

A MINOR PROJECT REPORT ON
(CRIME PREDICTION AND
ANALYSIS USING MACHINE
LEARNING)

SUBMITTED IN PARTIAL FULFILLMENT FOR THE AWARD OF
DEGREE OF
BACHELOR OF TECHNOLOGY IN
ELECTRONICS AND COMMUNICATION
ENGINEERING



Submitted By:

ESHA MAHENDRA (9916102024)
HARSHITA MADAN (9916102069)
SHIVANGI GUPTA (9916102126)

Under the Guidance Of:

DR. BAJRANG BANSAL

DEPARTMENT OF ELECTRONICS AND COMMUNICATION
ENGINEERING, JAYPEE INSTITUTE OF INFORMATION
TECHNOLOGY, NOIDA (U.P.) May, 2019

CERTIFICATE

This is to certify that the minor project report entitled, “CRIME PREDICTION AND ANALYSIS USING MACHINE LEARNING” submitted by Esha Mahendra, Shivangi Gupta, Harshita Madan in partial fulfillment of the requirements for the award of Bachelor of Technology Degree in **Electronics and Communication Engineering** of the Jaypee Institute of Information Technology, Noida is an authentic work carried out by them under my supervision and guidance. The matter embodied in this report is original and has not been submitted for the award of any other degree.

Signature of Supervisor:

Name of the Supervisor: Dr. Bajrang Bansal

ECE Department, JIIT, Sec-128, Noida-201304

Dated:

DECLARATION

We hereby declare that this written submission represents our own ideas in our own words and where others' ideas or words have been included, have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission.

Place: Noida

Date:

Name: Esha Mahendra

Enrollment: 9916102024

Name: Harshita Madan

Enrollment: 9916102069

Name: Shivangi Gupta

Enrollment: 9916102126

ABSTRACT

To be better prepared to respond to criminal activity, it is important to understand patterns in crime. The main aim of the project is to analyse the dataset which consist of numerous crimes happened in India(state-wise) and predicting the type of crime which may happen in future depending upon various conditions(like time, location, crime-type, arrested ratio etc.) using Machine Learning Algorithms. The dataset is extracted from the official website. The objective would be to train a model for prediction. Building the model will be done using better algorithms depending upon the accuracy. The use of ML in predicting crimes or an individual's likelihood for committing a crime has promise but is still more of an unknown. The biggest challenge will probably be “proving” that it works. When a system is designed to stop something from happening, it is difficult to prove the negative. Improvements in crime prevention technology will likely spur increased total spending on this technology.

ACKNOWLEDGEMENT

It is our privilege to express our sincerest regards to our project coordinator, Dr.Bajrang Bansal for his valuable inputs, guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of our project.

We deeply express our sincere thanks to our Head of Department for encouraging and allowing us to present the project on the topic “CRIME PREDICTION AND ANALYSIS USING MACHINE LEARNING” at our department premises for the partial fulfillment of the requirements leading to the award of B-Tech degree. We take this opportunity to thank all our lecturers who have directly or indirectly helped our project.

We pay our respects and love to our parents and all other family members and friends for their love and encouragement throughout our career. Last but not the least we express our thanks to our friends for their cooperation and support.

Signature:

Names:

Esha Mahendra (9916102024)

Harshita Madan (9916102069)

Shivangi Gupta (9916102126)

TABLE OF CONTENTS

Contents	Page No.
<i>Abstract</i>	<i>i</i>
<i>Acknowledgement</i>	<i>ii</i>
<i>Table of Contents</i>	<i>iii</i>
CHAPTER 1: INTRODUCTION	
1.1 Background Study	1
1.2 Motivation	2
1.3 Project Goal	2
CHAPTER 2: LITERATURE SURVEY	
2.1 PAASBAAN-Crime Prediction and Classification in Indore City	4
2.2 Machine Learning Approach to Predict Crime Using time and Location	4
2.3 Crime Prediction And Analysis using ML	4
CHAPTER 3: ALGORITHMS USED	
3.1 KNN	5
3.2 Decision Tree	6
3.3 Linear Regression	7
3.4 SVM	8
CHAPTER 4: DETAILED DESIGN	
4.1 Data Collection	9
4.2 Data Pre-Processing	10
4.3 Training and Testing the Model of Data	11
4.4 Implementation of the model	11
4.5 Visualization	12
CHAPTER 5: IMPLEMENTATION	
5.1 Working	14
5.2 Confusion Matrix of Algorithms	15
5.3 Advantages	17
CHAPTER 6: CONCLUSION & FUTURE SCOPE	18
References	19
Appendices	

CHAPTER 1: INTRODUCTION

1.1 Background Study

Our project is based on technology MACHINE LEARNING. Through many documentation and cases, it has come out that machine learning and data science can solve the crime cases easier and faster. Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed. It gives the computer that which makes it more similar to humans. One need DATA+OUTPUT which is run on machine during training and the machine creates its own program (logic) which can be evaluated while testing.

Machine learning is also widely used in scientific applications such as bioinformatics, medicine, and astronomy. One common feature of all of these applications is that, in contrast to more traditional uses of computers, in these cases, due to the complexity of the patterns that need to be detected, a human programmer cannot provide detailed specification of how such tasks should be executed. Machine learning tools are concerned with endowing programs with the ability to learn and adapt.

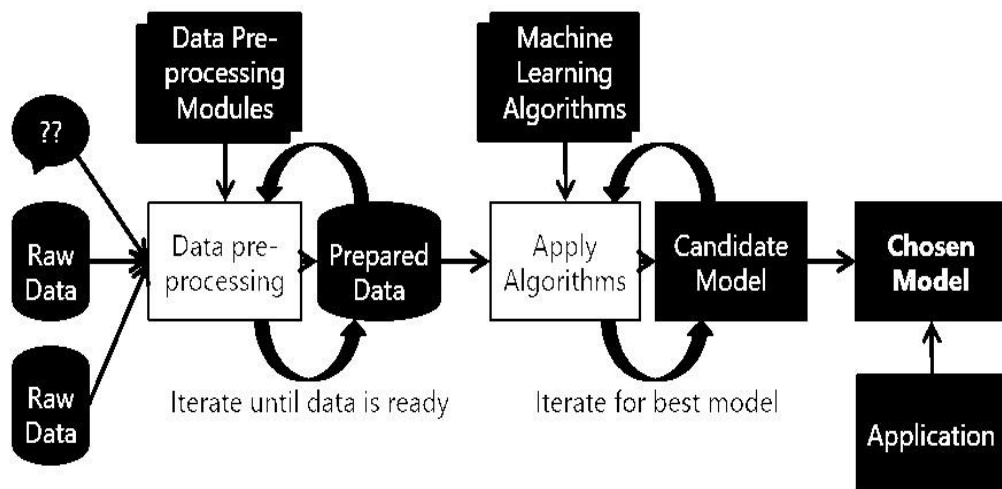


Fig 1.1 Machine Learning Process^[1]

1.2 Motivation

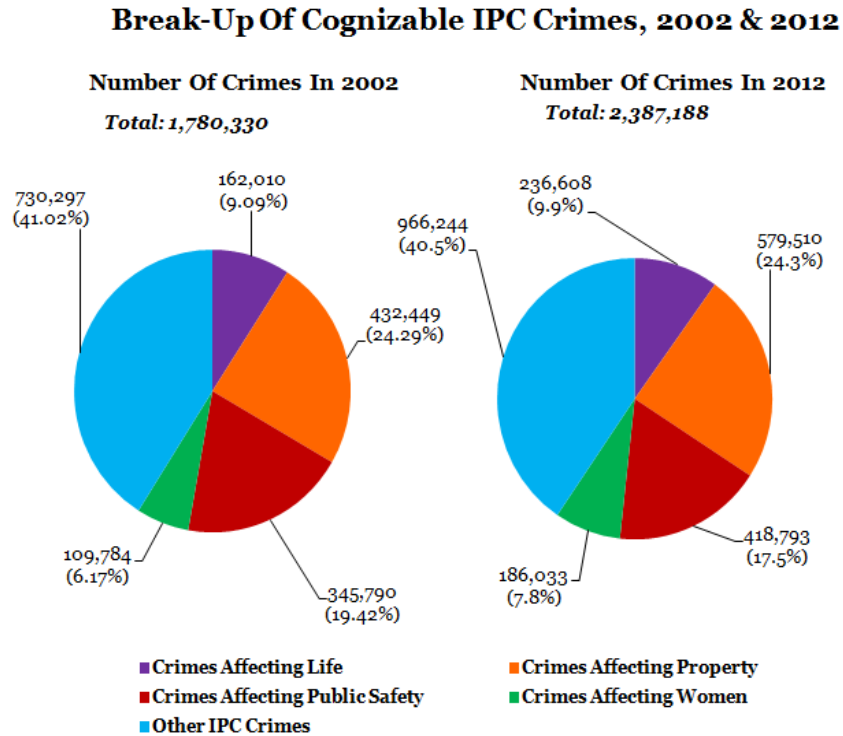


Fig.1.2 Report of Crimes^[2]

Crimes are the significant threat to the humankind. There are many crimes that happens regular interval of time. Perhaps it is increasing and spreading at a faster rate. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data that exist. There is a need of technology through which through which the case solving could be faster.

1.3 Project Goal^[3]

With the rapid urbanization and development of big towns and cities, the graph of the crimes is also on increase. This phenomenal rise in offenses and crime in the cities is a matter of great concern and alarm to us all. This requires keeping track of all the crimes and maintaining a database for same which may be used for future reference. Crime prediction and criminal identification are the major problems to the police department as there are tremendous amount of crime data

that exist. There is a need of technology through which the case solving could be faster. Through many documentation and cases, it came out that machine learning and data science can make the work easier and faster.

The soul purpose of the project to give a jest idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate. The current work is focused mainly in two directions: Predicting the hotspots of crime and understanding the criminal behaviour that could help in solving criminal investigation.

In this project, we will be using the technique of machine learning and data science for crime prediction of the state-wise crime dataset. The crime data is extracted from the official website Kaggle.com. It consists of the crime information like the location description, type of crime, time and the total no. of crimes that took place. Before training of the model data preprocessing will be done following this feature selection and scaling will be done so that accuracy obtain will be high. The K-Nearest Neighbor (KNN), Linear Regression, decision tree and SVM algorithms will be tested for crime prediction and one with better accuracy will be used for train data. The entries was done just to make the machine learn what all it has to do with the data and what actually the output is being demanded. As soon as the machine learnt the algorithms and the process, accuracy of different algorithms were measured & the algorithm with the most accuracy is used for the prediction. Visualization of dataset is done in terms of graphical representation of many cases for example at which time the criminal rates are high or in which the criminal activities are high.

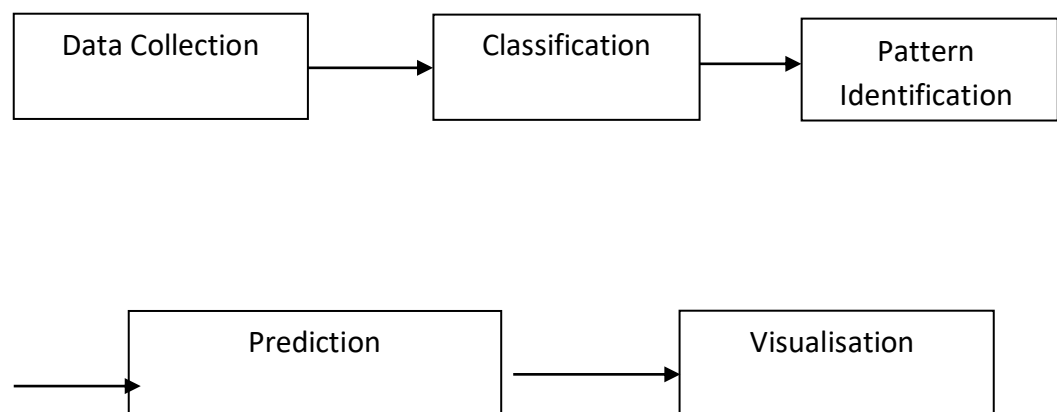


Fig 1.3 Flowchart of Machine Learning Process

CHAPTER 2: LITERATURE SURVEY^[4]

2.1 PAASBAAN – Crime Prediction and Classification in Indore City

Analysis of crime data from the city of Indore, scraped from publicly available website of Indore Police. The task was to predict which category of crime is most likely to occur given a time and place in Indore. The use of ML/AI to detect crime via sound or cameras currently exists, was proven to work, and expected to continue to expand. The objective of our work is to: Predicting crime before it takes place, Predicting hotspots of crime, Understanding crime pattern, Classify crime based on location and Analysis of crime in Indore.

2.2 A Machine Learning Approach to Predict Crime Using Time and Location Data

In this research, a dataset from San-Francisco Open Data is used which contains the reported criminal activities in the neighborhoods of the city San Francisco for a duration of 12 years. I used different classification techniques like Decision Tree, Naive Bayesian, Logistic Regression, k-Nearest Neighbor, Ensemble Methods to find hotspots of criminal activities based on the time of day. Results of different algorithms have been compared and most the effective approach has also been documented .

2.3 Crime Prediction and Analysis Using Machine Learning

The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending upon the accuracy. The K-Nearest Neighbor (KNN) classification and other algorithm will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country. This work helps the law enforcement agencies to predict and detect crimes in Chicago with improved accuracy and thus reduces the crime rate.

CHAPTER 3: ALGORITHMS USED

For the purpose of proper implementation and functioning several Algorithms are used.

3.1 K-NEAREST NEIGHBOUR^[5]

A powerful classification algorithm used in pattern recognition K nearest neighbors stores all available cases and classifies new cases based on a similarity measure (e.g. distance function). One of the top data mining algorithms used today. A non-parametric lazy learning algorithm (An Instance based Learning method).

KNN: Classification Approach

An object (a new instance) is classified by a majority votes for its neighbor classes.

The object is assigned to the most common class amongst its K nearest neighbors. (measured by distance function).

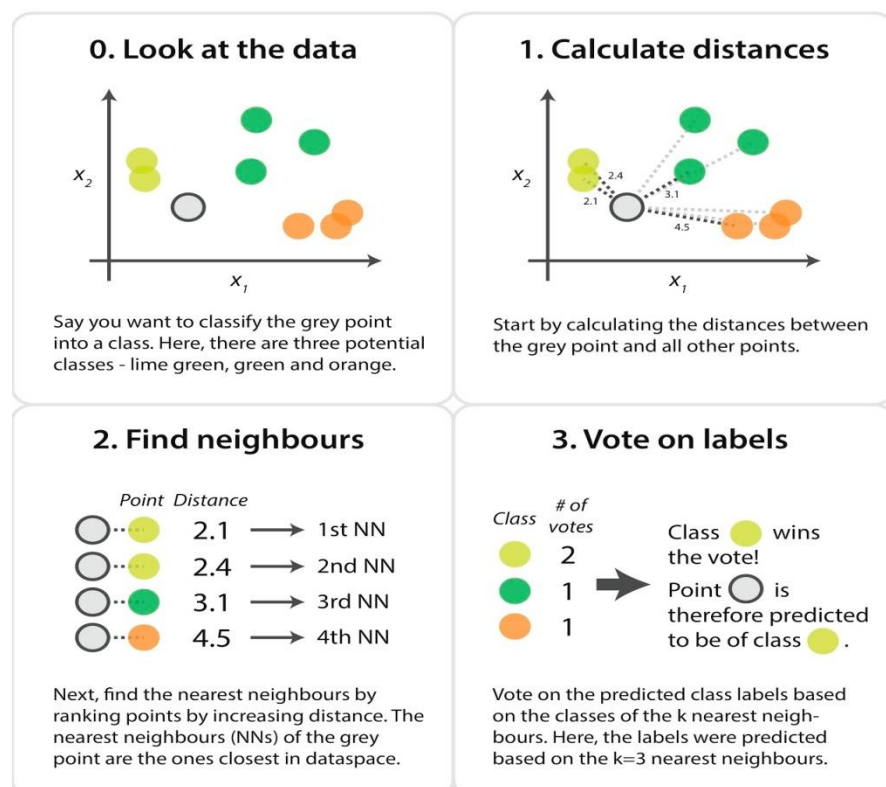


Fig 3.1 Representation of KNN

3.2 Decision Tree ^[6]

It is a tree which helps us by assisting us in decision-making. Used for both classification and regression, it is a very basic and important predictive learning algorithm. It is different from others because it works intuitively i.e., taking decisions one-by-one. It consists of nodes which have parent-child relationships. Decision tree considers the most important variable using some fancy criterion and splits dataset based on it. It is done to reach a stage where we have **homogenous subsets** that are giving predictions with utmost surety.

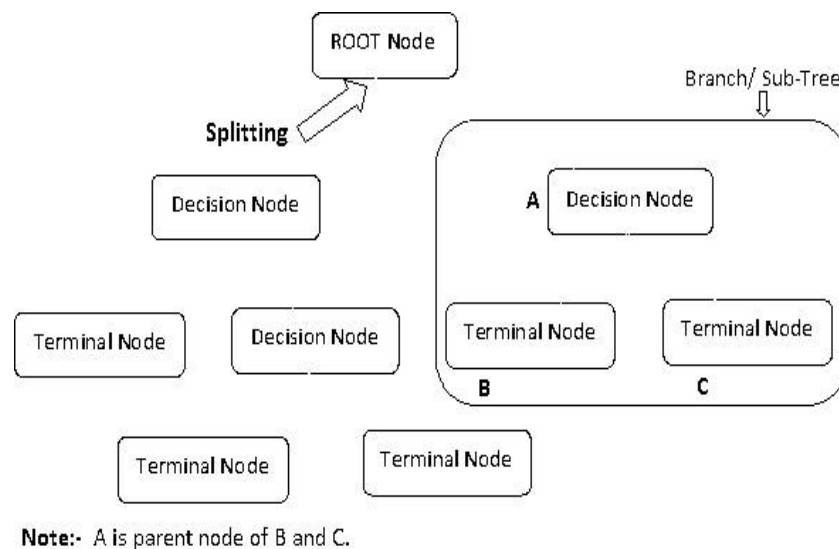


Fig 3.2 Decision Tree Example

3.3 Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y (output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

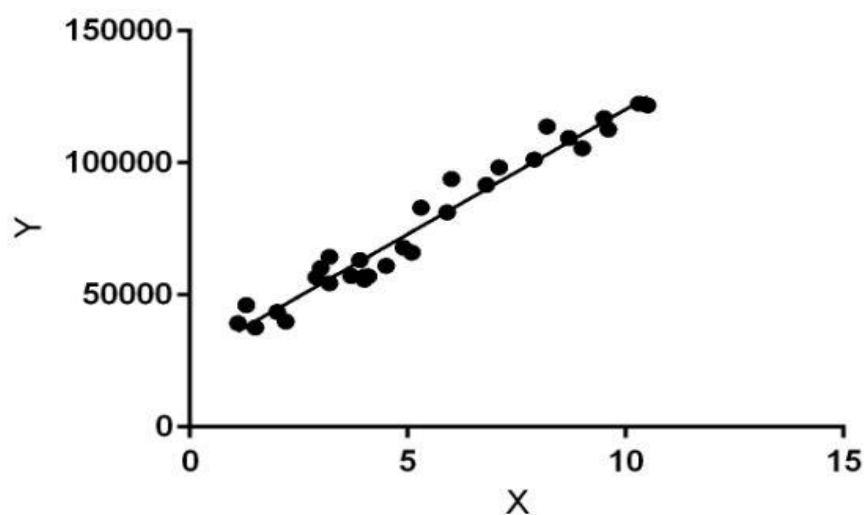


Fig 3.3 Linear Regression

3.4 SVM ^[7]

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, we 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. Then, we perform classification by finding the hyper-plane that differentiate the two classes. Support vectors are simply the coordinates of the individual observation.

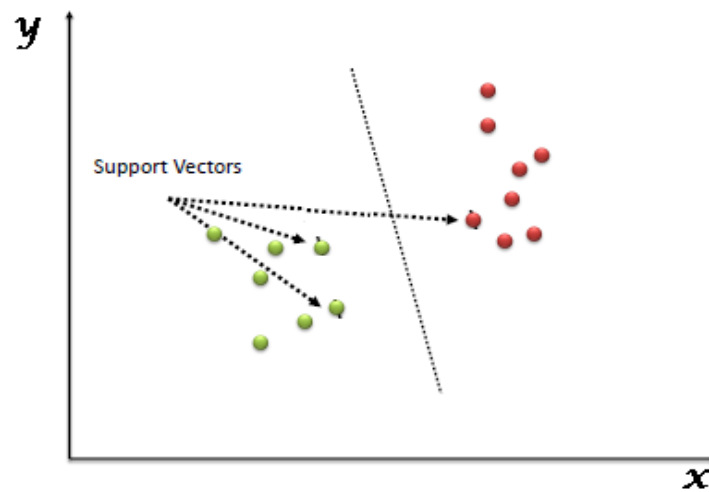


Fig 3.4 Plane Showing Support Vectors

CHAPTER 4: DETAILED DESIGN

4.1 DATA COLLECTION

The Crime data set which is used in project is collected from the official website kaggle.com that stores the data in CSV(Comma Separated Values) format which stores it in tabular form. The collected data cannot be used directly for performing the analysis process as there was lot of missing data, many large values and it was unorganized. Therefore, Data preparation was done. Among the various types of data, our dataset belong to the Categorical datatype. Below is the snapshot of our dataset.

Clipboard																		Font		Alignment		Number		Formatting		Table		Styles	
A701			DELHI UT																										
	Name Box	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q												
699	DAMAN & DIU		2001	1	0	0	0	0	0	0	0	0	0	0	0	4	6												
700	DAMAN & TOTAL		2001	7	5	0	0	0	0	5	3	2	0	0	0	43	40												
701	DELHI UT CENTRAL		2001	29	42	4	28	0	28	62	42	20	3	1	42	149	1595												
702	DELHI UT DELHI UT		2001	547	510	63	381	0	381	1627	964	663	48	74	624	3029	19276												
703	DELHI UT EAST		2001	56	51	15	24	0	24	163	39	124	4	15	47	270	1651												
704	DELHI UT G.R.P.(RLY)		2001	11	4	0	1	0	1	7	7	0	1	0	11	2	1430												
705	DELHI UT I.G.I. AIRP		2001	2	2	1	0	0	0	4	1	3	0	0	2	0	52												
706	DELHI UT NEW DELH		2001	11	14	0	9	0	9	27	19	8	1	0	22	75	1237												
707	DELHI UT NORTH		2001	28	33	5	28	0	28	99	53	46	2	5	52	177	1364												
708	DELHI UT NORTH EA		2001	62	79	5	36	0	36	157	123	34	3	6	74	157	961												
709	DELHI UT NORTH W		2001	134	107	13	103	0	103	518	298	220	14	9	122	619	2850												
710	DELHI UT S.T.F.		2001	0	0	0	0	0	0	0	0	0	0	0	0	0	0												
711	DELHI UT SOUTH		2001	83	75	7	65	0	65	265	170	95	9	10	120	803	4220												
712	DELHI UT SOUTH W		2001	61	52	1	42	0	42	174	113	61	9	4	74	359	1816												
713	DELHI UT WEST		2001	70	51	12	45	0	45	151	99	52	2	24	58	418	2100												
714	LAKSHADV LAKSHADV		2001	1	0	0	0	0	0	0	0	0	0	0	0	1	10												
715	LAKSHADV TOTAL		2001	1	0	0	0	0	0	0	0	0	0	0	0	1	10												
716	PUDUCHEI PONDICHE		2001	25	32	1	9	0	9	4	3	1	1	0	4	111	528												
717	PUDUCHEI TOTAL		2001	25	32	1	9	0	9	4	3	1	1	0	4	111	528												
718	ANDHRA P ADILABAD		2002	100	79	17	37	0	37	42	32	10	7	0	40	193	226												
719	ANDHRA P ANANTAPI		2002	166	113	7	28	0	28	65	49	16	13	0	21	196	356												
720	ANDHRA P CHITTOOR		2002	105	66	5	28	0	28	67	39	28	4	0	14	208	832												
721	ANDHRA P CUDDAPAI		2002	91	0	9	16	0	16	27	22	5	1	0	11	84	205												
722	ANDHRA P EAST GOD.		2002	97	102	2	25	0	25	55	5	50	3	0	12	468	1206												
723	ANDHRA P GUNTAKAI		2002	0	0	0	1	0	1	0	0	0	0	0	2	0	148												
724	ANDHRA P GUNTUR		2002	143	77	1	46	0	46	87	71	16	12	0	53	357	1182												
725	ANDHRA P HYDERAB		2002	117	109	3	57	0	57	115	61	54	11	2	68	1348	3777												
726	ANDHRA P KARIMNAC		2002	160	88	6	99	0	99	62	42	20	18	0	60	216	404												
727	ANDHRA P KHAMMAN		2002	106	55	1	48	0	48	70	55	15	5	0	11	164	431												

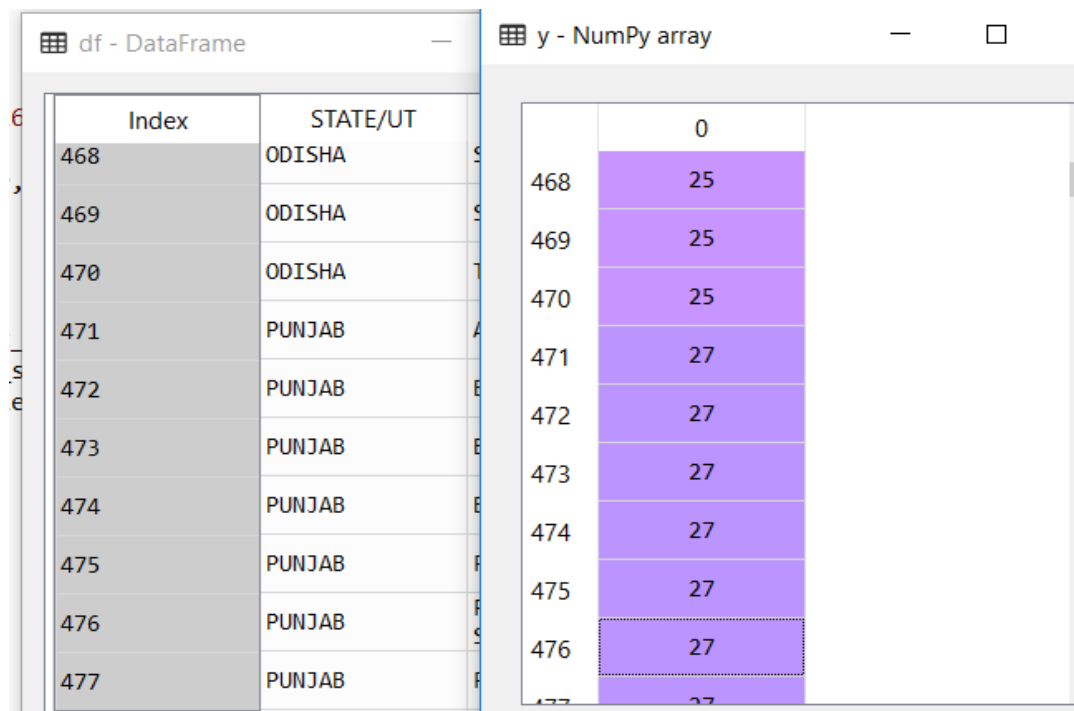
Fig 4.1 Crime Dataset^[8]

4.2 DATA PRE-PROCESSING^[9]

It is one of the most important steps in machine learning as it involves the cleaning of raw data into clean data. Basic pre-processing techniques that can be used to convert the raw data are:

- The categorical and raw data (Location, Area, Crime Type) must be converted into numeric data using the Label Encoder.
- The missing data can be removed by removing the row or column of data depending on the need.
- If the data is missing, one can add manually usually the mean, median or the highest frequency value.
- The data can also be predicted of the empty position by help of existing data using machine learning.

In Machine Learning, 80% of data is used to train the model and rest 20% for the testing.



Index	STATE/UT
468	ODISHA
469	ODISHA
470	ODISHA
471	PUNJAB
472	PUNJAB
473	PUNJAB
474	PUNJAB
475	PUNJAB
476	PUNJAB
477	PUNJAB

	0
468	25
469	25
470	25
471	27
472	27
473	27
474	27
475	27
476	27
477	27

Fig 4.2 Data before and after Pre-Processing

4.3 TRAINING AND TESTING THE MODEL OF DATA^[10]

For training of model, we initially split the model into 3 sections: Training Data, Validation Data and Testing Data.

- **TRAINING SET:** A set of data used for learning, that is to fit the parameters of classifier. It uses algorithms to perform the training part.
- **VALIDATION SET:** Cross validation is primarily used in applied machine learning model on unseen data. A set of unseen data is used from the training data to tune the parameters of a classifier.
- **TEST SET:** A set of unseen data used only to assess the performance of fully-specified classifier.

Once the data is divided into the 3 given sets, training process gets started.

4.4 IMPLEMENTATION OF THE MODEL^[11]

The main goal is to train the best performing data using the pre-processed data. The process of training an ML model involves providing an ML Algorithm with the training data to learn from. In our dataset each data was tagged with the correct label. When the AI Sytem is presented with the data which is labeled Supervised Learning is used. The model suited the Classification category under the Supervised Learning as the target variable is categorical. The classifications algorithms used in the project are:

- K-Nearest Neighbor (KNN)
- Decision Tree
- Support Vector Machine (SVM)
- Linear Regression

4.5 VISUALISATION ^[12]

Data visualization is an important skill in applied statistics and machine learning. Data visualizations can be used to express and demonstrate key relationships in plots and charts. Analysis of crime data set is done by plotting various graphs. Below are the graphs obtain in our projects showing the total IPC crimes and crimes against women.

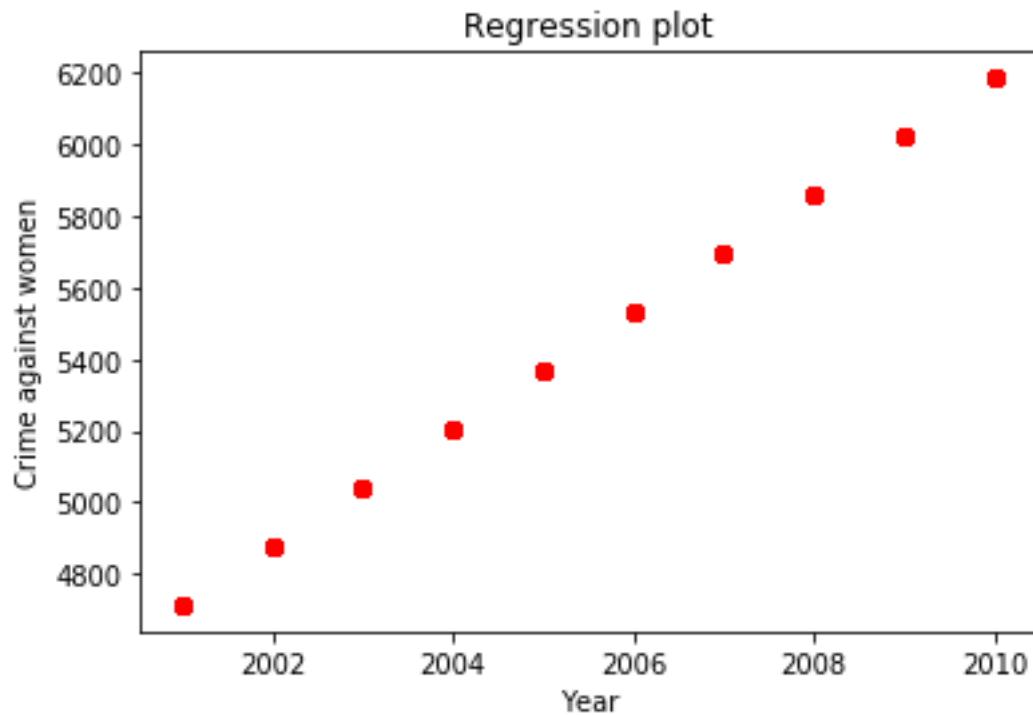


Fig 4.3 Graph showing crime against women

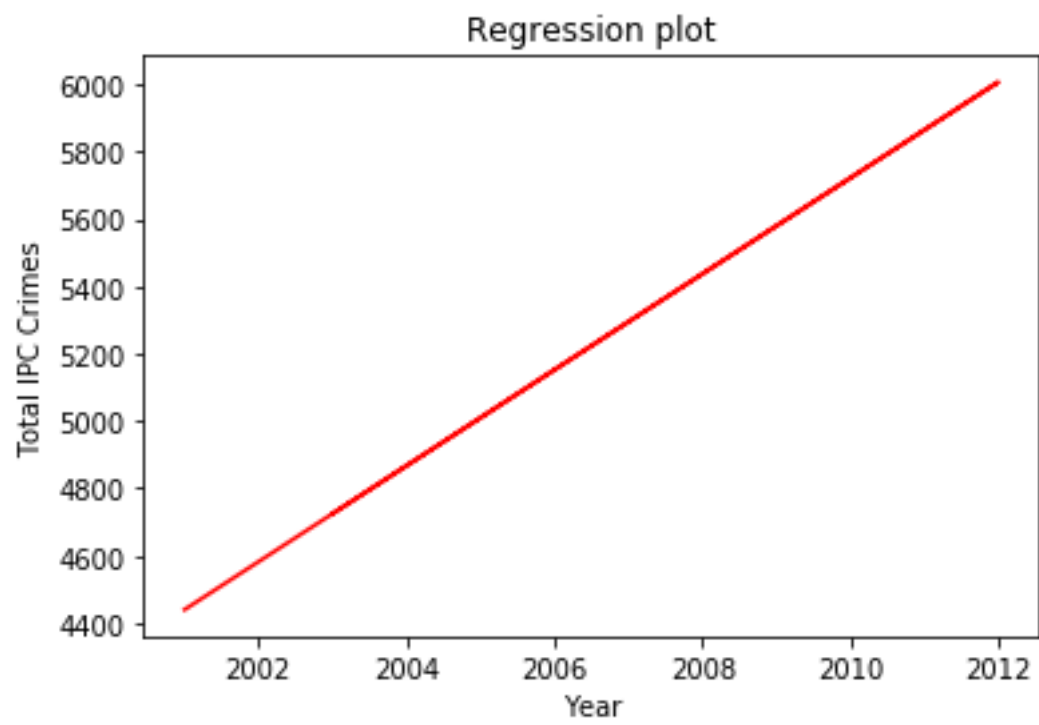


Fig 4.4 Graph showing the total IPC

CHAPTER 5: IMPLEMENTATION

5.1 WORKING ^[13]

The implementation of the project is done with the help of python language. To be particular, for the purpose of machine learning Anaconda is being used. The detailed design is already discussed above. Using the matplotlib library from sklearn plots were obtained. The analysis is performed by applying various algorithms and then checking the accuracy of each algorithm applied. The KNN algorithm had the highest accuracy. After dividing the data set into training set and testing set the model is trained using the algorithm having the highest accuracy. The table shows the accuracy obtained after applying algorithms.

K NEAREST NEIGHBOURS	86.5388816850331
SUPPORT VECTOR MACHINE	75.05543237250555
NAÏVE BAYES	12.250554323725056
DECISION TREE	72.11751662971175

Fig 5.1 Accuracy obtain after testing

The Confusion matrix obtain after implementing the algorithms help to describe the performance of classification model. Also known as error matrix, it has a specified table layout that allows the visualization of the performance of an algorithm. The number of correct and incorrect predictions are summarized with count values and broken down class. After implementation of various algorithms and training of data, linear regression gave the various plots that helped in analysis of crime dataset. Below are the confusion matrix of the algorithms applied. The array shows the error obtain after the testing and predicting the data.

5.2 CONFUSION MATRIX OF ALGORITHMS

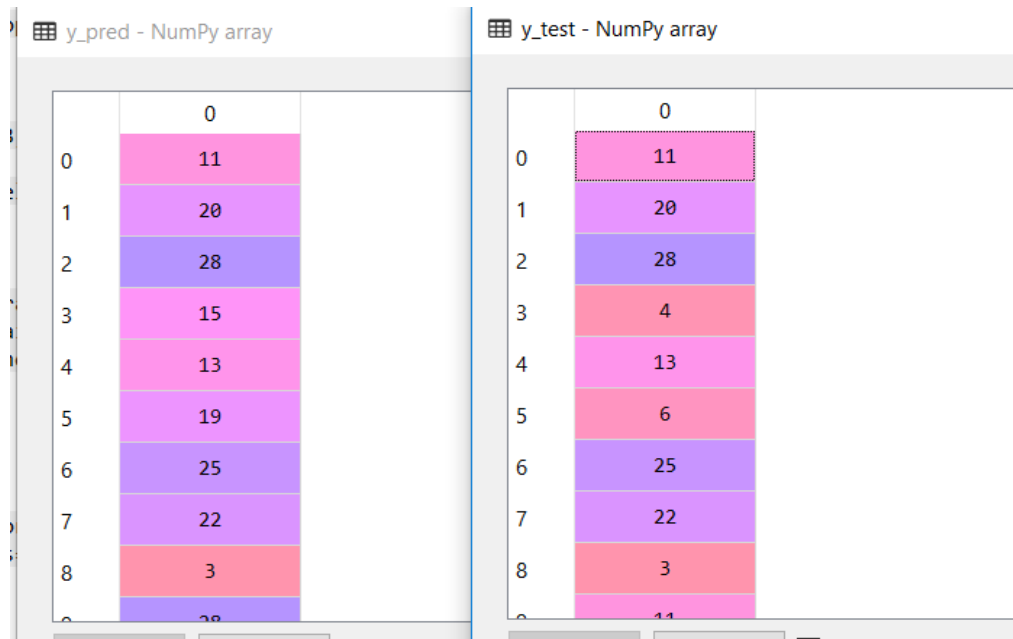


Fig 5.2 Decision Tree

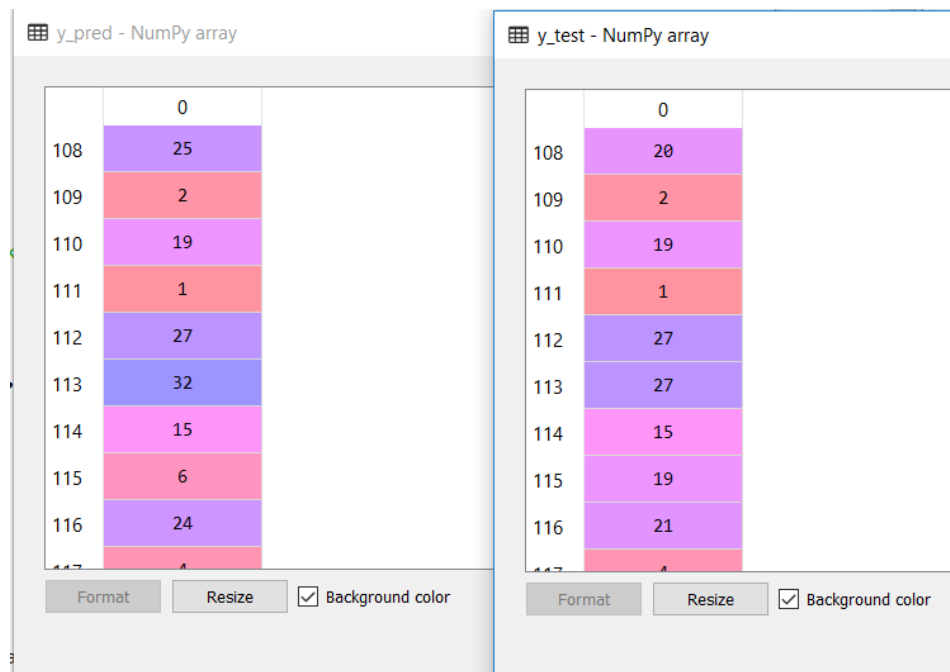


Fig 5.3 SVM

y_test - NumPy array	
	0
74	32
75	4
76	25
77	20
78	30
79	16
80	20
81	6
82	1

y_pred - NumPy array	
	0
74	32
75	4
76	25
77	20
78	30
79	16
80	20
81	19
82	2

Fig 5.4 Linear Regression

5.3 ADVANTAGES

- The idea behind this project is that crimes are relatively predictable; it just requires being able to sort through a massive volume of data to find patterns that are useful to law enforcement.
- Companies that are directly involved in providing governments with AI tools to monitor areas or predict crime will likely benefit from a positive feedback loop.
- The soul purpose of this project is to give a jest idea of how machine learning can be used by the law enforcement agencies to detect, predict and solve crimes at a much faster rate and thus reduces the crime rate.

CHAPTER 6: CONCLUSION & FUTURE SCOPE

With the help of machine learning technology, it has become easy to find out relation and patterns among various data's. The work in this project mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. The model predicts the type of crime with accuracy. Data visualization helps in analysis of data set. We generated many graphs and found interesting statistics that helped in understanding crimes datasets that can help in capturing the factors that can help in keeping society safe.

REFERENCES:

1. <https://machinelearningmastery.com/k-fold-cross-validation/>
2. Alkesh Bharati, DR Sarvanaguru RA.K (Sep 2018), "Crime Prediction Analysis Using Machine Learning", International Research Journal Of Engineering And Technology (IRJET), Volume: 05 Issue: 09 | Sep 2018,
3. <http://scikit-learn.org.stable/>
4. <https://app.dataquest.io/dashboard>
5. <https://pythonspot.com/k-nearest-neighbours/>
6. <https://www.datacamp.com/community/tutorials/decision-tree-classification-python>
7. <https://pythonspot.cpm/support-vector-machine/>
8. <https://www.kaggle.com/>
9. <https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85?gi=6cfcacc0cad8>
10. <https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/>
11. <https://machinelearningmastery.com/how-to-one-hot-encode-sequence-data-in-python/>
12. http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
13. <https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-result>

APPENDICES:

#IMPORTING LIBRARIES AND DATASET

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
df=pd.read_csv('C:/Users/Lenovo/Desktop/01_District_wise_crimes_committed_I
PC_2001_2012.csv')
df.shape
df.describe()
y=df.iloc[:,0].values
```

```
X=df.iloc[:,[3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,
28,29,30,31]].values
```

DATA PREPROCESSING

```
df.isnull().values.any()
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
le=LabelEncoder()
y=le.fit_transform(y)
```

SPLITTING THE DATASET

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,
random_state=0)
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
X_train=sc.fit_transform(X_train)
X_test=sc.transform(X_test)
```

#IMPLEMENTING KNN MODEL

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2)
knn.fit(X_train,y_train)
```

```
y_pred=knn.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix
```

```
cm_rf=confusion_matrix(y_test,y_pred)
```

```
#APPLYING K-FOLD CROSS VALIDATION.
```

```
from sklearn.model_selection import cross_val_score
```

```
accuracies=cross_val_score(estimator = knn, X=X_train, y=y_train, cv= 10,  
n_jobs=-1)
```

```
accuracy=accuracies.mean()*100
```

```
var=accuracies.std()*100
```

```
#IMPLEMENTING DECISION TREE
```

```
from sklearn.tree import DecisionTreeClassifier
```

```
dtree = DecisionTreeClassifier(max_depth=500, random_state=300)
```

```
dtree.fit(X_train,y_train)
```

```
y_pred=dtree.predict(X_test)
```

```
from sklearn.metrics import accuracy_score
```

```
acc=accuracy_score(y_test, y_pred)
```

```
dtree.score(X_test,y_test)
```

```
dtree.score(X_train,y_train)
```

```
#IMPLEMENTING SVM MODEL
```

```
from sklearn import svm
```

```
clf = svm.SVC(kernel='linear')
```

```
clf.fit(X_train, y_train)
```

```
y_pred = clf.predict(X_test)
```

```
from sklearn import metrics
```

```
acc=metrics.accuracy_score(y_test, y_pred)
```

```
# IMPLEMENTATION FOR LINEAR REGRESSION
```

```
import pandas as pd
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
df=pd.read_csv("C:/Users/Lenovo/Desktop/43_Arrests_under_crime_against_wo  
men.csv")
```

```

y=df.iloc[:,[15]].values
X=df.iloc[:,[1]].values
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)
from sklearn import linear_model
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
reg=linear_model.LinearRegression()
reg.fit(X_train, y_train)
y_pred=reg.predict(X_test)
acc=reg.score(X_test,y_test)
acc1=reg.score(X_test,y_pred)
plt.scatter(X_test,y_pred, color='red')
plt.title('Regression plot')
plt.xlabel('Year')
plt.ylabel('Crime against women')

```

ANALYSIS

```

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sn
get_ipython().magic(u'matplotlib inline')
sn.set_style("darkgrid")
df12 =
pd.read_csv("C:/Users/Lenovo/Desktop/01_district_wise_crimes_committed_ipc_
2001_2012 (2).csv")
df13 =
pd.read_csv("C:/Users/Lenovo/Desktop/01_district_wise_crimes_committed_ipc_
2013 (2).csv")
df14 =
pd.read_csv("C:/Users/Lenovo/Desktop/01_district_wise_crimes_committed_ipc_
2014 (2).csv")
df = pd.concat([df12, df13, df14])
df['state_ut'] = df.state_ut.str.replace('\s+&\s+', '&')

```

```

df['district_ut'] = df.district.apply(lambda x: x.replace('\s+&\s+', '&'))

no_total = df.loc[df.district != 'Total', :]
yr_totals = no_total.groupby('year')
yr_agg = yr_totals.sum().reset_index()
for cl in yr_agg.columns:
    if cl == 'year':
        continue
    if yr_agg[cl].isnull().sum() > 10:
        yr_agg = yr_agg.drop(cl, axis=1)
    else:
        mean_cl = np.mean(yr_agg[cl])
        yr_agg[cl] = yr_agg[cl].fillna(mean_cl)
        yr_agg[cl] = yr_agg[cl].apply(lambda x: x/10000)
yr_agg.describe()
cols = list(yr_agg.columns)
cols.remove('year')
cols.remove('total_ipc_crimes')
cols.remove('other_ipc_crimes')
fig = plt.figure()
fig.set_size_inches(15, 10)
ax = plt.subplot(111)
# ax.set_xlim([2001, 2014])
ax.set_title("Crime committed in India (in ten thousands)")
ax.set_xlabel("Year")
ax.set_ylabel("Number of cases filed")
for col in cols:
    ax.plot(yr_agg.year, yr_agg[col], label=col.replace('_', ' '))
ax.legend(loc=5, bbox_to_anchor=(1.5, .5))

group_by = ["state_ut"]
columns_of_interest = ["total_ipc_crimes"]
state_grp = no_total[columns_of_interest + group_by].groupby(["state_ut"])
state_agg = state_grp.sum().reset_index().sort_values(by='total_ipc_crimes')
state_agg['total_ipc_crimes'] = state_agg.total_ipc_crimes

```

```
fig = plt.figure()
fig.set_size_inches(20, 5)
ax = fig.add_subplot(111)
ax.set_xticklabels(state_agg.state_ut, rotation=90)
b = sn.barplot(x='state_ut', y='total_ipc_crimes', data=state_agg, ax=ax)
```