Bioscene

Bioscene
Volume- 21 Number- 03
ISSN: 1539-2422 (P) 2055-1583 (O)
www.explorebioscene.com

Novel Stacking Model for Clarifying Novel Relationship between Offenders and the Victims of Rape in India

Ria Pyne[1]; Suman Maji[2]; Avijit Kumar Chaudhuri[3]

[1,2] Department of Computer science Engineering, Brainware University, Barasat, Kolkata

[3]Associate Professor, Computer Science & Engineering, Brainware University ,Kolkata

Abstract: India, with a population of 1.4076 billion, holds the distinction of being the world's most populous country. Given the size of its population, managing crime rates is a complex task, influenced by a multitude of factors including poverty, peer pressure, drug abuse, politics, religious beliefs, societal background, and unemployment. Rape, in particular, has emerged as a prevalent crime in recent years. The pervasiveness of rape in India can be attributed to gender inequality, societal norms and attitudes, lack of awareness, insufficient law enforcement, a slow judicial process, social stigma, and victim-blaming. In this study, we have compiled a decade's worth of data for each state, categorizing each year into subgroups based on whether the offenders were known or unknown to the victims. The dataset comprises 10 independent features and one dependent feature. We have used various machine learning algorithms to train a model capable of categorizing the data into the aforementioned subgroups. The machine learning algorithms used for this analysis include ensemble algorithms comprising multiple machines learning methods, stacking and bagging of multiple machine learning algorithms both with and without feature selection. The machine learning methods used include decision trees, K-Nearest Neighbors (KNN), logistic regression, and random forests. So far, we have achieved an accuracy of 98% using novel stacking techniques.

Keywords: Data Mining, Machine Learning, Decision Tree, Crime Detection, Crime Prediction, Random Tree, Data Preprocessing,K-Nearest Neighbors (KNN)

Introduction

Rape is a terrible crime that violates basic human rights and continues to be a major problem in India. Over the years, there have been many cases of rape in the country, leading to widespread anger and highlighting the urgent need for comprehensive changes in society, laws, and culture to address this serious violation of human dignity. Understanding the issue of rape in India is complex because it has many causes and effects. To fully grasp the situation, we must carefully examine various factors such as cultural beliefs, legal systems,

economic inequalities, and power imbalances between genders. The number of rape cases reported in India is shockingly high, indicating the widespread nature of this crime across the nation. Unfortunately, many cases still go unreported due to factors like social shame, fear of retaliation, and lack of trust in the justice system. This silence surrounding rape makes it difficult to effectively fight against sexual violence. Rape in India is deeply rooted in patriarchal norms and gender inequalities, often reflecting power struggles and systematic oppression. Women and marginalized groups are disproportionately affected by this violence, showing how it intersects with other forms of discrimination based on factors like caste, class, and ethnicity. The normalization of rape culture in India perpetuates harmful stereotypes and attitudes that blame survivors while letting perpetrators off the hook. This further contributes to the problem and makes it harder to bring justice to survivors.

India's laws regarding rape have undergone significant changes in recent times with amendments aimed at better protecting survivors and ensuring stricter punishment for offenders. However, there are still challenges in implementing these laws effectively, especially when dealing with biases within institutions, limited resources, and delays in court proceedings. The impact of rape goes beyond just the immediate physical and emotional trauma experienced by survivors. It affects their families, communities, and society as a whole, creating a cycle of violence that is difficult to break. In India, perpetrators of rape can be divided into two main categories: those who are known to the victim and those who are unknown. Known offenders usually have some prior relationship or acquaintance with the victim, such as partners, family members, or acquaintances. These cases often involve elements of manipulation, coercion, or breach of trust. On the other hand, unknown offenders are individuals who have no previous connection to the victim and may commit rape through opportunistic or predatory behavior. This can include strangers, serial offenders, or online predators.

Both known and unknown perpetrators present different challenges when it comes to investigating, prosecuting, and supporting survivors. This highlights the importance of having comprehensive strategies in place to effectively address sexual violence in India. To understand whether most rape cases involve known or unknown offenders and clarify the motives behind these crimes, predictive models have been used to analyze the available data. These models, such as stacking, bagging, and ensemble algorithms, were combined with feature selection techniques for better accuracy.

Among these methods, bagging feature selection has shown significant promise in providing accurate predictions for categorizing rape cases into known and unknown offenders.

Literature review

The paper''Crime Prediction using Machine Learning'' explores the application of machine learning algorithms in predicting crime patterns and trends. By leveraging data from 2001 to 2016, the study aims to forecast crime cases from 2017 to 2020 across India. The methodology involves using Linear Regression and Random Forest algorithms to analyze crime data, with a focus on identifying trend-changing years to enhance prediction accuracy. The study employs machine learning techniques, including Linear Regression and Random Forest, to analyze historical crime data. The study highlights the importance of considering demographic and geographic factors in crime prediction models. Additionally, the research demonstrates the potential of using clustering techniques to identify hotspots for targeted law enforcement interventions. **[1]**

The ''Patterns of risk—Using machine learning and structural neuroimaging to identify pedophilic offenders'' explores the potential of machine learning and structural neuroimaging to identify pedophilic offenders (PO) by analyzing diffusion tensor imaging (DTI) data from a cohort of 14 PO individuals and 15 matched healthy control individuals. The study aims to discover neurobiological correlates that could enhance the early detection and risk assessment of PO individuals, potentially contributing to the prevention of child sexual abuse (CSA). **[2]**

The file'' Machine learning-based soft computing regression analysis approach for crime data prediction'' proposes a machine learning-based soft computing regression analysis approach to predict crime data in India amidst the growing crime rates. It utilizes various regression algorithms—Simple Linear Regression (SLR), Multiple Linear Regression (MLR), Decision Tree Regression (DTR), Support Vector Regression (SVR), and Random Forest Regression (RFR)—to analyze Indian Penal Code (IPC) crime counts by region and type. The study concludes that the RFR model offers the best accuracy in predicting regional crime statistics, revealing high potential for data-driven insights in law enforcement and crime prevention strategies. **[3]**

The paper ''An optimized machine learning and big data approach to crime detection'' presents a machine learning-based approach for crime detection and prediction that extracts features like time zones, crime probability, crime hotspots, and vulnerability analysis to improve the accuracy of crime prediction. The proposed feature generation method, which includes time zone classification, crime probability calculation, crime hotspot analysis, and vulnerability analysis, increased the performance of machine learning models for crime detection. The Naïve Bayes algorithm achieved the highest accuracy of 97.5% when applied to the San Francisco dataset using the proposed methodology. The study's unique contributions include the ability to analyze crime patterns across different time zones, predict crime probability for the next day, identify crime hotspots, and

perform vulnerability analysis to pinpoint locations prone to future criminal activities. **[4]**

The paper ‘‘Violence, and Triage: Complainant Identity and Criminal Justice in India’’ provides evidence that women in India face discrimination in accessing the criminal justice system, with their complaints more likely to be delayed and dismissed compared to men's complaints. Women's complaints are more likely to be delayed and dismissed at both the police station and the courthouse compared to men's complaints. - Suspects accused by female complainants are less likely to be convicted and more likely to be acquitted, even when accounting for cases of violence against women (VAW). Contrary to the view of policymakers and judges, VAW cases, including dowry-related abuse, often involve serious forms of violence like starvation, poisoning, and marital rape, rather than just "petty quarrels". The study uses an original dataset of 418,190 crime reports from Haryana, India, which was merged with 251,804 (60.2%) judicial records. The key methods are descriptive and OLS analyses, structural topic modeling, and topical inverse regression matching. **[5]**

The paper ‘‘An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: an ensemble approach. Ieee Access’’ proposes a novel ensemble-based crime prediction method called SBCPM that uses SVM algorithms and achieves 99.5% accuracy, outperforming previous research and being useful for predicting future crimes. The ensemble "assemble-stacking based crime prediction method (SBCPM)" outperformed individual machine learning models in accuracy, correlation, and error metrics. The SBCPM model achieved a 99.5% classification accuracy on the testing data. The SBCPM model was more predictive than previous research on crime datasets focused on violence. Data preprocessing (handling missing values, cleaning, transformation). Using 5 classifier algorithms: J48, SMO, Naive Bayes, Bagging, and Random Forest. Applying Support Vector Machines (SVM) - Using a crossover model combining J48 and C4.5 classifiers. Applying ensemble learning through stacking, using J48 and C4.5 as sub-models and SVM as an aggregator model. **[6]**

This research paper ‘‘Survey on crime analysis and prediction using data mining and machine learning techniques.’’ explores the application of data mining and machine learning techniques in analyzing and predicting crime. The authors present a comprehensive survey of various methodologies, including data pre-processing, feature selection, pattern identification, and classification, highlighting the potential of these techniques for enhancing crime prevention strategies.**[7]**

This research ‘‘Modelling and forecasting gender-based violence through machine learning techniques. Applied Sciences’’ explores the use of machine learning techniques to model and forecast Gender-Based Violence (GBV) in Spain. The authors compiled and prepared a comprehensive database of GBV-

related features spanning over a decade from official sources, then tested various feature selection methods and predictive algorithms. The results demonstrate that predicting the number of GBV complaints presented to a court within a six-month horizon is achievable with a high level of accuracy, particularly when using a Multi-Objective Evolutionary Search Strategy for variable selection and Random Forest as the predictive algorithm.**[8]**

This dissertation'' Stereotypes and phenotypes: using machine learning to examine racial implicit bias in sex offender criminal case processing'' examines racial/ethnic disparities in sentencing outcomes among publicly registered sex offenders in New York. It draws on two theoretical perspectives, an uncertainty and causal attributional approach and a normal crimes approach, to predict that offenders with Afrocentric or Hispanic facial features will receive harsher sentences. **[9]**

The paper'' A graph-based clustering approach for relation extraction from crime data'' proposes a graph-based clustering technique to extract relations between named entities in a corpus of crime reports against women in India, in order to identify crime patterns and insights that can aid criminal investigations and the criminal justice industry. The proposed graph-based clustering technique can extract relations between entities in a crime dataset, which can help analyze crime patterns and aid criminal investigations. The extracted relations and clusters are evaluated using various internal and external cluster validation metrics, and the proposed method is compared to other existing relation extraction techniques. Choosing entity pairs from three domains (PER-PER, PER-LOC, ORG-PER) for analysis. Measuring the similarity between entity pairs based on their intermediate context words. Constructing a weighted undirected graph where nodes are entity pairs and edge weights are similarity scores. Partitioning the graph into subgraphs based on a threshold (average edge weight), and further partitioning the subgraphs in an iterative, hierarchical manner. Using a cluster validation index (Score Function) to evaluate the quality of the partitions and continue the process as long as the quality improves. **[10]**

Author Name	Field of study
Rathod, S., & Talari, S. (2017).	Understanding Rape: A Criminological Study From the Perspective of Rape Offenders.[11]
Salunke, P. (2016, September 1)	In 99.3% rape cases, accused known to survivor: NCRB.'' Hindustan Times [12]
Plummer, M., & Cossins, A. (2016)	The cycle of abuse: when victims become offenders.'' Trauma,

	Violence & Abuse, 19(3), 286–304. [13]
Chaudhary, T. S.	Violence Against Women in Delhi (analyzing the Nature, Time, Place, Age and Relationship Between a Rape Victim and Offender [14]
Mitchell, D., Angelone, D., Kohlberger, B., & Hirschman, R. (2008)	Effects of offender motivation, victim gender, and participant gender on perceptions of rape victims and offenders." Journal of Interpersonal Violence, 24(9), 1564–1578. [15]
Greenfeld, L. A. (1997)	Sex offenses and offenders: An Analysis of Data on Rape and Sexual Assault: Executive Summary. [16]
Gudjonsson, G. H., & Sigurdsson, J. F. (2000)	Differences and similarities between violent offenders and sex offenders." Child abuse & neglect, 24(3), 363-372.[17]
Gartner, R., & Macmillan, R. (1995)	The effect of victim-offender relationship on reporting crimes of violence against women." Canadian Journal of Criminology, 37(3), 393–429 [18]
Gartner, R., & Macmillan, R. (1995b)	The effect of victim-offender relationship on reporting crimes of violence against women." Canadian Journal of Criminology, 37(3), 393–429 [19]
Ullman, S. E., & Siegel, J. M. (1993)	Victim-Offender relationship and sexual assault [20]

Material and methods

Data preparation

The dataset for this study was sourced from Kaggle. It encompasses data from 28 states and 8 union territories, with each state and union territory providing data spanning a decade from 2001 to 2010. The dataset includes features such as Area_Name, Year, Rape_Cases_Reported, Victims_Above_50_Yrs, Victims_Between_10-14_Yrs, Victims_Between_14-18_Yrs, Victims_Between_18-30_Yrs, and Victims_Upto_10_Yrs. The dependent variable in this dataset is the 'Subgroup', which we aim to classify into two categories: incest rape and others.

Certain modifications have been made to the dataset. For instance, under the 'Subgroup' category, instances of incest rape have been recoded as '1', while all other instances have been recoded as '0'. Additionally, the 'Area_Name' values have been transformed into numerical values for ease of analysis. The specifics of this transformation are as follows:

Table 1. Every state is identified as numbers

States	Number
Andaman Nicobar	1
Andhra Pradesh	2
Arunachal Pradesh	3
Assam	4
Bihar	5
Chandigarh	6
Chattisgarh	7
Dadara & Nagar Haveli	8
Daman & Diu	9
Delhi	10
Goa	11
Gujarat	12
Haryana	13
Himachal Pradesh	14
Jammu & Kashmir	15
Jharkhand	16
Karnataka	17
Kerala	18
Lakshadweep	19
Madhya Pradesh	20
Maharashtra	21
Manipur	22
Meghalaya	23
Mizoram	24
Nagaland	25
Odisha	26
Puducherry	27
Punjab	28
Rajasthan	29
Sikkim	30
Tamil Nadu	31
Tripura	32
Uttar Pradesh	33

Uttarakhand	34
West Bengal	35

After making above changes, the new dataset inserted in Weka to convert the data into true and false, then the weka generated data is been used in the stacking code mentioned below.
From the dataset, year also have been removed as it doesnot hold any value to our objective and also providing less accuracy.

Data flow

- The Dataset has been divided into different percentage of train-test split and 10-fold cross validation.
- Each train-test split and 10 cross validation data has been passed through the classifier.
- The various classifier has been used on data set for building a prediction model that wholly dependent on evaluation of provided the information of various metrics and statistical test.
- The classifier that has given best results from all the classifier will be taken as final prediction model.

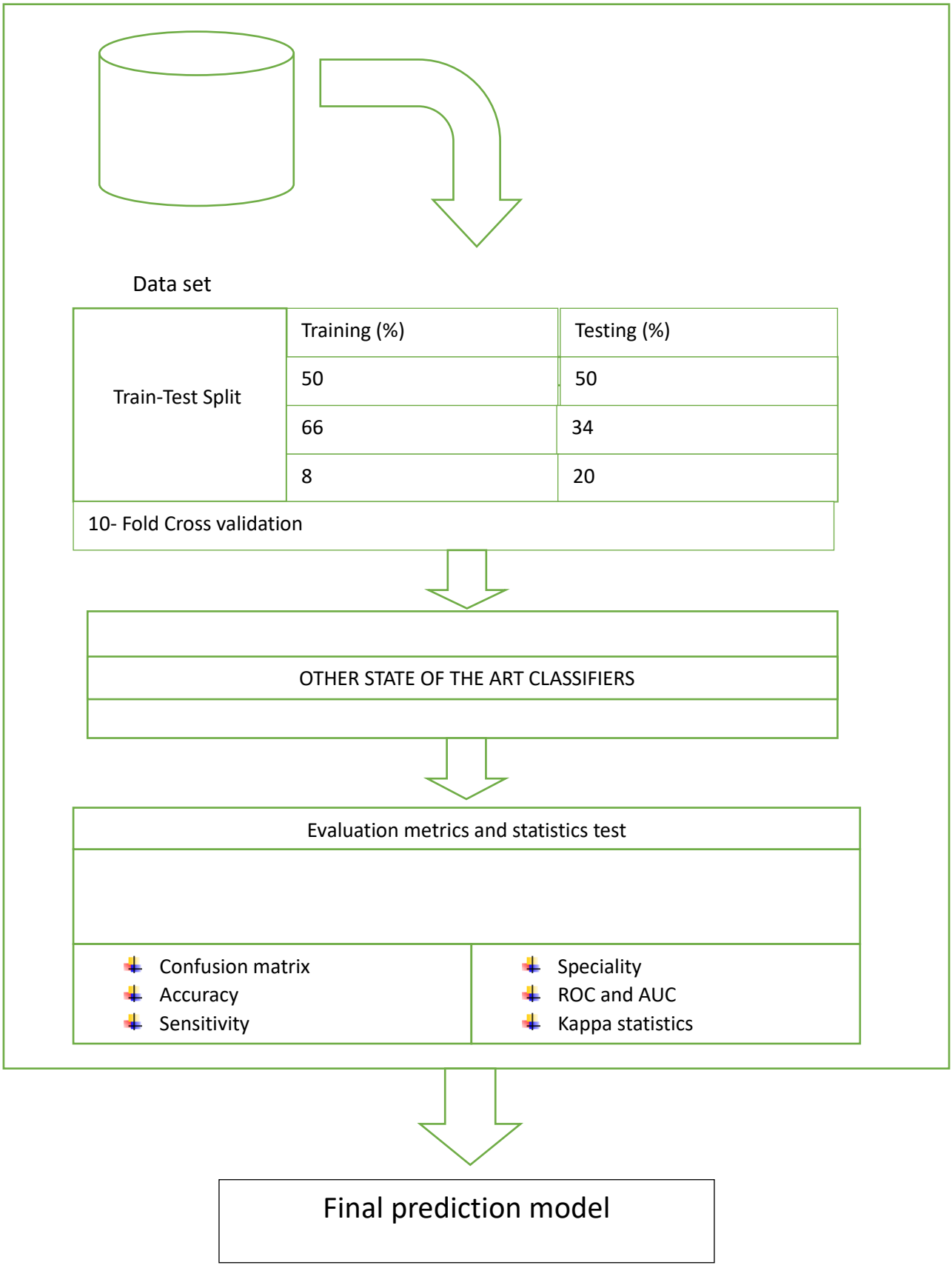

Data set
Train-Test Split
Training (%)
Testing (%)
50
50
66
34
8
20
10- Fold Cross validation
OTHER STATE OF THE ART CLASSIFIERS
Evaluation metrics and statistics test
Confusion matrix
Accuracy
Sensitivity
Speciality
ROC and AUC
Kappa statistics
Final prediction model

Figure 1: Analysis of Relationship between offenders and rape victim framework

Classifier

On this dataset, various ensemble algorithm used with bagging, stacking and adaboost algorithm.

The algorithm that has given maximum accuracy is stacking ensemble technique.

In machine learning, a stacking ensemble (also known as stacked generalization) is a technique that combines the predictions of multiple models (base models) to improve predictive performance.

Here's how it typically works:

1. **Base Models**: Different machine learning algorithms (such as decision trees, support vector machines, or neural networks) are trained on the same dataset to create diverse predictions.
2. **Meta-Model**: Another model, often called a meta-model or a blender, is trained on the predictions of the base models. Instead of using the original features of the dataset, this model uses the predictions made by the base models as its input features.
3. **Training and Prediction**: The base models are trained on a subset of the training data, and then each base model makes predictions on the validation set (or a subset of it). These predictions are used as input features for the meta-model, which then makes the final predictions.
4. **Final Prediction**: The meta-model aggregates the predictions from the base models and combines them into a single prediction. This final prediction is often more accurate than the predictions of any individual base model, as the stacking ensemble learns to correct the weaknesses of the base models.

Stacking ensembles are powerful because they can capture different aspects of the data and combine them effectively to make more accurate predictions. Still, they can also be computationally precious and bear careful tuning to avoid overfitting.

In Stacking model, three random foresthas been taken as base model and multiple perceptron as meta-model.

The work flow of stacking model given below:-

Input: Training data D = {xi, yi} (x; € R", y; € Y) Output: An ensemble classifier H

1: Step 1: Learn first-level classifiers

2: for 1 to T do

3:Learn a base classifier h, based on D 4: end for

5: Step 2: Construct new data sets from D

6: for i1 to m do

7:Construct a new data set that contains {x,y;}, where x = {hi (xi), h2(xi), . . . , hr (xi)} 8: end for

9: Step 3: Learn a second-level classifier

10: Learn a new classifier h' based on the newly constructed data set

11: return H(x) = h' (h1 (x), h2(x),...,hT (x))

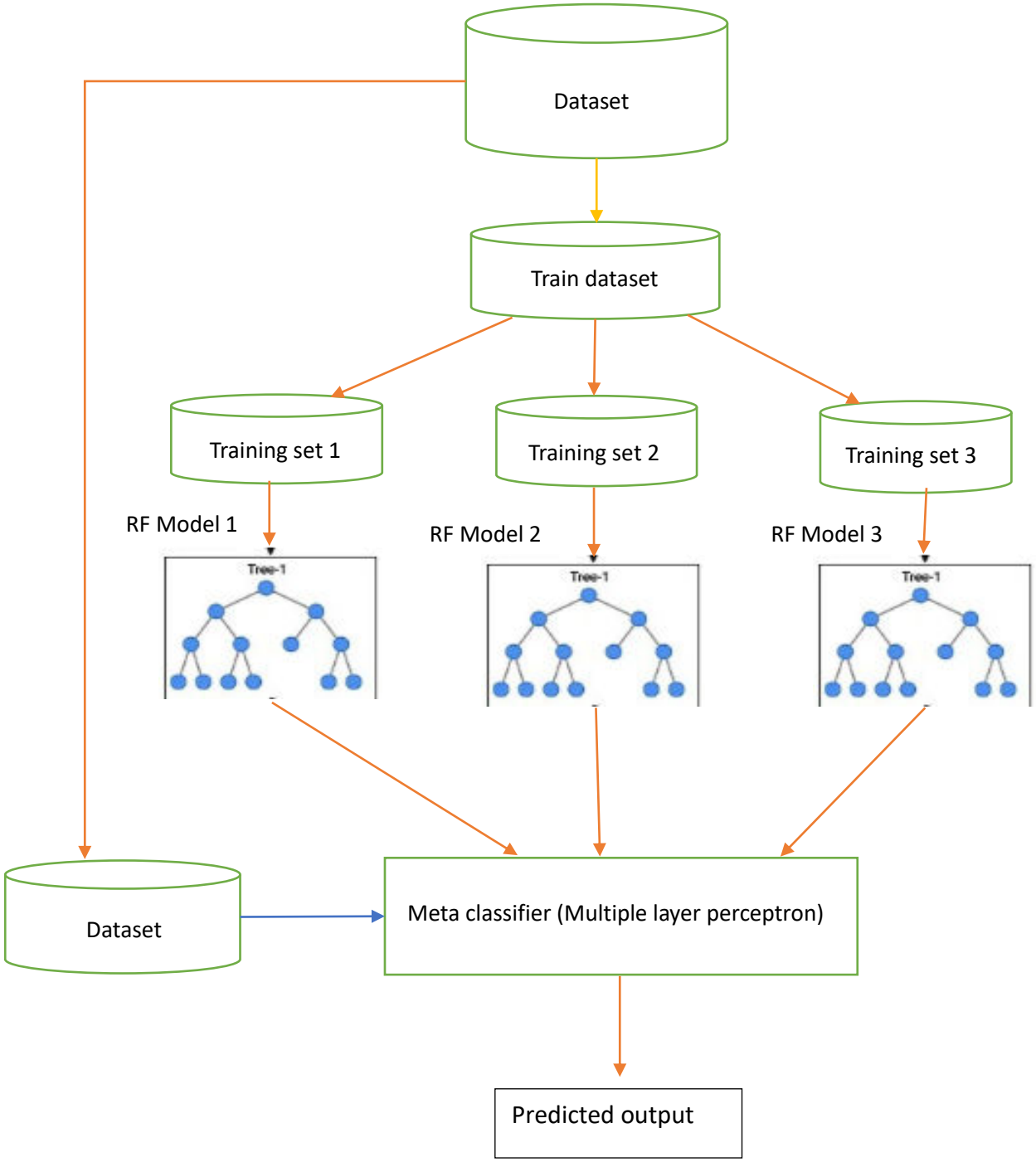


Figure 2: Stacking model

Ensemble learning

The concept of the "wisdom of crowds," which postulates that decision-making by a bigger group of individuals is often superior than that of a single expert, is supported by ensemble learning. In a similar vein, ensemble learning describes a collection of basic learners, or models, that collaborate to produce a more accurate final prediction.

A single model, sometimes referred to as a basic or weak learner, might not function well on its own because of significant bias or excessive variance. On the other hand, weak learners can be combined to create a strong learner by lowering bias or variance and improving model performance.

Decision trees are widely used as illustrations in ensemble methods. If this algorithm isn't pruned, it may overfit, exhibiting high variance and low bias. When it's really little, like a decision stump, which is a decision tree with one level, it might also favor underfitting, with low variance and significant bias.

Recall that an algorithm cannot generalize well to new data sets if it overfits or underfits to its training set. To mitigate this behavior, ensemble approaches are employed, enabling the model to be generalized to new data sets. Decision trees are not the only modeling strategy that uses ensemble learning to identify the "sweet spot" within the bias-variance tradeoff, despite the fact that they can show significant variance.

Stacking Ensemble Algorithm:

1. **Initialization**:
    - The training dataset D, which consists of feature vector pairs and their matching labels, is where we begin.
    - We select a base learner method (such as Decision Tree), specify a sampling rate (e.g., 0.8 for 80% of the dataset), and set the number of base learners T.
2. **Training**:
    - We randomly choose from D with replacement to build a bootstrap sample Dt of size n for each base learner from 1 to T.
    - Using Dt, we train a baseline learner ht. For later use, we keep the trained base learner ht stored.
3. **Prediction**:
    - We use the Dtest test dataset.
    - For every baseline student, ht:

We make predictions on Dtest to get a set of predictions Yt.

4. **Aggregation**:
    - When a classification problem is involved, we combine all base learners' predictions using majority voting.
    - If the task involves regression, the predictions are combined using averaging.

5. **Output**:
    - We provide the last forecast for every Dtest instance.

Result

1. Accuracy

Given table shows, Random Forest classifier is providing better accuracy in 10-fold cross validation.

Base Algorithm	10-Fold	50-50	66-34	80-20
Logistic regression	0.791	0.824	0.807	0.843
Naive bayes	0.79	0.801	0.803	0.829
SVM	0.786	0.7624	0.809	0.805
KNN classifier	0.963	0.881	0.915	0.94
Decision tree	0.897	0.867	0.854	0.85
Stacking model	0.976	0.923	0.944	0.979

Table 1. Accuracy

2. Precision

Precision is a metric that measures how often a machine learning model correctly predicts the positive class. You can calculate precision by dividing the number of correct positive predictions (true positives) by the total number of instances the model predicted as positive (both true and false positives).
Precision = True positives/ (True positives + False positives)
Given table shows, Random Forest classifier is providing better precision in 10-fold cross validation

Table 2. Precision

Base Algorithm	10-Fold	50-50	66-34	80 20
Logistic regression	0.7842	0.82	0.806	0.8428
Naive bayes	0.7885	0.797	0.802	0.828
SVM	0.7857	0.8171	0.806	0.8
KNN classifier	0.9628	0.8771	0.911	0.9357
Decision tree	0.89	0.84	0.84	0.85
Stacking model	0.98	0.92	0.94	0.98

3. Sensitivity

Random Forest classifier is providing better sensitivity in 10-fold cross validation

Sensitivity in Machine Learning can be described as the metric used for evaluating a model's ability to predict the true positives of each available category. In literature, this term can be also recognized as a true positive rate and it can be calculated with the following equation:

Table.3 Sensitivity

Base Algorithm	10-Fold	50-50	66-34	80-20
Logistic regression	0.784	0.82	0.807	0.843
Naive bayes	0.789	0.797	0.803	0.829
SVM	0.786	0.9775	0.807	0.8
KNN classifier	0.963	0.877	0.912	0.936
Decision tree	0.894	0.84	0.84	0.85
Stacking model	0.976	0.92	0.941	0.979

Base Algorithm	10-Fold	50-50	66-34	80-20
Logistic regression	0.5673	0.6388	0.61311	0.6836
Naive bayes	0.5756	0.593	0.6047	0.6566
SVM	0.5702	0.5882	0.6128	0.6013
KNN classifier	0.9256	0.7535	0.8232	0.8703
Decision tree	0.7977	0.6778	0.6795	0.6988
Random forest	0.9514	0.8396	0.8822	0.9569

Sensitivity = TP / TP+FN (True Positive/True Positive + False Negative)

4. F1 score

Given table shows, Random Forest classifier is providing better precision in 10-fold cross validation.

The F-score, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'.
Table 4. F1-score

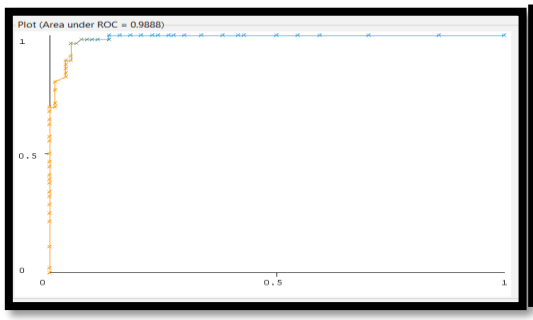
5. AUC ROC

Given table shows, Random Forest classifier is providing better precision in 10-fold cross validation.AUC ROC stands for "Area Under the Curve" of the "Receiver Operating Characteristic" curve. It's a way to measure the performance of a machine learning (ML) model.

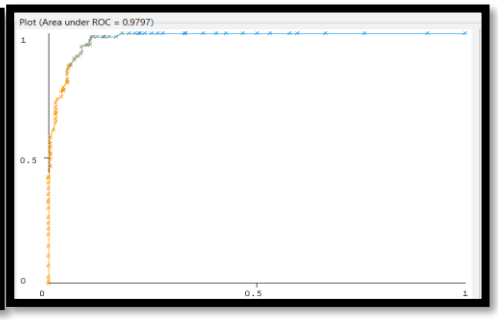
Table 5. AUC ROC

Base Algorithm	10-Fold	50-50	66-34	80-20
Logistic regression	0.784	0.819	0.807	0.842
Naive bayes	0.788	0.796	0.802	0.829
SVM	0.785	0.8447	0.806	0.8
KNN classifier	0.963	0.877	0.912	0.935
Decision tree	0.894	0.836	0.838	0.85
Stacking model	0.976	0.92	0.941	0.979

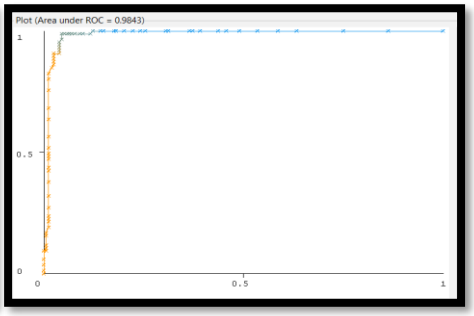
i. AUC ROCofStacking model



ROC train-test (80-20)

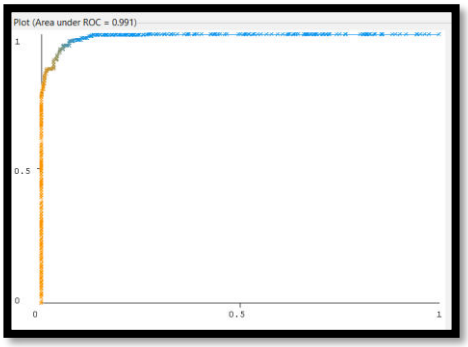


ROC train-test (50-50)

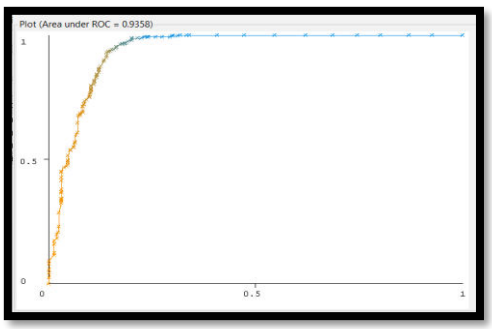


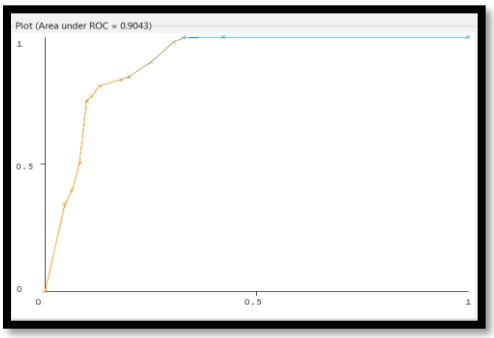
ROCtrain-test ROC train-test(66-34)

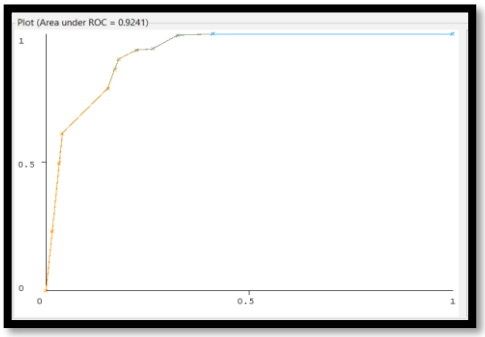
ROC 10- fold cross validation



ii. AUC ROC of Decision tree

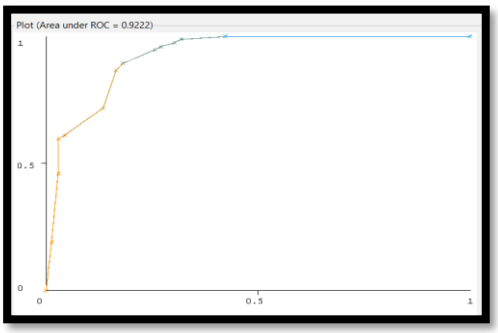






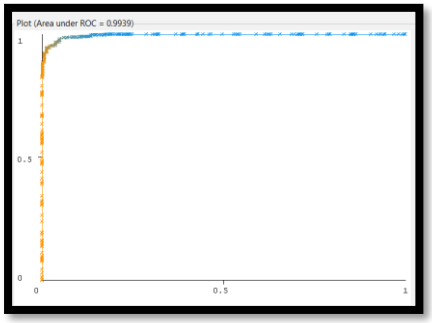
ROC 10- fold cross validation ROC train-test (50-50) ROC train-test (66-34)

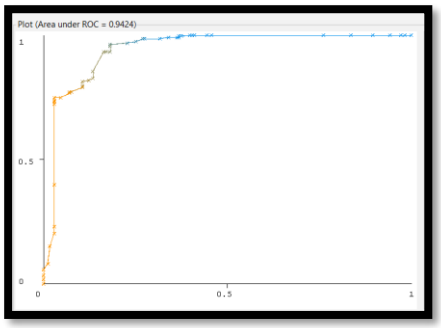
ROC train-test(80-20)

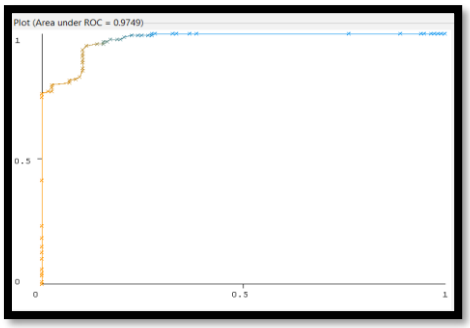


iii. AUC ROC of KNN

ROC 10- fold cross

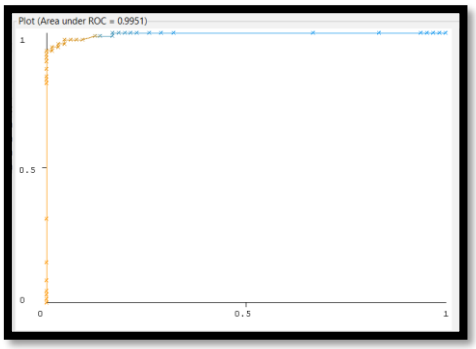




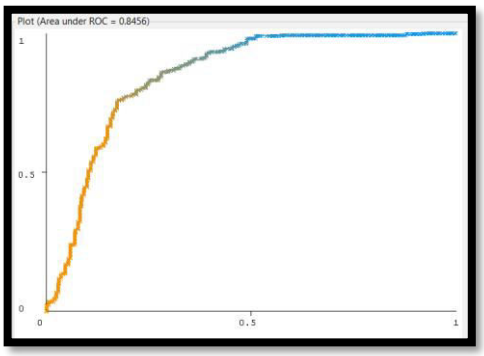


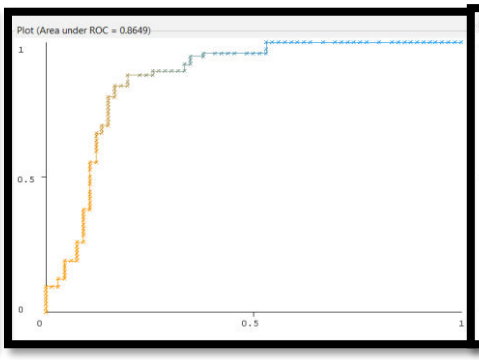
validationROCtrain-test(50-50)ROC train-test(66-34)

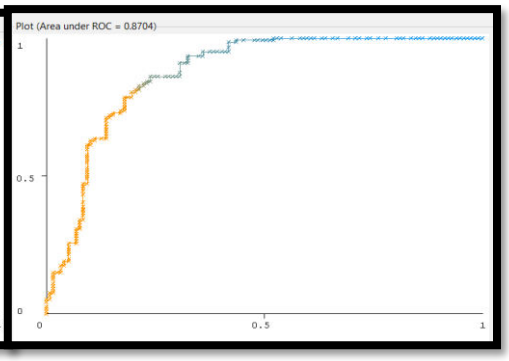
ROC train-test(80- 20)



iv. AUC ROC of Logistic

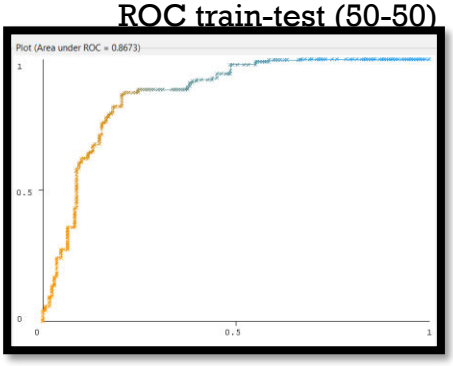




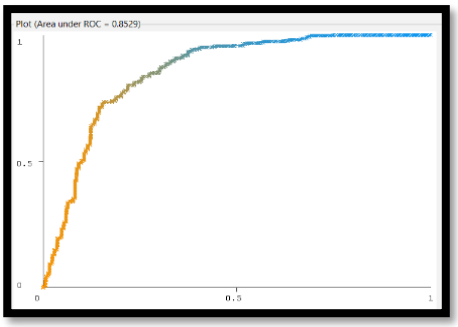


ROC 10-fold cross validation ROCtrain-test (80-20) ROC train-test (66-34)

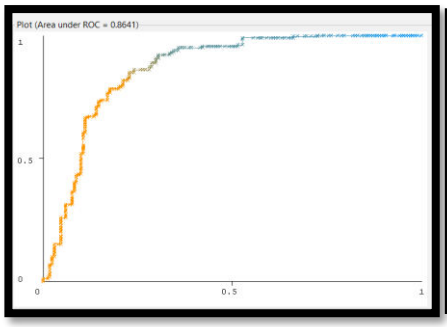
ROC train-test (50-50)



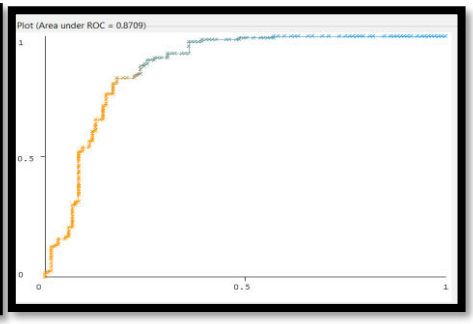
v. AUC ROC of naïve bayes



ROC 10-fold cross validation

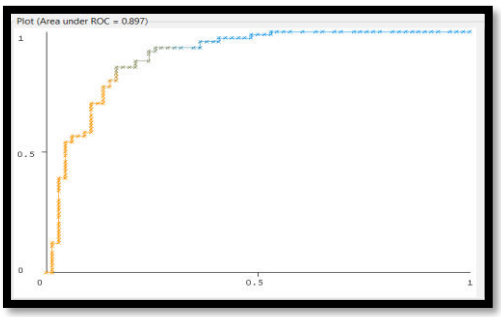


ROC test-train(50-50)

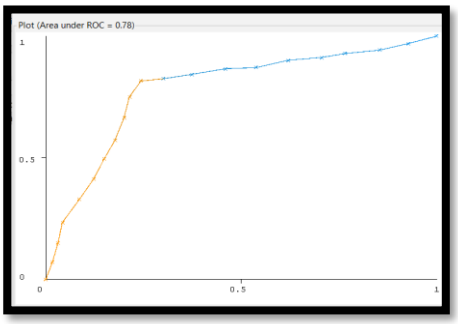


ROC 10-fold cross validation ROC test-train(66-34)

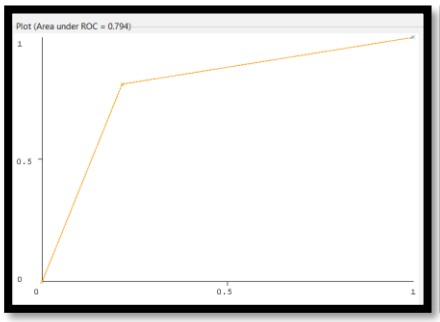
ROCtest-train (80-20)



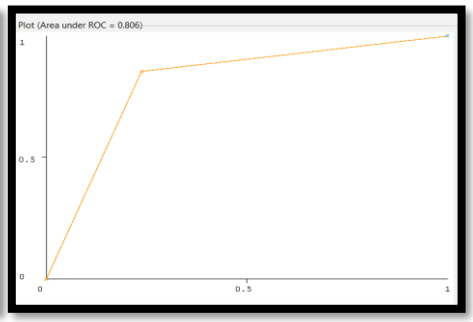
vi. AUC ROC of SVM(support vector machine)



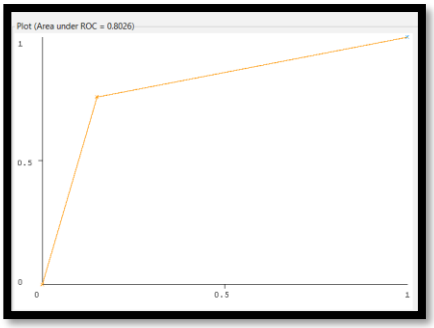
ROC 10-fold cross validation ROC test-train(66-34)



ROC test-train (50-50)



ROC test-train (80-20)



6. Kappa

Given table shows, Random Forest classifier is providing better precision in 10-fold cross validation.

Kappa is a statistic that measures the agreement between two dependent categorical samples. The range of possible values for kappa is from −1 to 1, but it usually falls between 0 and 1.
A kappa of less than 0.4 is generally considered poor. Kappa values of 0.4 to 0.75 are considered moderate to good. A kappa of greater than 0.75 represents excellent agreement.

Table 6. Kappa

Base Algorithm	10-Fold	50-50	66-34	80-20
Logistic regression	0.846	0.867	0.87	0.865
Naive bayes	0.853	0.864	0.871	0.897
SVM	0.78	0.794	0.806	0.803
KNN classifier	0.994	0.942	0.975	0.995
Decision tree	0.936	0.904	0.924	0.922
Stacking Model	0.996	0.971	0.99	0.997

7. Specificity

Here are the findings of Specificity
Specificity itself can be described as the algorithm/model's ability to predict a true negative of each category available. In literature, it is also known simply as the true negative rate. Formally it can be calculated by the equation below

Specificity = TN / TN + FP (True Negative/True Negative + False Positive)

Table 7. Specificity

Base Algorithm	**10-Fold**	**50-50**	**66-34**	**80-20**
Logistic regression	0.7783	0.79	0.8	0.833
Naive bayes	0.775	0.78	0.793	0.847
SVM	0.7867	0.795	0.786	0.854
KNN classifier	0.961	0.8434	0.8787	0.901
Decision tree	0.869	0.77	0.786	0.8533
Stacking Model	**0.972**	**0.891**	**0.908**	**0.961**

Conclusion

In every possible test option using different attribute selection, six features consistently demonstrated higher percentages in key properties such as accuracy, specificity, and kappa, compared to other sets of features. Among various cross-validation methods, 10-fold cross-validation emerged as the best, achieving an impressive 98% accuracy in the random forest classifier. This classifier not only excelled in accuracy but also outperformed other classifiers across different metrics. Moreover, the transformation of data into a binary true/false format further enhanced accuracy when employing stacking techniques.Analysing the results section reveals that, following the random forest classifier, the k-nearest neighbors classifier also performed admirably, achieving an accuracy of up to 96% and demonstrating superior performance in other properties as well. Additionally, the AUC ROC graph for the random forest classifier indicates a high-performing model, with the graph line ascending, signifying the classifier's capability to accurately distinguish between positive and negative instances. This study successfully classified data into subgroups, specifically distinguishing between incest rape and other categories. The findings of this paper offer valuable insights into the nature of crimes, the contributing factors, and the identification of those factors that are crucial in discerning the truth between defendants and rape victims.

Authors' Contribution

Ria Pyne: Data curation, Formal analysis, Investigation, Writing – original draft.

Suman Maji: Editing, Conceptualization, Supervision, Validation, Writingreview and editing

Avijit Kumar Chaudhuri: Data collection, Interpretation.

Reference

1. Majumdar, B., Laxminarayana, K., & Ghosh, P. (2020). Effect of integrated nutrient management on soil physical properties and crop productivity under maize–mustard crops in acidic soils of northeast India. Communications in Soil Science and Plant Analysis, 41(4), 2187–2200.
2. Sridharan, S., Srish, N., Vigneswaran, S., & Santhi, P. (2024). Crime prediction using machine learning. EAI Endorsed Transactions on Internet of Things, 10.
3. Popovic, D., Wertz, M., Geisler, C., Kaufmann, J., Lähteenvuo, M., Lieslehto, J., Witzel, J., Bogerts, B., Walter, M., Falkai, P., Koutsouleris, N., & Schiltz, K. (2023). Patterns of risk—Using machine learning and structural neuroimaging to identify pedophilic offenders. Frontiers in Psychiatry, 14.
4. Aziz, R. M., Hussain, A., Sharma, P., & Kumar, P. (2022). Machine learning-based soft computing regression analysis approach for crime data prediction. Karbala International Journal of Modern Science, 8(1), 1–19.
5. Palanivinayagam, A., Gopal, S. S., Bhattacharya, S., Anumbe, N., Ibeke, E., & Biamba, C. (2021). An optimized machine learning and big data approach to crime detection. Wireless Communications and Mobile Computing, 2021, 1–10.
6. Jassal, N. (2021). Gender, violence, and triage: Complainant identity and criminal justice in India.
7. Kshatri, S. S., Singh, D., Narain, B., Bhatia, S., Quasim, M. T., & Sinha, G. R. (2021). An empirical analysis of machine learning algorithms for crime prediction using stacked generalization: An ensemble approach. IEEE Access, 9, 67488–67500.
8. Saravanan, P., Selvaprabu, J., Raj, L. A., Khan, A. A. A., & Sathick, K. J. (2020). Survey on crime analysis and prediction using data mining and machine learning techniques. In Lecture notes in electrical engineering (pp. 435–448).
9. Rodríguez-Rodríguez, I., Rodríguez, J. V., Pardo-Quiles, D. J., Heras-González, P., & Chatzigiannakis, I. (2020). Modeling and forecasting gender-based violence through machine learning techniques. Applied Sciences, 10(22), 8244.
10. Walsh, C. M. (2019). Stereotypes and phenotypes: Using machine learning to examine racial implicit bias in sex offender criminal case processing.
11. Das, P., Das, A. K., Nayak, J., Pelusi, D., & Ding, W. (2019). A graph-based clustering approach for relation extraction from crime data. IEEE Access, 7, 101269–101282.
12. Rathod, S., & Talari, S. (2017). Understanding rape: A criminological study from the perspective of rape offenders.
13. Salunke, P. (2016, September 1). In 99.3% of rape cases, accused known to survivor: NCRB. Hindustan Times.

14. Plummer, M., & Cossins, A. (2016). The cycle of abuse: When victims become offenders. Trauma, Violence & Abuse, 19(3), 286–304.
15. Chaudhary, T. S. (2015). Violence against women in Delhi: Analyzing the nature, time, place, age, and relationship between a rape victim and offender (No. 10, June).
16. Mitchell, D., Angelone, D., Kohlberger, B., & Hirschman, R. (2008). Effects of offender motivation, victim gender, and participant gender on perceptions of rape victims and offenders. Journal of Interpersonal Violence, 24(9), 1564–1578.
17. Greenfeld, L. A. (1997). Sex offenses and offenders: An analysis of data on rape and sexual assault: Executive summary.
18. Gudjonsson, G. H., & Sigurdsson, J. F. (2000). Differences and similarities between violent offenders and sex offenders. Child Abuse & Neglect, 24(3), 363–372.
19. Gartner, R., & Macmillan, R. (1995). The effect of victim-offender relationship on reporting crimes of violence against women. Canadian Journal of Criminology, 37(3), 393–429.
20. Ullman, S. E., & Siegel, J. M. (1993). Victim-offender relationship and sexual assault. Violence and Victims, 8(2), 121–134.