**DAB 322 CAPSTONE PROJECT 1**

**911 Calls for Service 2021**

A logo for a college

Description automatically generated

**Section-A**

**Submitted to**

**Prepared For: Prof. Savita Seharawat**

**Submitted By**

**Navdeep Singh (0817650)**

**Nitika (0820943)**

**Navjot Kaur (0818071)**

**TABLE OF CONTENTS**

1. Abstract
2. Keywords
3. Research question
4. GitHub account
5. Introduction
6. Literature review
7. Methodology
8. Data details
9. Data pre-processing
10. Normalization
11. conclusion

**Abstract**

"911 Calls for Service 2021" is the capstone project. To increase safety, the 911 Calls Capstone Project examined statistics on the frequency of emergency 911 calls. Python with NumPy, Panda's data analysis libraries, together with Matplotlib and seaborn data visualization tools are used to construct this project on a Jupiter notebook. Advanced methods, including machine learning, are used to analyze the gathered data once it has been recorded in a central database. In this research, exploratory data analysis and data visualization are used to determine the frequency of 911 emergency calls throughout the year as well as the most frequent causes and occurrences that lead to 911 calls and when they typically occur.

Problem Understanding:

* Priority Based Approach: There has been a significant number of priority-based calls in the given dataset, which helps us to understand the need for help.
* Incident Location: With a population of 576498 the high population density in Baltimore contributes to increased incidents, increasing the emergency factors.
* Timing patterns: 911 Calls for Service 2021 pose a significant detailed concern for incident timings in Baltimore, leading to location, and other different factors.

The dataset used for the capstone project "911 Calls for Service 2021" is obtained from a data catalogue that specifically focuses on emergency calls data in the Baltimore City (USA). The dataset encompasses information about incidents occurring on several locations. The data is sourced from various law enforcement agencies such as Baltimore City police department,

The dataset provides detailed statistics about each 911 emergency call, including information about the location and incident type. It serves as a valuable resource for analyzing and understanding the characteristics of accidents and priority patterns.

The dataset was originally published in the year 2021, and has been periodically updated, with the latest modification made on June 30, 2022. It contains a total of 1048576 samples or instances and comprises 20 different features or variables.

Given the large number of observations, data cleaning and preprocessing steps will be necessary to handle missing values and remove irrelevant columns. The focus will then shift towards selecting specific factors that are likely to have a significant impact on incidents. These factors may include Incident type Priority, Community Statistical Areas, Location.

Questions

1. How many incidents can occur in future years?
2. What are the major reasons for emergency calls?
3. Are there any significant patterns for these emergency calls?
4. At what time do most of the incidents occur?
5. Predicting the severity of incidents based on different factors?

**Keywords**

Classification.

As our data is categorical, we must use a classification model in order to analyze the intensity of road rashes. We are going to implement these two models:

* K-nearest neighbor
* Support Vector Machine (SVM)
* Techniques & Theme

Following are the tools which we will use to perform these tasks:

* Python
* For visualization Tableau or Power Bi or MS Excel

**Introduction**

Emergency 911 services are essential for ensuring everyone's safety and saving lives. They were founded in 1968 and offer people a straightforward three-digit number to call for immediate aid in times of need. The system has developed through time and is now the go-to resource for reporting emergencies including accidents, fires, crimes, and medical situations. The difficulties that emergency dispatchers encounter and the significance of responding quickly and allocating resources wisely are emphasised. In many nations, the 911 call system has gained widespread acceptance and reliance, safeguarding the security and welfare of inhabitants in times of emergency.

**Literature Review**  
Comprehensive Analysis of Existing Research Related to 911 Call Centers:  
Numerous research studies have been conducted to understand the functioning and challenges of 911 call centers. These studies have explored various aspects, including call volume patterns, response times, call categorization, and dispatcher training. Call Volume Patterns: Studies have identified peak hours and days for emergency calls, allowing call centers to anticipate and prepare for high call volumes during specific periods. Call Categorization: Researchers have investigated the use of artificial intelligence and natural language processing to automate call categorization, streamlining dispatch and resource allocation processes. Public Education and Awareness: Research has emphasized the importance of public education about proper use of 911 services, as inappropriate calls can strain resources and hinder response to genuine emergencies. GPS and Location Tracking: Integration of GPS technology enables accurate location tracking of emergency callers, allowing for quicker and more precise dispatch of resources. Real-time Data Sharing: Technological solutions that enable seamless data sharing between emergency services, such as police, fire departments, and medical responders, facilitate faster and more coordinated responses.  
Identification of Best Practices and Success Stories from Other Regions:  
Comparative studies and success stories from different regions can offer valuable insights into effective emergency response strategies and practices. Regional Collaboration and Consolidation: Some regions have achieved success by consolidating multiple smaller call centers into a centralized hub, leading to improved resource coordination and better utilization.

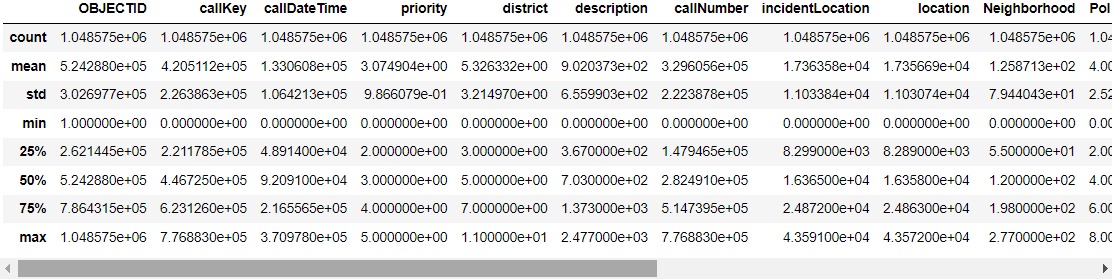
**Methodology**  
I do not currently have access to any datasets as of September 2021, however the following is a common technique for producing and utilising a dataset associated with the 911 call service: Data collecting include locating the sources, gaining legal and ethical access to the data, deciding on the format for the data, and pre-processing, cleaning, normalising, and anonymizing the data. Finding pertinent features, extracting them, analysing the data, and visualising trends are all steps in the process of feature extraction and data analysis. Use case development, model training, and model assessment are all aspects of data utilisation. Data privacy, prejudice reduction, and observing terms of use and data sharing agreements are only a few examples of ethical issues. A dataset should be created and used with caution, keeping ethical and privacy concerns in mind.

**Data Details**

**Categorical Data**

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Description | Number of levels | Names |
| priority | refers to the level of urgency or importance assigned to an emergency call. | 6 | Low  High  Non-emergency  Medium  Out of service  Emergency |
| district | refers to the geographical or administrative division within the emergency response system. | 9 | SW  WD  NE  ND  ED  SD  SE  NW  CD |
| location | refers to the geographical or spatial information associated with the incident reported during a 911 emergency call. | 9 | Eastern  Western  Southern  Northern  Central  Northeastern  Northwestern  Southeastern  Southwestern |

**Numerical Data**



**Data Preprocessing**

**Data Cleaning**: Before start working on data, we have to assure that our data is clean and appropriate to use. For this we have to deal with the missing and duplicated values.

**Step 1**: Duplicated Values: There was duplicated data in our data set. We just check the duplicates it checks for duplicate rows in the entire Data Frame and returns a Boolean Series where True indicates a duplicate row.

**Step 2**: Missing values:- After that we Identify and handle missing values in the dataset. This may involve imputation techniques, removing rows or columns with missing data, or using statistical methods to fill in missing values.

|  |  |
| --- | --- |
| Attribute Name | No. of missing values |
| Census\_Tracts | 10277 |
| VRIZones | 901631 |
| Zipcode | 5906 |

**Step 3:** Drop irrelevant data like 'Census\_Tracts','VRIZones','ZIPCode' from the dataset

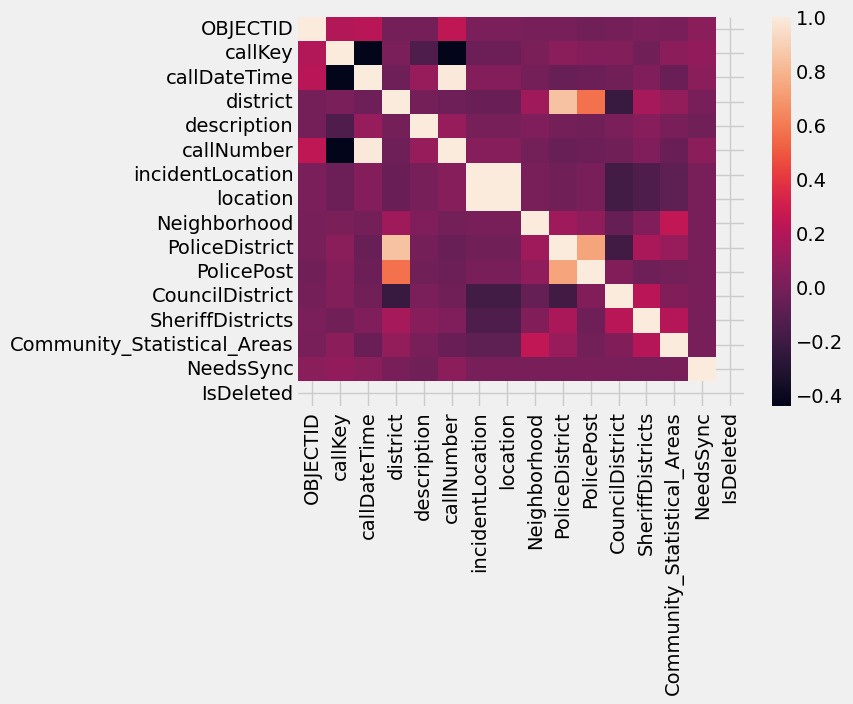
**Step 4:** : Once we dealed with missing data, duplicated data , and irrelevant data we proceeded to next step that is changing categorical data to numerical data. For that

• firstly, we checked data type of our frame.

• Secondly, we started working on it by converting object data to categorical data using As type.

• Lastly, we assigned cat codes to categorical data. To make it understandable.

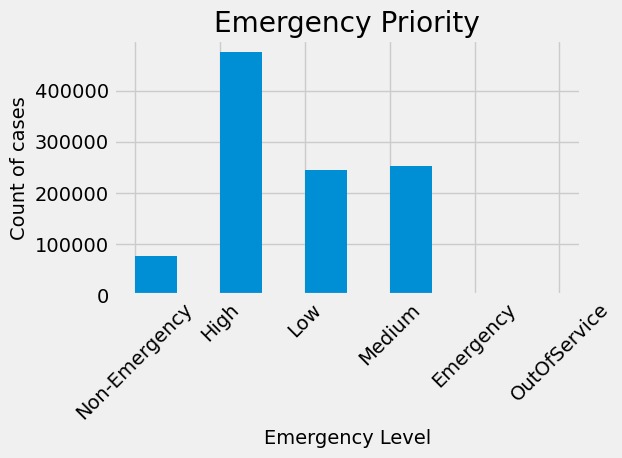
**Step 5:** Correlation Analysis: Import the library seaborn. Check the Co-relation between features using correlation graph:



This correlation gives the relation between variables and shows how the change in one variable impact the others. Its values vary between -1 to 1.

**Exploratory data analysis:**

**Graphical representation of emergency priority**:



**Bar graph of emergency priority**

In this bar graph, the count of cases in emergency level of high more than 400000.in addition, emergency level in high and medium the number of count lies between 2 lakhs to 3 lakhs .in case of non emergency count of cases nearly 80000.

**Normalisation the dataset:**

Normalisation is the process of transforming data into a normal form which implies creating the proper balance in data by mathematical approach. it is the process of transforming numerous types of data into a scale that is comparable.

**Min-Max Scaling**:

It is data normalisation approach that scale data between zero and one. It is simple technique to normalise data using python min-max functions.

**Experimental Design:**

**Cross validation: -**

A technique for testing the efficiency if machine learning model is cross validation. For modeling I implemented some model’s validation techniques as stated below: -

**Train and Test Split Approach (Hold out validation):-**

In this method the entire dataset is randomly portioned into training and test sets. I divided the information into two parts (training and test sets). The Training set contains 80% of the records in the datasets whereas the Test set contains 20% of dataset.

**KNN model: -**

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm used for both classification and regression tasks. Given a new data point, KNN finds the K-nearest data points from the training dataset based on a distance metric (such as Euclidean distance). For classification tasks, the class label of the new data point is determined by taking a majority vote among the K-nearest neighbors' class labels. For regression tasks, the output value for the new data point is determined by taking the average of the K-nearest neighbors' target values. KNN is a simple and intuitive algorithm but may suffer from high computational costs, especially for large datasets.

**Decision tree model: -**

A decision tree model is a predictive model that uses a decision tree algorithm for solving classification and regression tasks. The decision tree model is built using a tree-like structure, where each internal node represents a decision based on a specific attribute, each branch represents an outcome of that decision, and each leaf node represents the final prediction or target value.

**Confusion matrix: -**

In machine learning, a confusion matrix is a common evaluation tool used for assessing the performance of a classification model. It is especially useful in binary classification tasks, where the goal is to predict between two classes: positive and negative. The confusion matrix compares the model's predictions with the actual ground truth labels to measure the model's accuracy and its ability to correctly classify instances into their respective classes.

**Model implementation and evaluation**

I employed a variety of predictive modeling techniques on our dataset. The purpose of study is to figure out which emergency case needs high priority of help by showing the high incident no. in the particular class and we observed that after applying two modelling techniques than which emergency case required topmost priority.

**Accuracy Matrix:-**

|  |  |
| --- | --- |
| MODEL | ACCURACY |
| KNN | 37.0 |
| DECISION TREE | 94.0 |

The best model was shown is a decision tree method in the above table.

**Classification report:-**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LEVELS | PRECISION | RECALL | f-1 SCORE | SUPPORT |
| 0 | **0.94** | **0.97** | **0.96** | **35** |
| 1 | **0.86** | **0.86** | **0.86** | **15028** |
| 2 | **0.92** | **0.92** | **0.92** | **48549** |
| 3 | **0.90** | **0.90** | **0.90** | **50932** |
| 4 | **0.98** | **0.90** | **0.90** | **95122** |
| 5 | **0.90** | **0.76** | **0.82** | **49** |

**Conclusion**

The decision tree model demonstrates promising results, but there is potential for refinement. Continuously monitoring and improving the model's performance is crucial, especially in real-life scenarios where the accuracy of emergency predictions can significantly impact people's lives and safety.