, such as easy affordability, efficient computational processing, cost effective data storage and myriad other such benefits have made it feasible and advantageous to develop models which analyse large amounts of complex data expeditiously. Moreover, the automation of tasks, originally requiring manpower has been ensured by using such models, which is a highly desirable feature for various industries. Adherence to the use of Machine Learning has facilitated personalized services and differentiated products that meet and satiate the demands of the customers. Furthermore, machine learning and deep learning facilitates a prescient outlook to businesses, which can prove to be profitable in the long run. [3]

The Machine Learning Algorithms can be applied in various domains like: - Healthcare industry, Automotive Industry, Finance and Banking, Industrial Sector, etc. The various applications are Face Detection, Anti-virus and anti-spam, Weather forecast, etc. Deep Learning algorithms can be used in numerous domains like: - Agriculture Field, Food and Processing Sector, Object Detection NLP. Some of the major extensively used applications of

deep learning are listed below: - Image Classification, Self-Driving Cars, Fraud Detection, and Automatic Handwriting Generation. [4]

Deep learning is a study that is closely related to machine learning which uses extensive application of a hierarchical level of artificial neural networks. A neural network works in correlation with input, hidden and output layers called nodes. The main function is to minimize the difference between its prediction and expected output. There are various types of neural networks which exist for different applications; e.g. Convolutional NN for computer vision, Recurrent NN for NLP, etc. [5]



[6] Fig.1 [C]

Fig.2 [D] [7]

In this chapter we talk about one of the Applications of Machine Learning called the Prediction Model for Criminal Recidivism. It discusses the possibility and tendency of a felon to commit an offence again. The chapter flow begins by defining criminal recidivism and subsequently describing its effects on society. It is preceded by discussing the existing systems and the issues regarding it. A detailed account of the data analysis and data science techniques employed and the results obtained, has been given in the next phase of the chapter. The benefits and influence of data science and data analytics has also been discussed. Their use on the dataset employed for the study has also been incorporated and discussed in the next phase. Following that, the Machine Learning models used have also been documented and compared on basis of their accuracy for the purpose of classification of criminals into distinct categories of potential recidivists. [8] The chapter is eventually concluded by talking about the conclusion and future scope of the project.

CRIMINAL RECIDIVISM

Often, a criminal who is granted parole/bail, may be involved unpropitiously, in committing a crime. Criminal recidivism can be defined as the tendency of convicted criminals, to commit a crime again. There have been many cases, around the globe, where a criminal on bail/parole has committed a crime, owing to multiple reasons. Many judicial authorities employ Risk Assessment Instruments (RAIs) to make decisions of whether to grant parole/bail to the criminals. There are statistical models (called Pre-trial RAIs) used before the upcoming court dates, which assess the probability of a criminal to commit a crime, given their release on parole/bail. One of the existing systems is COMPASS RAI. Compass RAI is an efficient risk assessment instrument that predicts accurately a defendant's risk of committing a felony or same offence repeatedly. It takes into account the features of the individual and the individual's past criminal record. [9] The limitations of the COMPASS RAI were that some interpersonal nuances like demeanour, eye contact, body language cannot be picked up by the compass tool. Meta-analysis found only moderate levels of accuracy in decisions of preventative detention.

DATA ANALYSIS AND DATA SCIENCE: -

Data Analysis is defined as a process of various operations that needs to be performed on data in order to extract useful information for business decision-making. The various operations performed on the data are cleaning, transforming and modelling of data. There are various Data Analysis tools for processing and manipulating the data, analysing the relationships and also correlations among the data sets. It helps to identify and extract various unique patterns and trends for interpretation. The various tools used for performing data analysis are Python, R, MATLAB, Java, SAS, and SQL. [10]

Data analysis is important in business to understand problems encountered by an organization or company. It gives information to an organization in the form of numbers, figures and graphical representation. The tasks mainly performed are: - 1. To present technical insights 2. Contribute to decision making 3. To explore the meaning behind the numbers and figures in data. [11]

Data preprocessing is an important step that needs to be done before Machine Learning is performed. Data preprocessing is a data mining technique which is employed to transform the raw data in a useful and efficient format. The various steps involved in Data Pre-Processing are import libraries, import datasets, taking care of the missing values, encoding categorical data, splitting the dataset into training and testing and feature scaling. [12]

[13]

Fig.3 [E]

The purpose of Data Analysis and Data Science in the field of Criminal Recidivism is to give the model a complete and consistent data set. If the data set is not cleaned or complete, it will result in an inappropriate output. The data set for Criminal Recidivism is taken from Carnegie Mellon University (CMU) Data Repository. The data set consists of around 20 features and 60000 records. The data set tells us whether a criminal will commit a crime or not after the sentence or bail.

We performed various functions on the above mentioned data set for analysis and exploratory purpose. Firstly, we analysed the uniformity about the data set i.e. to know whether the data set is uniform. Through this we can achieve whether the accuracy parameter will decide the performance of the model or there should be some other decision making performance parameter. We analysed various features which will contribute to the better performance of the model. The next step we performed is cleaning of the data, removing the redundant data and dropping the unnecessary columns. After performing the above mentioned step, the dataset of 60000 records came down to 18000 records. The features that we have employed in our model are 'Score Text, Assessment Type, Record Supervision Level, Custody Status, and Legal Status'. The prediction is analysed by considering the above mentioned features and also few other features. Statistical calculation for race/ethnicity groups, ages, gender and marital status involved in the crime is shown using graphical representation.[14]

STATISTICAL REPRESENTATION: -

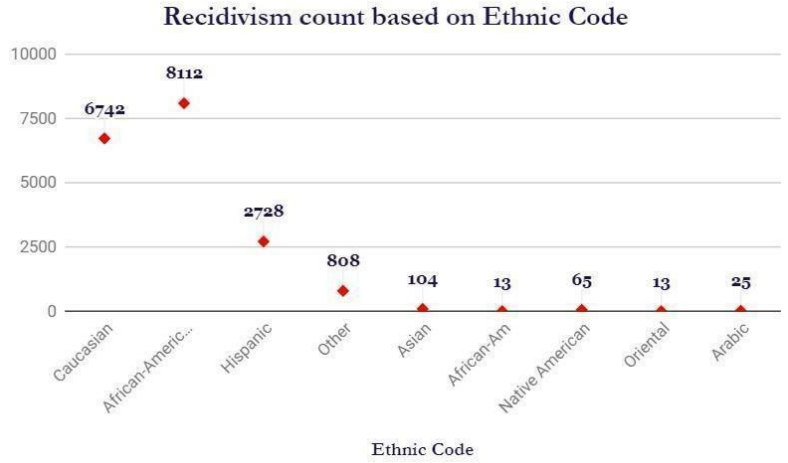

Recidivism count based on Ethnic Code
10000
7500
5000
2500
0
6742
8112
2728
808
104
13
65
13
25
Caucasian
African-Americ...
Hispanic
Other
Asian
African-Am
Native American
Oriental
Arabic
Ethnic Code

Fig.4

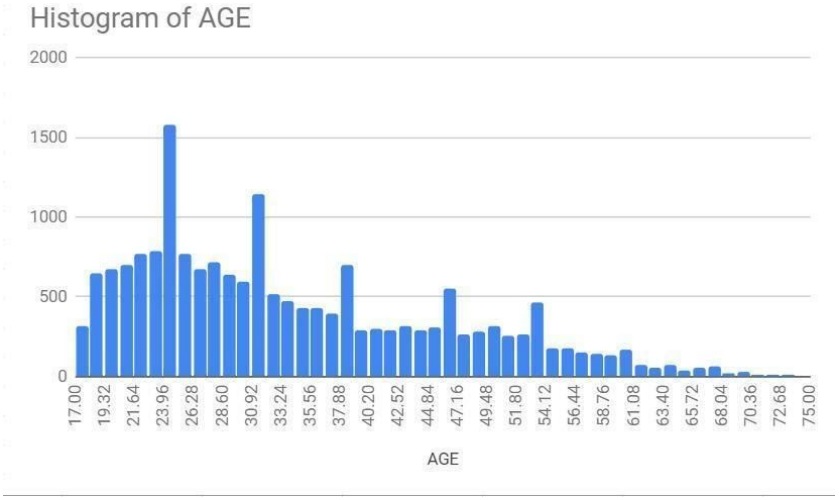

Histogram of AGE
2000
1500
1000
500
0
17.00
19.32
21.64
23.96
26.28
28.60
30.92
33.24
35.56
37.88
40.20
42.52
44.84
47.16
49.48
51.80
54.12
56.44
58.76
61.08
63.40
65.72
68.04
70.36
72.68
75.00
AGE

Fig.5

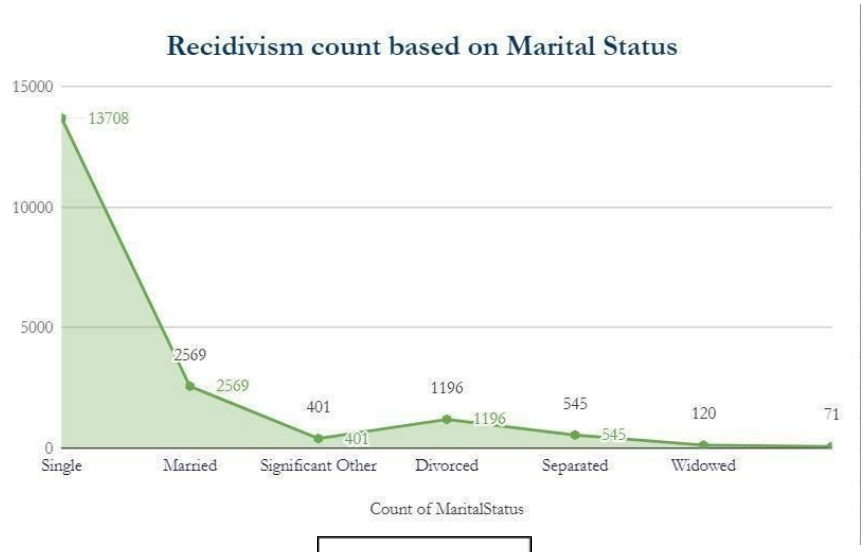


Fig.6

MACHINE LEARNING MODELS:-

In this chapter, we have used three models:-

1. Random Forest
2. K-Nearest Neighbours
3. Logistic Regression

1. Random Forest: -

The Random Forest algorithm is an algorithm used extensively in the emerging field of Machine Learning. This is mainly a supervised classification algorithm, which means that it is used for segregating or assigning the object to a particular class of instances where it might belong. An example of classification is to classify a cancer as benign or malignant. There are myriad such algorithms used for classification, like Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbors (KNN) and numerous others [15]. This algorithm in addition to being used for classification can also be seldom used for regression.

The Random Forest algorithms is a huge collection and cluster of numerous decision trees.

DECISION TREE:

The decision trees have a representation similar to the tree structure. A tree-like flowchart is drawn, where each node resembles an attribute, the branches denote a decision rule and the leaf nodes are the final outcomes. These trees, as suggested by the name help in decision making process by the numerous recursive branches of the tree. CART (Classification and Regression techniques) are applied in deriving and deciding the final outcome of any event.

The results are governed by various factors like probability and various indices. Parameters like Gini Index and Entropy are some of the factors which are used to determine the root node and construct the final decision tree. The outcomes are then drawn by traversing the branches according to the conditions and values posit in the data tuple. [16]

To construct the decision tree, to predict the outcome of a single tree, we can use the Gini Impurity or the Entropy.

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Using this above formula and multiplying this with the probability of the instances of the tuple, we obtain the Gini index. The attribute with the least value becomes the root node and the same procedure is now carried out again, once the tuples are classified according to the new root node.

This repeats till we get definite leaf nodes which signify the final outcome, when the tree is traversed using a given tuple.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Information Gain (Parent, Child)

*= Entropy(parent) - [p1(c1) * entropy(c1) + p(c2) * entropy(c2) + …]*

Entropy is calculated in a similar fashion, but the one with the largest value is made the root node in case of decision tree formation using entropy calculation. [17]

This algorithm falls under the category of Ensemble machine learning algorithm, which are meta- algorithms which combine several machine learning techniques into one single model which decreases variance, bias and improves predictions. [18]

The random forest being a collection of numerous decision trees depend on the outputs of these trees. The final output of the random forest algorithm is decided by choosing the outcome with maximum votes as decided by all the individual trees. Thus, this ensemble approach provides a highly efficient output, since in this manner the individual errors are compensated for by the other trees. [19]

The correlation between the trees present in the forest is very low and this plays a very paramount role in the decision of outcome.

A pictorial representation of the Random forest algorithm can be depicted by the image below. [20]

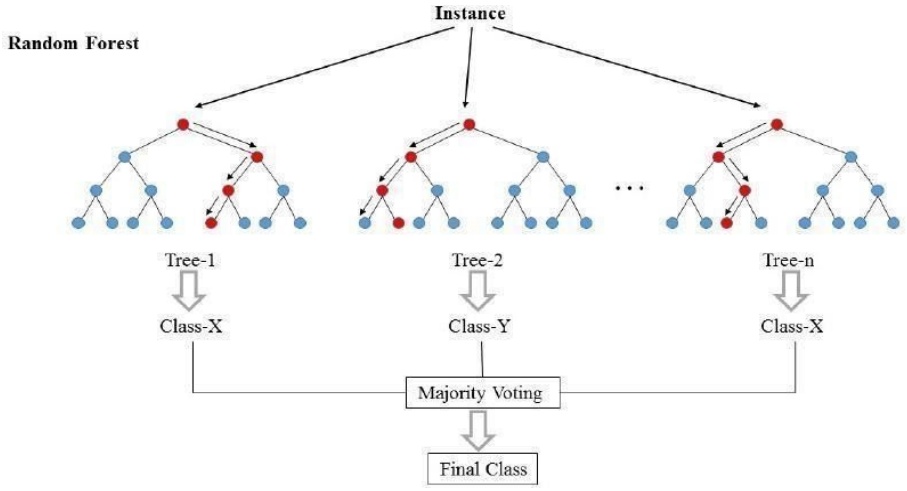


Fig.1 [21]

2. K-Nearest Neighbours: -

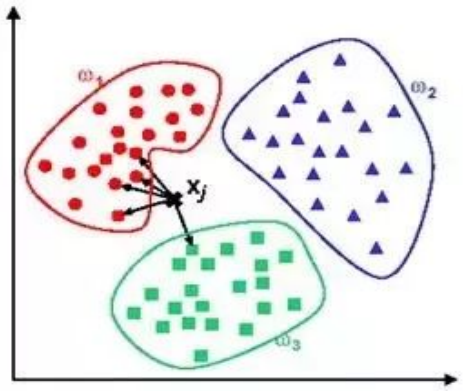
The kNN Algorithm of Supervised Machine learning stands for K-Nearest Neighbours. This is a classification algorithm, used for many applications in classifying the data tuples. To explicate this, we can consider a data tuple N, to be classified, among the retrieved k nearest neighbours to tuple N. The classification of the tuple N is dependent on the weight based factors (distance of the tuple from the neighbours) and is classified into the neighbourhood with minimum associated weight [22].

There is no predetermined value of k, which can certainly yield an optimum outcome. The value of k is changed each time, which reflects the change in accuracy of the model. The value of k, at which maximum accuracy is obtained, can be the right value for our purpose. [23] However, research is being carried out to determine the value of k using a k-Tree Method to determine the optimal value of k, as proposed by the paper mentioned in [24].

In the diagram below, the distance of attribute/tuple xj, is calculated from each of the 3 neighbours (k=3). The neighbours - w1,w2,w3 represent the classes, which denotes the general characteristics of all the data points included in the respective class. The closest class to xj, will determine the class of xj, thus completing and yielding an outcome of the classification problem.

The Euclidean distance between points A(x1,x2….xm) and B(y1,y2...ym), *m* being the dimensionality of the feature space is given by the formula below [25].

$$dist\,(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}}$$



3. Logistic Regression: -

The Logistic Regression algorithm is an algorithm used extensively in the emerging field of Machine Learning. This is mainly a supervised classification algorithm, which means that it is used for segregating or assigning the object to a particular class of instances where it might belong. y=f(x) is a method for fitting a regression curve in logistic regression as the function varies between zero and unity i.e success and failure. It requires iterative methods to fit a logistic regression model and we use the maximum likelihood function to maximise the probability using the values of alpha and beta concerned in the data set. Logistic regression deals with the relationship of multiple independent variables and dependent variables to analyse the probability of the event concerned.[26]

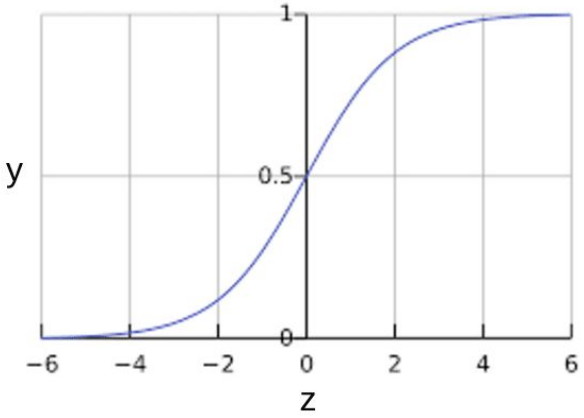
The model is used to predict a plethora of dependent variables distinguished from the pile of independent variables. Overall Evaluation of the model and goodness of fit-statistics are some important factors to be considered while fitting the logistic model efficiently. It's a unique type of multivariate which analyses variables used in high frequency.

We first find the regression coefficient b1 which we used to estimate the increase in the odds of the outcome for every increase in the value of the independent variable. The coefficients are the key representations in a logistic regression model. The odds play a vital role in the accuracy of the model as they signify the ratio of probability of success to failure which are mapped into the model. It tends to behave as a specialised case of the linear regression model but gives a better accuracy in most of the cases and fits the data efficiently.

The Sigmoid Function is a special characteristic function. It is a function which ranges between the values of zero and unity. It has a characteristic “S”-shaped curve which is extensively used in logistic regression [27]. A common example of sigmoid function is the logistic function shown in the figure.

The expression of sigmoid function is:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}.$$



[28]

Maximum likelihood estimation

We will find the β parameters that reach the global maximum of the log likelihood function using maximum likelihood estimation.

Using the maximum likelihood estimation we can fit the model by the equation:

$$\ell(\beta) = ln\left(\sum_{i=1}^{n}(P(X_i;\beta))^{(y_i)}(1-P(X_i;\beta))^{(1-y_i)}\right)$$
$$= \sum_{i=1}^{n}(y_i)\ ln(P(X_i;\beta))\ +\ (1-y_i)\ ln(1-P(X_i;\beta))$$

APPLICATIONS OF MACHINE LEARNING ALGORITHM ON CRIMINAL RECIDIVISM: -

To carry out the above mentioned algorithms, our target variable is the '*ScoreText'* attribute of the dataset, which throughout the dataset may have 3 values - Low, Medium and High. This is the measure of the tendency of the criminal to commit recidivism. Thus, we have set the '*ScoreText'* as the target attribute, since we need to classify the criminals based on their tendency of committing a crime again. The classification into Low, Medium and High risk provides a cogent perspective to the authorities while processing the given criminal for parole or bail. To explicate this, the criminal with a higher tendency, must not get a bail/parole, as compared to one with a lower risk. This solves the purpose of checking and curbing criminal recidivism in society, hence ensuring the safety of citizens and eschewing a potential crime. The attributes selected as features, for the algorithms, directly or indirectly affected and related to the recidivism tendency of the given criminals. Nearly 17 attributes of the dataset were selected for the training and testing of our machine learning models. For a thorough comparative study among all 3 algorithms used, the features of the model remained the same.

Some of them were as follows:-

1. Sex Code
2. Ethnic Code
3. Marital Status (Single, Married, Widowed, etc.)
4. Legal Status (Post Sentence, Conditional Release, Pre-trial, etc.)
5. Custody Status (Jail Inmate, Pre-trial Defendant, Probation, etc.)
6. Rec Supervision Level
7. Agency Text

Multiple such features were incorporated, for a comprehensive training and subsequent testing of the machine learning models. The training data was 70 % of the dataset, while testing was performed on the rest 30 % for an accurate best fit result. The number of training samples were 13020, while 5581 data samples were employed for testing.

COMPARISON OF KNN, RANDOM FOREST CLASSIFIER AND LOGISTIC REGRESSION: -

Parameters	Random Forest Algorithm	Logistic Regression Algorithm	K-Nearest Neighbours Algorithm
Library Used	Scikit Learn (Python)	Scikit Learn (Python)	Scikit Learn (Python)
Total No. of correctly classified tuples	4901	4202	4849
Total No. of falsely classified tuples	680	1379	732
Accuracy Score	87.815 %	75.29 %	86.88 %

The confusion matrix/crosstab matrix of all the above algorithms, can be tabulated using the metrics function of the Scikit Learn library (sklearn) available in python.

1. Confusion Matrix/Crosstab Matrix for Random Forest Algorithm -

Predicted Outcome	High	Low	Medium
Actual Outcome			
High	313	14	51
Low	1	3894	305
Medium	23	338	642

2. Confusion Matrix/Crosstab Matrix for K-Nearest Neighbours Algorithm -

Predicted Outcome	High	Low	Medium
Actual Outcome			
High	315	12	51
Low	4	3949	247
Medium	16	350	637

3. Confusion Matrix/Crosstab Matrix for Logistic Regression Algorithm -

Predicted Outcome	High	Low	Medium
Actual Outcome			
High	0	354	24
Low	2	4175	23
Medium	0	976	27

According to the paper '*Concepts of Recidivism in India*' [29], the two major causes of recidivism are:-

1. Difficulty of Social Adaptation of people released from Punishment -

This is a largely psychological problem faced by the convicted felons. Thus, there is an impediment in trying to ameliorate the repercussions of this cause. It is very arduous to restrain a potential recidivist, from committing recidivism, through psychological therapy. Hence, we cannot do much to eliminate this cause.[30]

2. Shortcomings of Law Enforcement -

This has been effectively dealt with, using our model of predicting and classifying recidivists. This system, when implemented would check and limit the occurrence of recidivism, by ensuring a statistically more reliable and analysed parole/bail hearing and granting process. [31]

CONCLUSION: -

The above model designed to categorise convicted criminals into low, medium and high risk of turning into recidivists, helps curb the increasing crime rates in the society, thus ensuring the welfare and well-being of its citizens. In this way, Machine Learning can be made of paramount importance to perpetuate the security and safety of innocent citizens, who might be potential victims of assaults or any such ordeal. The quest of revenge or mental instability on the part of the felons, might cause them to commit recidivism, endangering not only theirs, but also the lives of many individuals, who might not have any relation to it. Thus, our Machine Learning model aims at resolving one of the above mentioned causes, by overcoming the shortcomings of law enforcement practices, by enabling the authorities to make an informed, statistically and analytically cogent decision, in matters of granting bail/parole to the criminals, who might be potential recidivists.

REFERENCES

[1]:Osvaldo Simeone "A Very Brief Introduction to Machine Learning With Applications to Communication Systems" IEEE(2018)

Available: https://ieeexplore.ieee.org/document/8542764

[2]:https://www.outsource2india.com/software/articles/machine-learning-applications-how-it-works-who-uses-it.asp

[3]:Sheena Angra.Sachin Ahuja “Machine learning and its applications: A review” IEEE(2017)

[4]:Ping Wang, Rick Mathieu, Jie Ki, H.J.Cai “Predicting Criminal Recidivism with Support Vector Machine” IEEE(2010)

[5]:Introduction to Deep Learning - Towards Data Science (2020, March 11). [Online]

[6]:Applications of Machine Learning - Javatpoint [Image]

Source:https://www.google.com/search?q=applications+of+machine+learning&rlz=1C1GGRV_enIN751IN751&sxsrf=ALeKk03tKLrRsmBY01U356p71tlC11DRQA:1586231743989&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjn6IbdtdXoAhWb4zgGHf1vCvsQ_AUoAXoECBAQAw&biw=1366&bih=657#imgrc=rtwjdR9DjEMfqM

[7]:Six Top Applications of Machine Learning |Hacker Noon[Image]

Source:https://www.google.com/search?q=applications+of+machine+learning&rlz=1C1GGRV_enIN751IN751&sxsrf=ALeKk03tKLrRsmBY01U356p71tlC11DRQA:1586231743989&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjn6IbdtdXoAhWb4zgGHf1vCvsQ_AUoAXoECBAQAw&biw=1366&bih=657#imgrc=Ngk3BqG175ZOLM

[8]:Lin Song,Roxanne Lieb, “Recidivism: The Effect of Incarceration and Length of Time Served”

[9]: Pamela K. Lattimore Joanna R.Baker “The impact of recidivism and capacity on prison populations” Springer(1992)

Available: https://link.springer.com/article/10.1007/BF01066744

[10]: Nazak Dadashazar “Offender Recidivism: A Quantitative Study of Risk Factors and Counseling ” Walden University ScholarWorks(2017).

[11]: Claus Weihs,Katja Ickstadt, “Data Science: the impact of statistics” International Journal of Data Science and Analytics(2017)

[12]: Dr. Raja Sambandham”Application of Data Science in marketing through Big Data” IEEE(2015)

[13]: 5 Steps of the Data Analysis Process [Image]

Source:https://www.google.com/search?q=define+why+you+need+data+analysis&tbm=isch&ved=2ahUKEwiLqeaesdboAhXOK7cAHX-PCE0Q2-cCegQIABAA&oq=define+why+you+need+data+analysis&gs_lcp=CgNpbWcQAzoECAAQQzoCCAA6BQgAEIMBOgYIABAIEB46BAgAEBhQ78MBWM_uAWDt7wFoAHAAeACAAe0BiAH3IpIBBjcuMjIuNJgBAKABAaoBC2d3cy13aXotaW1n&sclient=img&ei=Q3uMXsv7Gc7X3LUP_56i6AQ&bih=754&biw=1536#imgrc=DZzeez2hTpCh1M

[14]: Schwartz Liam “Data,Data Science and Research University” IEEE(2016)

[15]:Understanding Random Forest - Towards Data Science (2020, March 12)

[16]: Decision Tree Classification in Python -datacamp (2020, March 12). [Online]

[17]: Gini Index For Decision Trees - Quantinsti.com (2020, March 13).

[18]: Somayeh Shojaee, Fatimah Sidi, Aida Mustapha, Marzanah A. Jabar “A study on Classification Learning Algorithms to predict Crime status” International Journal of Digital Content Technology and its applications(2013)

[19]: Jitendra Kumar Jaiswal; Rita Samikannu, “Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression” IEEE(2017)

[20]: Julia Andre, Luis Ceferino, Thomas Trinelle “Prediction algorithm for crime recidivism ” IEEE(2017)

[21]: Random Forest Explained (Image) (2020, March 14).

[22]: KNN Model-Based Approach in Classification - Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer.

Available : https://link.springer.com/chapter/10.1007/978-3-540-39964-3_62

[23]: Machine Learning Basics with the K-Nearest Neighbors Algorithm - Choosing the right value of k
Available :
https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761

[24]: Efficient kNN Classification With Different Numbers of Nearest Neighbors - Shichao Zhang, Senior Member, IEEE, Xuelong Li, Fellow, IEEE, Ming Zong, Xiaofeng Zhu, and Ruili Wang.
Available : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7898482

[25]: The distance function effect on k-nearest neighbor classification for medical datasets - Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, Chih-Fong Tsai.

Available : https://springerplus.springeropen.com/articles/10.1186/s40064-016-2941-7

[26]: Joanne Peng “An Introduction to Logistic Regression Analysis and Reporting” The Journal of Education Research(2002)

[27]: Park, Hyeoun-Ae “An Introduction to Logistic Regression: From Basic Concepts to Interpretation with Particular Attention to Nursing Domain”J Korean Acad Nurs Vol.43 No.2

[28]: Sigmoid function (Image)

[29]: Gupta, Isha & Yadav, Dr Raj. (2015). CONCEPT OF RECIDIVISM IN INDIA. Plebs Journal of Law. 1. 240-257.

Available - https://www.researchgate.net/publication/311922915_CONCEPT_OF_RECIDIVISM_IN_INDIA

[30]: James Bernard, Katie Haas, Brian Siler and Georgie Ann Weatherby, “Perceptions of Rehabilitation and Retribution in the Criminal Justice System: A Comparison of Public Opinion and Previous Literature” Journal of Forensic Science and Criminal Investigation(2017)

Available: https://juniperpublishers.com/jfsci/pdf/JFSCI.MS.ID.555669

[31]: UNITED NATIONS OFFICE ON DRUGS AND CRIME”Introductory Handbook on the Prevention of Recidivism and the Social Reintegration of Offenders”

Available:https://www.unodc.org/documents/justice-and-prison-reform/18-02303_ebook.pdf