International Journal of Innovation in Computer Science and Information Technology
Vol.2, NO.2 , P: 31 - 41
Received : 19 April 2020
Accepted : 24 June 2020

An Ensemble Learning Approch For Crime Analysis And Detection

Sina Dami*
Assistant Professor of Computer Department, West Tehran Branch, Islamic Azad University, Tehran, Iran
dami@wtiau.ac.ir
Maryam Kamravafar
Master Degree, West Tehran Branch, Islamic Azad University, Tehran, Iran
maryam.kamravafar@gmail.com

Abstract
The rate of crimes has substantially increased with the passage of time, and technologies that have helped people enjoy an easier life have contributed criminals to employ more accurate techniques in committing crimes. One of the key concerns for law enforcement officials is how to enhance the police investigative efficacy and attempt to improve it so that they could remain in the eternal contest between law enforcers and criminals. Data mining is the strongest and best method to extract basic knowledge and relationships between data and detect patterns in large amounts of data using various sciences. As a result, crime prediction, crime prevention, and crime detection, scientifically not just empirically with the help of data mining, is a strategy that causes the adoption of better decisions and strategic planning at the micro and macro levels.
In this paper, with the help of data mining algorithms and CRISP (Cross Industry Standard Process for Data Mining) methodology, we have dealt with the intelligent identification of criminals. Given that each of the data mining techniques and algorithms has different advantages and disadvantages, the use of ensemble methods that the jury actually constitutes and announces the final decision by a maximum voting will lead to attaining the best result. Our proposed model was trained by ensemble learning classifiers methods from three based classifier algorithms of RandomForest, Naive Bayes, and support vector machine (SVM) with weights of 2, 1 and 4, respectively. This technique has a greater efficiency than other methods and base algorithms and increases the evaluation criteria for the precision and accuracy of classification to 74% and 75%.

Keywords: Intelligent Crimnials Identifcation, Data Mining, Crime,Classification, Ensemble method.

1. Introduction

The rate of crimes has considerably increased over time, and technologies that have helped people enjoy an easier life have contributed offenders to employ more accurate techniques in committing crimes. One of the key concerns for law enforcement officials is how to enhance the effectiveness of police investigation and attempt to improve it so that they could remain in the eternal contest between law enforcers and criminals. Due to the growing volume of occurring crime, limited number of resources and police officers relative to the number of cases, analysis of large volumes of data to identify patterns of crime or to forecast future crimes by individuals is difficult. Hence, we need a tool for effective analysis so that it could refine the crime data effectively and quickly to analyze the crime. Also, when work is repeatedly done too much, most people make mistakes. It seems that computers have more accurate and faster performance than humans. Finally, benefitting from the tool, the investigation and discovery of the crime after committing is accurately carried out in the least possible time [1]. Data mining is the strongest and best method to extract basic knowledge and relationships between data

and detect patterns in large volumes of data using various sciences such as artificial intelligence, machine learning, statistics, database, and pattern recognition. As a result, crime prediction, crime prevention, and crime detection, scientifically not just empirically with the help of data mining, is a strategy that causes the making of better decisions and strategic planning at the micro and macro levels.

Considering the enormous and very valuable volume of data collected over time through the entry of information into the comprehensive system of police, in this article, we intended to improve the performance of police systems using the data mining techniques (DMT) as an essential solution. The most important of these is to help identify the offender or offenders with acceptable accuracy and percentage for use in the process of crime detection.

Although valuable theoretical articles have been presented in the domain of data mining and crime, they are limited to "manually collected information", or confined to "a type of crime". In many countries, including India [2], where crime statistics are very high compared to the number of police officers, the subject has been used in the crime detection employing "a type of algorithm".

In this study, using CRISP methodology [3] and data mining in the information of the Criminal Investigation Department (CID) Police comprehensive system (as a case study), we attempt to achieve valuable results in the field of crime prediction and detection. Our model includes the following characteristics:

• Our model is not limited to the detection of "a type of crime" and covers "all crimes".

• In our proposed model, feature extraction is carried out manually in consultation with experts via statistical results.

• In our proposed model, the ensemble learning method and several algorithms have been used to improve and increase accuracy in identifying offenders.

2. The proposed method

In this section, intending to present a proposed approach to "provide a model for crime analysis and intelligent identification of criminals", using CRISP methodology in data mining on the information of the Criminal Investigation Department (CID) Police comprehensive system as a case study, we attempted to achieve valuable results in the field of crime prediction and detection in order to enhance the speed of crime discovery.

Of course, several valuable theoretical articles have been provided in the domain of data mining and crime but

• They are limited to information collected manually or confined to a particular "type of crime"

• The models made have used one "type of algorithm" in the detection of crime section

It should be noted that, in this study, we intend to employ a new method of combining algorithms for improving and increasing the accuracy using the extracted features manually by the evaluation of the statistical results and consultations with experts.

Taking into account the research hypotheses, while the analysis of crimes occurred and the extraction of the behavioral characteristics specific to each offender, using the ensemble learning model developed with data mining techniques, we can identify the offender in future crimes.

The overview of the proposed method is in accordance with Figure 1.

The five major steps of the data mining process (green rectangles according to Figure 1) are as follows[4]:

- Data Extraction
- Data preprocessing
- Modeling
- Model evaluation
- Knowledge extraction and integration

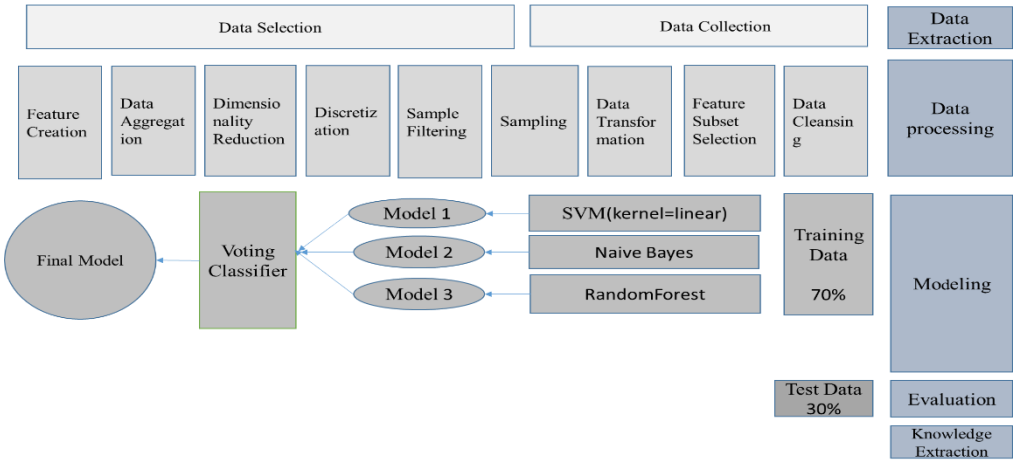

Figure 1: Proposed Method

The tools used

• Oracle12i has been used as a database

• In the first step, we used pl/sql Developer for the collection of dataset, analysis of crime, and large-scale decision making of the problem, and writing of composite queries. The extraction of histograms and statistics and the construction of the model and initial tests were done employing OracleDataMiner, two powerful products of Oracle Corporation.

• In the second step, which is related to the preparation of the hybrid and final model, we used the open-source Python library called sklearn in the spider environment.

2.1. The main steps of the data mining process

2.1.1. Data Extraction

Data extraction includes data collection and data selection.

2.1.1.1. Data Collection

Since access to the information of the year 2007 was issued with the arrangements made in isolated conditions, this information should be separated from the entire tables by maintaining communications and primary and foreign keys of the tables and transferred to the new database. The file table with the filter of date of convening (session formation) in 2017 was the basis of all the tables, and all tables, with/without the mediation, used an external key to extract the table.

2.1.1.2. Data Selection

A part of the cached central data repository, which will be the aim of exploring our data mining process, is selected. Due to the purpose of this problem, i.e. "construction of a model to identify the criminals of files", the datasets directly related to the problem under investigation were collected through complex queries. With respect to the need for the problem of data mining, to analyze the structure of tables, general statistical information and making macro decisions were performed on all data (data of year 2017). Then, a required dataset with fewer number of records was extracted as the test sample. Because processing all the data in the dataset is very time-consuming for data mining algorithms, we used sampling. Information of records includes the data of crime occurred and offender such as type of file, type of crime, state of occurrence, date of occurrence, address of occurrence, investigating unit, location of occurrence, techniques and tricks of crime happened, type

of stolen goods, type of damaged case (commodity, document, vehicle, person), area under the protection of the crime occurrence, number of men arrested, number of women arrested, date of occurrence, type of committing a crime, offender ID, gender, recidivism, whether or not the career of offender, is he/she professional (habitual), status of the offender, control level of offender, level of offender, age group of offender, nationality, education level, criminal record, marital status, and motivation. The data samples are similar to those in Table 3.

Table 1: Raw data sample

File Type ID	Name of occurrence mode	Crime occurrence mode ID	Offender ID	Row
141081012	Normal	101021001	146514	1752290
Crime occurrence interval ID	Name of the crime subject	Crime subject ID	Date of the crime occurrence	Name of file type
102581003	Forgery/ Counterfeiting banknotes and securities	506	1/5/2015	Occurrence
Address of the crime occurrence	Province of crime occurrence	Province of crime occurrence ID	Address of the crime occurrence ID	Name of the crime occurrence interval
City: Tehran – Neighborhood …	Tehran	16071015	15257814	10-12
Main techniques and tricks of crime ID	Name of place of crime occurrence	Place of crime occurrence ID	Name of the investigating unit	Name of the investigating unit ID
102111140	Banks / State Banks / National Banks	171631007	GREAT TEHRAN POLICE COMMAND/ Criminal Investigation Department (CID) Police	11826
Survival status	Nationality ID	Date of birth	Gender ID	Main techniques and tricks
1	102641019	22/10/1988	101061001	Making all the components of a ...
Record status	Professional (habitual)	Is the crime his career	Recidivism	Criminal record
2	1	1	1	

Code of the type of committing a crime	Motivation	Motivation ID	Field of techniques and tricks	Control level
102641001	Gaining material benefits	102551172	1E+228	1
Number of women arrested	Number of men arrested	Name of the area under the protection of the crime occurrence	Area under the protection ID	Name and type of committing a crime
0	3	GREAT TEHRAN POLICE COMMAND/ Criminal Investigation Department (CID) Police	11826	Gang member
Education level	Marital status	Age group of offender	Offender's year of birth	Date of the crime occurrence
3	1	3	1988	267
Crime case - commodity	Crime case – animal	Technique and trick…	Technique and trick 2	Technique and trick 1
1	0			1
Type of stolen goods…	Type of stolen goods 1	Crime case - automobile	Crime case - motorcycle	

2.1.2. Data processing

Data processing includes data cleansing, feature subset selection, sample filtering, sampling, data transformation, discretization, dimensionality reduction, data aggregation, and feature creation. This step may be repeated at all stages snail-like based on the needs, even in the final stages such as evaluation phase in which we find the reason for the inaccurate results due to the lack of accurate performance of data preprocessing because this part is the most important and time-consuming phase in conducting the operation, and it is the main feed of the model and the learning algorithm.

2.1.2.1. Data Cleansing

Problems that endanger data quality such as noise, outliers, missing values, and duplicate data. One-to-many connections in a series of tables and transfer them to a dataset were difficult in some respects. For example, since one file could have several techniques and tricks and because we wanted the model to find out the possibility of sharing of all files, three methods were taken into account to solve this problem. Finally, concerning the low number of records, the third case was considered, otherwise, the first case would be selected:

- We can add 600 columns to the number of techniques and tricks, which are very diverse and this causes the generation of columns with an empty value.
- We can create a 600-length field to the number of techniques and tricks that, for availability of each of them, have a bit with value one.
- Classification of all techniques and tricks according to the data into five general

groups, which led to the generation of five columns.

2.1.2.2. Feature Subset Selection (FSS)

Feature subset selection is one of the dimensionality reduction operations, in which plug-in features are removed so that the algorithm space does not become unnecessarily large, and we maintain the more favorable attributes.

2.1.2.3. Sample Filtering

We select a subset of the data assuming that offenders of the crimes happened are selected from among professional criminals (habitual offenders). According to our assumptions, a very high rate of crimes are committed by a certain series of the perpetrators, particularly in the crimes of theft.

2.1.2.4. Sampling

Because the processing of all data in the dataset is very time consuming for data mining algorithms, sampling is used. A total of 1322 records with 17 classes were selected.

2.1.2.5. Data Transformation

The mission of data transformation operation is to transfer feature values from one domain to another domain in the entire records of a dataset, such as:
new_field_value= abs(substr(field_value,0, length(field_value)-3)||'000'-field_value)
abs(substr(101061003,0, length(101061003)-3)||'000'- 101061003)=3

Decimal transformation of numbers such as:
to_decimal(01011111)= 95

Transformation of a date to a number using the following formula used such as:
New_date=
day+ (30*(month+ 12*(year-1398)))

We transformed all the information in the form of numerical data in a limited range (due to better training of the neural network).

2.1.2.6. Discretization

Date, which is a discrete amount, is discretized to sequential values of 1, 2, 3, 4, and 5 that is in the order of date of birth including 1 to 17, 18 to 25, 26 to 40, 41 to 55, and above 55 respectively.

2.1.2.7. Dimensionality Reduction

The higher the dimension (size) or the number of features of the problem being explored, the records in the search space will be more fragmented. One method is in a way that a subset of features whose information value is low for exploration will be selected for elimination. With the tests done, features such as career and criminal history do not have much impact on the accuracy of the model with regard to their recorded information.

2.1.2.8. Data Aggregation

is the combination of two or more features and generation of a new feature or reduction in the number of features or records such as the classification of stolen goods into three different categories

2.1.2.9. Feature Creation

It is defined as the creation of new features that can, along with other previous features, represent the most important information available in a dataset more effective and complete than the initial features. In general, there are three different approaches to create features, including feature extraction, feature construction, and data transformation. Feature extraction of year, season, which months of the year, and which days of the week from the date of crime occurrence can, for example, determine that the number of theft will increase in March and April.
All information was prepared in the form of numerical data in a limited range (due to better training of the neural network).
The features of preparation data including the data of crime occurred and offender such as type of file, type of crime, state of occurrence, investigating unit, place of crime occurrence, techniques and tricks of crime happened, type of stolen goods, type of damaged case

(commodity, document, vehicle, person), area under the protection of the crime occurrence, number of men arrested, number of women arrested, year of crime occurrence, month of crime occurrence, day of crime occurrence, season of crime occurrence, type of committing a crime, and offender ID were prepared as the target variables. The data samples are similar to those in Table 4.

Table 2: Sample data after pre-processing

Crime subject ID	File type ID	Crime occurrence mode ID	Offender ID	Row
506	12	1	514	1752290
Place of the crime occurrence ID	Investigating unit ID	Province of crime occurrence ID	Address of the crime occurrence ID	Crime occurrence interval ID
7	826	15	814	3
Date of the crime occurrence	Number of women arrested	Number of men arrested	Area under the protection ID	Code of the type of committing a crime
267	0	3	826	1
Which day of the week – crime occurrence	Which month - crime occurrence	Year of the crime occurrence	Season of the crime occurrence	Technique and trick ID
2	1	23	1	140
Technique and trick 1	Crime case - commodity	Crime case - animal	Crime case - automobile	Crime case - motorcycle
1	1	0	0	0
Type of stolen goods…	Type of stolen goods 2	Type of stolen goods 1	Technique and trick…	Technique and trick 2

2.1.3. Feature selection with Auto Encoder

The structure of Auto Encoder is divided into two parts of encoding and decoding. The input data are mapped to the feature space in the encoding section, and they again convert from the feature space to their initial state in the decoding section. Indeed, the main part of an Auto Encoder is the hidden middle layer, which is used as the extracted feature for classification. We add columns of all kinds of goods, documents, etc. without limitation. After that, using the Auto-Encoder network, we extract important features [5] and then transfer the result to the classification algorithms.

2.1.4. Modeling

We divide the entire data into two parts, in which 70% of the information for the training dataset and 30% of the information for the test dataset are used so that we could evaluate the reaction of the algorithms to information that has not been seen before.

Taking into account that any training model made by an algorithm has a percentage of error, disadvantages, and advantages in the real world, we attempted to provide a better and more powerful model employing the ensemble learning method by collecting the strengths of the best algorithms, minimizing and overlapping the weaknesses of one algorithm with the strengths of other algorithms [6].

To achieve this, we must select the best algorithms for the combination in a step earlier. We build various models using algorithms proposed in similar domains such

as decision tree, Kneighbors, simple naive Bayes classifier, random forest, logistic regression, and support vector machines (SVM) with different kernels (linear, polynomial, sigmoid, and rbf)[7][8] and, after optimizing and evaluating them, select the best algorithms for the main stage.
In the main stage of combining the best algorithms using
• The result of evaluating each of the algorithm in the first stage
• The best combination of algorithms in a valid papers presented [9] [10][11][12]
• Taking into account the behavior of the dataset on different combinations of algorithms as the most important item We provide the best combination as the final model for intelligent identification of offenders (by voting and weighing the selected algorithms, they are actually constituted by jury and the results are predicted with maximum vote).

3. Evaluation

In this section, we initially evaluate the models constructed by basic algorithms. Then, we evaluate the hybrid models with different weights and finally compare the results of the basic and hybrid models.

3.1. Dataset

We have worked on a dataset containing 1322 data records with 17 class types. We considered the data into two parts, 70% for training and 30% for testing. The training is first done on the dataset and then the test is taken of it.

3.2. Evaluation criteria

Evaluation can be conducted in two stages, first on the training data and, after being the logical result, in the final stage on the test data. Evaluation criteria in this article include precision, recall, accuracy, F1 measure, micro average, macro average displayed in Table 5.

Table 3: Evaluation criteria

Accuracy	$\text{Accuracy} = \frac{TP+TN}{N}$
Precision	$\text{Precision} = \frac{TP}{TP+FN}$
Recall	$\text{Recall} = \frac{TP}{TP+FN}$
F1-Score	$\text{F1-Score} = \frac{2}{\frac{1}{\text{Recall}}+\frac{1}{\text{Precision}}} = 2 * \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$
Macro averaging	$\text{Precision}_{\text{macro}} = \frac{\text{Precision}_1 + \cdots + \text{Precision}_k}{K}$
Micro averaging	$\text{Precision}_{\text{micro}} = \frac{TP_1 + \cdots + TP_k}{TP_1 + \cdots + TP_k + FP_1 + \cdots + FP_k}$

3.3. Analysis of the results

Evaluation criteria of the models contain precision, recall, F1 measure, accuracy, macro average, and micro average. One of the evaluation criteria including the accuracy of the models in the first phase of the proposed method is illustrated in Chart 1.

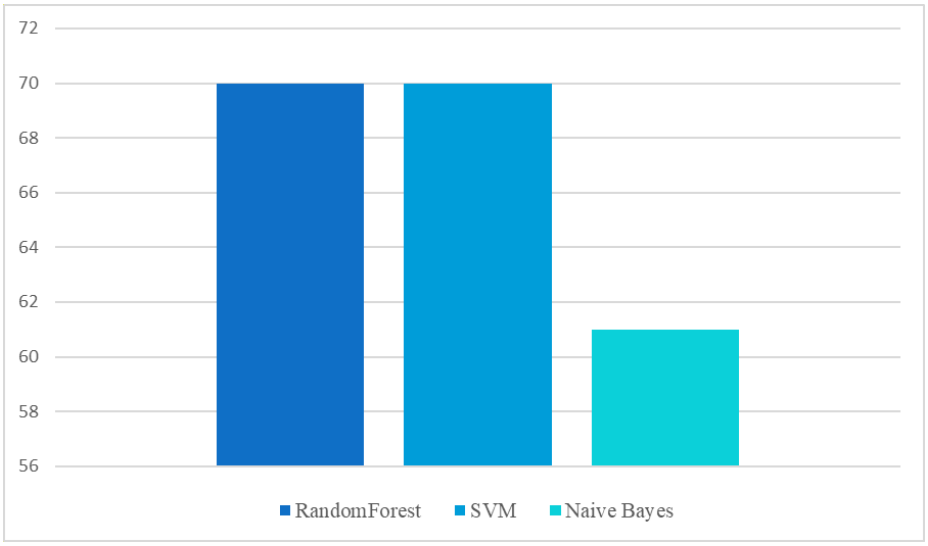

Chart 1: Comparison of the accuracy of the initial models

Then, in Chart 2, a comparison of the accuracy of hybrid algorithms with different weights is represented. As can be seen, the models constructed by random forest, naive Bayes, and support vector machine algorithms with the weights [2, 1, 4] had the highest accuracy, respectively, according to the available data.

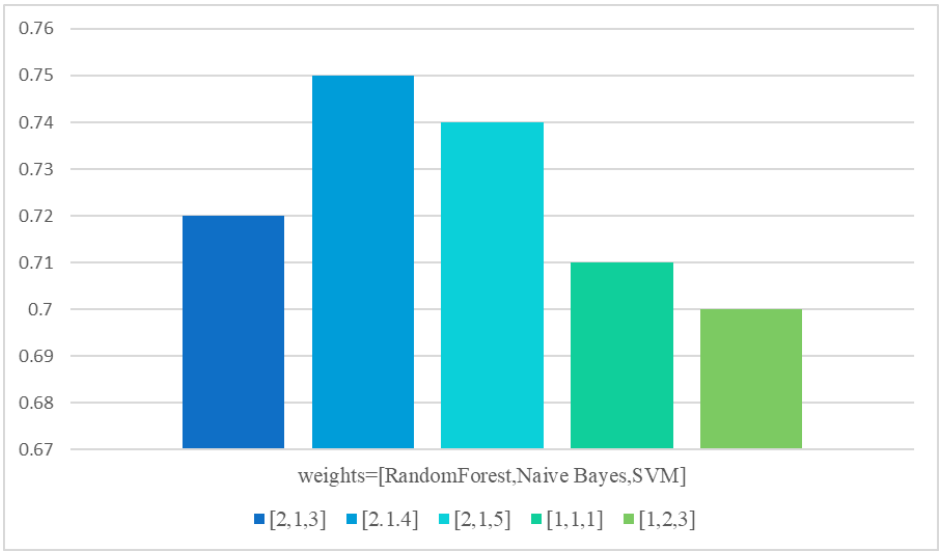

Chart 2: Comparison of the accuracy of hybrid algorithms with different weights

Finally, after selecting the best algorithms for the combination, in the third step, according to Chart 3, a comparison between the accuracy of the ensemble learning model with other initial models is indicated. As it is known, the model constructed by the hybrid algorithm has a higher accuracy than the models made by the basic algorithms.

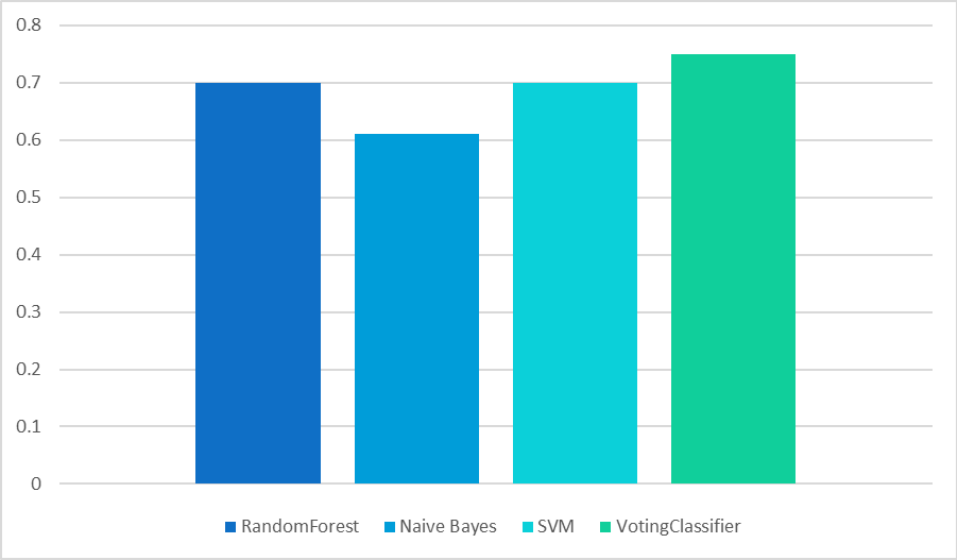

Chart 3: Comparison of the accuracy of the ensemble learning model with other models (basic algorithms)

In Table 6, the evaluation criteria of the final model are represented in detail.

Table 4: Evaluation criteria of the final model

VotingClassifier [2,1,4]	targe	precision	recall	f1-score	support
	19	0.92	0.92	0.92	12
	40	1.00	1.00	1.00	18
	239	1.00	1.00	1.00	58
	250	0.92	0.96	0.94	24
	251	0.86	0.95	0.90	19
	304	0.22	0.21	0.21	24
	498	0.87	0.93	0.90	28
	499	0.97	1.00	0.99	35
	514	0.37	0.53	0.44	32
	670	0.93	0.87	0.90	15
	690	0.23	0.12	0.16	24
	691	0.54	0.69	0.60	32
	692	0.71	0.91	0.80	11
	705	0.14	0.05	0.07	22
	781	1.00	1.00	1.00	14
	813	1.00	0.82	0.90	11
	939	0.94	0.83	0.88	18
accuracy				0.75	397
macro avg		0.74	0.75	0.74	397
weighted avg		0.73	0.75	0.74	397

4 . Conclusions and Suggestions

The objective of this paper was to predict the crime occurred in which criminal class is, by the implementation of the hybrid algorithm. We have worked on a dataset containing 1322 data records with 17 class types. We considered the data into two parts, 70% for training and 30% for testing. The training is first done on the dataset and then the test is taken of it. After evaluating the model, we reached 75% accuracy in identifying the offender.Use combinations of other algorithms such as ELM, RBF It can be concluded that using the hybrid (ensemble) model, we can deal with the identification of the offenders in the next crimes while analyzing the crimes occurred and extracting the behavioral characteristics specific to each offender. In light of the findings, the following items are recommended in the future: One can add columns of types of goods, documents, etc. without any restriction, then select and extract essential features employing the Deep Belief Network (DBN) and next transfer the result to the next classification algorithms.

Instead of the output of "a criminal," a list of criminals with the probability of being found guilty can be displayed **.**

References

[1].Saravanan, P., Selvaprabu, J., Raj, L. A., Khan, A. A. A., & Sathick, K. J. (2021). Survey on Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques. In Advances in Smart Grid Technology (pp. 435-448). Springer, Singapore.

[2].Jabeen, N., & Agarwal, P. (2021). Data Mining in Crime Analysis. In Proceedings of Second International Conference on Smart Energy and Communication (pp. 97-103). Springer, Singapore.

[3].Tasnim, S., Sarkar, P., Hossain, A., & Ali, M. A. A Classification Approach to Predict Severity of Crime on Boston City Crime Data.

[4].Yerpude, P. (2020). Predictive Modelling of Crime Data Set Using Data Mining. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol, 7.

[5].Aldossari, B. S., Alqahtani, F. M., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., & Aslam, N. (2020, January). A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. In Proceedings of 2020 the 6th International Conference on Computing and Data Engineering (pp. 34-38).

[6].Gahalot, A., Dhiman, S., & Chouhan, L. (2020, February). Crime Prediction and Analysis. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-6). IEEE.

[7].Farahnakian, Fahimeh, and Jukka Heikkonen. "A deep auto-encoder based approach for intrusion detection system." 2018 20th International Conference on Advanced Communication Technology (ICACT). IEEE, 2018.

[8].EKUNDAYO, O. EVALUATION OF MACHINE LEARNING ALGORITHMS FOR REGRESSION AND CLASSIFICATION PROBLEMS.

[9].Viloria, A., Lezama, O. B. P., & Hurtado, J. (2021). Combination of Support Vector Machine (SVM) and Bayesian Model to Identify Criminal Language. In Proceedings of International Conference on Intelligent Computing, Information and Control Systems (pp. 255-262). Springer, Singapore.

[10].Yadav, D. C., & Pal, S. (2020). Prediction of thyroid disease using decision tree ensemble method. Human-Intelligent Systems Integration, 2(1), 89-95

[11].Li, C., Ding, N., Zhai, Y., & Dong, H. (2021). Comparative study on credit card fraud detection based on different support vector machines. Intelligent Data Analysis, 25(1), 105-119.

[12].Razali, M. N. (2020, January). A Classification Approach for Crime Prediction. In Applied Computing to Support Industry: Innovation and Technology: First International Conference, ACRIT 2019, Ramadi, Iraq, September 15–16, 2019, Revised Selected Papers (Vol. 1174, p. 68). Springer Nature.