A logo for a university

Description automatically generated

**Applications of Data Science - A**

**Project Proposal**

**Project Title:**

Crime Hotspot Predictor and Safety Application

**Objective:**

To develop a comprehensive crime analysis and prevention solution to predict, monitor, and mitigate criminal activities, thereby enhancing public safety and aiding law enforcement efforts.

**Problem Statement:**

The increasing rate of crimes necessitates an intelligent system that not only detects crime hotspots but also predicts potential criminal activities. A solution is required to provide real-time alerts, personalized safety advice, and actionable insights for law enforcement and the public.

**Brazil Crime Dataset:**

This dataset contains structured data about all crime occurrences that have been acted upon by the PM, the main police force in Sao Paulo. The dataset is not consistent in its completeness, as some of the towns comprising the Greater Sao Paulo were slow in collecting full data. It also does not contain the actual historic of each crime report, as that would violate privacy.

1. **ID\_DELEGACIA**: Code of the police station registering the incident.
2. **NOME\_DEPARTAMENTO**: Department responsible for registering the incident.
3. **NOME\_SECCIONAL**: Sectional police station handling the case.
4. **NOME\_DELEGACIA**: Specific police station recording the incident.
5. **CIDADE**: City where the crime happened.
6. **ANO\_BO**: Year of the crime.
7. **NUM\_BO**: Incident report number (BO).
8. **NOME\_DEPARTAMENTO\_CIRC**: Jurisdiction department.
9. **NOME\_SECCIONAL\_CIRC**: Jurisdiction sectional police station.
10. **NOME\_DELEGACIA\_CIRC**: Jurisdiction-specific police station.
11. **NOME\_MUNICIPIO\_CIRC**: Municipality under jurisdiction.
12. **DESCR\_TIPO\_BO**: Type of incident document.
13. **DATA\_OCORRENCIA\_BO**: Date of the crime.
14. **HORA\_OCORRENCIA\_BO**: Time of the crime.
15. **DATAHORA\_COMUNICACAO\_BO**: Date and time of registering the crime.
16. **FLAG\_STATUS**: Status of the case (open, solved, ongoing).
17. **RUBRICA**: Legal classification of the incident.
18. **DESCR\_CONDUTA**: Action or behavior during the crime.
19. **DESDOBRAMENTO**: Developments or follow-ups of the case.
20. **DESCR\_TIPOLOCAL**: Type of crime location.
21. **DESCR\_SUBTIPOLOCAL**: Sub-category of the crime location.
22. **LOGRADOURO**: Street where the crime occurred.
23. **NUMERO\_LOGRADOURO**: Street number where the crime happened.
24. **LATITUDE**: Latitude of the crime location.
25. **LONGITUDE**: Longitude of the crime location.
26. **DESCR\_TIPO\_PESSOA**: Role of the person involved (e.g., victim, suspect).
27. **FLAG\_VITIMA\_FATAL**: Whether the victim died.
28. **SEXO\_PESSOA**: Gender of the involved person.
29. **IDADE\_PESSOA**: Age of the involved person.
30. **COR\_CUTIS**: Skin color of the involved person

**Proposed Solutions:**

#### **1. Detecting Crime Hotspots**

* **What it does:** Identifies locations where crimes are most frequent by analyzing geographical data.
* **How:** K-Means Clustering, that group location points into clusters based on their geographical proximity.
* **Columns used:**
  + **CIDADE** (City)
  + **LATITUDE** (Latitude)
  + **LONGITUDE** (Longitude)

#### **2. Predicting Crimes**

* **What it does:** Predicts the likelihood of crimes occurring at specific times and locations.
* **How:** Logistic Regression will be used to predict whether a crime will happen at a given time and location or not.
* **Columns used:**
  + **DATA\_OCORRENCIA\_BO** (Crime date)
  + **HORA\_OCORRENCIA\_BO** (Crime time)
  + **DESCR\_TIPO\_BO** (Type of crime)
  + **CIDADE** (City)

#### **3. Analyzing Crime Trends**

* **What it does:** Identifies patterns such as peak times for crimes or seasonal trends, helping to understand when crimes are more likely to occur.
* **How:** Uses **Time-Series Analysis** to forecast and analyze crime data over time.
  + **ARIMA**: if your data is stationary (or can be made stationary) and has linear trends.
  + **LSTM**: if your data has complex, non-linear trends or long-term dependencies that ARIMA cannot capture well.
* **Columns used:**
  + **DATA\_OCORRENCIA\_BO** (Crime date),
  + **HORA\_OCORRENCIA\_BO** (Crime time).

#### **4. Understanding Demographics of Crimes**

* **What it does:** Identifies which age groups, genders, or ethnicities are more affected by crimes.
* **How:** Uses **Descriptive Statistics** and **Chi-Square Tests**.
  + **Descriptive Statistics** will summarize the data.
  + **Chi-Square Test** will check if there is a **significant relationship** between demographic factors (like gender, age, and ethnicity) and the occurrence of crimes. It helps determine if the distribution of crimes is related to a specific gender or ethnicity, or if it is just random.
* **Columns used:**
  + **SEXO\_PESSOA** (Gender),
  + **IDADE\_PESSOA** (Age),
  + **COR\_CUTIS** (Ethnicity).

#### **5. Assessing Location Risk**

* **What it does:** Rates places based on how safe they are by identifying patterns in crime occurrence at different locations.
* **How:** **Clustering** techniques like **K-Means** will group locations with similar crime patterns. The algorithm looks at the types of locations (e.g., residential, commercial) and their subtypes to form clusters. Based on the clusters, we can assign a "risk score" to each location, indicating how safe or risky it is.
* **Columns used:**
  + **DESCR\_TIPOLOCAL** (Type of location),
  + **DESCR\_SUBTIPOLOCAL** (Subtype of location).

#### **6. Tracking Crime Movement**

* **What it does:** Maps how crimes spread geographically, showing patterns of crime movement across different regions.
* **How:** Uses **Heatmaps**.
  + **Heatmaps** will be used to visualize the density of crimes across various areas. The data will be plotted based on department, sectional, and police station locations, showing areas with higher crime rates.
  + The heatmap will represent crime frequency, allowing us to see how crimes are concentrated in specific regions and how they spread over time.
* **Columns used:**
  + **NOME\_DEPARTAMENTO** (Department name),
  + **NOME\_SECCIONAL** (Sectional name),
  + **NOME\_DELEGACIA** (Police station).

#### **7. Monitoring Crime Resolution**

* **What it does:** Tracks how quickly crimes are solved.
* **How:** Uses **Logistic Regression**.
* **Columns used:**
  + **FLAG\_STATUS** (Case status: solved or unsolved).

#### **8. Ranking Cities by Safety**

* **What it does:** This model ranks cities based on their safety levels by calculating a safety score. The safer cities are ranked higher, while cities with more reported crimes receive a lower ranking.
* **How:** The safety score is determined through an **Index Calculation**, which considers crime data such as the number of incidents reported in each city during a specific year. Cities with fewer crimes will have higher safety scores, while those with more crimes will have lower scores.
* **Columns used:**
  + **CIDADE** (City),
  + **ANO\_BO** (Year).

**Technical Implementation:**

* **Backend**:
  + **Languages**: Python
  + **Libraries**: Pandas, NumPy, Scikit-learn, Statsmodels, Keras/TensorFlow
  + **Machine Learning**: Classification (Logistic Regression, Random Forest), Clustering (K-Means), Time-Series Analysis (ARIMA, LSTM)
* **Visualization**:
  + **Libraries**: Matplotlib, Seaborn
  + **Geospatial**: Google Maps API for mapping crime hotspots and movement patterns
* **Real-Time Data**:
  + **Integration**: APIs for live crime data updates, such as from local law enforcement databases or news sources
* **Frontend**:
  + **Mobile App**: Flutter for creating a user-friendly mobile app interface providing real-time crime data, alerts, and safety recommendations.

References: https://www.kaggle.com/datasets/inquisitivecrow/crime-data-in-brazil