**Final Project**

**Development a Predictive Model for Identifying Future Crime Hotspots Using Historical and Geospatial Crime Data**

Business Understanding Report

Supervisor: Yoav Ziv

Ori Fogel 315729293

Iris Birman 208111377

14/12/2024

The project will focus on developing a predictive model aimed at identifying potential locations for future crime hotspots based on historical crime patterns. The model will enable predictions of both the locations of future crime incidents and the expected types of crime, using geographic, temporal, and other relevant data.

# Background

## Determining Business Objectives

The primary business goal of this project is to leverage data science to improve crime prevention strategies by predicting potential future crime hotspots and types of crimes.

This involves addressing key challenges, such as identifying high-risk crime areas, optimizing resource allocation, and improving the efficiency of crime prevention efforts. Currently, crime prevention lacks data-driven insights, leading to inefficient resource use and strain on manpower.

The primary goal is to develop a predictive model for identifying crime hotspots using historical and geospatial data. This will enable proactive crime prevention, optimized resource allocation, and hopefully even improve public safety. The project is not affiliated with any specific organization or law enforcement agency but aims to demonstrate how data science can enhance crime prevention efforts at a broader level.

To support the project’s goals, several key resources will be utilized, including high-quality crime datasets from the Los Angeles Police Department (LAPD) uploaded to the data.gov website, geospatial analysis tools, and machine learning methods such as Density-Based Clustering, KNN, and Decision Tree. Although no direct collaboration with law enforcement agencies is part of the project, the findings and methodologies can contribute to broader discussions about the role of data-driven approaches in crime prevention.

The project’s objectives focus on improving crime prevention, operational efficiency, and demonstrating the use of data-based information for informed decision-making, even without direct involvement from law enforcement bodies.

## Business Background

**Problem Area Description**

This project aims to develop a predictive model to identify potential future crime hotspots and predict crime types based on historical data. By analyzing geographic and time-based data, the model will help optimize resource allocation and improve public safety efforts. Law enforcement agencies often face challenges in crime prevention due to a lack of data-driven insights, resulting in inefficient resource allocation, such as over-deployment in low-risk areas and insufficient coverage in high-risk zones. By leveraging predictive analytics, this model seeks to demonstrate how to enhance crime prevention strategies, improve operational efficiency, and reduce crime rates.

**Prerequisites for the Project:**

1. **High-quality data** – Access to publicly available historical crime data, such as the LAPD dataset, including crime locations, dates, and types.
2. **Geospatial and machine learning tools** – Algorithms such as DBSCAN for clustering crime hotspots, KNN for evaluating crime likelihood, Decision Trees for classifying crime types, and Logistic Regression for spatial-temporal analysis.
3. **Technological infrastructure** – Sufficient computational resources and platforms to handle large datasets and implement machine learning models effectively.

**Motivations Behind the Project:** The project is motivated by the need to improve public safety and optimize resource allocation in law enforcement. Predicting crime hotspots and identifying crime patterns can enable proactive crime prevention, reduce crime rates, and enhance workforce efficiency. This data-driven approach ensures that resources are deployed where they are most needed, increasing the overall effectiveness of security efforts.

**Current Use of Data Science:** While data science is not yet widely adopted in law enforcement, the potential for its application is clear. This project aims to serve as a foundational step in demonstrating the value of predictive analytics and machine learning for public safety. By showing the impact of data-driven decision-making, the project can inspire further integration of these technologies in crime prevention strategies.

**Status of the Data Science Project:** As this project is not part of an ongoing initiative within a specific organization, it has not been officially approved or adopted. However, its findings could be broadly applied to enhance crime prevention strategies in law enforcement agencies. If needed, informational presentations will be prepared to highlight the potential benefits of data science in improving public safety and guiding strategic decision-making.

# Business Objectives and Success Criteria

## Business Objectives

The primary objective of this project is to enable law enforcement to optimize resource allocation and improve public safety efforts. By developing a predictive model based on historical and geospatial data, the project aims to address inefficiencies in current crime prevention strategies and reduce overall crime rates.

**Problem to Solve:**  
Law enforcement agencies currently allocate resources reactively, often relying on generalized historical trends rather than predictive frameworks. This results in inefficiencies, such as over-policing low-risk areas and under-resourcing high-risk zones. The goal is to create a proactive and data-driven approach that allows for smarter deployment of resources, reducing inefficiencies and improving public safety outcomes.

**Business Questions:**

1. What are the geographic areas most at risk for future criminal activity?
2. What types of crimes are likely to occur in specific areas during specific times?
3. How can law enforcement agencies improve resource allocation to maximize coverage in high-risk areas while maintaining efficiency?
4. How can data-driven insights reduce the overall crime rate and improve public safety?

**Business Requirements:**

1. Ensure the predictive model provides actionable and accurate insights that law enforcement can easily interpret and apply.
2. Integrate the model into existing workflows with minimal technological adjustments.
3. Align the solution with ethical standards and data protection regulations.

**Expected** **Benefits in Business Terms:**

1. Optimize law enforcement resource allocation for better coverage in high-risk areas.
2. Reduce crime rates through proactive and targeted prevention strategies.
3. Enhance workforce efficiency by directing efforts to areas and times of highest crime risk.
4. Improve public safety and strengthen community trust in law enforcement.

These objectives and requirements will guide the project’s development, ensuring its alignment with the strategic goals of law enforcement and its potential to improve public safety outcomes.

## Success Criteria

**Objective Criteria:**

Achieve a measurable improvement in resource allocation efficiency (e.g., at least 15%).

1. Increase the accuracy of predictive models to 85% or higher, based on validation metrics like precision and recall.
2. Reduce the time required for decision-making processes by automating key analytical tasks.
3. Provide clear and actionable insights through visualizations and reports.

**Subjective**:

1. Ensure that the predictive model provides meaningful and interpretable insights for optimizing resource allocation and decision-making.
2. Confirm that the model supports broader goals, such as showcasing the potential of data science in improving efficiency in improving public safety outcomes.

# Assessing the Situation

# The project will utilize publicly available LAPD datasets (2020–present) containing crime types, locations, dates, and times, alongside geospatial data for predictive modelling. The project team consists of two students with a focus on data science and machine learning. The project will not involve direct collaboration with law enforcement authorities, and the team does not include experts in the field of law enforcement. Instead, the team will rely on publicly available datasets and existing research to inform the project.

# The main risk factors analyzed by us are:

# Risk of missing, inaccurate, or biased data in the LAPD dataset, which may affect model reliability.

# Lack of clarity in the model's results and challenges in integrating them into a system that enables practical application in daily operations.

# Navigating privacy and ethical considerations, particularly when working with sensitive geospatial data related to crime.

# Risk of the model underperforming when scaled to larger datasets or applied in new regions.

# The contingency plans addressing those risks:

# Develop preprocessing steps to clean and validate data, ensuring consistency and addressing any identified biases.

# Use imputation techniques for missing data or consider integrating supplementary datasets if necessary.

# Incorporate explainable AI (XAI) techniques to ensure predictions are clear and actionable for non-technical stakeholders.

# Implement data anonymization techniques and ensure compliance with regulations, such as GDPR (General Data Protection Regulation) or CCPA (California Consumer Privacy Act).

# Design the model to be modular and flexible, allowing for easy integration of additional data sources

# Inventory of Resources

The primary data source for this project is the publicly available crime dataset from the Los Angeles Police Department (LAPD), which is continuously updated from 2020 to the present. This data set includes a total of 984,045 records and contains detailed information about reported crimes. The data includes various types of information, such as dates and times of occurrences, geographical locations (latitude and longitude), descriptions of crime categories and types, victim demographics, and the weapons used. The data is stored in a structured format, with columns such as DR\_NO (incident number), Date Rptd (date reported), Vict Sex (victim's sex), LOCATION, and LAT, LON (geospatial coordinates), among others. The data types include both numerical values (e.g., float64, int64) and categorical data (e.g., object-String), which will allow for various types of analyses, including geographic clustering, trend analysis, and classification of crime types. Access to this dataset is available through data.gov, and no live access to operational databases or data warehouses is required for this project. The data is provided in a CSV format, making it suitable for analysis with tools such as pandas in Python.We do not plan to purchase external data in this stage of project but if it will be necessary in more advanced aspects in favor of bringing processed and complementary data, we will consider it. There are no privacy restrictions preventing access to data.

In Personnel Resources aspects, we don’t have access to business and data experts. We do identify database administrators and other support staff that may be needed such as the LAPD via their connection mail.

# Requirements, Assumptions and Constraints

The primary requirement for this project is to develop a predictive model for identifying potential crime hotspots and forecasting crime types, based on historical data from the LAPD. There are no security or legal restrictions on the data itself, as it is publicly available, but it is essential to comply with all applicable data protection laws, such as GDPR or CCPA, when handling sensitive demographic information. The project schedule is aligned with academic deadlines. The results will be presented in a visual format, such as a dashboard and report, for easier interpretation.

In terms of assumptions, the project assumes that the available crime data from LAPD is accurate and of sufficient quality for analysis. It also assumes that the stakeholders, how will be interested in the model's results, are primarily interested in viewing the predictive model's outcomes, rather than understanding the intricacies of the model itself. Economic factors, such as potential costs related to software tools or computing resources, will have an impact on the project budget, but no significant external costs are anticipated at this stage.

Regarding constraints, access to the LAPD dataset is already in place, and no passwords or additional access credentials are required. There are no legal constraints related to data usage, as the data is publicly available.

# Risks and Contingencies

Several risks could potentially impact the success of the project.

**Scheduling risks** include the possibility that the project may take longer than anticipated due to unforeseen technical challenges, such as issues with data preprocessing or model complexity. To mitigate this, we will build buffer time into the project timeline and regularly assess progress against milestones.

**Financial risks** could arise if there are unexpected costs, such as the need for additional computational resources or software licenses. To address this, we will ensure that the project operates within its defined budget and explore free or low-cost alternatives if necessary.

**Data risks** involve the possibility that the data might be of poor quality, incomplete, or not as representative as expected, which could hinder the model's accuracy. In response, we will perform thorough data cleaning and validation, and, if necessary, supplement the data with publicly available external datasets to improve coverage.

**Results risks** include the scenario where the initial results are less impactful or less accurate than expected. To manage this, we will adjust the model based on iterative testing and feedback, and ensure that the results are communicated clearly, with recommendations for further refinement.

Each risk will be monitored regularly, and the team will remain agile to adjust strategies as needed.

# Terminology

To ensure clear communication, here is our key terms that require clarification for this project:

# Crime Hotspot: A geographical area identified as having a higher-than-average rate of criminal activity. In the context of this project, it refers to locations where crimes are most likely to occur based on historical data.

# Predictive Model: A machine learning model that uses historical data to make predictions about future events or outcomes. In this project, it refers to a model that forecasts future crime hotspots and types of crimes.

# Geospatial Analysis: The process of analyzing geographic data to identify patterns or trends. In this project, it involves the use of location data (e.g., latitude and longitude) to understand crime patterns across different areas.

# Accuracy: In machine learning, it refers to the percentage of correct predictions made by the model. It is a measure of the model's overall performance.

# Precision: A metric in machine learning that evaluates how many of the predicted positive outcomes (such as predicting a crime) are true positive cases. This is particularly important in identifying crime hotspots.

# Recall: Another metric that refers to the ability of the model to correctly identify all actual positive instances (e.g., identifying all real crime hotspots). It is important for understanding the model’s sensitivity.

# Latent Variables: Variables that are not directly observed but are inferred from other data. For example, factors like social conditions may not be explicitly in the dataset but can influence crime rates.

# Model Training: The process of teaching a machine learning model using historical data so that it can learn patterns and make predictions.

# Clustering (DBSCAN): A machine learning technique used to group similar data points together. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is particularly useful for identifying crime hotspots based on spatial data.

# KNN (K-Nearest Neighbours): A machine learning algorithm used to predict the likelihood of an event based on the nearest data points. In this project, it will be used to estimate the likelihood of a crime occurring in a specific area.

# Decision Tree: A machine learning model that splits data into branches to make decisions. In this project, it will be used to classify types of crimes based on input features like location and time.

# Logistic Regression: A statistical method used to model the probability of a binary outcome. In this project, it will be used to analyze time and location factors to predict whether a crime is likely to occur at a specific time and place.

# Costs and Benefits

In assessing the costs and benefits of the project, we must consider several factors.

**Costs**: The costs of this project may include acquiring additional data if needed, as well as hardware or software resources. In the future, we may integrate the predictive model from this project into law enforcement systems, which could lead to additional expenses not covered in this document.

**Benefits:** The primary benefit of this project is the successful achievement of its goal, which is accurately predicting crime hotspots and types of crime to optimize resource allocation and improve public safety efforts. Additionally, the project will provide valuable insights through data exploration, enhancing the understanding of crime patterns and trends. This deeper understanding can lead to long-term improvements in crime prevention strategies, better use of law enforcement resources, and overall operational efficiency.

# Data Science Goals and Success Criteria

## Data Science Goals

## The goal of this data science project is to develop a predictive model for crime prevention, specifically by identifying potential crime hotspots and forecasting the types of crimes likely to occur. This problem can be categorized as both classification (predicting the types of crimes) and clustering (identifying crime hotspots through spatial data). The technical objective is to build a model that can provide reliable predictions with a validity period of three to six months, enabling law enforcement agencies to make data-driven decisions about resource allocation. Desired outcomes include increasing the accuracy of crime predictions by at least 15% compared to current methods.

## Data Science Success Criteria

To assess the success of the predictive model, we will measure both accuracy and performance metrics. For crime type prediction, we will focus on achieving at least 85% accuracy in correctly classifying crimes based on historical data. For hotspot prediction, we will evaluate the precision of identified hotspots by comparing the model’s predictions to actual high-crime areas. Success will be defined by the ability to predict high-crime areas with at least 80% precision, compared to baseline random selection of areas. This will be measured by the proportion of predicted hotspots that correspond with areas having the highest crime density, using metrics like precision, recall, and F1-score.

# Project Plan

The project plan has been discussed and reviewed with all team members involved, ensuring alignment on the project tasks, resources, and timelines. Time estimates for each phase, including data collection, model development, and deployment, have been included.. Key decision points, such as model evaluation and deployment milestones, are highlighted, ensuring that reviews and adjustments can be made at critical stages. Additionally, phases like model training and refinement have been marked as iterative, acknowledging the need for multiple cycles of testing and improvement to achieve optimal results.

## The project stages and schedule

# Data understanding- 15/12/24-31/01/25

# Data preparation- 01/02/25-24/03/25

# Modelling- 25/03/25- 30/04/25

# Evaluation- 01/05/25-23/05/25

# Deployment-24/05/25-04/08/25

# Initial Assessment of Tools and Techniques

For this project, Python has been selected as the primary data science tool due to its versatility, ease of use, and robust ecosystem of libraries for each phase of the project. Python offers a wide range of tools, including pandas for data manipulation, scikit-learn for machine learning model development, and Matplotlib/Seaborn for data visualization. Additionally, Python integrates well with geospatial libraries such as GeoPandas and Folium, which will be essential for analysing crime data and visualizing potential crime hotspots. In terms of data science techniques, clustering algorithms (e.g., DBSCAN) will be used to identify hotspots, while classification models (e.g., Decision Trees) will help predict crime types. These techniques are appropriate given the nature of the problem and the available data, allowing for both spatial and temporal analysis of crime patterns.