CHAPTER 1

INTRODUCTION

In recent years, Big Data Analytics (BDA) has become an emerging approach for analyzing data and extracting information and their relations in a wide range of application areas. Due to continuous urbanization and growing populations, cities play important central roles in our society. However, such developments have also been accompanied byan increase in violent crimes and accidents. To tackle such problems, sociologists, analysts, and safety institutions have devoted much effort towards mining potential patterns and factors. In relation to public policy however, there are many challenges in dealing with large amounts of available data. As a result, new methods and technologies need to be devised in order to analyze this heterogeneous and multi-sourced data. Analysis of such big data enables us to effectively keep track of occurred events, identify similarities from incidents, deploy resources and make quick decisions accordingly. This can also help further our understanding of both historical issues and current situations, ultimately ensuring improved safety/security and quality of life, as well as increased cultural and economic growth.

The rapid growth of cloud computing and data acquisition and storage technologies, from business and research institutions to governments and various organizations, have led to a huge number of unprecedented scopes/complexities from data that has been collected and made publicly available. It has become increasingly important to extract meaningful information and achieve new insights for understanding patterns from such data resources. BDA can effectively address the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods As a fast growing and influential practice, DBA can aid organizations to utilize their data and facilitate new opportunities. Furthermore, BDA can be deployed to help intelligent businesses move ahead with more effective operations, high profits and satisfied customers.

CHAPTER 2

LITERATURE SURVEY

1.Title: BIG DATA ANALYTICS. Issues in Information Systems

Authors: Zakir J, Seymour T, et al

Abstract: Today Big Data draws a lot of attention in the IT world. The rapid rise of the Internet and the digital economy has fuelled an exponential growth in demand for data storage and analytics, and IT department are facing tremendous challenge in protecting and analyzing these increased volumes of information. The reason organizations are collecting and storing more data than ever before is because their business depends on it. The type of information being created is no more traditional database-driven data referred to as structured data rather it is data that include documents, images, audio, video, and social media contents known as unstructured data or Big Data. Big Data Analytics is a way of extracting value from these huge volumes of information, and it drives new market opportunities and maximizes customer retention. This paper primarily focuses on discussing the various technologies that work together as a Big Data Analytics system that can help predict future volumes, gain insights, take proactive actions, and give way to better strategic decision-making. Further this paper analyzes the adoption, usage and impact of big data analytics to the business value of an enterprise to improve its competitive advantage using a set of data algorithms for large data sets such as Hadoop and MapReduce.

Advantages:

□ To better strategic decision-making in big data analytics.

□ To improve its competitive advantage of an enterprise.

Disadvantages:

□ It requires huge data storage.

□ It requires huge computational power.

2.Title: An integrated big data analytics-enabled transformation model: Application to health care

Authors: Wang Y, Kung L A, et al.

Abstract: A big data analytics-enabled transformation model based on practice-based view is developed, which reveals the causal relationships among big data analytics capabilities, IT-enabled transformation practices, benefit dimensions, and business values. This model was then tested in healthcare setting. By analyzing big data implementation cases, we sought to understand how big data analytics capabilities transform organizational practices, thereby generating potential benefits. In addition to conceptually defining four big data analytics capabilities, the model offers a strategic view of big data analytics. Three significant path-to-value chains were identified for healthcare organizations by applying the model, which provides practical insights for managers.

Advantages:

□ To provide practical insights for managers.

□ To improve its advantage of health care.

Disadvantages:

□ It accuracy is very less.

□ It requires huge computational power.

3.Title: A survey of data mining techniques for analyzing crime patterns

Authors: Thongsatapornwatana U.

Abstract: In recent years the data mining is data analyzing techniques that used to analyze crime data previously stored from various sources to find patterns and trends in crimes. In additional, it can be applied to increase efficiency in solving the crimes faster and also can be applied to automatically notify the crimes. However, there are many data mining techniques. In order to increase efficiency of crime detection, it is necessary to select the data mining techniques suitably. This paper reviews the literatures on various data mining applications, especially applications that applied to solve the crimes. Survey also throws light on research gaps and challenges of crime data mining. In additional to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining.

Advantages:

- To find the patterns and trends in crime.
- To increase efficiency of crime detection.

Disadvantages:

- It accuracy is very less.
- It can't consider the all features to detect crime .

4.Title: Big data analytics in healthcare: promise and potential
Authors: Raghupathi W, Raghupathi V.

Abstract: To describe the promise and potential of big data analytics in healthcare. The paper describes the nascent field of big data analytics in healthcare, discusses the benefits, outlines an architectural framework and methodology, describes examples reported in the literature, briefly discusses the challenges, and offers conclusions. The paper provides a broad overview of big data analytics for healthcare researchers and practitioners. Big data analytics in healthcare is evolving into a promising field for providing insight from very large data sets and improving outcomes while reducing costs. Its potential is great; however there remain challenges to overcome.

Advantages:

- To improve outcomes while reducing costs.
- To increase efficiency of analytics in healthcare.

Disadvantages:

- It can't process the large amount of data.
- It can't consider the all features to analysis large data.

5.Title: A survey of big data analytics in healthcare and government

Authors: Archenaa J, Anita E A M.

Abstract: This paper gives an insight of how we can uncover additional value from the data generated by healthcare and government. Large amount of heterogeneous data is generated by these agencies. But without proper data analytics methods these data became useless. Big Data Analytics using Hadoop plays an effective role in performing meaningful real-time analysis on the huge volume of data and able to predict the emergency situations before it happens. It describes about the big data use cases in healthcare and government.

Advantages:

- To improve outcomes while reducing costs.
- To increase efficiency of analytics in healthcare.

Disadvantages:

- It requires huge data storage.
- It requires huge computational power.

CHAPTER 3

PROPOSED SYSTEM

BDA can effectively address the challenges of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods. Analyzing large amount of data identification of trends accuracy is less.

3.1 EXISTING SYSTEM

Consequently, BDA becomes increasingly crucial to organizations to address their developmental issues . As one of the fundamental techniques of BDA, data mining is an innovative, interdisciplinary, and growing research area, which can build paradigms and techniques across various fields for deducing useful information and hidden patterns from data. Data mining is useful in not only the discovery of new knowledge or phenomena but also for enhancing our understanding of known ones. With the support of such techniques, BDA can help us easily identify crime patterns which occur in a particular area and how they are related with time. The implications of machine learning and statistical techniques on crime or other big data applications such as traffic accidents or time series data, will enable the analysis, extraction and understanding of associated patterns and trends, ultimately assisting in crime prevention and management.

3.2 LIMITATIONS OF EXISTING SYSTEM

- It difficult to extract the trends accurately.
- Moreover, some unreasonable features may reduce the accuracy of detection.
- Huge complexity of predicting crime data.

3.3 PROPOSED SYSTEM

In this project, state-of-the-art machine learning and big data analytics algorithms are utilized for the mining of crime data from three Indian cities. After preprocessing, including data filtering and normalization, Google maps based geo mapping of the features are implemented for visualization of the statistical results. Various approaches in machine learning, deep learning, and time series modelling are utilized for future trends analysis. The major contribution of this project can be summarized as follows:

1) A series of investigative explorations are conducted to explore and explain the crime data in three Indian cities;

2) We propose a novel visual representation which is capable of handling large datasets and enables users to explore, compare, and analyze evolutionary trends and patterns of crime incidents;

3) A combination and comparison of different machine learning, deep learning and time series modeling algorithms to predict trends with the optimal parameters, time periods and models.

3.4 ADVANTAGES OF PROPOSED SYSTEM

- Efficient prediction of crime trends and forecasting.
- It will support large amount of data.
- It takes less computational time.

CHAPTER 4

. REQUIREMENT ANALYSIS AND FEASIBILITY STUDY

4.1 FUNCTIONAL REQUIREMENTS:

This section describes the functional requirements of the system for those requirements which are expressed in the natural language style.

1. Create an desktop application.
2. Load the crime dataset.
3. System should preprocess and extract the deep features from the dataset.
4. Applying the LSTM with ARIMA model to train and forecast crime trends.
5. Application should provides high accuracy of forecasting the crime trends.

4.2 NON FUNCTIONALITY REQUIREMENTS

These are requirements that are not functional in nature, that is, these are constraints within which the system must work.

- The program must be self-contained so that it can easily be moved from one Computer to another. It is assumed that network connection will be available on the computer on which the program resides.

- **Capacity, scalability and availability**.

The system shall achieve 100 per cent availability at all times. The system shall be scalable to support additional clients and volunteers.

- **Maintainability.**

The system should be optimized for supportability, or ease of maintenance as far as possible. This may be achieved through the use documentation of coding standards, naming conventions, class libraries and abstraction.

- **Randomness, verifiability and load balancing.**

The system should be optimized for supportability, or ease of maintenance as far as possible. This may be achieved through the use documentation of coding standards, naming conventions, class libraries and abstraction. It should have randomness to check the nodes and should be load balanced.

4.3 HARDWARE REQUIREMENTS

□ System	:	Pentium i3.
□ Hard Disk	:	120 GB.
□ Monitor	:	15” LED
□ Input Devices	:	Keyboard, Mouse
□ Ram	:	4 GB

4.4 SOFTWARE REQUIREMENTS

□ Operating system	:	Windows 7.
□ Coding Language	:	Python3.6
□ Tools	:	Python IDLE

4.5 FEASIBILITY STUDY

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

Three key considerations involved in the feasibility analysis are

- ♦ ECONOMICAL FEASIBILITY
- ♦ TECHNICAL FEASIBILITY
- ♦ SOCIAL FEASIBILITY

4.5.1 ECONOMICAL FEASIBILITY

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

4.5.2 TECHNICAL FEASIBILITY

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

4.5.3 SOCIAL FEASIBILITY

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

5. System Design:

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

CHAPTER 5

SYSTEM DESIGN

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development.

5.1 System Architecture:

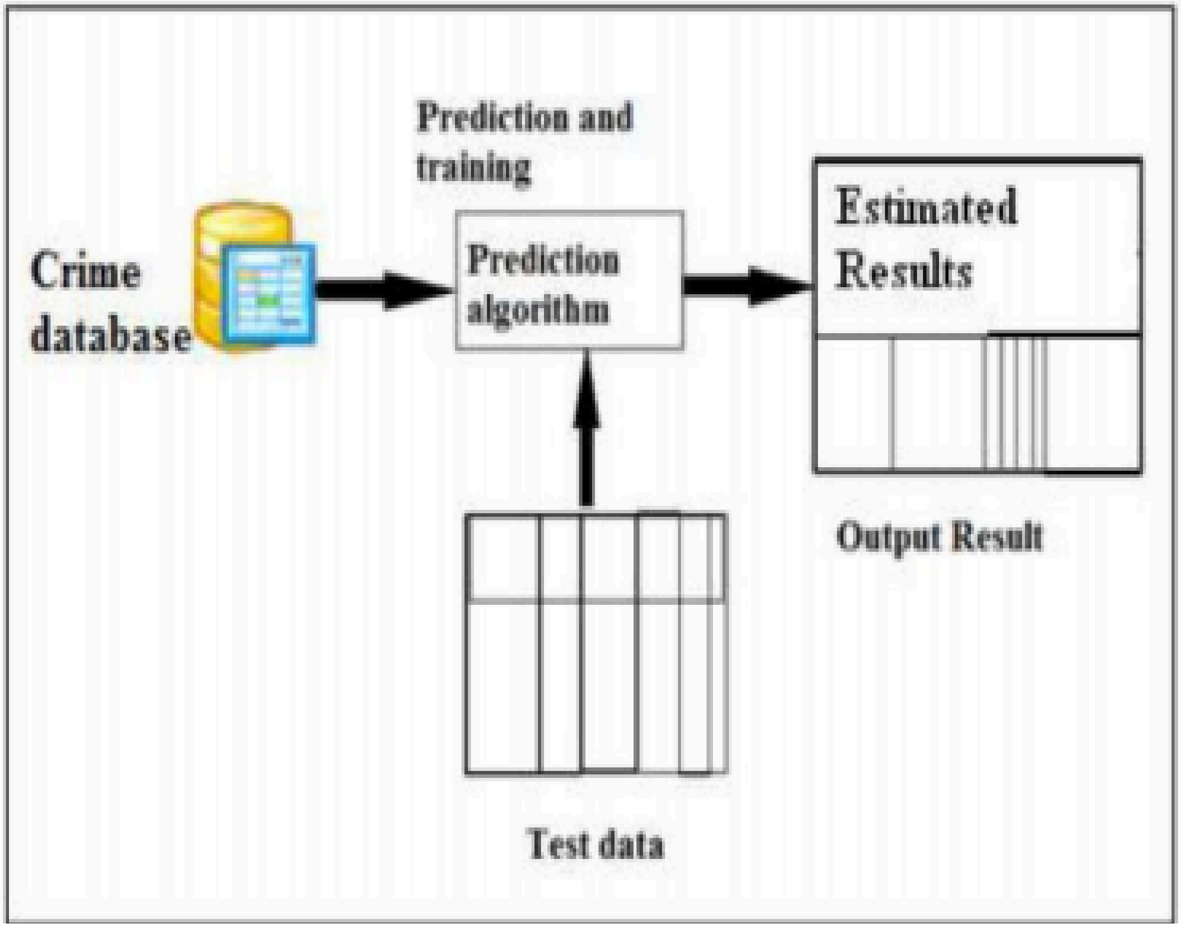
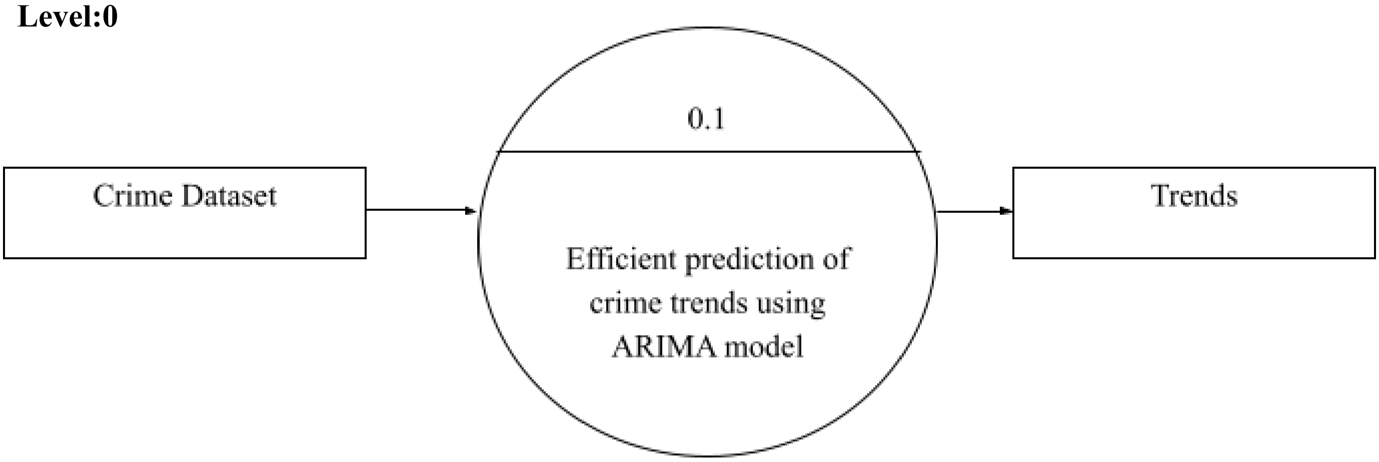


Fig 5.1:System Architecture

The framework of our proposed approach as shown in above system architecture. The first is the prediction algorithm module, which contains data preprocessing, feature extraction and classification. The data consists of a large number of crime data. The system will split into testing and training and classify the crime trends.

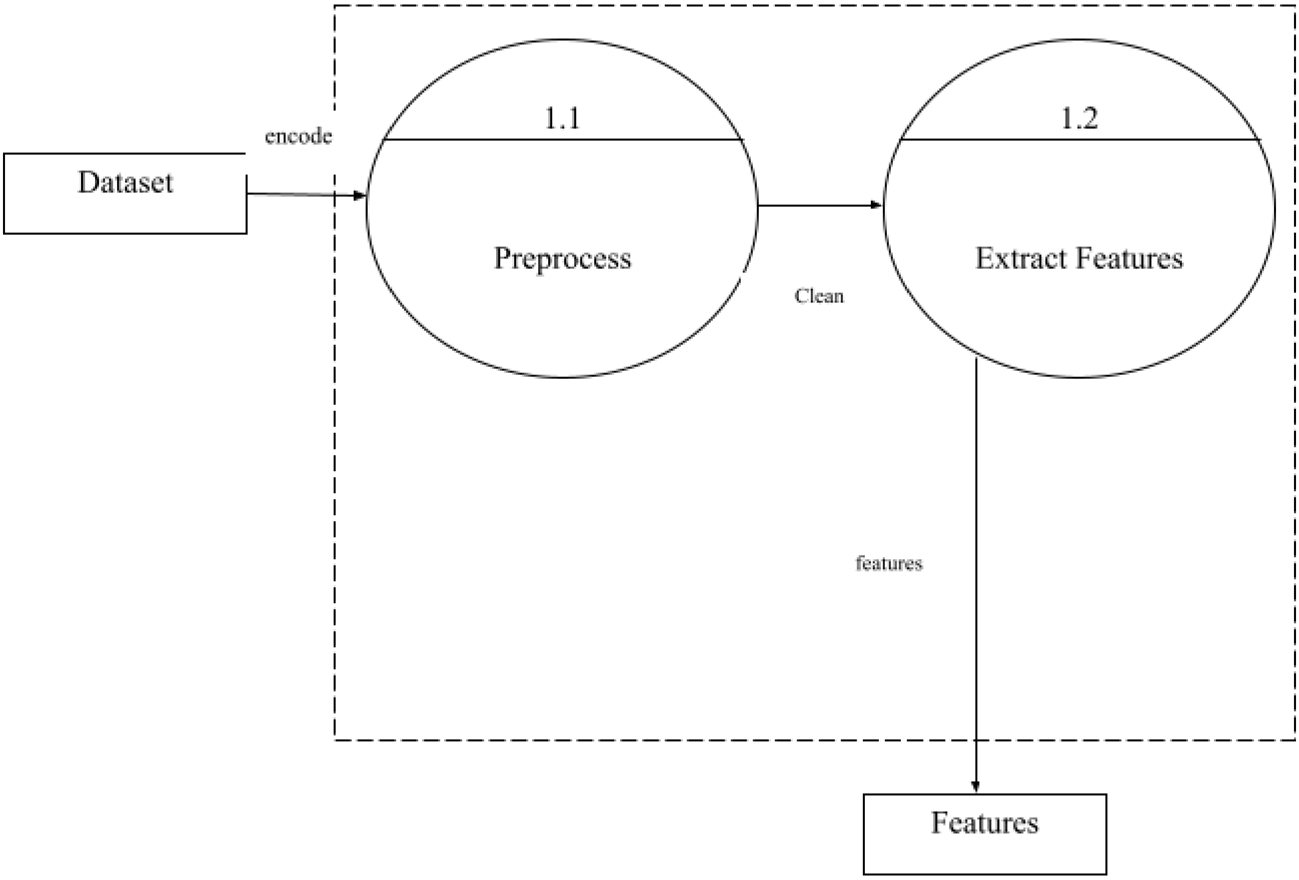
5.2 Data Flow Diagram:

Level:0



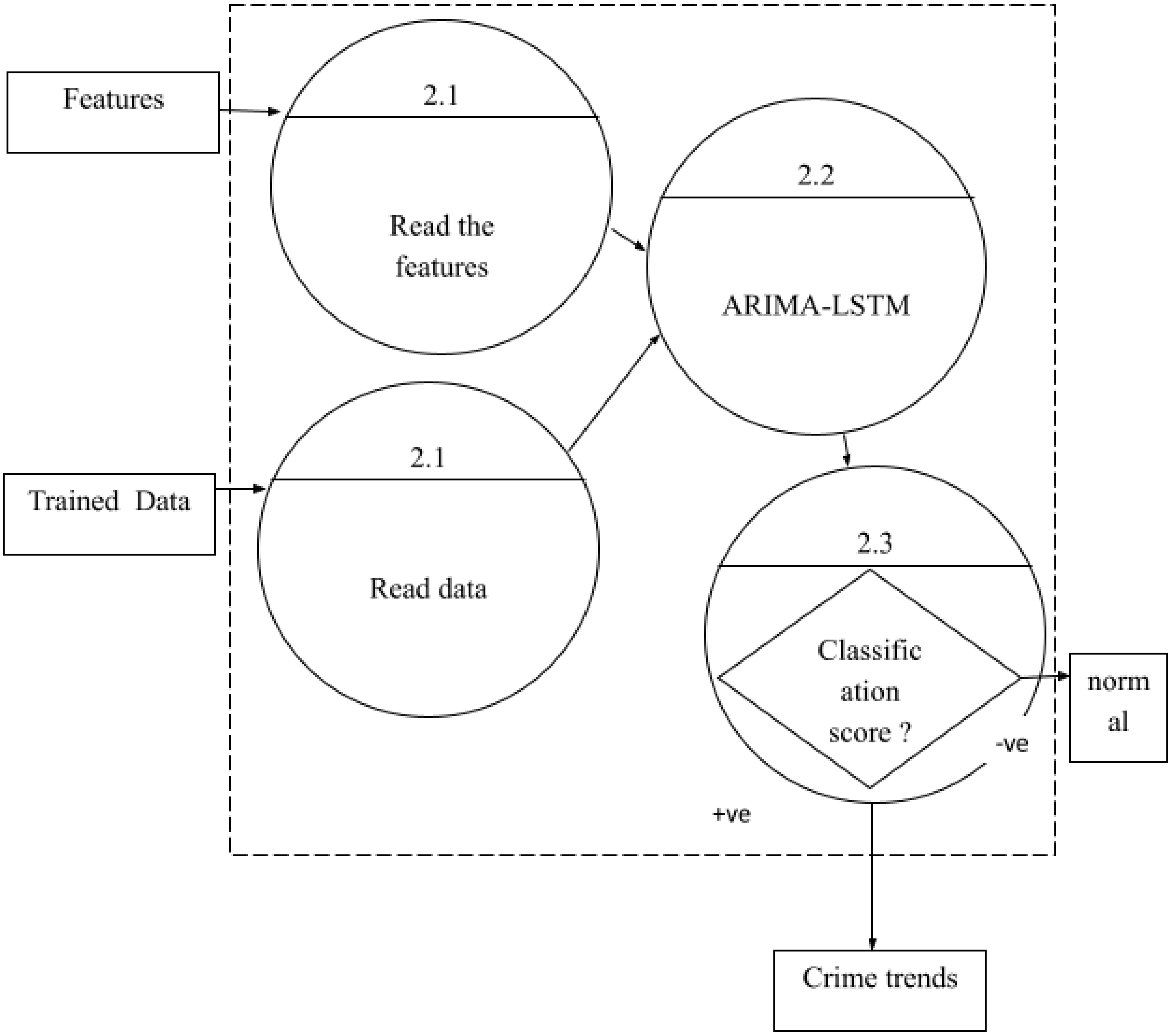
Level 0 Describes the overall process of this project. we are passing crime dataset as a input the system will predict crime trends using ARIMA model.

Level 1



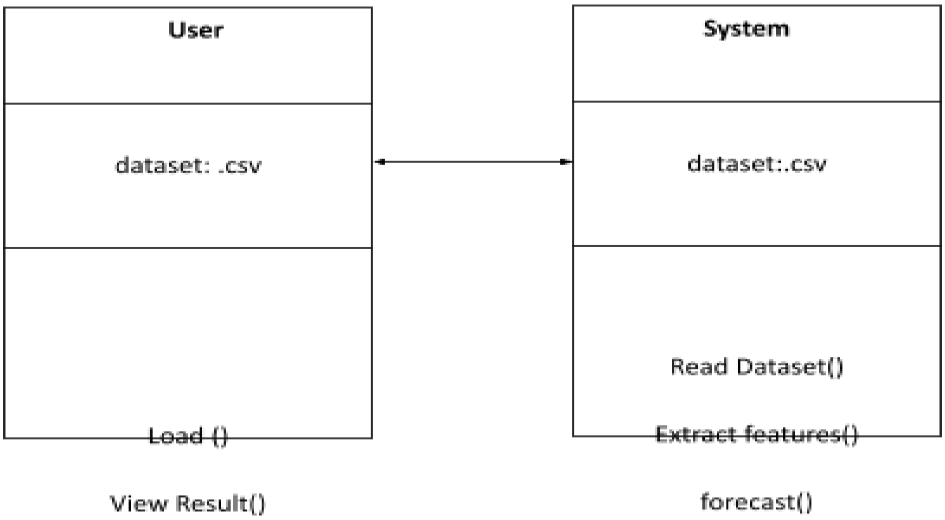
Level 1 Describes the first stage process of this project. we are passing crime dataset as a input the system will perform the preprocess and extract the important features.

Level 2:

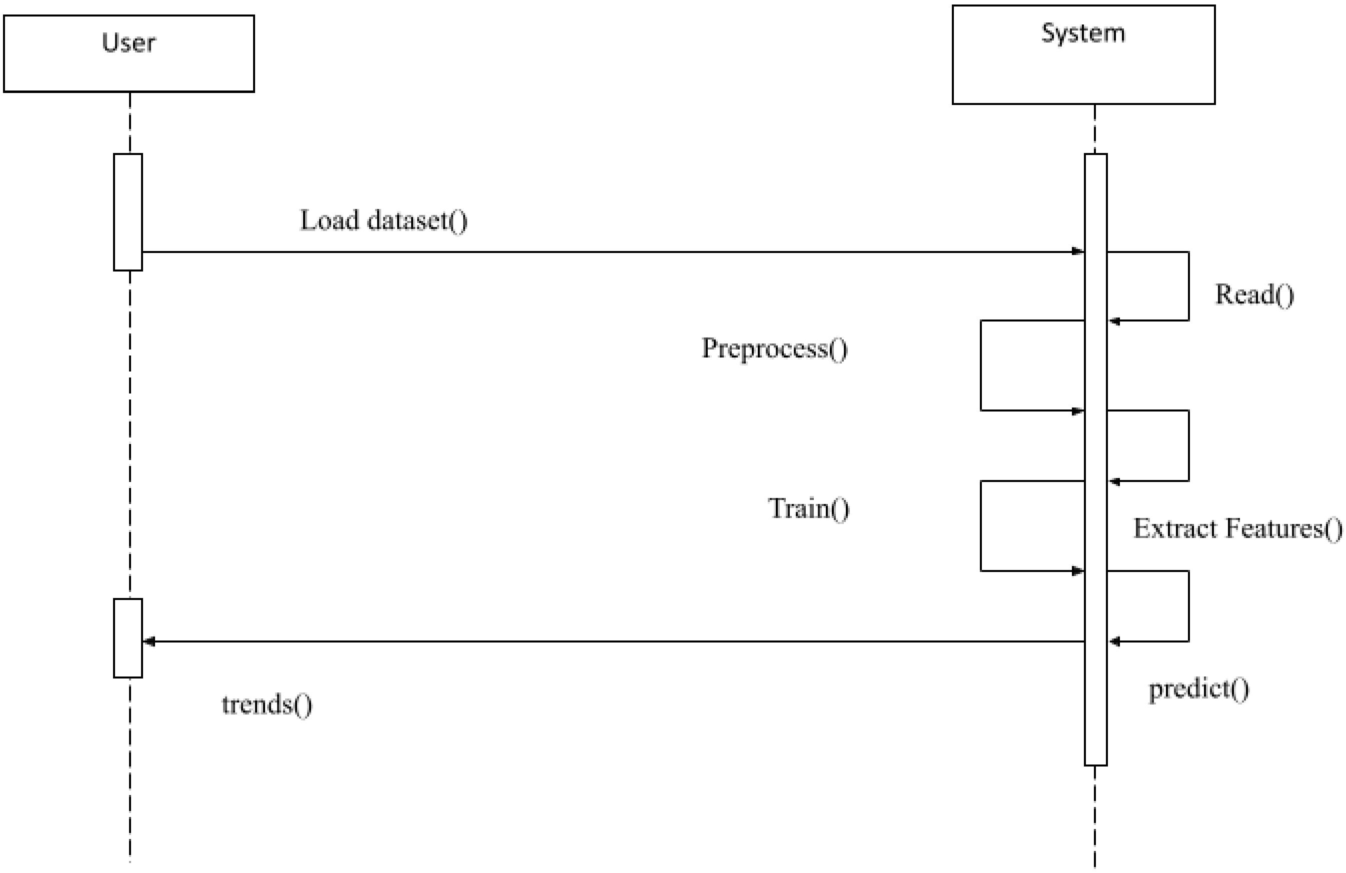


Level 2 Describes the final stage process of this project. we are passing extracted features from level 1and trained data as a input the system will predict the trends on crime data using ARIMA model.

5.3 Class Diagram:



5.4 Sequence Diagram



The sequence diagram will determine the users states (active/ inactive) with our applications.

CHAPTER 6

IMPLEMENTATION

6.1 Algorithm:

6.1.1 LSTM

LSTM model is a powerful type of recurrent neural network (RNN), capable of learning long-term dependencies. For time series involves autocorrelation, i.e. the presence of correlation between the time series and lagged versions of itself, LSTMs are particular useful in prediction due to their capability of maintaining the state whilst recognizing patterns over the time series. The recurrent architecture enables the states to be persisted, or communicate between updated weights as each epoch progresses. Moreover, the LSTM cell architecture an enhance the RNN by enabling long term persistence in addition to short term

$$f(t) = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Where, ft is a sigmoid function to indicate whether to keep the previous state, C_{t-1} is the old cell state, Ct is the updated cell state, W_f, Wi, and WC are the previous value in each layer, h_{t-1} and x_t, is the input value, b_f, b_i,and b_c are constant values, it decides which value will be used to update the state, C_t stands for the new candidate values.

6.1.2 ARIMA forecast model:

ARIMA is an acronym that stands for Auto Regressive Integrated Moving Average. It is a class of model that captures a suite of different standard temporal structures in time series data.

6.1.2.1 Components of ARIMA:

- ARIMA has three components – AR (autoregressive term), I (differencing term) and MA (moving average term). Let us understand each of these components
- AR term refers to the past values used for forecasting the next value. The AR term is defined by the parameter ‘p’ in arima.
- MA term is used to defines number of past forecast errors used to predict the future values. The parameter ‘q’ in arima represents the MA term. ACF plot is used to identify the correct ‘q’ value.
- Order of differencing specifies the number of times the differencing operation is performed on series to make it stationary. Test like ADF and KPSS can be used to determine whether the series is stationary and help in identifying the d value.

6.1.2.2 Steps of ARIMA

- Load the data: This step will be the same. Load the data into your notebook
- Preprocessing data: The input should be univariate, hence drop the other columns
- Fit Auto ARIMA: Fit the model on the univariate series
- Predict values on validation set: Make predictions on the validation set
- Calculate RMSE: Check the performance of the model using the predicted values against the actual values.

6.1.2.3 ARIMA Flow Chart:

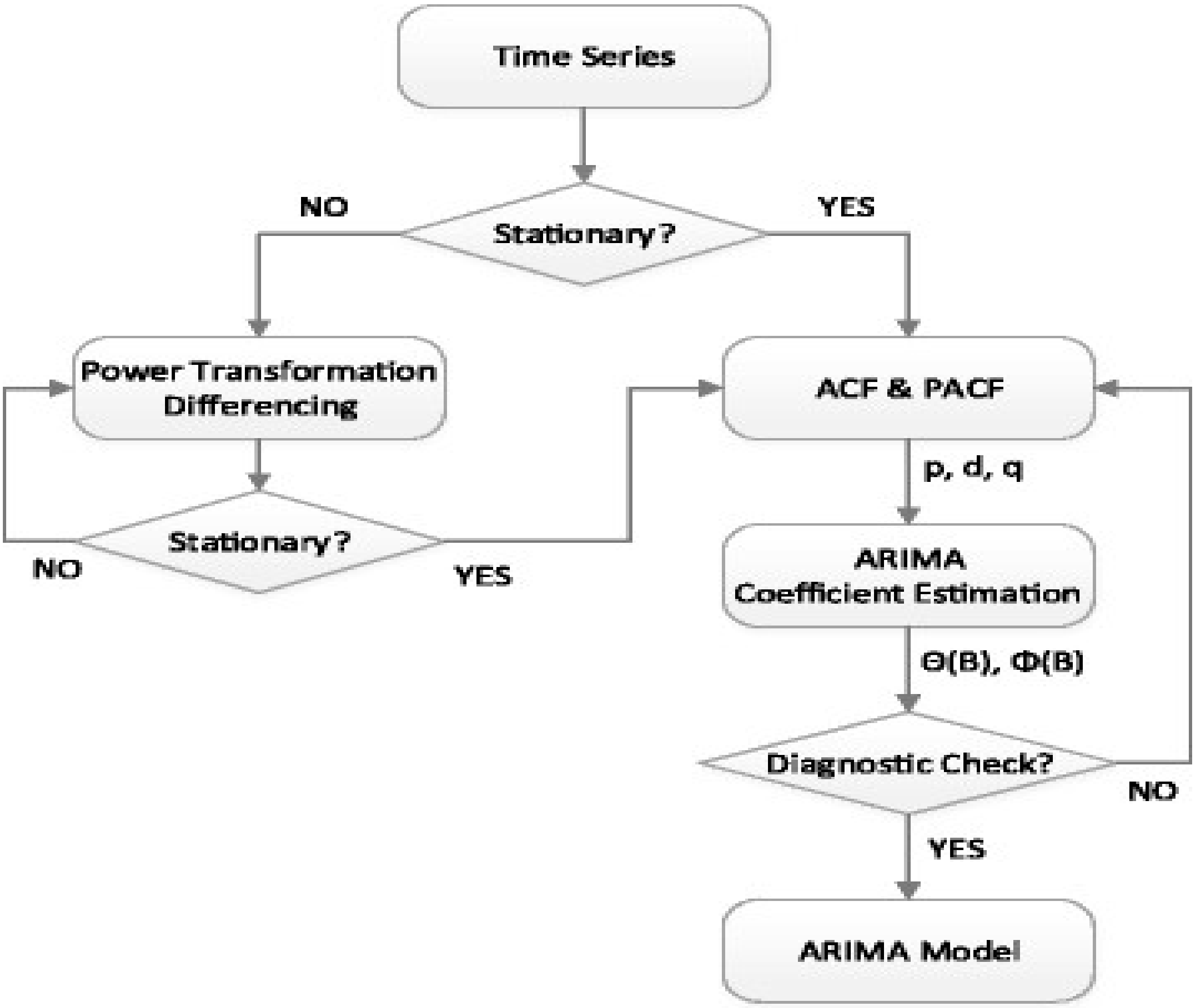


Fig 6.1.2.3: ARIMA Flow Chart

6.1.3 Random Forest

Random Forest algorithm is a supervised classification algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.

There are two stages in Random Forest algorithm, one is random forest creation, the other is to make a prediction from the random forest classifier created in the first stage. The whole process is shown below, and it's easy to understand using the figure.

Here the author firstly shows the Random Forest creation pseudocode:

1. Randomly select "**K**" features from total "**m**" features where **k << m**
2. Among the "**K**" features, calculate the node "**d**" using the best split point
3. Split the node into **daughter nodes** using the **best split**
4. Repeat the **a to c** steps until "l" number of nodes has been reached
5. Build forest by repeating steps **a to d** for "n" number times to create **"n" number of trees**

6.1.4 Decision Tree Algorithm

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data).

In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

Important Terminology related to Decision Trees

1. **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
2. **Splitting:** It is a process of dividing a node into two or more sub-nodes.
3. **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
4. **Leaf / Terminal Node:** Nodes do not split is called Leaf or Terminal node.
5. **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.

6. **Branch / Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.
7. **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

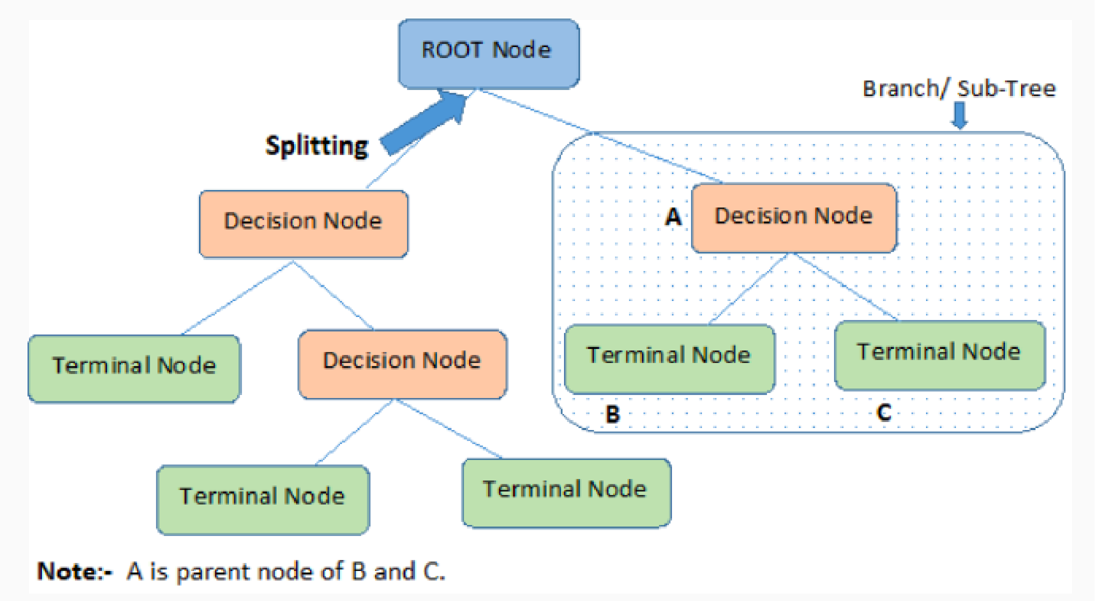


Fig 6.1.4:Decision Tree

Decision trees classify the examples by sorting them down the tree from the root to some leaf/terminal node, with the leaf/terminal node providing the classification of the example. Each node in the tree acts as a test case for some attribute, and each edge descending from the node corresponds to the possible answers to the test case. This process is recursive in nature and is repeated for every subtree rooted at the new node.

6.2 Configuration:

Our project is configured with the following modules.

6.2.1 Modules:

- Featured Attributes
- Data Preprocessing
- Narrative Visualization
- Prediction Models

6.2.2 Module Descriptions:

6.2.2.1 Featured Attributes:

For each entry of crime incidents in the datasets, the following 13 featured attributes are included:

1) Incident Num- Case number of each incident;

2) Dates-Date and timestamp of the crime incident;

3) Category - Type of the crime.

4) Descript - A brief note describing any pertinent details of the crime;

5) DayOfWeek- Day of the week that crime occurred;

6) PdDistrict - Police Department District ID where the crime is assigned;

7) Resolution - How the crime incident was resolved (with the perpetrator being, say, arrest or booked);

8) Address - The approximate street address of the crime incident;

9) X-Longitude of the location of a crime;

10)Y-Latitude of the location of a crime;

11) Coordinate- Pairs of Longitude and Latitude;

12) Dome- whether crime id domestic or not;

13) Arrest-Arrested or not

6.2.2.2 Data Pre Processing

Before implementing any algorithms on our datasets, a series of preprocessing steps are performed for data conditioning as presented below:

1) Time is discretized into a couple of columns to allow for time series forecasting for the overall trend within the data.

2) For some missing coordinate attributes in Chicago and Philadelphia datasets, we imputed random values sampled from the non-missing values, computed their mean, and then replaced the missing ones.

3) The timestamp indicates the date and time of occurrence of each crime, we deduced these attributes into five features: Year (2003-2017), Month (1-12), Day (1-31), Hour(0-23),and Minute(0-59).

4) We also omit some features that unneeded like incident Num, coordinate.

6.2.2.3 Narrative Visualization

Considering the geographic nature of the crime incidents, an interactive map based on Google map was used for data visualization, where crime incidents are clustered according to their latitude/longitude information.

6.2.2.4 Prediction Modules

In order to tackle the problem of crime trends forecasting we explored several state-of-the-art machine learning and deep learning algorithms and time series models. A time series is a sequence of numerical data points successively indexed or listed/graphed in the time order. Usually, the successive data points within a time series are equally spaced in time, hence these data are discrete in time.

LSTM model is a powerful type of recurrent neural network (RNN), capable of learning long-term dependencies. For time series involves autocorrelation, i.e. the presence of correlation between the time series and lagged versions of itself, LSTMs are particular useful in prediction due to their capability of maintaining the state whilst recognizing patterns over the time series. The recurrent architecture enables the states to be persisted, or communicate between updated weights as each epoch progresses. Moreover, the LSTM cell architecture can enhance the RNN by enabling long term persistence in addition to short term.

6.3 CODE

```
import tkinter as tk
from tkinter import Message ,Text
import shutil
import csv
import numpy as np
from PIL import Image, ImageTk
import pandas as pd
import datetime
import time
import tkinter.font as font
from tkinter import filedialog
import tkinter.messagebox as tm
import preprocess as pr

import ARIMAALG as AR
import DTALG as DT
import RFALG as RF
import LSTMALG as LST
from tkinter import ttk

from_date = datetime.datetime.today()
currentDate = time.strftime("%d_%m_%y")

fontScale=1
fontColor=(0,0,0)
cond=0

bgcolor="#FF5733"
fgcolor="white"

window = tk.Tk()
window.title("Crime Data Analysis and Prediction")


window.geometry('1280x720')
window.configure(background=bgcolor)
#window.attributes('-fullscreen', True)

window.grid_rowconfigure(0, weight=1)
window.grid_columnconfigure(0, weight=1)
crim=['BATTERY','OTHER OFFENSE','ROBBERY','NARCOTICS','CRIMINAL DAMAGE','WEAPONS
VIOLATION','THEFT','BURGLARY','MOTOR VEHICLE THEFT','PUBLIC PEACE
VIOLATION','ASSAULT','CRIMINAL TRESPASS','CRIM SEXUAL ASSAULT','INTERFERENCE WITH
PUBLIC OFFICER','ARSON','DECEPTIVE PRACTICE','LIQUOR LAW
VIOLATION','KIDNAPPING','SEX OFFENSE','OFFENSE INVOLVING
CHILDREN','PROSTITUTION','GAMBLING','INTIMIDATION','STALKING','OBSCENITY','PUBLIC
INDECENCY','HUMAN TRAFFICKING','CONCEALED CARRY LICENSE VIOLATION','OTHER NARCOTIC
VIOLATION','HOMICIDE','NON-CRIMINAL']

message1 = tk.Label(window, text="Crime Data Analysis and Prediction" ,bg=bgcolor
,fg=fgcolor  ,width=50  ,height=3,font=('times', 30, 'italic bold underline'))
message1.place(x=100, y=10)

lbl = tk.Label(window, text="Select Dataset",width=20  ,height=2  ,fg=fgcolor
,bg=bgcolor ,font=('times', 15, ' bold ') )
lbl.place(x=10, y=200)

txt = tk.Entry(window,width=20,bg="white" ,fg="black",font=('times', 15, ' bold
'))
txt.place(x=300, y=215)


lbl1 = tk.Label(window, text="Latitude",width=20  ,height=2  ,fg=fgcolor
,bg=bgcolor ,font=('times', 15, ' bold ') )
lbl1.place(x=10, y=300)

lat = tk.Entry(window,width=20,bg="white" ,fg="black",font=('times', 15, ' bold
'))
lat.place(x=300, y=315)

lbl1 = tk.Label(window, text="Longitude",width=20  ,height=2  ,fg=fgcolor
,bg=bgcolor ,font=('times', 15, ' bold ') )
lbl1.place(x=500, y=300)

lon = tk.Entry(window,width=20,bg="white" ,fg="black",font=('times', 15, ' bold
'))
lon.place(x=750, y=315)


lbl1 = tk.Label(window, text="DATE AND TIME",width=20  ,height=2  ,fg=fgcolor
,bg=bgcolor ,font=('times', 15, ' bold ') )
lbl1.place(x=10, y=400)


txt2 = ttk.Combobox(window,width=10,font=('times', 15, ' bold '),values=["2015"])
txt2.place(x=300, y=415)
txt2.current(0)

txt3 = ttk.Combobox(window,width=10,font=('times', 15, ' bold
'),values=["01","02","03","04","05","06","07","08","09","10","11","12"])
txt3.place(x=430, y=415)
txt3.current(0)

txt4 = ttk.Combobox(window,width=10,font=('times', 15, ' bold
'),values=["01","02","03","04","05","06","07","08","09","10","11","12","13","14","
15","16","17","18","19","20","21","22","23","24","25","26","27","28","29","30","31
"])
txt4.place(x=560, y=415)
txt4.current(0)

txt5 = ttk.Combobox(window,width=10,font=('times', 15, ' bold
'),values=["01","02","03","04","05","06","07","08","09","10","11","12","13","14","
15","16","17","18","19","20","21","22","23","24"])
txt5.place(x=690, y=415)
txt5.current(0)

txt6 = ttk.Combobox(window,width=10,font=('times', 15, ' bold
'),values=["00","01","02","03","04","05","06","07","08","09","10","11","12","13","
14","15","16","17","18","19","20","21","22","23","24","25","26","27","28","29","30
","31","32","33","34","35","36","37","38","39","40","41","42","43","44","45","46",
"47","48","49","50","51","52","53","54","55","56","57","58","59"])
txt6.place(x=820, y=415)
txt6.current(0)

lbl4 = tk.Label(window, text="Notification : ",width=20  ,fg=fgcolor,bg=bgcolor
,height=2 ,font=('times', 15, ' bold underline '))
lbl4.place(x=10, y=500)

message = tk.Label(window, text="" ,bg="white"  ,fg="black",width=30  ,height=2,
activebackground = bgcolor ,font=('times', 15, ' bold '))
message.place(x=300, y=500)

def clear():
   txt.delete(0, 'end')
   lat.delete(0, 'end')
   lon.delete(0, 'end')
   res = ""
   message.configure(text= res)

def browse():
   path=filedialog.askopenfilename()
   print(path)
   txt.insert('end',path)
   if path !="":
      print(path)
   else:
      tm.showinfo("Input error", "Select Dataset")

def preprocess():
   sym=txt.get()
   if sym != "" :
      pr.process(sym)
      res = "Preprocess Finished Successfully"
      message.configure(text= res)
      tm.showinfo("Input", "Preprocess Finished Successfully")
   else:
      tm.showinfo("Input error", "Select Dataset")


def arima():
   res=AR.process()
   print(res)
   res1=crim[int(res[0])]
   message.configure(text= res1)

   tm.showinfo("Input", "arima Fininshed Successfully")

def lstm():
   sym1=lat.get()
   sym2=lon.get()
   if sym1 != "" and sym2 !="":
      s=[]
      s.append(float(sym1))
      s.append(float(sym2))
      s.append(int(txt2.get()))
      s.append(int(txt3.get()))
      s.append(int(txt4.get()))
      s.append(int(txt5.get()))
      s.append(int(txt6.get()))
      print(s)
      res=LST.process(s)
      print(res[0])
      res=crim[int(res[0])]
      message.configure(text= res)
   else:
      tm.showinfo("Input error", "Enter Latitude,Longitude values")

def decisiontree():
   sym1=lat.get()
   sym2=lon.get()
   if sym1 != "" and sym2 !="":
      s=[]
      s.append(float(sym1))
      s.append(float(sym2))
      s.append(int(txt2.get()))
      s.append(int(txt3.get()))
      s.append(int(txt4.get()))
      s.append(int(txt5.get()))
      s.append(int(txt6.get()))
      print(s)
      res=DT.process(s)
      res=crim[int(res)]
      message.configure(text= res)
      tm.showinfo("Input", "DecisionTree Finished Successfully")
   else:
      tm.showinfo("Input error", "Enter Latitude,Longitude values")

def randomforest():
   sym1=lat.get()
   sym2=lon.get()
   if sym1 != "" and sym2 !="":
      s=[]
      s.append(float(sym1))
      s.append(float(sym2))
      s.append(int(txt2.get()))
      s.append(int(txt3.get()))
      s.append(int(txt4.get()))
      s.append(int(txt5.get()))
      s.append(int(txt6.get()))
      print(s)
      res=RF.process(s)
      res=crim[int(res)]
      message.configure(text= res)
      tm.showinfo("Input", "RandomForest Finished Successfully")
   else:
      tm.showinfo("Input error", "Enter Latitude,Longitude values")


clearButton = tk.Button(window, text="Clear", command=clear  ,fg=fgcolor
,bg=bgcolor  ,width=20  ,height=2 ,activebackground = "Red" ,font=('times', 15, '
bold '))
clearButton.place(x=960, y=200)

browse = tk.Button(window, text="Browse", command=browse  ,fg=fgcolor  ,bg=bgcolor
,width=15  ,height=1, activebackground = "Red" ,font=('times', 15, ' bold '))
browse.place(x=530, y=205)

pre = tk.Button(window, text="Preprocess", command=preprocess  ,fg=fgcolor
,bg=bgcolor  ,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, '
bold '))
pre.place(x=10, y=600)

texta = tk.Button(window, text="Decision Tree", command=decisiontree  ,fg=fgcolor
,bg=bgcolor  ,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, '
bold '))
texta.place(x=200, y=600)

texta1 = tk.Button(window, text="RandomForest", command=randomforest  ,fg=fgcolor
,bg=bgcolor  ,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, '
bold '))
texta1.place(x=400, y=600)


pred = tk.Button(window, text="ARIMA", command=arima  ,fg=fgcolor,bg=bgcolor
,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, ' bold '))
pred.place(x=600, y=600)

lst = tk.Button(window, text="LSTM", command=lstm  ,fg=fgcolor,bg=bgcolor
,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, ' bold '))
lst.place(x=800, y=600)

quitWindow = tk.Button(window, text="QUIT", command=window.destroy  ,fg=fgcolor
,bg=bgcolor  ,width=18  ,height=2, activebackground = "Red" ,font=('times', 15, '
bold '))
quitWindow.place(x=1000, y=600)

window.mainloop()
```

CHAPTER 7

TESTING

7.1 TESTING:

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

7.1.1 TYPES OF TESTS

7.1.1.1 UNIT TESTING

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

7.1.1.2 INTEGRATION TESTING

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

7.1.1.3 FUNCTIONAL TEST

Functional tests provide systematic demonstrations that functions tested are available as specified by the business and technical requirements, system documentation, and user manuals.

Functional testing is centered on the following items:

Valid Input : identified classes of valid input must be accepted.

Invalid Input : identified classes of invalid input must be rejected.

Functions : identified functions must be exercised.

Output : identified classes of application outputs must be exercised.

Systems/Procedures: interfacing systems or procedures must be invoked.

Organization and preparation of functional tests is focused on requirements, key functions, or special test cases. In addition, systematic coverage pertaining to identify Business process flows; data fields, predefined processes, and successive processes must be considered for testing. Before functional testing is complete, additional tests are identified and the effective value of current tests is determined.

7.1.1.4 SYSTEM TEST

System testing ensures that the entire integrated software system meets requirements. It tests a configuration to ensure known and predictable results. An example of system testing is the configuration oriented system integration test. System testing is based on process descriptions and flows, emphasizing pre-driven process links and integration points.

7.1.1.5 WHITE BOX TESTING

White Box Testing is a testing in which in which the software tester has knowledge of the inner workings, structure and language of the software, or at least its purpose. It is purpose. It is used to test areas that cannot be reached from a black box level.

7.1.1.6 BLACK BOX TESTING

Black Box Testing is testing the software without any knowledge of the inner workings, structure or language of the module being tested. Black box tests, as most other kinds of tests, must be written from a definitive source document, such as specification or requirements document, such as specification or requirements document. It is a testing in which the software under test is treated, as a black box .you cannot “see” into it. The test provides inputs and responds to outputs without considering how the software works.

Unit Testing Tables

Test Case#	UTC01
Test Name	File import format
Test Description	To test whether an excel file with comma delimited format is accepted or not
Input	A comma separated excel file with valid dataset
Expected Output	The file should be read by the program and few lines of the file should be displayed
Actual Output	The file is read and contents are displayed accordingly
Test Result	Success

Table 7.1.1.6: Unit Testing1

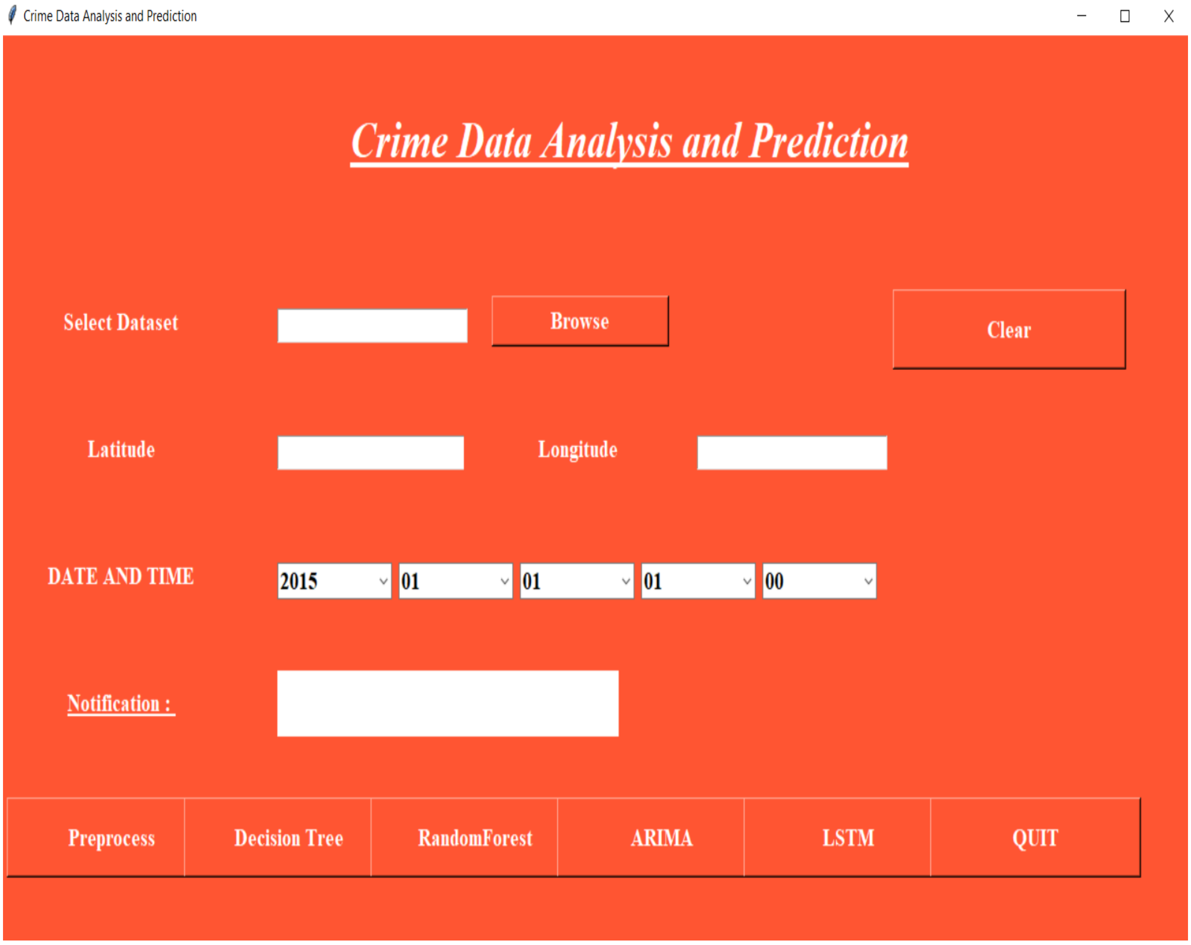
Test Case#	UTC02
Test Name	File import format
Test Description	To test whether an excel file with comma delimited format is accepted or not
Input	A text file with valid dataset
Expected Output	It Should show the alert Message Select Only CSV file
Actual Output	Shown alert message
Test Result	Success

Table 7.1.1.6: Unit Testing2

CHAPTER 8

RESULT ANALYSIS

8.1 SNAPSHOTS

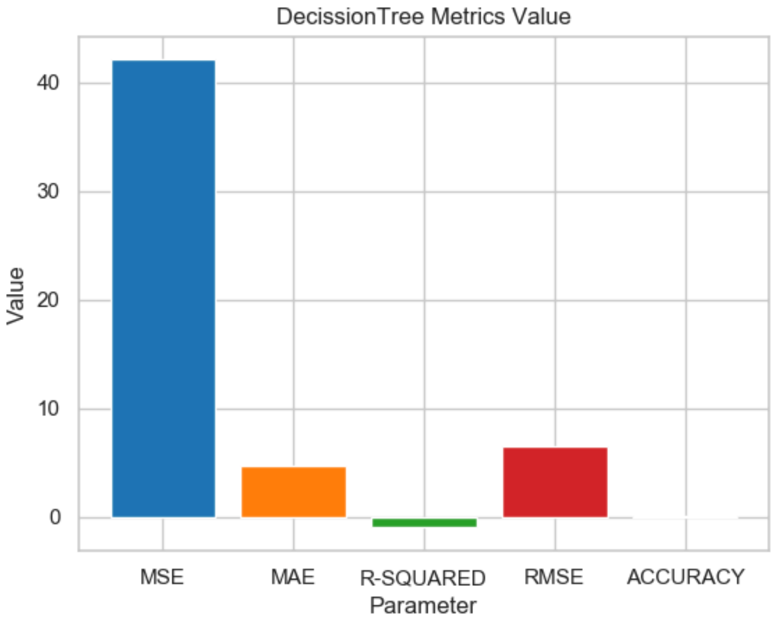


DATE AND TIME 2015 01 01 01 00

8.1.1 Home Page

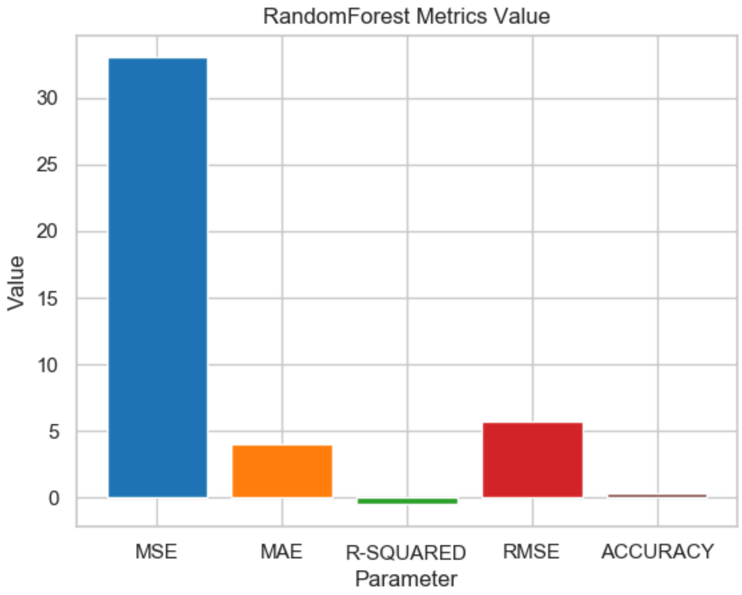
The figure 8.1.1 shows the front page.It consist of uploading of dataset, Latitude, Longitude, Preprocess, Date and time, DecisionTree, RandomForest, ARIMA, LSTM algorithm, and quit.

8.2 GRAPH



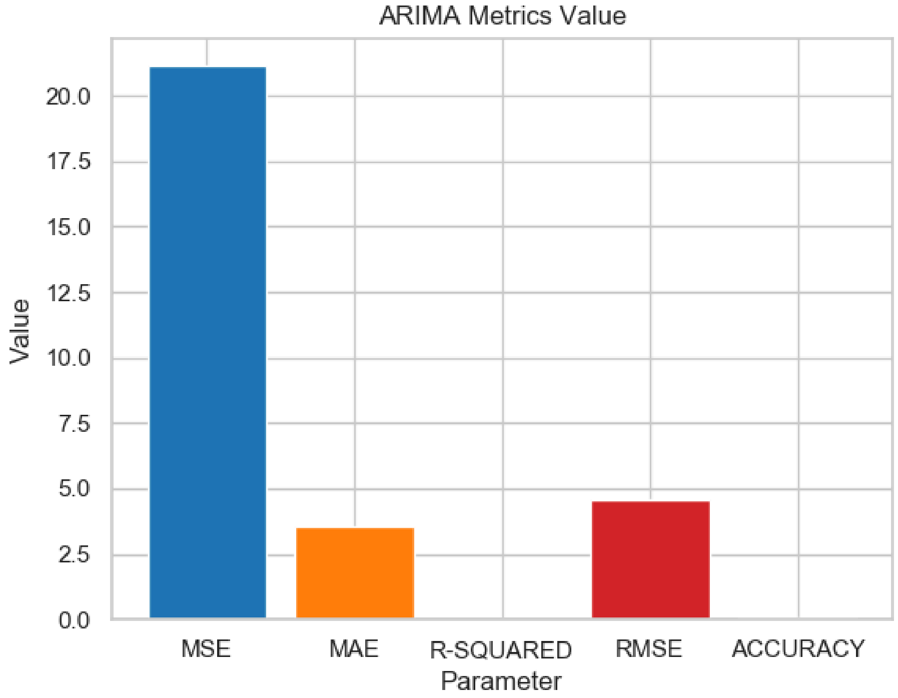
8.2.1 DecisionTree

The figure 8.2.1 It calculates the MSE, MAE, R-SQUARED Parameter, RMSE, and accuracy of the given dataset using DecisionTree algorithm. It consists of MSE (mean square error), MAE (mean absolute error), R^2 parameter, RMSE (root man square error), Accuracy



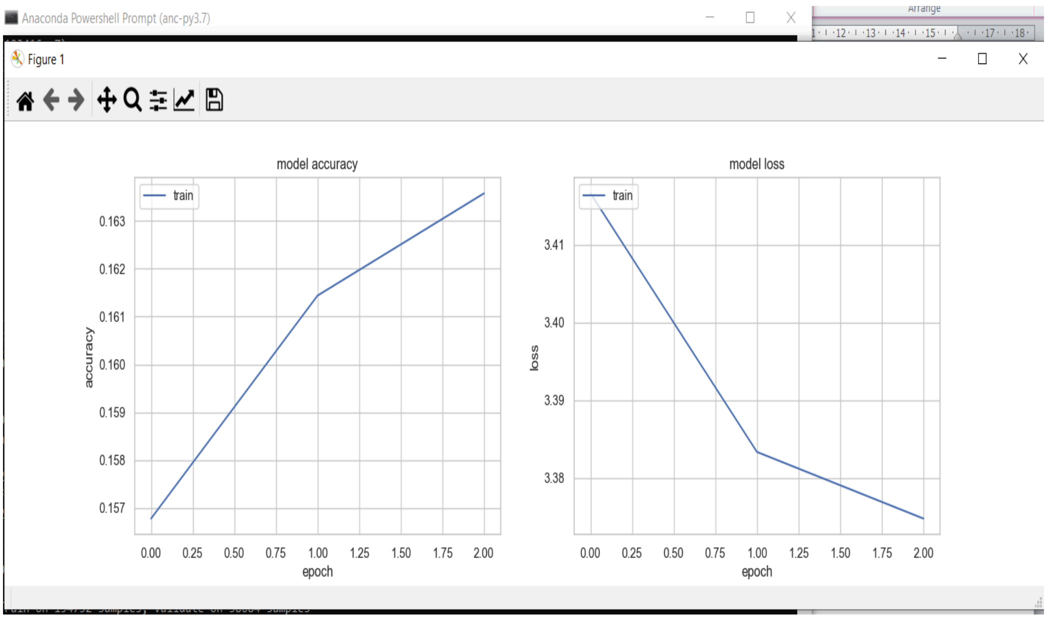
8.2.2 RandomForest

The figure 8.2.2 It calculates the MSE, MAE, R-SQUARED Parameter, RMSE, and accuracy of the given dataset using RandomForest algorithm



8.2.3 ARIMA

The figure 8.2.3 It calculates the MSE, MAE, R-SQUARED Parameter, RMSE, and accuracy of the given dataset using ARIMA algorithm



8.2.4 LSTM

This LSTM loss graph basically shows the model loss of the training that has been trained by the LSTM algorithm. It basically shows the model accuracy against the epoch.

CHAPTER 9

CONCLUSION AND FUTURE WORK

9.1 Conclusion

In this project a series of state-of-the-art big data analytics and visualization techniques were utilized to analyze crime big data from three Indian cities, which allowed us to identify patterns and obtain trends. By exploring the Prophet model, a neural network model, and the deep learning algorithm LSTM, we found that both the Prophet model and the LSTM algorithm perform better than conventional neural network models. We also found the optimal time period for the training sample to be 3 years, in order to achieve the best prediction of trends in terms of RMSE and spearman correlation. Optimal parameters for the Prophet and the LSTM models are also determined. Additional results explained earlier will provide new insights into crime trends and will assist both police departments and law enforcement agencies in their decision making.

9.2 Future Work

In future, we plan to complete our on-going platform for generic big data analytics which will be capable of processing various types of data for a wide range of applications. We also plan to incorporate multivariate visualization, graph mining techniques and fine grained spatial analysis to uncover more potential patterns and trends within these datasets. Moreover, we aim to conduct more realistic case studies to further evaluate the effectiveness and scalability of the different models in our system.

REFERENCES

[1] GandomiA,HaiderM.,vol.35,no.2,pp.137-144,Apr.2015.

[2] Wang Y, Kung L A, et al. An integrated big data analytics-enabled transformation model: Application to health care. Information & Management,vol.55,no.1,pp.64-79,Jan.2018.

[3] Thongsatapornwatana U. A survey of data mining techniques for analyzing crime patterns. In: 2nd Asian Conf. On Defence Technology, ChiangMai,Thailand,2016,pp.123-128.

[4] Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health information science and systems, vol.2,no.1,pp.1-10,Feb.2014.

[5] Londhe A., Rao P., Platforms for big data analytics: Trend towards hybrid era, 2017 Int. Conf. on Energy, Communication, Data Analytics and Soft Computing (ICECDS), Chennai, 2017, pp. 32353238.

[6] Grady W., Payne A. et al., Agile big data analytics: Analytics Ops for data science, In2017IEEEInt.Conf.OnBigData,Boston,MA, USA,2017,pp.2331-2339.

[7] VatrapuR., R.MukkamalaR., HussainA.etal., Social Set Analysis: A Set Theoretical Approach to Big Data Analytics, IEEE Access, vol.4,pp.2542-2571,Apr.2016.

[8] Zhang Y, Ren S, Liu Y, et al. A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. Journal of Cleaner Production, vol. 142, no. 2, pp. 626641,Jan.2017.

[9] Ngai E W T, Gunasekaran A, et al. Big data analytics in electronic markets. Electronic Markets, vol.27, no.3, pp.243-245,Aug.2017.

[10] Liu Y Y, Tseng F M, Tseng Y H. Big Data analytics for forecasting tourism destination arrivals with the applied Vector Auto regression model. Technological Forecasting and Social Change, vol. 130, pp. 123-134,May2018.

[11]] Joshi A., SabithaA. S., et al.: Crime Analysis Using K-Means Clustering. In: 2017 3rd Int.Conf. on Computational Intelligence and Networks,Odisha,2017,pp.33-39.

[12] Wang S., Wang X., et al.: Parallel Crime Scene Analysis Based on ACP Approach. IEEE Transactions on Computational Social Systems,vol.5,no.1,pp.244-255, Jan.2018.

[13] Baloian N. et al.: Crime prediction using patterns and context. In: 21st IEEE Int. Conf. on Computer Supported Cooperative Work in Design,Wellington,NewZealand,2017,pp.2-9.

[14] ZhaoX.,TangJ.: Exploring Transfer Learning for Crime Prediction. In Proc. IEEE Int. Conf. on Data Mining Workshops, New Orleans, LA,2017,pp.1158-1159.

[15] Vineeth K.,PandeyA., etal.: Anovel approachfor intelligent crime pattern discovery and prediction. In: Int. Conf. on Advanced Communication Control and Computing Technologies, Ramanathapuram,Ramanathapuram,India,2016,pp.531-538.

[16] Rodríguez C., Gomez D., et al.: Forecasting time series from clustering by a memetic differential fuzzy approach: An application to crime prediction. IEEE Symposium Series on Computational Intelligence,Honolulu,HI,2017,pp.1-8.

[17] Noor N., Ghazali A., et al.: Supporting decision making in situational crime prevention using fuzzy association rule. In: Int. Conf. on Computer, Control, Informatics and Its Applications (IC3INA),Jakarta,2013,pp.225-229.