**Project Report On**

**HORIZON-2020: A STEP TOWARDS CRIME FREE INDIA**

**Submitted in partial fulfilment for the award**

**Of**

**PG DIPLOMA IN BIG DATA ANALYTICS**



CENTRE FOR DEVELOPMENT OF ADVANCED COMPUTING (C-DAC) ADVANCED COMPUTING TRAINING SCHOOL (ACTS)

Knowledge Park, Bengaluru - 560 038

Submitted By

ABHISHEK PALO

ASHWINI CHOURE

DEBJIT MISHRA

SUNNY KUMAR

PRN: 170850125002

PRN: 170850125007

PRN: 170850120311

PRN: 170850125040

Guided by

Mr. MOHIT VED

SENIOR TECHNICAL OFFICER (CDAC R&D BENGALURU)

CERTIFICATE

This is to certify that, the project report entitled

**HORIZON-2020: A STEP TOWARDS CRIME FREE INIDA**

Submitted By

ABHISHEK PALO PRN: 170850125002

ASHWINI CHOURE PRN: 170850125007

DEBJIT MISHRA PRN: 170850125011

SUNNY KUMAR PRN: 17085012040

is the record of bonafide work carried out by them in partial fulfilment of the requirement for the award of PG Diploma in Big Data Analytics prescribed by Centre For Development Of Advanced Computing (C-DAC).

**Mr. MOHIT VED**

**Project Guide**

**Mrs. Janaki C.H. Mrs.M.Savitri**

**Project Co-ordinator Course Co-ordinator**

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to all the people without whom this project would have been highly impossible.

We would like to devote our first vote of thanks to our guide Mr.Mohit Ved for their constant support and encouragement. They have a great hand in the firm foundation of this project. We are deeply indebted for their valuable suggestions, scholarly guidance and constructive criticisms along with constant encouragement at each and every step for successful completion of the project.

We would also like to thank our Project Co-ordinator Mrs. Janaki for inspiring us towards completion of this project.

Last but not the least we would like to thank all those who assisted us directly or indirectly for their valuable time and help.

# TABLE OF CONTENTS

Contents

[ABSTRACT 5](#_Toc504920501)

[INTRODUCTION 6](#_Toc504920502)

[AIMS AND OBJECTIVES 7](#_Toc504920503)

[DATASETS 8](#_Toc504920504)

[PROBLEM STATEMENTS 9](#_Toc504920505)

[METHODOLOGY 10](#_Toc504920506)

[CODES 11](#_Toc504920507)

[CONCLUSION 11](#_Toc504920508)

# ABSTRACT

Law and justice is one division in India where advancements in technologies have not been utilized to the fullest potential. Be it organized or unorganized crime, the security, intelligence & record keeping systems even today follow very old practices that can be made lot better using statistical analysis techniques like ML and Data Visualization and easily made available at a granular levels.

In this project we take up one such segment of the law and Justice sector, which is to measure and predict the crime pattern across states in India and propose a fault proof system for enhancing awareness among end users. In the existing system, the crime data available has no structured and measurable form which can provide meaningful observations and insights to public and law enforcing agencies. All this makes it a difficult information gathering and sharing model.

In this project we propose a ML and DV based analytic model and application which combines many features such as descriptive, comparative and predictive study of the crime pattern across district and states in India broadly categorized into crime against women, children and oppressed classes. The statistical analytic techniques and graphical representation makes it relatively easier to comprehend and take precautionary or corrective measures as appropriate.

# INTRODUCTION

* The last few decades has seen a big increase in India’s population owing to better health care, ever expanding economy and a stable administration. There has been noticeable movement of masses from rural heartland to bigger cities in search of better opportunities. As such, there has been a greater focus to carry out independent comprehensive study on the crime pattern in changing times and demography which would in turn provide more visibility and awareness for initiating corrective and precautionary measures.
* Horizon 20-20 aims at taking a closer look and doing an in-depth-study at the current crime patterns, possible factors and investigating the possibility to predict criminal events for specific areas. The application will make use of new generation software technologies as ML and Big Data tools to analyze and predict.

# 

# AIMS AND OBJECTIVES

Objectives are stated below:-

* Descriptive summary of crime patterns region-wise.
* Groupings of high disturbed, moderate disturbed and low disturbed regions.
* Prediction of criminal activities for specific regions
* Comparative study of criminal activities over the years (decline/increase)
* The output will be in form of graphical representation for quick and easier understanding.

# DATASETS

The datasets contains complete information about various aspects of crimes happened

in India from 2001-14.

Reference- <https://www.kaggle.com/rajanand/crime-in-india>

I: SC/ST Cases Reported and their Disposal by Police and Court

* Crime against SCs
* Crime against STs

II: Children Cases Reported and their Disposal by Police and Court

* Buying of Girls for Prostitution (Section 373 IPC)
* Child Marriage Restraint Act, 1929
* Foeticide (Section 315 and 316 IPC)
* Murder (Section 302, 315 IPC)
* Other Crimes against Children
* Total Crimes against Children

III: Cases under Crime against Women

* Rape
* Kidnapping & Abduction of Women & Girls
* Dowry Deaths
* Cruelty by Husband and Relatives
* Importation of Girls
* Assault
* Total Crimes against Women
* Importation of Girls

Some of the datasets contains district level information.

The districts are police district headquarters or special police unit.

Therefore these may be different from revenue districts.

Most of the data is from 2001 to 2010.

But there are few files which has data only from 2011 and few are having 2001-14.

# 

# PROBLEM STATEMENTS

**Descriptive analysis**

* District wise analysis
* State wise analysis
* Nature and density of crime in a particular locality for people to take corrective and precautionary measures.
* Overall crime awareness for an immigrant to a new city (depending on the data the person can take measures to choose and reside in a locality of his/her choice.
* Administration efficacy in dealing with criminal incidents (how efficiently police functions in a district/state).
* Crime against women.
* Crime against oppressed classes.

**Comparative analysis**

* District wise
* State wise
* Crime against women vs. all cases
* Crime against oppressed classes vs. crime against general class
* Number of conviction under trials vs. no. of cases.
* Crime by Juvenile vs. regular cases

**Predictive analysis**

* District wise crime
* State wise crime
* Graphical representation of the data predicting the future trend taking an account of the current trend that can be used by law enforcing agencies and media.
* Future prediction of crime against oppressed classes with respect to current scenario.

# METHODOLOGY

eXtreme Programming (XP) technique has been used to develop various modules of the application.

**Planning**

1. Identification of problem statement
2. Infrastructure Requirements(s/w and h/w)
3. Collection of crime datasets for 2001 till 2014

**Analysis**

1. Analysis of available datasets
2. Prioritize stories to prepare the model
3. Scrubbing of stories for estimation
4. Resource planning for DV and ML

**Design**

1. Break down of tasks (WBS)
2. System requirement specification document containing the flow.

**Execution**

1. Data prreprocessing using excel for DV through tableau on descriptive statistics
2. Data preprocessing through python for restructuring the data sets
3. Application of python pandas and ML techniques (Arima, Regression) to build the summary and models.
4. Usage of various python libraries such as MatPlotLib,Seaborn,Altair for representing the data through graphs and charts.
5. Unit Testing through Jupyter notebook.
6. Execution of Manual test scenarios using Jupyter notebook
7. Mid Iteration peer review
8. End of Iteration peer review

**Wrapping**

1. Release through Tableau and python framework
2. Demos and final review

**Closure**

1. Pilot Launch

# CODES

1. Data Preprocessing

* Merge similar datasets containing data for different years
* Cleaning of bad data,
* Renaming of columns to appropriate standards
* Merge separate notations of categorical values(state/district) to unique.

*# Considering data between 2001 to 2013*

crimes1 = pd.read\_csv('/home/java/Downloads/HORIZON2020/DATASETS/women/42\_District\_wise\_crimes\_committed\_against\_women\_2001\_2012.csv')

crimes2 = pd.read\_csv('/home/java/Downloads/HORIZON2020/DATASETS/women/42\_District\_wise\_crimes\_committed\_against\_women\_2013.csv')

*# Concat data sets from 2 files*

crimes = pd.concat([crimes1,crimes2], ignore\_index=**False**, axis=0)

*# rename the STATE/UT column to STATE*

crimes.rename(columns={'STATE/UT':'STATE'}, inplace=**True**)

1. Descriptive Approach

* State wise categorical description

crimes\_total\_women = crimes\_total\_women[crimes\_total\_women['DISTRICT'] == 'TOTAL']

crimes\_total\_women.drop('DISTRICT', axis=1, inplace=**True**)

crimes\_total\_women['Total Crimes']= crimes\_total\_women.iloc[:, -9:-1].sum(axis=1)

crimes\_total\_women = crimes\_total\_women.groupby(['STATE'])['Total Crimes'].sum()

fig, ax = plt.subplots()

crime\_rape = crimes\_total\_2001['STATE'].values

y\_pos = np.arange(len(crime\_rape))

performance = crimes\_total\_2001['Rape'].values

ax.barh(y\_pos, performance, align='center',color='brown', ecolor='black')

ax.set\_yticks(y\_pos)

ax.set\_yticklabels(crime\_rape)

ax.invert\_yaxis() *# labels read top-to-bottom*

ax.set\_xlabel('Rapes')

ax.set\_title('RAPE VS STATE')

fig.set\_size\_inches(20, 18, forward=**True**)

plt.show()

* Year wise descriptive summary using MatPlotLib and Seaborn

cols = list(year\_agg.columns)

cols.remove('Year')

fig = mplt.figure()

fig.set\_size\_inches(15, 10)

ax = mplt.subplot(111)

ax.set\_title("Crime committed in India (in ten thousands)")

ax.set\_xlabel("Year")

ax.set\_ylabel("Number of cases occured")

**for** col **in** cols:

ax.plot(year\_agg.Year, year\_agg[col], label=col.replace('\_', ' '))

ax.legend(loc=5, bbox\_to\_anchor=(1.5, .5))

* Correleation charts between types of crime against women

corr = year\_agg[cols].corr()

mask = np.array(corr)

mask[np.tril\_indices\_from(mask)] = **False**

fig, ax = mplt.subplots()

fig.set\_size\_inches(15, 15)

sn.heatmap(corr, mask=mask, ax=ax, square=**True**)

* Frequency distribution of crimes by states

fig = mplt.figure()

fig.set\_size\_inches(20, 5)

ax = fig.add\_subplot(111)

ax.set\_xticklabels(state\_agg\_rape.STATE\_UT, rotation=90)

b = sn.barplot(x='STATE\_UT', y='Rape', data=state\_agg\_rape, ax=ax)

* fig2 = mplt.figure()

fig2.set\_size\_inches(20, 5)

ax = fig2.add\_subplot(111)

ax.set\_xticklabels(state\_agg\_dowry.STATE\_UT, rotation=90)

c = sn.barplot(x='STATE\_UT', y='Dowry', data=state\_agg\_dowry, ax=ax)

fig3 = mplt.figure()

fig3.set\_size\_inches(20, 5)

ax = fig3.add\_subplot(111)

ax.set\_xticklabels(state\_agg\_assault.STATE\_UT, rotation=90)

d = sn.barplot(x='STATE\_UT', y='Assault', data=state\_agg\_assault, ax=ax)

1. Comparative Approach

* Using MatPlotLib,Pysal ,Seaborn,Prettyplotlib,Altair
* Number of dacoity state wise

**def** sub\_frame\_total(dataset, group\_cols=['STATE\_UT','YEAR'], col\_name = 'PLACE\_OCCURENCE', category="TOTAL (MANUAL)"):

*'''Subsets the original dataframe to a category'''*

subset\_frame = dataset[dataset[col\_name] == category]

*#print(type(subset\_frame))*

*#print(subset\_frame.head)*

*'''Group by group\_cols and category, returns df with number of occurrences'''*

*#group = subset\_frame.groupby(group\_cols)[col\_name].count().reset\_index()*

group = subset\_frame[['STATE\_UT','YEAR','DACOITY\_NUM']]

*#print(group)*

chart = alt.Chart(group)

chart.configure\_cell(height=400, width=600)

graph = chart.mark\_line().encode(

X('YEAR:T', axis=Axis(title='States')),

Y('DACOITY\_NUM:Q', axis=Axis(title='Number of total Dacoity')),

color='STATE\_UT',)

display(graph)

name = 'Trends\_'+category

path = '/home/dbda/Desktop/'

chart.savechart(path+name+'.html')

*#return graph*

print(type(group['DACOITY\_NUM']))

* Area chart for total crime vs. states

**def** sub\_frame\_areatotal(dataset,crime, group\_cols=['STATE\_UT','YEAR'], col\_name = 'PLACE\_OCCURENCE', category="TOTAL (MANUAL)"):

*'''Subsets the original dataframe to a category'''*

subset\_frame = dataset[dataset[col\_name] == category]

*#print(type(subset\_frame))*

*#print(subset\_frame.head)*

*'''Group by group\_cols and category, returns df with number of occurrences'''*

group = subset\_frame[['STATE\_UT','YEAR',crime]]

print(group.head)

chart = alt.Chart(group)

chart.configure\_cell(height=400, width=1000)

graph = chart.mark\_area().encode(X('YEAR:T', axis=Axis(axisWidth=0.5,

*#format='%M',*

labelAngle=0.0,title='Places')),

Y(crime, axis=Axis(title='Total '+crime+'at different places')),color='STATE\_UT')

display(graph)

name = 'Trends\_'" in "+crime+" "

path = '/home/dbda/Desktop/'

chart.savechart(path+name+'.html')

1. Predictive Approach

* sklearn.linear\_model LogisticRegression
* sklearn.cross\_validation KFold
* sklearn.ensemble RandomForestClassifier
* sklearn.tree DecisionTreeClassifier, export\_graphviz

Classification Model

*# Define a function for making a classification model and accessing performance:*

**def** classification\_model(model, data, predictors, outcome):

*# Fit the model*

model.fit(data[predictors],data[outcome])

*# Make predictions on training set:*

predictions = model.predict(data[predictors])

*# Print accuracy*

accuracy = metrics.accuracy\_score(predictions,data[outcome])

print ('Accuracy : **%s**' % '**{0:.3%}**'.format(accuracy))

*# Perform k-fold cross-validation with 10 folds*

kf = KFold(data.shape[0], n\_folds=10)

error = []

**for** train, test **in** kf:

*# Filter training data*

train\_predictors = (data[predictors].iloc[train,:])

*# The target we're using to train the algorithm.*

train\_target = data[outcome].iloc[train]

*# Training the algorithm using the predictors and target.*

model.fit(train\_predictors, train\_target)

*# Record error from each cross-validation run*

error.append(model.score(data[predictors].iloc[test,:], data[outcome].iloc[test]))

print('Cross-Validation Score : **%s**' % '**{0:.3%}**'.format(np.mean(error)))

*# Fit the model again so that it can be refered outside the function:*

model.fit(data[predictors],data[outcome])

Regression Model

*# Logistic regression*

outcome\_var = 'Rape'

model = LogisticRegression()

predictor\_var = ['STATE\_UT', 'DISTRICT', 'Year', 'Kidnapping and Abduction',

'Dowry', 'Assault', 'Insult to modesty of Women',

'Cruelty by Husband or his Relatives', 'Importation of Girls']

classification\_model(model, no\_total\_F,predictor\_var,outcome\_var)

Random forest

model = RandomForestClassifier(n\_estimators=100)

classification\_model(model, no\_total\_F,predictor\_var,outcome\_var)

# CONCLUSION

In summary, the application aims at providing a healthy picture of the crime trend using descriptive and forecasting techniques which would make the society healthier in terms of better visualized information and help law enforcing agencies to understand facts and figures and take necessary steps.