Investigation Assistant with Suspect Anticipating Intelligence

By

Saqueeb Abdullah
18241029
Farah Idid Nibir
15301080
Suraiya Salam
14101130
Akash Dey
15201029

A thesis submitted to the Department of CSE in partial fulfillment of the requirements for the degree of
BSC in CSE

Department of Computer Science and Engineering
BRAC University
August 2019

Declaration

It is hereby declared that

1. The Thesis report submitted is our own original work while completing degree at BRAC University.
2. The report does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The report does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. I/We have acknowledged all main sources of help.

Student's Full Name & Signature:

Farah Idid Nibir
15301080

Saqueeb Abdullah
18241029

Akash Dey
15201029

Suraiya Salam
14101130

Ethics Statement:

I testify on behalf of all co-authors that our Thesis article submitted to the Department of CSE in partial fulfillment of the requirements for the degree of BSC in Computer Science & Engineering.

Title: Investigation Assistant with Suspect Anticipating Intelligence

We hereby state that:

1) This material has not been published in whole or in part elsewhere;
2) The manuscript is not currently being considered for publication in another journal;
3) All authors have been personally and actively involved in substantive work leading to the manuscript, and will hold themselves jointly and individually responsible for its content.

Approval

The thesis titled “Investigation Assistant with Suspect Anticipating Intelligence” submitted by

1. Farah Idid Nibir (15301080)
2. Saqueeb Abdullah (18241029)
3. Suraiya Salam (14101130)
4. Akash Dey (15201029)

Of Summer, 2019 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on 28th of August,2019.

Supervisor: ________________________________

Dr. Md. Ashraful Alam
Assistant Professor, Department of
Computer Science and Engineering.

Departmental Head:
(Chair) ________________________________

Mahbubul Alam Majumdar
Professor & Chairperson, Department of
Computer Science and Engineering.

Acknowledgement

Foremost, we might want to offer our genuine thanks to our Advisor Assistant Prof. Dr.Md. Ashraful Alam for the consistent help for our research, and for his understanding, inspiration, eagerness, and monstrous information also. His direction helped us in all the way of research and composing of this theory. We couldn't have envisioned having a superior consultant and guide for our undergrad research and postulation. A very special gratitude to our families, words can't express that we are so appreciative to our families for the majority of the penances that they have made for our sake. Your supplications for us were what supported us up to this point. We might likewise want to thank the majority of our companions who persuaded us to endeavor towards our objective. In conclusion, we offer our respects and favors to those who upheld us in any regard during the halfway fruition of our proposition.

Abstract

Truly settling crimes have been the privilege of the criminal justice and law enforcement specialists. With the expansion of the utilization of the computerized system to track violations and follow culprits, computer application can be an incredible path for helping the law enforcement officers and analysts to accelerate the way toward fathoming crimes. Criminology is a procedure that is utilized to recognize crime and criminal qualities. The lawbreakers and the crime event probability can be surveyed with the assistance of criminology systems. The criminology helps the police division, the analyst offices and crime branches in distinguishing the genuine attributes of a criminal. The criminology division has been utilized in the procedures of crime detection as far back as 1800. Crimes are a social aggravation and create a reasonable doubt among general people in a few different ways. For quicker explanation of violations, we have chosen to build up a framework called "Investigation Assistant with Suspect Anticipating Intelligence". This system keeps information related to crime situation such as place, method, gender and so on. The principal task of the proposed model is to analyze all the information and provide a short overview of the criminal’s characteristics. The main objective is to discover and anticipate the plausible suspect for the unsolved cases from the criminal records present at the database in the system. Thus, we are attempting to focus on foreseeing the criminal, by the assistance of how the crime had occurred.

Keywords: Computerized system; Machine Learning; Track violations; Suspect Prediction; Criminal Prediction

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Criminal investigation is a multi-faceted, problem-solving challenge. Arriving at the location of a crime, an official is regularly required to quickly settle on basic choices, in some cases including life and death, in view of constrained data in a dynamic domain of dynamic and as yet advancing occasions. After a criminal occasion is finished, the agent is relied upon to save the wrongdoing scene, gather the proof, and devise an insightful arrangement that will prompt the framing of sensible grounds to distinguish and capture the individual or people in charge of the wrongdoing. To address these difficulties, police examiners, through preparing and experience, learn insightful procedures to create analytical plans and organize reactions. [1] Criminal investigation isn't only a lot of errand skills; it is similarly a lot of reasoning aptitudes. To turn into a compelling investigator, these abilities should be consciously comprehended and created to the point where they are purposely drawn in to work through the critical thinking process that is criminal investigation. Prepared reasoning and reaction can be hard to adjust into our own collections since we are altogether molded to be substantially less formal and less proof driven in our regular reasoning. In perspective on this investigation of criminal activity the whole process ends up being incredibly problematic. Our proposed system is going to transform this extensive investigation process into an exceptionally speedy solution ever. In view of the past records of crime and the information base our fabricated framework will foresee the Age Range, Sex, Race and Method that could have been utilized while that criminal behavior event was going on.

1.2 Objective

The aim of this thesis is to compare and evaluate the whole proposed model with 8 classifiers which are Decision Tree, Linear-SVM, Logistic Regression, Naïve-Bayes, K-nearest neighbors (KNN),AdaBoost, Random forest, Neural network. For each model we are indicating correlation with 4 classifiers to foresee the specific key terms that is going to help the investigation officers to explore the case and venture into a final choice of that specific crime activity.

1.3 Thesis Orientation

This thesis book is made out of a sum of seven chapters. Chapter 1 is the present section and presents the point of the proposal, the effects of the criminal investigation and what we have done so far to speed up the process. Chapter2 portrays the past commitments in this field. It portrays various calculations seeing utilized as prescient models for criminal investigation. It additionally portrays the latest works in this field. The constraints of this field are likewise depicted in this section. Chapter3 expresses the proposed model of our exploration; it confirms the dataset we utilized in our examination, the prescient models we chose and how we created results for both when applying Principal Component Analysis. Chapter4 depicts the dataset we chose which is collected from Bangladesh Police Department, subtleties of the dataset alongside the genuine esteemed highlights of every cell is given here. Chapter5 shows the framework execution. It discusses sub segments, for example, Train Test split; the proportion where the dataset utilized in our exploration was part into preparing and testing models Includes determination and gives a concise record of Principal Component Analysis. Chapter6 presents our trial settings and results. A concise record of the exhibition measurements utilized in our examination and the outcomes acquired of different execution measurements of every calculation are outlined and analyzed here, both for with and without applying Principal Component Analysis. Chapter7 outlines our examination and furthermore features the confinements of our exploration. A concise record of things to come works, or steps we mean to take to improve our models or research is likewise expressed here.

1.4 The Basics of Machine Learning

Machine Learning is a process inside man-made consciousness which has a place with the science and designing of making intelligent machines. Mechanized information obtaining centered by AI through the plan and execution of calculations where experimental information is required by calculations. Fundamentally, Systems for learning of a machine is educated by AI relying upon the utilization of likelihood. There are diverse sort of ways have a place with AI. Regulated learning: In directed getting the hang of beginning with the datasets which contains preparing models, which can distinguish themselves through the related level those have. It does it by running information through a learning calculation. The objective of managed learning is, effectively distinguish the new information given to it through the directed learning and utilizing the past informational index and learning calculations can become familiar with the strategy to recognize the information. The calculations working underneath directed learning takes the sources of info that the yield is as of now known for the reason all together that the calculations will make the machine to discover by holding it contrast the specific yield and the definitely realized yield to test for to any degree further mistakes. The machine is then demonstrated thusly. The well-known managed learning calculations incorporate order, slope boosting, expectation and relapse. At that point the model is altered by it subsequently. With such calculations, a machine makes a utilization of managed figuring out how to attempt to do the forecast of name esteems on unlabeled data by misuse proper examples. Regulated learning finds the apparatus in such territories any place the more extended term occasions are normal through the recorded data. Solo learning: Unaided learning considers anyway frameworks will figure out how to speak to explicit information designs in a way that mirrors the connected math structure of the collection of info designs. By stand out from directed learning or fortification learning, there aren't any express objective yields or ecological assessments identified with each information; rather the unattended student acquires contact past inclinations on what parts of the structure of the information should be caught inside the yield. A particular yield isn't having by unaided learning. Finding the structures and examples in the information is pointed by the learning specialist. Semi-Managed learning: Under this AI sort, the machine is shaped.

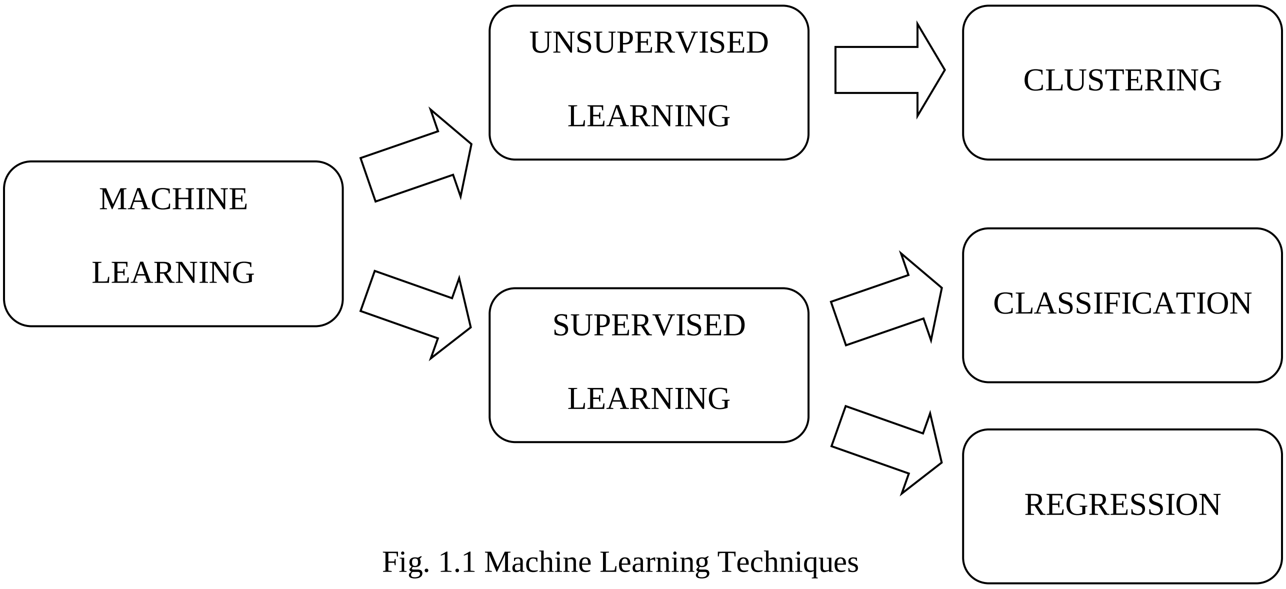


Fig. 1.1 Machine Learning Techniques

Chapter 2

Literature Review

Previously, research regarding classification and predicting criminal activities are not that much available, there are only few of them which using several data mining techniques specially. The procedure of examination and track status of continuous case by foreseeing out the essential suspects based on the records which comprises of summary of the individuals related to the case, previous criminal foundation evidences recuperated from criminal area, and so on. This digitized framework makes the work simple for an official to check the status of the case on the web and even permits him to include the new significant data identified with the case as it's when required. The proposed framework comprises of suspect expectation calculation to foresee and recommend the suspects in the legitimate request [2] On the other hand according to [3] a lot of data is gathered during criminal investigation, data mining is a methodology which can be valuable in this viewpoint. Data mining essentially removes valuable data from enormous measure of criminal information with the goal that potential suspects of the criminal can be distinguished proficiently. Quantities of data mining procedures are accessible for criminal Investigation. Utilization of specific Data mining strategy has more prominent effect on results. So the exhibition of three data mining methods – J48, Credulous Bayes and JRip will be looked at against test criminal and criminal database and best performing calculation will be utilized against definite criminal and criminal database to recognize potential suspects of the criminal.

With due respect to all related work alluded over, this paper compares the performance of the Algorithm; Decision Tree, Linear-SVM, Logistic Regression, Naïve-Bayes, K-nearest neighbors (KNN),AdaBoost, Random forest, Neural network along with the datasets in both determination and examination to get the final result. The objective is to accomplish the most productive calculation to predict the culprit for particular scenery. To do as such, we analyze proficiency and adequacy of those methodologies as far as specific criteria, for example, accuracy, specification of the crime, complexity, eliminating the limitations while predicting the final output with a standardized framework. The main catch of our system is, it is going to predict not only the primary features to find the actual criminal but also a secondary search method to get the probable criminal list to help the officer to find the actual culprit and this leads to a huge boost to the investigation system we strongly believe.

Chapter 3

Proposed Model

Since the world’s bad guys is increasing significantly, we need to step up to stop the occurrences or at least try to minimize where possible. Although the law enforcement organization is working heart and soul, the crime rates going up day by day. Keeping in mind, technology always steps up to the rescue, hence we have come up with a dynamic model. We collected the dataset from Bangladesh Police under the Ministry of Home Affairs of the Government of Bangladesh. The purpose of our model is to predict necessary information about a crime analyzing the initial data of victim, more importantly the crucial clue about the criminal behind it. Our model pre-processes the data with Label Encoder as well as One Hot Encoder, extracts the features and labels, divide the data into 80:20 ratios for training and test data. We are going to train 80 percent of the data and we will test over 20percent of the data to get the higher accuracy in final results. Additionally, we have implemented eight machine learning algorithms. Such as: Logistic Regression, Support Vector Classifier (SVC), AdaBoost Classifier, Naïve Bayes, Random Forest Classifier, Decision Tree Classifier, Artificial Neural Network and K-Nearest Neighbor to find the best fit for our dataset. However, to enhance the result of our model we applied Hyper Parameter Tuning to find the best parameters which deliver the best outcome. Finally, we compare the classifiers with the confusion matrix and accuracy.

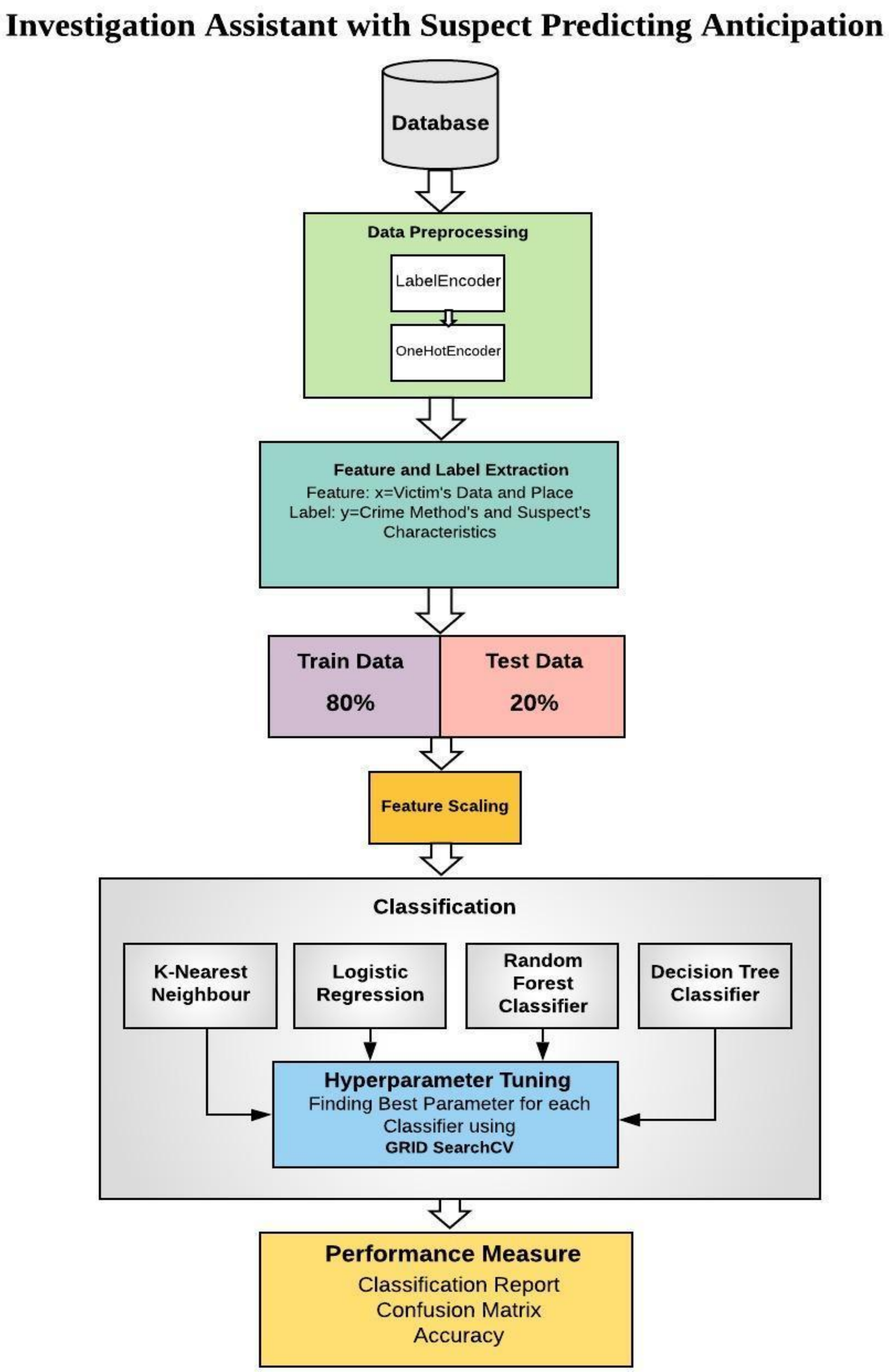


Fig. 3.1 Proposed Model

Chapter 4

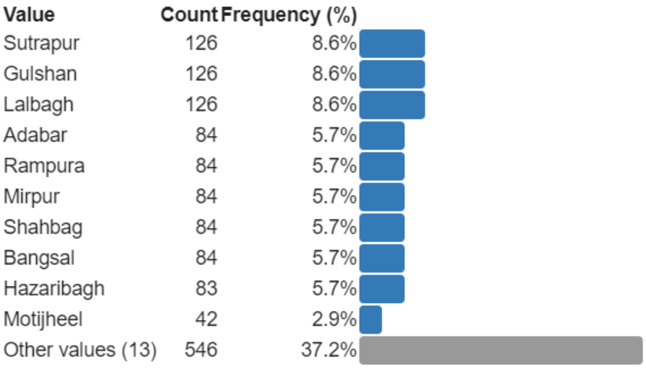
Dataset

Dataset used in this model made our work remarkable, whereas this data has been collected directly from Bangladesh Police under the Ministry of Home Affairs of the Government of Bangladesh. Dataset was difficult to collect as it was classified data and enriched with subtle information. Although, the amount of our data was not colossal, we managed to figure out our model works good enough. Additionally, dataset is consisting of thirteen attributes where the crime records from 1977 are listed snugly. Let's visualize the data with graphical illustrations:

Dataset Demonstration:

Area of the crimes:

Categorical

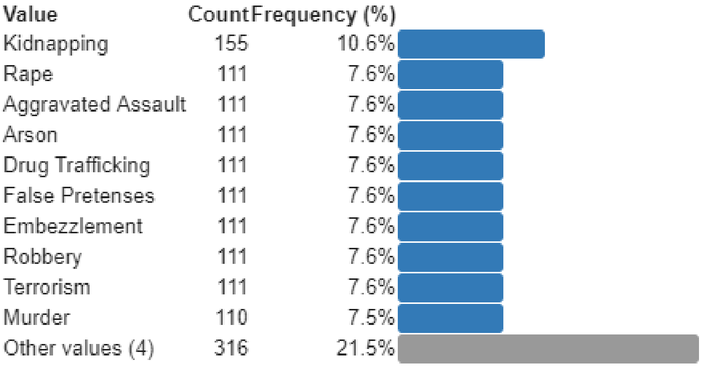


Value	Count	Frequency (%)
Sutrapur	126	8.6%
Gulshan	126	8.6%
Lalbagh	126	8.6%
Adabar	84	5.7%
Rampura	84	5.7%
Mirpur	84	5.7%
Shahbag	84	5.7%
Bangsal	84	5.7%
Hazaribagh	83	5.7%
Motijheel	42	2.9%
Other values (13)	546	37.2%

Fig. 4.1 Area of the Crimes (Categorical)

Crimes:

Categorical

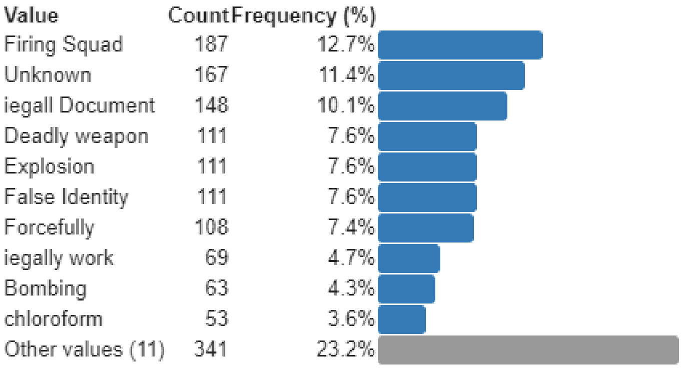


Value	Count	Frequency (%)
Kidnapping	155	10.6%
Rape	111	7.6%
Aggravated Assault	111	7.6%
Arson	111	7.6%
Drug Trafficking	111	7.6%
False Pretenses	111	7.6%
Embezzlement	111	7.6%
Robbery	111	7.6%
Terrorism	111	7.6%
Murder	110	7.5%
Other values (4)	316	21.5%

Fig. 4.2 Crimes (Categorical)

Method of those Crimes:

Categorical



Value	Count	Frequency (%)
Firing Squad	187	12.7%
Unknown	167	11.4%
iegall Document	148	10.1%
Deadly weapon	111	7.6%
Explosion	111	7.6%
False Identity	111	7.6%
Forcefully	108	7.4%
iegally work	69	4.7%
Bombing	63	4.3%
chloroform	53	3.6%
Other values (11)	341	23.2%

Fig. 4.3 Method of Crime (Categorical)

Race:

Categorical

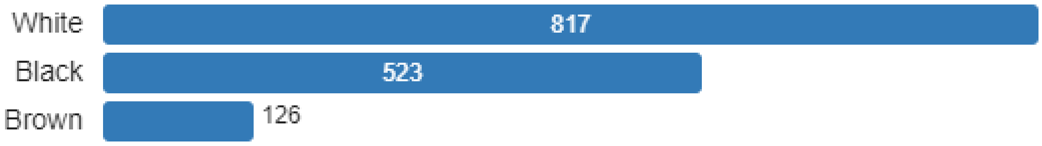
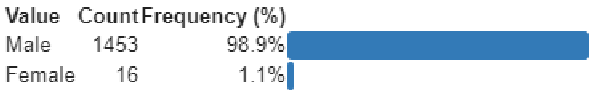


Fig. 4.4 Race (Categorical)

Sex:

Categorical



Value	Count	Frequency (%)
Male	1453	98.9%
Female	16	1.1%

Fig. 4.5 Sex (Categorical)

Criminal Name:

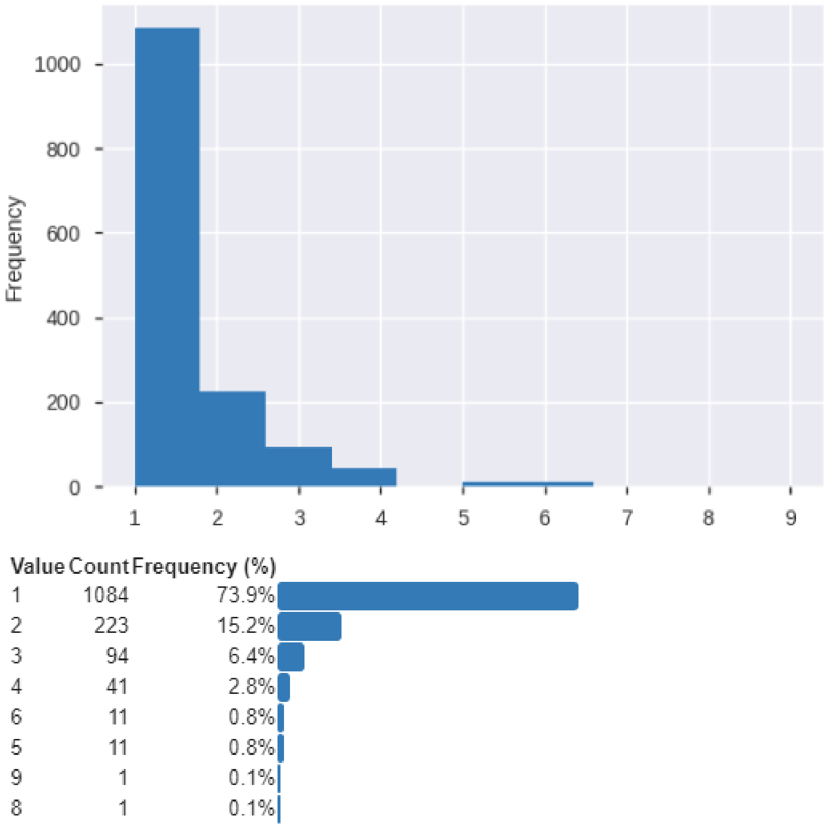
Categorical

Distinct count	1452
Unique (%)	98.8%
Missing (%)	0.0%
Missing (n)	0

Fig. 4.6 Criminal Name (Categorical)

Victim Count:

Numeric Value



Value	Count	Frequency (%)
1	1084	73.9%
2	223	15.2%
3	94	6.4%
4	41	2.8%
6	11	0.8%
5	11	0.8%
9	1	0.1%
8	1	0.1%

Fig. 4.7 Victim Count (Categorical)

Age:

Categorical

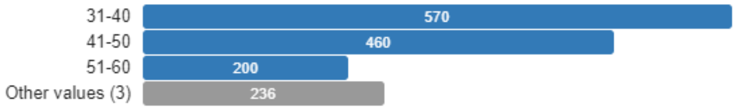


Fig. 4.8 Age (Categorical)

Correlations:

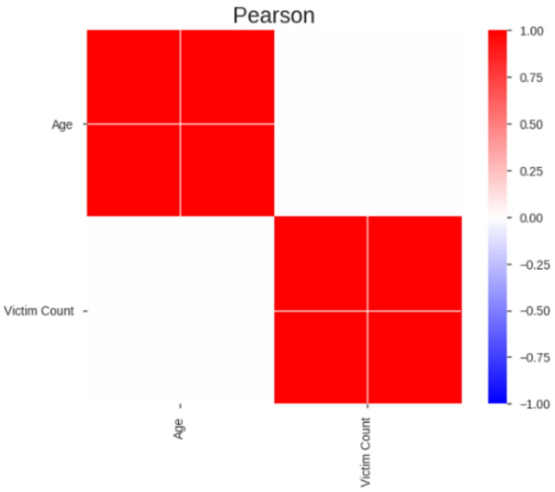


Fig. 4.9.1 Pearson Correlation

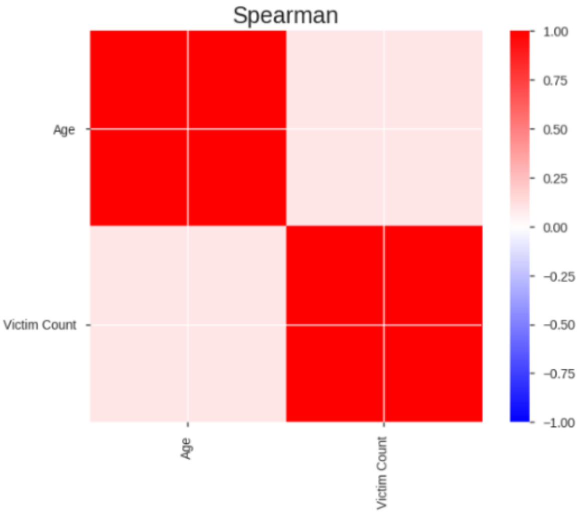


Fig. 4.9.2 Spearman Correlation

Chapter 5

System Implementation

1.1 Feature Extraction

Being up-to-date with all the necessary information of crimes and related data of victims as well as criminals, law enforcement organizations truly try to figure out the black sheep behind any occurrence. Despite of having all the data to some extent, highest efficiency has not been actualized. Machine Learning has some outstanding capability of analyzing data and Supervised Learning gave that opportunity to foretell cabalistic indication which helps us lift ourselves to another dimension in information technology. In this situation, whenever a criminal activity discovered people go to nearest law enforcement organization to report and first the information gathers are mostly; where did it occurred, how many victims got affected, what is the crime that occurred, what was the race and sex of the victims. Reckoning the dataset, we have chosen this useful information to find out a pattern, additionally forecast characteristics of the criminal behind this perpetration.

Features: (Data here is for example)

Victim Count	Victim Sex	Victim Race	Crime	Area
1	Male	Black	Murder	Badda
2	Female	Latino	Terrorism	Lalbagh

Labels: (Criminal’s Characteristics and Method)

Age Range	Sex	Race	Method
41-50	Male	Black	Electrocution
21-30	Male	White	Deadly Weapon

5.2 Data Preprocessing:

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing [4].

There are many ways we can preprocess data before implementation, however we have applied Data Cleaning method. Illustration of Data Cleaning:

5.2.1 Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

1. **Ignore the tuples:**

   This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. **Fill the Missing values:**

   There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value [4].

Moreover, we applied **Label Encoder** and **One Hot Encoder** to some extent.

5.3 Label Encoder:

Label Encoding in Python can be achieved using Sklearn Library. Sklearn provides a very efficient tool for encoding the levels of categorical features into numeric values. Label Encoder encode labels with a value between 0 and n_classes-1 where n is the number of distinct labels. If a label repeats it assigns the same value to as assigned earlier [5].

In our model, we utilized the importance of Label Encoder and applied it to the categorical features.

5.4 One Hot Encoder:

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction [6].

We exert these two techniques to four categorical features; Victim Sex, Victim Race, Crime and Area

5.5Train-Test Split:

Data, in machine learning, in most scenarios are split into training data and testing data (and sometimes to three: train, validate and test), and fit our model on the train data, in order to make predictions on the test data. Training dataset is a part of the actual dataset that we use to train the model. The model sees and learns from this data. Test data, on the other hand, is the sample of data used to provide an unbiased analysis of a final model fit on the training dataset. The Test dataset provides the ideal standard used to evaluate the model. It is used once the model is completely trained [7].

A certain ratio is selected for this training and testing. Generally, the ratio is 80:20. Moreover, we have chosen 80:20 ratios as well, 80% data for training and 20% for testing. As a result, 1176 data is allocated for training and 294 data is allocated for testing. Furthermore, we trained this data with 8 significant classifiers and tested carefully with the test data which delegate our work robustly.

5.6 Algorithms:

5.6.1 Decision Tree

Characterization calculations, The Decision tree calculation attempts to take care of the issue, by utilizing tree portrayal. Each interior hub of the tree relates to a characteristic, and each leaf hub compares to a class name [8].

Decision Tree Calculation Pseudo code

1. Place the best quality of the dataset at the foundation of the tree.

2. Split the planning set into subsets. Subsets should be made so every subset contains data with a comparative motivating force for an attribute.

3. Rehash stage 1 and stage 2 on each subset until you find leaf center points in all of the pieces of the tree. In decision trees, for envisioning a class name for a record we start from the base of the tree. We take a gander at the estimations of the root property with record's characteristic. In view of assessment, we seek after the branch identifying with that value and jump to the accompanying center.

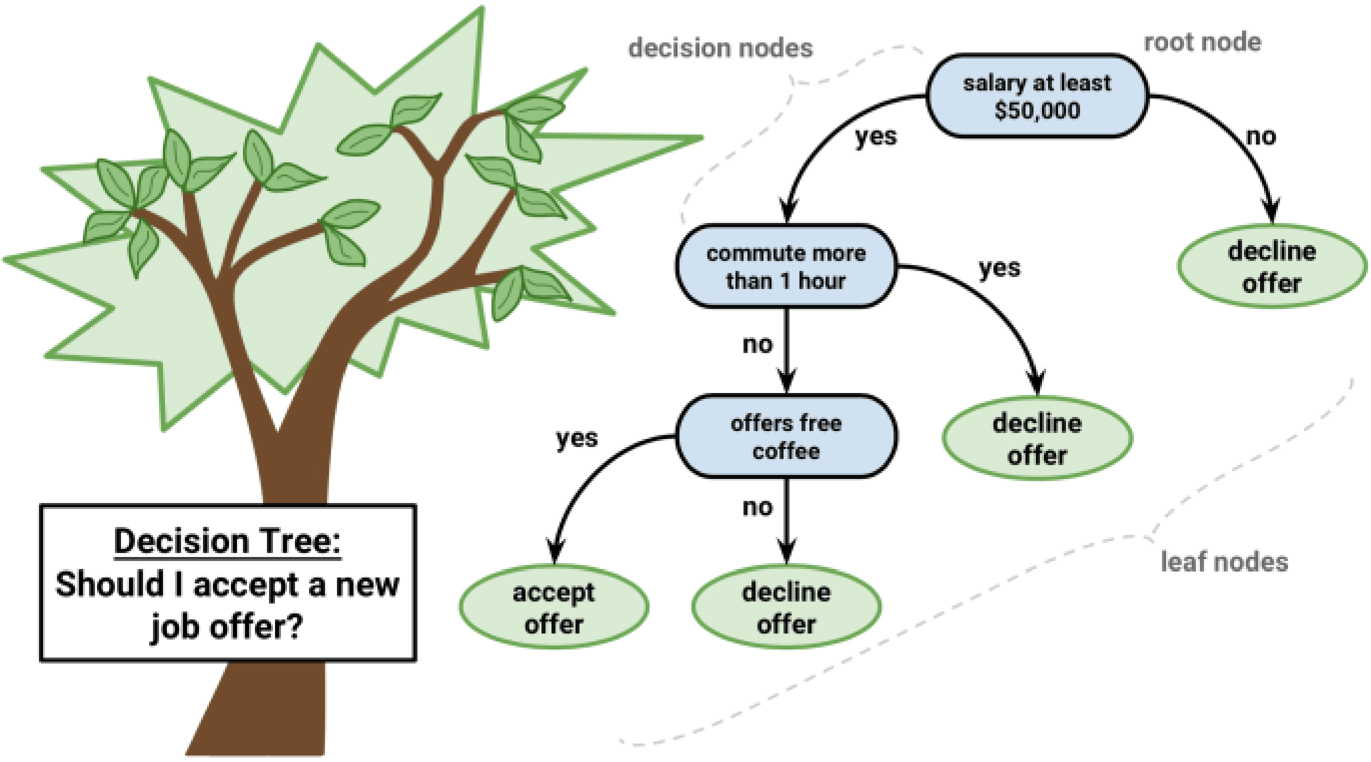


Fig. 5.1 Decision Tree Classifier

We keep contrasting our record's characteristic qualities and other inward hubs of the tree until we arrive at a leaf hub with anticipated class esteem. As we probably are aware how the showed decision tree can be used to anticipate the target class or the value. By and by we should perceiving, how we can settle on the Decision tree model.

5.6.2 Linear SVM

Linear vector model is one of the quickest AI model for grouping multiclass datasets. It fits fine for taking care of numerous down to earth issues. The shortsighted thought is working behind the calculation, which is finding a best fitted line or hyper plane to isolate the information into two classes.

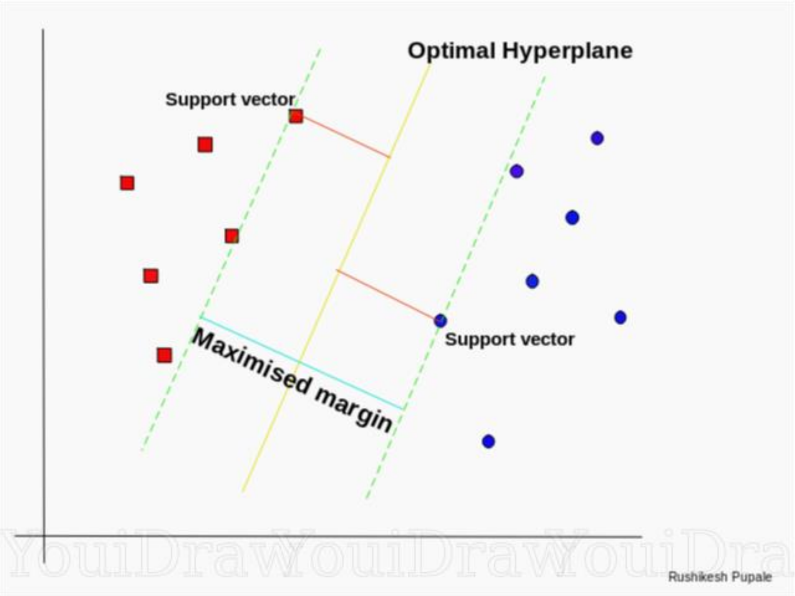


Fig. 5.3 Linear SVM Techniques

In Linear SVM, one can think about that simply drawing a line and separate the class is simple. In any case, we need to remember that to isolate the information into two classes there a large number of line can be drawn, yet we need to discover the perfect one for the specific informational index to isolate the two classes in a most ideal manner conceivable [9]. As per linear SVM to discover the best fitted line, it relies upon Help vector. Bolster

Vectors are the focuses that are nearest to the line and by computing the good ways from support vectors to line we get the Ed**g**e. We will pick ideal hyper plane by taking the greatest Edge among the other hyper planes. Linear SVM can draw the answer for multiclass order with any number of classes.it can likewise manage huge information and multi measurement information great.

5.6.3 Logistic Regression

Logistic Regression is one of the most well-known relapse models. It is additionally broadly utilized calculation in AI field for arranging the datasets. It pursues regulated learning strategy for arrangement. Despite the fact that it is fundamentally the same as with another well-known relapse model (Linear Regression), the significant distinction lies in creating wanted yield between the calculations. In Linear relapse, ordinarily we give free qualities and anticipate the journalist ward esteems as a result via preparing the model with preparing datasets. Interestingly, same methodology is performed in Logistic relapse to anticipate a parallel choice (subordinate qualities). Here, free qualities or elements can be numerical or all out. Besides, in Multinomial Logistic relapse, it utilized for downright ward factors with multiple classes. The general condition of Logistic relapse is:

$$\mathbf{log(p(X)/(1-p(X))) = \beta 0+\beta 1X}$$

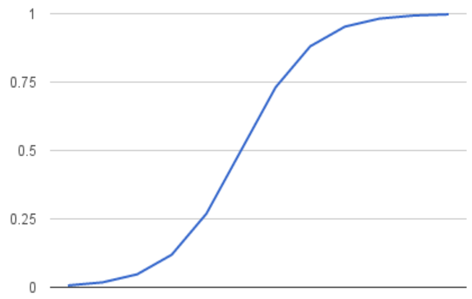


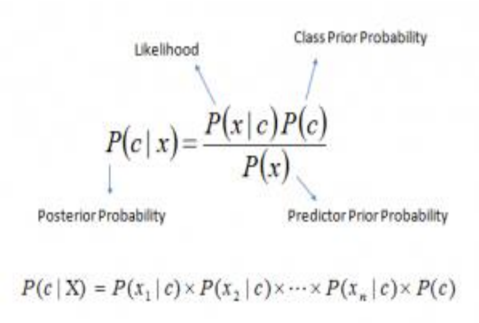
Fig. 5.4 Logistic Regression

Where, p(X) speaks to the needy variable, X speaks to autonomous variable. β0 speaks to the capture and β1 speaks to the slant co-proficient.

Logistic Regression utilizes a more mind boggling capacity than straight called Sigmoid Function. It makes an S-formed bend, which can take any genuine worth number and produce a twofold yield somewhere in the range of 0 and 1, yet never precisely at those points of confinement. The coefficients (Beta qualities b) of the logistic regression calculation need to appraise from preparing information. We utilize most extreme probability estimation for that. Here the needy variable pursues Bernoulli distribution. The best coefficients would foresee a worth exceptionally near 1 for the default class and a worth near 0 for different classes in the model. In Logistic relapse we additionally use Gradient Descent to lessen the cost worth and make it increasingly proficient one.

5.6.4 Naïve Bayes

Naive Bayes is a basic yet successful calculation for forecast of models. This calculation originates from Bayes' hypothesis of likelihood to order protests. Naive Bayes classifiers accept solid, or credulous, autonomy between characteristics of information focuses. Naïve Bayes calculation can be connected for parallel and multi classification order. It could be prepared on little uninformed set, which could end up being a major in addition to point. In addition, it is a lot quicker and climbable. There we discovered many propelled connections between the element variable. Ventures to ascertain Prediction:



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Fig. 5.5 Naïve Bayes Equation

Step1: Converting the informational index into a recurrence table

Step2: Creating probability table by finding the probabilities

Step3: Using Naive Bayesian agreement to compute the back probability for each class. The class with the most surprising back probability is the consequence of figure.

5.6.5 Random Forest

Random Forest is a regulated arrangement calculation settled on of numerous choice trees. They are sets of choice trees, every decision tree produced using a subset of attributes used to characterize a predefined sub trees. Those Decision trees vote on the most proficient method to arrange input information in a predefined model, and the Random Forest bootstraps those votes to choose the best estimate. To counteract over fitting, a typical blemish of decision tree, Random Forest is finished. Random Forest makes a forest where it arranges their nodes and parts haphazardly. The more trees in the forest, the better the outcomes it can create. On the off chance that a preparation dataset is contribution with targets and highlights in the decision tree, some arrangement of principles will be planned that can be utilized to perform expectations.

5.6.6 AdaBoost

AdaBoost, short for Adaptive Boosting is a machine learning boosting algorithm which solves two-class problems and is used to train the weak classifiers to build a strong one. The instances that are misclassified by the previous classifiers, AdaBoost is adaptive in the sense that subsequent classifiers built are tweaked in favor of those.

A training set **S = (X1, Y1),…..,(Xm, Ym)** is taken as input, where each instance **Xi**, belongs to a domain or instance space X, and each label **Yi** belongs to a finite label space Y.

In every round m = 1,…,M, a weak or base learning algorithm is called by AdaBoost which accepts a sequence of training examples S as an input along with a distribution or set of weights over the training example. A weak classifier is computed by a weak learner given such an input, **$\{-1, +1\} \in ht$** where **ht** has the form **$ht : X \rightarrow R$**. Once the weak classifier has been received, AdaBoost chooses a parameter that intuitively measures the importance that it assigns to ht.

Boosting is for using the weaker learner to form a highly accurate prediction rule by calling the weaker ones repeatedly over the training examples on different distributions. All the weights are set equally initially, but the weights of incorrectly classified examples are increased in each round so that on the next iteration, the observations that the previously classifier poorly predicts receive greater weight.

5.6.7 K-Nearest Neighbor

The classifier utilized by KNN works as pursues. Preparing is a clear task to reserve all preparation models. The predefined steady K indicates the quantity of stored passages to be utilized when playing out the order, which are closest in the region to the model being characterized. Through casting a ballot the grouping happens. Indicated range restrains the persistent information space of the area of every factor.

For showcase on working graphical board, the class id that is legitimately returned by the classifier is changed over to a mark. K-NN classifier returns straightforwardly esteem which recognizes a mark.

5.6.8 Artificial Neural Network

ANN was developed as generalizations of biological nervous system mathematical models. Neural network's basic processing elements are called artificial neurons, or just neurons or nodes. In a simplified neuron mathematical model, the impacts of the synapses are represented by link weights that modulate the impact of the related input signals, and a transfer function represents the non-linear trait displayed by the neurons. The impulse of the neuron is then determined as the weighted aggregate of the input signal, converted by the function of transfer. An artificial neuron's learning ability is accomplished by changing the weights to the learning algorithm selected. The basic neural network has three layers: input, hidden and output layers but the data processing can extend over multiple layers of units. When a dataset is given as input, the input nodes turn that into something that is usable for output layer. The computation of the output of each artificial neuron is done by some non-linear function of the sum of its inputs.

Chapter 6

Result Analysis

In order to get the result, we applied above Machine Learning algorithms. As mentioned, we allocated 20% of the data to perform the testing as well as determine the performance. However, we embedded Hyper Parameter Tuning by GridSearchCV to find out the best parameter that outcome better accuracy. For analyzing the result, we considered Classification Report where we can examine with Precision, Recall, F1-Score and Support and then with performance matrices, Such as: Confusion Matrix.

6.1 Classification Report:

The classification report demonstrates a portrayal of the primary order measurements on each class premise. This gives a more profound instinct of the classifier conduct over worldwide precision which can veil utilitarian shortcomings in a single class of a multi class issue. Visual characterization reports are utilized to contrast arrangement models with select models that are "redder", for example have more grounded order measurements or that are progressively adjusted. The metrics are characterized as far as true and false positives, and true and false negatives. Positive and negative for this situation are conventional names for the classes of a binary classification issue.

True positive is the point at which the real class is as positive as the evaluated class. A false positive is the point at which the real class is negative yet the estimated class is positive. Utilizing this terminology, the metrics are characterized as pursues:

6.2 Precision

Precision is the capacity of a classifier not to mark an instance positive that is really negative. For each class it is characterized as the proportion of true positives to the entirety of true and false positives. Said another way, “for all instances classified positive, what percent was correct?”

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

6.3 Recall

Recall is the capacity of a classifier to locate every single positive occurrence. For each class it is characterized as the proportion of true positives to the aggregate of true positives and false negatives. Said another way, "for all occasions that were really positive, what percent was characterized accurately?"

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

6.4 F1 Score

The F1 score is a weighted symphonic mean of accuracy and review with the end goal that the best score is 1.0 and the most exceedingly clobber is 0.0. As a rule, F1 scores are lower than accuracy measures as they implant accuracy and recall into their calculation. As a standard guideline, the weighted normal of F1 ought to be utilized to analyze classifier models, not worldwide exactness.

$$F_1 = \left(\frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}\right) = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} .$$

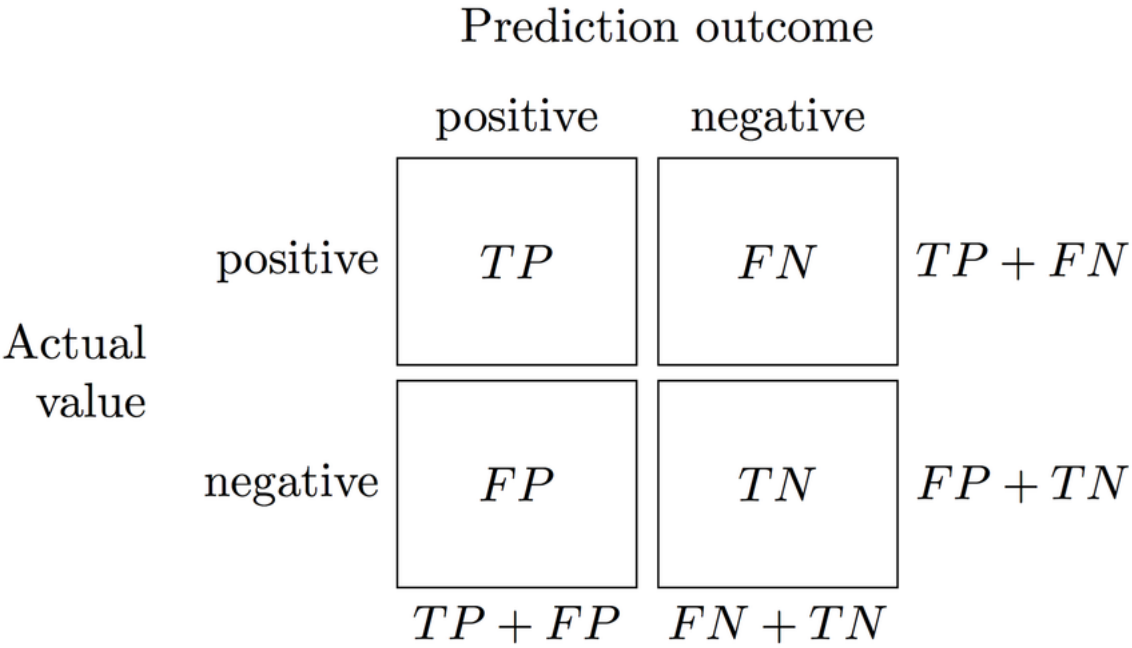
6.5 Support

Support is the quantity of actual events of the class in the predetermined data set. Imbalanced support in the training data may demonstrate auxiliary shortcomings in the announced scores of the classifier and could show the requirement for stratified sampling or re balancing. Backing doesn't change between models however rather analyze the assessment procedure [10].

6.6 Performance Metrics:

6.6.1 Confusion Matrix:

Confusion Matrix, which is also known as error matrix is used in Machine Learning especially in supervised learning, provides a specific table layout. In this layout, we can visualize the performance of an algorithm used in the model. It is apparently the least demanding approach to control the presentation of a classification model by looking at what number of positive occasions is effectively/mistakenly ordered and what number of negative occurrences is accurately/inaccurately grouped. Here, as appeared, the lines speak to the real label while the sections speak to the anticipated label.



		Prediction outcome		
		positive	negative	
Actual value	positive	TP	FN	$TP + FN$
	negative	FP	TN	$FP + TN$
		$TP + FP$	$FN + TN$	

Figure 6.1: Understanding Confusion Matrix

6.7 Model Performance:

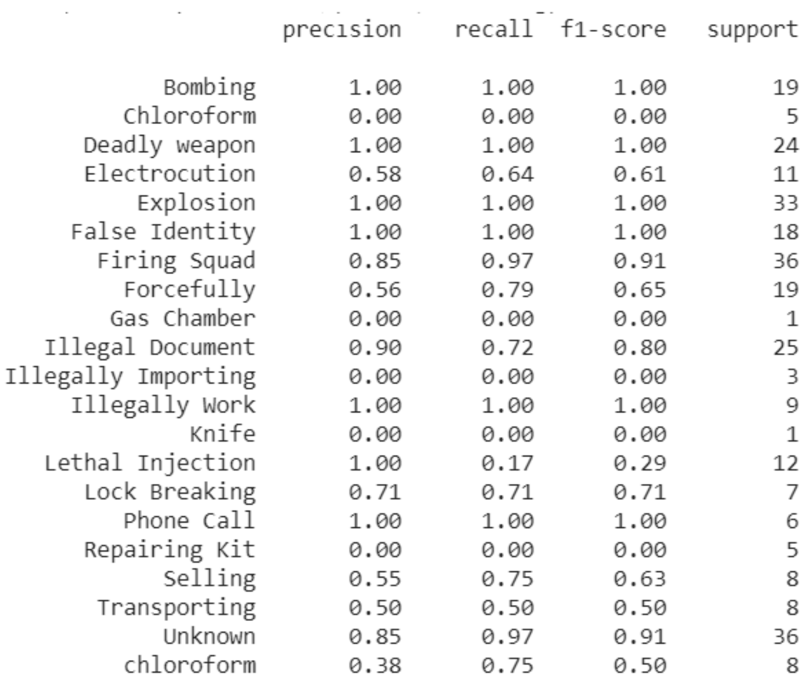
As we have mentioned we have applied eight classifiers with Hyper Parameter Tuning by GridSearchCV for four labels, it has produced four models and we are going to analyses the results with four Classifiers which gave us best results. Our labels are Method, Criminal's Sex, Criminal's Race and Age Range.

6.8 Method Prediction:

6.8.1 Logistic Regression:

Best Parameters: 'max_iter': 50, 'multi_class': 'auto', 'solver': 'newton-cg'

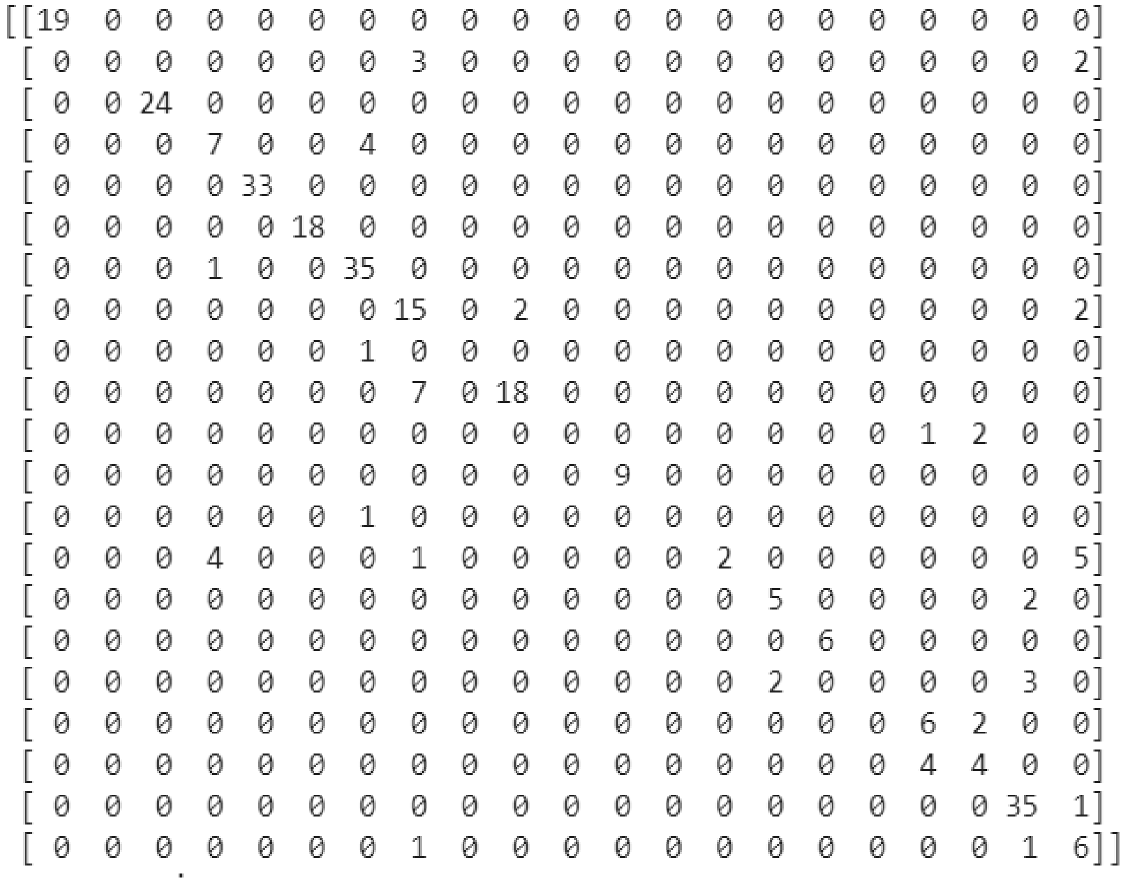
Classification Report:



```
                   precision    recall  f1-score   support

            Bombing       1.00      1.00      1.00        19
         Chloroform       0.00      0.00      0.00         5
      Deadly weapon       1.00      1.00      1.00        24
       Electrocution      0.58      0.64      0.61        11
          Explosion       1.00      1.00      1.00        33
     False Identity       1.00      1.00      1.00        18
       Firing Squad       0.85      0.97      0.91        36
         Forcefully       0.56      0.79      0.65        19
        Gas Chamber       0.00      0.00      0.00         1
    Illegal Document       0.90      0.72      0.80        25
 Illegally Importing       0.00      0.00      0.00         3
     Illegally Work       1.00      1.00      1.00         9
              Knife       0.00      0.00      0.00         1
    Lethal Injection       1.00      0.17      0.29        12
       Lock Breaking       0.71      0.71      0.71         7
         Phone Call       1.00      1.00      1.00         6
       Repairing Kit       0.00      0.00      0.00         5
            Selling       0.55      0.75      0.63         8
       Transporting       0.50      0.50      0.50         8
            Unknown       0.85      0.97      0.91        36
         chloroform       0.38      0.75      0.50         8
```

Figure 6.2: Logistic Regression Classification Report

Confusion Matrix:



```
[[19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0  0  2]
 [ 0  0 24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  7  0  0  4  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 33  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  1  0  0 35  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 15  0  2  0  0  0  0  0  0  0  0  0  0  2]
 [ 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  7  0 18  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  1  0  0  0  0  0  2  0  0  0  0  0  0  5]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  0  0  0  0  2  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  6  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  0  0  0  3  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  6  2  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  4  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 35  1]
 [ 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  6]]
```

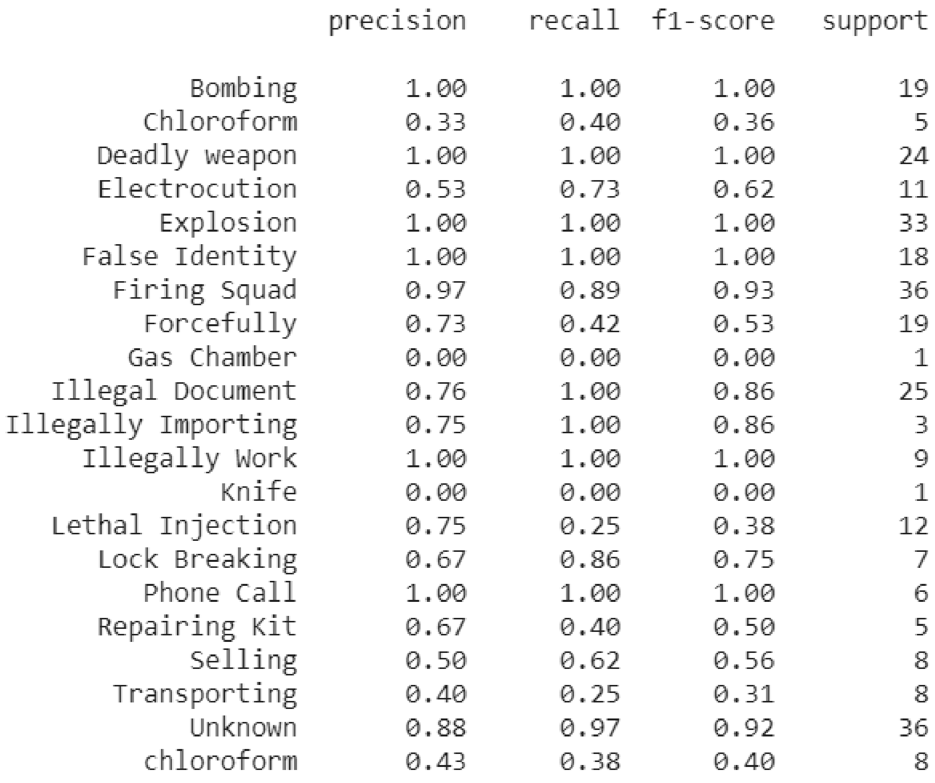
Figure 6.3: Logistic Regression Confusion Matrix

Accuracy: 82.327%

6.8.2 K-Nearest Neighbor:

Best Parameters: 'algorithm': 'auto', 'leaf_size': 40, 'n_neighbors': 7, 'weights': 'distance'

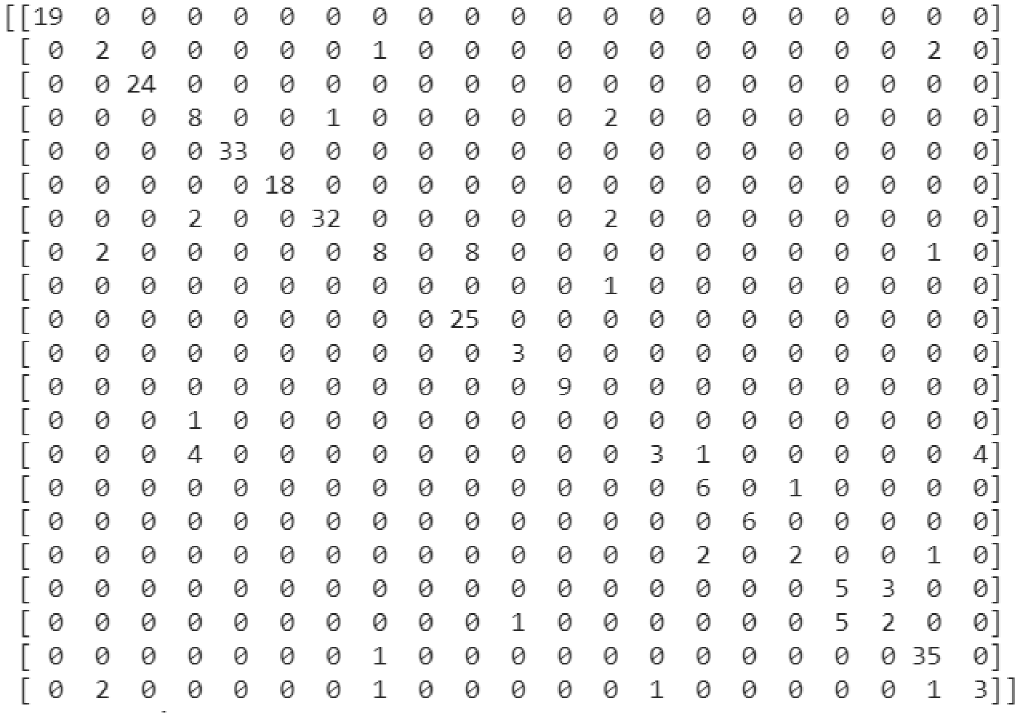
Classification Report:



```
                     precision    recall  f1-score   support

            Bombing       1.00      1.00      1.00        19
         Chloroform       0.33      0.40      0.36         5
      Deadly weapon       1.00      1.00      1.00        24
      Electrocution       0.53      0.73      0.62        11
          Explosion       1.00      1.00      1.00        33
     False Identity       1.00      1.00      1.00        18
       Firing Squad       0.97      0.89      0.93        36
         Forcefully       0.73      0.42      0.53        19
        Gas Chamber       0.00      0.00      0.00         1
   Illegal Document       0.76      1.00      0.86        25
 Illegally Importing       0.75      1.00      0.86         3
     Illegally Work       1.00      1.00      1.00         9
              Knife       0.00      0.00      0.00         1
   Lethal Injection       0.75      0.25      0.38        12
      Lock Breaking       0.67      0.86      0.75         7
         Phone Call       1.00      1.00      1.00         6
      Repairing Kit       0.67      0.40      0.50         5
            Selling       0.50      0.62      0.56         8
       Transporting       0.40      0.25      0.31         8
            Unknown       0.88      0.97      0.92        36
         chloroform       0.43      0.38      0.40         8
```

Figure 6.4: K-Nearest Neighbor Classification Report

Confusion Matrix:



```
[[19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  2  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  2  0]
 [ 0  0 24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  8  0  0  1  0  0  0  0  0  2  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 33  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  2  0  0 32  0  0  0  0  0  2  0  0  0  0  0  0  0  0]
 [ 0  2  0  0  0  0  0  8  0  8  0  0  0  0  0  0  0  0  0  1  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0 25  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  4  0  0  0  0  0  0  0  0  0  3  1  0  0  0  0  0  4]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  6  0  1  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  6  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0  2  0  0  1  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  3  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  5  2  0  0]
 [ 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0 35  0]
 [ 0  2  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0  0  1  3]]
```

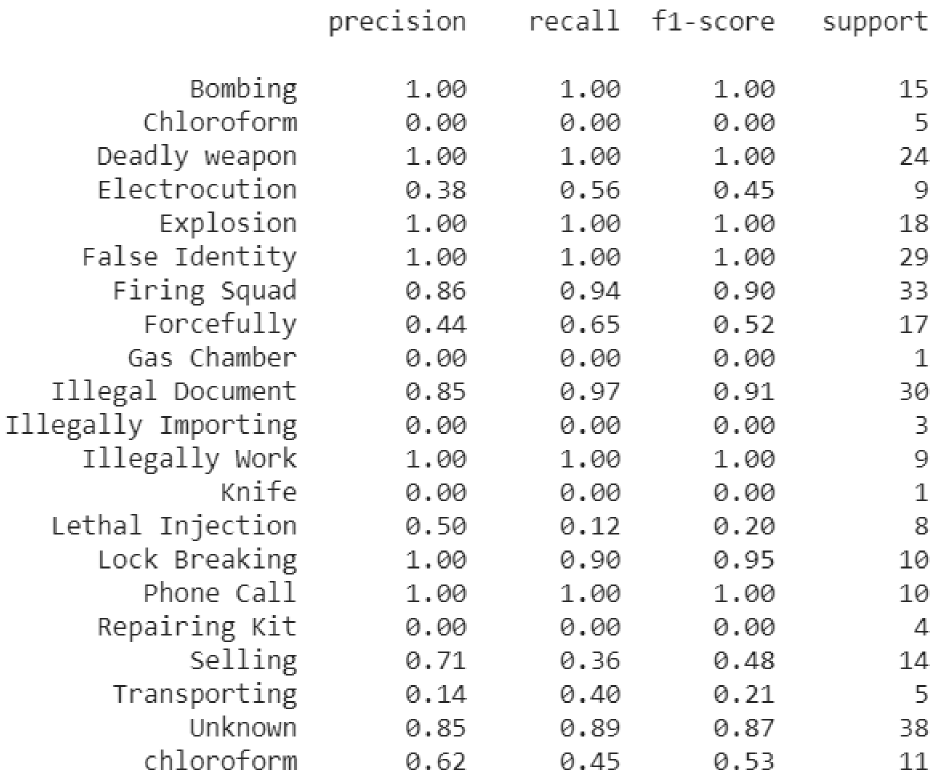
Figure 6.5: K-Nearest Neighbor Confusion Matrix

Accuracy: 82.653%

6.8.3 Artificial Neural Network:

Best Parameters: 'activation': 'identity', 'learning_rate': 'invscaling', 'max_iter': 300, 'solver': 'lbfgs'

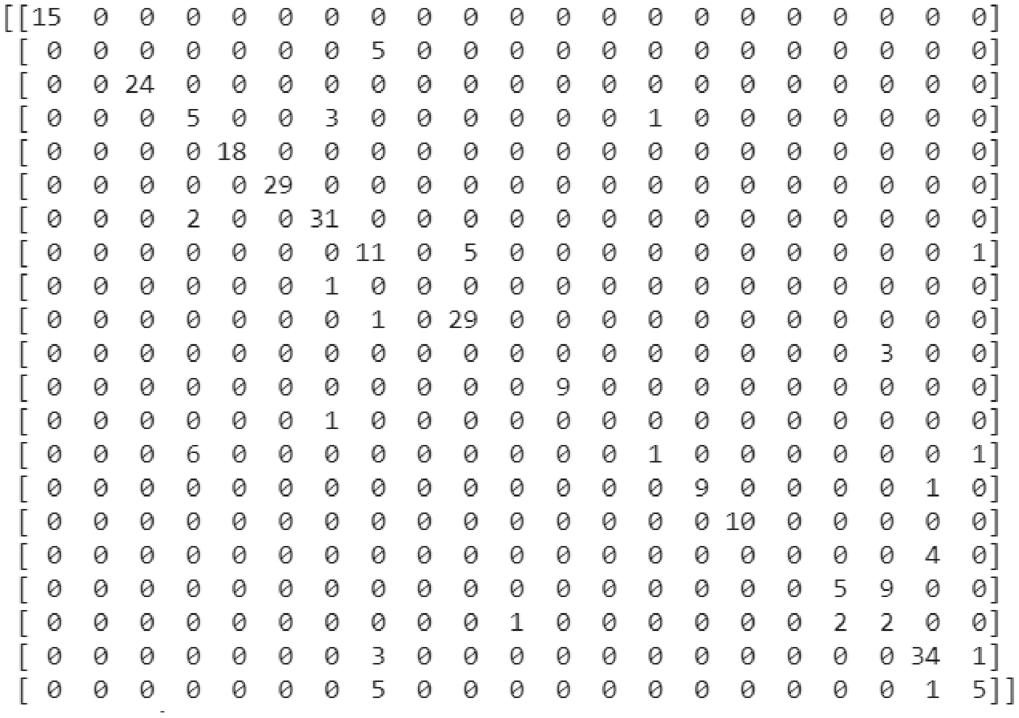
Classification Report:



```
                     precision    recall  f1-score   support

             Bombing       1.00      1.00      1.00        15
          Chloroform       0.00      0.00      0.00         5
       Deadly weapon       1.00      1.00      1.00        24
       Electrocution       0.38      0.56      0.45         9
           Explosion       1.00      1.00      1.00        18
      False Identity       1.00      1.00      1.00        29
        Firing Squad       0.86      0.94      0.90        33
          Forcefully       0.44      0.65      0.52        17
         Gas Chamber       0.00      0.00      0.00         1
    Illegal Document       0.85      0.97      0.91        30
 Illegally Importing       0.00      0.00      0.00         3
      Illegally Work       1.00      1.00      1.00         9
               Knife       0.00      0.00      0.00         1
     Lethal Injection      0.50      0.12      0.20         8
       Lock Breaking       1.00      0.90      0.95        10
          Phone Call       1.00      1.00      1.00        10
       Repairing Kit       0.00      0.00      0.00         4
             Selling       0.71      0.36      0.48        14
        Transporting       0.14      0.40      0.21         5
             Unknown       0.85      0.89      0.87        38
          chloroform       0.62      0.45      0.53        11
```

Figure 6.6: ANN Classification Report

Confusion Matrix:



```
[[15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  5  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 24  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  5  0  0  3  0  0  0  0  0  0  1  0  0  0  0  0  0  0]
 [ 0  0  0  0 18  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 29  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  2  0  0 31  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 11  0  5  0  0  0  0  0  0  0  0  0  0  1]
 [ 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  1  0 29  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  3  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  6  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  1]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  9  0  0  0  0  1  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 10  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  4  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  9  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  2  2  0  0]
 [ 0  0  0  0  0  0  0  3  0  0  0  0  0  0  0  0  0  0  0 34  1]
 [ 0  0  0  0  0  0  0  5  0  0  0  0  0  0  0  0  0  0  0  1  5]]
```

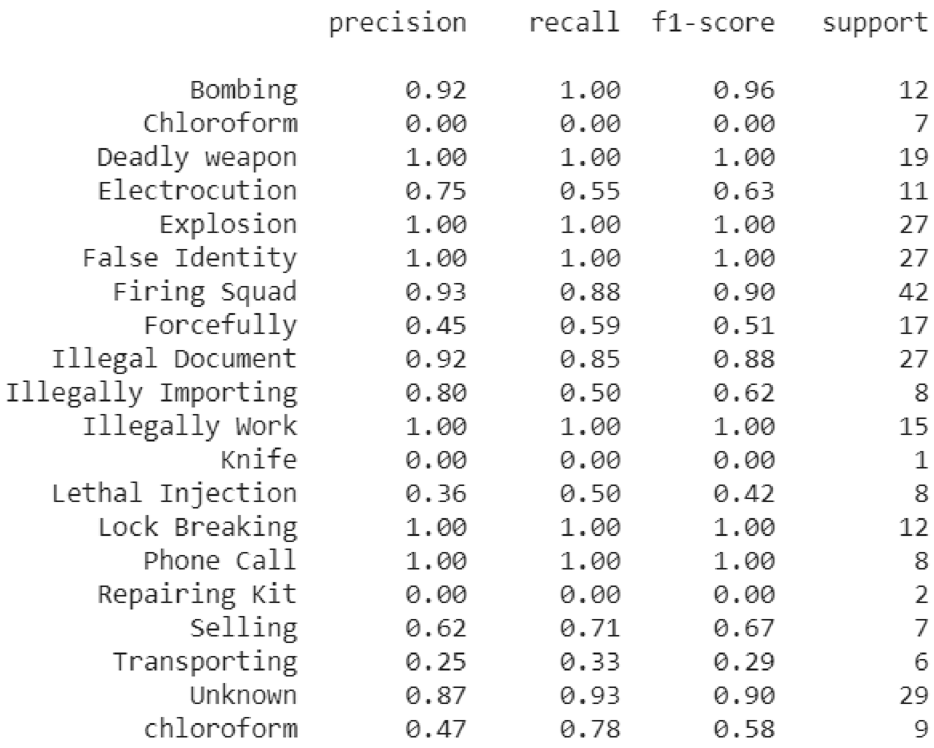
Figure 6.7: ANN Confusion Matrix

Accuracy: 80.612%

6.8.4 Random Forest Classifier:

Best Parameters: 'criterion': 'entropy', 'n_estimators': 30

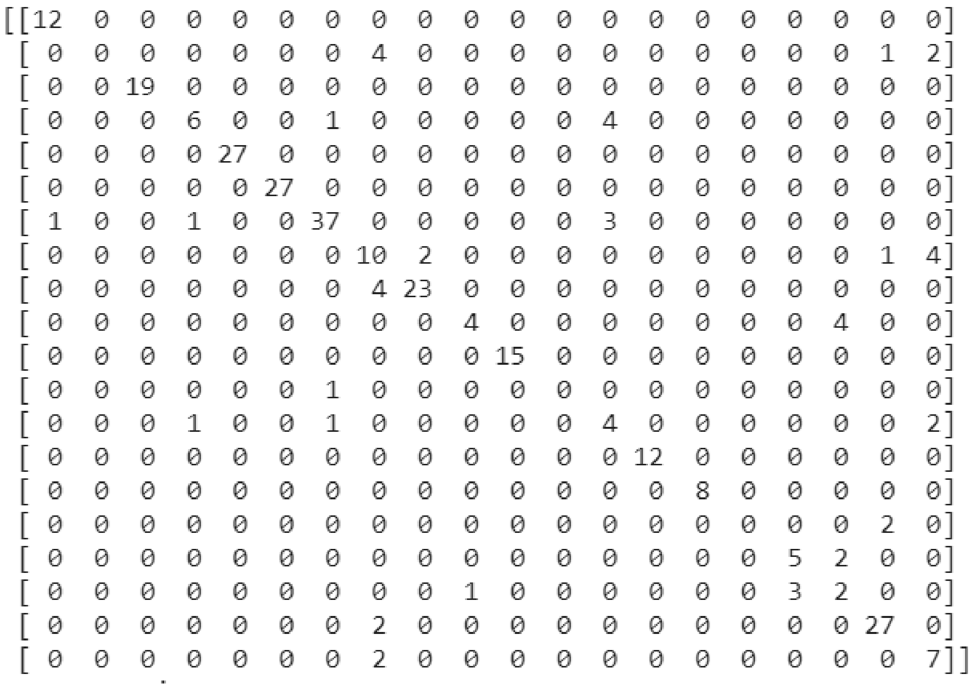
Classification Report:



```
                   precision    recall  f1-score   support

           Bombing       0.92      1.00      0.96        12
        Chloroform       0.00      0.00      0.00         7
     Deadly weapon       1.00      1.00      1.00        19
     Electrocution       0.75      0.55      0.63        11
         Explosion       1.00      1.00      1.00        27
    False Identity       1.00      1.00      1.00        27
      Firing Squad       0.93      0.88      0.90        42
        Forcefully       0.45      0.59      0.51        17
  Illegal Document       0.92      0.85      0.88        27
Illegally Importing       0.80      0.50      0.62         8
    Illegally Work       1.00      1.00      1.00        15
             Knife       0.00      0.00      0.00         1
  Lethal Injection       0.36      0.50      0.42         8
     Lock Breaking       1.00      1.00      1.00        12
        Phone Call       1.00      1.00      1.00         8
     Repairing Kit       0.00      0.00      0.00         2
           Selling       0.62      0.71      0.67         7
      Transporting       0.25      0.33      0.29         6
           Unknown       0.87      0.93      0.90        29
        chloroform       0.47      0.78      0.58         9
```

Figure 6.8: Random Forest Classification Report

Confusion Matrix:



```
[[12  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  0  0  0  1  2]
 [ 0  0 19  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  6  0  0  1  0  0  0  0  0  4  0  0  0  0  0  0  0]
 [ 0  0  0  0 27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 27  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 1  0  0  1  0  0 37  0  0  0  0  0  3  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 10  2  0  0  0  0  0  0  0  0  0  1  4]
 [ 0  0  0  0  0  0  0  4 23  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  4  0  0  0  0  0  0  0  4  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 15  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  1  0  0  1  0  0  0  0  0  4  0  0  0  0  0  0  2]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0 12  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  8  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  2  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  5  2  0  0]
 [ 0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  3  2  0  0]
 [ 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0 27  0]
 [ 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  7]]
```

Figure 6.9: Random Forest Confusion Matrix

Accuracy: 83.333%

Comparison:

Classifiers	**Logistic Regression**	**K-Nearest Neighbour**	**Artificial Neural Network**	**Random Forest Classifier**
Accuracy	**82.653%**	**82.327%**	**80.612%**	**83.333%**

Visualization (Bar Chart):

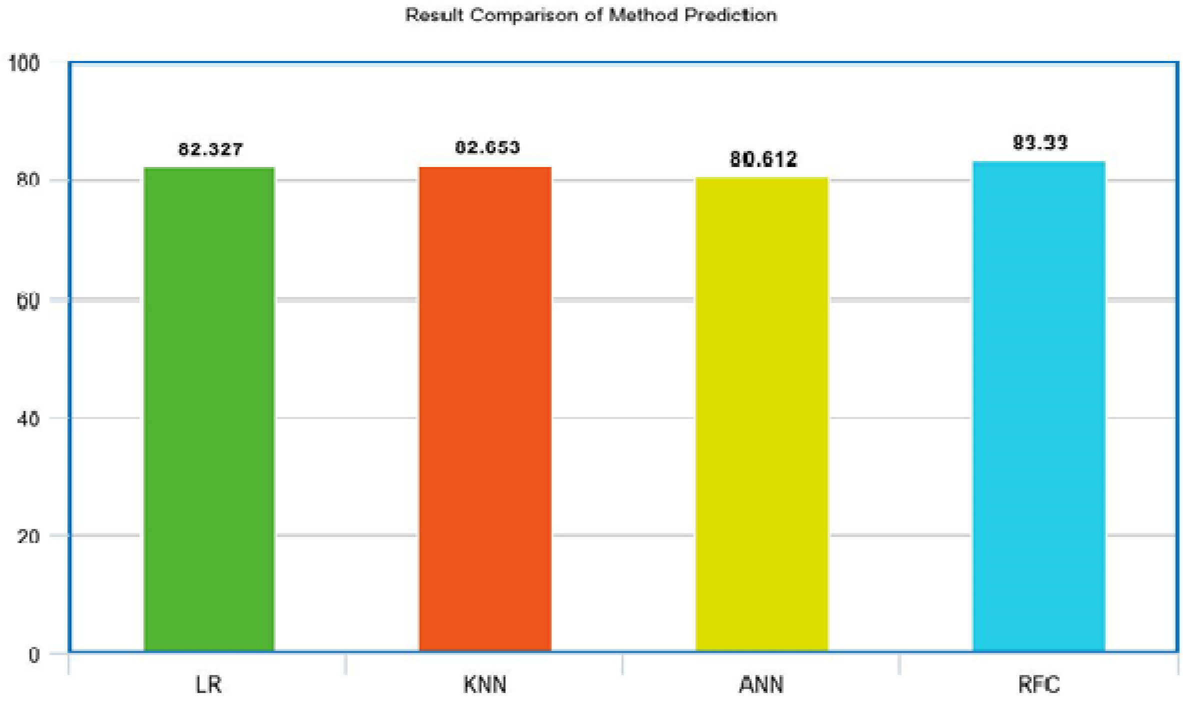


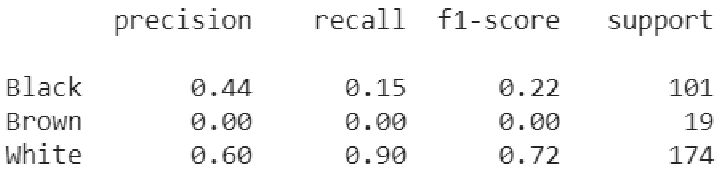
Figure 6.10: Bar Chart of Method Prediction

6.9 Criminal’s Race Prediction:

6.9.1 Logistic Regression:

Best Parameter:'max_iter': 50, 'multi_class': 'ovr', 'solver': 'liblinear'

Classification Report:



```
              precision    recall  f1-score   support

       Black       0.44      0.15      0.22       101
       Brown       0.00      0.00      0.00        19
       White       0.60      0.90      0.72       174
```

Confusion Matrix:

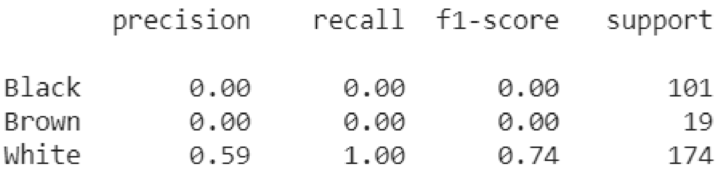
[[45   0 46]
[2   0 7]
[7   0 187]]

Accuracy: 73.275%

6.9.2 Support Vector Machine:

Best Parameter: 'C': 0.1, 'decision_function_shape': 'ovo', 'kernel': 'sigmoid'

Classification Report:



```
              precision    recall  f1-score   support

       Black       0.00      0.00      0.00       101
       Brown       0.00      0.00      0.00        19
       White       0.59      1.00      0.74       174
```

Confusion Matrix:

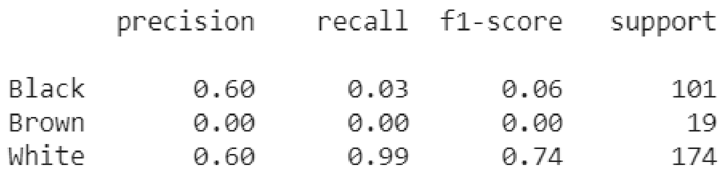
[[0   0 51]
[0   0 9]
[0   0 234]]

Accuracy: 74.359%

6.9.3Artificial Neural Network:

Best Parameter:'activation': 'relu', 'learning_rate': 'adaptive', 'max_iter': 150, 'solver': 'sgd'

Classification Report:



```
              precision    recall  f1-score   support

       Black       0.60      0.03      0.06       101
       Brown       0.00      0.00      0.00        19
       White       0.60      0.99      0.74       174
```

Confusion Matrix:

[[3   0 48]

[0   0 19]

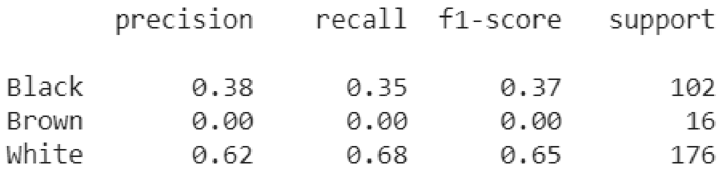
[2   0 222]]

Accuracy: 76.531%

6.9.4 K-Nearest Neighbor:

Best Parameter: 'algorithm': 'ball_tree', 'leaf_size': 20, 'n_neighbors': 10, 'weights': 'uniform'

Classification Report:



```
              precision    recall  f1-score   support

       Black       0.38      0.35      0.37       102
       Brown       0.00      0.00      0.00        16
       White       0.62      0.68      0.65       176
```

Confusion Matrix:

[[56   3 23]

[5   0 11]

[13   3 160]]

Accuracy: 63.888%

Comparison:

Classifier	Logistic Regression	Support Vector Machine	Artificial Neural Network	K-Nearest Neighbour:
Accuracy	**73.275%**	**74.359%**	**76.531%**	**63.888%**

Visualization (Bar Chart):

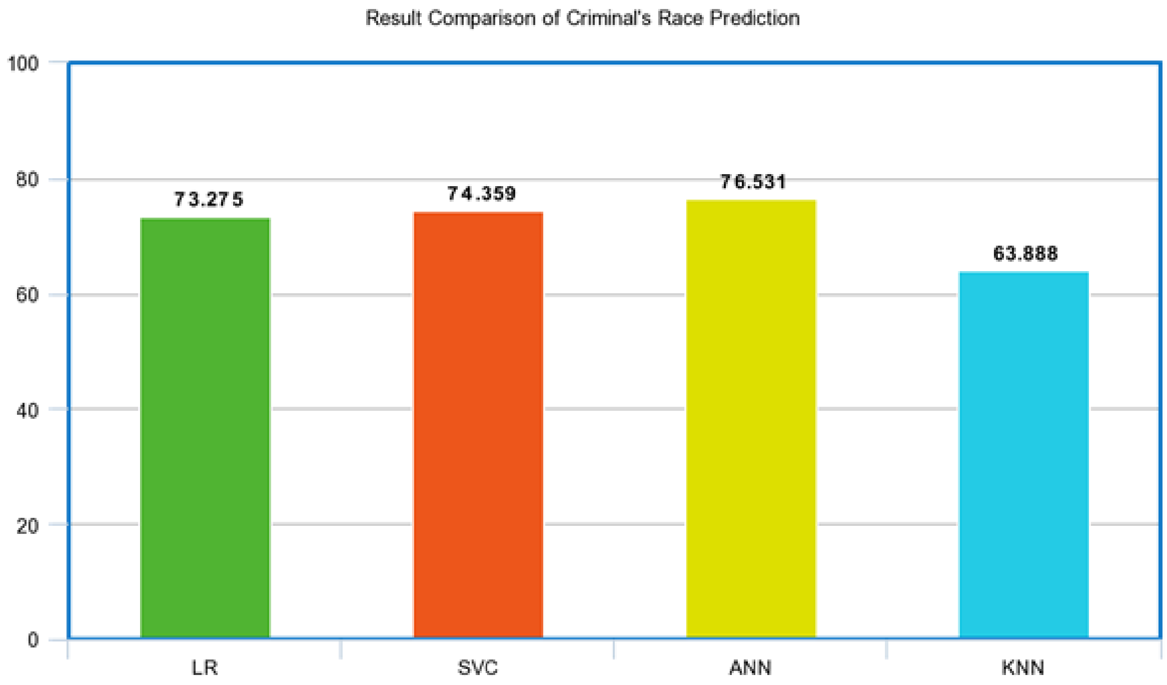


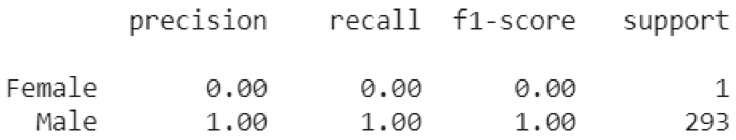
Figure 6.11: Bar Chart of Criminal’s Race Prediction

6.10 Criminal’s Sex Prediction:

6.10.1 Support Vector Machine:

Best Parameter: 'C': 0.1, 'decision_function_shape': 'ovo', 'kernel': 'linear'

Classification Report:



```
              precision    recall  f1-score   support

      Female       0.00      0.00      0.00         1
        Male       1.00      1.00      1.00       293
```

Confusion Matrix:

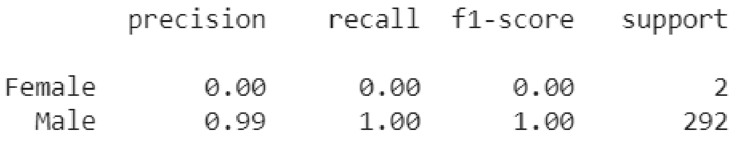
```
[[  0   1]
 [  0 293]]
```

Accuracy: 99.660%

6.10.2 AdaBoost Classifier:

Best Parameter: 'algorithm': 'SAMME', 'n_estimators': 10

Classification Report:



```
              precision    recall  f1-score   support

      Female       0.00      0.00      0.00         2
        Male       0.99      1.00      1.00       292
```

Confusion Matrix:

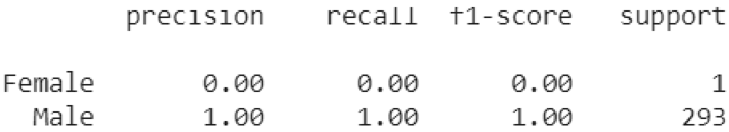
```
[[  0   2]

 [  0 292]]
```

Accuracy: 99.320%

6.10.3 Decision Tree Classifier:

Best Parameter: 'criterion': 'gini', 'max_depth': 60, 'min_samples_split': 5, 'splitter': 'random'

Classification Report:



```
              precision    recall  f1-score   support

      Female       0.00      0.00      0.00         1
        Male       1.00      1.00      1.00       293
```

Confusion Matrix:

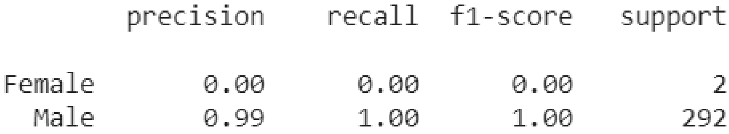
```
[[  0   1]
 [  0 293]
```

Accuracy: 99.660%

6.10.4 Random Forest Classifier:

Best Parameter: 'criterion': 'gini', 'max_depth': 10, 'max_features': 'auto', 'min_samples_split': 5

Classification Report:



```
              precision    recall  f1-score   support

      Female       0.00      0.00      0.00         2
        Male       0.99      1.00      1.00       292
```

Confusion Matrix:

```
[[  0   2]
 [  0 292]]
```

Accuracy: 99.320%

Comparison:

Classifier	Support Vector Machine	AdaBoost Classifier	Decision Tree Classifier	Random Forest Classifier
Accuracy	99.660%	99.320%	99.660%	99.320%

Visualization (Bar Chart):

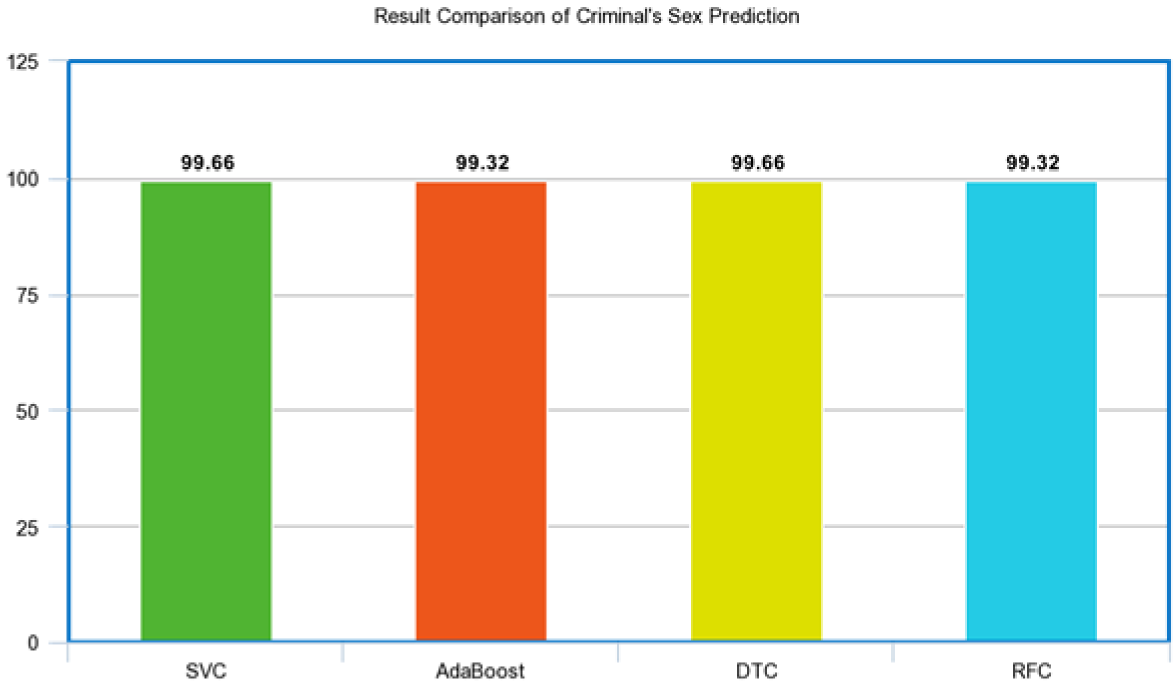


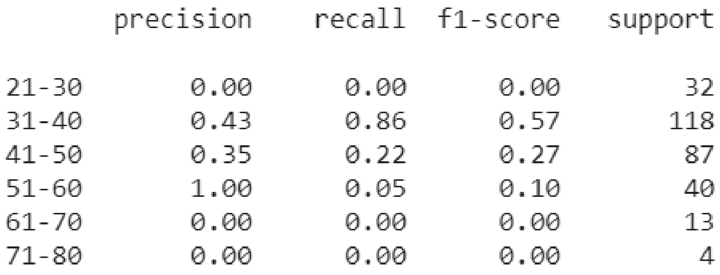
Figure 6.12: Bar Chart of Criminal’s sex Prediction

6.11 Age Range Prediction:

6.11.1 Logistic Regression

Best Parameters: {'max_iter': 50, 'multi_class': 'ovr', 'solver': 'newton-cg'}

Classification Report:



```
              precision    recall  f1-score   support

       21-30       0.00      0.00      0.00        32
       31-40       0.43      0.86      0.57       118
       41-50       0.35      0.22      0.27        87
       51-60       1.00      0.05      0.10        40
       61-70       0.00      0.00      0.00        13
       71-80       0.00      0.00      0.00         4
```

Confusion Matrix:

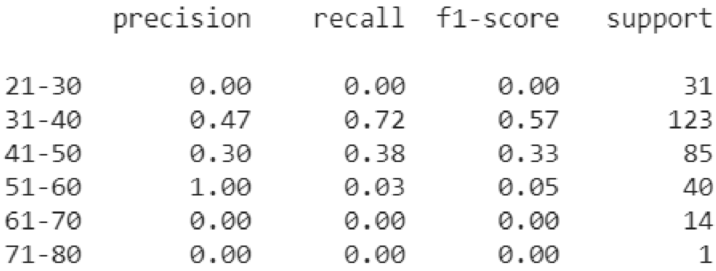
```
[[  019   3   0   0   0]
[  0 102  16   0   0   0]
[  028  59   0   0   0]
[  019   9   12   0   0]
[  0   8   5   0   0   0]
[  0   2   2   0   0   10]]
```

Accuracy: 62.585%

6.11.2Artificial Neural Network:

Best Parameter: 'activation': 'tanh', 'learning_rate': 'adaptive', 'max_iter': 150, 'solver': 'sgd'

Classification Report:



```
             precision    recall  f1-score   support

       21-30       0.00      0.00      0.00        31
       31-40       0.47      0.72      0.57       123
       41-50       0.30      0.38      0.33        85
       51-60       1.00      0.03      0.05        40
       61-70       0.00      0.00      0.00        14
       71-80       0.00      0.00      0.00         1
```

Confusion Matrix:

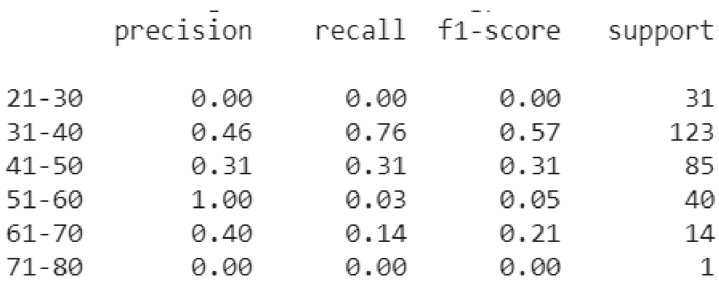
```
[[ 0 11 0  0  0  0]
 [ 0 88 35  0  0  0]
 [ 0 23 52  0  0  0]
 [ 0 18 21  31  0  0]
 [ 0  6  8  0  0  0]
 [ 0  0  1  0  0  0]]
```

Accuracy: 58.163%

6.11.3Support Vector Machine:

Best Parameter:'C': 0.3, 'decision_function_shape': 'ovo', 'kernel': 'poly'

Classification Report:



```
       precision    recall  f1-score   support

21-30       0.00      0.00      0.00        31
31-40       0.46      0.76      0.57       123
41-50       0.31      0.31      0.31        85
51-60       1.00      0.03      0.05        40
61-70       0.40      0.14      0.21        14
71-80       0.00      0.00      0.00         1
```

Confusion Matrix:

```
[[ 0 25  6  0  0  0]
 [ 0 94 8  0  1  0]
 [ 0 39 46  0  0  0]
 [ 0 0 8  31  1  0]
 [ 0  7  5  0  22  0]
 [ 0  0  0  0  1  0]]
```

Accuracy: 65.646%

6.11.4Decision Tree Classifier:

Best Parameters: 'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 5, 'splitter': 'random'

Classification Report:

```
[[ 5 13  4  1  0]
 [ 4 87 10  2  0]
 [ 3 30 54  2  0]
 [ 1 1 13  44  0]
 [ 0  9  9  0  2]]
```

Accuracy: 65.306%

Comparison:

Classifier	Logistic Regression	Artificial Neural Network	Support Vector Machine	Decision Tree Classifier
Accuracy	**62.585%**	**58.163%**	**65.646%**	**65.306%**

Visualization (Bar Chart):

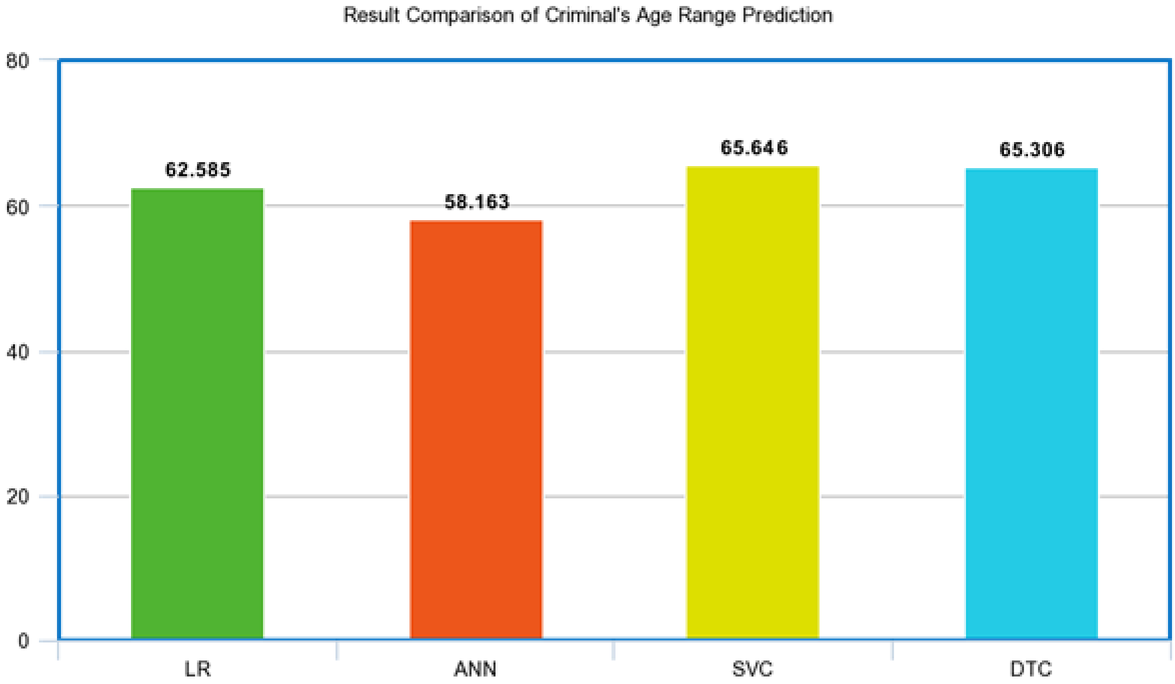


Figure 6.13: Bar Chart of Criminal’s Age Prediction

Chapter 7

Conclusion

1.1 Summary

This paper is based on utilization of information and technology system for identifying crime patterns and predict the suspect which concentrated on building up a system call " Investigation Assistant with Suspect Anticipating Intelligence " for the law and enforcement agencies of Bangladesh. An endeavor has been made in this paper to propose a tracker framework system which adequate to perceive the criminal through the exchange with the unfortunate casualties to help the law authorization exercises. Exactness and precision of the proposed framework is very reliant on the view of the onlookers.

7.2 Limitations

Despite of getting significantly good results out of this research, we got some drawbacks, solving which could enhance the outcome more lucrative in an illustrious way. Firstly, large dataset could have been a dazzling addition to our project, though we got average of 80% accuracy predicting four labels. As we mentioned the data from a law enforcement organization like Bangladesh Police, was hard to collect as it is classified for security reasons. More observation over a large time frame allows more approaches to analysis, particularly enabling predictive analytics and machine learning approaches. It could contain statistical soundness of the results provided by any statistical analysis. One of the major limitations of our framework is, if we do not get the victim’s characteristics or features that we are working with to predict the final result then our system cannot predict the result. Furthermore, lack of some classifiers not working with all other parameters has been seen working for this research while Hyper Parameter Tuning. This could ennoble our research outcome with a huge margin. However, Hyper Parameter Tuning is a time-consuming process as it searches for every possible parameter by resulting in the accuracy to determine best parameter.

7.3 Future Work

As technology infiltrates each part of our lives, it is no big wonder that comprehending crime with predicting the suspect can be another Futuristic System for us. Right off the beginnings we are planning to set up the project appropriately and test it with larger amount of genuine records of crime. At the point we will get the expected outcome, we will try to make an arrangement with the Major Law and enforcement offices of our country to give our System a decent business esteems. From that point forward, we have a plan to incorporate "Picture Search" option into our system with storing the Finger prints of the specific criminal into our database. Moreover, processing crime scenes pictures with the help of image processing and autopsy report analysis will be another major step to take this system to another dimension. In addition, to make out application more productive, we will likewise dispatch a Web-Based Application as well as an Android application for the Investigation officers to continue a case in Emergency premise.

Reference

[1] Plecas, R. G. (2017). *Introduction to Criminal Investigation: Processes, Practices and Thinking*. Darryl Plecas, University of the Fraser Valley: BCcampus.

[2] Akash Desai, G. (2017). Criminal Investigation Tracker with Suspect Prediction. *IJESC*.

[3] Nitin Sakhare, S. J. (2014). Criminal Identification System Based On Data Mining. *Recent Trends in Engineering & Technology* (p. 1). 2014: Elsevier Publication.

[4] *Data Preprocessing in Data Mining - GeeksforGeeks*. (n.d.). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[5] Raheel, S. (2018, November 9). *Choosing the right encoding method- Label vs OneHot Encoder*. Retrieved from Towards Data Science: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b

[6] *What is One Hot Encoding? Why and when do you have to use it?* (n.d.). Retrieved from Hackernoon: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

[7] Bronshtein, A. (2017). Train/Test Split and Cross Validation in Python. Understanding Machine Learning.

[8] *Dataaspirant.* (2017). Retrieved from Dataaspirant: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[9] Pupale, R. (2018). *Support Vector Machines(SVM) — An Overview*. Retrieved from Towards data science: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[10] *Classification Report - yellowbrick 0.9.1 documentation*. (2016). Retrieved from Yellow Brick: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

[11] *Classification Report - yellowbrick 0.9.1 documentation*. (2016). Retrieved from Yellow Brick: https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

[12] *Dataaspirant.* (2017). Retrieved from Dataaspirant: https://dataaspirant.com/2017/01/30/how-decision-tree-algorithm-works/

[13] Akash Desai, G. (2017). Criminal Investigation Tracker with Suspect Prediction. *IJESC*.

[14] Bhandari, I., Colet, E., Parker, J., Pines, Z., Pratap, R., & Ramanujam, K. (1997, March). Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery*, 121-125.

[15] Brian, S. (2012, January 25). The Problem of Shot Selection in Basketball. *PLoS One*.

[16] Bronshtein, A. (2017). Train/Test Split and Cross Validation in Python. Understanding Machine Learning.

[17] Choudhury, R. D., & Bhargava, P. (2007). Use of Artificial Neural Networks for Predicting the Outcome. *International Journal of Sports Science and Engineering, 1*(2), pp. 87-96.

[18] *Data Preprocessing in Data Mining - GeeksforGeeks*. (n.d.). Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[19] Duckworth, F., & Lewis, T. (1999). *Your Comprehensive Guide to the Duckworth/Lewis Method for Resetting Targets in One-day Cricket.* University of the West of England.

[20] Nitin Sakhare, S. J. (2014). Criminal Identification System Based On Data Mining. *Recent Trends in Engineering & Technology* (p. 1). 2014: Elsevier Publication.

[21] Plecas, R. G. (2017). *Introduction to Criminal Investigation: Processes, Practices and Thinking.* Darryl Plecas, University of the Fraser Valley: BCcampus.

[22] Pupale, R. (2018). *Support Vector Machines(SVM) — An Overview.* Retrieved from Towards data science: https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[23] Raheel, S. (2018, November 9). *Choosing the right encoding method- Label vs OneHot Encoder*. Retrieved from Towards Data Science: https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b

[24] Sankaranarayanan, V. V., Sattar, J., & Lakshmanan, L. S. (2014). Auto-play: A data mining approach to ODI cricket simulation and prediction. *SIAM International Conference on Data Mining*, (p. 1064).

[25] Tulabandhula, T., & Rudin, C. (2014). Tire Changes, Fresh Air, and Yellow Flags: Challenges in Predictive Analytics for Professional Racing. *Big data*.

[26] *What is Data Preprocessing? - Definition from Techopedia*. (n.d.). Retrieved from Techopedia: https://www.techopedia.com/definition/14650/data-preprocessing

[27] *What is One Hot Encoding? Why and when do you have to use it?* (n.d.). Retrieved from Hackernoon: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f