**CHAPTER 1**

**INTRODUCTION**

**1.1 Introduction**

The project focuses in applying machine learning methods to datasets regarding crime (crime statistics in particular cities) and possible related factors. Specifically, we are interested in investigating if it is possible to predict criminal events for a specific time and place in the future (for example, assigning a risk level for a shooting within the next week to different neighbourhoods).

There is an abundance of data to incorporate into the model. Most notably, numerous cities, such as Chicago, have large datasets detailing crimes over the past decade. This data is broken down into type of crime (e.g. theft, murder, narcotics, etc.), time of day and year, geographic location, and more.

Furthermore, cities have released a myriad of other datasets that may possibly relate: median income, location of police stations, etc. there is a copious amount of data to engineer features from and incorporate. This problem area is interesting because of the dynamics that drive crime and the level of granularity needed.

Furthermore, this problem area is useful and novel. Being able to predict where crime will be most prevalent is of obvious benefit to a city and its citizens. The police force could allocate more resources to the predicted high-crime area in order to pre-empt criminal activity, and citizens would be safer by knowing what parts of a city they should avoid. However, in order to create actionable insights, the model created needs to be able to predict crime within a small time frame, for example around a day or week, and a small geographic region, e.g. a block or a neighbourhood.

Having described the available data and the possible difficulties, we can now renew the questions to investigate in this project. First and foremost, we are interested in seeing if we can predict the criminal incidents, perhaps for a specific type of crime, for a small time frame and geographic region. Second, we are interested in learning which features have the most predictive power with respect to crime. Having an understanding of driving factors, cities can better work to mitigate the risk factors for crime.

**CHAPTER 2**

**LITERATURE REVIEW**

**2.1 Literature Review**

The paper is written by M.Saravanan, Rakhi Thayyil and Shwetha Narayanam they wrote the model paper which prediction using mobile call detail records of suspects and victims to understand their presence in crime. It records of cell tower near crime scene have been analysed to track the real perpetrators. Most of the data mining technique are used by them as prediction techniques these techniques are clustering, classification, pattern mining, event extraction from textual documents, nearest neighbour method etc. Also they use classifiers applied to crime cases CRF classifier, Naive Bayes and JRip classifiers.[11]

The paper is written by Xinyu Chen, Youngwoon Cho and Suk Young jang they wrote the model paper which prediction using twitter sentiment and weather. They tell that when statistical analysis is applied to unstructured data of twitter reveal valuable insights. In this paper they predict the time and location in which specific type of crime will be occur. They only take the previous weather data and also at this weather which type of crime will be occur according to this data they will be able to predict the crime with very low accuracy. In the case of sentimental analysis the researchers employ semantic analysis on the contextual contents of each tweet and draw the predictive response of the selected group of people.[10]

A.Wang and L.Perez proposed apply machine learning methods to datasets regarding crime (crime statistics in particular cities). Specifically, we are interested in investigating if it is possible to predict criminal events for a specific time and place in the future (for example, assigning a risk level for a shooting within the next week to different neighbourhoods). They applied Gaussian Process on abundance data such as Chicago and New York crime datasets over the past. This data is broken down into type of crime (e.g. Theft, Murder, Narcotics, etc.), Time of day and year, geographic location, and more. They applied the historical data of crimes and took high crime areas to calculate crime hotspots. [8]

PEVNET tool are used to visualize, A.Rasheed and U.K.Wiil applied PEVNET as IA tool by demonstrating its utilization with the perspective of visualization. They used PEVNET features like similar node feature, clustering of sub-groups, detecting collaborating sub-cluster feature, trend analysis. We are inspired by this paper for using Chicago datasets which is available in excelle format. By using PEVNET the analyst can reveal the hidden interactions or patterns in their relations which might lead to linkage of activities that gives a strong clue of involvement of certain nodes in criminal activities. [7]

The paper is written by P. Thongtae and S. Srisuk they will give the information about how the data mining use in crime domain. They use data mining technique such as association mining, classification and prediction, clustering they can proposed a framework for regional crime analysis program (Re-CAP) in which they use meta-data, process and display the useful informative data. [6]

The paper is written by shiju sathyadevan,devan M.S.,Surya Gangad-haran.S will predict crime using classification and prediction data mining technique in that paper they get the data from news, blogs, social media etc. and stored in the database as raw data and then classify the collected raw data by naive bayes-classifier and statistical method for classi\_cation.by using this classified data they will predict the crime using decision tree.[5]

The paper is written by Amer Rasheed and U\_e Kock Wiil will visualize the criminal network by using PEVNET technique.in this technique they will use clustering to find the network of attacker they noted the one attacker person and then find all links related to him to visualize the whole network. This technique will be useful to find the entire criminal network related to that attack. [4]

In the paper purposed by J. Chae, J. Zhang,W. Hatton and A. Malik, different Data Mining techniques are used such as several clustering techniques based on different characteristics , including k-means, considering different parameters such as movement data and communication data and activities for analysis of criminal event and using heat maps and line graphs.[1]

**CHAPTER 3**

**ANALYSIS**

**3.1 Project plan**

**3.2 Requirement analysis**

**3.3 Problem Analysis**



**Figure 1: Analysis**

**Initial Analysis** -

This contins main analysis phase where all the raw data is processed and all the datasets with miss-spellings are handled by the given phase for analysis. This is the \_rst stage of data preprocessing where datasets are processed. For all the categorical variables dummy variable are generated where as while doing this precostions are taken that there will be no dummy trap condition will occur.

**Handling Missing Values** -

All the missing values in any dataset record are handled in pre-processing phase all the attributes can be removed from the dataset which is one of the easiest method but it could effect thr accurricy of the given model hence there is another method which is more sutable. By calculating the mean value of given column we can replace the missing value by that mean value.

**Feature Reduction** -

Feature reduction is the process of reducing the features which are dependent of n other independent variable as well as those feature which dosent contribute in prediction model.Some features could cause the prediction module to decrease the prediction value as well as accuracy

**Feature Extraction** -

Feature extraction is the process of generating no of feature from the given data field from the dataset.Only those features are extracted the canbe used in predictive model/machine learning model to predict the future values. Feature extraction or feature selection is complex as well as important process beacause the whole model is dependent on that.selection of right features inceases the accuricy of model where wrong features could misslead the prediction model.

**XGBoost Model** -

XGBoost Model stands for Xtream Gradient Boosting Model.It is the sklearn library in python which provides a prediction model based on gaining maximum prediction values by combining more than one Classification and Regration trees(CART).It enhances the probabilty of correct prediction by boosting the gradient of the model.

**CHAPTER 4**

**DESIGN**

**4.1 Architecture :**



**Figure 2: Basic Architecture**

Step 1: Data Acquisition and Preprocessing

* Removing missing values.
* Handling categorical variable
* Feature Reduction

Step 2: XGBoot Classi\_cation and Prediction

* fit the model to dataset
* Classi\_cation of Data using xgbclassi\_er.
* Prediction of crime in future.

Step 3: Training and Testing Set

* splitting dataset into training and testing set
* fit testing set with dependent and independent variable
* Backward elimination.

Step 4: Model testing

* Apply testing dataset to predictive model.
* Evaluate model comparing predicted value on exact value.

**CHAPTER 5**

**MODELLING**



**Figure 3: UML Diagram**

**CHAPTER 6**

**CODING**

**6.1 Algorithum**

Step 1:-

Import the data from Chicago website

Input the dataset as matrix of features and dependent variables Dataset (Xi,Yi);

Step 2:-

Initialize the XGBoost classifier from the XGBoost package set testing models to true when pre-processing the data and false when final phase

Step 3:-

Load the pre-processed data

Step 4:-

Split the training and test set

Step 5:-

Standard scale the x,y initialize raw x,y for test and train data add some new features based on latitude and longitude rotate coordinates by 30,45, and 60 degrees compute the radial distance from centre.

Step 6:-

To train model, we need to optimize a loss function (Loss function measures the performance of a classification model) Generation and ensemble of classification and regression trees Rooted Mean Squared Error for regression



Log Loss for binary classification



Step 7:-

To optimize it by gradient descent, we need to calculate the gradient. The performance can also be improved by considering both first and second other gradient.



where KK is the number of trees, is a function in the functional space

FF, and FF is the set of all possible CARTs.

Step 8:-

XGBoost can handle missing values in the data. For each node, we guide all the data points with a missing value to the left sub node, and calculate the maximum gain To the right sub node, and calculate and maximum gain Choose the direction with larger gain Finally, every node has a default direction for missing values.

Step 9:-

Parameter selection for CART



Step 10:-

Reduce the number of dimensions now we can re-construct the data set

Step 11:-

Optimization goal for new tree



6.2 Software used

Editor - Spyder

GUI - HTML, CSS, D3

Database - CSV

Library - Scikit Learn

Platform - Window

Language - Python

6.3 Hardware specification

Processor - PC with minimum 4 cores @ 3.1GHz

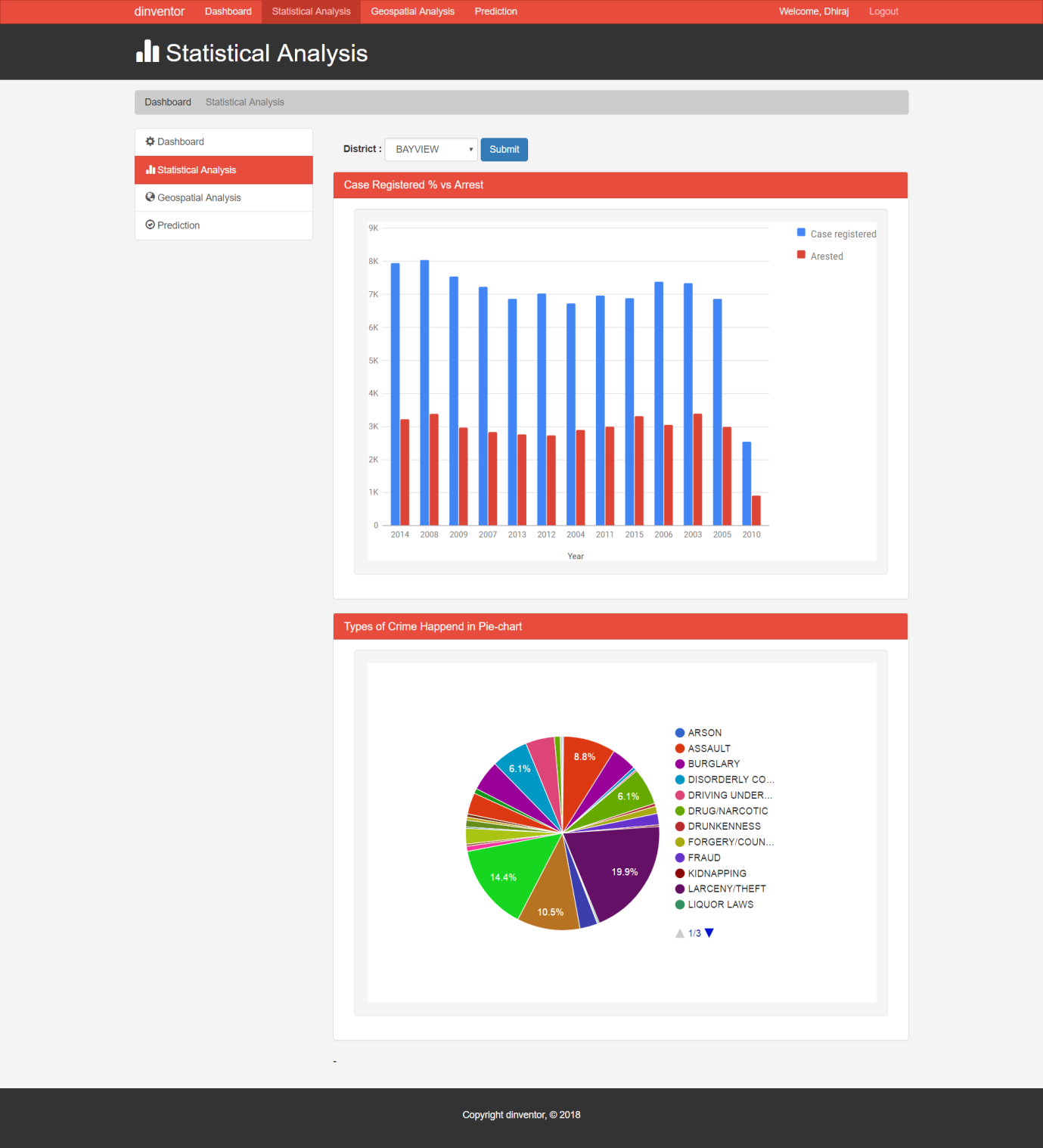
RAM - 8GB.

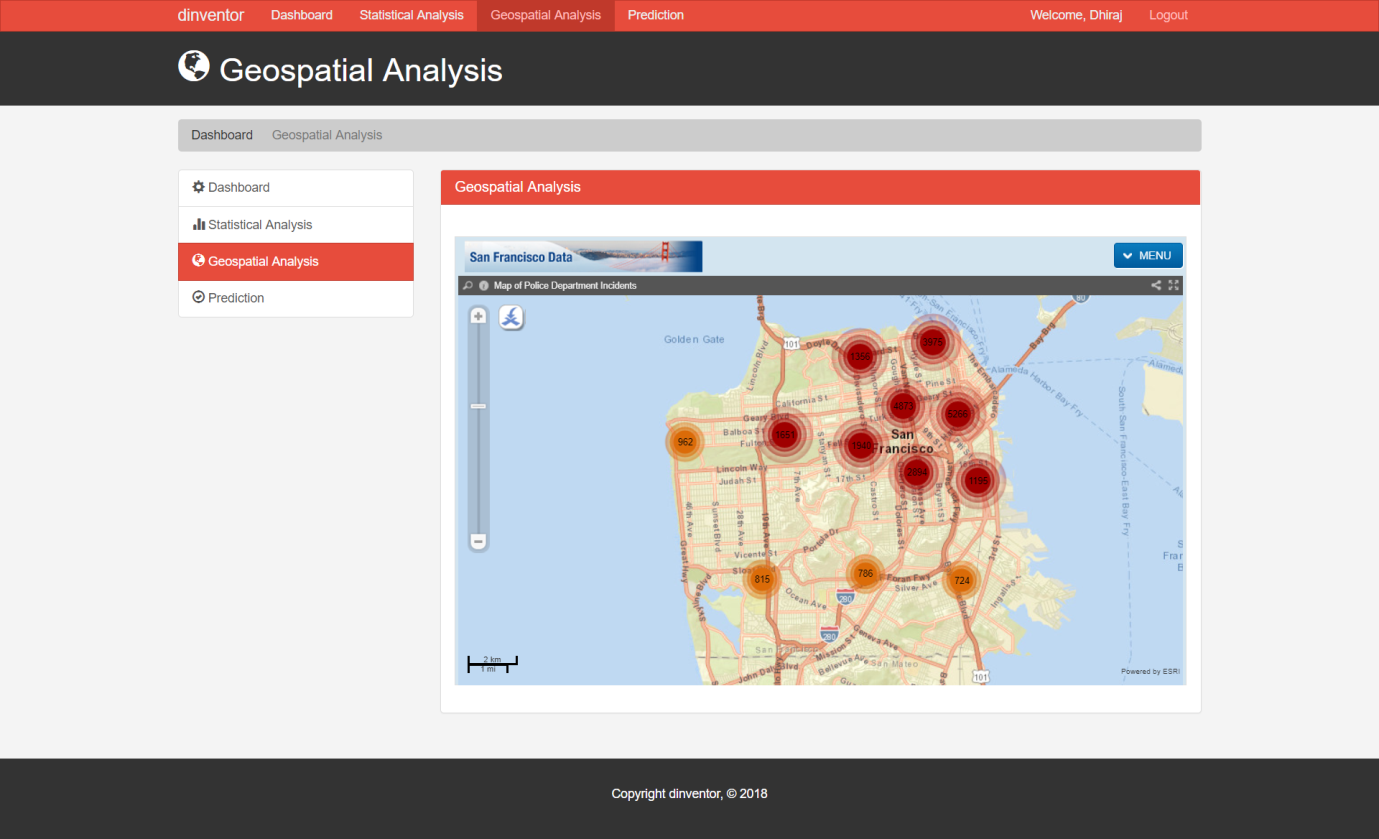
Graphics Card - 2GB Graphics.

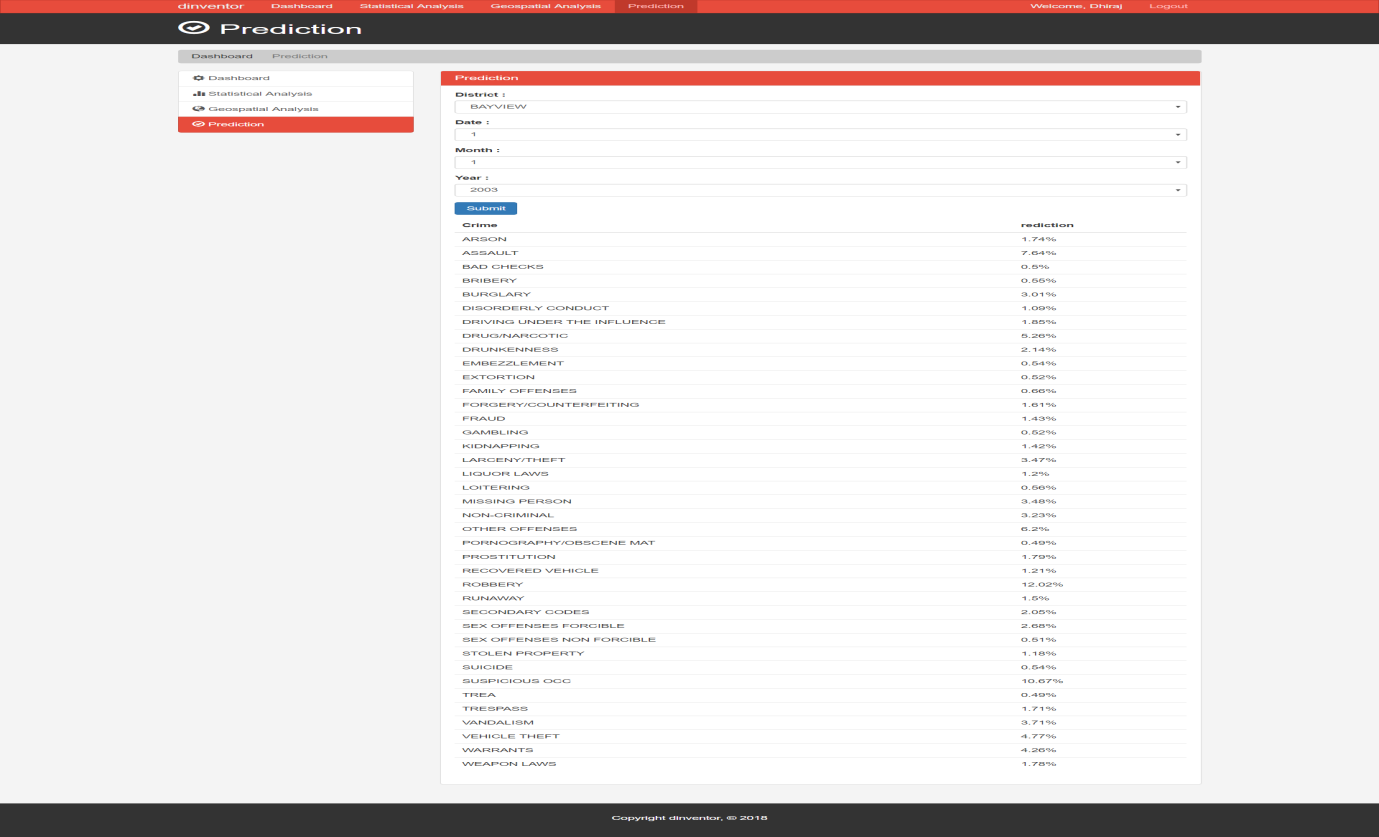
**CHAPTER 7**

**RESULT SETS**

****

****

****

****

**CHAPTER 8**

**CONCLUSION**

**8.1 Conclusion :**

These we have apply the machine learning algorithm regarding crime (crime statistics in particular cities) and possible related factors. Specifically, we are investigated predict criminal events for a specific time and place in the future (for example, assigning a risk level for a shooting within the next week to different neighbourhoods). We aim to assign law enforcement in crime hotspot areas to prevent future potential crimes.

**CHAPTER 9**

**REFERENCES**

**9.1 Reference :**

[1] J.Chae,Jiawei Zhang,Abish Malik,William Hatton,"Visual Analytics of Heterogeneous Data for Criminal Event Analysis".

[2] Isuru Jayaweera, Chamath Sajeewa, Adeesha Wijayasiri, Crime Analytics: Analysis of Crimes through Newspaper Articles.

[3] Chih-Hao Ku, James H. Nguyen, Gondy Leroy, TASC Crime Report Visualization for Investigative Analysis: A Case Study.

[4] Shiju Sathyadevan, Devan M.S, Surya Gangadharan. S, Crime Analysis and Prediction Using Data Mining.

[5] Amer Rasheed, U\_e Kock Wiil, PEVNET: A Framework for Visualization of Criminal Network2014.

[6] P. Thongtae and S. Srisuk, An Analysis of Data Mining Applications in Crime Domain .

[7] Amer Rasheed, U\_e Kock Wiil, PEVNET: A Framework for Visualization of Criminal Network 2015.

[8] Luis Perez and Alex Wang, Gaussian Processes for Crime Prediction.

[9] Jazeem Azeez and D. John Aravindhar, Hybrid Approach to Crime Prediction using Deep learning.

[10] Xinyu Chen, Youngwoon Cho, and Suk young Jang, Crime Prediction Using Twitter Sentiment and Weather.

[11] M. Saravanan,Rakhi Thayyil and Shwetha Narayanan, Enabling Real Time Crime Intelligence Using Mobile GIS and Prediction Methods.