# 

**IS424 Data Mining and Business Analytics**

**Project Proposal**

**Title: Predictive Analysis of Crime Hotspots for Enhanced Resource Allocation**

**Section : G3**

**Group : Team 8**

**Team Members :**

* **Bryan Bramaskara**
* **Nanda Gian Yong Xin**
* **Nguyen Tuan Minh**
* **Nguyen Vu Truong An**
* **Sebastian Hong Kai Xuan**

**Table of Content**

[**1. Project Overview**](#_nnquehhvps0j) **3**

[1.1 Problem Statement](#_tmztv275vjrm) 3

[1.2 Motivation](#_jg2vbzmbkskk) 3

[**2. Literature Review**](#_z90wrxwg4nnf) **4**

[**2.1 Source 1: Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures (Al-Janabi, Kadhim & Fatlawi, Hayder, 2010)**](#_z83lgue0ixxl) **4**

[**2.2 Source 2: A Review of Data Mining Applications in Crime (Hassani, Hossein & Huang, Xu & Silva, Emmanuel & Ghodsi, Mansi, 2016)**](#_hwg5z0ha8lkb) **4**

[**2.3 Source 3: An Adaptive Approach For Analyzing And Predicting The Crime Location (R, S., & D, E., 2017)**](#_smytk1agvska) **4**

[**3. Datasets**](#_d0k5q3yrrmlq) **4**

[3.1 Crime Dataset](#_tu6tib7rl943) 4

[**4. Tools and Resources**](#_z22n1yuocnzz) **5**

[**5. Preliminary Results and Evaluation**](#_6ab3h2hi2os3) **5**

[**6. Project Milestones**](#_m865oku8rnh3) **6**

[**7. References (APA 7 Format)**](#_99f7fl59o9rb) **7**

# 

# **1. Project Overview**

The various communities in the US have diverse demographic makeups. This project aims to do an in-depth analysis of the characteristics of the residents of the communities i.e. social economic data and law enforcement resource allocation, and predict the potential emergence of crime through ensemble learning methods.

## 1.1 Problem Statement

Through utilizing various data mining techniques, we want to take on the challenge of manipulating the crime dataset to discover embedded patterns in different types of attributes present. This enables insight generation to form the basis of predictive modelling so that there can be better allocation of manpower and resources to help in lowering crime rates.

Below are a few tasks which we have defined when exploring this dataset (non-exhaustive list):

* To understand the contributing factors that affect crime rate and study crime patterns.
* To discover inter and intra-relationships present in individual crime records by mapping them to their respective locations based on key factors..
* Predict how likely a crime will occur in a particular area based on relevant attributes (Predictive Policing).
* Prescript a better allocation of manpower in areas that have higher crime rates to help in crime prevention.

## 1.2 Motivation

Looking at global crime rates, the Americas have a higher rate than the global average, with the US prison population making up more than 20% of the global prison population, in spite of the fact that the US population accounts for less than 5% of the global population (Statistica Research Department, 2020). These offenders come from various backgrounds and have different crime scenes. With such variability in crime records, there is a need to uncover hidden patterns and use them to identify areas prone to criminal activity to prevent crimes from occurring (“Predictive Policing: Stopping Crime Before it Starts”, 2020).

Although there has been rigorous work dedicated to improve the accuracy of predictive policing, several drawbacks have been identified, such as being overly data-driven so much so that accuracy is compromised. Hence, our project aims to address this drawback by using data mining techniques and methodologies to inculcate context into our predictions. This way, we hope to improve resource deployment and reduce crimes, which will translate into lower costs for the police.

# **2. Literature Review**

## 2.1 Source 1: *Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures (Al-Janabi, Kadhim & Fatlawi, Hayder, 2010)*

This conference paper puts forward a data warehouse framework for crime data analysis and detection using different data mining techniques, such as Classification, Prediction, Link Analysis and Association. The paper aims to assist in making criminal forecasts, finding reasons and relationships behind crimes, and mapping criminal networks. We can leverage on the Data Mart architecture highlighted in this paper to further explore the crime dataset thoroughly.

## 2.2 Source 2: *A Review of Data Mining Applications in Crime (Hassani, Hossein & Huang, Xu & Silva, Emmanuel & Ghodsi, Mansi, 2016)*

This paper presents a concise review and efficiency of various data mining techniques that have been implemented for crime analysis. It offers insights on data mining techniques most frequently adopted for crime analysis, against the backdrop of Big Data. With reference to this paper, we can apprise ourselves on research that has already been carried out and use it to further our understanding of the crime dataset, along with rigorous analysis.

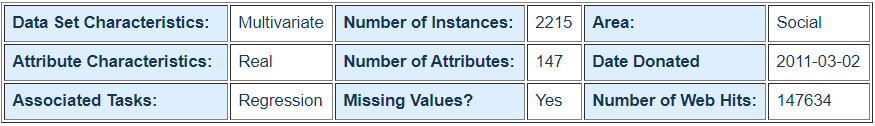
## 2.3 Source 3: *An Adaptive Approach For Analyzing And Predicting The Crime Location (R, S., & D, E., 2017)*

This paper divides data mining approaches commonly employed in crime analysis into different categories and provides in-depth explanation of each. It lists out the different approaches used in past researches and also includes expert opinions on some methodologies. From this, we can use it to avoid pitfalls in our own research and draw on insights stated in this paper for novel ways to explore the dataset by using methodologies not explicitly touched on in class.

# **3. Datasets**

## 3.1 Crime Dataset

The dataset for this project is retrieved from the following link: <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>. It describes the distribution of crime rates in different states in the US, along with attributes covering socioeconomic status such as median income, and the number of violent crimes per population. The table below illustrates some basic characteristics of the dataset:



*Figure 1: Overview of Crime Dataset*

# **4. Tools and Resources**

Our project will use the data mining techniques shared in class, such as Cluster Analysis, Random Forest, Outlier Analysis, Regression and more if required. Python packages such as Pandas, NumPy, Scikit-Learn, MatPlotLib etc. will also be used for data handling and data visualization purposes. Other open-source tools that we will utilize include Google Colab and Google Drive to facilitate real-time dissemination of information and codes. Additionally, we will leverage on out-of-class techniques such as the cost matrix to determine the efficiency of our analysis.

# **5. Preliminary Results and Evaluation**

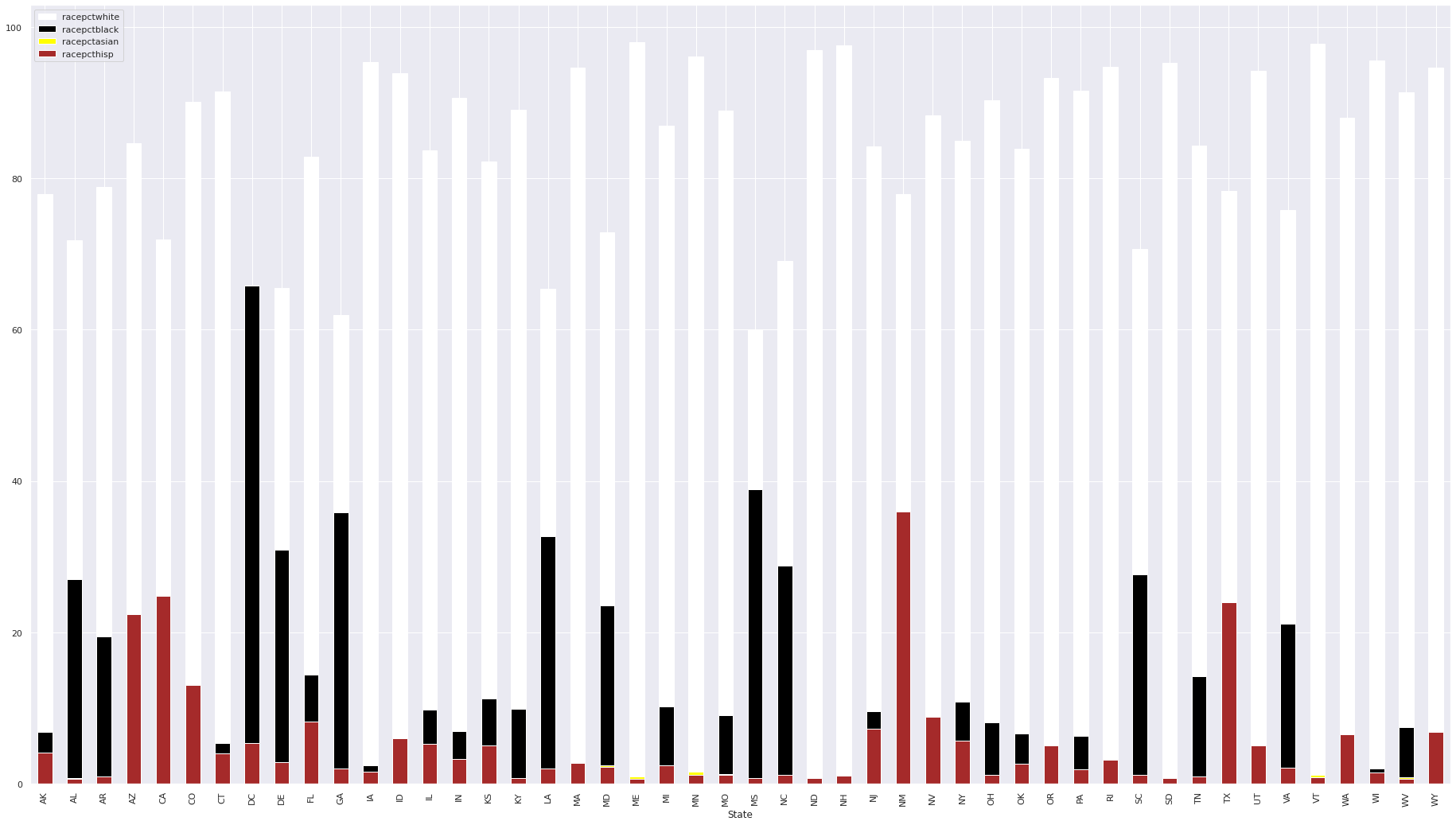
We found a total of 147 columns, with 2215 rows. Of the 147 columns, 6 columns are nominal variables, 138 columns are interval variables, and 3 columns are ordinal variables. We will mainly focus on more relevant interval variables with few missing values, such as race percentage, age groups percentage and income status.

In addition, columns with more than 50% of missing values (see Figure 2) will be excluded from our analysis to ensure a considerable amount of accuracy obtained in our final results.

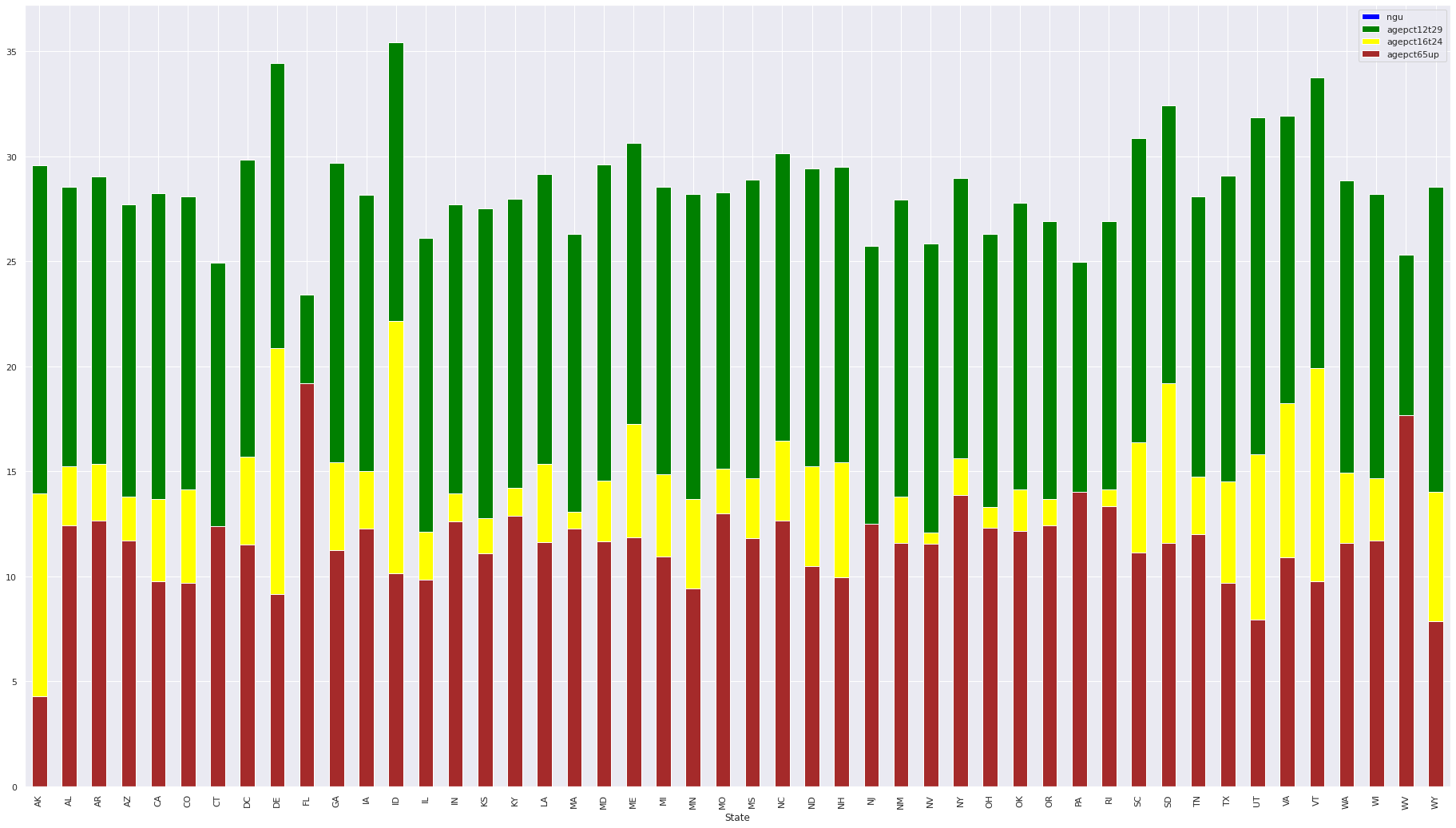
|  |  |
| --- | --- |
| PctPolicBlack: 84.51  CountyCode: 55.12  CommunityCode: 55.26  LemasSwornFT: 84.51  LemasSwFTPerPop: 84.51  LemasSwFTFieldOps: 84.51  LemasSwFTFieldPerPop: 84.51  LemasTotalReq: 84.51  LemasTotalReqPerPop: 84.51  PolicReqPerOffic: 84.51  PolicPerPop: 84.51  RacialMatchCommPol: 84.51  PctPolicWhite: 84.51 | PctPolicHisp: 84.51  PctPolicAsian: 84.51  PctPolicMinor: 84.51  OfficAssgnDrugUnits: 84.51  NumKindsDrugsSeiz: 84.51  PolicAveOTWorked: 84.51  PolicCars: 84.51  PolicOperBudg: 84.51  LemasPctPolicOnPatr: 84.51  LemasGangUnitDeploy: 84.51  PolicBudgPerPop: 84.51 |

*Figure 2: Variables With Many Missing Values*

Below are 2 sample visualizations produced by conducting Exploratory Data Analysis (EDA) on the dataset in Python Jupyter Notebook, which gives an overview of the dataset that we are handling:



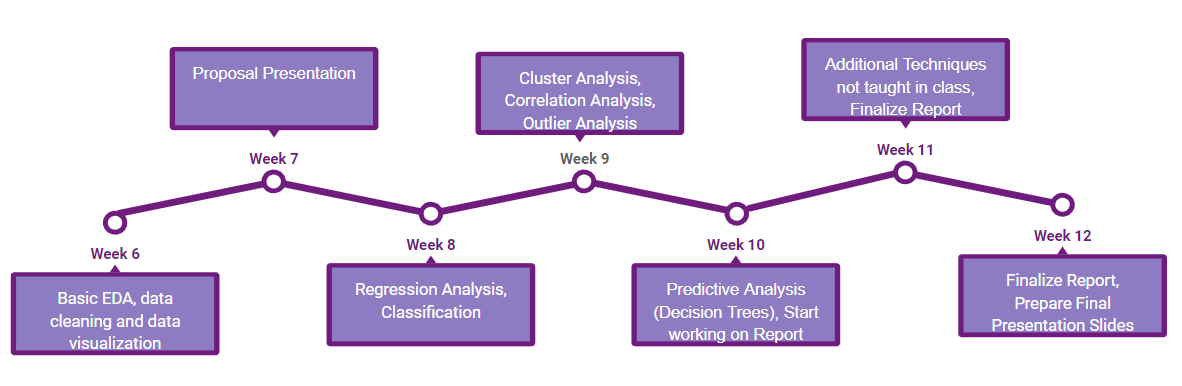
*Figure 3: Population Distribution of Race Groups Across States*



*Figure 4: Population Distribution of Age Groups Across States*

# **6. Project Milestones**

Our team has come up with a project timeline. For reference, our week starts on Mondays and ends on Sundays, hence weekly tasks should be completed by end of week.



*Figure 5: Proposed Project Timeline*

# **7. References (APA 7 Format)**

Hassani, H., Huang, X., Silva, E. S., & Ghodsi, M. (2016). A review of data mining applications in crime. Statistical Analysis and Data Mining: The ASA Data Science Journal, 9(3), 139–154. <https://doi.org/10.1002/sam.11312>

Kaur, S., & Singh, W. (2017). Systematic Review of Crime Data Mining, 8(5), 1336–1342.

R, S., & D, E. (2017). An Adaptive Approach For Analyzing And Predicting The Crime Location (thesis). Shodhganga, India.

Redmond, M. (2011, March 2). Communities and Crime Unnormalized Data Set. <https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized>.

Sanders, J. (2020, April 3). Predictive Policing: Stopping Crime Before it Starts. Pappalardo & Pappalardo LLP. <http://pappalardolaw.com/2019/02/predictive-policing-stopping-crime/>.

Statistica Research Department. (2020, January 28). Topic: Crime and punishment around the world. Statista. <https://www.statista.com/topics/780/crime/>.

Swadi Al-Janabi, K. B., & Fatlawi, H. K. (2010). Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures.