Classification of Criminal Recidivism Using Machine Learning Techniques

Article · June 2020

CITATIONS
0

READS
144

4 authors, including:

Pratik Kanani
Dwarkadas J. Sanghvi College of Engineering
45 PUBLICATIONS **31** CITATIONS

SEE PROFILE

International Journal of Advanced Science and Technology
Vol. 29, No.4, (2020), pp. 5110 – 5122

Classification of Criminal Recidivism Using Machine Learning Techniques

Heeket Mehta[1*], Shanay Shah[2*], Neil Patel[3*], Pratik Kanani[4*]

[1,2,3,4] *Department of Computer Engineering, Dwarkadas J. Sanghvi College of Engineering, University of Mumbai, Maharashtra, India.*

** All authors have contributed equally.*
[1]*heeketmehta@gmail.com*, [2]*shanayshah131@gmail.com*, [3]*neilpatel1222@gmail.com*, [4]*pratikkanani123@gmail.com*

Abstract

There are numerous cases in the recent times, where a criminal commits a crime, immediately after being granted parole, this is called Criminal Recidivism. The act of recidivism poses a great threat to the society and thus needs to be checked. This paper posits a machine learning approach to detect and predict the tendency of a criminal to commit recidivism. The proposed system helps classify the criminals into Low, Medium, and High risk of committing recidivism. Features like 'Ethnic code', 'Marital Status', 'Age', 'Sex Code', 'Legal Status' and many more are considered while training the model on the dataset. Supervised Classification Algorithms are implemented, and voting is subsequently done, to select the algorithm with the highest accuracy. The Random Forest Algorithm provides the highest accuracy score followed by KNN and lastly Logistic Regression. Moreover, the data is analyzed using visualization charts, where various attributes are deeply analyzed in relation to the target variable 'Score Text'. Graphs between these attributes and the target variable highlight trends, which may provide useful insights to parole granting authorities while assessing a criminal for parole. Stratified K-Fold Cross Validation is used to bolster the results of the algorithms, which gives us accuracy score similar to the above algorithms. Thus, it validates and renders the algorithms unbiased and fair.

Keywords: *Criminal Recidivism, Random Forest Algorithm, K-Nearest Neighbor, Logistic Regression, Isometric Feature Mapping (ISOMAP), Stratified K-Fold Cross Validation, Risk Assessment Instrument (RAI).*

1. Introduction

A convicted offender is statistically very much likely to commit an offence after being granted a clement parole or bail. When the above phenomenon occurs, it is termed as Criminal Recidivism. Criminal Recidivism is a very widely spread and ubiquitously occurring phenomenon across the globe, which must be mitigated to ensure harmony and peace in the society. Criminal recidivism can be eschewed by implementing some Risk Assessment Instruments, to ensure that potential recidivists, do not get granted a parole. Criminals with a high probability of turning into recidivists must be identified by certain means and abstained from being granted parole. Filtering and classifying indicted criminals into categories of potential recidivists helps the authorities mitigate recidivism and curb the increasing crime rate.

The system proposed below provides a machine learning based solution to classify the criminals based on historic trends by evaluating on various parameters and thus swaying the decision, based on the outcome. To train the machine learning model, the dataset employed contains around 18,000 records about criminals with various traits. Machine learning models enables the system to classify based on the processed data and provides

output, which can be extrapolated to various cases. A detailed comparison between the results of three Machine Learning Models has been achieved. Data analysis and Data Visualization on certain important features have been performed on the associated dataset such as the age group involved, marital status, ethnic code, etc. A small description on how the ML models work along with the developed system. Stratified K-Fold Cross Validation is performed on all above-mentioned ML algorithms. Therefore, this model can serve as a RAI (Risk Assessment Instrument) to classify the criminals based on their tendency to commit recidivism.

The paper is structured as follows: In section 3, the proposed methodology has been explained followed by Section 4, where we have provided a detailed account about the data pre-processing and cleaning steps. Further, the dataset was cleaned to bring uniformity. In section 5, we have elaborated on the machine learning models like Random Forest classifier, K-Nearest Neighbor and Logistic Regression and discussed about the biases while implementing the models. Section 6 provides a visual understanding of the data using statistical and graphical representation. The results of implementing the classification algorithms have been depicted in the section 7, along with comparison among all three algorithms. Stratified K-Fold cross validation has also been implemented and explained in the section 7. Finally, the conclusion and future scope have been discussed subsequently in the paper.

2. Literature Survey

Criminal Recidivism is an issue faced across the globe, where acquitted criminals tend to commit an offense again. There are various systems and software developed, to serve the purpose of stemming this phenomenon. Many models are proposed, to deal with various specific cases like substance abuse, sexual offenses, offenses by mentally disordered offenders and numerous other factors. The aim of this paper is to carry out a generalized approach for classifying the criminals as potential recidivists, with a holistic approach and not limiting it to specific cases like Substance Abuse, etc. General traits and characteristics are considered for our model, without any restrictions on age, sex, ethnicity and many such attributes [1].

M.O. Franke et al. presented a research paper involving prediction of general criminal recidivism for mentally disordered offenders. The authors posit a novel approach to risk assessment of potential recidivists, based on mentally unstable offenders. The approach takes into consideration 259 cases, to gauge and determine risk factors of potential recidivism. This was carried out using the Random Forest algorithm. This research influenced our paper, to conduct a similar assessment, but on a more generalized data, not only on the data of mentally disordered offenders [2].

J. Rossegger A. et al. proposes the research aimed at examining the sensitivity of prognostic models to the test data sample. The number of cases recorded and analyzed was 773, using the bivariate logistic regression algorithm. The authors contend that models are very sensitive to the features and calibration samples. The inference drawn by them suggests that given some of the attributes to be the same for any case, a different value of another attribute may alter the result, during risk assessment. The bivariate logistic regression used, provided us with an insight to use the same for our model. Hence we involve a plethora of features in training our model and also tried logistic regression on it [3].

Steinhausen, H. et al. present a novel measure of research on juvenile male substance users, for criminal recidivism. The paper sheds light on the effects of substance abuse on juvenile males, subsequently contributing to criminal recidivism. Cox Regression played a paramount role to predict the behavior of offenders. The paper focuses specifically on juvenile males, subject to substance abuse, which limits the features of the model. Thus, we incorporate a more openly admissible approach, which is not restricted by fixating a gender,

age or any specific cause. Thus, a more heuristic approach, with less constraints, serves to provide a better analysis for a wider range of criminals [4].

Dr Raj Yadav et al. contend with facts regarding recidivism in various countries. The paper posited by them gives an overview of the factors, characteristics, causes and various other terms related to criminal recidivism. One of the causes mentioned is the shortcoming in law enforcement system. To eliminate/abate the cause we propose a system, where criminals would be classified into categories of potential recidivists, using machine learning algorithms [5].

P. Wang et al. employs the Support Vector Machine Algorithm to predict criminal recidivism in their research. The implementation of a classification algorithm influences our model, to employ supervised classification algorithms like Random Forest Algorithm, K-Nearest Neighbors and other such algorithms [6]. According to the paper 'Concepts of Recidivism in India' [18], 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.

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 analyzed parole/bail hearing and granting process.

3. Proposed Methodology

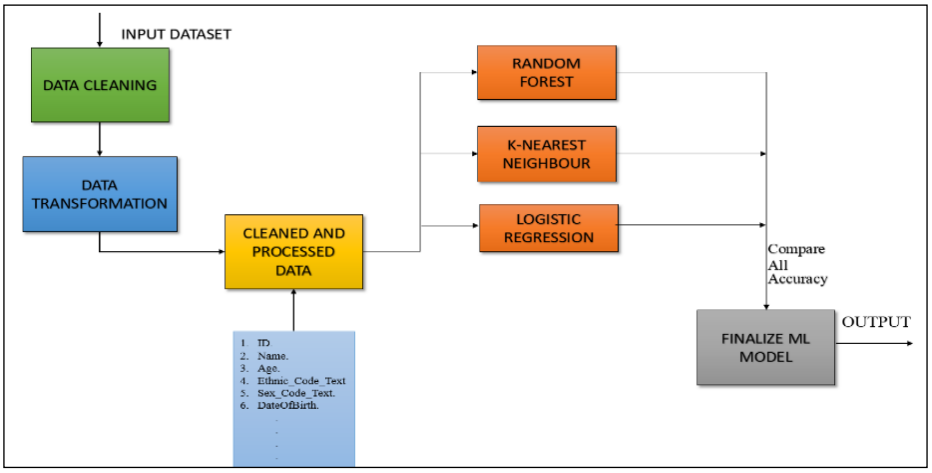

Figure 1. System Architecture

In the above system, the dataset used is pre-processed, cleaned and transformed using various data mining techniques. Further, the result of this phase gives us a cleaned compatible dataset free of anomalies which can be fed into the next phase where Machine Learning Algorithms would be applied onto the processed dataset. The next phase entails Machine Learning Implementation where various Supervised Classification Algorithms are employed to classify the records of criminals into three categories namely Low, Medium and High. We train our models using three important classification algorithms namely Random Forest, K-Nearest Neighbor and Logistic Regression. Subsequently, Voting takes place to determine which algorithm provides us with the best outcome. This process of Voting is based on the Accuracy Score provided by each algorithm and considering the one with the best accuracy score to be the ideal algorithm for this system. Now, we finalize the

model with the highest accuracy and given any tuple for a criminal record the finalized algorithm fetches the best results of classifying the criminal into the respective target categories.

4. Data Science and Preprocessing

4.1 Understanding the Dataset

The dataset considered for implementation is from Carnegie Mellon University (CMU). In this phase, we understood the features of the dataset and the type of data present in the dataset.
The Dataset contains approximately 60000 tuples consisting of around 22 attributes. It contains the records of criminals in the United States of America, with their various traits and characteristics which may help us classify them as recidivists.
Moreover, we also checked for the uniformity of the dataset to decide the performance parameter for the Machine Learning Model.

4.2 Elimination of Null Values

The initial dataset consisted of three tuples with null values throughout all attributes. These three tuples were dropped using Pandas and Numpy libraries of Python.
Thus, out of 60000 records dropping three tuples full of Null values would negligibly affect the performance of Machine Learning Models. Thus, we drop and eliminate these three tuples. Other than three tuples, the data set consisted of few null values which were replaced by the mean value of that specific column.

4.3 Elimination of Duplicate Data

The Dataset contained multiple duplicate values for all tuples. Hence, to avoid overfitting we had to eliminate all repeated records. This scaled down the dataset from 60000 repeated to 18000 unique records.

4.4 Label Encoding

The dataset contains various attributes with values in the string datatype. The String datatype is not compatible with the Machine Learning Models. To convert the data into compatible format (Integer, Float), we perform Label Encoding Technique. In the associated dataset, label encoding is done manually to reduce the biases on the dataset which is one of the most important factors that needs to be take care of.

4.5 Features Taken into Consideration

The dataset taken into consideration consists of 20 features. In total, 16 important features are considered out of 20 for the implementation of prediction of the models. The features mainly considered are ‘Ethnic code’, ‘Marital Status’, ‘Age’, ‘Sex Code’, ‘Legal Status’ and many more are considered while training the model on the dataset. The target variable is ‘Score_Text’ with values low, medium and high.
The table given below describes the integer type attributes of the dataset, providing with statistical values like Mean, Standard Deviation, Minimum Value, and Maximum Value. To explicate this, let’s consider the attribute ‘Age’, we observe that the Mean age of the criminals is between 34 and 35 years. The Standard Deviation is 12.20 years. The Youngest criminal in the dataset is 16 years whereas, the oldest criminal is 84 years old. Likewise, inferences can be made and extrapolated to other such attributes as well.

ISSN: 2005-4238 IJAST

Table 1. Summary Table of Various Attributes

Features	Mean	Standard Deviation	Min	Max
Assessment Id	67858	7219.64	1310	79677
Decile Score	3.20	2.38	1	10
RecSupervision Level	1.59	0.91	1	4
Age	34.31	12.20	16	84
Scale Id	7.00	0.00	7	7
Scaleset Id	21.91	0.68	17	22

5. Supervised Classification Algorithms

5.1 Random Forest Algorithm

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 class of instances where it might belong. The Random Forest algorithms is a huge collection and cluster of numerous decision trees.

Decision Trees: 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 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 [7]. To construct the decision tree, to predict the outcome of a single tree, we can use the Gini Impurity or the Entropy.

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

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.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (2)$$

$$\text{Information Gain (Parent, Child)} = \text{Entropy(parent)} - [p1(c1)*\text{entropy}(c1) + p2(c2)*\text{entropy}(c2)+...] \quad (3)$$

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 [8]. This

ISSN: 2005-4238 IJAST

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 [9, 10].

5.2 K-Nearest Neighbor Algorithm

The KNN Algorithm of Supervised Machine learning stands for K-Nearest Neighbors. 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 neighbors to tuple N. The classification of the tuple N is dependent on the weight-based factors (distance of the tuple from the neighbors) and is classified into the neighborhood with minimum associated weight [11].
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 [12]. For instance, the distance of attribute/tuple xj, is calculated from each of the 3 neighbors (k=3). The neighbors - 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 [13].

$$\text{Distance (A,B)} = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}} \quad (4)$$

5.3 Logistic Regression Algorithm

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 (failure and success). It requires iterative methods to fit a logistic regression model and we use the maximum likelihood function to maximize 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 analyses the probability of the event concerned [14]. 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. 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 [15]. A common example of sigmoid function is the logistic function shown in the figure.
The expression of sigmoid function is:

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} \quad (5)$$

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

5.4 Application of Machine Learning for 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.

5.5 Biases in Machine Learning Algorithms

A bias can be encountered in any Machine Learning model, this basically may influence the outcome generated by the machine learning model [16]. Biases in models must be removed since they provide us with an impartial outcome. A bias-free machine learning model cannot exist as it requires a certain amount of bias to model the data and to analyze predictions. However, the aim is to reduce these biases occurring in our model. In the case of training models for Criminal Recidivism, there can be various bias which may be encountered. To explicate this, some of the biases may be against certain races, where people of a particular race may be impartially evaluated for gauging the recidivism score. Another such bias may occur in the gender of any offender, where a person of a particular gender may be more biased/likely to be categorized as a recidivist.
There can likely be the following biases -
1) Sample Bias - This is an inevitable type of bias arising due to the randomness and irregularities in the data samples. This occurs during the training phase of the model. This can be an intuitive way of the model to pick up the more frequently occurring values of a particular attribute [17]. For instance, in the dataset used to train our model, the number of cases, where the tuple has value "Single" in the "Marital Status" column is much higher than other values (shown in fig. 4). This inevitably creates an occurrence of sample bias, where the model would be inclined to categorize most of the "single" values to higher recidivism.
2) Algorithmic Bias - This is the type of bias which is introduced by the algorithmic phase of the machine learning model and is not present due to anomalies in data samples. Data Scientists strive to attain a perfect balance between high variance and high bias. Here, in our model, the Random Forest Classifier introduces bias, when training and testing the given dataset. This is an inherent property of the algorithm.
3) Measurement Bias - It occurs when we select the features we wish to incorporate in the model. It may be the way these features/attributes are used in the machine learning phase [17]. A striking example of this, is the use of this for criminal recidivism, where any priory committed crimes or crimes committed by relatives/friends may also taint the outcome of the model. Thus, attributes like "Agency Type", "Custody Status", "Legal Status", etc. may create a measurement bias for criminals in evaluating their score text.
4) Prejudice Bias - This type of bias is mainly due to the influence of social stereotypes and orthodox opinions. It mainly occurs on training data, where prejudice against a particular culture, gender, ethnicity or any such factor may make the model biased while generating the output. If the algorithm is exposed to a more even-handed data distribution, then the statistical relationship between such potentially prejudiced attributes can be avoided.

ISSN: 2005-4238 IJAST

6. Data Visualization and Graphical Representation

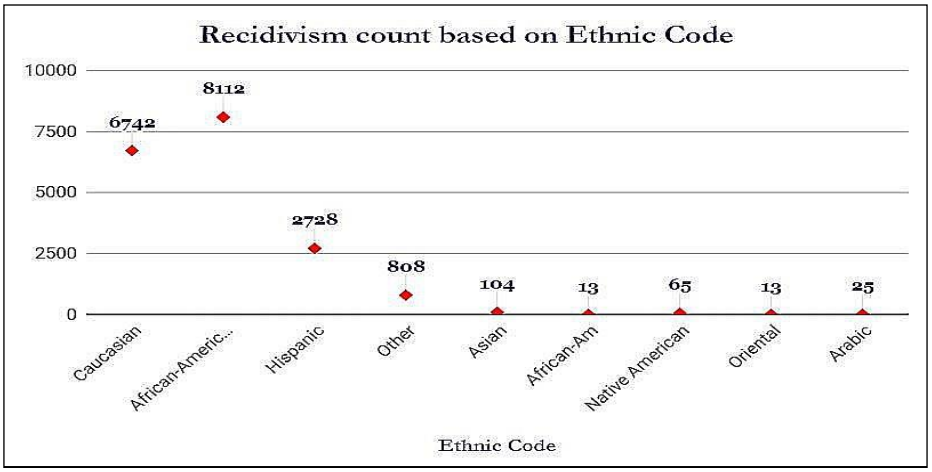

Figure 2. Recidivism Count vs. Ethnic Code

In Figure. 2 we observe that criminals of the ethnic code African-American are most likely to be recidivists followed by Caucasian, Hispanic, Asian and other such ethnic groups. Thus, the jury in-charge of granting parole or bail must be very careful when dealing with offenders belonging to the above groups. These studies are purely based on trend and number of recidivists belonging to respective ethnic codes. The number of African-American offenders are 8112 out of a total of 18,598 cases which is about 43% of the total recidivism cases. The Caucasian Ethnic Group has the second largest percentage which is about 36% of the total.

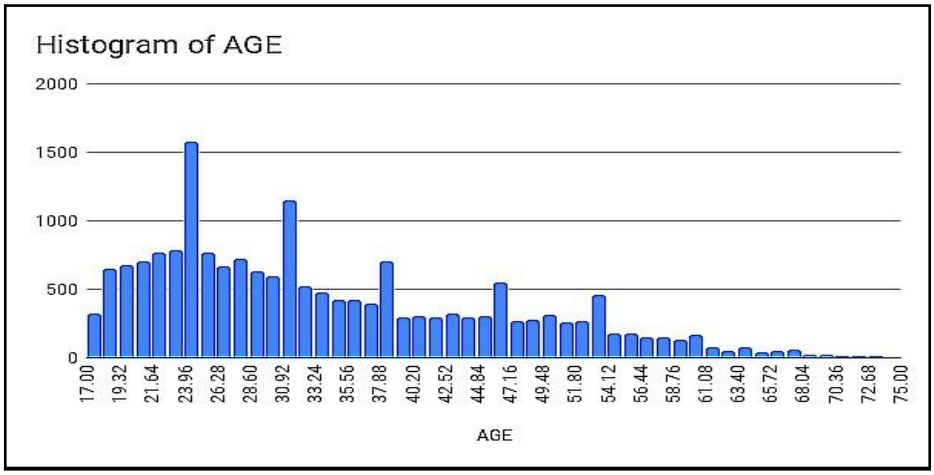

Figure 3. Recidivism Count vs. Age

In the above histogram, we observe that the age where the approximate age of most recidivists is around 23-24 years old. This is followed by 30-31 years of age and subsequently 37-38 years. It is observed that youngsters and fledglings are much more likely to commit offenses again rather than the more mature counterpart (around 40+ years of age). A general trend can be inferred that as the age increases the number of recidivism cases decreases. The above graph also documents juvenile cases where age is around 17-19 years.

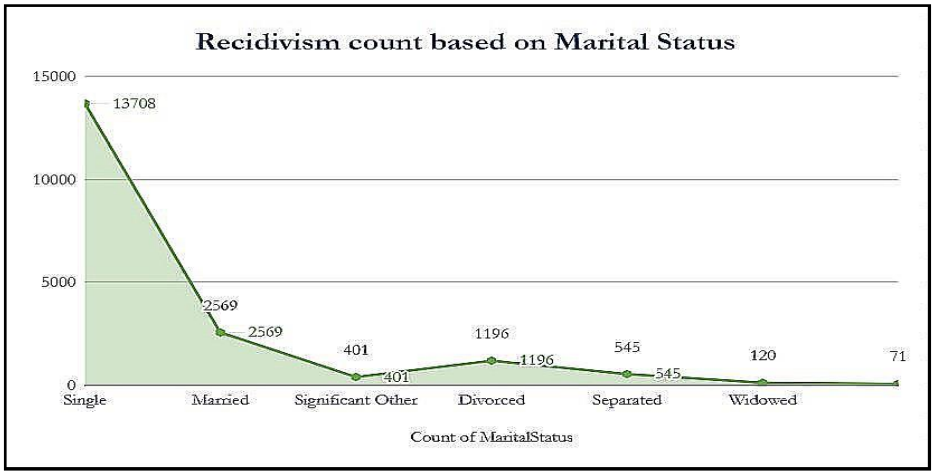

Figure 4. Recidivism Rate vs. Marital Status

The above Line Graph depicts the trend of marital status corresponding to the recidivism rate. It is evident from the above visualization that criminals who were single have a higher tendency of committing recidivism. Criminals who were single accounted for a striking 73.70 % of the total cases. It is also seen that this trend was followed by criminals who were married, they accounted for 13.49 % of the total cases. In this way a jury would judge and relate the marital status with the tendency of the criminal to commit recidivism.

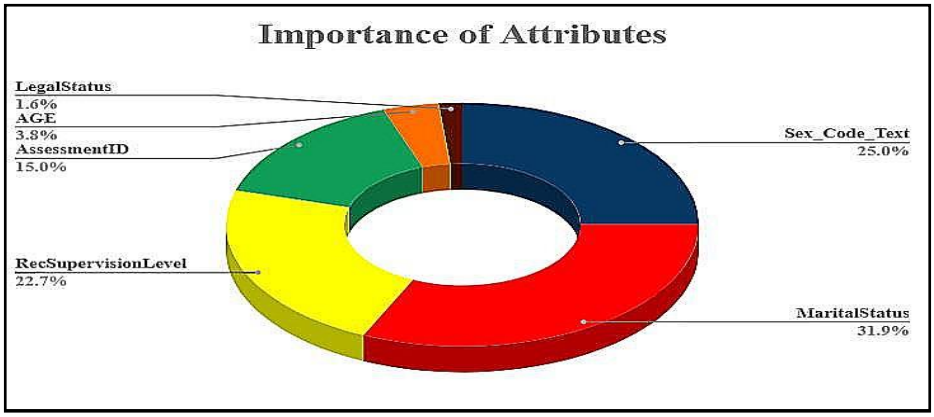

Figure 5. Importance of Attributes

The Pie chart displayed above represents the six most important features taken into consideration during the implementation of the Random Forest Algorithm. These 6 features play a vital role when the data is trained and tested against the target variable 'ScoreText'. On keen evaluation, we can observe that 'MaritalStatus' accounts for the largest percentage (31.9%). Hence, the MaritalStatus influences and sways the output of the random forest algorithm the most. The second most important feature to influence the output is 'Sex_Code_Text' (25%), followed by 'RecSupervisionLevel' (22.7%). The 3 other important features which affect the outcome are 'AssessmentID' (15%), 'AGE' (3.8%) and 'LegalStatus' (1.6%).

ISSN: 2005-4238 IJAST

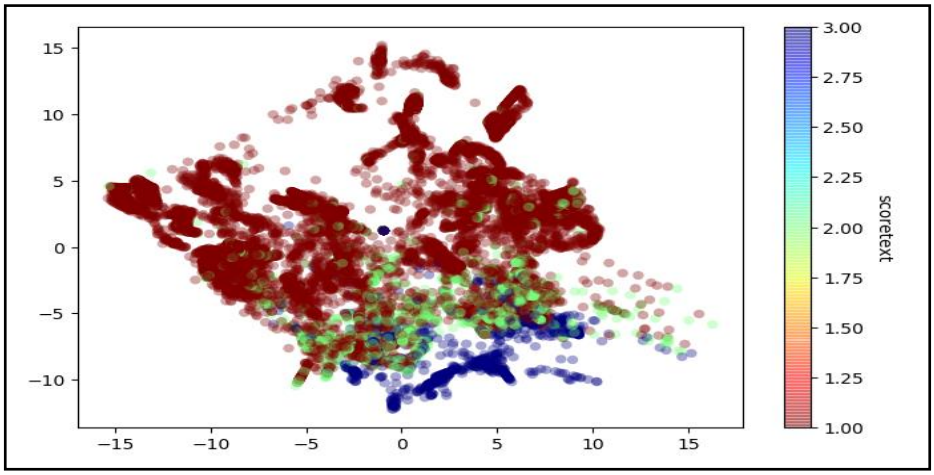

Figure 6. Isometric Feaure Mapping (ISOMAP)

ISOMAP is a non-linear dimensionality reduction method applied on high-dimensional data. This technique helps preserve Euclidean distances while transforming from a larger to smaller metric space. The distance over the manifold is larger than the neighboring data points, which may consider two data points as near points. ISOMAP overcomes the issue, by using a pairwise geodesic distance between the data points. The geodesic distance is calculated by using various algorithms like Dijkstra's Shortest Path algorithm. In the above figure, we have taken into consideration some important features like 'MaritalStatus', 'LegalStatus', 'Sex_Code_Text', 'RecSupervisionLevel', 'AGE', 'AssessmentID' and mapped them against the target variable 'ScoreText' [19].

7. Results and Analysis

7.1 Confusion Matrix and Cross Tab Analysis

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.

7.1.1 Confusion Matrix and Cross Tab for K-Nearest Neighbor Algorithm

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

Figure 7. Confusion Matrix for K-NN

$$\text{Accuracy} = \frac{Total\ number\ of\ Correctly\ \ Classified\ Tuples}{Total\ Number\ of\ Tuples}$$

$$\text{Accuracy} = \frac{313+3894+642}{5581} = \frac{4849}{5581} = 0.8688$$

Accuracy Score = 0.8688 * 100 (percent) = 86.88% (6)

7.1.2 Confusion Matrix and Cross Tab for Logistic Regression Algorithm

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

Figure 8. Confusion Matrix for Logistic Regression

Accuracy = $\frac{0+4175+27}{5581} = \frac{4202}{5581}$ = 0.7529

Accuracy Score = 0.7529*100 = 75.29% (7)

7.1.3 Confusion Matrix and Cross Tab for Random Forest Algorithm

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

Figure 9. Confusion Matrix for Random Forest

Accuracy = $\frac{315+3949+637}{5581} = \frac{4901}{5581}$ = 0.8781

Accuracy Score = 0.8781*100 = 87.81% (8)

7.2 Comparison Among Classification Algorithms

Table 2. Comparison Among Algorithms

Parameters	Random Forest	Logistic Regression	K-Nearest Neighbors
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 %

7.3 Stratified K-Fold Cross Validation

Table 3. Results of Stratified K-Fold Cross Validation

ISSN: 2005-4238 IJAST

No. of Folds	Random Forest	K-Nearest Neighbor	Logistic Regression
K = 2 (Leave One Out)	87.13 %	85.21 %	74.15 %
K = 5	86.93 %	85.77 %	74.07 %
K = 10	86.99 %	85.96 %	74.63 %
K = 15	86.88 %	85.97 %	74.68 %
K = 20	86.91 %	85.90 %	74.30 %

Stratified K-Fold Cross Validation is employed to bolster and corroborate our findings. Here, the folds are selected based on the mean response value in a way that it is similar in all folds. This technique ensures that each fold represents all strata of the dataset so that irrespective of the training and testing sample, the results of the algorithm are unbiased and uniformly calculated. The above table represents the accuracy obtained by implementing the Stratified K-Fold Cross Validation Technique on the training dataset using each algorithm, for various values of k. The Accuracy obtained is almost equal to the accuracy obtained by running the individual algorithms on the entire dataset. This similarity of accuracy infers that the algorithms yield thorough and cogent results irrespective of the training and testing samples. Different Folds (values) of k split the data into k-1 folds for training and the remaining for testing. The Accuracy obtained by each algorithm, being similar to equation 7, 8 conform our findings and renders our research cogent and reliable.

8. Conclusion

The above model designed to categorize 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. 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 parole to the criminals, who might be potential recidivists. An accuracy of 87.81% obtained by Random Forest, 86.88 % by K-nearest Neighbors algorithm and 75.29% by Logistic Regression helps classify the criminals and this technique, when implemented in parole hearings, should result in decrease in the crime rate. Based on the accuracy score, implementing a Random Forest Classifier would provide the best outcome, for any criminal with given set of characteristics.

Future Scope: The above research methodology system should be extrapolated for more localized prisons/facilities so that the results are more specific and pertaining to the respective location. Moreover, the research should be conducted for a longer period of time, thus generating more records of criminals to obtain better accuracy. The proposed model should help in reducing criminal recidivism in the upcoming years. The results of this research should also be merged with the oral statements provided by criminal on day of trial to judge for more authenticity using Natural Language Processing techniques. Extensive analysis should be carried out locally and subsequently on a larger scale to gain insightful trends. This system should help the parole granting authorities to abstain from granting parole to criminals with a medium to high risk of recidivism.

ISSN: 2005-4238 IJAST

9. References

[1] Rhodes, W. Predicting criminal recidivism: A research note. J Exp Criminal 7, 57–71 (2011).
[2] Pflueger, M.O., Franke, I., Graf, M. et al. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. BMC Psychiatry 15, 62 (2015).
[3] Urbaniok, F., Endrass, J., Rossegger, A. et al. The prediction of criminal recidivism. Eur Arch Psychiatry Clin Neurosci 257, 129–134 (2007).
[4] Aebi, M., Bessler, C. & Steinhausen, H. A Cumulative Substance Use Score as a Novel Measure to Predict Risk of Criminal Recidivism in Forensic Juvenile Male Outpatients. Child Psychiatry Hum Dev (2020).
[5] Gupta, Isha & Yadav, Dr Raj. (2015). CONCEPT OF RECIDIVISM IN INDIA. Plebs Journal of Law. 1. 240-257.
[6] P. Wang, R. Mathieu, J. Ke and H. J. Cai, "Predicting Criminal Recidivism with Support Vector Machine," 2010 International Conference on Management and Service Science, Wuhan, 2010, pp. 1-9.
[7] Decision Tree Classification in Python -data camp (2020, March 12).
[8] Gini Index for Decision Trees - Quantinsti.com (2020, March 13).
[9] 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).
[10] Jitendra Kumar Jaiswal; Rita Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression" IEEE(2017).
[11] Julia Andre, Luis Ceferino, Thomas Trinelle "Prediction algorithm for crime recidivism " IEEE(2017).
[12] KNN Model-Based Approach in Classification - Gongde Guo, Hui Wang, David Bell, Yaxin Bi, Kieran Greer.
[13] Machine Learning Basics with the K-Nearest Neighbors Algorithm - Choosing the right value of k.
[14] 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.
[15] Joanne Peng "An Introduction to Logistic Regression Analysis and Reporting" The Journal of Education Research(2002).
[16] Gordon, Diana & desJardins, Marie. (1995). Evaluation and Selection of Biases in Machine Learning. Machine Learning. 20. 5-22. 10.1007/BF00993472.
[17] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." ArXiv abs/1908.09635 (2019): n. pag.
[18] Gupta, Isha & Yadav, Dr Raj. (2015). Concept of Recidivism in India. Plebs Journal of Law. 1. 240-257.
[19] Parekh, Vishwa & Jacobs, Jeremy & Jacobs, Michael. (2014). Unsupervised Non Linear Dimensionality Reduction Machine Learning methods applied to Multiparametric MRI in cerebral ischemia: Preliminary Results. Progress in Biomedical Optics and Imaging - Proceedings of SPIE. 9034. 90342O. 10.1117/12.2044001.
[20] United Nations Office On Drugs And Crime "Introductory Handbook on the Prevention of Recidivism and the Social Reintegration of Offenders".
[21] 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.
[22] Understanding Random Forest - Towards Data Science (2020, March 12).
[23] The distance function effect on k-nearest neighbor classification for medical datasets - Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, Chih-Fong Tsai.
[24] 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).

ISSN: 2005-4238 IJAST