Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

CSE4250 | Thesis-II

Group No 3903

Presented by:

Nakib Uddin Ahmed 16.01.04.137
Abu Ubaida Akash 17.01.04.060
Md Atique Ahmed Ziad 17.01.04.087
Faisal Tareq Shohan 17.01.04.105

Supervised by:

Dr. Md. Shafiul Alam
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Development of Machine Learning Models for Crime Prediction using Historical Data

Literature Review

1. Predicting Crime Using Time and Location Data (2019)

Yuki, Jesia & Sakib, Md. Mahfil & Zamal, Zaisha & Habibullah, Khan & Das, Amit

The major aim of this study was to expect which category of crime is most probably to take place at a detailed time and places in Chicago.

Algorithms:

- Decision Tree
- Random Forest
- Bagging
- AdaBoost
- ExtraTree Classifier

Dataset:

- Chicago Crime Dataset (over 16 years)

Result:

Algorithm	Accuracy
Random Forest	95.99%
Decision Tree	99.88%
AdaBoost	74.78%
Bagging	99.92%
Entra Tree	91.10%

Literature Review

2. Building a Learning Machine Classifier with Inadequate Data for Prediction (2017) *Nguyen, Trung & Hatua, Amartya & Sung, Andrew*

A crime predicting method which forecasts the types of crimes that will occur based on location and time.

Algorithms:

- Support Vector Machine
- Random Forest
- Gradient Boosting
- Multilayer Neural Network

Dataset:

- Portland Police Bureau (PPB)
- Public Government Source American FactFinder

Result:

Algorithm	Accuracy
Support Vector Machine	67.095%
Random Forest	67.088%
Gradient Boosting	76.42%
Multilayer Neural Network	50.2%

Literature Review

3. Crime Analysis Through Machine Learning (2018)

Suhong Kim , Param Joshi, Parminder Singh Kalsi, and Pooya Taheri

This paper investigates machine-learning-based crime prediction.

Algorithms:

- K-Nearest Neighbor
- Boosted Decision Tree

Dataset:

- Vancouver Police Dataset

Result:

Algorithm	Accuracy
K-Nearest Neighbor	39%
Boosted Decision Tree	44%

Literature Review

4. Crime Prediction and Analysis Using Machine Learning(2018)

Bharati and D. S. RA

Result:

Algorithm	Accuracy
K-Nearest Neighbor	78.91%
Gausian Naive Bayes	64.60%
Multinomial Naive Bayes	45.60%
Bernouli naive Bayes	31.35%
SVC	31.35%
Decision Tree	78.60%

Algorithms:

- K-Nearest Neighbor
- Gausian Naive Bayes
- Multinomial Naive Bayes
- Bernouli Naive Bayes
- SVC
- Decision Tree

Literature Review

5. Crime Prediction Using Spatio-Temporal Data(2020)

S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker

Algorithms:

- Decision Tree
- K-Nearest Neighbor
- Random Forest

Dataset:

- San Francisco PD Crime Dataset (over 16 years)

Result:

Algorithm	Accuracy	Log Loss
Decision Tree	31.17%	3.312
KNN (N=50)	28.50%	5.04
KNN (N=500)	27.91%	2.62
Random Forest (T=10)	31.22%	2.34
Random Forest (T=50)	31.70%	2.28
Random Forest (T=100)	31.71%	2.28

Proposed Methodology

Steps for developing any machine learning models:

- Data Collection
- Data Preprocessing
- Feature Extraction
- Classification Strategy

Data Collection

Possible Sources

Bangladesh Police	Bangladesh Newspapers

Data Collection

Selected Source

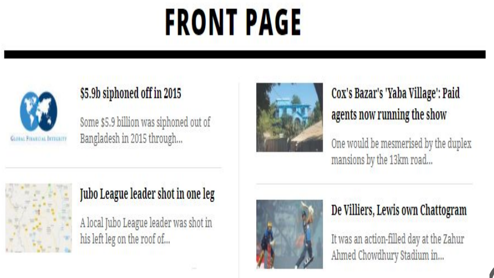
The Daily Star

Bangladesh Newspaper

- Gathering crime headlines and links
- Extracting the informations

Data Collection

Paper Sections



FRONT PAGE

$5.9b siphoned off in 2015

Some $5.9 billion was siphoned out of Bangladesh in 2015 through...

Jubo League leader shot in one leg

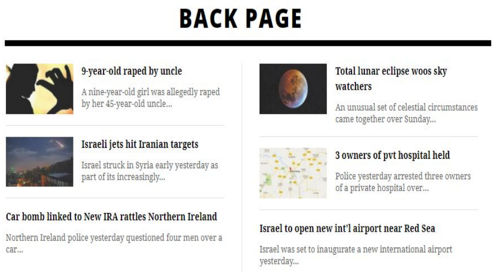
A local Jubo League leader was shot in his left leg on the roof of...

Cox's Bazar's 'Yaba Village': Paid agents now running the show

One would be mesmerised by the duplex mansions by the 13km road...

De Villiers, Lewis own Chattogram

It was an action-filled day at the Zahur Ahmed Chowdhury Stadium in...



BACK PAGE

9-year-old raped by uncle

A nine-year-old girl was allegedly raped by her 45-year-old uncle...

Israeli jets hit Iranian targets

Israel struck in Syria early yesterday as part of its increasingly...

Car bomb linked to New IRA rattles Northern Ireland

Northern Ireland police yesterday questioned four men over a car...

Total lunar eclipse woos sky watchers

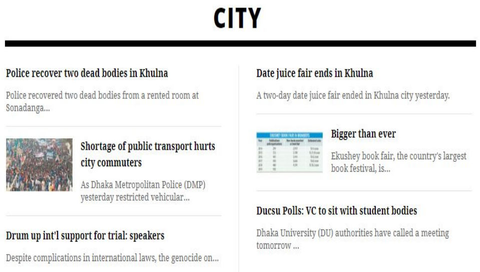
An unusual set of celestial circumstances came together over Sunday...

3 owners of pvt hospital held

Police yesterday arrested three owners of a private hospital over...

Israel to open new int'l airport near Red Sea

Israel was set to inaugurate a new international airport yesterday...



CITY

Police recover two dead bodies in Khulna

Police recovered two dead bodies from a rented room at Sonadanga...

Shortage of public transport hurts city commuters

As Dhaka Metropolitan Police (DMP) yesterday restricted vehicular...

Drum up int'l support for trial: speakers

Despite complications in international laws, the genocide on...

Date juice fair ends in Khulna

A two-day date juice fair ended in Khulna city yesterday.

Bigger than ever

Ekushey book fair, the country's largest book festival, is...

Ducsu Polls: VC to sit with student bodies

Dhaka University (DU) authorities have called a meeting tomorrow ...

COUNTRY

Bridge work on Ratnai river misses second deadline

Several thousand villagers of 10 ten villages under Mogholhat union...

Ex-UP member found murdered

A former union parishad (UP) member was found murdered in...

Woman stabbed dead by husband

A woman was stabbed to death allegedly by her husband over a family...

Candidates of 3 parties belonging to grand alliance to contest

Candidates of three parties belonging to the grand alliance will...

Selection of Crimes

Narcotics	112549
Woman & Child Repression	16253
Theft	5561
Murder	3830
Smuggling	4501
Theft	5561
Kidnapping	444
Robbery	562
Assault	811

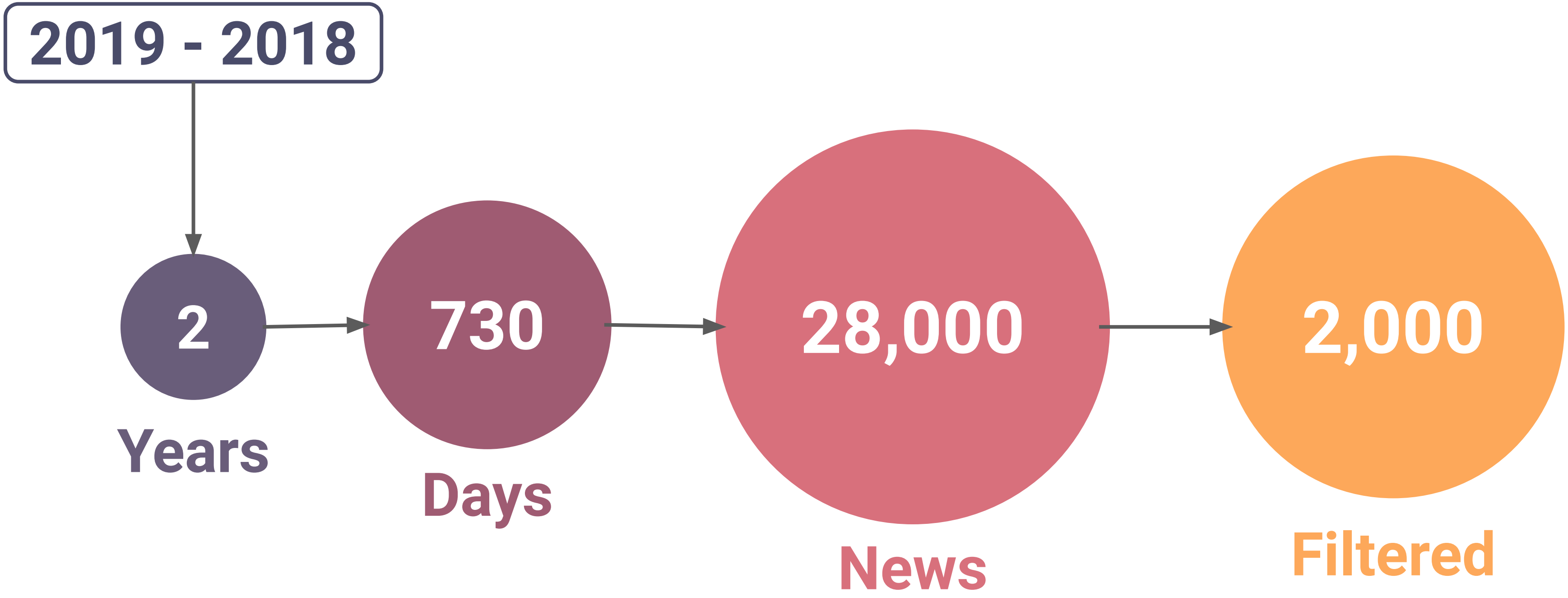
Mostly Occurred
in
Bangladesh
(2018)

Selection of Crimes

01	02	03	04	05	06
Murder	Rape	Assault	Body Found	Kidnapping	Robbery

Rationality for Newspaper

Data Collection - Manual



Data Collection - Manual

BACK PAGE

Two workers killed

Two workers were killed in a landslide at an illegal stone quarry...

One killed as AL men clash

A man was killed and at least two others were injured in a clash...

Einstein vs Modi

The organisers of a major Indian science conference distanced...

WikiLeaks tells journos 140 things not to say about Assange

WikiLeaks on Sunday advised journalists not to report 140 different...

RMG workers' wage demo continues

Vehicles start to move on the busy Airport Road in Dhaka after the...

Of a polls observer group

An EC-registered organisation which observed the December 30...

Schoolgirl murdered after 'rape'

An eight-year-old girl was killed after being "raped" in Gabtala...

UN observers welcome to Xinjiang, with conditions: China

China said yesterday it would welcome UN officials to the restive...

Myanmar asks army to crush Rakhine rebels

Myanmar government leader Aung San Suu Kyi yesterday discussed...

British citizen of Bangladesh origin found dead in rehab

A Bangladesh-born UK citizen on Sunday was found dead in front of a...

Rape News

Murder News

Accident News (False Positive)

News out side of Bangladesh (false Positive)

12:00 AM, June 12, 2018 / LAST MODIFIED: 12:06 AM, June 12, 2018

Schoolboy found dead

Our Correspondent, Pirojpur

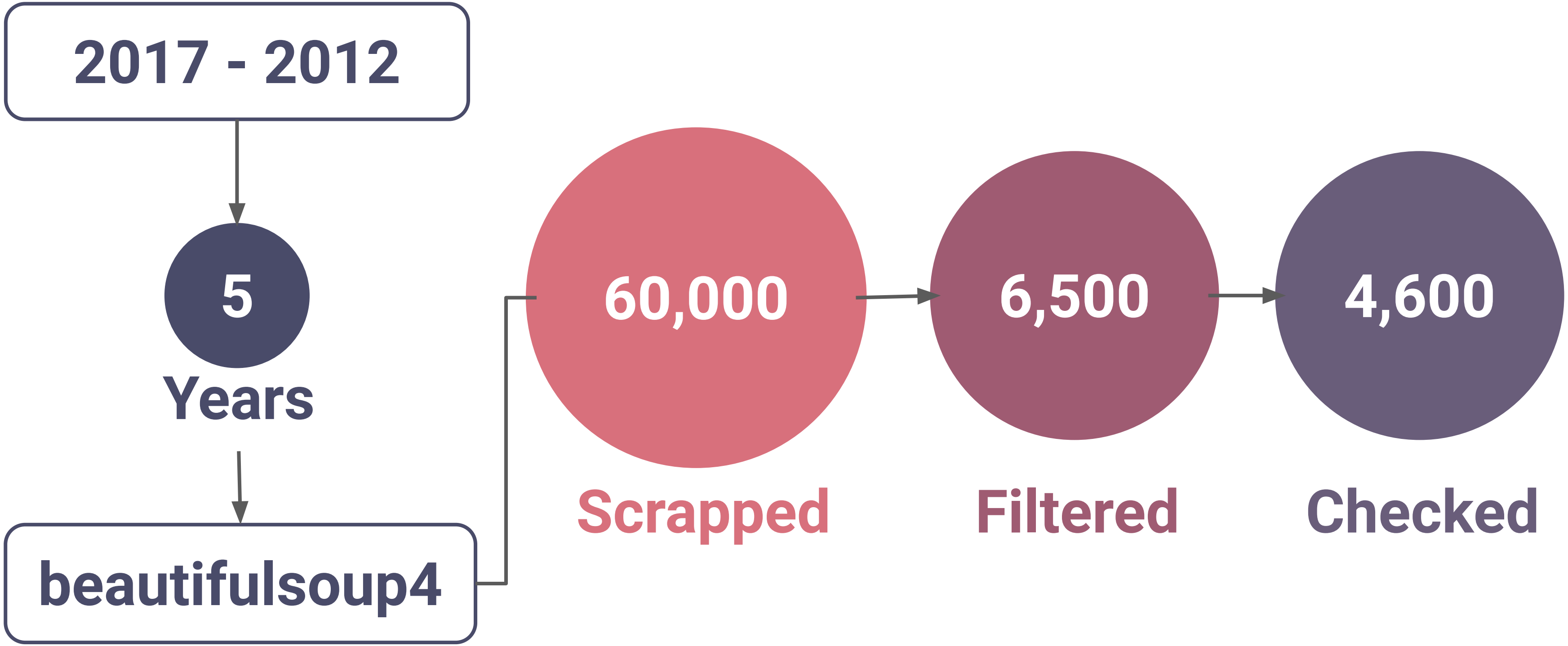
A schoolboy was found dead in Bhandaria upazila of the district on Sunday.

The deceased was Yamin Hossain Hridoy, 14, son of Md Shahjahan Hawlader of Darulhuda village, and a Class VIII student of Pasharibuniya High School in the upazila.

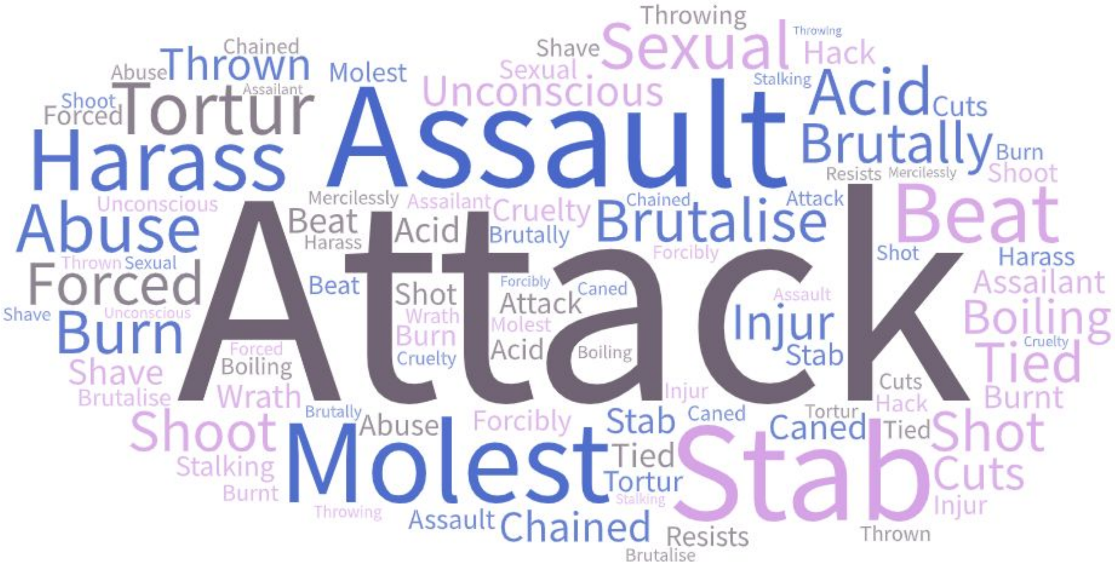
The victim's father Shahjahan said Yamin went to sleep at a decorator shop of one of his relatives on Saturday night and did not return home.

Later, locals found the body floating on a ditch near the shop on Sunday evening and informed police.

Data Collection - Scrapped



Data Collection - Scrapped - Filtering

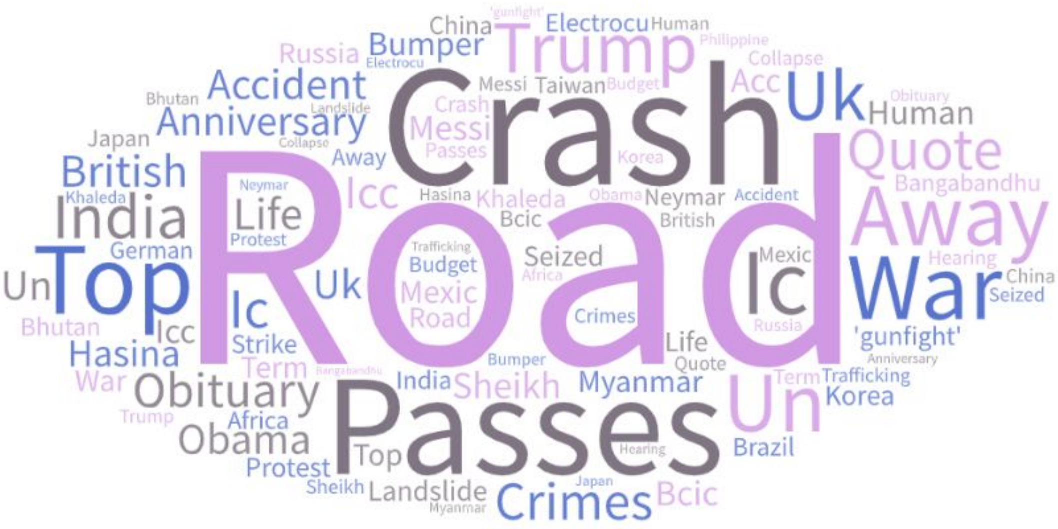


- Filtering crime news from scrapped news link with most frequent keywords

Data Collection - Scrapped - Filtering

- ❑ Filtering crime news from scrapped news link with most frequent keywords

Data Collection - Scrapped - Filtering



❑ Removing redundant news based on keyword not related to crime

Feature Selection

Country

Youth kills uncle, lands in jail

Our Correspondent, Mymensingh

Sun Apr 7, 2019 12:00 AM Last update on: Sun Apr 7, 2019 12:06 AM 1

A man was stabbed 2 to death allegedly by his nephew 3 in Monipur Ghat area of Kishoreganj town 4 on Friday evening. 5

The victim was day labourer 6 Habibur Rahman Habib, 55, 7 son of Abdur Rashid of Brahmankandi in Sadar upazila. 8

Quoting locals, Officer in Charge of Kishoreganj Model Police Station Abu Bakar Siddiq said there had been a feud 10 between Habib and his nephew Shamim Ahmed Akash, 23, 9 son of Dulal Mia, over killing of Akash's sister-in-law Rawshan Ara in 2017.

Both Habib and Dulal are stepbrothers, police said.

A case was lodged, accusing Rawshan's husband Hazrat Ali, his parents, brother Akah and Habib's son Humayun.

1. News Date
2. Criminal Approach
3. Relation between Victim and Criminal
4. Incident Place
5. Incident Time
6. Victim Profession
7. Victim Age
8. Victim's Address
9. Criminal Age
10. Motive behind the Crime.

Feature Selection

Since the murder, Akash and his brother Hazrat had been blaming Habib for implicating them in the case and the feud had been prevailing, said the OC.

Over the feud, Akash stabbed Habib when he was having tea at a stall of Monipur Ghat. Habib died on the spot.

Hearing screams, local people rushed to the spot and caught Akash red-handed. Later he was handed over to police. 11

On information, police recovered the body and sent it to Kishoreganj General Hospital for autopsy.

The victim's wife, Kalpana Akter, lodged a case with the police station, accusing Akash.

Akash was produced before a Kishoreganj court that sent him to jail yesterday.

11. If Criminal Arrested

Feature Selection

News Date	Incident Date	Part of the day	Victim Age
Victims Injured	Incident Area	Victim Tribal	Criminal Age
Criminal Address	Victim Religion	Victim Profession	Criminal Social Status
Victim Address	Criminal Gender	No of Criminals	Criminal Religion
Criminal Tribal	Victims Injured	Victim Gender	Relation Between Victim and Criminal

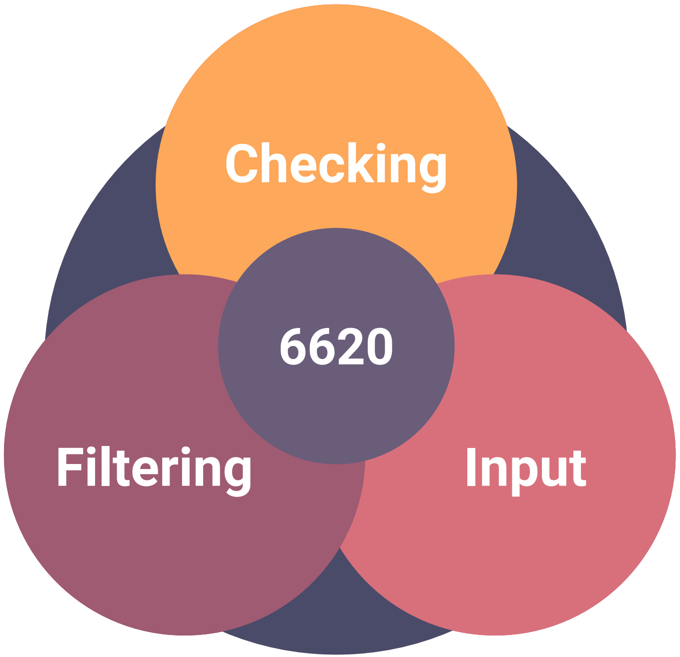
Common Features

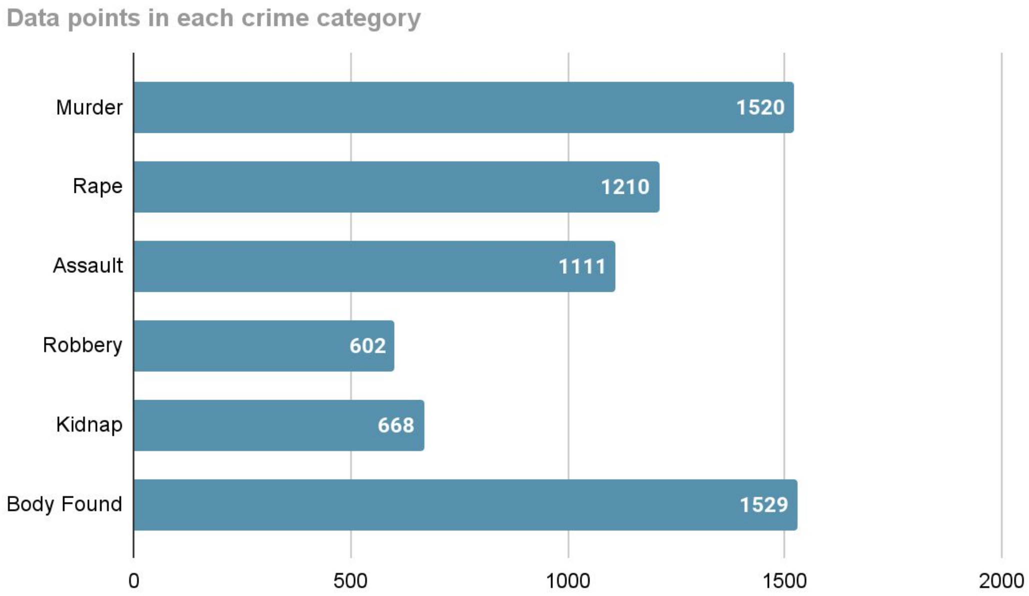
Feature Selection

Murder	Kidnapping	Assault	Robbery	Body Found
Approach	Rescue Date	Reason	Mass Victim	Body State
Weapon	Rescued Area	Physically Harmed	Victime State	
Motive			Mass Criminal	

Category Special Features

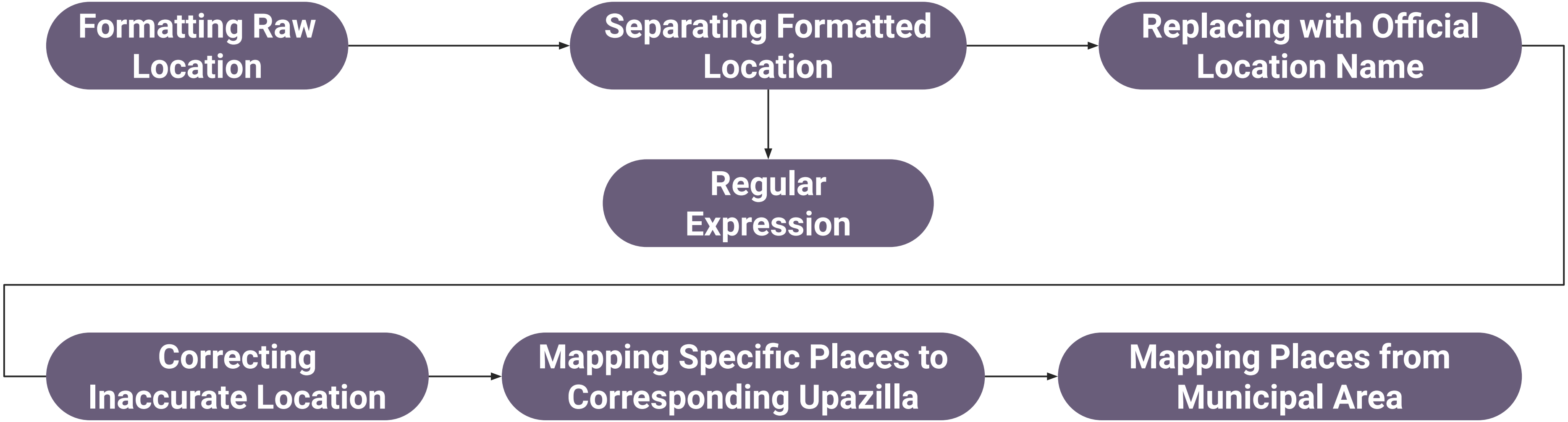
Obtained Dataset





Data Preprocessing

Incident Place



Data Preprocessing

Incident Date

- Manually formatted the dates of different forms, such as - Friday, yesterday, the day before, before an event, etc.
- Python module **“Datetime”** used to extract day number, week day, week number, etc.

Data Preprocessing

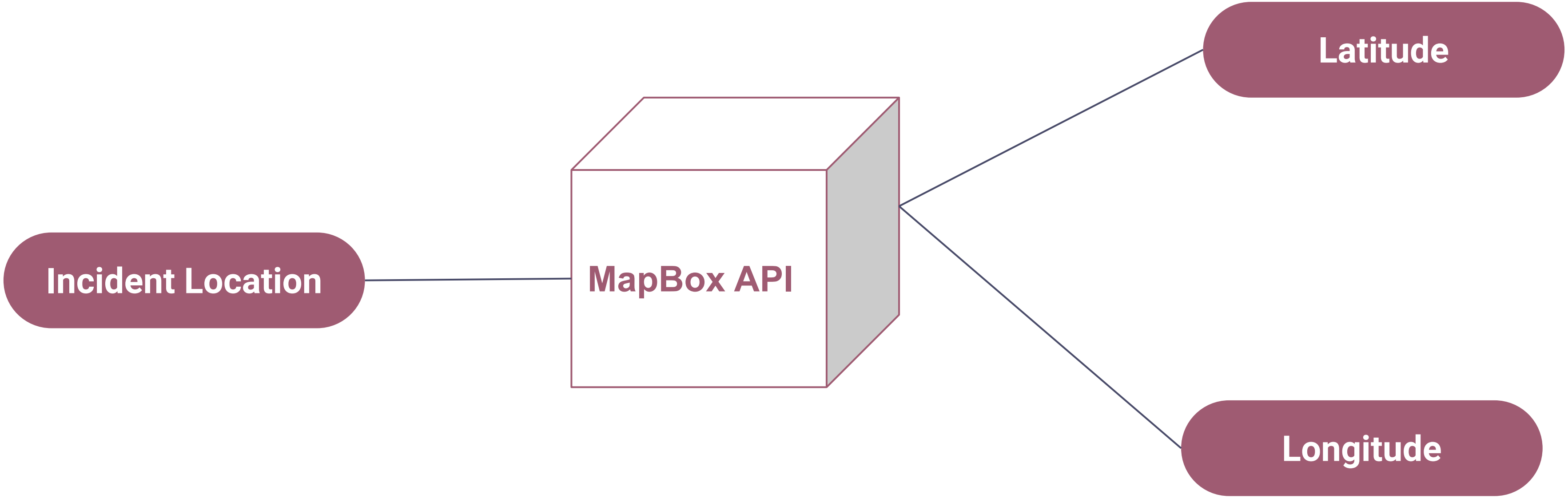
Incident Time

Determines certain part of the day.

Time frames	Part of the day
6 am - 11:59 am	Morning
12 pm - 3:59 pm	Noon
4 pm - 5:59 pm	Afternoon
6 pm - 7:59 pm	Evening
8 pm - 5:59 am	Night

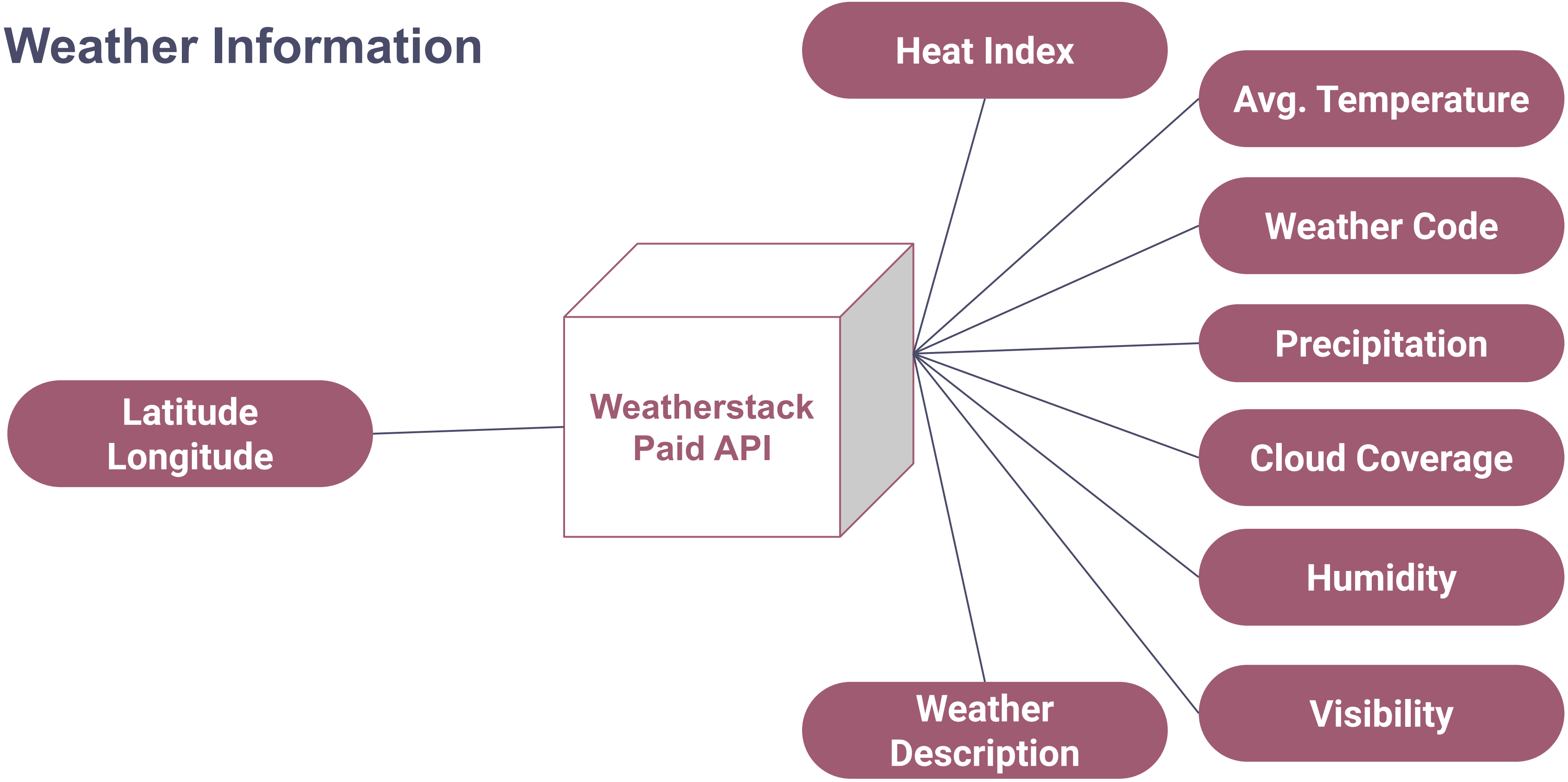
Feature Engineering - Feature Addition

Geocoding



Feature Engineering - Feature Addition

Weather Information



Feature Extraction

Data Adjustment

- Precision of latitude and longitude ranged from 6 to 8 digits in floating point precision.
- Latitude and longitude values were rounded to the nearest six digit floating point precision.

Weather

- Contained in six features: precipitation, humidity, visibility, cloud cover, heat index, and average temperature in the json format.
- The values of all features were converted to Int and Float.

Feature Extraction

Unique Values, Floating Numbers and Sequentiality

- Categorical
  - Few unique values
  - Contains character, string, integers
- Numerical
  - Significant amount unique values
  - Contains integers, floating point numbers

Feature Extraction

Date

- Using Python’s **“Datetime”** module, the incident date was converted to a Datetime Object.
- **Year**, **Month**, **Day**, **Weekday** were extracted from this Datetime object

Season

- Season is another categorical feature.

  Hot: March-May.

  Rainy: June-October.

  Winter: November-February.

Weekend

- This feature returns: True and False depending on weekdays - Friday and Saturday.

Feature Extraction

Correlation of the Weather Features

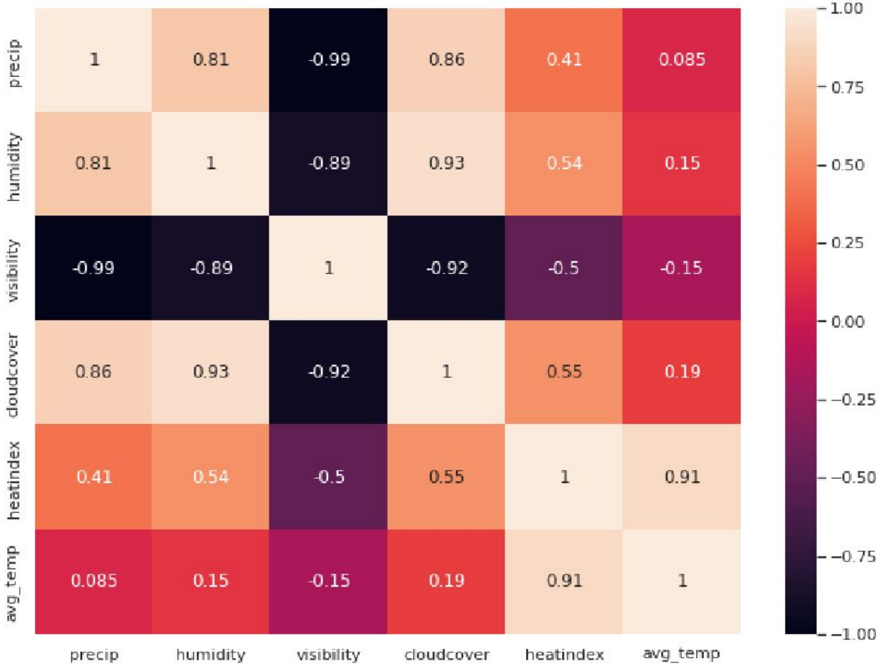


Figure 7.2: Correlation of Six Weather Features

Feature Extraction

Weather features were grouped into three categories:

- Average Temperature and Heat Index.
- Cloud Cover and Humidity.
- Visibility and Precipitation.

Importance of Weather Feature:

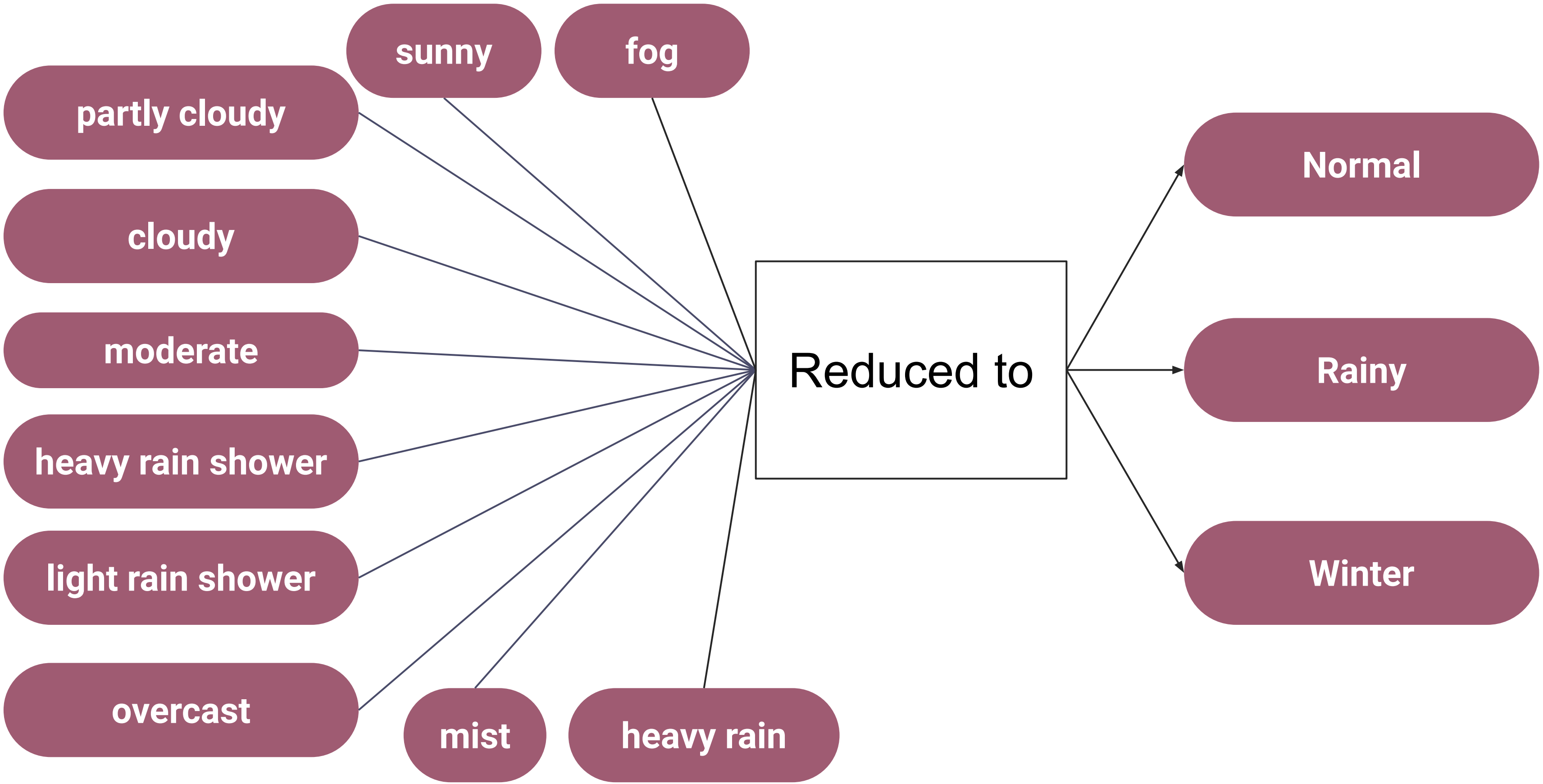
- Feature importance was calculated before choosing one from each category.
- Min-Max normalization was used for this purpose.

Feature Extraction

Importance of Weather Features

Feature	Importance
Cloud Cover	0.258
Humidity	0.206
Precipitation	0.202
Heat Index	0.166
Average Temperature	0.134
Visibility	0.031

Feature Extraction



Feature Extraction

Converting Features from Numerical to Categorical

Precipitation

Intensity	Precipitation Rate
No Rain	rate = 0.0
Light Rain	0.0 <rate <2.5
Moderate Rain	2.5 <rate <10
Heavy Rain	10 <rate <50
Violent Rain	50 <rate

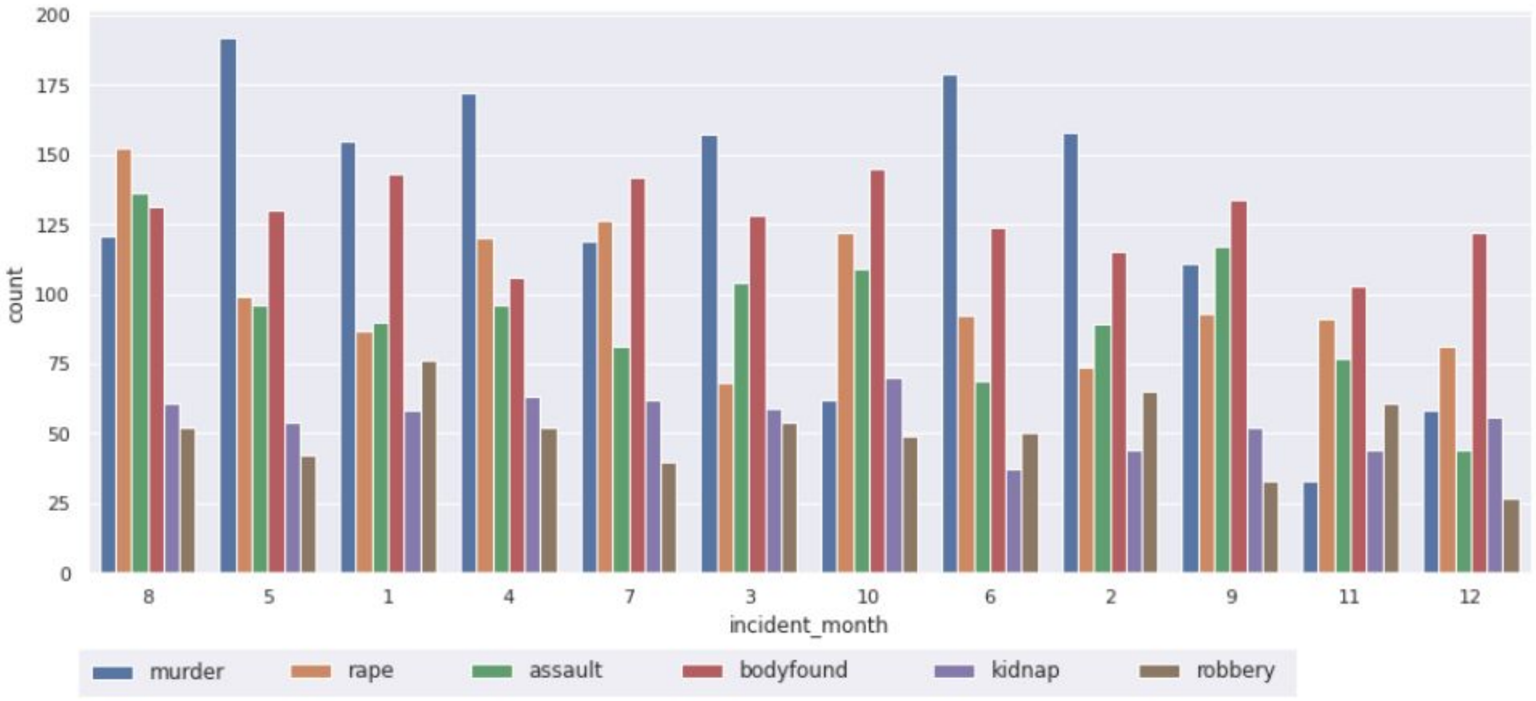
Cloud Cover

Type	Cloud Cover
Clear	cover < 10
Scattered	10 < cover < 50
Broken	50 < cover < 90
Overcast	90 < cover

Heat Index

Shade	Temperature (Celsius)
Normal	below 26
Cautious	26 - 32
Extreme Cautious	33 - 41
Danger	42 - 54
Extreme Danger	over 54

Feature Extraction



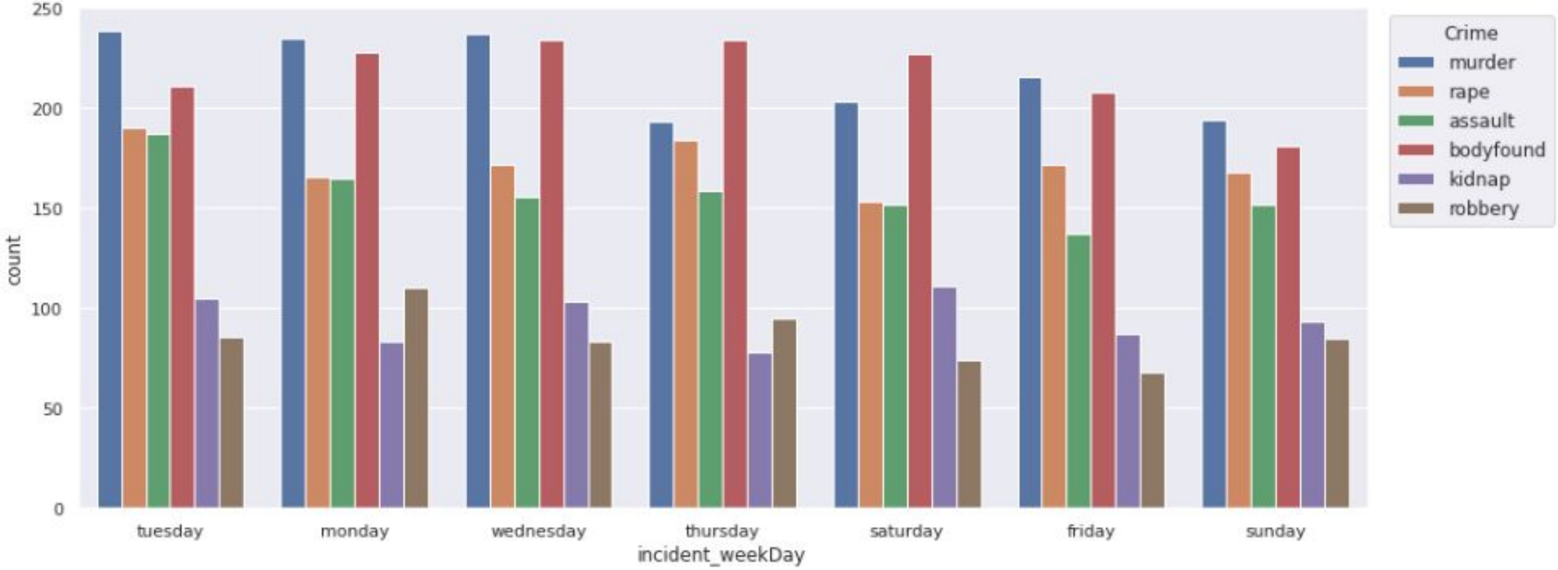
Crime Count Per Month

Feature Extraction

Class	Months	Frequency Order
1	8	Rape >Assault >Body Found >Murder >Kidnap >Robbery
2	5	Murder >Body Found >Rape >Assault >Kidnap >Robbery
3	1,2	Murder >Body Found >Assault >Rape >Robbery >Kidnap
4	4	Murder >Rape >Body Found >Assault >Kidnap >Robbery
5	7	Body Found >Rape >Murder >Assault >Kidnap >Robbery
6	3	Murder >Body Found >Assault >Rape >Kidnap >Robbery
7	10	Body Found >Rape >Assault >Kidnap >Murder >Robbery
8	6	Murder >Body Found >Rape >Assault >Robbery >Kidnap
9	9	Body Found >Assault >Murder >Rape >Kidnap >Robbery
10	11	Body Found >Rape >Assault >Robbery >Kidnap >Murder
11	12	Body Found >Rape >Murder >Kidnap >Assault >Robbery

Crime Frequency Order on Months

Feature Extraction



Crime Count Per Weekday

Feature Extraction

Class	Weekdays	Frequency Order
1	Tuesday, Wednesday Friday, Sunday	Murder >Body Found >Rape >Assault >Kidnap >Robbery
2	Monday	Murder >Body Found >Rape >Assault >Robbery >Kidnap
3	Thursday	Body Found >Murder >Rape >Assault >Robbery >Kidnap
4	Saturday	Body Found >Murder >Rape >Assault >Kidnap >Robbery

Crime Frequency Order on Weekdays

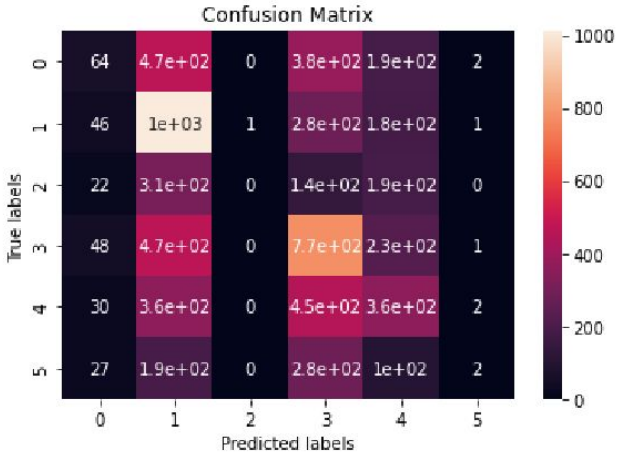
Feature Extraction

Distance: Incident area, District, Division

- Distance between the incident area and the corresponding District city was added to the dataset.
- Distances between the District city and the corresponding divisional city were introduced to the dataset.
- Haversine formula was used to calculate distance from latitude and longitude.

Result and Performance Analysis

Logistic Regression



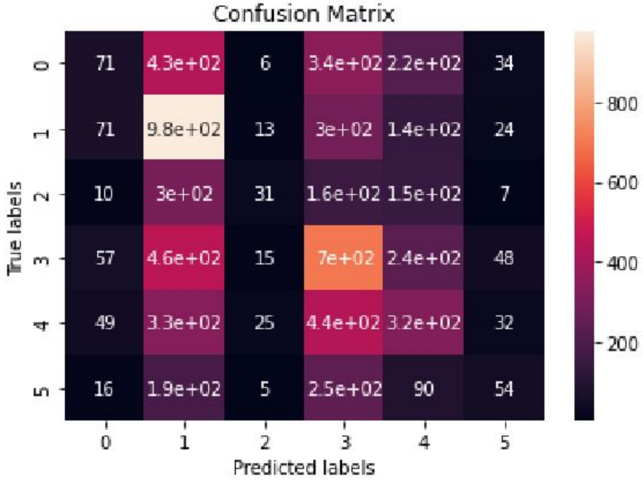
Confusion Matrix for Logistic Regression

Crime	Precision	Recall	F1	Accoracy
Assault	0.27	0.06	0.10	37.97
Body Found	0.36	0.66	.047	
Kidnap	0.00	0.00	0.00	
Murder	0.33	0.51	0.40	
Rape	0.29	0.30	0.30	
Robbery	0.25	0.00	0.01	

Performance of Logistic Regression

Result and Performance Analysis

Naive Bayes



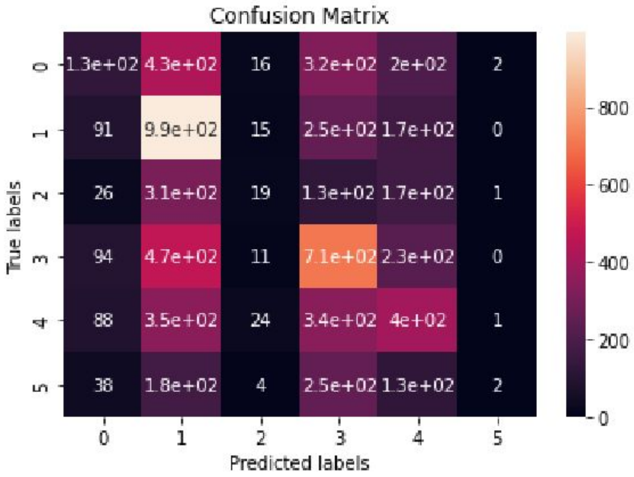
Confusion Matrix for Naive Bayes

Crime	Precision	Recall	F1	Accuracy
Assault	0.26	0.06	0.10	36.00
Body Found	0.36	0.64	0.46	
Kidnap	0.33	0.05	0.08	
Murder	0.32	0.46	0.38	
Rape	0.28	0.27	0.27	
Robbery	0.27	0.09	0.14	

Performance of Naive Bayes

Result and Performance Analysis

SVM



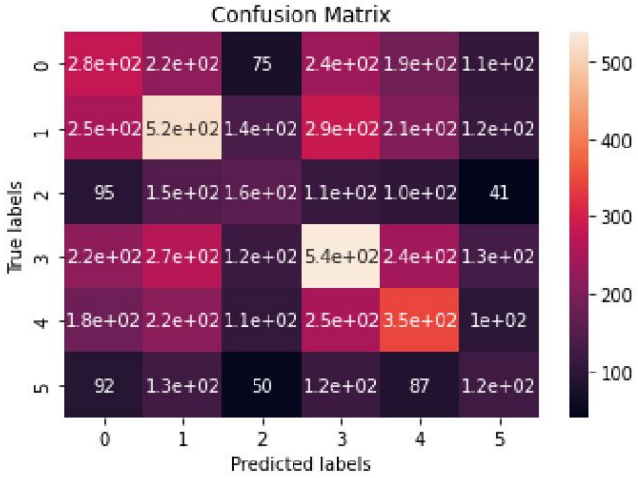
Confusion Matrix for SVM

Crime	Precision	Recall	F1	Accuracy
Assault	0.28	0.12	0.17	
Body Found	0.36	0.65	0.47	
Kidnap	0.21	0.03	0.05	
Murder	0.35	0.47	0.40	
Rape	0.31	0.33	0.32	
Robbery	0.33	0.00	0.01	38.12

Performance of SVM

Result and Performance Analysis

Decision Tree



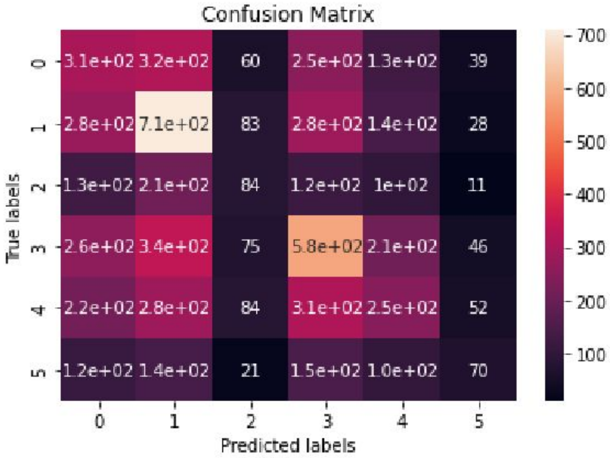
Confusion Matrix for Decision Tree

Crime	Precision	Recall	F1	Accuracy
Assault	0.25	0.25	0.25	34.74
Body Found	0.36	0.35	0.36	
Kidnap	0.24	0.24	0.24	
Murder	0.35	0.35	0.35	
Rape	0.28	0.28	0.28	
Robbery	0.20	0.22	0.21	

Performance of Decision Tree

Result and Performance Analysis

KNN



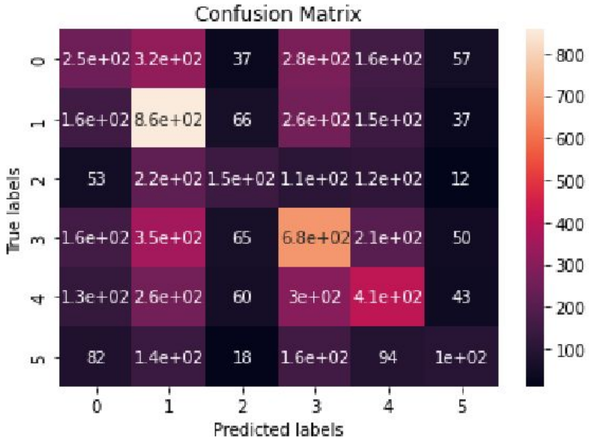
Crime	Precision	Recall	F1	Accuracy
Assault	0.23	0.28	0.25	33.05
Body Found	0.35	0.46	0.40	
Kidnap	0.21	0.13	0.16	
Murder	0.34	0.38	0.36	
Rape	0.26	0.21	0.23	
Robbery	0.28	0.12	0.17	

Confusion Matrix for KNN

Performance of KNN

Result and Performance Analysis

Random Forest



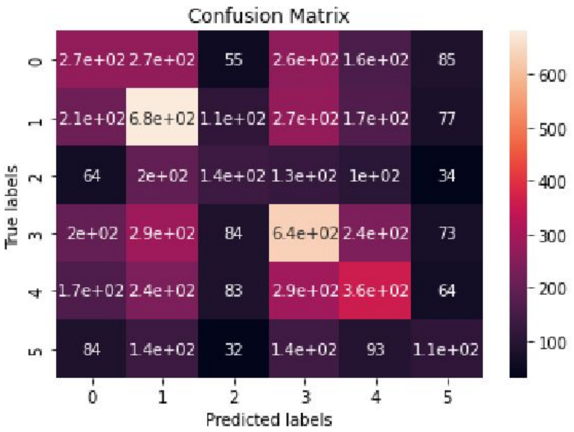
Confusion Matrix for Random Forest

Crime	Precision	Recall	F1	Accuracy
Assault	0.30	0.23	0.26	40.33
Body Found	0.40	0.56	0.46	
Kidnap	0.36	0.21	0.27	
Murder	0.38	0.45	0.41	
Rape	0.34	0.32	0.33	
Robbery	0.33	0.17	0.23	

Performance of Random Forest

Result and Performance Analysis

Extra Tree



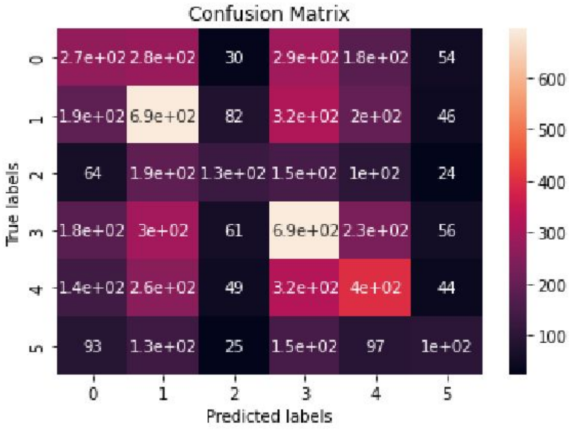
Confusion Matrix for Extra Tree

Crime	Precision	Recall	F1	Accuracy
Assault	0.28	0.26	0.27	36.60
Body Found	0.37	0.43	0.40	
Kidnap	0.28	0.22	0.24	
Murder	0.36	0.42	0.39	
Rape	0.31	0.28	0.29	
Robbery	0.25	0.19	0.21	

Performance of Extra Tree

Result and Performance Analysis

Adaboost



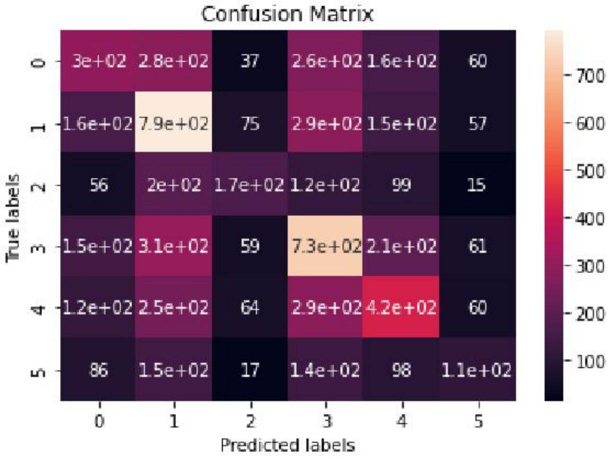
Confusion Matrix for Adaboost

Crime	Precision	Recall	F1	Accuracy
Assault	0.30	0.24	0.27	39.57
Body Found	0.38	0.52	0.44	
Kidnap	0.42	0.17	0.24	
Murder	0.37	0.49	0.42	
Rape	0.37	0.34	0.35	
Robbery	0.37	0.16	0.23	

Performance of Adaboost

Result and Performance Analysis

XGBoost



Confusion Matrix for XGBoost

Crime	Precision	Recall	F1	Accuracy
Assault	0.34	0.27	0.30	41.50
Body Found	0.40	0.52	0.45	
Kidnap	0.40	0.26	0.32	
Murder	0.40	0.48	0.44	
Rape	0.37	0.35	0.36	
Robbery	0.30	0.18	0.23	

Performance of XGBoost

Result and Performance Analysis

Comparative analysis

Algorithm	Accuracy
Logistic Regression	37.97
Naive Bayes	36.00
SVM	38.12
Decision Tree	34.74
KNN	33.05
Random Forest	40.33
Extra Tree	36.60
AdaBoost	39.57
XGBoost	41.50

Conclusion and Future Works

Limitations

- Only 6600 criminal records in this dataset.
- Only those crimes that were reported in the newspaper were gathered.
- It was difficult to collect socioeconomic data.
- The dataset is imbalanced. From 2019 to 2012, there were approximately 600 kidnap and robbery criminal records available. However, around 1500 murder criminal records were added during the same time period.

Conclusion and Future Works

Future Works

- Collection of more criminal records in this dataset.
- Inclusion of human trafficking, narcotics, smuggling, and other criminal records.
- Fine-tuning and more customization of applied models.
- Exploration of algorithms like CatBoost, LogitBoost, LGBM, Bagging, etc.
- Handling missing values in other ways, such as by clustering them or making predictions based on dataset.
- Balancing the dataset with oversampling and undersampling techniques.

Recommendation

Closing

Thank You