CAPSTONE PROJECT

Final Report

Crime Analysis: Unveiling Patterns and Enhancing Safety in South Africa

Prepared by:

Siyabonga Mnyango Mlambo

202205704

Department of Computer Science & Information Technology

Sol Plaatje University

Submitted to:

Dr Silas Verkijika

Lecturer: Capstone Project

Department of Computer Science & Information Technology

Sol Plaatje University

24-26 Scanlan St, New Park, Kimberley, 8301

03 October 2024

EXECUTIVE SUMMARY

This project aimed to understand and predict crime trends in South Africa using historical crime data. With crime being a major concern for communities and law enforcement, our goal was to develop a reliable system that could help identify where and when crimes are likely to occur.

Key Deliverables and Findings:

- Understanding Crime Patterns: We analyzed crime data from various provinces over the years, revealing that some areas have consistently high crime rates, especially in urban locations. Certain types of crimes, such as violent crimes and property crimes, were identified as the most common.
- Predictive Modeling: Using advanced statistical techniques and machine learning methods, we created models that can forecast future crime occurrences. For example, our models suggest that while crime counts might decline slightly in the coming years, they are expected to stabilize at a high level.
- Visualizing Trends: We developed a Dashboard that allow law enforcement and community leaders to see trends in crime data easily. This helps them understand which areas are at higher risk and which types of crimes are prevalent.

Practical Impact:

The insights gained from this project provide valuable information for law enforcement agencies, enabling them to allocate resources more effectively and develop targeted crime prevention strategies. By understanding where and when crimes are most likely to happen, police and community organizations can work proactively to enhance public safety and improve the overall quality of life in South African communities.

BUSINESS DOMAIN

This predictive crime analysis system is specifically designed for use within the law enforcement and public safety sector. It can be implemented at various levels, including municipal police departments, national law enforcement agencies, and community safety organizations.The main client for this project is the South African Police Service (SAPS). SAPS is responsible for maintaining public order, preventing crime, and ensuring the safety of citizens across South Africa.
The primary user group consists of:

- Law Enforcement Officers: Police officers and detectives who need access to real-time data and insights for effective patrolling and crime prevention strategies.

- Crime Analysts: Specialists who analyze crime data to understand trends, assess risks, and develop strategies for crime reduction. They can leverage the predictive models to identify hotspots and allocate resources efficiently.
- Policymakers: Government officials and community leaders who require data-driven insights to formulate effective public safety policies and interventions.
- Community Organizations: Local community groups focused on crime prevention and safety can utilize the findings from the predictive models to engage with residents, promote safety programs, and foster collaboration with law enforcement.

1. INTRODUCTION

This project aimed at understanding and forecasting crime trends in South Africa. With crime being a significant issue affecting the safety and well-being of communities, the project seeks to provide law enforcement agencies and policymakers with the tools and insights necessary to make informed decisions about crime prevention and resource allocation.

Key Components of the Project:

- The project begins with a comprehensive analysis of historical crime data, which includes various types of crimes, their occurrences across different provinces, and trends over time. This analysis helps identify patterns and factors that influence crime rates.
- Utilizing advanced statistical techniques and machine learning algorithms, such as ARIMA, Random Forest, and XGBoost, the project develops models that can predict future crime occurrences. These models assess various factors, including geographic location, population density, and time of year, to identify high-risk areas.
- The project incorporates data visualization tools which is Power BI that present crime trends of South Africa. This allows stakeholders, including law enforcement officers and community leaders, to quickly grasp the information and take appropriate actions.
- The ultimate goal is to understand where and when crimes are likely to occur, they can implement targeted prevention strategies, allocate resources more effectively, and foster safer communities.

2. PROBLEM AND OBJECTIVES

2.1. Problem Statement

Despite concerted efforts by the South African government and law enforcement agencies, South Africa continues to battle with alarmingly high crime rates. These rates pose significant threats to public safety, hinder socio-economic development, and affect the overall well-being of society. Official crime statistics from the

South African Police Service (SAPS) indicate an increase in household crimes such as housebreaking and home robbery compared to the previous year1 (Maluleke, 2023). Moreover, the crime rate statistics for Page 4 of 14 2021 showed a 23.26% increase from 20202 (Macrotrends, 2024), underscoring the severity of the issue. . Furthermore, the significant gap between reported crimes and actual victimization experiences underscores the importance of integrating victimization surveys and qualitative research methods to capture the full scope of crime and its impact on individuals and communities (Graan, 2021). This project aims to bridge this gap by analyzing a combination of crime statistics, victimization surveys, and qualitative research to provide a comprehensive picture of crime patterns (Faull, 2022). The insights gained will inform the development of targeted crime prevention strategies, ultimately enhancing the safety and well-being of South African communities.

2.2. Business Objectives

- Examine Crime Variations:
    - Analyze variations in crime occurrences across different regions, urban-rural divides, and time periods to understand crime distribution dynamics. Success will be measured by producing a comprehensive report that details trends and patterns in crime data, identifying at least three significant factors influencing these variations within the first four months of the project.
- Identify Common Crime Types:
    - Determine prevalent crime types such as violent, property, and white-collar crimes, and investigate their socio-economic drivers. This will involve qualitative and quantitative analysis of the data, aiming to produce a detailed summary of the most common crime types , it was done within six months of project initiation.
- Implement Innovative Approaches:
    - Explore innovative crime analysis approaches utilizing technology, data analytics, and predictive modeling to gain deeper insights. This includes conducting a feasibility study on at least two advanced analytical techniques (e.g., machine learning algorithms, and implementing them to enhance crime analysis capabilities, it was completed within five months.
- Develop Predictive Model:
    - Create a predictive model to identify high-risk areas for proactive policing. The model will aim for a minimum accuracy of 80% when forecasting crime occurrences and will be deployed to relevant stakeholders within six months. Success was evaluated by comparing the model's predictions against actual crime incidents over a subsequent three-month period.

3. SCOPE

3.1. In-Scope

- Data Sources:
    - The primary data source for this project is the 2022-2023 Annual Crime Statistics Report, specifically from the South African Police Service (SAPS). This report provides comprehensive crime data, including types of crimes reported from 2013 to 2023. The data was extracted from pages 59 to 76 of the PDF document, which can be found here.
    - Additional demographic information, including population data and urban/rural status for each province in South Africa, was obtained from the Government of South Africa's website at Gov. za. This data helps categorize provinces based on whether they are predominantly urban or rural.

- Techniques and Methodologies:

- Exploratory Data Analysis (EDA): To visualize crime trends and identify patterns across different regions and periods.
- Statistical Analysis: Including correlation analysis to examine relationships between crime rates and socio-economic factors (e.g., population density, urban/rural status).
- Machine Learning Techniques:
    - ARIMA: For predict number of crime occurrences in the next 10 years.
    - Random Forest and XGBoost: For classification and regression tasks to predict crime types and counts.
    - Logistic Regression: For binary classification of crime types.
    - Gradient Boosting: For improved prediction accuracy in regression tasks.
- Data Visualization: Create dashboards and visual tools to present trends and predictions in an accessible format.
- Deliverables:

- A comprehensive report detailing findings from the data analysis, including trends in crime occurrences, and predictions generated by the model.
- A predictive model capable of forecasting crime occurrences in different regions.
- A dashboard created to visualize the trend of crime in South Africa, allowing stakeholders to interactively explore crime data and insights.

- Stakeholder Engagement:
    - Regular meeting with the relevant stakeholders to ensure the project meets their needs and incorporates their insights

3.2. Out-of-Scope

- Methods Not Used:

Deep Learning Techniques: Methods like neural networks will not be utilized due to the limited size of the dataset and the focus on interpretability for law enforcement purposes.

Advanced Statistical Techniques: Techniques such as Bayesian analysis or sophisticated econometric models will not be included, as the focus is on more straightforward predictive models.

- Data Sources Not Explored:

Private Crime Data: Data from private security firms or unverified sources will not be included to maintain the integrity and reliability of the analysis.

International Crime Data: The project will be limited to South African data, and comparisons with other countries will not be made.

- Types of Deliverables Not Considered:

The project did not produce real-time crime monitoring systems or applications, as the focus is on analysis and forecasting rather than operational deployment.

Implementation of the predictive model within a live system will not be included; instead, the project will provide recommendations for future implementation.

4 SOLUTION

4.1 Overview of Methodology

The project will employ the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, which is a robust and iterative approach to data science projects.

- Business Understanding:
    - This phase involves clearly defining the project's goals and objectives. The primary aim is to analyzed crime patterns in South Africa and develop effective strategies to prevent crime and enhance public safety.
    - Created a project plan that outlines the project's scope, key deliverables, and timelines to ensure all stakeholders have a clear understanding of the project's direction and milestones.
- Data Understanding:
    - Gather relevant data, including historical crime data from the South African Police Service (SAPS), victimization surveys, and findings from qualitative research to get a comprehensive view of crime dynamics.

    - Conduct exploratory data analysis (EDA) to understand the structure, quality, and types of data collected. This step includes identifying missing values, understanding data distributions, and visualizing trends.

- Data Preparation:
    - Process the collected data to handle missing values, remove duplicates, and ensure consistency in data formats.
    - Convert categorical variables into numerical formats (e.g., one-hot encoding) and normalize or standardize features as necessary for model development.
- Modeling:
    - Choose appropriate modeling techniques such as ARIMA for time series forecasting and machine learning models like Random Forest, XGBoost, and Logistic Regression for classification and regression tasks.
    - Train the selected models using the prepared data and optimize their parameters to improve predictive performance.
- Evaluation:
    - Evaluate the performance of the models using metrics such as accuracy, precision, recall, mean squared error (MSE), and mean absolute error (MAE). This step helps determine if the models meet the project's objectives and identifies any areas for improvement.
- Deployment:
    - Show the crime variations and common crime in each provinces to relevant stakeholders, such as law enforcement agencies, to facilitate data-driven decision-making.
    - Compile a comprehensive report detailing the methodologies used, findings, model performances, and actionable recommendations for law enforcement agencies.
    - Present the findings and insights to stakeholders, emphasizing how the developed strategies can enhance public safety and prevent crime.

4.2 Process Flowchart and Implementation

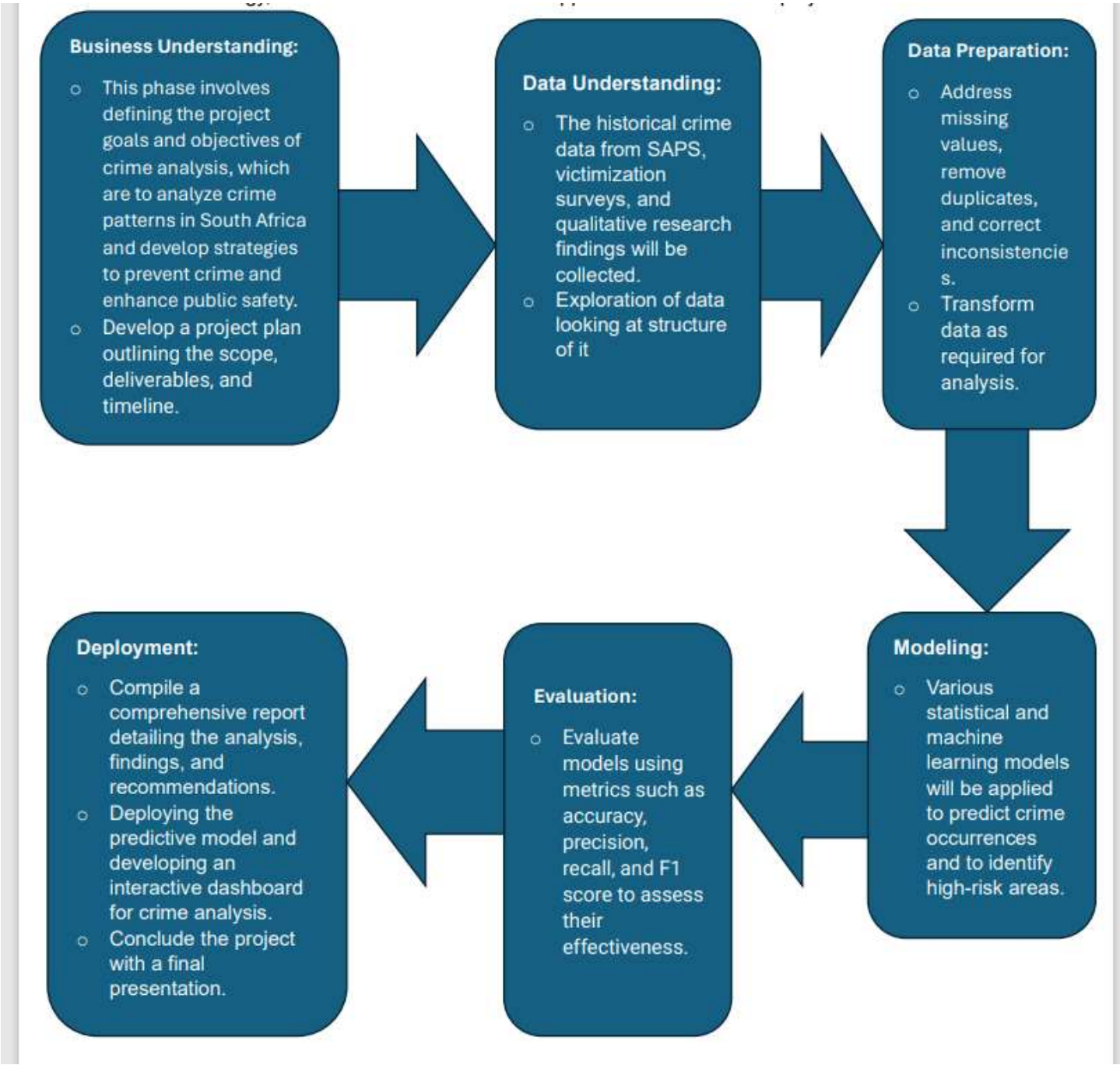


Figure 1: Flowchart of my methodology

Based Figure 1: I did the following

- Business Understanding:
  - Objective Identification: Define the primary goal of analyzing crime patterns to enhance public safety and support law enforcement.
  - Project Planning: Develop a project plan outlining the scope, deliverables, and timeline.
- Data Understanding:
  - Data Collection: Gather historical crime data from SAPS
  - Perform exploratory data analysis (EDA) to understand the nature of the data, and identify patterns.

- Quality Assessment: Evaluate the quality of the data, checking for completeness, accuracy, and consistency.

```
print("Number of Rows", crime_data.shape[0])
print("Number of Columns", crime_data.shape[1])
```

```
Number of Rows 2790
Number of Columns 8
```

```
print("Basic Information of this dataset:")
crime_data.info()
```

```
Basic Information of this dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2790 entries, 0 to 2789
Data columns (total 8 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Type of Crime         2790 non-null   object
 1   Crime Count           2790 non-null   object
 2   Province              2790 non-null   object
 3   Start Year Date       2790 non-null   int64
 4   Urban/Rural           2790 non-null   object
 5   End Year Date         2790 non-null   int64
 6   Crime Category        2790 non-null   object
 7   Province Population   2790 non-null   object
dtypes: int64(2), object(6)
memory usage: 174.5+ KB
```

Figure 2: Looking at the dimensions and information of the dataset.

- Figure 2 shows the number of columns and rows in the dataset, and the names of the columns with the data type.

- Data Preparation:
  - Data Transformation: Normalize, aggregate, or transform data as required for analysis.
  - Feature Engineering: Create new data features that could be significant predictors of crime patterns.

```
def convert_to_numeric_preserving_zeros(column):
    return pd.to_numeric(column, errors='coerce').fillna(0)


crime_data['Crime Count'] = convert_to_numeric_preserving_zeros(crime_data['Crime Count'])
crime_data['Province Population'] = convert_to_numeric_preserving_zeros(crime_data['Province Population'])
```

```
crime_data['Crime Rate'] = (crime_data['Crime Count'] / crime_data['Province Population']) * 100000
```

```
crime_data['Start Year Date'] = pd.to_numeric(crime_data['Start Year Date'], errors='coerce')
```

```
X = crime_data.drop('Crime Count', axis=1)
y_classification = (crime_data['Crime Count'] > 150).astype(int)  # Example binary target for Logistic Regression
y_regression = crime_data['Crime Count']
```

```
categorical_cols = ['Type of Crime', 'Province','Crime Category']
numeric_cols = ['Province Population']

preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), numeric_cols),
        ('cat', OneHotEncoder(), categorical_cols)
    ])
```

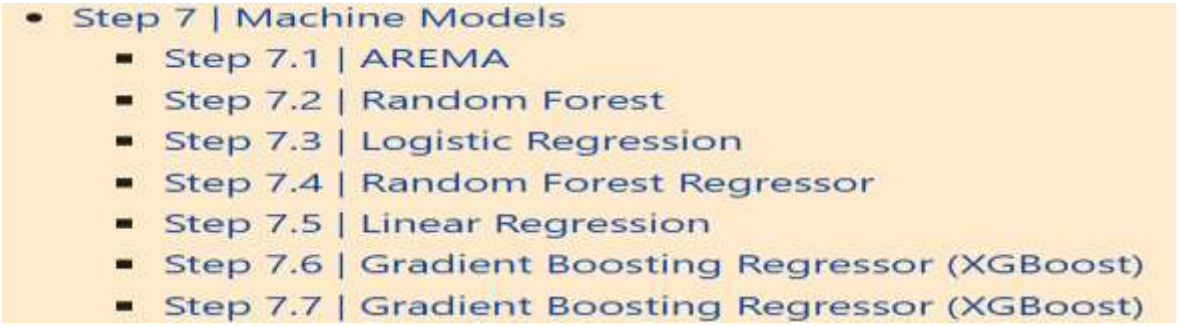
```
X_train_class, X_test_class, y_train_class, y_test_class = train_test_split(X, y_classification, test_size=0.2, random_state=42)

X_train_reg, X_test_reg, y_train_reg, y_test_reg = train_test_split(X, y_regression, test_size=0.2, random_state=42)

classification_pipeline = Pipeline(steps=[('preprocessor', preprocessor)])

X_train_class_transformed = classification_pipeline.fit_transform(X_train_class)
X_test_class_transformed = classification_pipeline.transform(X_test_class)
```

Figure 3: Preparing the dataset for creating models

- On figure 3 I was encoding the categorical variables to numerical. Splitting the dataset into training and testing. A training dataset is used to create the model and fit it. The testing dataset is used to test whether the model is learning.

- Modeling:

- Model Selection: Choose appropriate statistical and machine learning models based on the data characteristics and project objectives.
- Model Training: Train models using the prepared dataset.
- Performance Metrics: Evaluate models using metrics such as accuracy, precision, recall, and F1 score to assess their effectiveness.




```
Random Forest Classification Accuracy: 0.8781362007168458
Logistic Regression Classification Accuracy: 0.7634408602150538
XGBoost Classification Accuracy: 0.8781362007168458
Random Forest Regression MSE: 26957123.94016414
Linear Regression MSE: 26802065.43524434
XGBoost Regression MSE: 26908278.616554506
```

Figure 4 : Machine Model I did.

Figure 4 shows the models I chose to do after I did train each models. It show the performance of how much is classifying and predicting the number of crime.

- Deployment:
  - Reporting: Compile a comprehensive report detailing the analysis, findings, and recommendations.
  - Deployed the predictive model and developed an interactive Power BI dashboard for crime monitoring and analysis.
  - Project Closure: Conclude the project with a final presentation to stakeholders.

5 FINAL MODEL RESULTS

The final deliverable of this project consists of multiple predictive models that have been developed and refined throughout the project lifecycle. These models are designed to analyze crime patterns and forecast future occurrences, providing valuable insights for law enforcement agencies in South Africa. The key models to be delivered include:

1. ARIMA Model for Time Series Forecasting:
   - The ARIMA (AutoRegressive Integrated Moving Average) model will be utilized for forecasting the total number of crimes based on historical data.
   - It captures the underlying trends and seasonality of crime occurrences over time, allowing for short- to medium-term predictions of future crime counts.

- The model will generate forecasts for the next 5 to 10 years, indicating potential future crime counts, which can help in resource allocation and strategic planning for law enforcement.

2. Random Forest Classifier:
    - This ensemble learning model will be employed to classify different types of crimes based on various features, such as geographic location, socio-economic factors, and time of year.
    - The Random Forest model will aggregate predictions from multiple decision trees, improving accuracy and reducing the risk of overfitting.
    - The model will provide probabilities for different crime types occurring in specific areas, enabling law enforcement to focus on high-risk crime categories.
3. XGBoost Classifier:
    - Similar to the Random Forest, the XGBoost (Extreme Gradient Boosting) model will be used for crime classification, leveraging gradient boosting techniques to enhance predictive performance.
    - XGBoost is known for its efficiency and performance, especially in handling complex relationships in the data.
    - This model will yield classification results indicating the likelihood of various crime types, providing additional insights alongside the Random Forest classifier.
4. Logistic Regression Model:
    - A logistic regression model will be included to perform binary classification, specifically to predict whether a crime will occur based on identified features.
    - This model provides interpretability and straightforward probabilities for crime occurrence, making it valuable for stakeholders.
    - The logistic regression model will provide a clear understanding of the factors influencing crime and their corresponding weights.
5. Random Forest Regressor:
    - This model will be used to predict the number of crimes based on features such as population density, urban/rural classification, and socio-economic indicators.
    - The Random Forest Regressor will allow for continuous predictions of crime counts, providing insights into the expected number of crimes in specific areas.
    - The model will produce predicted crime counts, aiding in proactive policing efforts by anticipating the volume of crime.
6. XGBoost Regressor:

- This model will serve a similar purpose as the Random Forest Regressor but will utilize the benefits of XGBoost for enhanced prediction accuracy.
- It will capture complex interactions between variables more effectively, providing better performance in regression tasks.
- The XGBoost Regressor will deliver predicted crime counts with higher accuracy, allowing for informed decision-making by law enforcement.

5.1 Final Data

The final dataset used for this project consists of historical crime data collected from the South African Police Service (SAPS), along with demographic information and other relevant features that influence crime patterns. Below is a summary of the descriptive statistics and discussions regarding sampling and segmentation of the data.

Descriptive Statistics

The following statistics provide an overview of the final dataset, including key metrics related to crime occurrences:

Basic statistics for numeric columns:

	count	mean	std	min	25%	50%	75%	max
Start Year Date	2790	2017.5	2.8728	2013	2015	2017.5	2020	2022
End Year Date	2790	2018.5	2.8728	2014	2016	2018.5	2021	2023

Basic statistics for non-numeric columns:

	count	unique	top	freq
Type of Crime	2790	31	Murder	90
Crime Count	2790	2325	0	58
Province	2790	9	Eastern Cape	310
Urban/Rural	2790	2	Rural	1860
Crime Category	2790	7	CONTACT CRIMES (CRIMES AGAINST THE PERSON)	630
Province Population	2790	317	16098571	311

Figure 5: Statistic of dataset.

Based on figure 5 Most of the crime incidents in the dataset started and ended between 2013 and 2023, with the majority happening around 2017 and 2018. The spread of the data is relatively even, with a difference of about 3 years between when the incidents started and ended for most of the cases. The earliest crime incident started in 2013 and ended in 2014, while the most recent one started in 2022 and ended in 2023. The dataset includes crime data from 9 provinces is South Africa, with the Eastern Cape being the most common region.There are 32 unique types of crimes, with Murder being the most frequent. Contact Crimes are the most prevalent category, occurring more frequently than other categories. A significant portion of the data is from rural areas, indicating that rural crime might be a major focus of this dataset. The dataset has a wide variety of

crime counts, but a notable number of entries have 0 as the count. The population data varies, but one province has a population of over 16 million, and this population value appears frequently.

5.2 Features

The final dataset includes a variety of raw features that provide essential information for analyzing crime patterns. The following are the key raw features present in the dataset:

Data types of this dataset:

Column	Dtype
Type of Crime	object
Crime Count	object
Province	object
Start Year Date	int64
Urban/Rural	object
End Year Date	int64
Crime Category	object
Province Population	object

Figure 6: Features in dataset.

- Type of Crime: Categorical variable representing the specific category of crime (e.g., murder, theft, assault).
- Crime Count: Numerical variable indicating the number of reported incidents for each type of crime.
- Province: Categorical variable indicating the geographic region where the crime occurred (e.g., Gauteng, Western Cape).
- Urban/Rural: Categorical variable indicating whether the crime occurred in an urban or rural area.
- Start Year Date: Temporal variable representing the year the crime incident was reported.
- End Year Date: Temporal variable representing the year the crime incident concluded.
- Crime Category: Categorical variable that groups crimes into broader categories (e.g., violent crime, property crime).
- Province Population: Numerical variable representing the population of the province in which the crime occurred.

Derived Features

- Crime Rate: Derived by calculating the ratio of total crimes to the population of the province, providing insights into crime prevalence relative to the population size.

Formula: Crime Rate=Total Crime CountProvince Population×100,000\text{Crime Rate} = \frac{\text{Total Crime Count}}{\text{Province Population}} \times 100,000Crime Rate=Province PopulationTotal Crime Count×100,000

The importance of features can be assessed using techniques such as Random Forest feature importance or other model-specific methods. Here is a ranking of features based on their importance in predicting crime occurrences and types, determined from the models developed in the project:

- Province Population: High importance as it provides context on crime density relative to the population.
- Type of Crime: Crucial for classification tasks as it directly influences the model's predictions.
- Crime Rate: Significant in understanding crime prevalence and identifying high-risk areas.
- Urban/Rural: Important for differentiating crime patterns in urban vs. rural settings.
- Province: Provides geographic context, influencing regional crime trends and patterns.
- Crime Category: A broader classification that helps in understanding crime dynamics.

5.3 Algorithm

The final models used in the project, focusing on the algorithms implemented for both classification and regression tasks related to crime prediction. Each model's architecture, purpose, hyperparameters, and rationale for selection will be discussed.

- ARIMA Model for Time Series Forecasting:
  The ARIMA (AutoRegressive Integrated Moving Average) model is designed for time series forecasting. It effectively captures trends, seasonality, and cyclic patterns in data, making it ideal for predicting future crime counts based on historical data.
  Key Components:
    - AutoRegressive (AR) Term: Utilizes past values in the time series.
    - Integrated (I) Term: Represents the number of differences needed to make the series stationary.
    - Moving Average (MA) Term: Uses past forecast errors in a regression-like model.

Hyperparameters:

- p (AR term): The number of lag observations (e.g., 1).
- d (Differencing): The number of times the data is differenced (e.g., 1).
- q (MA term): The size of the moving average window (e.g., 1).

Example Configuration and Training Example:

```
arima_model = ARIMA(annual_data, order=(1, 1, 1)).fit()
arima_forecast = arima_model.forecast(steps=10)
```

Figure 7: Creating and training ARIMA Model

- Random Forest Classifier

The Random Forest classifier is an ensemble method that combines multiple decision trees to improve classification accuracy and robustness. It is particularly well-suited for handling high-dimensional datasets with

complex interactions. The was not hyperparameters used.

Example Configuration:

```
rf_classifier = RandomForestClassifier()
rf_classifier.fit(X_train_class_transformed, y_train_class)
```

```
RandomForestClassifier
RandomForestClassifier()
```

Figure 8: Creating and training Random Forest classification model.

- XGBoost Classifier

XGBoost (Extreme Gradient Boosting) is a highly efficient implementation of gradient boosting designed for speed and performance. It captures complex patterns in the data and is particularly effective for classification tasks. Consider looking at the example below how it was made.

```
xgb_classifier = XGBClassifier(use_label_encoder=False, eval_metric='logloss')
xgb_classifier.fit(X_train_class_transformed, y_train_class)
```

```
XGBClassifier
colsample_bytree=None, device=None, early_stopping_rounds=None,
enable_categorical=False, eval_metric='logloss',
feature_types=None, gamma=None, grow_policy=None,
importance_type=None, interaction_constraints=None,
learning_rate=None, max_bin=None, max_cat_threshold=None,
max_cat_to_onehot=None, max_delta_step=None, max_depth=None,
max_leaves=None, min_child_weight=None, missing=nan,
monotone_constraints=None, multi_strategy=None, n_estimators=None,
n_jobs=None, num_parallel_tree=None, random_state=None, ...)
```

Figure 9: Making the XGBoost classifier model

Figure 9 it shows the model also evaluates its predictions using metrics like logloss to ensure it is on the right track, helping law enforcement make informed decisions based on accurate predictions.

- Logistic Regression

Logistic regression is a statistical model that predicts the probability of a binary outcome based on one or more predictor variables. It is interpretable and straightforward, making it valuable for understanding the impact of different features on crime occurrence.

Training Example:

```
logistic_classifier = LogisticRegression()
logistic_classifier.fit(X_train_class_transformed, y_train_class)
```

```
LogisticRegression
LogisticRegression()
```

Figure 10: Training Logistic Regression Model.

In figure 10 it is the model that classifies whether that type of crime happened or not.

5.4 Results

The results obtained from the various machine learning models implemented in the project, including their performance metrics and relevant visualizations illustrate the findings.

- Model Performance Metrics: We evaluated the performance of our classification and regression models based on specific metrics:

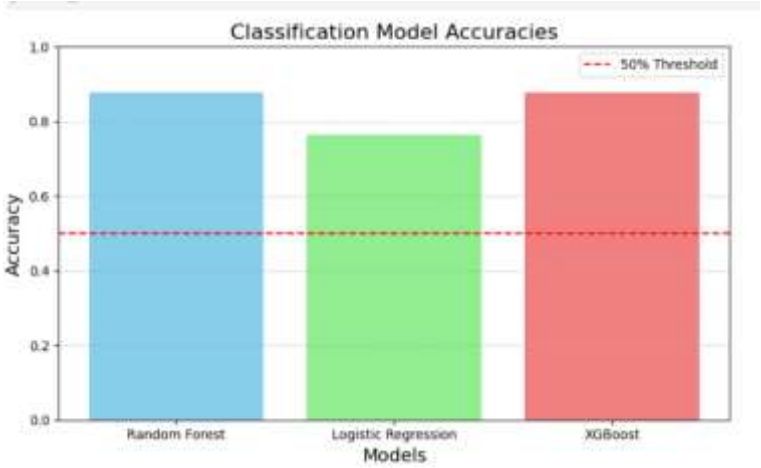


Figure 11: Accuracies of the classification models

In figure 11 shows classification models: The Random Forest and XGBoost models demonstrate strong performance in classification tasks, with XGBoost slightly edging out Random Forest. Logistic Regression, while decent, underperforms compared to these ensemble methods.

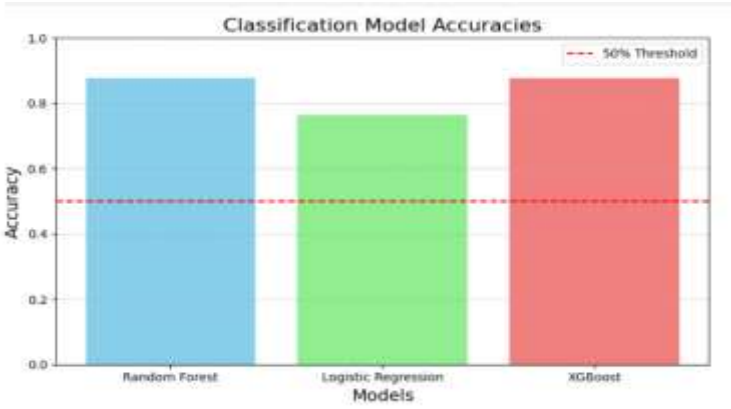


Figure 11: Mean Squared Errors of regression models

- In figure 11 Regression Models: The Mean Squared Errors indicate that the predictions are somewhat off, with all models showing high MSE values. However, Linear Regression performs best in this

scenario, followed closely by Random Forest, while XGBoost's performance in regression is slightly less favorable.

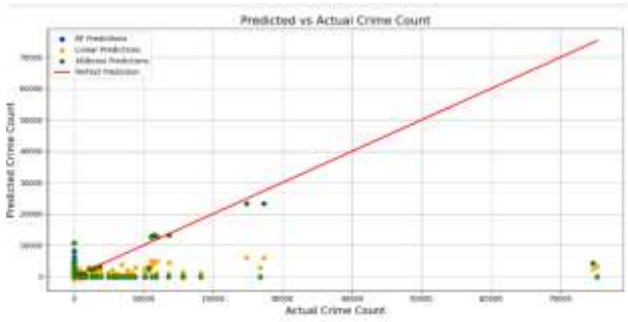


Figure 12: Predicted vs. Actual Crime Count Visualization

In figure 12 shows scatter plot below compares the predicted crime counts from different models against the actual crime counts. A perfect prediction line is included for reference

6 RECOMMENDATIONS

6.1 Actionable Insights

Based on the analysis and predictive modeling conducted in this project, the following recommendations are made to enhance crime prevention strategies and improve resource allocation for law enforcement agencies in South Africa:

- Targeted Resource Allocation
    - Recommendation: Utilize the predictive models to identify high-risk areas for specific crime types, allowing for focused deployment of police resources.
    - Implementation:
        - Use insights from the XGBoost and Random Forest classifiers to determine regions with a higher likelihood of violent crimes or property crimes.
        - Allocate additional police patrols and resources to these identified hotspots, especially during peak crime periods as indicated by historical trends.
- Utilize Advanced Data Analytics
    - Recommendation: Incorporate advanced analytics and machine learning techniques in ongoing crime analysis efforts to stay ahead of crime trends.
    - Implementation:
        - Invest in training for crime analysts in machine learning techniques to refine and enhance predictive capabilities.
        - Regularly update models with new data to improve accuracy and adjust to changing crime patterns.

Using machine learning models can even help in predicting whether the number of crimes wil increase or decrease consider example where I predict using the Arima model.

```
ARIMA Forecast for next 5 years: 2023-12-31    1.596097e+06
2024-12-31    1.519462e+06
2025-12-31    1.502231e+06
2026-12-31    1.498357e+06
2027-12-31    1.497486e+06
2028-12-31    1.497290e+06
2029-12-31    1.497246e+06
2030-12-31    1.497236e+06
2031-12-31    1.497234e+06
2032-12-31    1.497233e+06
```

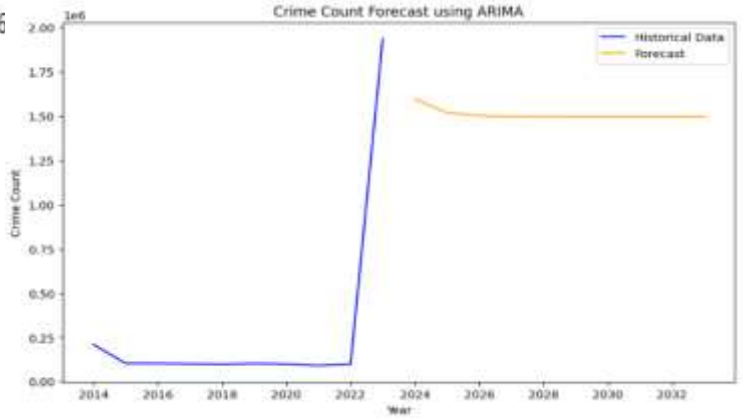


Figure 13 : Prediction of number of crimes in the next 10 year.

Based on figure 13 the findings from the ARIMA model:

- Proactive Crime Prevention: Given the expected stable trend in crime counts, the SAPS can allocate resources effectively to maintain safety levels in areas projected to have high crime counts.
- Policy Interventions: Implementing targeted interventions in specific years based on forecasted increases can help mitigate future crime rates.
- Regular Model Updates: Continuously update the model with new data to refine predictions and adapt to changing crime trends.

6.2 Potential Risks and Mitigation Strategies

Potential Risks and Limitations:

• Data Quality and Reliability: Official crime data from SAPS may suffer from issues such as underreporting, misclassification, and data manipulation, which could affect the accuracy and reliability of crime analysis findings. Similarly, victimization surveys Page 12 of 14 may be subject to response biases, memory errors, and sampling limitations, compromising the validity of the results.

• Access to Data: Access to comprehensive and up-to-date crime data, especially at a granular level (e.g., sub-regional or neighborhood level), may be restricted due to data privacy concerns, bureaucratic hurdles, and legal constraints. Limited access to relevant datasets could hamper the completeness and depth of the analysis.

• Overfitting of predictive models to the training data could lead to poor generalization performance on unseen data. The models were not learning well especially those for regression

Mitigation Strategies:

• Data Quality and Reliability: Perform cross-checking with other sources to minimize errors and inconsistencies. Incorporate uncertainty and error estimates in the analysis and reporting to reflect the limitations of the data.

• Access to Data: Explore alternative data sources, such as open-source datasets, social media, and other public records, to supplement the official data.

• Overfitting of Predictive Models: Perform model selection based on out-of-sample performance metrics, such as F1 score, to ensure the model's generalizability.

6.3 Challenges

- Accessing up-to-date and detailed crime data from reliable sources was a significant challenge. Crime data from SAPS was often incomplete, lacked proper standardization, or missed key information at the sub-regional level.
- The machine learning models, particularly for regression, displayed signs of overfitting, meaning they performed well on training data but struggled to generalize to unseen data.
- Certain crime categories were underrepresented, leading to model bias and reduced performance in predicting crimes from these categories.

7 REFERENCES

- bank, T. w. (2023, November 22). South Africa Economic Update: Raising South Africa's Economic Prospects by Curbing Crime. Retrieved from The world bank: https://www.worldbank.org/en/country/southafrica/publication/raising-south-africa-s□afe-1123-economic-prospects-by-curbingcrime#:~:text=The%20challenge%20of%20high%20crime%20rates%20undermines%2 0the%20country's%20economic,percent%20of%20GDP%20every%20yea
- Carlos Rosa, I. G. (2020). Data Analytics in Public Safety. European Emergency Number Association, 21. Retrieved from https://eena.org/knowledge-hub/documents/data□analytics-in-public-safety/
- Dajao, A. M. (2021). Crime Mapping Approach for Crime Pattern Identification: A Prototype for the Province of Cavite. In N. D. Xin-She Yang, Proceedings of Sixth International Congress on Information and Communication Technology. Lecture Notes in Networks and Systems (pp. 899-909). Springer, Singapore. Retrieved from https://doi.org/10.1007/978-981-16-2380-6_79
- Faull, A. (2022, March 28). More questions than answers in South Africa’s latest victim survey? Institution for security studies, 9. Retrieved from https://issafrica.org/iss-today/more□questions-than-answers-in-south-africas-latest-victim-survey
- G. Sivapriya, B. V.-A. (2023). Crime Prediction and Analysis Using Data Mining and Machine Learning: A Simple Approach that Helps Predictive Policing. FMDB Transactions on Sustainable Computer Letters, Vol. 1(No.2), 12. Retrieved from https://www.researchgate.net/publication/375799201_Crime_prediction_and_analysis_using_Data_mining_and_Machine_learning_An_approach_that_helps_Predictive_policing#read

- Graan, J. v. (2021). Perspectives on the Violent Nature of Crime Victimisation in South Africa. In H. C. Adjorlolo, Crime, Mental Health and the Criminal Justice System in Africa (pp. 39– 62). Palgrave Macmillan, Cham. doi: https://doi.org/10.1007/978-3-030-71024-8_3
- Jean-Claude Manaliyo, P.-F. M. (2013). Community Participation in Crime Prevention: Informal Social Control Practices in Site B, Khayelitsha Township. Mediterranean Journal of Social Sciences , 4(No.2), 9. doi:Doi:10.5901/mjss.2013.v4n3p121
- Karabo Jenga, ·. C. (2023). Machine learning in crime prediction. Journal of Ambient Intelligence and Humanized Computing, 28. doi:https://doi.org/10.1007/s12652-023-04530-y
- Macrotrends. (2024, April 23). South Africa crime rate & statistics 1960-2024. Retrieved from Macrotrends: https://www.macrotrends.net/global-metrics/countries/ZAF/southafrica/crime-rate-statistics.
- Maluleke, R. (2023, August 24). Victims of Crime 2022/23. Governance, Public Safety and Justice Survey , 34. Retrieved from https://www.statssa.gov.za/publications/P0341/GPSJS%202022_23%20Final.pdf
- Omowunmi Isafiade, B. N. (2021). Predictive Policing Using Deep Learning: A Community Policing Practical Case Study. In A. P. Rafik Zitouni, Towards new e-Infrastructure and e-Services for Developing Countries. AFRICOMM 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (Vol. 7, pp. 269–286). Springer, Cham. Retrieved from https://doi.org/10.1007/978-3-030-70572- 5_1

8 APPENDIX:

- LINKS:
  - https://github.com/Syabonga04/Mlambo.git
  - https://github.com/Syabonga04/Mlambo