Revealing Cyber Criminals with Machine Learning

Raveesha N[1], Dr. Sanjay Kumar Tiwari[2]

[1]Research Scholar, Sunrise University, Alwar
[2]Associate Professor, Sunrise University, Alwar

Abstract

The National Crime Records Bureau is responsible for maintaining the availability of all case records, which include annual crime statistics from all throughout the country. In most cases, the obtained data has been left unprocessed and contains errors or blanks. Pre-processing of data is crucial for correcting and normalizing this information. This requires the pre-processing and cleaning of data. Part of this procedure is picking out recurring criminal themes. The pattern of criminal activity in a given location is the product of pattern recognition. Conditions such as climate, significance of nearby events, sensitivity of the region, and the presence of criminal gangs are among the many site-specific factors considered. support vector machine knn algorithm research, a supervised learning technique.

Keywords: Security, Fuzzy Logic, Data Mining, Machine Learning, Cybercrime, E-Crime and Supervised.

1. INTRODUCTION

Cybercriminals are persons or groups who use computers and other electronic devices to perpetrate illegal acts against other people or organisations for financial gain. The dark web is home to several underground marketplaces frequented by cybercriminals who deal in illegal products and services including hacking tools and stolen data. Cybercrime's dark web marketplaces are notorious for selling just a select few goods and services.

Cybercrime laws are still developing and changing in many parts of the globe. Finding cybercriminals, arresting them, prosecuting them, and establishing their guilt are all ongoing difficulties for law enforcement. Not all hackers commit cybercrime, since hacking itself is not always considered a crime. Whereas hackers are interested in discovering novel and useful applications for existing systems, cybercriminals hack and penetrate them for harmful purposes.

Differences between cybercriminals and threat actors include, first and foremost, the motivation behind each group's actions. Individuals that undertake targeted assaults, seeking out and compromising a specific organization's systems, are known as threat actors. Cybercriminals seldom target a single organisation, but rather target large groups of people who share characteristics like their internet habits or the software they use. Two, they have different operational procedures. The six-step method that threat actors use to compromise a system begins with gathering information about their intended victims and continues with them moving laterally across the network. The opposite is true of cybercriminals, who are not likely to use a certain set of procedures to get their goals from their victims.

Understanding and developing 'learning' techniques, or methods that use data to enhance performance on some set of tasks, is the focus of machine learning (ML), a subfield of computer science. It is considered a subset of AI. In order to draw inferences or choices without being explicitly taught, machine learning algorithms construct a model using sample data. This data is referred to as training data. In many fields, where it would be impractical or impossible to create custom algorithms to accomplish the desired results, machine learning algorithms are utilized instead. This is the case in medicine, email filtering, voice recognition, agriculture, and computer vision, to name a few.

Not all machine learning is statistical learning, but a portion of it is strongly connected to computational statistics, which focuses on generating predictions using computers. Mathematical

optimization provides the field of machine learning with new tools, theoretical frameworks, and potential application areas. Exploratory data analysis utilizing unsupervised learning is at the heart of data mining, a related area of research. The utilization of data and neural networks in some forms of machine learning is reminiscent to how the human brain processes information. Machine learning, when applied to business issues, is also known as predictive analytics.

2. CRIME ANALYSIS

Data Collection: The police department keeps extensive information on criminal activity. The National Crime Bureau of Records is responsible for maintaining the accessibility of all recorded crime data from around the country in the form of cases. In most cases, the obtained data has been left unprocessed and contains errors or blanks. Pre-processing of data is crucial for correcting and normalizing this information. It requires pre-processing, or the act of cleaning, and processing the data beforehand.

Classification: Each subset of the dataset is based on a unique combination of characteristics of the data item. Criminal activity may be categorized by state and city. During the classification process, crimes are categorized according to their many characteristics. Data with shared characteristics may be clustered with the help of the K-mean technique.

Pattern Identification: Part of this procedure is picking out recurring criminal themes. The pattern of criminal activity in a given location is the product of pattern recognition. Factors like climate, major events, region sensitivity, the presence of criminal gangs, and so on are considered in accordance with the specifics of each site. This kind of trend data helps police personnel do their jobs more efficiently.

Prediction: In each case, a model is developed to fit the specific environment. The current time and qualities are entered into the prediction engine to get the most dangerous neighborhoods. Data visualization allows for the representation of findings.

Visualization: A heat map depicting the intensity of criminal activity provides a visual depiction of these trouble spots. For example, a dark colour scheme indicates little activity, whereas a bright colour scheme indicates great activity. Phases of criminal activity are shown in Figure 4.1. analysis.

3. SUPERVISED LEARNING METHOD

The goal of the supervised learning approach is to construct a model that can make predictions in conditions of uncertainty using the available data. Patterns in data may be identified using adaptive algorithms, allowing computers to "learn" or "understand" from their experiences. It has been shown that as the number of observations increases, the computer's predicting ability improves. A supervised learning algorithm takes into account the set of input data and produces an answer to a question based on the data. Next, the algorithm "trains" a model to make accurate predictions in light of fresh information. The input data set is structured as a matrix, with rows representing different occurrences, observations, or examples, and columns representing different qualities, predictors, or features. Each user's measurements are represented by a set of values in both the row and column.

The results are represented as a column vector, with each row containing the corresponding output for one observation in the input data.

A supervised learning model is trained by feeding it input and output data using an appropriate method.

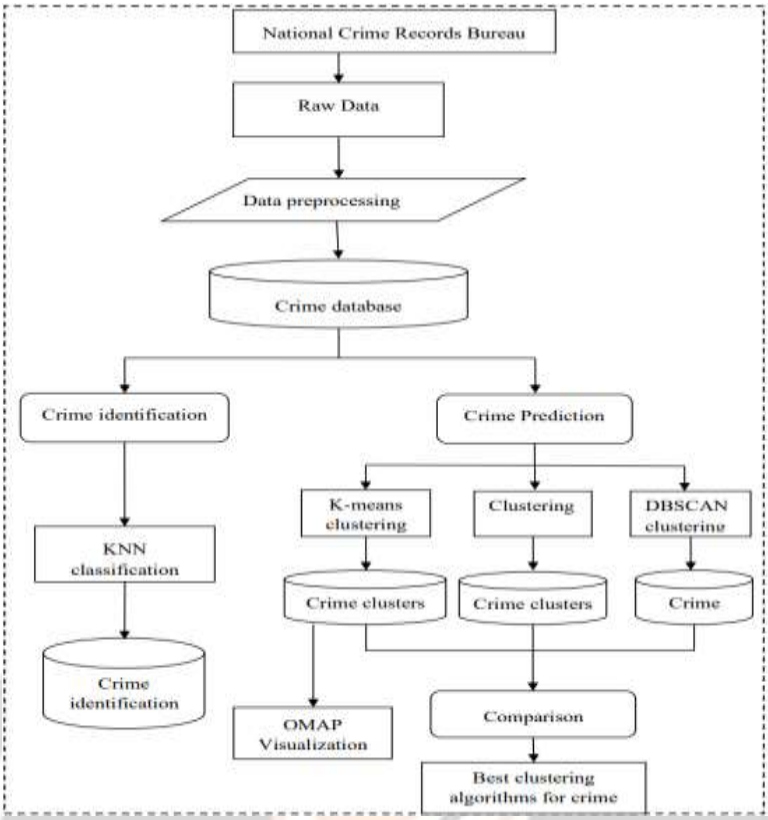

Figure 1 Crime Analysis

4. SUPPORT VECTOR MACHINE

In this study, we categories the CBS data stat line retrieval using the cybercrime detection model implemented as a support vector machine. https://www.cbs.nl/en-gb/our-services/open-data. Training may be provided with the aid of a support vector machine. When the process is complete, a user's expected status as a Genuine or Crime User is based on a number of factors.

Steps:

Step 1: Input of real-time data set.

Step 2: Classification is performed using the Clustering techniques

Step 3: Classification carried out through SVM.

Step 4: Based on the average acquired from the data, Cluster Classification is performed. Also based on new Classes SVM Classifier is conducted fordifferent attributes / predictors/ features is 10

Step 5: For carrying out performance evaluation, various performance metrics are employed such as TP, FP, TN and FN, FAR, ACC, DR, Specificity, Sensitivity, Precision, Recall and Fowlkes-Mallows scoresfor different attributes.

Step 6: Thereafter, following are determined for the training data: cv MSE - mean-squared error for regression via10-fold cross validation, cv MCR - misclassification rateviastratified 10-fold cross validation, cfMat confusion matrix via stratified 10-fold cross validation. SVM Struct. Support Vectors, SVM Struct. Alpha, SVM Struct. Bias, SVMStruct. Support Vectorization is obtained for the training data along withfinding min and max values for the training attributes.

Step 7: With the use of SVM, classification accuracy of 89% is achieved.

SVM Algorithm

Input: Train Data Set - Train, Test Data Set – Test

Output: Cyber threats Classification

Step 1: Read Train Data Set

Step 2: Apply SVM algorithm

Step 3: Generate SVM Model for kernel function

Step 4: Read Test Data Set

Step 5: For each Characteristic in Test Data

Step 6: Extract all the features **Step 7:** Apply SVM algorithm **Step 8:** Return Result of Test Data

Step 9: End

5. SVM CLASSIFIER TRAINING DATA

The SVM classifier uses machine learning techniques to mine cybercrime detection datasets. Information used to train the SVM classifier is shown in Table 1. Table 2 displays the TP, TN, FP, and FN classifications for characteristic 1. This table classifies users into two broad categories: criminal and honest. Classification by TP, TN, FP, and FN for attribute 1 is shown below. The following table classifies each user as either a Criminal User or a Legitimate User.

Table 1: Training Data for SVM classifier

Sl. No.	Training Data set-Average	SVMStructSupportVectors	SVM StructScaleDatashift	SVMStructScaleDatascaleFactor	SVM StructAlpha	SVM Struct Bias	SVMStructSupportVectorization	mean-squared errorfor regression using10-fold cross cvMSE	misclassification rate using stratified 10-foldcross validation: cvMCR	Confusious matrix using stratified 10-foldcross validation: cfMat
1	7.850	0.314	-7.102	0.260	0.625	1.958	6.000	0.003	0.150	0 1 0 12 69 2
2	8.040	0.403	-7.102	0.260	0.625	1.958	10.000	0.003	0.150	0 0 16
3	8.000	0.379	-7.102	0.260	0.625	1.958	11.000	0.003	0.150	
4	7.750	0.317	-7.102	0.260	0.625	1.958	13.000	0.003	0.150	
5	7.730	0.462	-7.102	0.260	0.625	1.958	14.000	0.003	0.150	
6	8.310	0.384	-7.102	0.260	0.625	1.958	15.000	0.003	0.150	
7	7.780	0.743	-7.102	0.260	0.625	1.958	49.000	0.003	0.150	
8	7.670	0.936	-7.102	0.260	-1.250	1.958	50.000	0.003	0.150	
9	8.220	0.749	-7.102	0.260	0.625	1.958	51.000	0.003	0.150	

10	8.650	0.837	-7.102	0.260	-2.500	1.958	52.000	0.003	0.150
11	8.560	0.533	-7.102	0.260	0.625	1.958	53.000	0.003	0.150
12	8.170	0.689	-7.102	0.260	0.625	1.958	54.000	0.003	0.150
13	8.320	0.754	-7.102	0.260	0.625	1.958	55.000	0.003	0.150
14	8.880	0.790	-7.102	0.260	-2.500	1.958	56.000	0.003	0.150
15	8.580	0.319	-7.102	0.260	0.625	1.958	62.000	0.003	0.150
16	8.180	0.907	-7.102	0.260	-2.500	1.958	97.000	0.003	0.150
17	7.320	0.561	-7.102	0.260	0.625	1.958	98.000	0.003	0.150
18	7.230	0.343	-7.102	0.260	0.625	1.958	100.00	0.003	0.150

Table 2: The TP, TN, FN, FP Classification for Attribute

UID	GROUP		Attribute 1					
	ClusterClassification based on Average	GD=0 CD=1	New ClassesSVM Classifier Attribute1	GD=0 CD=1	TP 0	TN 1	FP 11	FN 10
1	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
2	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
3	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
4	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
5	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
6	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
7	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
8	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
9	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
10	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
11	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
12	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
13	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
14	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
15	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
16	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE

17	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
18	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
19	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
20	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
21	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
22	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
23	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
24	Genuine User	0	'Crime User '	1	FALSE	TRUE	FALSE	FALSE
25	Genuine User	0	'Genuine User'	0	TRUE	FALSE	FALSE	FALSE

6. KNN CLASSIFIER

One example of a non-parametric approach used for classification and regression is the KNN algorithm. The input in both cases is the k-th most similar training exemplar in the feature space. KNN regression uses the KNN technique to estimate a continuous variable. Another technique uses a weighted average of the k neighbors closest to the labelled samples when sorted by the inverse of distance. The algorithm operates as follows: Find the metric distance between the query example and the instances with labels using the Euclidean distance. When K is set to 3, the algorithm will look at the three closest neighbors to the new data point to determine what category it belongs to.

KNN Algorithm

Input: Crime Data, Watermark Data

Output: Modified Crime Observation Data

Step 1: Add Crime Profiles (P).

Step 2: Add Crime Observation Data (O).

Step 3: Enter watermark content (W).

Step 4: Convert the watermark data to bytes and find the length of watermarkdata (L).

Step 5: Sort Crime Observation Data (O) Crime wise.

Step 6: I=0

Step 7: For Each Crime Observation Set in (O)

Step 8: Alter the Observation Data's third value such that OD(3) =301 + W(I)

Step 9: Change the OD(1) position = OD(1) position + W(I)

Step 10: I=I+1 11.

If I>=L ThenBreak

End If

1. Next

2. Output the New Crime Data Set.

Methodology Used for KNN Classifier

Classifying Rows into One of Two Groups: Each row of the dataset is randomly assigned to one of the two classes used for training. Table 3 and picture 3 detail the instructions used to train a matrix, cluster the variable group, and plot training rows into two groups, respectively.

Classifying Rows Using the Three Nearest Neighbors: Classifying sample rows using three nearest neighbors rather than one makes use of the same data as in Classifying Rows into One of Two Groups.

Construct a KNN Classifier: For the data sent in from CBS Stat line, we construct a k-nearest neighbor classifier. Classified information is shown in table 4.4 in terms of real data and crime data. The classifier is constructed by using a Classification KNN fit. Names for the predictors are shown as x1, x2, x3, x4, x5, x6, x7, x8, x9, and x10, and the class is denoted by CD, Genuine_ User. There was no Score Transform in the data we gathered. There were one hundred observations, and the distance measure was called "Euclidean" with a neighbor count of 1. Four closest neighbors are used to accomplish the classification, and the resulting reconstitution loss equaled rloss = 0.2000.

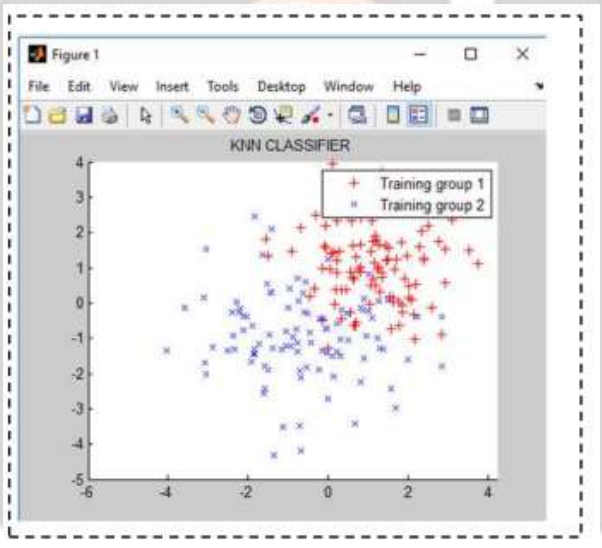

Figure 2: Classifying rows into one of two groups

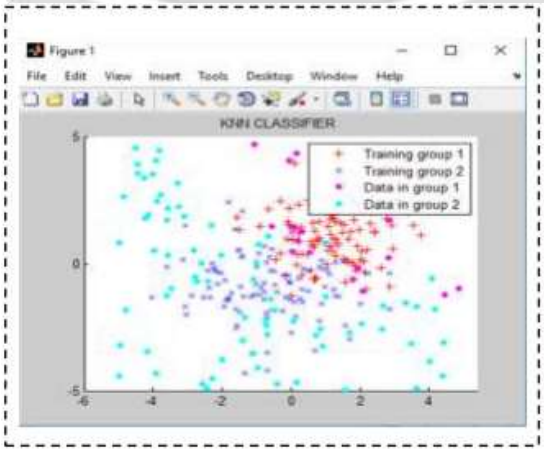

Figure 3: Classifying rows using three nearest neighbors

About 4% of the training data was misclassified by the classifier. The model is then used to construct a cross-validated classifier, which is defined as the classifier with the lowest average loss while making predictions using data that was not utilized during training. kloss = 0.2000. The re-substitution accuracy may be seen as the cross-validated classification accuracy. Therefore, if the new data follows the same distribution as the training data, around 5% of the new data will be misclassified.

Table 3: Classifying rows as group 1 or group 2

UserID	CD/GD	UserID	CD/GD	UserID	CD/GD	UserID	CD/GD	User ID	CD/GD
UID1	1	UID21	1	UID41	2	UID61	2	UID81	1
UID2	2	UID22	2	UID42	2	UID62	2	UID82	2
UID3	2	UID23	2	UID43	1	UID63	2	UID83	2
UID4	2	UID24	2	UID44	2	UID64	1	UID84	2
UID5	1	UID25	2	UID45	2	UID65	1	UID85	2
UID6	1	UID26	2	UID46	1	UID66	2	UID86	1
UID7	2	UID27	2	UID47	2	UID67	1	UID87	1
UID8	2	UID28	1	UID48	1	UID68	2	UID88	1
UID9	2	UID29	2	UID49	2	UID69	1	UID89	2
UID10	1	UID30	2	UID50	2	UID70	2	UID90	2
UID11	1	UID31	2	UID51	2	UID71	2	UID91	1
UID12	1	UID32	2	UID52	1	UID72	2	UID92	1
UID13	2	UID33	1	UID53	1	UID73	2	UID93	2
UID14	2	UID34	2	UID54	2	UID74	2	UID94	2
UID15	2	UID35	2	UID55	1	UID75	2	UID95	2
UID16	1	UID36	1	UID56	2	UID76	1	UID96	2
UID17	2	UID37	2	UID57	2	UID77	2	UID97	1
UID18	1	UID38	2	UID58	1	UID78	2	UID98	2
UID19	1	UID39	2	UID59	1	UID79	2	UID99	2
UID20	2	UID40	1	UID60	2	UID80	2	UID100	2

Table 4: Classification of data as Genuine data or Crime data

UserID	GD/CD	UserID	GD/CD	UserID	GD/CD	UserID	GD/CD	User ID	GD/CD
UID1	GD	UID21	GD	UID41	CD	UID61	GD	UID81	GD

UID2	GD	UID22	GD	UID42	CD	UID62	GD	UID82	GD
UID3	GD	UID23	GD	UID43	CD	UID63	GD	UID83	GD
UID4	GD	UID24	GD	UID44	CD	UID64	GD	UID84	GD
UID5	GD	UID25	GD	UID45	CD	UID65	GD	UID85	GD
UID6	GD	UID26	GD	UID46	CD	UID66	GD	UID86	GD
UID7	GD	UID27	GD	UID47	CD	UID67	GD	UID87	GD
UID8	GD	UID28	GD	UID48	CD	UID68	GD	UID88	GD
UID9	GD	UID29	GD	UID49	GD	UID69	GD	UID89	GD
UID10	GD	UID30	GD	UID50	CD	UID70	GD	UID90	GD
UID11	GD	UID31	GD	UID51	GD	UID71	GD	UID91	GD
UID12	GD	UID32	GD	UID52	CD	UID72	GD	UID92	GD
UID13	GD	UID33	CD	UID53	GD	UID73	GD	UID93	GD
UID14	GD	UID34	CD	UID54	GD	UID74	GD	UID94	GD
UID15	GD	UID35	CD	UID55	GD	UID75	GD	UID95	GD
UID16	GD	UID36	CD	UID56	CD	UID76	GD	UID96	GD
UID17	GD	UID37	CD	UID57	GD	UID77	GD	UID97	CD
UID18	GD	UID38	CD	UID58	GD	UID78	GD	UID98	GD
UID19	GD	UID39	CD	UID59	GD	UID79	GD	UID99	GD
UID20	GD	UID40	CD	UID60	GD	UID80	GD	UID100	GD

Table 5: Performance of different KNN Classification Learner

Total features = 25, Predictors =1, PCA is on							
Classification Learner	Fine KNN	Medium KNN	Coarse KNN	Cosine KNN	CubicKNN	Weighted KNN	Optimizable KNN
Accuracy	86.2%	88.7%	87.5%	51.8%	88.7%	87.5%	87.3%
Total Misclassifications	138	113	125	482	113	125	127

Prediction speed (obs/sec)	~16000	~11000	~14000	~16000	~14000	~11000	~9500
Training Time (sec)	0.80359	0.58836	0.5122	0.46068	0.43864	0.5543	85.099
Number of neighbours	1	10	100	10	10	10	3
Distancemetrics	Euclidean			Cosine	Min Kowski (cubic)	Euclidean	Mahalanobis
Distanceweight	Equal	Equal	Equal	Equal	Squared Inverse	Squared Inverse	Squared Inverse
Standardizedata	True	True	True	True	True	True	True
Optimizeroptions	Hyperparameter options disabled						Hyper parameter options enabled
Feature selection	All features used in the model before PCA						
Misclassifications costs	Cost matrix: default						
For PCA	For PCA training, 2 of 1000 observations were ignored because they contain Infs or NaNs. PCA is keeping enough components to explain 95% variance.After training, 1 component were kept Explained variance per component (in order) :99.9%						

KNN classifier and numerous classification learners are shown in table 4.5. One thousand people made use of a total of 25 qualities. Medium KNN and Cubic KNN achieved 88.7 percentage accuracy, while Optimizable KNN achieved 88.6 percentage accuracy.

Table 6: Performance of Optimizable KNN Classification Learner

Classification Learner	Optimizable KNN
Accuracy	87.3%
Total Misclassifications	127
Prediction speed (obs/sec)	~9500
Training Time (sec)	85.099
Number of neighbours	3
Distance metrics	Mahalanobis
Distance weight	Squared Inverse
Standardize data	True
Hyperparamater searchranges	Number of neighbours: 1-500

Distance metrics	City block, Chebyshev, Correlation, Cosine, Euclidean,Hamming, jaccard, Mahalanobis, Minkowski(cubic), Spearman
Distance Weight	Equal, Inverse, Squared Inverse
Standardized data	True, False
Optimizer Options	Bayesian Optimization
Acquisition function	Expected improvement per second plus
Iterations	30
Training Limit time	False
Feature selection	All features used in the model before PCA
Misclassifications costs	Cost matrix: default
For PCA	For PCA training, 2 of 1000 observations were ignoredbecause they contain Infs or NaNs PCA is keeping enough components to explain 95%variance. After training, 1 component were kept Explained variance per component (in order) :99.9%

Data regarding Optimizable KNN's performance as a Classification Learner (table 6) reveals that it achieves an accuracy of 88.6%.

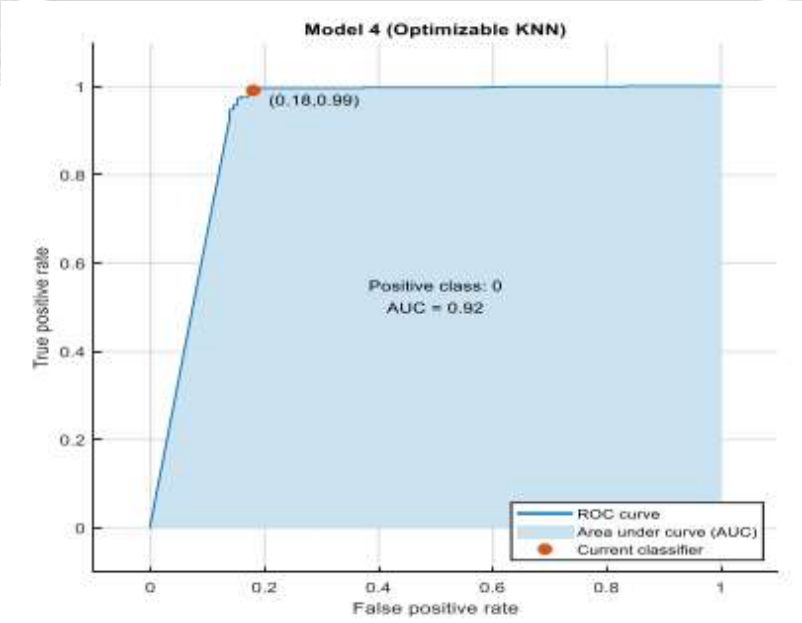

Figure 4: ROC curve for the optimizable KNN classifier

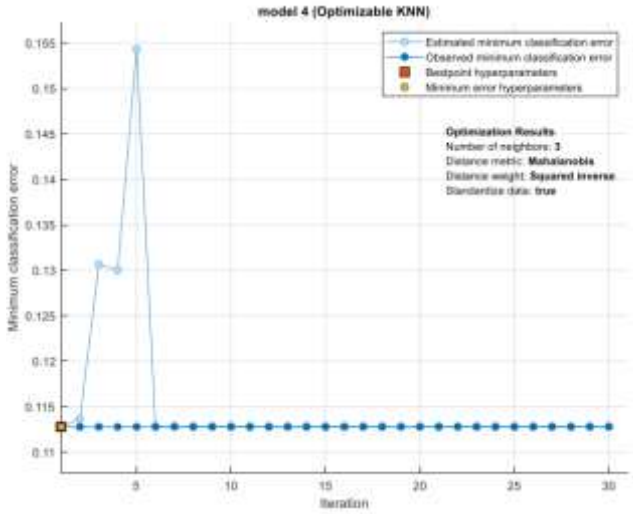

Figure 5: The minimum classification error plot

7. CONCLUSION

In order to identify cybercrime, the proposed study applies both traditional approaches and machine learning techniques to user profiles with different combinations of features. The cybercrime data was categorized using supervised classification techniques, such as the SVM model and the K -NN models. In the same vein, the unsupervised technique of classification used the Gaussian Mixture Model, Cluster Quasi Random through Fuzzy C mean as Clustering, and K Means Clustering.

8. REFERENCE

1. Dr. Farzana begum (2019) beware of cybercrime with awareness - a review. International journal of science and healthcare researchvol.4; issue: 3; july-sept. 2019website: www.ijshr.com
2. Sabillon, regner & cano m., jeimy & serra-ruiz, jordi & cavaller, víctor. (2016). Cybercrime and cybercriminals: a comprehensive study. International journal of computer networks and communications security. 4. 165-176.
3. Damian odunze (2018) cyber victimization by hackers: a criminological analysis. Public policy and administration research www.iiste.org issn 2224-5731(paper) issn 2225-0972(online) vol.8, no.1, 2018
4. Biswal, chandra & pani, dr. Subhendu. (2020). Cyber-crime prevention methodology. 10.1002/9781119711629.ch14.
5. Tabassum, lubna & baker, syed. (2020). Cybersecurity and safety measures.
6. Pascal pouani tientcheu (2021) security awareness strategies used in the prevention of cybercrimes by cybercriminals. August 2021. Https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=12290&context=dissertations
7. Mohamed s. Salhein elbelekia (2020) attitudes of employees towards cybersecurity. Computer information systems, nicosia, 2020. Http://docs.neu.edu.tr/library/6849837766.pdf
8. Arshad, ayesha & ur rehman, attique & javaid, sabeen & ali, tahir & sheikh, javed & azeem, muhammad. (2021). A systematic literature review on phishing and anti-phishing techniques.
9. Markelj, blaz & zgaga, sabina. (2018). Cyber security and cyber criminality of mobile device users in slovenia. Revija za kriminalistiko in kriminologijo. 69. 15-29.