**CRIME DATA ANALYSIS AND PREDICTING VICTIMS USING MACHINE LEARNING**

**A PROJECT REPORT**

***Submitted by***

**SILAMBARASAN K (810015205076)**

**UMAPATHY B (810015205092)**

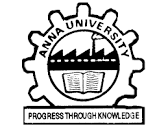
***in partial fulfillment for the award of the degree***

***of***

**BACHELOR OF TECHNOLOGY**

***in***

**INFORMATION TECHNOLOGY**



**UNIVERSITY COLLEGE OF ENGINEERING, BIT CAMPUS,**

**TIRUCHIRAPPALLI-620024**

**ANNA UNIVERSITY::CHENNAI 600 025**

**APRIL 2019**

**BONAFIDE CERTIFICATE**

Certified that this project report “**CRIME DATA ANALYSIS AND PREDICTING VICTIMS USING MACHINE LEARNING**” is the bonafide work of “**SILAMBARASAN K (810015205076)** and **UMAPATHY B (810015205092)**“ who carried out the project work under my supervision.

**Dr. D.VENKATESAN Mrs. K. AMBIKA**

Assistant Professor & Head Assistant Professor

Department of CSE&IT Department of CSE

Anna University, BIT Campus Anna University, BIT Campus

University College of Engineering University College of Engineering

Tiruchirappalli - 620 024 Tiruchirappalli - 620 024

Submitted for the ANNA UNIVERSITY Viva-voce Examination held on ……………….

**Internal Examiner External Examiner**

**DECLARATION**

We hereby declare that the work entitled **“CRIME DATA ANALYSIS AND PREDICTING VICTIMS USING MACHINE LEARNING”** is submitted in partial fulfillment of the requirements for the award of the degree in B.Tech.-Information Technology, University College of Engineering, BIT Campus, Tiruchirappalli, is a record of our own work carried out by us during the academic year 2018-2019 under the guidance of **Mrs. K.AMBIKA,** Assistant Professor, Department of Computer Science and Engineering, University College of Engineering, BIT Campus, Tiruchirappalli. The extent and source of information are derived from the existing literature and have been indicated through the dissertation at the appropriate places. The matter embodied in this work is original and has not been submitted for the award of any other Degree, either in this or any other University.

**SILAMBARASAN K (810015205076)**

**UMAPATHY B (810015205092)**

I certify that the declaration made above by the candidate is true.

Signature of the Project Guide

**Mrs. K.AMBIKA**

Assistant Professor

Department of Computer Science and Engineering

University College of Engineering (BIT Campus)

Tiruchirappalli – 620 024

**ACKNOWLEDGEMENT**

First and foremost of all, we would like to thank our beloved Parents and God Almighty for giving us the strength, knowledge, ability and opportunity to undertake this project study and to persevere and complete it satisfactorily.

We would like to convey our thanks to our honorable Dean **Dr. T.SENTHIL KUMAR**, Associate Professor, for having provided us with all required facilities not only to complete our project without hurdles but also for the entire course of study.

We extremely indebted to our Head of the Department **Dr. D.VENKATESAN**, Assistant Professor, Department of CSE & IT, for the devoted attention shown upon us and making the project a grant success.

We whole heartedly express our gratitude to our Internal Project Guide **Mrs. K. AMBIKA**, Assistant Professor, Department of Computer Science and Engineering for her valuable guidance and help for the successful completion of the project.

We also express our sincere thanks to our Project Coordinator  
 **Mr. M .PRASANNA KUMAR**, Teaching Fellow, Department of Computer Science and Engineering, for her constant inspiration and support.

We extend our thanks to all other Teaching and Non-Teaching Staff members for their constant encouragement and moral support.

And we would like to thank all our well-wishers who have encouraged us to do this project.

**TABLE OF CONTENTS**

**CHAPTER TITLE PAGE**

**NO NO**

**LIST OF FIGURES** III

**LIST OF TABLES** IV

**ABSTRACT** V

1. **INTRODUCTION**

1.1 Data Mining 1

1.2 Machine Learning 4

1.3 Data Mining In Crime 7

1.4 Objective 9

1. **LITERATURE SURVEY** 10
2. **SYSTEM ANALYSIS**

3.1 Existing System 15

3.2 Proposed System 16

**4 SYSTEM SPECIFICATION**

4.1 Software Requirements 17

4.2 Hardware Requirements 17

* 1. About The Software 17

5 **SYSTEM DESIGN**

5.1 System Architecture 22

5.2 Module Description 24

5.3 Experimental Results 33

6 **CONCLUSION AND FUTURE WORK**

6.1 Conclusion 34

6.2 Future Work 34

APPENDIX-I **SAMPLE CODE** 35

APPENDIX-II **SCREENSHOTS** 38

**REFERENCES** 50

**LIST OF FIGURES**

**FIGURE NO DESCRIPTION PAGE NO**

Figure 1.1 Process of data mining 2

Figure 1.2 Types of machine learning 5

Figure 5.1 System Architecture 22

Figure 5.2 Data flow diagram 23

**LIST OF TABLES**

**TABLE NO DESCRIPTION PAGE NO**

Table 5.1 Experimental results of 33

classification Algorithms.

Table 5.2 Experimental results of 33

regression Algorithms.

**ABSTRACT**

Huge collections of crime data are available at crime investigation departments. These data become a valuable resource for the investigators. They help them in predicting crime incidents. However, such data areoften unorganized leading to information navigation and knowledge acquisition. The current system cannot handle large data and it has high time complexity. Decision tree suffers with over fitting and more modern approaches are developed now. In proposed system, classification and regression which comes under supervised learning are used. Machine learning is a subset of artificial intelligence which provides machines the ability to learn automatically and improve from experience without being explicitly programmed. The dataset contain aggregated counts of crime and crime related events categorized by the police department. Here an analysis made on variety of classification methods such as random forest, support vector machine, naive bayes to determine which is best for predicting crime. The parameters used for comparing are accuracy, precision, f-measure, recall. The vulnerable victims for crime incidents are predicted based on results of various algorithms. The experimental results of various algorithms shown that support vector machine has better accuracy and performs well in comparisons to various other algorithms.

**CHAPTER 1**

**INTRODUCTION**

**1.1 DATA MINING**

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), a field at the intersection of computer science and statistics, is the process that attempts to discover patterns in large data sets. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and database systems The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use Aside from the raw analysis step, it involves database and data management aspects, data preprocessing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

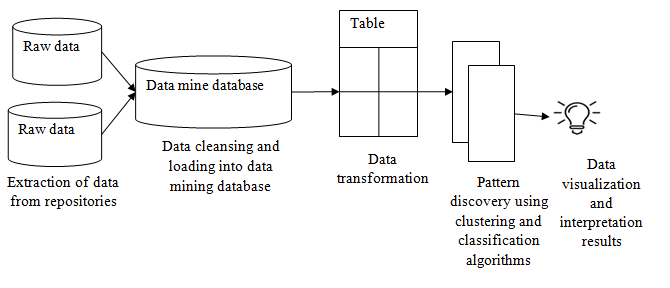


Fig 1.1: Process of data mining

**Data**

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

* Operational or transactional data such as, sales, cost, inventory, payroll, and accounting
* Non operational data, such as industry sales, forecast data, and macro economic data

**Meta data**

data about the data itself, such as logical database design or data dictionary definitions .

**Information**

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling.

**Knowledge**

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

**Data Warehouses**

In computing, a data warehouse (DW or DWH) is a database used for reporting and data analysis. It is a central repository of data which is created by integrating data from multiple disparate sources. Data warehouses store current as well as historical data and are commonly used for creating trending reports for senior management reporting such as annual and quarterly comparisons. The data stored in the warehouse are uploaded from the operational systems (such as marketing, sales etc., shown in the figure to the right). The data may pass through an operational data store for additional operations before they are used in the DW for reporting. The typical ETL-based data warehouse uses staging, integration, and access layers to house its key functions.

Data mining is also known as Knowledge Discovery in Data (KDD). The Data Mining Process: Technological Infrastructure Required:

1. **Database Size**: For creating a more powerful system more data is required to processed and maintained.
2. **Query complexity**: For querying or processing more complex queries and the greater the number of queries, the more powerful system is required.

* Data mining techniques are useful in many research projects, including mathematics, cybernetics, genetics and marketing.
* With data mining, a retailer could manage and use point-of-sale records of customer purchases to send targeted promotions based on an individual’s purchase history. The retailer could also develop products and promotions to appeal to specific customer segments based on mining demographic data from comment or warranty cards.

**1.2 Machine learning**

Machine learning is a data analytics technique that teaches computers to do what comes naturally to humans and animals: learn from experience. Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model. The algorithms adaptively improve their performance as the number of samples available for learning increases. With the rise in big data, machine learning has become a key technique for solving problems in areas, such as:

* Computational finance, for credit scoring and algorithmic trading.
* Energy production, for price and load forecasting.
* Automotive, aerospace, and manufacturing, for predictive maintenance.
* Natural language processing, for voice recognition applications.

The above Figure 1.1 describes the types of Machine Learning Techniques. Unsupervised Learning uses Clustering whereas Supervised Learning technique uses Classification and Regression

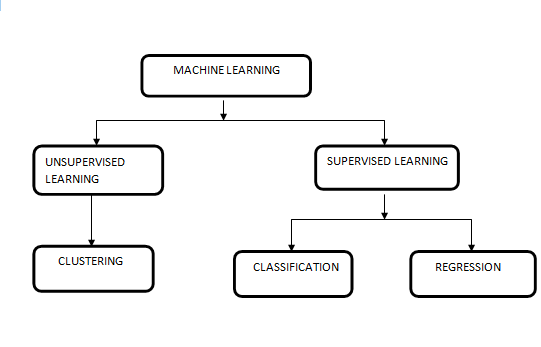


Figure 1.2 Types of Machine Learning

**1.2.1 Supervised Learning**

This algorithm consist of a target variable which is to be predicted from a given set of predictors (independent variables). Using these set of variables, generate a function that map inputs to desired outputs. The training process continues until the model achieves a desired level of accuracy on the training data.

**1.2.1.1 CLASSIFICATION**

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedure is recognized method for repeatedly making such decisions in new situations. Classification techniques predict discrete responses—for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories. Typical applications include medical imaging, speech recognition, and credit scoring. Common algorithms for performing classification include support vector machine (SVM), boosted and bagged decision trees, k-nearest neighbor, Naïve Bayes, discriminant analysis, logistic regression, and neural networks.

**Decision Tree**

It is a type of supervised learning algorithm that is mostly used for classification problems. Surprisingly, it works for both categorical and continuous dependent variables. In this algorithm, here split the population into two or more homogeneous sets. This is done based on most significant attributes/ independent variables to make as distinct groups as possible.

**Support vector Machine**

Support Vector Machine (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However,  it is mostly used in classification problems. In this algorithm, plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.

It is a classification method. In this algorithm, plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

**Bayesian Classification**

Naive Bayes is a simple, yet effective and commonly-used, machine learning classifier. It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network. Naive Bayes classifiers have been especially popular for text classification, and are a traditional solution for problems such as spam detection.

**1.2.1.2 Regression**

Regression techniques predict continuous responses—for example, changes in temperature or fluctuations in power demand. Typical applications include electricity load forecasting and algorithmic trading. Common regression algorithms includes linear model, nonlinear model, regularization, stepwise regression.

**1.2.2 Unsupervised learning**

In this algorithm,  do not have any target or outcome variable to predict / estimate.  It is used for clustering population in different groups, which is widely used for segmenting customers in different groups for specific intervention. Examples of Unsupervised Learning: Apriori algorithm, K-means.

**1.2.3 Reinforcement Learning**

Using this algorithm, the machine is trained to make specific decisions. It works this way: the machine is exposed to an environment where it trains itself continually using trial and error. This machine learns from past experience and tries to capture the best possible knowledge to make accurate business decisions. Example of Reinforcement Learning: Markov Decision Process.

**1.3 Data Mining in crime prediction**

An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. However, in the present scenario, the following major challenges are encountered. Increase in the size of crime information that has to be stored and analysed. Problem of identifying techniques that can accurately and efficiently analyse these growing volumes of crime data Different methods and structures used for recording crime data. The data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult. Investigation of the crime takes longer duration due to complexity of issues. Finding the patterns and trends in crime is a challenging .

To identify a pattern, crime analysts takes a lot of time, scanning through data to find whether a particular crime fits into a known pattern. If it does not fit into an existing pattern then the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, anticipate and prevent crime. The classification algorithm works in a supervised learning manner in which the training and testing phase is required in order to train the classifier to identify the new unknown crime record. This is known as prediction.

The data was taken from the SanFransico crime data set from the kaggle website. Initially do data pre-processing such as data cleaning, filling missing values by taking mean of all the values and followed by applying classification and regression algorithms and comparing the accuracy of the various algorithms and pick up the algorithm that has most accuracy. The parameters picked up are F-Measure, accuracy, precision, recall and found the attribute significance. The present research work proposes the use of data mining techniques that are linked with a common aim of developing such a crime analysis tool. For this purpose, the following specific objectives were formulated.

**1.4 Objective**

To develop a data cleaning algorithm that cleans the crime dataset, by removing unwanted data. Use techniques to fill missing values in an efficient manner. To explore and enhance clustering algorithms to identify crime patterns from historical data o To explore and enhance classification algorithms to predict future crime behaviour. To contrast various data mining algorithm and predict the outcome whether crime occurs on a particular day. Also predict who are vulnerable for the crime. Compare the accuracy of various algorithms and predict the best algorithm for our dataset.

**CHAPTER 2**

**LITERATURE SURVEY**

In this chapter focus is on the literature survey. The literature survey gives as the clear idea about the existing system.

**Malathi. A** et al [1] imparted “Enhanced Algorithm to Predict a Future Crime using Data Mining” .In this paper they present analysis of Crime data mining using MV Algorithm, and Apriori Algorithm. Two clustering techniques, Kmeans and DBScan (Density-Based Spatial Clustering Application with Noise) algorithm are considered for this purpose. The basic method involves cluster the states having the same crime trend and then using ”next year” cluster information to classify records. This is combined with the state poverty data to create a classifier that will predict future crime trends. To the clustered results, a classification algorithm was applied to predict the future crime pattern.

**Chung-Hsien A** et al [2] proposed “Crime Forecasting Using Data Mining Techniques” In this paper they discuss the preliminary results of a crime forecasting model developed in collaboration with the police department of a United States city in the Northeast. First discuss our approach to architecting datasets from original crime records. The datasets contain aggregated counts of crime and crime related events categorized by the police department. The location and time of these events is embedded in the data. Additional spatial and temporal features are harvested from the raw data set. Second, an ensemble of data mining classification techniques is employed to perform the crime forecasting. The result of their research is a model that takes advantage of implicit and explicit spatial and temporal data to make reliable crime predictions.

**D.E. Brown** et al [3] intimated “Crime Pattern Analysis, Prediction Using Data Mining ” imparted a software framework called ReCAP(Regional Crime Analysis Program) for mining data in order to catch professional criminals using data mining and data fusion techniques. To detect suspects, the system will process these inputs through four integrated components: geographic profiling, social network analysis, crime profile, and physical matching. Essentially, geographic profiling determines where" the suspects are, while other components determine the suspects. De Bruin introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly came up with a project which helps crime against women. They followed a efficient technique to fill up the missing values. Various data has been missed. And such data has been replaced by using Random forest algorithm. Random forest algorithm basically handles the missing values well.

**Mugdha Sharma** et al [4] submitted their work on “A data mining tool for the detection of suspicious criminal activities based on decision tree ”. For detecting the suspicious criminal activities, he has concentrated on the importance of data mining technology and designed a proactive application for that purpose. The prediction can be made based on the Textual information or the Geospatial information or even the prisoner records which were manually recorded. By using the real open data such as internet, social feeds and messages the researcher can use the text processing or NLP techniques .

**Donald E. Brown** et al [5] imparted “Data Mining Time Series with Applications to Crime Analysis “All data sets were split into training and test sets. The training sets were used to build the model and the test sets were used to evaluate the results. For the precincts, the last year was used as test data. For the grid cells, only the last three months were used. A final mean squared error was obtained by averaging the results of all grid cells over the entire city. Each precinct was evaluated individually by looking at the mean squared error of the test set. Several dynamic regression models were analyzed to determine if an improvement could be made over the baseline models previously described. For precinct data, the best equation for each precinct was used. For each model for grid cell data, three separate scenarios were analyzed. In the first scenario, the best equation for each grid cell was developed. The second scenario found one equation for all grid cells, but did not use information about B&Es that occurred in surrounding neighborhoods during the previous time period. The fiial scenario used one equation for all grid cells, but this time the information from surrounding neighborhoods was used.

**Suresh Y** [6] intimated “Quality of a software product being designed, has a critical role in software process management”. Detection and prediction of faults in software with huge lines of code is a very tedious task. So it is very essential, as to reduce the maintenance cost and in turn increase the software reliability. Many object-oriented metrics have found to be suitable for software fault prediction. Using data mining techniques, design of prediction and classification models can be incorporated to give insight of the systems quality to the developing team to effectively tackle the quality problems. Chidamber and Kemerer Metric suite, along with classifiers which have immense classification capacity have been used in predicting software fault classification accuracy. Finally, they concluded that Logistic classifier is able to obtain better fault classification accuracy when compared to Naive Bayes approach.

**Ashkan Sadeghi-Mobarakeh** et al [7] submitted their work on forecasting model using random forest algorithm. In classification view, a random forest classifier was applied on the training set and then the performance of the model was evaluated on the test dataset using MAE and MAPE. In classification perspective, used two different approaches. In the first one, trained an RF regressor and predict the price values on the test set. Then, compared values with the class labels using pre-specified threshold. In the second one, converted all price features to pre-specified class labels and created an RF classifier to obtain the class labels directly. The performance of both approaches was evaluated by MPCE. Random forest works good against over fitting and it following bagging principle. The parameters are F-Measure, accuracy, precision, recall and found the attribute significance. The present research work proposes the use of data mining techniques that are linked with a common aim of developing such a crime analysis tool.

**Khushabu A** et al [8] proposed a model for crime analysis using K-Means Clustering algorithm. . In this paper k means clustering technique of data mining used to extract useful information from the high volume crime dataset and to interpret the data which assist police in identify and analyse crime patterns to reduce further occurrences of similar incidence and provide information to reduce the crime. In this paper k mean clustering is implemented using open source data mining tool which are analytical tools used for analysing data .Among the available open source data mining suite such as R, Tanagra ,WEKA ,KNIME ,ORANGE ,Rapid miner Naïve Bayes to predict possible suspects from the criminal records. His approach includes developing a multi-agent for crime pattern identification. There are agents for the place, time, role trademark and substance of criminals which separates the role of the criminals in components.. Classifying the criminals/ suspects is based on the Naïve Bayes classifier for identifying most possible suspects from crime data. Clustering the criminals is based on the model to help to identify patterns of committing crimes.

**Paul S** et al [9] represented “A simultaneous feature selection and weighting–an evolutionary multi-objective optimization approach”. They introduced a new feature selection and weighting method aided with the decomposition based evolutionary multi-objective algorithm called

MOEA/D. The feature vectors are selected and weighted or scaled simultaneously to project the data points to such a hyper space, where the distance between data points of non-identical classes is increased, thus, making them easier to classify. The inter-class and intra-class distances are simultaneously optimized by using MOEA/D to obtain the optimal features and the scaling factor associated with them. Finally, k-NN (k-Nearest Neighbor) is used to classify the data points having the reduced and weighted feature set. The proposed algorithm was tested with several practical datasets from the well-known data repositories like UCI and LIBSVM. The results are compared with those obtained with the state-of-the-art algorithms to demonstrate the superiority of the proposed algorithm.

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 EXISTING SYSTEM**

For detecting the suspicious criminal activities, concentrated on the importance of data mining technology and designed a proactive application for that purpose. In this system, applies an enhanced Decision Tree Algorithm to detect the suspicious people about the criminal activities. An improved ID3 Algorithm with an enhanced feature selection method and attribute-importance factor is applied to produce a better and faster Decision Tree based on the information entropy which is explicitly derived from a series of training data sets from several classes. The system provides us informative data for classification by choosing effective feature selection strategy recognizing sentiment sentence. It is a combination of Advanced ID3 classification algorithm and enhanced feature selection method for the better efficiency of the algorithm.

**Limitation:**

* The model suffers with overfitting.
* The model consumes lot of time and accuracy compared to modern methods is very low.
* The model sometimes become inaccurate and a small change in the data lead to the large change in the structure of the tree.

**3.2 PROPOSED SYSTEM**

Crime rate is one of the biggest concern now as days. Predicting whether crime can occur on a particular day can be useful and deploying all police resource in that area. And also predicting who are the victim of the crime is a major problem. Identification of victims of the crime helps in of classifying people is an easy task. For this we are using decision tree, random forest, linear regression, Support vector machine, Naïve Bayes algorithms. Applying those algorithms and checking the experimental results of those algorithms. And finding the similarities of those algorithms and also contrast the accuracy of various algorithms and predicting the best one for our model. Data set has been taken from www.kaggle.com/datasets. Initially classification algorithms and regression algorithms are applied to cleaned data set and later with raw datasets. The data has split into training and testing data using training data and train the model and check the performance of the model using the testing data. The model will predict whether a crime will occur on a particular day and who are major victims of the crime scene. Finally contrast these results as well and finally predict most vulnerable victims among the given dataset.

**CHAPTER 4**

**SYSTEM SPECIFICATION**

**4.1 SOFTWARE REQUIREMENTS**

Operating System : Windows Pro

Front End : Juypter Notebook

Back End : SQL

Language : Python

**4.2 HARDWARE REQUIREMENTS**

Processor : Intel Pentium

Hard Disk : 40 GB

RAM : 4 GB

**4.3 ABOUT THE SOFTWARE**

**Python**

Python is a widely used general-purpose, high level programming language. It was initially designed by Guido van Rossum in 1991 and developed by Python Software Foundation. It was mainly developed for emphasis on code readability, and its syntax allows programmers to express concepts in fewer lines of code. Python is a programming language that lets you work quickly and integrate systems more efficiently. There are two major Python versions- Python 2 and Python 3. Both are quite different. Some of the advantages of using python:

* Emphasis on code readability, shorter codes, ease of writing
* Programmers can express logical concepts in fewer lines of code in comparison to languages such as C++ or Java.
* Python supports multiple programming paradigms, like object-oriented, imperative and functional programming or procedural.

There exist inbuilt functions for almost all of the frequently used concepts.

**Features**

* Interpreted
* Platform Independent
* Free and open source, Redistributable
* Embeddable
* Robust
* Rich Library support

**Data Analysis in Python**

Python is a great language for doing data analysis, primarily because of the fantastic ecosystem of data-centric Python packages. Pandas is one of those packages, and makes importing and analyzing data much easier.

**NLTK**

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to over 50 corpora and lexical resources. Such as WordNet, along with a suite of text processing libraries for classification, tokenization,stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active discussion forum.

NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

**Anaconda:**

Anaconda is a free and open source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment. Package versions are managed by the package management system conda. The Anaconda distribution is used by over 12 million users and includes more than 1400 popular data-science packages suitable for Windows, Linux, and MacOS.

**Anaconda Navigator**

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS and Linux. Anaconda® is a package manager, an environment manager, a Python/R data science distribution, and a collection of over 1500+ open source packages. Anaconda is free and easy to install, and it offers free community support. Navigator is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window.

**Anaconda Cloud:**

Anaconda Cloud is a package management service by Anaconda where you can find, access, store and share public and private notebooks, environments, and conda and PyPI packages. Cloud hosts useful Python packages, notebooks and environments for a wide variety of applications. You do not need to log in or to have a Cloud account, to search for public packages, download and install them.

You can build new packages using the Anaconda Client command line interface (CLI), and then manually or automatically upload the packages to Cloud.

**Conda**

Conda is an open source package management system and environment management system that runs on Windows, macOS and Linux. Conda quickly installs runs and updates packages and their dependencies. Conda easily creates, saves, loads and switches between environments on your local computer. It was created for Python programs, but it can package and distribute software for any language.

Conda can be combined with continuous integration systems such as Travis CI and AppVeyor to provide frequent, automated testing of your code. The conda package and environment manager is included in all versions of Anaconda and Miniconda, Anaconda Repository. Conda is also included in Anaconda Enterprise , which provides on-site enterprise package and environment management for Python, R, Node.j.

Applications Available in Navigator

* JupyterLab
* Jupyter Notebook
* QTConsole
* Spyder
* VSCode
* GlueViz
* Orange 3 App
* Rodeo

**Jupyter Notebook**

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

**CHAPTER 5**

**SYSTEM DESIGN**

**5.1 SYSTEM ARCHITECTURE**

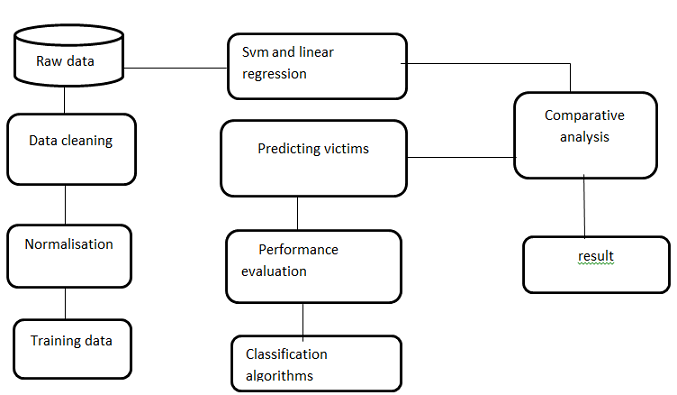


Figure 5.1: Architecture of the proposed system

The above Figure 5.1 shows the Architectural diagram of the proposed system.

**DATA FLOW DAIGRAMS**

Crime dataset

result

Figure 5.2 Level 0 DFD

Crime dataset

result

**5.2 MODULE DESCRIPTION**

The Proposed System consists of the following modules:

* Dataset Collection
* Data Preprocessing
* Applying Decision tree
* Applying Guassian NB
* Applying Svm
* Linear Regression

**5.2.1 Dataset Collection**

. The data is collected from kaggle dataset. It contains sanfransico crime data set. The dataset contains 128 attributes with 2049 records. These data are real crime details. Initially data are in They are pre-processed and converted into csv format. Later the csv format is loaded into Jupter notebook.

**5.2.2 Data Preprocessing**

The following Figure 5.2 shows the steps of preprocessing. This preprocessing involves transforming raw data into an understandable format. Real world data often contains noise, incomplete, inconsistent data, lacking in certain patterns. These issues are resolved by Data Preprocessing.

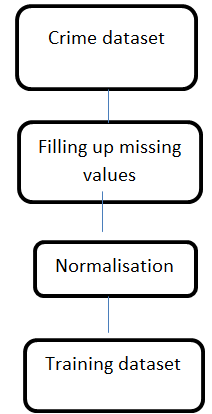


Figure 5.2: Preprocessing steps

Preprocessing steps involves Data Cleaning, Data Integration and Data Transformation.

**5.2.3 Applying Decision tree**

Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

**Gini Index**

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure.

Steps to Calculate Gini for a split:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split.

**Information Gain:**

Information theory is a measure to define this degree of disorganization in a system known as Entropy. If the sample is completely homogeneous, then the entropy is zero and if the sample is an equally divided (50% – 50%), it has entropy of one.

Entropy can be calculated using formula:-

Entropy=-plog2p-qlog2q

Here p and q is probability of success and failure respectively in that node. Entropy is also used with categorical target variable. It chooses the split which has lowest entropy compared to parent node and other splits. The lesser the entropy.

Steps to calculate entropy for a split:

1. Calculate entropy of parent node
2. Calculate entropy of each individual node of split and calculate weighted average of all sub-nodes available in split.

**Confusion matrix**

A confusion matrix is a summary of prediction results on a classification problem.  
The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.

• True Positive (TP) : Observation is positive, and is predicted to be positive.  
• False Negative (FN) : Observation is positive, but is predicted negative.  
• True Negative (TN) : Observation is negative, and is predicted to be negative.  
• False Positive (FP) : Observation is negative, but is predicted positive.

**Accuracy:**

Accuracy defines the how efficient the model is

**Accuracy=(TP+TN)/(TP+TN+FN+FP)**

**Recall:**   
Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (small number of FN).

Recall is given by the relation:

**Recall= TP/(TP+FN)**

**Precision**   
To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labeled as positive is indeed positive (small number.  
Precision is given by the relation:

**Precision=TP/(TP+FP)**

**F-measure**   
Having two measures (Precision and Recall) it helps to have a measurement that represents both of them. Calculate an F-measure which uses Harmonic Mean in place of Arithmetic Mean as it punishes the extreme values.   
The F-Measure will always be nearer to the smaller value of Precision or Recall.

**F-measure= 2\* recall\* precision/recall + precision**

The decision tree gives accuracy and other parameters and then compare all accuracy for better results. PctKids2Par, racePctWhite, racePctHisp, HousVacant, blackPerCap are the most significant variables in this model. PctKids2Par- percentage of kids in family housing with two parents (numeric - decimal) racePctWhite- percentage of population that is caucasian (numeric - decimal) racePctHisp- percentage of population that is of hispanic heritage HousVacant- number of vacant households (numeric - decimal) blackPerCap- per capita income for african americans (numeric - decimal)

Out of the above 5 factors, the first (PctKids2Par), makes sense. The higher percentage of kids in family housing could mean higher targets for crime and eventually higher crime rate

**5.2.4 Applying Gaussian NB:**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes’ Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other.

Bayes’ Theorem finds the probability of an event occurring given the probability of another event that has already occurred. Bayes’ theorem is stated mathematically as the following equation:

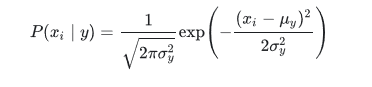
P(A/B)=P(B/A) P(A)/P(B)

where A and B are events and P(B) ? Basically, are trying to find probability of event A, given the event B is true. Event B is also termed as evidence.

P(A) is the priori of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).

P(A|B) is a posteriori probability of B, i.e. probability of event after evidence is seen.

GuassianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian:



The parameters σy and μy are estimated using maximum likelihood.

PctKids2Par, PctFam2Par, racePctWhite, PctIlleg,FemalePctDiv are the most significant variables identified PctKids2Par- percentage of kids in family housing with two parents (numeric - decimal) PctFam2Par - percentage of families (with kids) that are headed by two parents (numeric - decimal) racePctWhite - percentage of population that is caucasian (numeric - decimal) PctIlleg - percentage of kids born to never married (numeric - decimal) FemalePctDiv - percentage of females who are divorced (numeric - decimal)

Out of the above 6 factors, the first (PctKids2Par) and the 2nd (PctFam2Par) make sense and is consistent with what was observed in the earlier decision tree. The higher percentage of kids in family housing with 2 parents could mean higher targets for crime and eventually higher crime rate. And higher the percentage of such households would linearly increase/decrease the targets.

It would be nice to mention the formula, and discuss why it make sense to calculate

**5.2.5 Applying Linear svc**

Similar to SVC with parameter kernel=’linear’, but implemented in terms of liblinear rather than libsvm, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples.

This class supports both dense and sparse input and the multiclass support is handled according to a one-vs-the-rest scheme.

**random\_state :**

The seed of the pseudo random number generator to use when shuffling the data for the dual coordinate When the underlying implementation of linearsvc is not random and has no effect on the results. If int, random\_state is the seed used by the random number generator; If RandomState instance, random\_state is the random number generator.

**Decision function(X)**

Predict confidence scores for samples.

The confidence score for a sample is the signed distance of that sample to the hyperplane

**densify()**

Convert coefficient matrix to dense array format.

Converts the coef member to a numpy.ndarray. This is the default format of coef\_ and is required for fitting, so calling this method is only required on models that have previously been sparsified; otherwise, it is a no-op.

**Score**

Returns the mean accuracy on the given test data and labels.In multi-label classification, this is the subset accuracy which is a harsh metric since you require for each sample that each label set be correctly predicted. MalePctDivorce, racepctblack, PctOccupMgmtProf, agePct12t21, racePctHisp are the most significant variables identified. MalePctDivorce - percentage of population who are divorced (numeric - decimal) racepctblack - percentage of population that is african american (numeric - decimal) PctOccupMgmtProf - percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal) agePct12t21 - percentage of population that is 12-21 in age (numeric - decimal) racePctHisp- percentage of population that is of hispanic heritage (numeric - decimal)

Out of the above features identified, the ones that make sense are PctOccupMgmtProf and agePct12t21. This is because, people in this age range and occupations are likely to be outside their homes during the day, hence they could be potential targets. Difference in the values of these features will make an impact on the predcited variable.

**5.2.6 Applying Linear Regression**

Regression analysis is a very widely used statistical tool to establish a relationship model between two variables. One of these variable is called predictor variable whose value is gathered through experiments. The other variable is called response variable whose value is derived from the predictor variable.

In Linear Regression these two variables are related through an equation, where exponent (power) of both these variables is 1.

Y=ax+b

**fit\_intercept:**

Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations

**normalize:**

his parameter is ignored when fit\_intercept is set to False

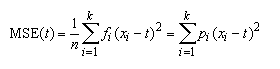
**coef**

Estimated coefficients for the linear regression problem. If multiple targets are passed during the fit (y 2D), this is a 2D array of shape (n\_targets, n\_features), while if only one target is passed, this is a 1D array of length n\_features.

**Mean Square Error**

In a sense, any measure of the center of a distribution should be associated with some measure of error. Say that the number t is a good measure of center, then presumably we are saying that t represents the entire distribution better, in some way, than other numbers.

In this context, suppose that we measure the quality of t, as a measure of the center of the distribution, in terms of the mean square error.



**5.3 Experimental results**

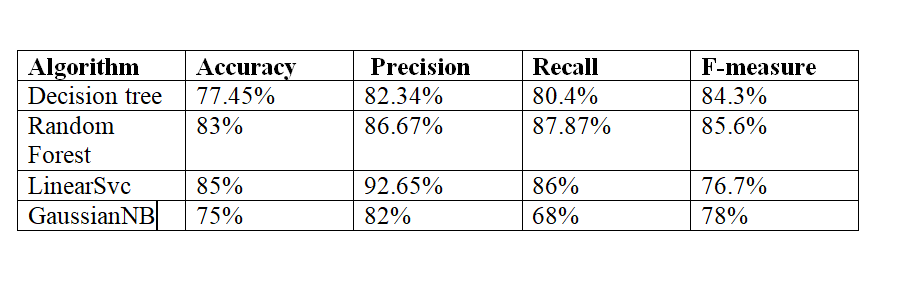


Table 5.1 Experimental results for the classification algorithms

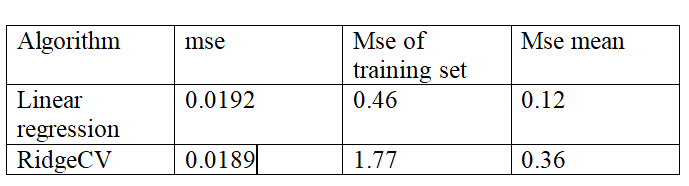


Table 5.2 Experimental results for the regression algorithms

Experimental results confirms that LinearSvc is best among classification algorithms and linear regression is better than the quadratic regression ridgeCV for this dataset.

MSE for Linear regression: 0.0192 MSE for Ridge regression: 0.0189 .The mean squared error can be interpreted as the closeness of the data points with the predcited regression line. This explains that Ridge regression is able to redcue overfitting to some extent compared to the result of linear regression. No parameters between this model and Gaussian Naive Bayes models are similar. Only one paramter, which is racePctHisp, is common between this model and decision tree.

**CHAPTER 6**

**CONCLUSION AND FUTURE WORKS**

**6.1 CONCLUSION**

Crime analysis has become a promising technique for industry and academia and it widely applied in the investigation. The proposed system uses decision tree, LinearSVC, GuassianNB, Linear regression and ridgeCV. Various mentioned algorithm has been applied and their accuracy and various parameters are compared and the best model has been interpreted. The existing systems uses Kmeans algorithm. Whereas SVM uses a hyperplane which separates training data into two classes which results in good model for this datasets.

**6.2 FUTURE WORKS**

To develop an android application which takes current data as input and predict whether crime occur or not and list those who are vulnerable to crime attack. Further the model must be also trained with additional parameters and model need to handle huge amount of dataset and with a lesser time complexity. They should also consider the issues related to classification. And also this scheme will be tested using various classifiers such as Decision tree and Neural Networks.

**APPENDIX-I**

**SAMPLE CODE**

**Reading csv file**

import pandas as pd

import numpy as np

df = pd.read\_csv("communities-crime-clean.csv")

**Considering high crime**

df['highCrime'] = (df['ViolentCrimesPerPop'] > 0.1)

**Dividing data into train and test set**

X = df.drop(['communityname','ViolentCrimesPerPop','highCrime'],axis=1)

y = df['highCrime']

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split( X, y, test\_size=0.2, random\_state=42)

**Building Decision tree classifier**

from sklearn import tree

dtc = tree.DecisionTreeClassifier(criterion = 'entropy')

dtc = dtc.fit(X\_train, y\_train)

y=dtc.predict(X\_test)

**Vulnerable crime Identifying**

dtc.feature\_importances\_

significance = dtc.feature\_importances\_

f=zip(significance, dimensions)

l=sorted(f, reverse=True)

l[:5]

**GaussianNB classifier**

from sklearn.naive\_bayes import GaussianNB

gnb=GaussianNB()

gnb.fit(X\_train, y\_train)

y=gnb.predict(X\_test)

l= X.columns

**Significant attributes**

d = pd.DataFrame({"Dimension": [], "Significance": []})

for i in l:

means = df.groupby('highCrime')[i].mean()

variance = df.groupby('highCrime')[i].var()

d = d.append({"Dimension": i, "Significance": abs(means[0] - means[1])/ sum(np.sqrt(variance))}, ignore\_index=True)

d.iloc[np.argsort(d['Significance'])[::-1][0:5], ]

**LinearSVC model**

from sklearn.svm import LinearSVC

svc =LinearSVC()

svc.fit(X\_train, y\_train)

Finding vulnerable victims

dimensions =list(X\_train)

weights =svc.coef\_

weights=weights[0]

f= zip(weights, dimensions)

l=sorted(f, reverse=True)

l[:5]

**Linear regression model**

X = df.drop(['communityname','ViolentCrimesPerPop','highCrime'],axis=1)

y = df['ViolentCrimesPerPop']

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split( X, y, test\_size=0.2, random\_state=42)

Identifying feature importance

coeffecients = lm.coef\_

s = coeffecients >= 0

abs\_coeffecients=abs(lm.coef\_)

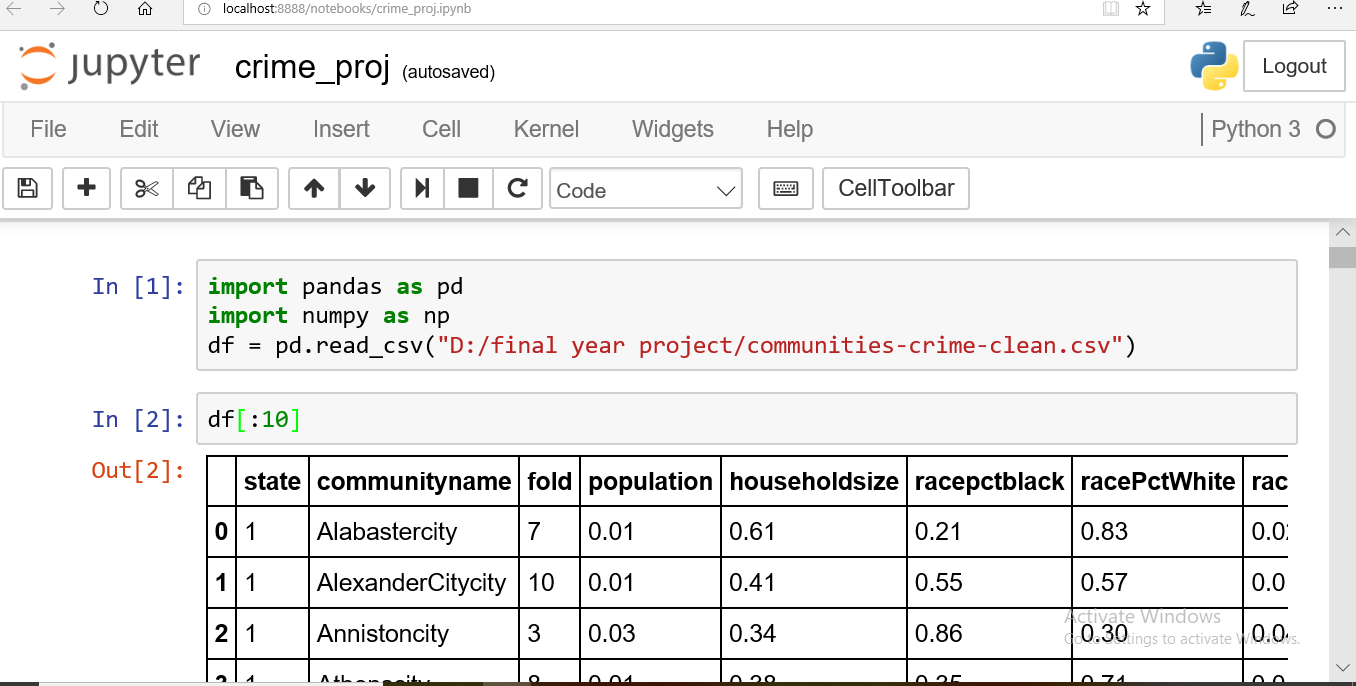
i = np.argsort(abs\_coeffecients)[::-1]

for f in range(10):

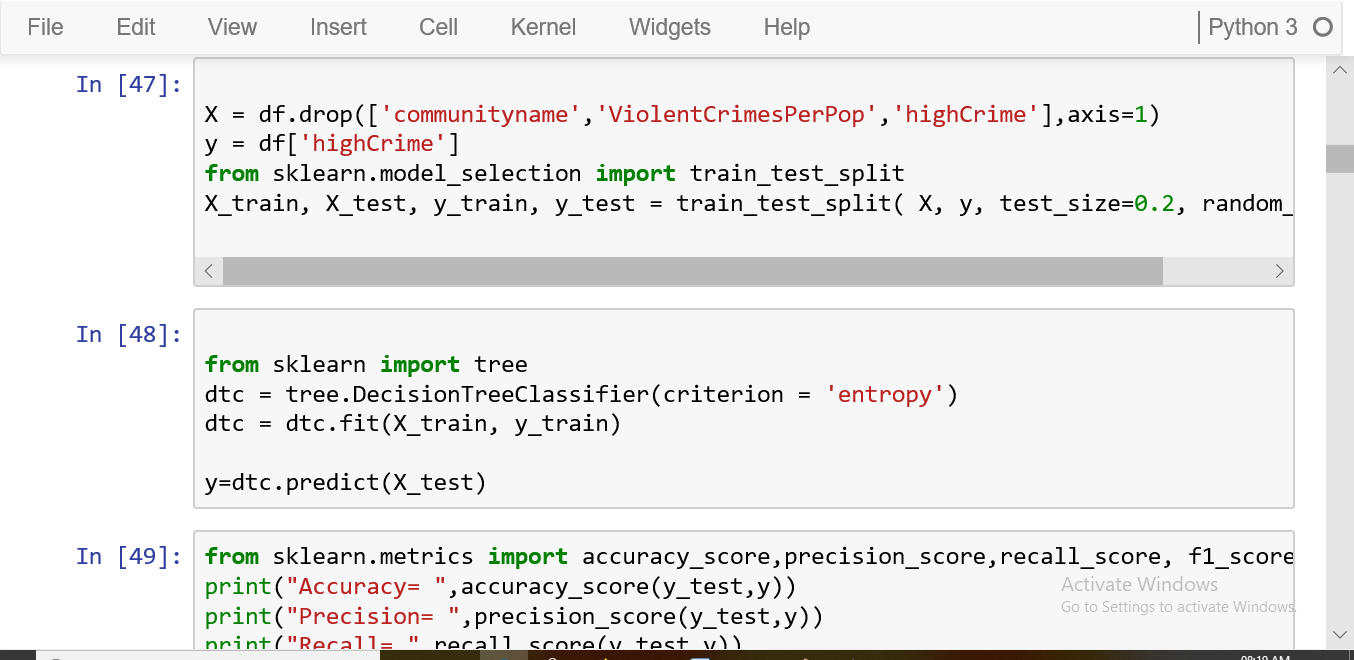
print("%s- %f - %s" % (X\_train.columns[i[f]], abs\_coeffecients[i[f]], s[i[f]]))

**APPENDIX-II**

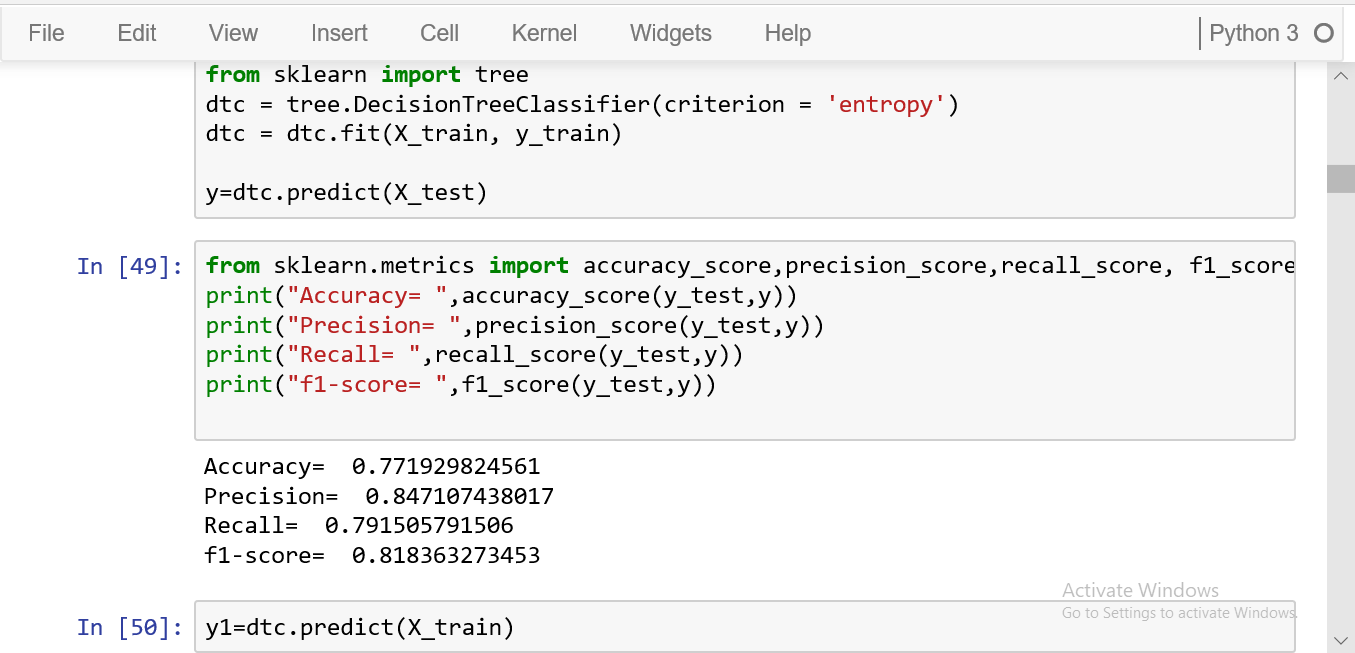
**SCREENSHOTS:**



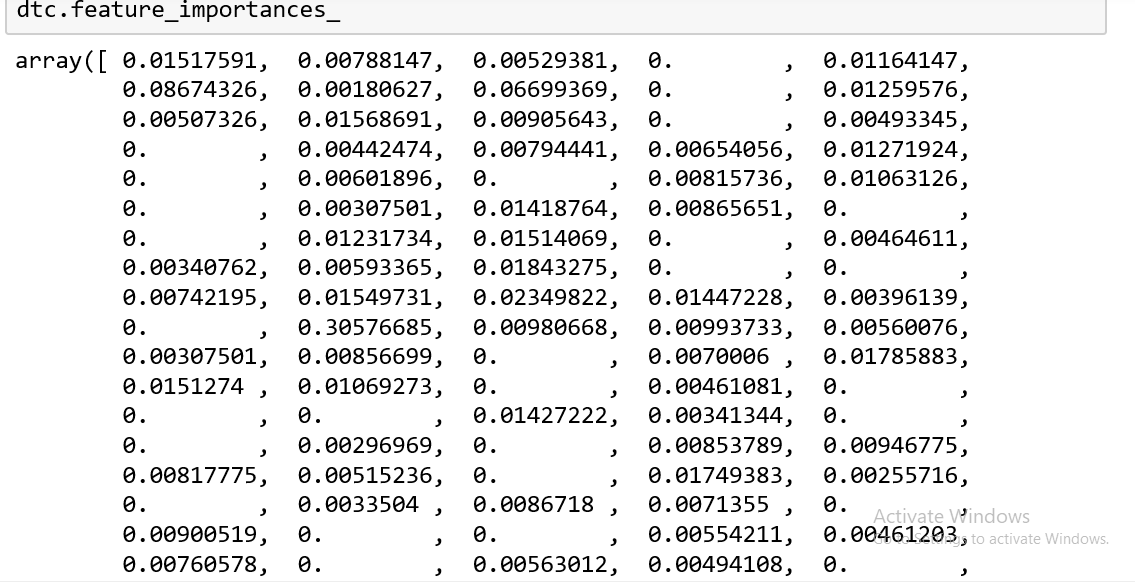
**S1 : Reading Dataset**



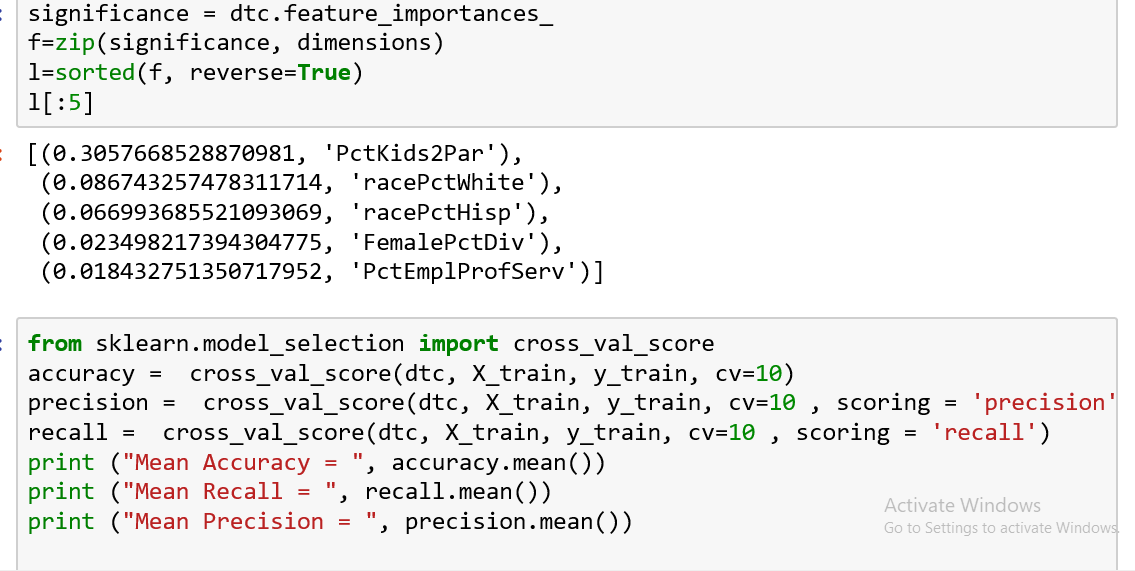
**S2: Decision tree classifier**



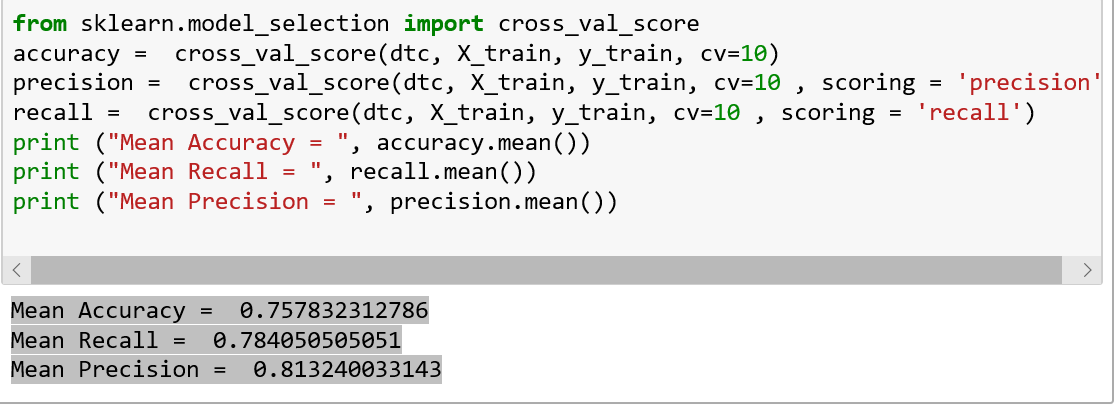
**S3: Decision tree accuracy**

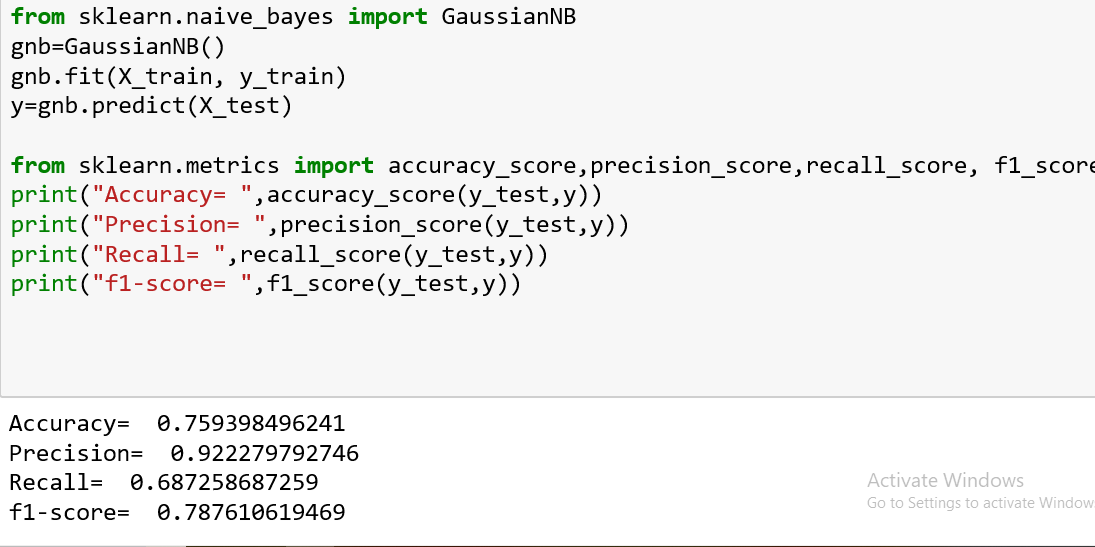


**S4: Feature importance**

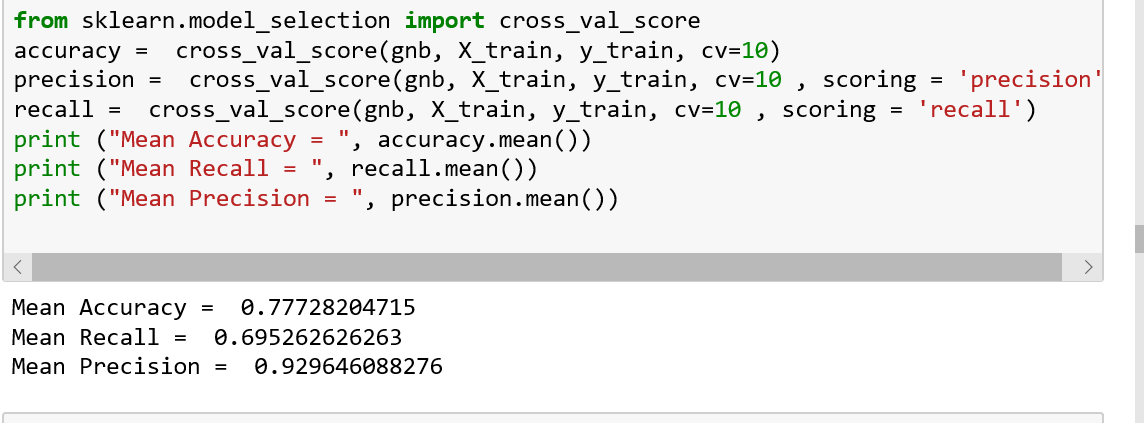


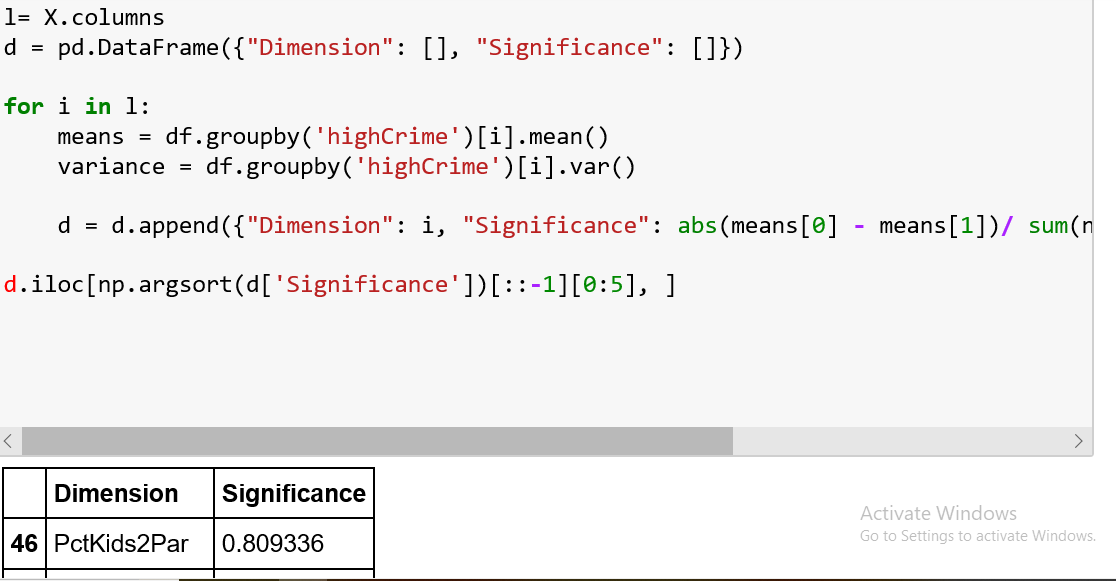
**S5: victims identification**



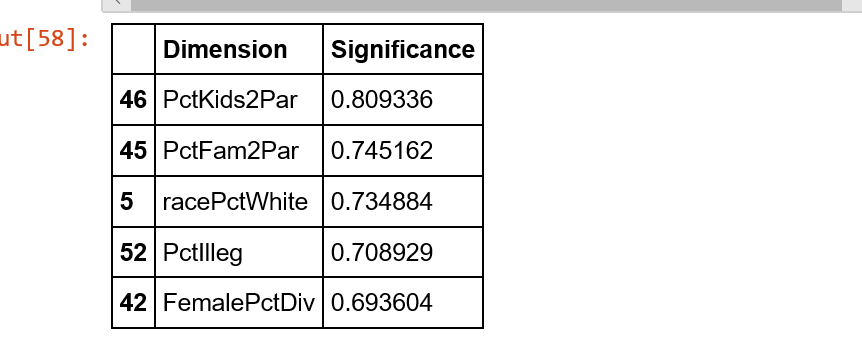


**S6: GuassianNB classifier**

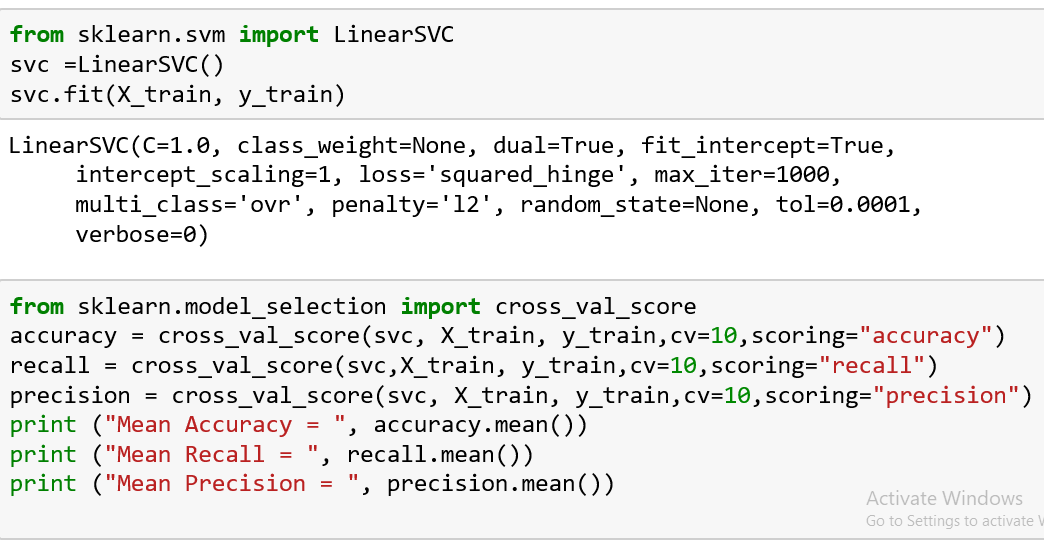




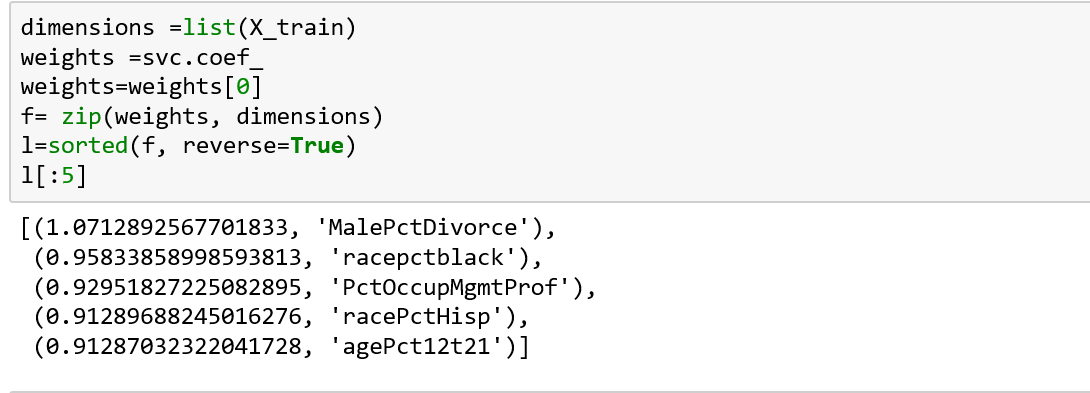
**S7:Guassian cross validation**



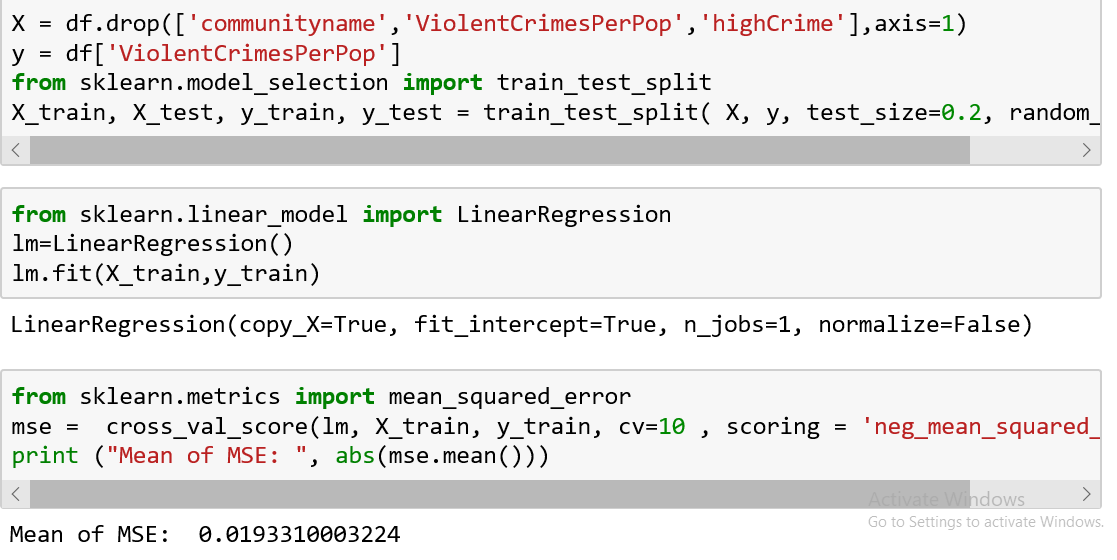
**S8: Significant attributes**



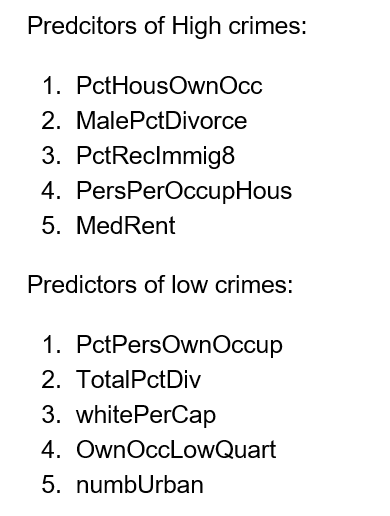
**S9: LinearSVC accuracy**

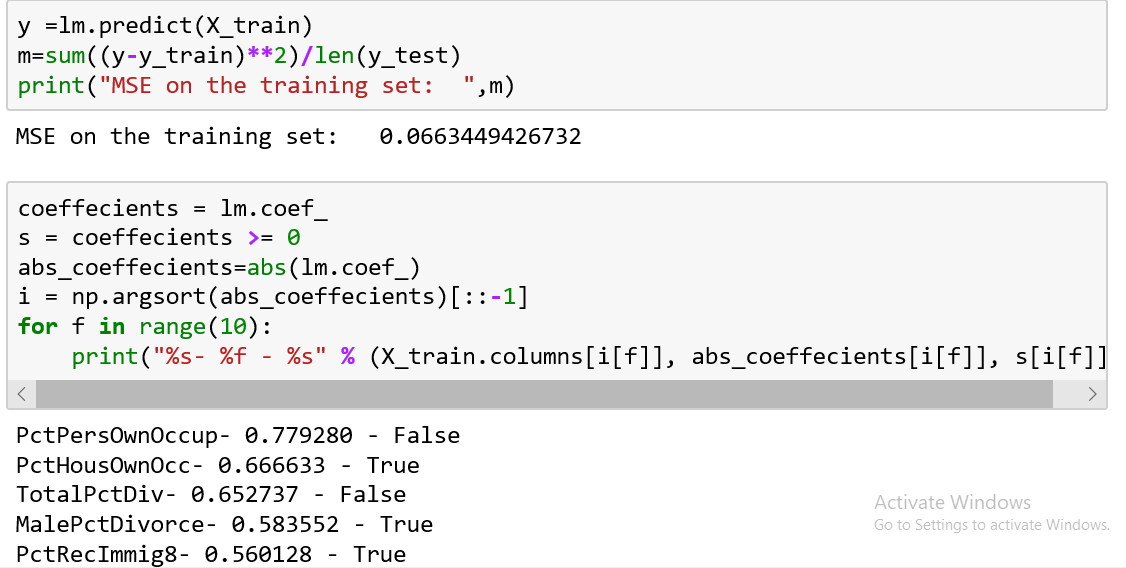


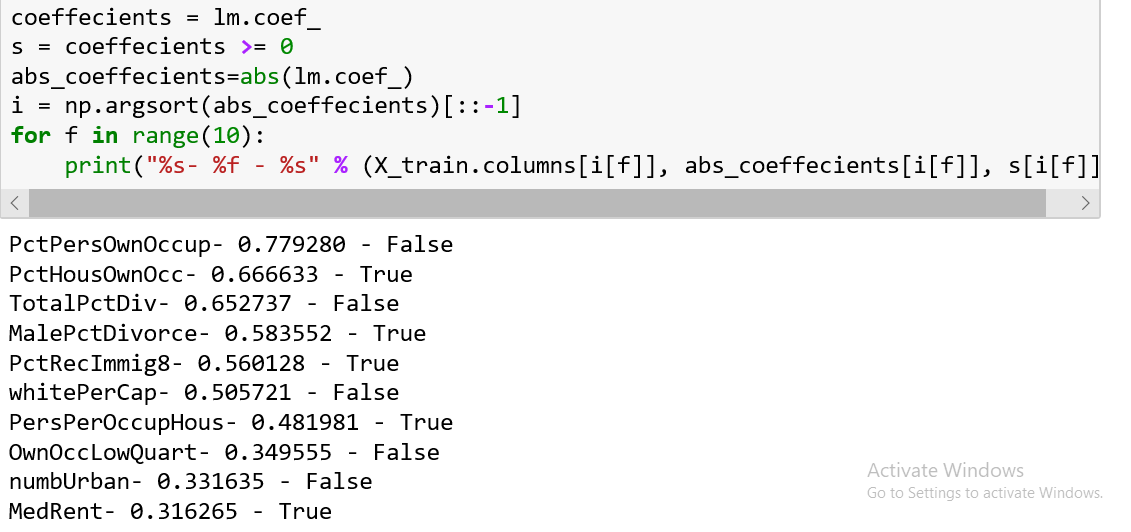
**S10: Vulnerable Crimes**

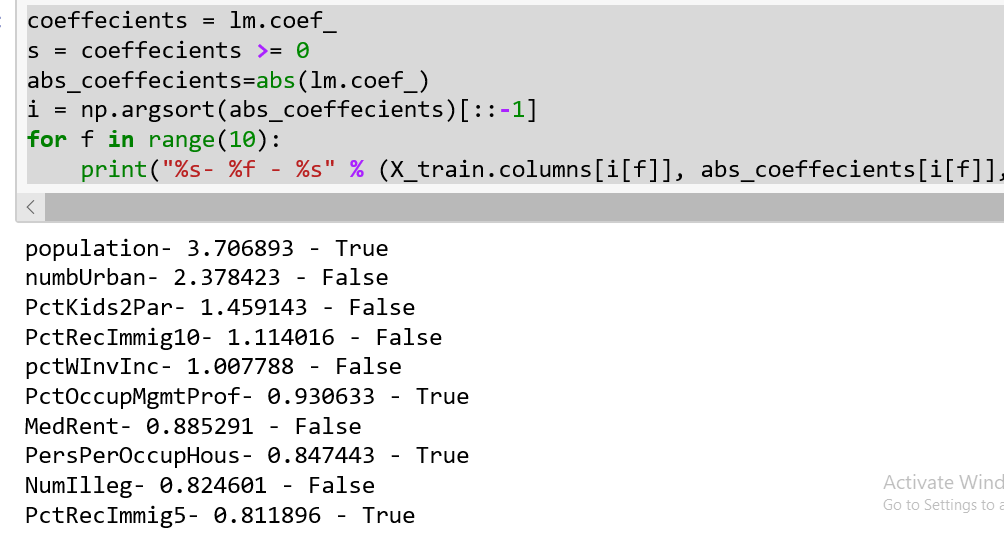


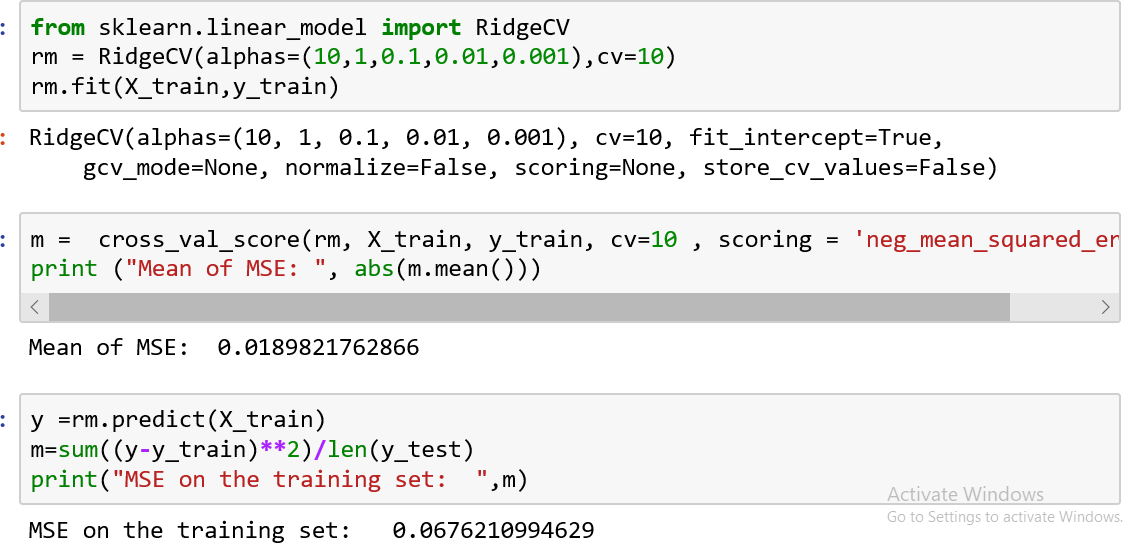
**S11: Linear Regression**



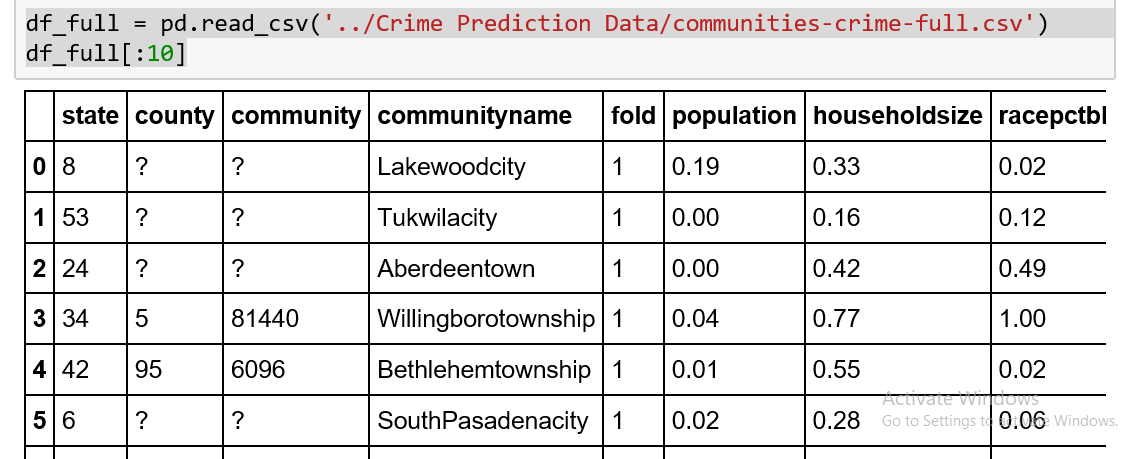
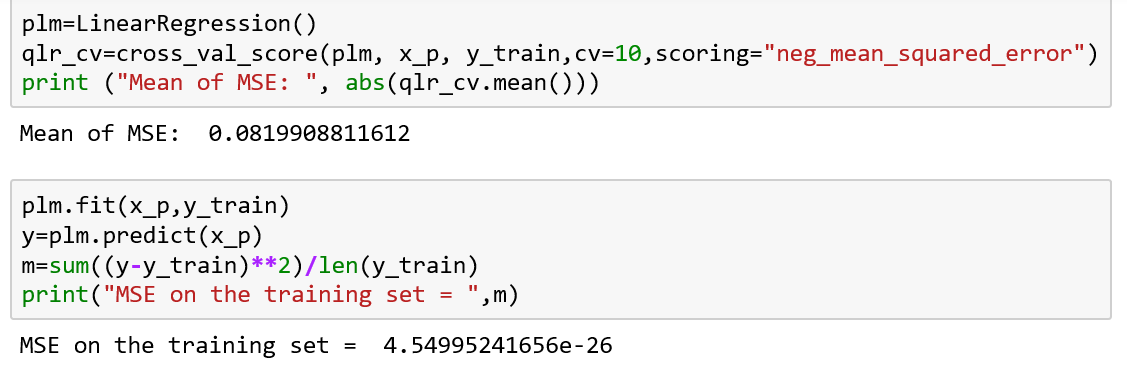


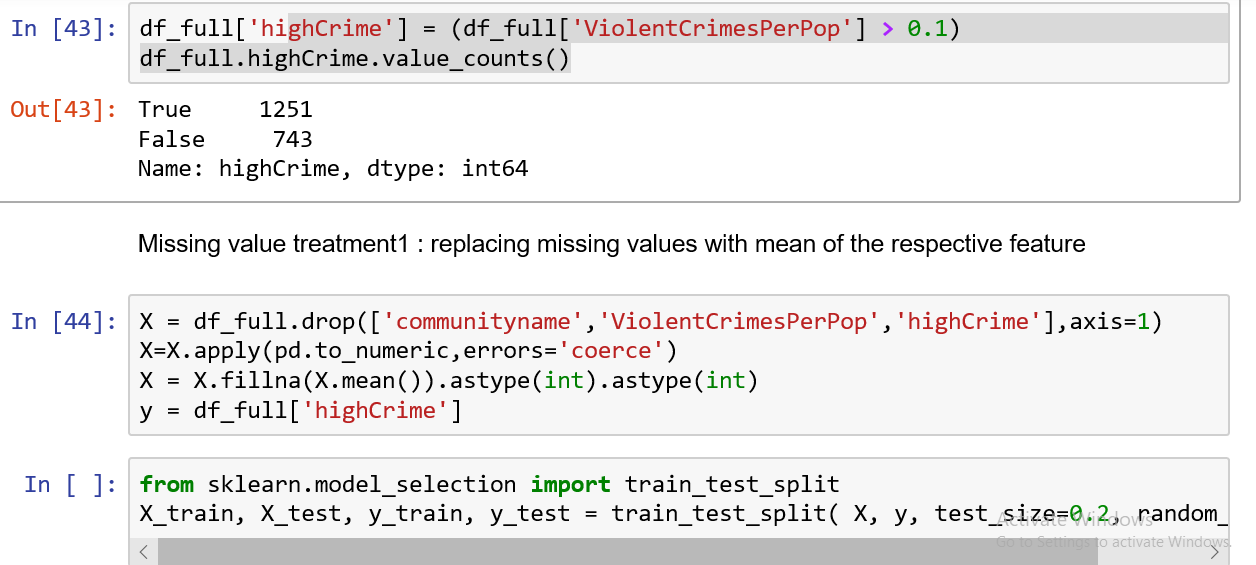


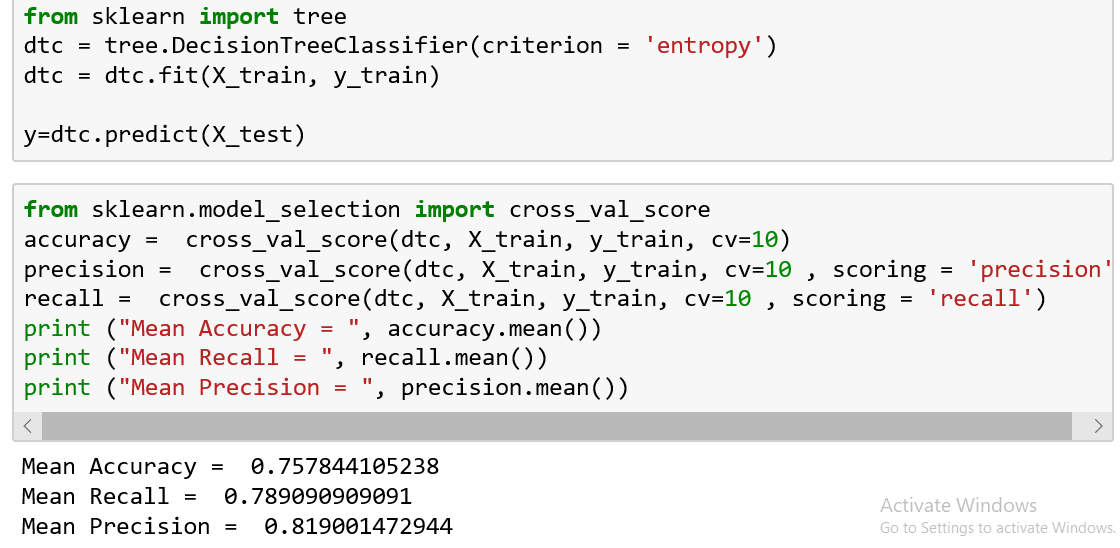




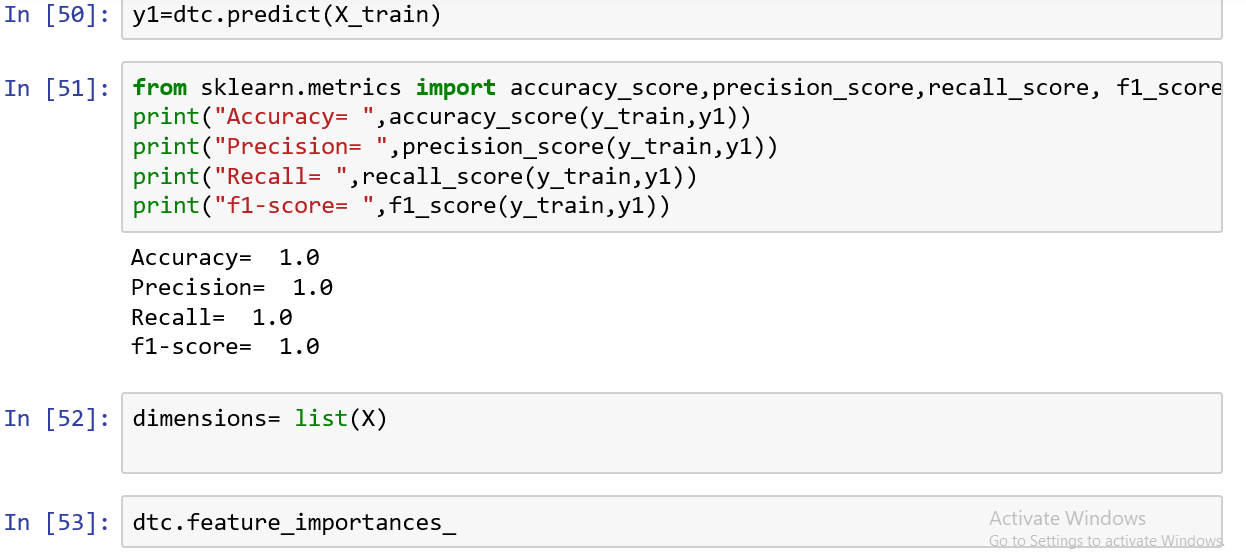


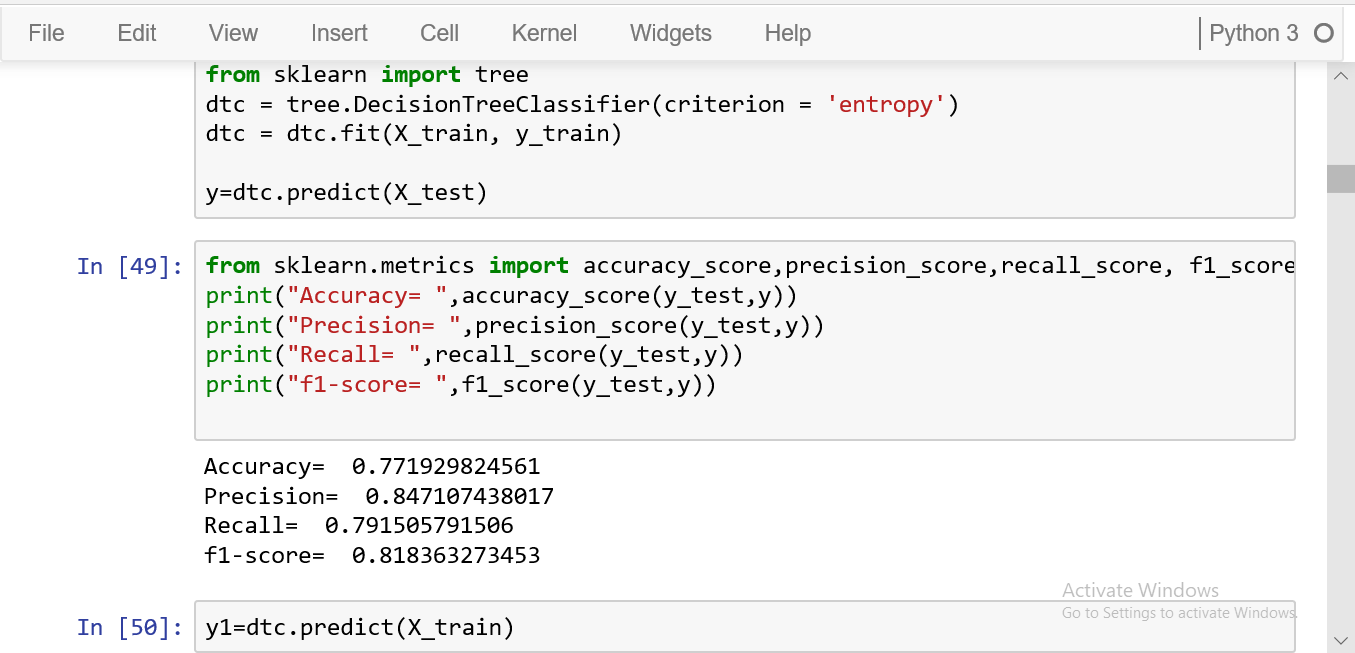






**S13: Dirty data decision tree**





**REFERENCES**

[1] G.Ateniese, R.D.Pietro, L.V.Manchini, and G.Tsudik, “Crime Investigation in a Novel way,”in Proc.16th ACM Conf,Secur.Privacy Commun.Netw. 2008, Art.no.9.

[2] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, and Addission Wesley. Introduction to Data Mining(2006) ISBN: 0-321-321136-7

[3] Wang Yan. Application of time series. Renmin University 0f china Publisher. 2005.

[4] George Kelling and Catherine Coles, Fixing Broken Windows : Restoring order and reducing crime in our society ,ISBN:0-684-83738-2.

[5] De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data mining approaches to criminal carrer analysis ,”in Proceedings of the Sixth International Conference on Data Mining (ICDM”06) ,Pp. 171-177

[6] M. B. Mitchell, D. E. Brown, and J. H. Conklin, “A Crime Forecasting Tool for the Web-Based Crime Analysis Toolkit,” 2007 IEEE Syst. Inf. Eng. Des. Symp.,pp.1– 5,2007.

[7] S. Ismail and N. Ramli,“Short-Term Crime Forecast ting In Keedah,” Procedia-Soc. Behav. Sci., vol.91,pp.654– 660,2013. [8] W. Gorr and R. Harries, “Introduction to crime forecasting,” vol. 19, pp. 551–555, 2003.

[9] K. H. Vellani, “Crime Analysis: for Problem Solving Security Professionals in 25 Small Steps,” Facilities,pp.1– 56,2010.

[10] W. Gorr, A. Olligschlaeger, and Y. Thompson, “Short-term forecasting of crime,” Int. J. Forecast., vol. 19, no. 4, pp. 579–594, 2003.

[11] Shiju Sathyadevan, M.S. Devan and S. Surya Gangadharan, “CrimeAnalysis and Prediction using Data Mining”, Proceedings of IEEE 1st International Conference on Networks and Soft Computing, pp. 406-412, 2014.

[12] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, “Mining Location-based Social Networks for Criminal Activity Prediction”, Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.

[13] Chung-Hsien Yu, Max W. Ward, Melissa Morabito and Wei Ding, “Crime Forecasting using Data Mining Techniques”, Proceedings of 11th IEEE International Conference on Data Mining Workshops, pp. 779-786, 2011.

[14] P. Thongtae and S. Srisuk, “An Analysis of Data Mining Applications in Crime Domain”, Proceedings of IEEE 8th International Conference on Computer and Information Technology Workshops, pp. 122-126, 2008.

[15] Yu-Yueh Huang, Cheng-Te Li and Shyh-Kang Jeng, “Mining Location-based Social Networks for Criminal Activity Prediction”, Proceedings of 24th IEEE International Conference on Wireless and Optical Communication, pp. 185-190, 2015.