**CHAPTER 5**

**DATA ACQUISITION**

The dataset for the proposed system is true and authentic and is acquired from the Chicago City Online Portal. The dataset is updated daily and contains all the criminal records for the past 17 years from (2001-2017).The Dataset has a total of 22 attributes and 6.4 million rows. The next dataset contains the images of all the criminal people recorded in the city of Chicago. This Dataset is primarily used for Visual verification of the criminal since the system works in a real time environment. Further Datasets includes the addresses of all the apartments, cafes, abandoned buildings and vehicles, grocery stores, parking locations etc located in the city of Chicago. These datasets are analyzed in order to form the training sets for the prediction model.

**5.1 Dataset**

The Following are the datasets are being used for training and testing

* Crimes in Chicago(2005-2007) – (460 MB)
* Crimes in Chicago(2008-2011) – (660 MB)
* Crimes in Chicago(2012-2017) – (235 MB)
* Criminal Photo Dataset(2001-2017) – (4.1 GB)
* Chicago Enterprise Block Address(2000-2017) – (325 MB)

**5.2 Dataset Pre-Processing**

The Datasets must be pre-processed in order to fit the data for further classification. We must first clean the data, integrate them and then transform it to the desired processing format. The Not Applicable values are removed from the dataset in order to make the data suitable for all types of processing.

|  |  |  |
| --- | --- | --- |
| **Attributes** | **Data type** | **Description** |
| Primary Type | Plain Text | Type Of Crime |
| Arrest | Checkbox | Whether the criminal is arrested or not |
| Description | Plain Text | Cause of Crime |
| Location Description | Plain Text | Description of the location of the crime |
| Block | Plain Text | Address of the place of the crime |
| Latitude | Number | Latitude of the location where the crime happened |
| Longitude | Number | Longitude of the location where the crime happened |
| Ward | Number | City Council District where the crime happened |
| Date | Date & Time | Date & Time when the crime happened |

**Table 1. Chicago Crime Dataset Attributes**

'Latitude' and 'Longitude' are two primary attributes which contains a lot of NA values and these values must be certain attributes must be neglected like 'ID','IUCR','FBI CODE','UPDATED', 'LOCATION'. These attributes increase the size of the dataset and reduce processing speed.