# Modelling Voting Patterns at the Ward and Municipal Levels in South Africa Using Remote Sensing and Machine Learning

*A Dissertation in fulfillment of the requirements*

*for the degree of Master of Science in Engineering*

*by*

Joseph Baggott

FACULTY OF ENGINEERING AND THE BUILT ENVIRONMENT

UNIVERSITY OF THE WITWATERSRAND

September 2025

# Declaration

I, **Joseph Baggott**, student number **2169705**, declare that this dissertation is my own unaided work. It is being submitted to the **Master of Science in Engineering** to the **University of the Witwatersrand**. It has not been submitted before for any degree or examination to any other University.

Joseph Baggott

(Signature of Candidate)

11th day of September year 2025

# *Abstract*

Name of the student: **Joseph Baggott** Student Number: **2169705**

Degree for which submitted: **Master of Science in Engineering**

Thesis title: **Modelling Voting Patterns at the Ward and Municipal Levels in South Africa Using Remote Sensing and Machine Learning**

Thesis supervisors: **Dr Martin Bekker**

Month and year of thesis submission: **September 2025**

This study investigates whether voting patterns in South Africa can be modelled using open-source data, satellite imagery, and Google Street View images. It responds to a context of declining voter turnout, growing political disillusionment, and a research gap in understanding how physical living conditions influence electoral behaviour. Existing studies largely rely on aggregated census or survey data and overlook the spatial and infrastructural environments that shape daily life and political decision-making.

To address this gap, the study collected over 157 000 Google Street View images and satellite-derived land use data at both ward and municipal levels. Using computer vision, it extracted features such as potholes, illegal dumping, people, and vehicles. These were supplemented with open-source crime and financial data, and modelled using machine learning and statistical regression techniques. The models performed strongly, with classification F1 scores reaching 0.93 and regression R-2 values up to 0.86, demonstrating a measurable relationship between living conditions and voting outcomes.

The findings reveal that no single voting theory can fully explain voter behaviour in South Africa. While rational choice theory accounts for non-voting in areas of poor service delivery, social structure and psychological theories better explain support for parties like uMkhonto we Sizwe, where regional identity and historical loyalty dominate. Ultimately, the study shows that voting behaviour emerges from a complex mix of infrastructure, identity, and lived experience. These insights have implications for political modelling, service delivery accountability, and understanding democratic participation in contexts marked by inequality and transition.

# *Acknowledgements*

I would like to extend my heartfelt gratitude to all those who contributed to the completion of this work.

Firstly, I am deeply thankful to my supervisor, Dr. Martin Bekker, for his unwavering support, guidance, encouragement throughout, and friendship this journey. His expertise and mentorship were instrumental in the successful completion of my dissertation.

I would also like to acknowledge the late Professor Barry Dwolatzky for his profound influence on the field and the creation of a fund that continues to support research and innovation. His legacy in education and technology will always be remembered and cherished.

Lastly, I am incredibly grateful to my parents for their endless love and support. Their belief in me has been a constant source of motivation, and without them, this achievement would not have been possible.

Thank you all for your invaluable contributions.

# Contents


**Acknowledgements** iv

**List of Figures** viii

**List of Tables** xii

**Abbreviations** xiv

**1 Introduction** 1
- 1.1 Context and Background . . . 1
- 1.2 Research Question . . . 2
  - 1.2.1 Supporting Research Questions . . . 3
- 1.3 Philosophical Orientation . . . 3
- 1.4 Contribution to Theory and Practice . . . 4
- 1.5 Dissertation Roadmap . . . 5
  - 1.5.1 Chapter 2: Literature Review . . . 5
  - 1.5.2 Chapter 3: Methodology . . . 5
  - 1.5.3 Chapter 4: Results and Model Performance . . . 5
  - 1.5.4 Chapter 5: Findings: Key Features Affecting Voting Patterns . . . 6
  - 1.5.5 Chapter 6: Discussion . . . 6
  - 1.5.6 Chapter 7: Conclusion . . . 6

**2 Literature Review** 7
- 2.1 Voting In South Africa . . . 7
- 2.2 Lived Experiences and Voting Outcomes . . . 9
- 2.3 Theories of Voting Behaviour . . . 10
  - 2.3.1 Rational Choice Theory . . . 11
  - 2.3.2 Clientelism . . . 11
  - 2.3.3 Social Structure Determinants . . . 12
  - 2.3.4 Psychological Theories of Voting . . . 13
- 2.4 Machine Learning (ML) and Statistical Modelling . . . 14
  - 2.4.1 Use Cases: Social and Political Sciences . . . 15
- 2.5 Computer Vision . . . 16

2.5.1 Use Cases . . . 17

**3 Methodology 18**

3.1 System Overview . . . 18
3.2 Voting Patterns and Feature Collection . . . 20
3.3 Structure of the Remainder of the Methodology Chapter . . . 22
3.4 Google StreetView Image Collection . . . 23
3.4.1 Exploratory Methodology . . . 25
3.4.2 Sourcing Methodology . . . 30
3.5 Google StreetView Sample Size Calculation . . . 39
3.6 Additional Data Sources . . . 44
3.6.1 Crime . . . 45
3.6.2 Normalising Crime Data . . . 47
3.6.3 Financial Data . . . 47
3.7 Land Use Data . . . 48
3.7.1 Converting From Raster to Vector Information . . . 49
3.7.2 Normalising Land Use Features . . . 50
3.8 Computer Vision . . . 51
3.8.1 Potholes . . . 52
3.8.2 People . . . 55
3.8.3 Vehicles . . . 56
3.8.4 Graffiti . . . 56
3.8.5 Illegal Dumping . . . 57
3.8.6 Inference . . . 59
3.9 machine learning (ML) and Statistical Regression . . . 60
3.9.1 Ward Administrative Level (ML Models) . . . 64
3.9.2 Municipal Administrative Level (Statistical Models) . . . 68

**4 Results and Model Performance 70**

4.1 Image Collection and Computer Vision . . . 71
4.1.1 Potholes . . . 71
4.1.2 People . . . 73
4.1.3 Vehicles . . . 75
4.1.4 Illegal Dumping . . . 77
4.2 Modelling Voting Patterns at The Ward Level . . . 79
4.2.1 ML Regression Model Performance . . . 81
4.2.2 Visualising Model Performance . . . 86
4.2.3 ML Classification Model Performance . . . 92
4.3 Modelling Voting Patterns at The Municipal Level . . . 99
4.3.1 Model Performance Post Feature Reduction . . . 99

**5 Findings: Key Features Affecting Voting Patterns 103**

5.1 ML Correlations and Feature Analysis At The Ward Level . . . 103
5.1.1 Non Vote Percentage Feature Analysis . . . 104
5.1.2 ANC Percentage Feature Correlations and Importance Plots . . . 106

5.1.3 Democratic Alliance (DA) Percentage Feature Correlations and Importance Plots . . . 108
5.1.4 uMkhonto we Sizwe (MK) Percentage feature correlations and importance plots . . . 111
5.1.5 Economic Freedom Fighters (EFF) Percentage Feature Analysis . . . 113
5.2 Feature Analysis at The Municipal Level . . . 114
5.2.1 Non Vote Importance Plots . . . 115
5.2.2 African National Congress (ANC) Importance Plots . . . 116
5.2.3 DA Importance Plots . . . 118
5.2.4 MK Importance Plots . . . 119
5.2.5 EFF Importance Plots . . . 121

**6 Discussion** **123**

**7 Conclusion** **126**
7.1 Introduction . . . 126
7.2 Literature Review . . . 127
7.3 Findings . . . 128
7.4 Limitations and Future Work . . . 128
7.5 Summary of Discussion and Findings . . . 130

**Note on appendix structuring** **142**

**3.8 Computer Vision Model Performance** **143**
3.8.1 Potholes . . . 143
3.8.2 People . . . 146
3.8.3 Vehicles . . . 149
3.8.5 Illegal Dumping . . . 152

**4.1 Choropleth Outputs** **156**
4.1.2 People Choropleth . . . 156
4.1.3 Vehicle Choropleth . . . 157
4.1.4 Illegal Dumping Choropleth . . . 157

**4.2 Actual and Predicted Distributions** **158**

# List of Figures


3.1 A macro view: flow diagram of the system highlighting data collection, feature generation, model building and visualisation . . . . . . . . . . . . . . 19
3.2 Outcome of the 2024 national elections . . . . . . . . . . . . . . . . . . . . . . 20
3.3 Comparison of the administrative units between the ward (left) and municipal (right) level in South Africa . . . . . . . . . . . . . . . . . . . . . . . . 22
3.4 Output of location generation of the grid (left) and random (right) approach on a ward in the City of Tshwane Metropolitan Municipality . . . . . . . . 26
3.5 Output of location generation of the grid (left) and random (right) approach when only valid locations are retained in the City of Tshwane Metropolitan Municipality . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
3.6 Road network South Africa (left), the City of Johannesburg (middle), and a ward in the City of Johannesburg (right) . . . . . . . . . . . . . . . . . . 28
3.7 Snapping random locations in green (left) to hypothetical roads in red (right) using Geopandas . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
3.8 Result of snapping locations to Murraysburg town in Beaufort West Local Municipality, illustrating points being snapped to the periphery of the town 29
3.9 Map showing residential and non-residential roads in Beaufort West Local Municipality . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
3.10 Screenshot of Google Maps highlighting roads that are available in Google StreetView (GSV) in Murraysburg, South Africa . . . . . . . . . . . . . . . 33
3.11 Screenshot of Google Maps highlighting that only a tertiary road is available while residential roads are not in a rural town in KwaZulu-Natal, a province in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
3.12 Output of getting four images per location, showing that at least one image or combination of two will point in the direction of a road (Parkview in City of Johannesburg). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 36
3.13 Output of four images using the bearing approach showing the camera pointing in the direction of the road . . . . . . . . . . . . . . . . . . . . . . . . 38
3.14 Correlation between normalised population size [3] and number of roads per municipality in South Africa [74], showing a strong correlation between roads and eligible voters . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
3.15 Line graph showing the number of roads available for each municipality in South Africa, represented as a index from least roads to most roads and a static horizontal line at 385 showing the infinite sample size . . . . . . . . . 42
3.16 Line graph showing the scaling factor for each municipality in South Africa represented as a index from least roads to most roads . . . . . . . . . . . . 44

3.17 Location of police stations across South Africa as provided in the ArcGIS dataset [83], which can then be aggregated at the municipal level . . . . . . 46
3.18 South African land use raster file obtained from the Department of Environmental Affairs [87] . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
3.19 Manhole covers as a cause for false positives in pothole detection . . . . . . 54
3.20 Example images from trash object detection set . . . . . . . . . . . . . . . . 58
3.21 Image showing the available wards to be used in ML models as a result of availability in GSV . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 61

4.1 Distribution of pothole detection confidence scores across collected GSV images in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 72
4.2 Images showing variability in confidence scores due to pothole non-conformity and distance from the camera . . . . . . . . . . . . . . . . . . . . . . . . 72
4.3 Pothole ratio density across three major municipalities in South Africa at the ward level, City of Tshwane (left), City of Cape Town (middle), ethekwini Metropolitan Municipality (right) . . . . . . . . . . . . . . . . . . . . 73
4.4 Distribution of people detection confidence scores across collected GSV images in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 74
4.5 Images showing variability in confidence scores due to people non-conformity and distance from the camera . . . . . . . . . . . . . . . . . . . . . . . . 75
4.6 Distribution of vehicle detection confidence scores across collected GSV images in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 76
4.7 Images showing high vehicle detection confidence regardless of image variation 77
4.8 Distribution of illegal dumping detection confidence scores across collected GSV images in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . 78
4.9 Images showing variability in confidence scores due to variation in illegal dumping sites . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 79
4.10 Spatial distribution of high-density urban and rural municipalities in South Africa . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 81
4.11 Actual and predicted distribution of ANC vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . . . . . . . . . . . . . . . . . . . . . . . . 85
4.12 Actual and predicted distribution of MK vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . . . . . . . . . . . . . . . . . . . . . . . . 85
4.13 Actual vs predicted scatter plot (left) and residual distribution (right) for non-vote percentage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87
4.14 Actual vs predicted scatter plot (left) and residual distribution (right) for ANC vote percentage using the final random forest model . . . . . . . . . . 88
4.15 Actual vs predicted scatter plot (left) and residual distribution (right) for DA vote percentage using the final random forest model . . . . . . . . . . . 89
4.16 Actual vs predicted scatter plot (left) and residual distribution (right) for MK vote percentage using the final random forest model . . . . . . . . . . . 90
4.17 Actual vs predicted scatter plot (left) and residual distribution (right) for EFF vote percentage using the final random forest model . . . . . . . . . . 91

4.18 Distribution of number of wards won by the dominant parties and the summation of the remaining parties in the national dataset . . . . . . . . . . . . 93
4.19 Distribution of number of wards won by the ANC, DA, MK in the national test dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 93
4.20 Confusion matrix of the logistic regression model in predicting majority vote on national test dataset . . . . . . . . . . . . . . . . . . . . . . . . . . 95
4.21 Confusion matrix of the random forest model in predicting majority vote on national test dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . 96
4.22 Confusion matrix of logistic regression model in predicting majority vote on national test dataset without including provinces . . . . . . . . . . . . . 98
4.23 Confusion matrix of random forest model in predicting majority vote on national test dataset without including provinces . . . . . . . . . . . . . . . 98

5.1 Feature importances of final random forest model in predicting non-vote percentage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 104
5.2 Feature importances of final random forest model in predicting ANC percentage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
5.3 Feature importances of final random forest model in predicting DA percentage 109
5.4 Feature importances of final random forest model in predicting MK percentage 111
5.5 Feature importances of final random forest model in predicting EFF percentage . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 113
5.6 Feature importance plot for the non vote elastic net regression model . . . . 115
5.7 Feature importance plot for the ANC elastic net regression model . . . . . . 116
5.8 Feature importance plot for the DA elastic net regression model . . . . . . . 118
5.9 Feature importance plot for the MK elastic net regression model . . . . . . 119
5.10 Feature importance plot for the EFF elastic net regression model . . . . . . 121

3.8.1 Normalised confusion matrix of final pothole object detection on test set 143
3.8.2 F1 curve of pothole object detection model . . . . . . . . . . . . . . . . . 144
3.8.3 Precision-confidence curve of final pothole detection model . . . . . . . . 144
3.8.4 Precision-recall curve of final pothole detection model . . . . . . . . . . . 145
3.8.5 Recall-confidence curve of final pothole detection model . . . . . . . . . . 145
3.8.6 Training results of final pothole detection model . . . . . . . . . . . . . . 146
3.8.7 Normalised confusion matrix of final pothole object detection on test set 146
3.8.8 F1 curve of pothole object detection model . . . . . . . . . . . . . . . . . 147
3.8.9 Precision-confidence curve of final pothole detection model . . . . . . . . 147
3.8.10 Precision-recall curve of final pothole detection model . . . . . . . . . . . 148
3.8.11 Recall-confidence curve of final pothole detection model . . . . . . . . . . 148
3.8.12 Training results of final pothole detection model . . . . . . . . . . . . . . 149
3.8.13 Normalised confusion matrix of final car object detection on test set . . . 149
3.8.14 F1 curve of car object detection model . . . . . . . . . . . . . . . . . . . 150
3.8.15 Precision-confidence curve of final car detection model . . . . . . . . . . . 150
3.8.16 Precision-recall curve of final car detection model . . . . . . . . . . . . . 151
3.8.17 Recall-confidence curve of final car detection model . . . . . . . . . . . . 151
3.8.18 Training results of final car detection model . . . . . . . . . . . . . . . . 152

3.8.19 Normalised confusion matrix of final illegal dumping object detection on test set . . . 152
3.8.20 F1 curve of illegal dumping object detection model . . . 153
3.8.21 Precision-confidence curve of final illegal dumping detection model . . . 153
3.8.22 Precision-recall curve of final illegal dumping detection model . . . 154
3.8.23 Recall-confidence curve of final illegal dumping detection model . . . 154
3.8.24 Training results of final illegal dumping detection model . . . 155

4.1.1 People ratio density across three major municipalities in South Africa at the ward level . . . 156
4.1.2 Vehicle ratio density across three major municipalities in South Africa at the ward level . . . 157
4.1.3 Vehicle ratio density across three major municipalities in South Africa at the ward level . . . 157

4.2.1.1 Actual and predicted distribution of DA vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . 158
4.2.1.2 Actual and predicted distribution of MK vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . 159
4.2.1.3 Actual and predicted distribution of EFF vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . 159
4.2.1.4 Actual and predicted distribution of Non Vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category . . . 160

# List of Tables


3.1 Number of geographical units available for electoral analysis at different administrative levels . . . 21
3.2 Table showing the length, random probability, and weighted probability of three roads . . . 30
3.3 Selected examples of road types in OpenStreetMap . . . 31
3.4 Statistical characteristics of tertiary and secondary roads in South Africa . 35

4.1 $R^2$ values of the linear regression (linear regression (LR)) and random forest (Random Forest (RF)) models using cross-validation with 10 folds on all wards on the train and test set . . . 82
4.2 $R^2$ values of the linear regression (LR) and random forest (RF) models using cross-validation with 10 folds on urban wards on the train and test set . . . 82
4.3 Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) of the linear regression (LR) model using cross-validation with 10 folds on all wards on the test set . . . 83
4.4 MAE, MSE, RMSE of the linear regression (LR) model using cross-validation with 10 folds on urban wards on the test set . . . 83
4.5 MAE, MSE, RMSE of the random forest (RF) model using cross-validation with 10 folds on all wards on the test set . . . 84
4.6 MAE, MSE, RMSE of the random forest (RF) model using cross-validation with 10 folds on urban wards on the test set . . . 84
4.7 Performance metrics of the logistic regression classifier on classifying majority vote on the test national dataset . . . 94
4.8 Performance metrics of the random forest classifier on classifying majority vote on the test national dataset . . . 95
4.9 Performance metrics of logistic regression classifier on classifying majority vote on the test national dataset without including provinces . . . 97
4.10 Performance metrics of random forest classifier on classifying majority vote on the test national dataset without including provinces . . . 97
4.11 Ordinary Least Squares (OLS) model performance summary for non-vote . 101
4.12 OLS model performance summary for the ANC . . . 101
4.13 OLS model performance summary for the DA . . . 102
4.14 OLS model performance summary for the MK . . . 102
4.15 OLS model performance summary for the EFF . . . 102

5.1 Top 10 most important Pearson and Spearman correlation scores for the non-vote percentage . . . 104
5.2 Top 10 most important Pearson and Spearman correlation scores for the ANC vote percentage . . . 106
5.3 Top 10 most important Pearson and Spearman correlation scores for the DA vote percentage . . . 108
5.4 Top 10 most important pearson and spearman correlation scores for the MK vote percentage . . . 111
5.5 Top 10 most important pearson and spearman correlation scores for the EFF vote percentage . . . 113

# Abbreviations

| | |
|---|---|
| **ANC** | African National Congress |
| **DA** | Democratic Alliance |
| **EFF** | Economic Freedom Fighters |
| **MK** | uMkhonto we Sizwe |
| **ML** | Machine Learning |
| **SAPS** | South Africa Police Service |
| **IEC** | Independent Electoral Commission |
| **UIFW** | Unauthorised, Irregular, Fruitless, and Wasteful Expenditure |
| **YOLO** | You Only Look Once |
| **YOLOv8** | You Only Look Once, Version 8 |
| **RFECV** | Recursive Feature Elimination with Cross-Validation |
| **MSE** | Mean Squared Error |
| **RMSE** | Root Mean Squared Error |
| **MAE** | Mean Absolute Error |
| **HTTP** | Hypertext Transfer Protocol |
| **API** | application programming interface |

| | |
|---|---|
| **mAP** | mean average precision |
| **ML** | machine learning |
| **VIF** | variance inflation factor |
| **OLS** | Ordinary Least Squares |
| **LR** | linear regression |
| **RF** | Random Forest |
| **GSV** | Google StreetView |
| **FOV** | Field of View |

# Chapter 1

# Introduction

## 1.1 Context and Background

Voting is a cornerstone of democracy, enabling citizens to influence leadership, shape policy, and legitimise political authority [1, 2]. It facilitates peaceful power transitions and fosters political stability by providing a structured mechanism for public participation and accountability.

In South Africa, the vote holds symbolic and practical importance, particularly given the country's apartheid legacy. However, recent elections have revealed declining trust in the political system. The 2024 general election marked a historic shift, with the African National Congress (ANC) losing its majority for the first time since 1994 [3]. While opposition parties gained traction, voter turnout continued to fall, reflecting disillusionment, especially among urban and younger voters [4, 5].

This trend signals not just apathy but a broader crisis in democratic legitimacy and representation. Many voters feel elections no longer deliver meaningful change, leading some to disengage entirely and others to seek alternatives to established parties [4].

Understanding voting patterns is thus essential for strengthening democratic participation [6]. While previous studies have examined socioeconomic influences through surveys or census data, generally relying on rational choice theory [7, 5], this perspective overlooks the full complexity of electoral behaviour. Decisions are shaped by interrelated factors, economic, social, cultural, and historical, making modelling inherently challenging [6].

In the South African context, one overlooked dimension is the impact of physical living conditions on voter behaviour. Some research has shown that service delivery and infrastructure quality can shape electoral outcomes, as citizens often vote based on their lived experiences and access to basic services. This relationship suggests that it may be possible to model voting patterns by directly observing environmental conditions on the ground. As a result, this study argues that living conditions affect voting considerations and voting considerations impact voting patterns. In this study, voting patterns are understood as the outcomes of electoral processes linked to specific geographical areas, reflected both in shifts in party support (e.g., ANC, Democratic Alliance (DA), uMkhonto we Sizwe (MK)) and in levels of voter abstention. Voting considerations, by contrast, refer to the psyche of the electorate, while living conditions represent the physical environments in which the electorate resides.

Current studies, however, rely on proxies for service delivery, such as access to water, sanitation, and refuse collection, drawn from census and survey data [8, 9, 10]. These models are typically constrained by scale, often conducted at the municipal level, which lacks the spatial granularity needed to capture localised variations in voter behaviour.

Given South Africa's persistent inequalities and diverse socio-spatial contexts, there is a clear need for richer, more granular data that directly represents physical living conditions. Integrating spatial analysis, machine learning (ML) and statistical regression, and alternative data sources, such as satellite and street-level imagery, offers a promising path forward in capturing the complexity of voting dynamics.

## 1.2 Research Question

The research seeks to address the following question: "To what extent can voting patterns be modelled in South Africa using open-source data, satellite, and Google StreetView imagery?" Voting patterns denote electoral outcomes, while open-source resources refer to "free to use" data and commonly used statistical, deep, and ML models and tools. This aligns with the broader issues identified earlier, addressing the need for granular and rich data that directly represents living conditions in South Africa. This question reflects a broader ambition to bridge the gap between lived experience and political analysis through data-driven methods.

### 1.2.1 Supporting Research Questions

To answer the primary research question, this study explores several sub-questions related to identifying voting patterns, determining appropriate spatial resolution, acquiring and processing open-source data, and evaluating feature extraction methods. The following supporting questions guide the research:

i. What are the most relevant and interpretable voting patterns in the South African context?

ii. At which spatial resolution can voting patterns be effectively modelled?

iii. To what extent can Google StreetView and satellite imagery be accessed and collected at scale?

iv. How effectively can computer vision models extract meaningful features from these images?

v. What additional open-source data sources can be integrated to enrich the analysis?

vi. What are the key features that affect voting patterns and how do these relate to theoretical frameworks?

## 1.3 Philosophical Orientation

This study adopts an inductive research approach, where patterns and relationships are identified from observed data rather than being driven solely by pre-defined theoretical models. Instead of beginning with a fixed set of hypotheses, the research starts by collecting a broad range of potentially relevant variables, such as environmental, socioeconomic, and infrastructural indicators, based on prior literature, exploratory analysis, and domain intuition.

These variables, often referred to as features in data science, are systematically tested and refined through empirical analysis to determine which best explain or predict voting behaviour. This flexible, data-driven methodology allows the study to uncover insights that may not be captured through traditional theory-first approaches, and is particularly well-suited for exploring complex, context-specific social phenomena such as voting patterns in South Africa.

## 1.4 Contribution to Theory and Practice

This study contributes to both the theoretical understanding of voting behaviour and the practical development of data-driven tools for political analysis.

**Theoretical Contribution:** The research advances voting behaviour theory by introducing a novel spatial and visual dimension, satellite and street-level imagery, into electoral modelling in South Africa. While previous studies in South Africa have largely focused on socioeconomic proxies and survey data, this study explores how direct observations of the built environment, captured via Google StreetView (GSV) and satellite imagery, relate to voter preferences and turnout. By integrating physical living conditions into models of political behaviour, the research complements and extends existing frameworks such as rational choice theory and social structure theories, providing a more holistic account of how context shapes democratic engagement.

**Practical Contribution:** From a methodological perspective, this study demonstrates the feasibility of using open-source computer vision tools, statistical models and ML to model voting patterns at a fine-grained spatial resolution. It introduces scalable approaches for extracting features from imagery and linking them to electoral data, offering a proof of concept for using remote sensing in political science research. Practically, the insights generated have potential applications in electoral analysis, public sector planning, and governance accountability, helping policymakers better understand areas of disengagement or shifting support using publicly available data.

## 1.5 Dissertation Roadmap

This dissertation is structured into seven chapters, each designed to address the core research question and its supporting sub-questions. The following sections outline the purpose and content of each chapter.

### 1.5.1 Chapter 2: Literature Review

This chapter reviews the literature relevant to the study. It begins by outlining the importance of voting in South Africa's democratic context and explores trends of declining voter participation and shifting party loyalties. It then examines key theoretical frameworks, rational choice, clientelism, social structure determinants, and psychological models, with particular attention to how physical living conditions shape political engagement. The chapter concludes by tracing the development of statistical modelling, ML, and computer vision in social science research, highlighting their potential to link lived experiences and infrastructure to voting outcomes.

### 1.5.2 Chapter 3: Methodology

This chapter details the methodological framework and data pipeline used in the study. It explains the collection and processing of GSV images, satellite imagery, and additional open-source datasets. The development and training of computer vision models to extract features from images are described, alongside the creation of a unified feature store. The chapter then outlines the design and implementation of both statistical and ML models used to analyse voting patterns.

### 1.5.3 Chapter 4: Results and Model Performance

This chapter presents the results of the ML and statistical models. It compares model performance in predicting voting patterns across different resolutions and evaluates each model's predictive accuracy and robustness. The results highlight the viability of using imagery and open-source data to model electoral outcomes in the South African context.

### 1.5.4 Chapter 5: Findings: Key Features Affecting Voting Patterns

This chapter analyses the most important features identified by the models. It explores how various socioeconomic, environmental, and infrastructural factors influence voting behaviour and turnout. These findings are then interpreted in light of the theoretical frameworks discussed in the literature review, offering new perspectives on how physical living conditions and spatial context shape democratic engagement.

### 1.5.5 Chapter 6: Discussion

This chapter reflects on the empirical findings and theoretical insights generated by the study. It evaluates how features extracted from satellite and GSV imagery, such as infrastructure, land use, and crime, relate to voter turnout and party support across South Africa. It discusses the performance of the ML and statistical models, the strengths of image-based analysis, and the explanatory power of different voting theories. In doing so, it highlights the need for a multidimensional approach to understanding electoral behaviour, showing that no single theory fully captures the complexity of South African voting patterns.

### 1.5.6 Chapter 7: Conclusion

The final chapter brings together the study's central insights, revisiting the research questions and reflecting on the key findings, limitations, and contributions. It highlights how physical living conditions, captured through satellite and street-level imagery, can offer powerful indicators of voting behaviour, and how ML methods can model these patterns with high accuracy. Importantly, the chapter argues that no single voting theory can fully account for South Africa's complex political behaviour, and that multiple frameworks must be considered together. It concludes by outlining avenues for future research and emphasising the value of new data sources in understanding democratic participation.

# Chapter 2

# Literature Review

## 2.1 Voting In South Africa

Voting is the cornerstone of democratic governance, serving as the primary means through which citizens influence political decisions, hold leaders accountable [2, 1], and shape the policies that govern society [1]. It ensures the peaceful transfer of power and legitimises governmental authority by granting elected officials a mandate to govern [1].

In the South African context, voting plays a particularly vital role in sustaining democratic legitimacy and ensuring inclusive governance [11]. Following the legacy of apartheid, where political participation was racially restricted, the right to vote now symbolises equality, agency, and the ability of all citizens to shape national outcomes [12]. Beyond its symbolic value, voting is used to hold elected officials accountable and influence the direction of government policy, especially in areas such as service delivery and social development [13, 14, 15].

The 2024 South African general election marked a pivotal moment in the country's democratic trajectory, as the African National Congress (ANC) lost its parliamentary majority for the first time since 1994, securing just over 40% of the national vote [3]. This shift underscores a broader trend of political realignment, with voters increasingly moving away from traditional party loyalties and exploring alternative political options [4]. The decline in support for the ANC has opened space for opposition parties and lends support to the notion of the ‘volatile voter,’ characterised by a growing willingness to shift political allegiance between elections [16], especially for younger generations [17].

While research suggests that increased competition results in increased engagement [18], this electoral volatility has not been matched by greater civic engagement. Instead, studies highlight a growing trend of voter apathy, particularly among young and urban populations [17, 19], driven by disillusionment with the effectiveness of elections and a lack of trust in political institutions [20, 4, 5, 21]. This has resulted in a steady decline in voter turnout over the past three decades. Notably, the 2021 municipal elections marked a historic low, with fewer than 50% of eligible voters participating for the first time in the country’s democratic era [3, 11, 16].

This trend of declining voter turnout is not merely a sign of political indifference but may signal a broader crisis of democratic legitimacy [12, 16]. As voter participation weakens, the representativeness and accountability of elected governments are called into question [12]. Despite these challenges, researchers emphasise that active electoral participation remains one of the most accessible and powerful tools for citizens to express dissatisfaction, drive change, and promote a more responsive and accountable state [11, 12].

Understanding the factors that influence voting patterns, both in terms of changing vote choices and declining turnout, is crucial for policymakers, social scientists, and governments seeking to strengthen democratic participation and improve electoral outcomes [17, 6, 22, 11]. As traditional loyalties erode and more citizens opt out of voting altogether, identifying the drivers behind these shifts becomes essential for designing effective interventions that rebuild trust, foster inclusion, and ensure that political institutions remain representative and responsive to the public.

A significant body of research on voting behaviour in South Africa has relied on survey and census data to identify the demographic, socioeconomic, and political factors that shape electoral outcomes [17, 12, 18, 22]. Survey methodologies typically involve collecting responses directly from individuals, capturing self-reported insights into attitudes, motivations, and past voting behaviour. Census data, on the other hand, provides broad demographic and economic indicators at the population level, enabling researchers to identify patterns across regions and social groups.

## 2.2 Lived Experiences and Voting Outcomes

Beyond demographic and attitudinal factors, a growing body of research in South Africa has explored the relationship between service delivery and voting patterns. This approach is grounded in the idea that citizens assess government performance based on their lived experiences — especially the accessibility and quality of public services. In this framework, voting becomes a response to tangible, everyday outcomes rather than abstract political platforms.

Studies have demonstrated a connection between living conditions and voter behaviour, often framed through rational choice theory, which posits that individuals weigh costs and benefits when making political decisions [9, 8, 10, 7, 5, 13]. However, the direct influence of physical living conditions remains relatively understudied in the South African context [10, 23]. Existing models tend to rely on broad proxies — such as access to piped water, sanitation, and refuse removal, often using highly aggregated census data. While informative, such approaches struggle to capture localised variation and the complexity of voter decision-making [10, 23].

In general, advances in computer vision have demonstrated the feasibility of richer, more granular approaches. A notable U.S. study showed that voting district outcomes could be predicted with 80% accuracy by classifying vehicles in Google StreetView (GSV) imagery [24]. These findings highlight the potential of using visual data linked to physical living conditions to model political behaviour. However, no equivalent research has yet been conducted in South Africa, despite the growing availability of GSV imagery [25].

Importantly, GSV itself is not neutral. Its coverage is shaped by several interrelated factors, including commercial, infrastructural, and social considerations, which often mirror broader socio-economic inequalities. As a for-profit service maintained by Google, GSV prioritises areas that drive user engagement, typically those with higher economic or commercial value [26]. Early GSV coverage focused primarily on major U.S. cities, illustrating its market-driven logic [27].

Furthermore, while Google does not publicly disclose its coverage criteria, studies suggest that GSV prioritises accessible, navigable roads — usually in urban environments with existing infrastructure [28]. Areas with limited road access or challenging terrain are less likely to be mapped, exacerbating gaps in representation.

Safety concerns also shape GSV deployment. Research shows that regions perceived as dangerous or affected by high crime rates tend to be underrepresented in GSV imagery, reflecting and reinforcing spatial stereotypes [29]. Together, these factors result in uneven GSV coverage, often systematically excluding areas that are poorer, rural, or marginalised.

This spatial bias has important implications for political research. If there is a link between physical living conditions and voter behaviour and GSV coverage is itself shaped by those same physical and socio-economic conditions, then GSV coverage can serve as a proxy for understanding political outcomes. At the same time, it also reflects broader patterns of exclusion and inequality, highlighting the need for engagement with both the data and the processes that produce it [30].

## 2.3 Theories of Voting Behaviour

Voting behaviour is shaped by a complex interplay of individual, social, and economic factors, leading to the development of various theoretical frameworks that attempt to explain electoral decision-making. Scholars have proposed different frameworks to understand why people vote the way they do, each emphasising distinct influences such as rational decision-making, social structures, economic incentives, and psychological factors. While no single theory can fully account for the nuances of voter behaviour, these frameworks offer valuable insights into the motivations behind electoral participation and political preferences. The following subsections explore key theoretical frameworks, including rational choice theory, clientelism, and social structure determinants. Together, these perspectives contribute to a more comprehensive understanding of the factors that drive electoral behaviour in South Africa and beyond.

### 2.3.1 Rational Choice Theory

Rational choice theory posits that individuals make political decisions by weighing costs and benefits to maximise personal utility [31, 7]. In voting, this means citizens support the party or candidate they believe will best serve their interests, particularly regarding issues like service delivery, economic opportunity, and governance [31].

In South Africa, rational choice theory helps explain how voters respond to material conditions such as unemployment, inequality, and access to basic services. Research shows that these factors influence vote choice, especially among younger and urban voters who increasingly evaluate parties based on perceived performance and economic promises [32, 19].

However, rational choice theory also raises the well-known paradox of voting: if the cost of voting outweighs the likelihood of influencing the outcome, rational individuals should choose to abstain [33, 31]. Yet, empirical evidence shows that many people still vote, suggesting that factors such as identity, civic duty, and social norms often override purely utilitarian calculations. In the South African context, low turnout, particularly among disillusioned youth, indicates that abstention can itself be a rational response when individuals believe their participation will not produce meaningful change [19]. Some studies go further, suggesting that abstention may serve as a deliberate form of political expression, with citizens "withholding their vote as a sign of protest" against the political system or available options [20].

### 2.3.2 Clientelism

Clientelism refers to a political strategy in which resources or services are distributed in exchange for electoral support, typically fostering a reciprocal relationship between politicians and voters [34]. This practice is especially prevalent in contexts marked by poverty and inequality, where voters may prioritise immediate material benefits over programmatic policy platforms [35]. In South Africa, clientelist patterns are often discussed in the context of the ANC and South Africas extensive social grant system. Critics argue that the distribution of grants may function as a tool to consolidate political support, particularly in impoverished communities where state assistance significantly influences livelihoods [35, 32]. This claim is supported by research showing a correlation between grant receipt and ANC support, however, the relationship may not be so straightforward.

Recent literature emphasises this understanding that voters are not passive recipients of benefits, but rather navigate complex trade-offs between loyalty, service expectations, and perceptions of state performance [36]. Moreover, studies highlight regional variation in the clientelist bargain, where expectations of service delivery and identity-based appeals intersect with historical loyalties and dissatisfaction with state performance [37]. These dynamics suggest that clientelism in South Africa may not merely be a top-down strategy of voter manipulation, but part of a broader negotiation shaped by local needs and national political narratives [36, 37].

Furthermore, clientelism often creates conditions that may foster corruption, as its informal exchanges can blur the line between legitimate resource distribution and illicit practices. The mechanisms that secure electoral support through clientelist networks can also facilitate rent-seeking and misappropriation of state resources. This dynamic may encourage officials to prioritise personal loyalty over institutional accountability, thereby undermining transparency and eroding public trust in governance [38].

### 2.3.3 Social Structure Determinants

Social structure theories emphasise that voting decisions are shaped by collective identities, group affiliations, and community contexts rather than by individual rational calculations alone. In South Africa, race remains a key determinant of voter preference. Despite some evidence of racial fluidity, aggregate trends consistently show that racial identity continues to predict party alignment: Black South Africans largely support the ANC, while Coloured, Indian, and white voters are more likely to favour the Democratic Alliance (DA) [39]. This racialised pattern has persisted despite generational shifts, with socioeconomic inequalities reinforcing historical divisions [39].

Gender also plays a role. Although findings are nuanced, studies suggest women are more inclined to support incumbent parties, while men are more likely to back opposition parties [40]. These patterns suggest that identity factors such as race and gender remain embedded in electoral behaviour.

Beyond identity, socioeconomic context further shapes voting. Voters in poor communities are often mischaracterised as politically passive, yet research reveals that they engage with elections in nuanced and context-specific ways, often informed by local service delivery and lived experiences of marginalisation [41]. Social context, including peer influence, family networks, and community leadership, also plays a role, where local disillusionment can normalise disengagement and abstention [18].

Youth electoral behaviour reflects these broader structures. Disillusioned by institutional failures and shaped by socioeconomic exclusion, many young people view voting as ineffective and instead express political discontent through abstention or alternative forms of engagement [19, 17].

### 2.3.4 Psychological Theories of Voting

Psychological theories of voting, such as those developed by the Michigan School [42], emphasise the internal, affective dimensions of voter behaviour. These models focus on party identification, perceptions of political candidates, and the salience of specific issues as key determinants of voting choices. Unlike rational or structural approaches, psychological theories highlight how emotions, long-term loyalties, and cognitive biases shape how individuals interpret political information and make decisions. For example, voters may continue to support a party out of habit or identity alignment, even when its policies no longer serve their material interests.

In the South African context, party loyalty, particularly to the ANC, has historically been shaped by emotional ties to the liberation struggle, often persisting across generations [43, 44]. The 2024 elections have shown a shift in this dynamic, with the uMkhonto we Sizwe (MK) taking the away the significant proportion of ANC votes [3], however the support for the MK can still be explained through phycological theories of voting as the Zulu ethnic population identify with the MK, and in particular the MK's leader Jacob Zuma [45]. As new parties emerge and historical bonds weaken, candidate charisma, personal trust, and issue-based evaluations are becoming increasingly influential, particularly among younger and urban voters [46]. These dynamics are important for interpreting emerging volatility in voting behaviour, where affective detachment may contribute to both vote switching and abstention.

## 2.4 Machine Learning (ML) and Statistical Modelling

machine learning (ML) and statistical modelling have developed along parallel, yet increasingly intertwined trajectories. Statistical modelling has roots in classical mathematics, dating back to the 18th and 19th centuries with the development of probability theory and regression analysis [47]. These models were originally used to understand and infer relationships between variables, underpinned by strong assumptions and an emphasis on interpretability and causal inference. The typical workflow involves specifying a model based on theoretical assumptions, fitting it to data, and interpreting the resulting coefficients to test hypotheses or draw conclusions about relationships. Typical models include Lasso Regression, Ridge Regression, and Elastic Net Regression [48, 49, 50], where Lasso performs feature selection by applying L1 regularisation to shrink some coefficients to zero, Ridge uses L2 regularisation to shrink all coefficients evenly without eliminating features, and Elastic Net combines both L1 and L2 penalties to balance feature selection and coefficient shrinkage — making it particularly effective when dealing with correlated predictors.

ML, by contrast, emerged in the mid-20th century as a subfield of artificial intelligence [51]. Early work focused on pattern recognition and decision trees, eventually expanding with the advent of neural networks and the rise of computational power. Unlike traditional statistical models, ML prioritises predictive accuracy and scalability, often relying on large datasets and high-dimensional features. Its typical workflow includes data preprocessing, splitting into training and testing sets, model training using algorithms such as decision trees or support vector machines, and performance evaluation using metrics like accuracy or mean squared error. Key milestones include the development of support vector machines in the 1990s [52], ensemble methods like random forests in the 2000s [53], and more recently, deep learning architectures that power natural language processing and image recognition tasks [54].

Today, the distinction between statistical modelling and ML is increasingly blurred. Many techniques, such as regularised regression (e.g. Lasso, Ridge), decision trees, and clustering algorithms, are used in both fields. This convergence is particularly valuable for social science applications, where researchers must balance the need for interpretability with the complexity of real-world data. In modelling voting behaviour, for example, both approaches are now used to extract insights from diverse datasets.

### 2.4.1 Use Cases: Social and Political Sciences

Statistical modelling has long served as a foundational tool in the social sciences, particularly for analysing voting behaviour, electoral outcomes, and patterns of political participation. These models have been used to explain voter turnout, party preference, and policy support using variables such as income, education, race, and geographic location [55, 56, 57]. In the study of political behaviour, statistical methods like logistic regression, multinomial logic models, and hierarchical modelling have been applied to examine the influence of demographic and socioeconomic factors on electoral choices. In the South African context, statistical models have been used to link service delivery metrics to voting outcomes, often incorporating census and survey data to understand how infrastructure, inequality, and governance affect political engagement [8, 9, 10]. Despite their interpretability and theoretical grounding, these models are often constrained by data sparsity and the assumptions required for inference, underscoring the need for more flexible, data-rich approaches. [8, 9].

ML has emerged as a complementary approach with the potential to overcome some of these limitations. In the social sciences, ML has proven useful for modelling and predicting collective action and political behaviour [58]. It has been applied to phenomena ranging from activism and consumer movements [58] to protest forecasting using both socioeconomic indicators and real-time social media data [59, 60]. In electoral research, ML techniques have been employed to estimate individuals' propensity to vote, model party switching, and infer political attitudes from digital behaviour [61, 62, 63, 64]. However, in the South African context, the application of ML to electoral modelling remains limited. Existing work, such as Ledwaba et al.'s semi-supervised classification of political sentiment on Twitter during the 2021 local elections [65], offers a noteworthy start, but there remains a need for more comprehensive, multi-source datasets and applied research to capture the full complexity of voting patterns in the country.

## 2.5 Computer Vision

Computer vision is a field within artificial intelligence that focuses on enabling machines to interpret and understand visual information from the world. Its origins trace back to the 1960s, with early efforts focused on basic image processing tasks such as edge detection and object recognition in constrained environments [66]. As computing power and algorithmic techniques advanced, so did the capabilities of computer vision systems. The 2010s saw a major breakthrough with the rise of deep learning, particularly convolutional neural networks, which dramatically improved performance in tasks like image classification, segmentation, and object detection [67]. These advances enabled computer vision to transition from theoretical experimentation to real-world applications across fields such as healthcare, agriculture, and urban planning.

One of the most influential developments in modern computer vision is the emergence of the You Only Look Once (YOLO) framework [68]. Unlike earlier object detection methods that separated region proposal and classification into multiple stages, YOLO performs both tasks simultaneously in a single forward pass of the network [69]. This architecture allows for real-time object detection with high accuracy, making it particularly effective for large-scale visual analysis. Over successive versions, and more recently with You Only Look Once, Version 8 (YOLOv8), the model has improved in speed, precision, and generalisation capability. In this study, YOLOv8 is used to extract structured features from GSV imagery, enabling the measurement of visual indicators of infrastructure and service delivery.

### 2.5.1 Use Cases

Computer vision has emerged as a powerful tool for analysing the built environment and its relationship with social and political outcomes. Leveraging sources such as satellite and GSV imagery, researchers have applied deep learning models to detect and quantify physical features of neighbourhoods — such as building types, roads, cross walks, and signs of physical disorder and link these to broader socioeconomic indicators like income, health, and inequality [70, 71, 72, 30]. For example, Nguyen et al. [70] processed over 31 million GSV images across the United States to assess built environment quality and its association with health outcomes, finding links between infrastructure indicators (like cross walks or visible utility wires) and measures such as obesity, physical inactivity, and premature mortality. Similarly, Abitbol and Karsai [71] predicted socioeconomic status across French cities using high-resolution aerial imagery and convolutional neural networks, interpreting visual model activations to uncover associations between urban form and wealth. Finally, one notable exception is a study from the United States, which demonstrated that classifying vehicles in GSV imagery could predict district-level political preferences with up to 80% accuracy [24]. These approaches provide a scalable and spatially detailed alternative to traditional surveys, enabling fine-grained mapping of social inequalities.

# Chapter 3

# Methodology

## 3.1 System Overview

The goal of this system is to develop models for understanding and predicting voting patterns by processing data that either directly captures, or serves as a proxy for the factors influencing voter behaviour. Based on the considerations from the literature review and the gaps identified, this system integrates a diverse range of data sources to improve the accuracy and, to some degree, granularity of electoral analysis. In particular, Google StreetView (GSV) imagery, satellite imagery, and publicly available datasets, provide complementary insights into the physical environment, socioeconomic conditions, and infrastructure quality that may influence voter behaviour. By incorporating multi-modal data, this approach extends beyond traditional survey-based electoral modelling, enabling a data-driven and spatially aware approach to modelling voting patterns. The system consolidates all data into a single database, facilitating input into machine learning (ML) and statistical models to uncover patterns and trends at different geographical levels.

As shown in Figure 3.1, the system consists of three main components: data/feature collection, feature storage and modelling, and result visualisation.

i. **Data/Feature collection:** The system collects data from three primary sources, each offering distinct but complementary insights into voter behaviour:

a) Satellite imagery — provides information on land use.

b) GSV imagery — captures ground-level visual characteristics that may reflect living conditions and service delivery.

c) Publicly available datasets — includes financial data and crime statistics.

ii. **Feature storage and modelling:** Once collected, the data is processed and integrated into a single feature store, where feature engineering is applied to derive additional meaningful variables. The processed features are used to model voting patterns at different geographical levels.

iii. **Result visualisation:** The final stage of the system involves interpreting and presenting the model outputs, enabling a clearer understanding of voting patterns across different regions. The results are visualised to highlight key insights, trends, and relationships between features and voting outcomes.

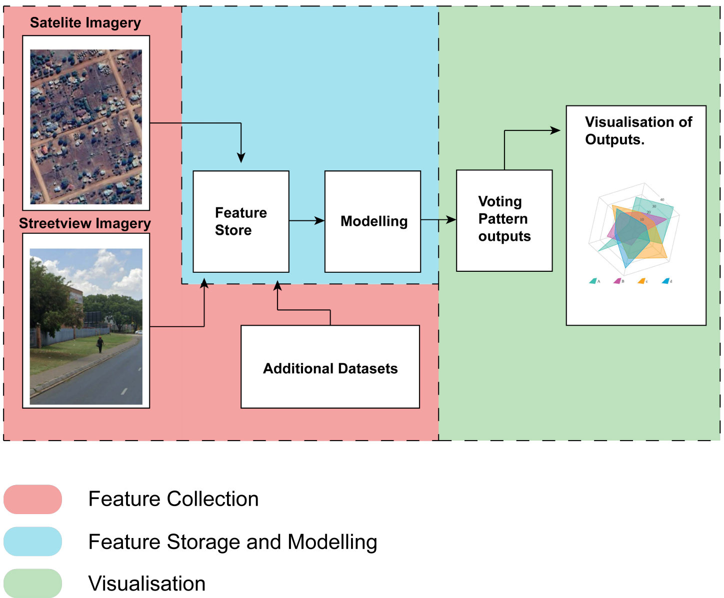


FIGURE 3.1: A macro view: flow diagram of the system highlighting data collection, feature generation, model building and visualisation

## 3.2 Voting Patterns and Feature Collection

The system will model two key voting patterns: party percentage and non-vote percentage. For party percentage, Figure 3.2 illustrates the relative vote share of the four main political parties in South Africa and the summation of the remaining parties, where the African National Congress (ANC), Democratic Alliance (DA), Economic Freedom Fighters (EFF), and uMkhonto we Sizwe (MK) obtained majority votes. Based on this, the will model the vote percentage for the ANC, DA, EFF, and MK. Additionally, the system will model the non-vote percentage, capturing the proportion of registered voters who abstained from voting.

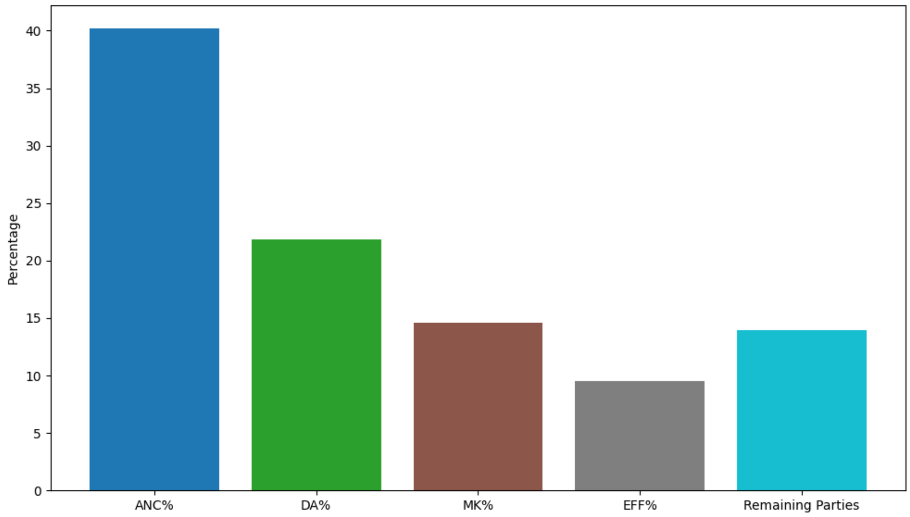


FIGURE 3.2: Outcome of the 2024 national elections

When modelling voting patterns in South Africa, the geographical resolution of the analysis is determined by the level at which the Independent Electoral Commission (IEC) releases election results [3]. These results are available at five main administrative levels: national, provincial, municipal, ward, and voting district. Each level offers a different balance of granularity and coverage, with varying advantages and limitations. Table 3.1 shows the number of units that South Africa is split into at each administrative level.

The national and provincial levels provide a high-level overview, making them less suitable for detailed analysis. Conversely, the voting district level is too granular, resulting in a resolution that does not capture enough information or detail per administrative units.

For practical modelling, a balance must be struck, with the municipal or ward level offering an optimal trade-off between detail and number of divisions.

Table 3.1: Number of geographical units available for electoral analysis at different administrative levels

| Level | Number of Administrative Units |
|---|---:|
| National | 1 |
| Provincial | 9 |
| Local Municipal | 213 |
| Ward | 4 468 |
| Voting District | 22 503 |

1. **Municipal Level:** While the municipal level provides a broader overview as compared to the ward level, it does not offer enough adminstrative units (213) for highly granular analysis. However, one can obtain more data and information at this level compared as to the ward level. Due to the relatively lower number and larger administrative units, this level is more suited for statistical regression models, which require fewer observations while still capturing general trends.

2. **Ward Level:** The ward level offers finer granularity, allowing for more detailed spatial analysis. Unlike the municipal level, the ward level provides a significantly higher number of administrative units 4 468, making it more suitable for ML models. The larger dataset at this level enables the models to be more reliable, as they are trained on a greater number of data locations, improving the generalisability and robustness of predictions.

Additionally, Figure 3.3 presents a map outlining the municipal and ward boundaries, illustrating the different levels of geographical resolution available for analysis. At the municipal level the units are composed of local and metro municipalities and at the ward level, the units are the geographical subdivisions of municipalities used for electoral purposes.

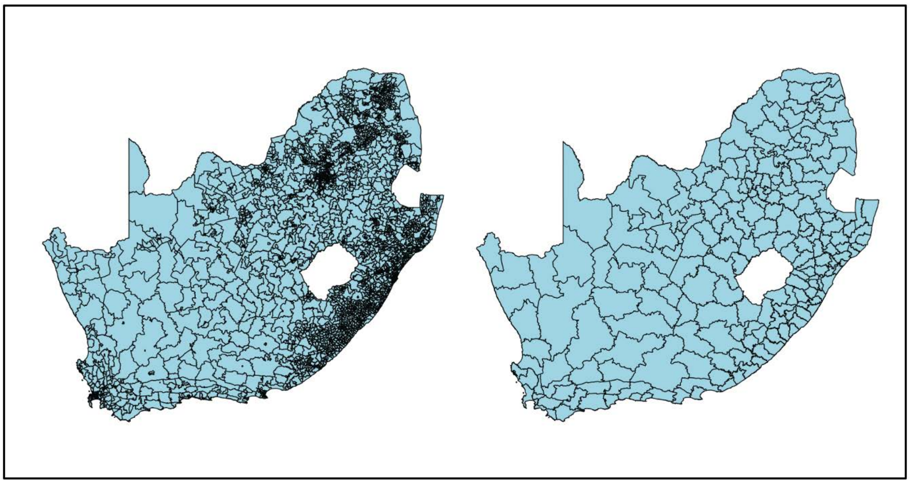

FIGURE 3.3: Comparison of the administrative units between the ward (left) and municipal (right) level in South Africa

At each of these levels, voting results are represented as a percentage of valid votes. For example, at the national level, which encompasses the entire country, the ANC secured 40.18% of valid votes, whereas in the Overstrand local municipality, the ANC secured 17,5% and the DA secured 63.5% of valid votes. Similarly, just as party vote percentages can be obtained for each of the levels, the non-vote percentage, which excludes both valid and spoilt votes as a proportion of registered voters, can also be calculated. This is expressed in Equation (3.1) .

$$\text{Non Vote Percentage} = (1 - (\frac{\text{Spoilt Votes} + \text{Valid Votes}}{\text{Registered Voters}})) \times 100 \tag{3.1}$$

## 3.3 Structure of the Remainder of the Methodology Chapter

The rest of this chapter will explain the methods including data acquisition, preprocessing, feature engineering, model selection, and evaluation. Listed below are the sections covered.

i. **GSV Image Collection:**

This section details how the GSV application programming interface (API) is utilised to collect GSV images across South Africa. It also describes the methodology for selecting image locations (see Section 3.4).

ii. **GSV sample size calculation:**

This section details the sample size calculation for the GSV images. It also explains how additional samples can be incorporated into the system following the initial round of image collection (see Section 3.5).

iii. **Additional data sources:**

This section details the external datasets explored and integrated into the system, including financial and crime data (see Section 3.6).

iv. **Land use data:**

This section describes various potential techniques for extracting land-use information from satellite imagery and explains the approach adopted in this thesis (see Section 3.7).

v. **Computer vision:**

This section provides an overview of the development of the different computer vision models used to extract features from GSV images. Including those designed to detect potholes and illegal dumping (see Section 3.8).

vi. **ML and statistical regression:**

This section outlines the methodology for developing ML models for ward-level analysis and statistical regression models for municipal-level analysis (see Section 3.9).

## 3.4 Google StreetView Image Collection

The GSV Static API [73] is a service provided by Google that allows developers to embed street-level imagery into their applications. The GSV API provides two primary endpoints for retrieving street-level imagery:

a) **Get Image URL**: This endpoint is used to retrieve the actual GSV image for a specified location. Each request incurs a cost of USD \$7 per 1 000 images. However, Google provides a monthly credit of USD \$200, which allows users to request approximately 28 500 images before incurring charges.

b) **Get Metadata URL**: This endpoint allows users to check whether a GSV image exists at a given location. Unlike the image retrieval endpoint, metadata requests are free and can be called an unlimited number of times. The metadata response includes details such as image availability, the closest image location, and capture date.

Given the cost structure of the API, an efficient approach is to first query the metadata endpoint to check whether an image exists at a specified location. If an image is available, the system can then proceed to request the actual image using the Get Image URL. This practice reduces unnecessary paid requests, optimising both cost and API usage.

The GSV API retrieves images through Hypertext Transfer Protocol (HTTP) requests, where developers specify parameters in the request URL to control key aspects of the image, such as location, size, direction, and Field of View (FOV).

The core parameters of the API include:

i. **Location**: Specifies the geographic location for the requested image, using latitude and longitude coordinates or an address. The API searches within a 50m radius for the closest available imagery. If no image is available within this radius, the API returns an error.

ii. **Size**: Defines the crop of the returned image, for example 650X650 returns the full uncropped image and 325x650 returns a image cropped across the horizontal line.

iii. **Heading**: Determines the compass direction the camera is facing, where 0° is north, 90° is east, 180° is south, and 270° is west.

iv. **FOV**: Controls the horizontal viewing angle of the image, with a range from 10° to 120°. A smaller FOV results in a zoomed-in effect, while a larger FOV provides a wider perspective. Usage of GSV through Google Maps utilises a FOV of 120°.

v. **Pitch**: Adjusts the vertical angle of the camera. A value of 0° keeps the camera level with the horizon, positive values tilt the view upwards, and negative values tilt it downwards.

vi. **Key**: A required API key that authenticates requests and ensures access to the Google Maps platform.

Since the system will collect a large number of images, locations will be specified using geographic coordinates rather than addresses to ensure precise and consistent retrieval. To maximise the amount of visual data captured, the full image resolution is used, the FOV is set to 120° for the widest possible perspective, and the pitch is set to 0° to maintain a level viewpoint.

### 3.4.1 Exploratory Methodology

Before finalising the methodology for image collection, several approaches were explored to identify the most effective way to acquire GSV imagery at scale. This exploratory phase involved testing different methods for selecting image locations, optimising image requests, and determining the appropriate camera heading. While some approaches proved unsuccessful, they provided valuable insights that helped refine the final methodology, ensuring it is both scalable and efficient.

*Grid and Random Location Generation*

Since the system requires collecting a large number of images, it needs a scalable and dynamic method for determining coordinate locations. Initially, two approaches were explored: a grid location and random location generation approach.

In the grid approach, a structured grid is overlaid on the area of interest, and image locations are retrieved at the intersections of the grid. The primary advantage of this approach is its systematic nature, allowing for uniform distribution of image locations. However, once an existing grid has already been established, challenges arise when attempting to increase the number of locations. This results in either an exponential increase in the number of locations or the breakdown of the spacial relationship between the locations.

In contrast, the random approach places coordinate locations randomly within the area of interest. This method introduces variability and may increase the likelihood of capturing diverse locations, but it lacks the structured coverage provided by the grid approach. However, an advantage of the random approach is that it allows for a linear increase in the number of locations, making it easier to scale without affecting spatial relationships. But, the lack of a systematic structure may lead to uneven coverage and inefficiencies in data collection.

Both approaches have limitations, particularly in cases where available GSV imagery does not align with the generated coordinate locations. Figure 3.4 shows the output of these two methods for a Ward in the City of Tshwane Metropolitan Municipality.

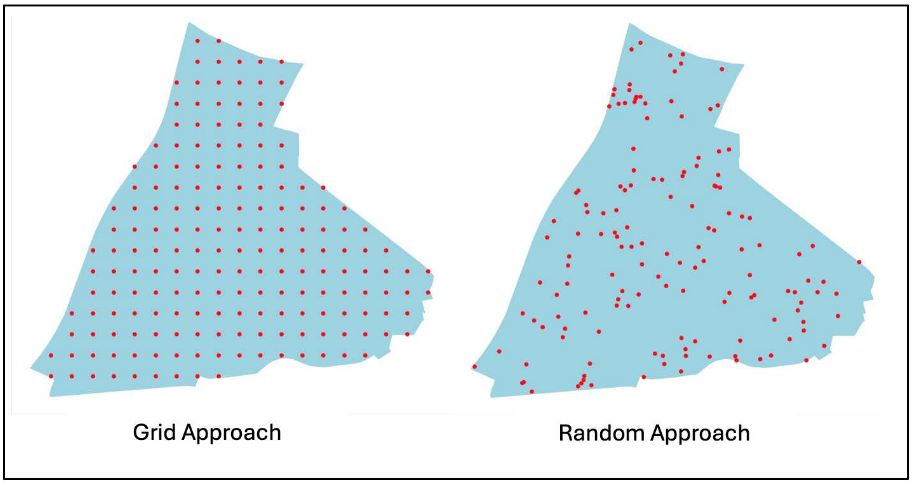


FIGURE 3.4: Output of location generation of the grid (left) and random (right) approach on a ward in the City of Tshwane Metropolitan Municipality

Once the locations have been generated, there is no guarantee that each one will be within 50 m of a road covered by the GSV API. To address this, each coordinate can first be queried using the GSV Metadata API to check whether imagery is available at that location. Only locations that return a valid response are retained for image retrieval. The resulting set of valid locations, representing locations where GSV images can be obtained in Figure 3.4, is shown in Figure 3.5, which shows a reduction of locations from the original image.

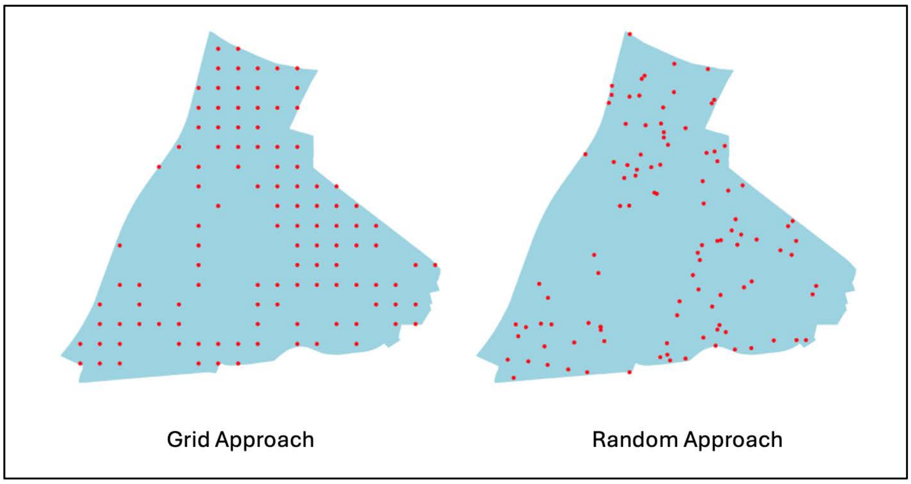


Figure 3.5: Output of location generation of the grid (left) and random (right) approach when only valid locations are retained in the City of Tshwane Metropolitan Municipality

Both approaches have a major drawback in that they are computationally intensive and wasteful. Each generated point must be individually queried against the GSV API to check for image availability. Given that a significant proportion of locations will not correspond to locations covered by GSV, a large number of unnecessary requests are made, leading to considerable processing overhead. To address this inefficiency, a second approach was explored, where locations are first linked to known road networks before being tested for availability in the GSV API. This significantly reduces the number of queries and improves the overall efficiency of the image collection process.

*Linking Locations To Roads*

Building on these findings, the random search approach was selected as the most effective method for generating locations. This decision was driven by its ability to scale linearly, allowing additional locations to be incorporated without disrupting the spatial distribution of existing locations. Unlike the grid search approach, which requires an exponential increase in locations to refine coverage, the random search approach enables straightforward expansion; if N additional locations are needed, they can simply be generated independently. Since these locations are placed randomly, they do not introduce bias or alter the spatial significance of previously selected locations.

To further improve this approach, once the locations are randomly generated, each point is adjusted to align with the nearest road to that point as shown in 3.7, regardless of whether that road is covered by the GSV network. The road network is sourced from OpenStreetMap [74], which includes 1 877 451 distinct roads across South Africa as shown in 3.6, which shows road networks at different administrative levels. By ensuring that generated locations are positioned on actual roads, this method increases the likelihood of successful image retrieval while maintaining a structured and scalable methodology.

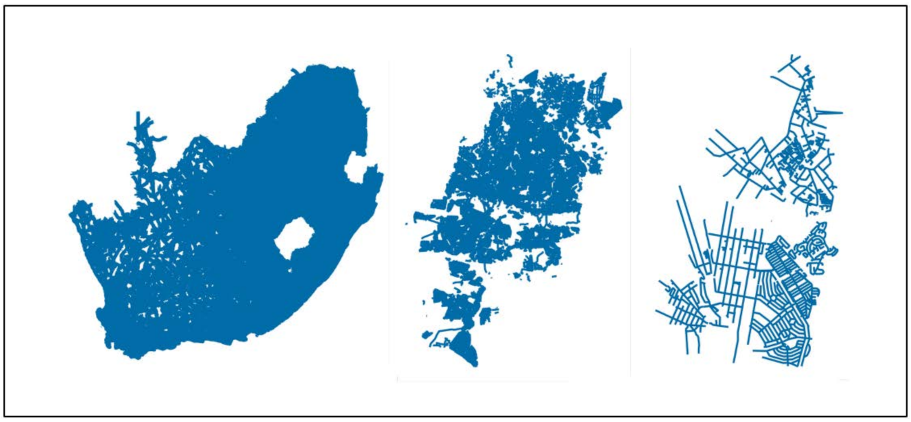

FIGURE 3.6: Road network South Africa (left), the City of Johannesburg (middle), and a ward in the City of Johannesburg (right)

The approach of snapping locations to roads significantly improves computational efficiency by reducing the number of unnecessary queries to the GSV API. However, applying this technique in rural areas or regions with highly variable road densities presents certain challenges. Figure 3.8 illustrates these challenges in Beaufort West local Municipality, which includes vast open spaces alongside small, dense road networks. For example, in Murraysburg, a town within Beaufort West local Municipality, the method tends to concentrate locations on the outskirts rather than within the town centre, where road networks are denser. This can result in an under-representation of central areas, which are often more reflective of the lived experiences of residents.

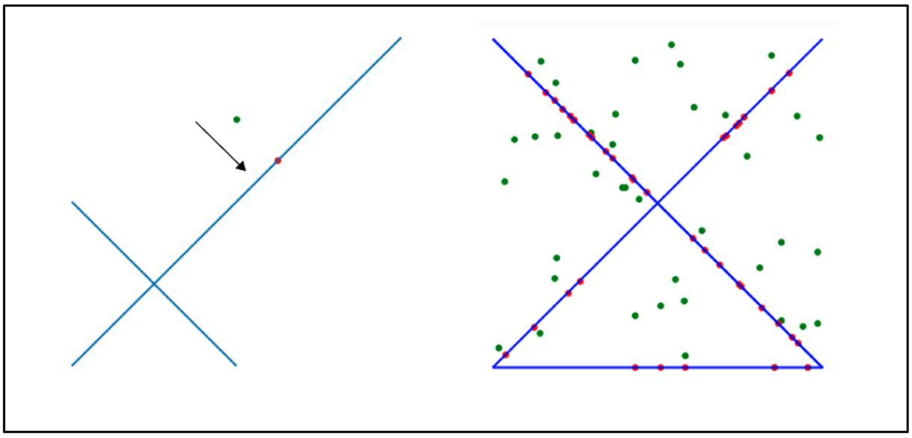

FIGURE 3.7: Snapping random locations in green (left) to hypothetical roads in red (right) using Geopandas

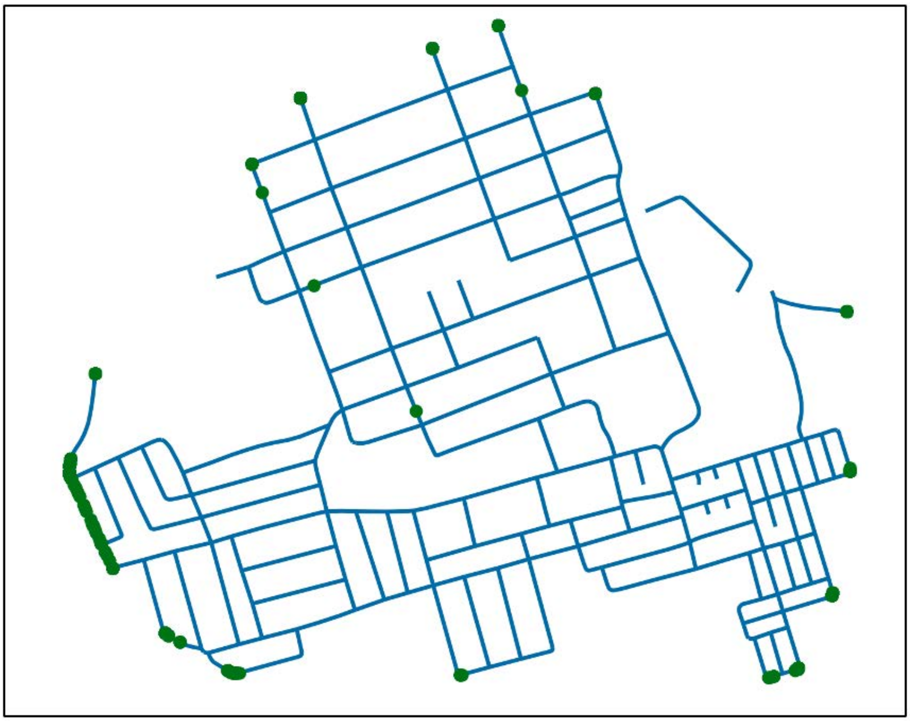

FIGURE 3.8: Result of snapping locations to Murraysburg town in Beaufort West Local Municipality, illustrating points being snapped to the periphery of the town

The under-representation of central areas occurs because randomly generated locations are initially distributed across wide open areas. When these locations are snapped to the nearest roads, they often align with the closest available road segment, which in low-density areas, like Beaufort West Local Municipality, is likely to be on the periphery of settlements rather than within the town centres. As a result, only a limited number of locations end up within the core urban areas, while the majority cluster around the town boundaries. This skews data collection and reduces the effectiveness of the system in capturing features within town centres.

*Random Roads*

To address the issue of locations concentrating on peripheries, a more effective approach was implemented. Instead of randomly generating locations across an area and then snapping them to the nearest roads, the method was refined to first randomly select roads and then place locations randomly along those selected roads. This new methodology offers two key benefits, first, it reduces computational complexity by eliminating the need for an additional snapping step, and secondly, it ensures a more balanced selection of roads, preventing the disproportionate placement of locations on peripheral roads while improving representation within town centres.

To achieve this, roads are randomly selected with a weighting proportional to their length. For example, if an area contains two roads, one of which is twice as long as the other, the longer road is twice as likely to be selected. This method preserves the spatial distribution characteristics of the original random placement approach (see 3.4.1) while mitigating the bias toward peripheral roads. Table 3.2 shows the weighted probabilities for three hypothetical roads based on their length.

TABLE 3.2: Table showing the length, random probability, and weighted probability of three roads

| Road | Length | Random Probability | Weighted Probability |
|---|---:|---:|---:|
| 1 | 10 | 1/3 | 1/6 |
| 2 | 20 | 1/3 | 1/3 |
| 3 | 30 | 1/3 | 1/2 |

### 3.4.2 Sourcing Methodology

While Section 3.4.1 refined the process for determining image locations, additional improvements were made to ensure the methodology is scalable, efficient, and reliable. The final approach optimises both the selection of image locations and the retrieval process, reducing computational overhead and minimising unnecessary API requests. This is achieved by filtering out inapplicable roads, such as footpaths and national highways, removing roads not covered by GSV, and introducing parallel processing to improve efficiency.

By integrating the lessons learned from exploratory approaches, this methodology ensures that GSV imagery is collected in a way that maximises coverage while maintaining efficiency. This section outlines the step-by-step process used to collect images, detailing how locations are selected, validated, and retrieved using the GSV API.

*Road Selection*

The road network obtained from OpenStreetMap [74] includes all roads available in South Africa. However, for the purpose of this study, only residential roads are considered applicable, as they are the most representative of the lived experiences of voters. Roads such as primary roads and motorways are excluded because they are maintained by the provincial government rather than the municipal government, making them less relevant to the analysis of local voting patterns.

Table 3.3 provides an overview of the different road classifications included in the dataset, highlighting the need to filter out non-residential roads to ensure meaningful and representative data collection.

Table 3.3: Selected examples of road types in OpenStreetMap

| Category | Example Road Types |
|---|---|
| Major Roads | Motorway, Primary, Trunk, Secondary, Tertiary |
| Residential Roads | Residential, Living Street, Unclassified |
| Pedestrian | Footway, Pedestrian, Steps, Cycleway, Path |
| Service Roads | Service, Track, Bridleway, Busway |
| Highway Links | Primary Link, Secondary Link, Tertiary Link |

The benefits of this filtering approach can be seen in Figure 3.9, which visualises the residential and non-residential roads in Beaufort West local Municipality, showing that the majority of roads are non residential. Without filtering, the majority of images would be collected from highways and major roads, which do not typically reflect the local environment where voters live. Since these roads are designed for transit rather than residential life, they lack the visual and infrastructural characteristics that are most relevant to understanding voter experiences. By restricting image collection to residential roads, the dataset more accurately captures the physical and socioeconomic conditions that influence voting behaviour, arguably, improving the quality and relevance of the analysis.

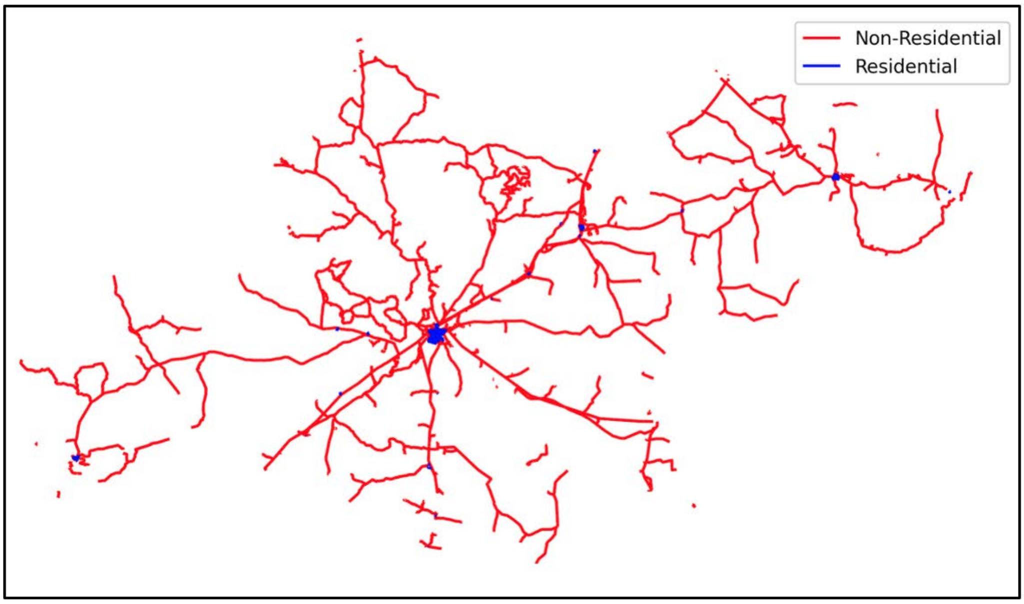


FIGURE 3.9: Map showing residential and non-residential roads in Beaufort West Local Municipality

*Google StreetView Road Pre-Selection*

To further enhance the efficiency of the system, a more targeted approach was implemented for selecting image locations. Instead of randomly selecting roads, generating locations, and only then checking their availability in GSV, a pre-filtered dataset was created containing only roads that are available in GSV. This refinement significantly reduces computational overhead and minimises unnecessary API requests, ensuring that generated locations are more likely to yield usable imagery.

Figure 3.10 illustrates the necessity of this approach by showing a screenshot of the Google Maps interface, which illustrates the roads available in GSV, highlighted in blue.

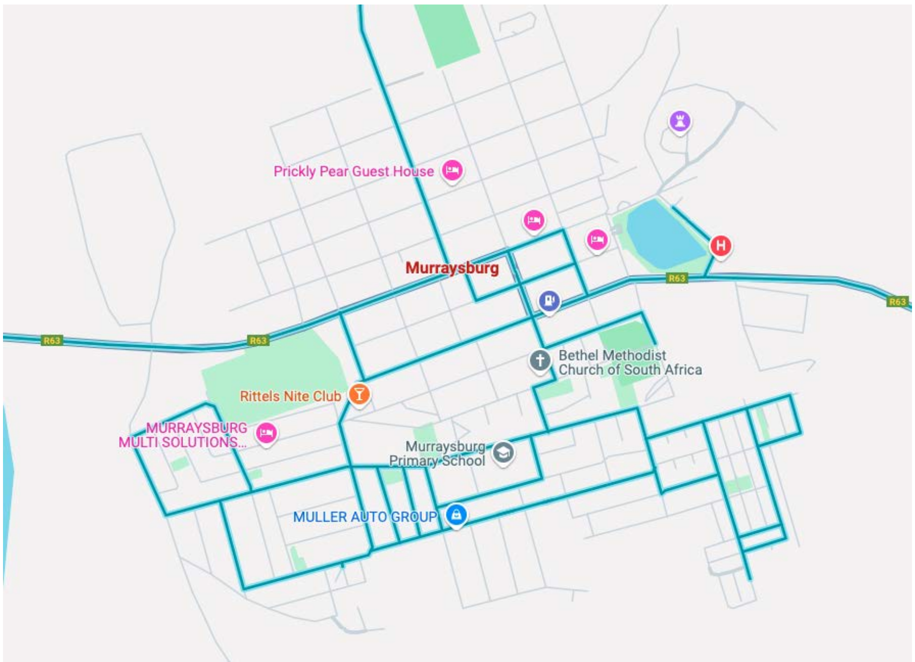


FIGURE 3.10: Screenshot of Google Maps highlighting roads that are available in GSV in Murraysburg, South Africa

To identify roads accessible via GSV, two locations are placed on each road segment from the OpenStreetMap dataset [74]. The first point was positioned a quarter of the way along the road, while the second was placed three-quarters of the way down. If both locations were found in the GSV Metadata API, the road segment was classified as available in GSV. Consequently, any point placed along such roads would reliably yield an image, streamlining the selection process and reducing unnecessary API queries. This resulted in a refined dataset containing 274 168 residential roads, down from the initial 967 688.

However, in more rural areas, the only roads consistently represented in GSV are the secondary and tertiary roads that pass through towns, as illustrated in Figure 3.11, which shows that a tertiary road is available in GSV, while residential roads are not available. This posed a challenge, as relying solely on residential roads would exclude a significant portion of available GSV coverage in these areas. To address this urban bias, it became necessary to incorporate tertiary and secondary roads into the dataset, ensuring broader coverage and a more representative distribution of sampling locations.

FIGURE 3.11: Screenshot of Google Maps highlighting that only a tertiary road is available while residential roads are not in a rural town in KwaZulu-Natal, a province in South Africa

The selection of secondary and tertiary roads for inclusion in the dataset was based on two key characteristics: their proximity to residential centres and their overall length. The goal was to exclude roads that were too far from residential areas or did not pass through them, ensuring that selected roads were more likely to yield relevant GSV imagery.

To select relevant tertiary and secondary roads, a function was developed to measure the proximity of secondary and tertiary roads to residential roads. Specifically, the function calculated the number of residential roads within a 500 m radius of each secondary or tertiary road. Through trial and error, it was determined that secondary and tertiary roads with more than ten residential roads nearby were likely to pass through residential areas and should be included in the dataset.

However, proximity alone was not a sufficient criterion. Many roads that briefly passed through residential areas extended far beyond them, meaning that most locations placed on such roads would fall outside areas of interest. To address this, a length-based filter was applied to further refine the selection.

Table 3.4 presents the mean, median, standard deviation, and range of secondary and tertiary road lengths in South Africa. Based on this analysis and further empirical testing, it was found that roads longer than 6 000 m, while sometimes passing through towns or residential areas, were predominantly located outside such areas. As a result, these longer tertiary and secondary roads were excluded to ensure that selected roads primarily contributed relevant GSV imagery.

Table 3.4: Statistical characteristics of tertiary and secondary roads in South Africa

| Road Type | Mean (M) | Median (M) | Standard Deviation (M) | Range (M) |
|---|---|---|---|---|
| Secondary | 1 166 | 94 | 5 210 | 213 172 |
| Tertiary | 2 082 | 175 | 5 288 | 158 928 |

This resulted in a dataset containing a total of 327 154 residential, tertiary and secondary roads, an increase of 152 986 secondary and tertiary roads.

*Location Proximity*

To further refine the location selection process, an additional filtering criterion was introduced to ensure that selected locations were not too close to one another. This was necessary to prevent redundancy, as images captured from locations in close proximity often contained similar information.

To ensure that images were not too close to one another, a proximity filter was implemented. When generating new locations, any point falling within 50 m of an existing selected point was excluded. Instead, an alternative location was chosen that maintained a minimum separation of 50 m from all previously selected locations.

The threshold of 50 m was determined through visual inspection of GSV images. It was observed that a single GSV image could capture meaningful information within a 50 m radius. Thus, enforcing this spacing ensured that each image contributed unique and valuable information while avoiding duplication.

*Image Bearing*

The heading parameter in the GSV API determines the compass direction the camera faces when capturing an image. The heading is specified in degrees, where 0° is north, 90° is east, 180° is south, and 270° is west. Selecting an appropriate heading is crucial for ensuring that images capture as much relevant information as possible.

One approach to ensuring that the camera would head in the direction of a road is to capture four images at each location, with each image oriented in a different and opposite direction (e.g., 0°, 90°, 180°, and 270°). This guarantees that at least one image, or a combination of two images, aligns with the road's direction, as illustrated in Figure 3.12.


FIGURE 3.12: Output of getting four images per location, showing that at least one image or combination of two will point in the direction of a road (Parkview in City of Johannesburg).

This approach comes with a significant drawback: it quadruples the number of images required per location. Given the GSV API's monthly limit of 28 500 images, capturing four images per location would reduce the total number of unique locations by a factor of four, limiting overall dataset coverage. Therefore, an alternative, more efficient method was needed to determine the correct heading ultimately reducing the number of images at each location to a single image.

To ensure that each GSV image is captured in the correct direction, the bearing (or heading) of each image location is calculated based on the road network from OpenStreetMap [74]. Since roads are represented as LineString objects in GeoPandas [75], each LineString can be decomposed into its constituent vertices, which define the shape of the road. The bearing is then computed using the nearest road vertex to each image location, ensuring that the camera is aligned with the road's direction. This process is described below:

i. **Extracting road vertices:** Each road segment is represented as a LineString containing a sequence of vertices (coordinate locations).

ii. **Finding the nearest road vertex:** For each image location, the closest vertex on the road segment is identified by computing the Euclidean distance between the image location and all vertices along the road.

iii. **Computing the bearing:** The bearing, $B$, between the image location $(latitude1, longitude1)$ and the nearest road vertex $(latitude2, Longitude2)$ is calculated using the following trigonometric functions (3.2) (3.3) (3.4) (3.5). Where $x$ and $y$ are in radians.

iv. **Normalising the bearing:** Since bearings are typically measured in degrees from 0° to 360°, the final value is converted to degrees and normalised to fall within this range as shown in these equations (3.6) (3.7).

$$\Delta longitude = longitude2 - longitude1 \tag{3.2}$$

$$x = \sin(\Delta longitude) \cdot \cos(latitude2) \tag{3.3}$$

$$y = \cos(latitude1) \cdot \sin(latitude2) - (\sin(latitude1) \cdot \cos(latitude2) \cdot \cos(\Delta longitude)) \tag{3.4}$$

$$B^c = \tan^{-1}\left(\frac{x}{y}\right) \tag{3.5}$$

$$B^\circ = \frac{B^c \times 180}{\pi} \tag{3.6}$$

$$B^\circ = (B^\circ + 360) \mod 360 \tag{3.7}$$

This approach ensures that each image is aligned along the direction of the road, providing the most informative and relevant perspective.

Figure 3.13 illustrates the results of applying the bearing technique to four random locations across South Africa, showing the camera pointing in the direction of the road for all four images.

FIGURE 3.13: Output of four images using the bearing approach showing the camera pointing in the direction of the road

*Parallel Programming*

Given the scale of the dataset, parallel processing was implemented to significantly reduce computation time. The dataset contains 327 154 roads, and multiple locations may be placed along each road. While processing a small number of locations, such as adding 100 locations to a dataset of 1 000 roads and retrieving corresponding images, remains computationally feasible, scaling to the full dataset and retrieving over 1 000 images results in excessive runtime.

To address this, Python's `concurrent.futures` module [76] was used. This approach enabled the simultaneous execution of multiple tasks, distributing the workload across multiple threads. By parallelising the processes of generating locations using 8 cores of a CPU (2021 MacbookPro M1 pro), checking GSV availability, and retrieving images, computation for the entire system time was reduced from several hours to under 30 minutes.

This parallel execution framework ensures that large-scale image collection remains scalable and practical, allowing the system to process the full dataset without excessive delays.

## 3.5 Google StreetView Sample Size Calculation

Determining the correct sample size is crucial to ensuring that the collected GSV images are representative and statistically reliable. Given the vast road network in South Africa and the GSV API's monthly image limits, capturing images for every road is impractical. A well-defined sampling approach is therefore necessary to balance statistical validity, computational efficiency, and cost-effectiveness.

A sample size that is too small risks loss of statistical power [77], increasing the likelihood of missing key relationships between environmental factors and voting behaviour. It may also lead to overestimated effects, where trends appear exaggerated due to skewed or limited data. Conversely, an excessively large sample size strains computational resources, incurs unnecessary costs, and may not yield proportionate improvements in accuracy [78].

Due to the variability between government, population, road networks, and living conditions in municipalities [79], it is not feasible to determine a single sample size for the entire country. Instead, each municipality is treated as a distinct sample group, requiring an individual sample size calculation. To achieve this, a stratified sampling approach is used [80], where a sample size is determined for each municipality, and the total sample size for South Africa is obtained by summing the sample sizes across all municipalities.

This ensures that data collection is proportional and accounts for regional differences in road networks, population distributions, land usage and socioeconomic variables.

When calculating the sample size for each municipality, two methodologies can be considered:

i. **Adjusted population sample size:** The sample size is scaled based on the municipality's population, meaning that smaller municipalities have proportionally smaller sample sizes.

ii. **Infinite population sample size:** Assumes an infinitely large population, resulting in a maximum possible sample size.

To ensure a statistical representation and minimise the risk of under-sampling, this study employs the infinite population approach, which allows for a larger dataset while maintaining methodological consistency across municipalities. The formula to calculate the infinite sample size is shown in equation (3.8).

$$n = \frac{z^2 \times \hat{p}(1-\hat{p})}{\varepsilon^2} \tag{3.8}$$

where $n$ is the infinite sample size, $z$ is the z-score, $\varepsilon$ is the margin of error, $\hat{p}$ is the population proportion.

To ensure statistical representation, the sample size is maximised by setting the $\hat{p}$ value to 0.5. The z-score is set to 1.96, corresponding to a 95% confidence level, and the margin of error is set at 5%, following standard research practices [80, 81]. Using these parameters, the required sample size per municipality is 385 images, leading to a total of 82 390 images for South Africa.

*Roads As A Proxy For Population Size*

Since direct population sampling is not feasible within the constraints of this study, roads are used as a proxy for population distribution. Road networks tend to be denser in urban areas, where population density is higher, and sparser in rural areas with lower population concentrations. This relationship is supported by figure 3.14, which illustrates a strong correlation between the normalised population size and the number of roads per municipality [74, 3].

By leveraging this correlation, the sampling methodology ensures that data collection is proportionally distributed across municipalities. The use of road networks as a population proxy provides a practical means of ensuring that the selected sample reflects variations in both infrastructure and population density. This approach enhances the representativeness of the dataset while maintaining consistency across diverse geographic regions.

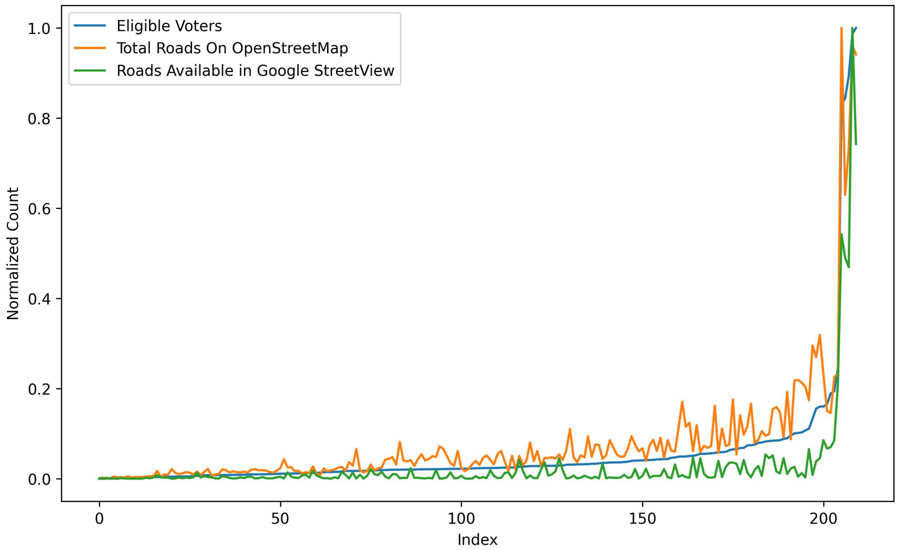


FIGURE 3.14: Correlation between normalised population size [3] and number of roads per municipality in South Africa [74], showing a strong correlation between roads and eligible voters

*Adjusting Sample Size for Municipalities with Limited Road Coverage*

While a sample size of 385 images per municipality is ideal, some municipalities have fewer than 385 roads available in GSV, as illustrated in Figure 3.15. In such cases, maintaining the standard sample size would lead to oversampling specific roads, reducing the diversity of captured environments and increasing the danger of statistical overfitting.

To mitigate oversampling roads, an adaptive sampling approach is applied. If a municipality has fewer than 192 roads available in GSV, the sample size is adjusted to twice the number of available roads (i.e., sample size = number of GSV roads $\times$ 2). This ensures that each road contributes multiple images while preventing excessive repetition of individual locations (note: 3.4.2). This approach maintains a balance between data availability and representativeness, ensuring that municipalities with limited road coverage are still meaningfully included in the dataset.

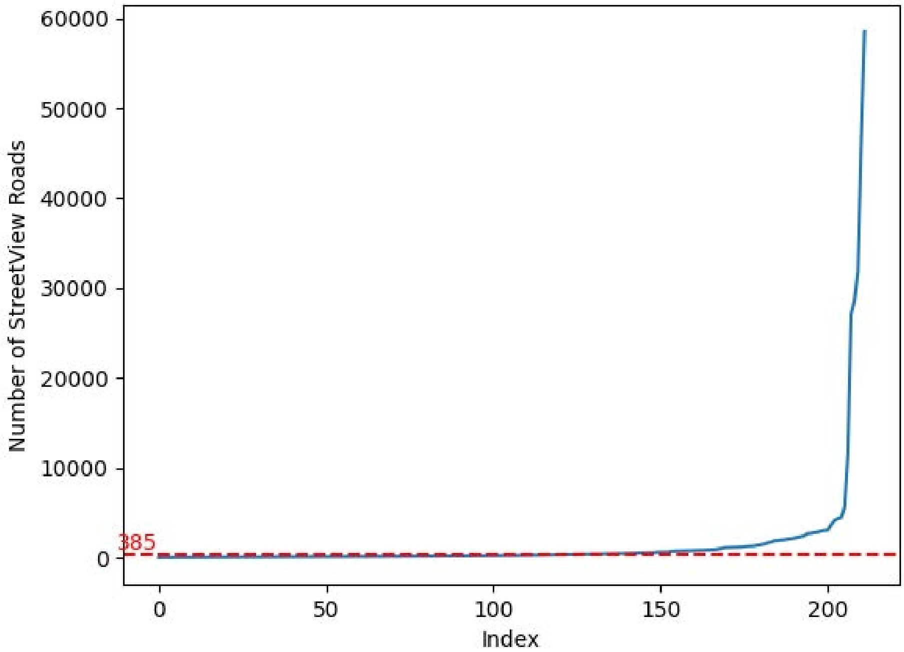


FIGURE 3.15: Line graph showing the number of roads available for each municipality in South Africa, represented as a index from least roads to most roads and a static horizontal line at 385 showing the infinite sample size

*Adding Additional Samples*

Once the initial sample size has been met, the system allows for incremental expansion, with up to 28 500 additional images each month.

However, as illustrated in Figure 3.15, municipalities vary significantly in the number of roads available in GSV. Assigning an equal number of additional samples to each municipality would not be an optimal strategy, as some municipalities have far more roads, and as a result a higher population, than others. To address this, additional samples are distributed proportionally to the number of roads in each municipality. This ensures that municipalities with larger road networks receive a greater number of samples, maintaining a balanced and representative dataset while avoiding over representation of areas with fewer roads.

To ensure a balanced distribution of additional samples across municipalities, a scaling factor is applied. This allows the number of added samples per municipality to be proportional to its road network size, while preventing under- or over-representation of specific regions.

The scaling factor for each municipality is calculated using equation 3.9.

$$k = \frac{\sqrt{R_m}}{\sum_{i=1}^{N} \sqrt{R_i}} \tag{3.9}$$

where $k$ is the scaling factor for municipality $m$, $R_m$ is the number of roads in municipality $m$, $N$ is the total number of municipalities, $\sum_{i=1}^{N} \sqrt{R_i}$ is the sum of the square roots of the number of roads across all municipalities. Figure 3.16 visually represents the scaling factor for the 213 local and metro municipalities.

Using this scaling factor, the number of additional samples allocated to each municipality when 28 500 samples are added is given by equation 3.10.

$$r = 28\ 500 \times k \tag{3.10}$$

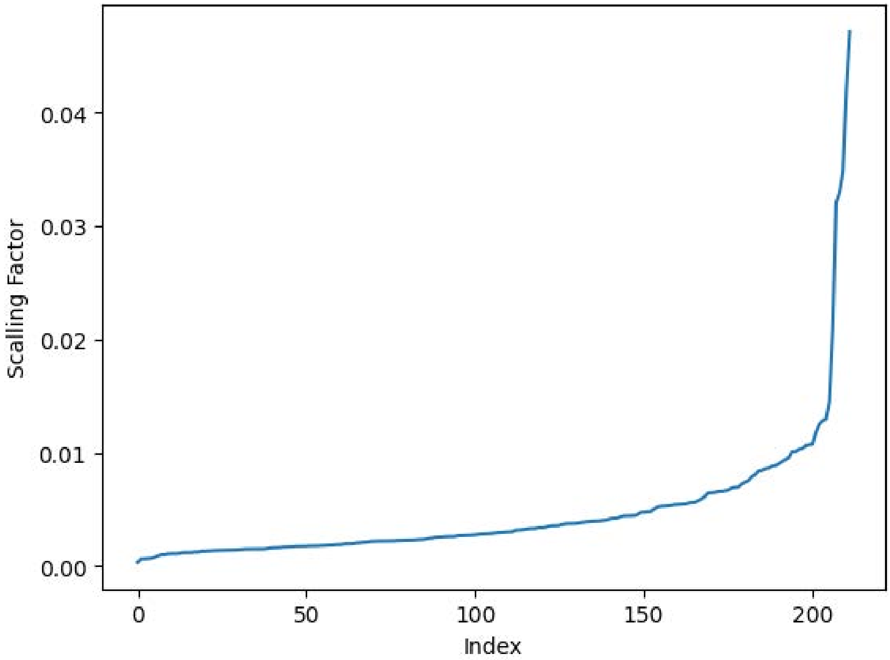


FIGURE 3.16: Line graph showing the scaling factor for each municipality in South Africa represented as a index from least roads to most roads

## 3.6 Additional Data Sources

In addition to GSV imagery, publicly available datasets were explored to enrich the modelling of voting patterns. This section details the external data sources investigated, the selection process, and the datasets ultimately incorporated into the study.

It is important to note that while correlating datasets with election dates is valuable, it is infeasible to determine the precise creation date of GSV imagery at scale, as Google does not provide consistent temporal metadata across the full image corpus. However, for the additional datasets incorporated into this study, data was collected for the year 2024, which coincides with the election results presented in this dissertation, thereby ensuring temporal alignment between external data and the electoral outcomes analysed.

Various socioeconomic and infrastructural datasets were considered, but only those with sufficient coverage and relevance at the municipal level were included. After evaluating multiple options including education and census data, the two key datasets used in this study are financial data and crime data as these were complete datasets that could be obtained at the municipal level in South Africa. The following subsections describe the rationale behind selecting these datasets, their sources, and how they were integrated into the analysis.

### 3.6.1 Crime

The South Africa Police Service (SAPS) publishes quarterly crime statistics [82], reporting the number of crimes recorded at each police station across the country. These crime reports provide a detailed breakdown of 40 crime categories, which are further grouped into seven major categories:

i. **Contact crimes (Crimes Against the Person):** Includes offences such as murder, assault, and attempted murder.

ii. **Sexual offences:** Encompasses crimes such as rape and sexual assault.

iii. **Aggravated robberies:** Captures violent crimes such as hijacking and armed robbery.

iv. **Contact-related crimes:** Includes offences such as arson and malicious property damage.

v. **Property-Related Crimes:** Captures burglaries, theft, and related offences.

vi. **Other serious crimes:** Includes commercial crime, fraud, and drug-related offences.

vii. **Crimes detected as a result of police action:** Consists of crimes like illegal possession of firearms, drug-related offences, and driving under the influence, which are primarily identified through police operations.

These crime statistics provide valuable insights into the safety and security conditions within municipalities, offering a potential link between crime levels and voter behaviour. The next section details how this data was processed and integrated into the study.

*Linking Crime To Area*

The primary challenge in integrating crime data into this study is that crime statistics are reported at the police station level, whereas electoral and socioeconomic data are structured at administrative levels such as wards and municipalities. To meaningfully incorporate crime data into the analysis, it must be aggregated to an appropriate administrative level.

Since there are too few police stations to generate crime statistics for each ward, the municipal level is chosen as the unit of aggregation. This ensures that crime data retains sufficient granularity while aligning with other datasets used in the study.

To assign police stations to municipalities, the ArcGIS [83] dataset of police station coordinates across South Africa is utilised. Each police station is then linked to the municipality in which it is located, allowing for the grouping and summation of crime statistics at the municipal level. Figure 3.17 shows the location of police stations across South Africa

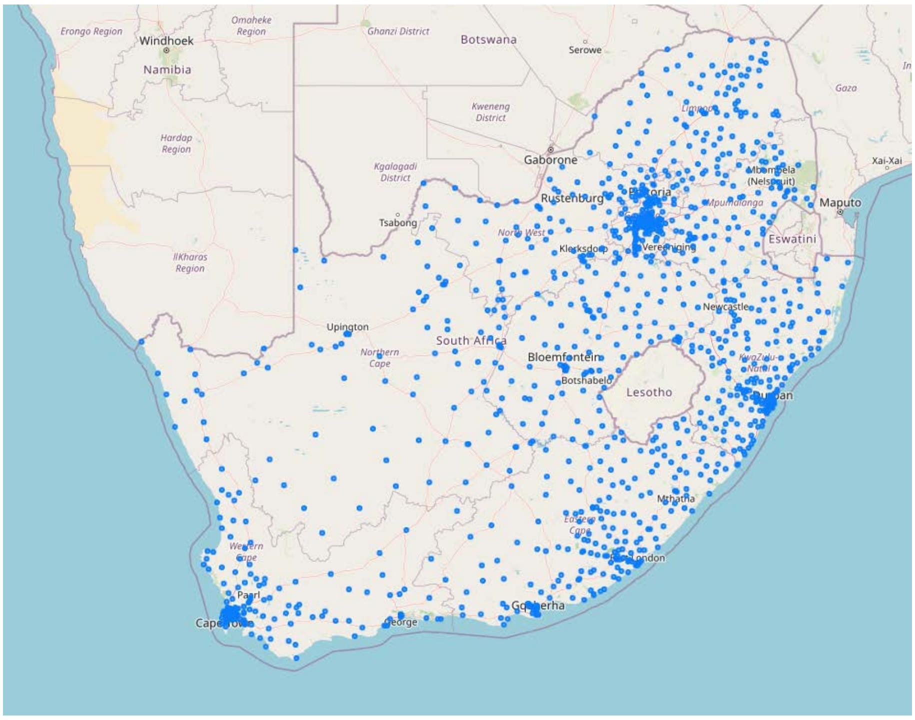


FIGURE 3.17: Location of police stations across South Africa as provided in the ArcGIS dataset [83], which can then be aggregated at the municipal level

### 3.6.2 Normalising Crime Data

To enable meaningful comparisons of crime levels across municipalities with varying population sizes, all crime features were standardised by calculating the number of crimes per capita. This adjusts for differences in population density and ensures that crime statistics reflect relative safety or risk, rather than absolute counts that could be biased by the size of the population.

For example, if a municipality records 1 000 instances of a particular crime and has a population of 100 000 people, the standardised value for that crime category would be 0.01. The general formula used for standardised is shown in Equation 3.11:

$$\text{crime ratio} = \frac{\text{number of crimes}}{\text{total population of municipality}} \tag{3.11}$$

This standardised approach allows for fair comparison of crime prevalence between municipalities, making it possible to investigate how relative crime exposure may influence voting behaviour.

### 3.6.3 Financial Data

To further enhance the understanding of voting patterns, municipal financial data from the South African National Treasury's Municipal Finance Data Portal [84] is incorporated into the study as a proxy, albeit limited, for government efficiency. This dataset provides a detailed breakdown of municipal financial health, including income, expenditure, debt, grants, and financial management indicators. By integrating this data, the analysis can explore whether economic conditions and governance performance influence voter behaviour.

However, many municipalities have incomplete financial records, with significant gaps in the available data. This limitation restricts the number of financial indicators that can be used in the analysis, as missing data could introduce bias or reduce the robustness of statistical models. However, the Unauthorised, Irregular, Fruitless, and Wasteful Expenditure (UIFW) dataset from the National Treasury's Municipal Finance Data Portal

[84] appears to provide complete coverage across all municipalities, making it a suitable financial metric for inclusion in this study.

The UIFW dataset tracks financial mismanagement within municipalities by reporting expenditures that do not comply with legal, procedural, or budgetary guidelines. It is divided into four categories:

i. **Unauthorised expenditure:** Spending that exceeds the approved municipal budget.

ii. **Irregular Expenditure:** Expenditure that does not comply with procurement processes or regulations.

iii. **Fruitless and wasteful expenditure:** Spending that could have been avoided if proper financial management had been in place.

iv. **Total expenditure:** The sum of all non-compliant expenditures.

## 3.7 Land Use Data

In addition to GSV imagery, crime, and municipal financial data, satellite imagery is incorporated into the study to further enhance the prediction of voting patterns. Satellite imagery provides insights into land use, infrastructure quality, and environmental conditions, which may possibly have some bearing on voter behaviour.

Three approaches were explored for obtaining satellite imagery:

i. **Google Cloud Console:** This platform offers high-resolution satellite imagery; however, given that the system already relies on Google Cloud credits for GSV imagery, using additional credits for satellite data would be costly.

ii. **Planet Labs:** This service allows researchers to apply for a research account [85], which provides 3 000 square kilometres of imagery per month. However, this is insufficient for South Africa's total land area of 1 221 037 square kilometres. Even with sampling, the available coverage would be severely limited, leading to an under-representation of vast regions.

iii. **South African National Land-Cover Dataset:** Provides a GIS Albers [86] dataset covering the entire country. This dataset, available from the Department of Environmental Affairs' Environmental Geographic Information System (EGIS) platform [87], offers a categorised land-cover map, detailing different land-use types such as urban, agricultural, and natural landscapes. Given its nationwide coverage and accessibility, this dataset was selected as the most viable option for integrating land-use data into the analysis.

### 3.7.1 Converting From Raster to Vector Information

The South African National Land-Cover Dataset is provided in raster format as a Tag Image File Format. However, to integrate it with the existing GeoPandas DataFrame, it needs to be converted into a vector format. Raster data consists of a grid of pixels, each representing a land-cover classification, whereas GeoPandas vector data consists of polygons, lines, or locations [75].

To achieve this conversion, QGIS [88] was used along with the "Raster to Vector" function. This process transformed the pixel-based land-cover classifications into polygon features, allowing for integration with the GeoPandas dataset. Once in vector format, the dataset can be processed, analysed, and linked to the relevant administrative boundaries within the study. Figure 3.18 Shows the raster file of the South African land use data.

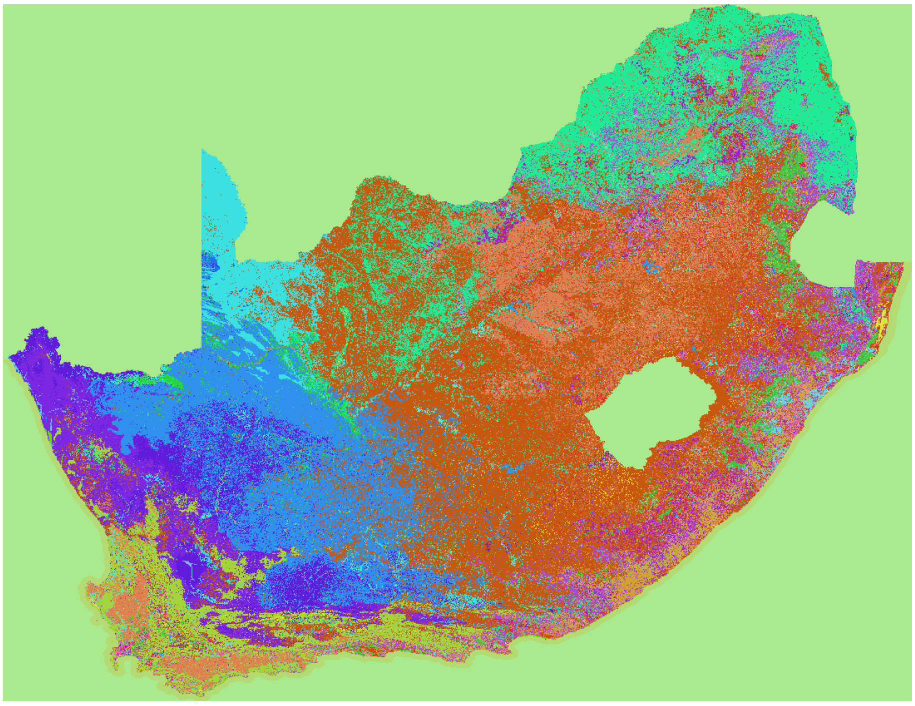

FIGURE 3.18: South African land use raster file obtained from the Department of Environmental Affairs [87]

### 3.7.2 Normalising Land Use Features

To ensure comparability across administrative units of varying sizes, all land-use features were normalised by calculating their proportion relative to the total area of the respective ward or municipality. This approach accounts for differences in spatial extent and allows for meaningful interpretation of land-use intensity within each area.

For instance, if an administrative unit covers 1 000 square kilometres, and 100 square kilometres of that area are classified as villages, the normalised value for the village category would be 0.1. The general formula used for normalisation is shown in Equation 3.12.

$$\text{land use ratio} = \frac{\text{area of land use}}{\text{total area of administrative unit}} \tag{3.12}$$

This standardisation ensures that the land-use data accurately reflects the proportion of each land-cover type, enabling fair comparisons between different wards or municipalities regardless of their absolute size.

## 3.8 Computer Vision

To extract meaningful features from the GSV imagery, this study employs the You Only Look Once, Version 8 (YOLOv8) [68] object detection model. Unlike traditional object detection models that require multiple passes over an image, YOLOv8 processes the entire image in a single forward pass, making it significantly faster.

YOLOv8 was chosen for this study because of its performance in detecting multiple objects in complex urban environments [89]. Its ability to generalise well across different lighting conditions, camera angles, and environmental variations ensures consistent and reliable feature extraction from the diverse range of images obtained through GSV [90]. This makes it an ideal choice for identifying key visual indicators, such as potholes, illegal dumping, vehicles, and urban infrastructure, which may serve as proxies for socioeconomic conditions influencing voter behaviour.

To extract relevant features from GSV imagery, five distinct YOLOv8 models were trained to detect illegal dumping, graffiti, vehicle types, potholes, and pedestrians. These models were developed using a two-stage training process, detailed below, to maximise accuracy and ensure that they generalise well to the diverse range of images collected.

Initially, a model was trained using publicly available datasets to establish a baseline for detecting the target objects. This preliminary model was then used to perform inference on the collected GSV imagery, identifying a subset of images likely to contain the object of interest, for example an illegal dumping site. This subset was manually annotated to create a refined dataset composed exclusively of GSV images, ensuring a more representative and contextually relevant dataset.

The refined dataset was then used to retrain the model, either by training from scratch or fine-tuning the weights of the initial model, depending on the initial model's performance on GSV imagery. This iterative process ensured that the final models were optimised for detecting objects in the collected GSV imagery.

Each of the following sections provides a detailed breakdown of the methodology used for training and evaluating each computer vision model, including dataset selection, annotation processes, and performance metrics.

*Note On Number Of Epochs Used In Computer Vision Training*

The number of training epochs varied across the different YOLOv8 models, depending on factors such as dataset size, the use of transfer learning, validation accuracy, and empirical observations from trial and error. For preliminary training using publicly available datasets, models were trained with a lower number of epochs (e.g., 50), ensuring a reasonable balance between computational efficiency and model convergence.

For full model training on the refined GSV datasets, the number of epochs was adjusted based on the complexity of the dataset and model performance during validation. In cases where transfer learning was applied, fewer epochs were typically required, as the model had already learned generalisable features from prior training. Conversely, when training a model from scratch or on variable datasets, longer training durations (e.g., 100 epochs) were necessary to achieve optimal performance.

The selection of epoch counts was guided by monitoring validation loss and accuracy scores, ensuring that models were trained sufficiently without overfitting. This adaptive approach allowed each model to reach an optimal state while accounting for the specific characteristics of the object detection task.

### 3.8.1 Potholes

For the preliminary model, an initial YOLOv8 model was trained using a dataset of 3 393 annotated pothole images from Roboflow [91] over 50 epochs. This model was then used to perform inference on the entire GSV dataset, identifying a subset of images likely to contain potholes for further manual annotation and refinement.

From the subset of images identified by the preliminary model, 4 000 images were randomly selected to ensure a geographically diverse representation across South Africa, preventing bias toward any specific region. This dataset was then manually annotated using CVAT

[92]. To further refine the dataset, 2 000 random images were removed that were guaranteed to contain potholes, ensuring a balanced and representative distribution for model training.

The refined dataset was then split into an 80/20 training and validation set, and a new YOLOv8 model was trained over 100 epochs as the initial model did not generalise well on GSV Imagery.

The training results indicate that the model performed well, with decreasing loss values and increasing precision, recall, and mean average precision (mAP). These trends suggest that the model learned to classify and localise potholes over time, demonstrating typical signs of improvement during training.

However, while the performance metrics indicate strong results on the training and validation sets, testing the model on unseen data revealed limitations. Upon visual inspection of inference results, a high number of false positives were observed, particularly with manhole covers, which were frequently misclassified as potholes. This issue was exacerbated by the fact that manhole covers are not evenly distributed across South Africa; they are predominantly found in metropolitan areas, particularly in Cape Town. As a result, the dataset was assumed to be likely skewed, leading the model to overfit to these urban features and increasing the likelihood of misclassification. Figure 3.8.9 presents examples of manhole covers that were mistakenly identified as potholes by the model, highlighting a common source of false positives.

To counteract this issue, 200 background images were added to the dataset, following best practices outlined in [93]. Of these, 100 images contained manhole covers, while the remaining 100 images were images where the model had previously detected false positives for potholes, despite no actual potholes being present, the model was then retrained using 100 epochs and this updated dataset. This augmentation helped the model differentiate between potholes and other visually similar features, such as manhole covers. Upon visual inspection of the updated inference results, the inclusion of background images significantly reduced false positives caused by manhole covers and other misclassified objects.

Figure 3.19: Manhole covers as a cause for false positives in pothole detection

The performance of the refined model shows significant improvement over the initial version, demonstrating enhanced learning and generalisation. The training and validation losses exhibit a consistent downward trend, indicating that the model is minimising error over successive epochs. The precision and recall metrics show steady increases, suggesting that the model is becoming more adept at correctly identifying potholes while reducing false negatives. The mAP curves, particularly at higher confidence thresholds, indicate improved localisation accuracy and overall detection performance. The learning curves suggest only mild overfitting, as validation losses flatten while training losses continue to decrease, but overall the model maintains good generalisation with stable precision, recall, and mAP scores. The detailed performance metrics and associated graphs are provided in Appendix 3.8.1.

### 3.8.2 People

To develop a pedestrian detection model, an initial dataset containing 1 700 annotated images of pedestrians was obtained from Roboflow. This dataset was used to train a preliminary YOLOv8 model over and additional 50 epochs, establishing a baseline model for pedestrian detection. Following this, inference was performed on the collected GSV imagery to identify a subset of images likely to contain pedestrians.

Once this subset was identified, a new dataset was created exclusively from GSV images, following the same methodology used for pothole detection. However, unlike the pothole model, manual inspection revealed that the initial pedestrian detection model already demonstrated strong performance on GSV imagery. Given this, rather than training a new model from scratch using only the refined dataset, the pretrained weights from the initial model were utilised and fine-tuned using the GSV-specific dataset over 50 epochs. This approach ensured that the model retained its generalisation capabilities while being optimised for pedestrian detection in GSV Imagery.

The pedestrian detection model demonstrated strong performance across key evaluation metrics. The training loss steadily decreased over the 50 epochs, indicating effective learning and convergence. Precision and recall values improved over time, with precision reaching high values, suggesting that the model confidently identified pedestrians while minimising false positives. The precision-recall and F1-confidence curves further reinforce the model's reliability, with a balanced trade-off between identifying pedestrians correctly and reducing false detections. The confusion matrix illustrates that the model correctly classified a high proportion of pedestrian instances, with relatively few misclassifications. The training and validation learning curves further indicate that the model is not underfitting, as both losses declined consistently, and only mild signs of overfitting are visible where validation losses plateau while training losses continue to decrease. Importantly, the precision, recall, and mAP scores remain stable, confirming that the model maintains good generalisation performance. Overall, these results suggest that the model generalises well to new data, making it suitable for large-scale inference on GSV imagery. Detailed performance metrics and additional evaluation graphs are provided in Appendix 3.8.2.

### 3.8.3 Vehicles

The vehicle detection model was developed using a dataset containing 7 500 annotated images of various vehicle types, including cars, motorbikes, buses, and trucks, obtained from Roboflow. The model was trained over 100 epochs, leveraging transfer learning to enhance detection accuracy and generalisation.

Unlike the previous models that required fine-tuning with GSV-specific imagery, this model demonstrated strong performance when tested on GSV images without requiring additional retraining. The training and validation learning curves show consistent downward trends across all loss functions, with no divergence between training and validation losses, indicating that the model is not overfitting and has achieved effective convergence. Precision, recall, and mAP metrics continue to improve steadily throughout training, suggesting robust learning and strong generalisation across different vehicle classes. The model identified vehicles across diverse urban and rural settings, achieving high precision and recall as shown in Appendix 3.8.3.

### 3.8.4 Graffiti

For the graffiti detection model, a dataset containing 3 500 images was obtained from Roboflow and trained over 100 epochs. This initial model was then used to identify GSV images that potentially contained graffiti, creating a subset of images for further annotation and training.

However, upon evaluating the subset, it was found that the model detected very few instances of graffiti in the GSV dataset. Moreover, the detected instances were concentrated almost exclusively in metropolitan areas, limiting the model's ability to provide meaningful insights across diverse geographic regions. Given the insufficient representation of graffiti in the dataset and its strong urban bias, the decision was made not to proceed with this model, as it would not contribute significantly to the overall analysis.

### 3.8.5 Illegal Dumping

Initially, a model was trained using 100 epochs to detect individual pieces of litter. When applied to GSV imagery, the model successfully identified instances of litter across South Africa. However, a significant issue arose: in cases where piles of litter were present, the model would detect each individual piece, leading to an over representation of litter in certain images, as a single image with a large trash pile would contain disproportionately more detections compared to images with scattered litter.

To address this, a new model was developed using a dataset obtained from Roboflow specifically designed to detect trash in a more structured way. Example images from this dataset are shown in Figure 3.20. Once this refined model was used to create a dataset exclusively from GSV imagery, three distinct types of litter were consistently identified, each reflecting different local conditions. The first category consisted of garbage bins left on the roadside for municipal refuse collection. The second included neatly placed trash bags, also indicating scheduled waste collection. The third, and most ostensibly significant, category was illegal dumping sites, which served as a key indicator of service delivery failures and weak enforcement of waste management regulations. Given their strong correlation with inadequate municipal services [94], illegal dumping sites were identified as the most relevant target for further analysis.

FIGURE 3.20: Example images from trash object detection set

To specifically detect illegal dumping, the initial model was used to generate a dataset of 2 000 images, containing illegal dumping sites, sourced exclusively from GSV imagery. This ensured the dataset focused solely on illegal dumping, excluding trash meant for municipal collection.

A new model was then trained on this dataset using 75 epochs. Training a separate model was necessary to prevent misclassification, as reusing weights from the previous model could lead to garbage bins and neatly placed trash bags being incorrectly identified as illegal dumping. By training from scratch, the model was optimised to reliably identify illegal dumping sites.

The performance metrics of the final illegal dumping detection model indicate moderate success in identifying illegal dumping sites. The precision-confidence and precision-recall curves suggest that the model distinguishes between illegal dumping sites and background elements, achieving relatively high precision at lower recall levels. The recall-confidence curve highlights a steady decrease in recall with increasing confidence thresholds, indicating that while the model can detect illegal dumping, it may struggle to generalise across diverse environments. The training and validation curves support these findings, as both sets of losses decrease consistently across epochs, showing that the model is learning without significant underfitting or overfitting. Precision, recall, and mAP values improve steadily, although they plateau at moderate levels, reflecting the inherent difficulty of generalising across the highly variable appearances of illegal dumping.

Compared to previous models, the illegal dumping detector exhibited lower performance, particularly its recall and F1-score. This is expected, due to the labour-intensive nature of creating a sufficiently large annotated dataset for illegal dumping and the high variability in its appearance. Unlike objects such as cars, which have consistent visual characteristics, illegal dumping sites vary significantly in composition, scale, and context, making it more challenging for the model to generalise effectively. Despite these challenges, the model provides a valuable tool for detecting illegal dumping, and future improvements could focus on expanding the annotated dataset and incorporating additional contextual features to enhance robustness. The final results are presented in Appendix 3.8.5.

### 3.8.6 Inference

The inference process was designed to ensure that detected objects in GSV images could be linked back to their respective geographic locations. Given the scale of the dataset and the fact that images alone do not inherently contain spatial information, a structured approach was required to maintain the association between images and their corresponding wards or municipalities.

To address this, each image file name was encoded with the ward and municipality ID before inference was performed. This approach eliminated the need for extensive folder structures, which would have been infeasible given the 4 468 wards across South Africa. Instead, the image filename itself contained the necessary geographic identifiers, allowing for efficient data retrieval and analysis after inference.

By embedding geographic metadata directly into the image filenames, the results of object detection can be linked back to their respective locations. This method ensured that once inference was completed, detected features could be aggregated at various administrative levels, facilitating spatial analysis without requiring a separate database or manual re-matching of images to locations.

## 3.9 ML and Statistical Regression

This section outlines the ML and statistical modelling approaches employed in this study to analyse voting patterns. As detailed in Section 3.2, voting patterns are represented as percentage shares for the ANC, DA, EFF, MK, and non-voting populations at both the ward and municipal administrative levels.

Table 3.1 presents the number of divisions at each level, with 213 municipalities and 4 468 wards. From an ML perspective, this means that at the ward level, there are sufficient data locations to develop robust predictive models. However, at the municipal level, the dataset is considerably smaller, making ML techniques less viable due to the risk of overfitting and insufficient generalisation.

A key reason for employing statistical regression at the municipal level is the imbalance between the number of observations and the number of features. A common rule of thumb in statistical modelling suggests that for reliable ML models, the number of observations should be at least ten times greater than the number of features [95]. Given the limited number of municipalities, this condition is not met, making ML approaches prone to overfitting. This challenge was further confirmed through preliminary ML modelling at the municipal level, where cross-validation results showed substantial disparity between train and test set performance. Specifically, the R-2 value on the training set was significantly higher than on the test set, indicating poor generalisation.

To address these limitations, ML models were applied at the ward level, where a larger dataset allows for capturing complex relationships and patterns, while statistical regression techniques were used at the municipal level to provide interpretable and generalisable insights.

*Note On Availability In GSV Imagery*

Before proceeding to the ML models, it is important to acknowledge the availability of data at the ward level. While both GSV and land-use data can be collected at this resolution, certain wards lack sufficient GSV coverage due to the absence of mapped roads in GSV. As a result, the total number of wards with available data is reduced from 4 468 to 3 636. The distribution of the remaining wards included in the analysis is presented in Figure 3.21, which shows a reduction of wards in the Eastern Cape and Northern Cape Provinces.


FIGURE 3.21: Image showing the available wards to be used in ML models as a result of availability in GSV

*Feature Engineering*

This section outlines the feature engineering process applied to the GSV-derived features. While the raw counts of these features provide valuable insights, they must be feature engineered to ensure they are suitable for ML and statistical regression models.

Given the sample size calculations, it must be accounted for that GSV images collected per ward unit is not uniform. This variation can introduce bias, as areas with more images are more likely to have higher raw counts of detected features. To address this, the first step in the feature engineering process is computing the feature ratio, shown in equation 3.13, which standardises the raw counts of features by the total number of images collected in each area.

$$\text{feature ratio} = \frac{\text{Instances of Feature}}{\text{Total Images Collected in Area}} \tag{3.13}$$

where the feature can be potholes, illegal dumping, or vehicles, ect.

After standardising feature counts, additional transformations are applied to capture spatial dependencies and variations across different areas. The feature lag is computed as the average feature ratio of the three nearest neighboring areas feature ratio, providing a measure of regional trends. This helps identify broader spatial patterns by reflecting how an area compares to its surroundings. The feature lag is calculated using 3.14:

$$\text{feature lag} = \frac{\sum_{i=1}^{3} \text{feature ratio}_i}{3} \tag{3.14}$$

Finally, the feature lag difference quantifies the localised deviation of an area from its surroundings. This is particularly salient for identifying anomalies, such as areas with disproportionately high or low instances of a feature compared to it's neighbouring regions. The feature lag difference is calculated using 3.15:

$$\text{feature lag difference} = \text{feature lag} - \text{feature ratio} \tag{3.15}$$

By first computing the feature ratio to control for differences in image density and then applying spatial transformations through feature lag and feature lag difference, this approach ensures that the dataset captures both localised and regional variations, improving the model's ability to identify spatial patterns.

To ensure that each area has three nearest neighbours, a k-nearest neighbours (KNN) approach is employed. Rather than relying solely on geographical adjacency, which may result in an insufficient number of neighbours, KNN ensures that the closest available areas are consistently selected for comparison.

A BallTree [96] data structure is used to perform spatial queries and find the three nearest areas for each location. The process is as follows:

i. **Centroid calculation:** The centroid of each area's geometry is computed to serve as its representative location.

ii. **BallTree construction:** A BallTree is built using the centroids, allowing for fast spatial lookups.

iii. **Nearest neighbors query:** For each area, the three nearest Neighbors are identified using the BallTree.

This approach ensures that the feature lag and feature lag difference captures spatial patterns while maintaining computational efficiency, even for large datasets.

In addition to generating normalised and spatially relative features, an additional indicator, the people to car ratio, was created to provide insight into local mobility patterns. As shown in Equation 3.16, this ratio is calculated by dividing the people ratio by the car ratio. It serves as a proxy for personal mobility within an area, where higher values may suggest limited access to private transport and lower levels of affluence.

$$\text{people to car ratio} = \frac{\text{people ratio}}{\text{car ratio}} \tag{3.16}$$

Finally, a road ratio feature is computed to capture digital mapping coverage across each area using 3.17. This variable is defined as the number of roads available on GSV divided by the total number of roads in the area. The resulting ratio provides a normalised measure of GSV coverage, with higher values indicating greater visibility or representation on the platform. As with other features, this normalisation helps account for underlying differences in infrastructure and digital inclusion, which may influence both image availability and electoral patterns.

$$\text{road ratio} = \frac{\text{roads available on GSV}}{\text{all available roads}} \tag{3.17}$$

### 3.9.1 Ward Administrative Level (ML Models)

At the ward level, several datasets are available for modelling voting patterns. Features derived from GSV imagery, such as the prevalence of potholes, illegal dumping, vehicle types, and pedestrian presence, provide granular insights into the built environment and socioeconomic conditions. Additionally, land-use features extracted from the South African National Land-Cover dataset, further enhance the analysis by capturing the spatial distribution of residential, commercial, and undeveloped land within each ward.

While these high-resolution features are available, data from additional sources such as crime statistics and municipal financial indicators are not reported at the ward level. These datasets are only aggregated at the municipal level, limiting their direct use in ward-level ML models. As a result, the analysis at this resolution primarily relies on visual and spatial data sources.

*ML Pipelines (Ward Administrative Level)*

Many of the collected features exhibited significantly skewed distributions, which could negatively impact model performance by violating the assumptions of normality in certain ML algorithms. To address this, a transformation step was incorporated into the pipeline to standardise feature distributions.

Several transformation techniques, such as logarithmic transformations, were considered. However, the Yeo-Johnson transformation was ultimately selected due to its ability to handle both positive and negative values while optimally normalising data distributions.

Given the large number of features derived from GSV imagery and land-use data, it was impractical to manually inspect and transform each feature. Instead, a dynamic approach was implemented to identify and transform only those features exhibiting substantial skewness. The process of dynamically transforming features is described below.

i. **Identification of numeric features:** The dataset was first filtered to retain only numerical features, as transformations are only applicable to continuous variables.

ii. **Calculation of skewness:** The skewness of each numeric feature was computed to determine the degree of asymmetry in its distribution.

iii. **Feature selection for transformation:** A threshold-based approach was applied, where features with an absolute skewness value greater than 0.5 were flagged for transformation. This threshold was chosen based on empirical studies that suggest a skewness value exceeding 0.5 may warrant transformation for improved model performance [97] and through trail and error.

Once the necessary transformations were applied to correct skewness, the numerical features were standardised using Standard Scaler [98]. Standardisation ensures that all features have a mean of zero and a standard deviation of one, preventing models from being biased toward features with larger numerical ranges.

To ensure compatibility with ML models, the province variable, which is categorical, was one-hot encoded. Since most ML algorithms require numerical inputs, categorical data must be transformed in a way that preserves its informational value without introducing unintended ordinal relationships. One-hot encoding achieves this by creating separate binary variables for each province, allowing the model to interpret the data without assuming any inherent ranking or hierarchy.

A range of ML algorithms were evaluated to determine the most effective approach for modelling voting patterns. These included support vector machines, decision trees, random forest, and linear regression. Each algorithm was tested on the dataset to assess its predictive performance and suitability for the structured data available.

After comparative analysis, linear regression [99] and random forest [100] emerged as the most effective models. Linear regression was well-suited for capturing linear relationships within the data, while random forest provided robustness in handling non-linear patterns and feature interactions [101, 102]. These models demonstrated the best balance between interpretability and predictive accuracy, making them the optimal choices for the final analysis.

The above work not withstanding, an analysis of the features derived from satellite imagery revealed that they did not contribute meaningful predictive value to the ML models. However, instead of enhancing model performance, these features appeared to introduce additional noise, reducing both accuracy and interpretability, while adding unnecessary model complexity. To optimise performance, streamline the feature set, and enable better interpretability these satellite-derived variables were ultimately removed from the ML models.

For the linear regression model, feature reduction was necessary to improve model performance and prevent over fitting. Given the number of features derived from GSV imagery, selecting the most relevant predictors was essential. To achieve this, Recursive Feature Elimination with Cross-Validation (RFECV) was used, optimising feature selection based on the R-2 score. RFECV systematically eliminates the least important features while evaluating model performance through cross-validation, ensuring that only the most predictive variables are retained.

In contrast, feature reduction was not required for the random forest model. As a tree-based ensemble method, random forest inherently handles feature selection by assigning importance scores to each variable. The algorithm is well-suited to high-dimensional data, reducing the risk of overfitting without the need for explicit feature elimination. To optimise model performance, hyperparameter tuning was conducted using 10-fold cross-validation, ensuring the best combination of parameters was selected. This flexibility, along with its ability to model non-linear relationships, enables random forest to capture complex interactions between features and voting patterns more effectively than linear regression in certain cases.

Model performance was evaluated using k-fold cross-validation, a robust technique that measures accuracy across multiple data splits to minimise bias and variance. This approach ensured a more reliable estimate of the models' generalisability. The evaluation was conducted using the following metrics: R-2 score, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). These metrics provided a comprehensive assessment of model accuracy and reliability.

To complement these metrics, visual tools were also used to assess performance. Predicted versus actual vote percentage plots provided an intuitive sense of model fit, while residual plots helped identify systematic prediction errors or biases — offering insights into where models may be improved or where patterns may require further investigation.

To explore relationships between features and voting outcomes, Spearman and Pearson correlation tables were created. These tables enabled the assessment of both linear and monotonic associations, offering a preliminary understanding of how individual variables relate to electoral results.

Building on this, feature importances were extracted from the final random forest models to determine which variables most strongly influenced voting predictions. Unlike correlation analysis, which captures pairwise relationships, random forest importances reflect complex, multi-dimensional interactions between features. By visualising the highest-ranked variables, the analysis offers interpretable insights into the key drivers of voting behaviour and enables comparison with established theoretical frameworks.

Together, these approaches allow for a richer understanding of the data — revealing both linear trends and the more complex, non-linear associations that shape model performance.

*Classification*

While the previous section focused on predicting vote percentages using regression models, this section explores an alternative approach by framing the problem as a classification task. Instead of estimating the exact percentage of votes received by each party, the objective here is to determine which political party, ANC, EFF, DA, or MK, secured the majority vote in each ward. This classification-based approach follows the methodology established in foundational work by Gebru et al., which used GSV to predict dominant political outcomes across geographic regions in the United States [24].

Both logistic regression and random forest classifiers are employed for this task. Logistic regression is selected for its interpretability and ability to model class probabilities, making it useful for understanding the influence of different features on party dominance. However, since logistic regression assumes linear relationships between the input features and the target variable, a random forest classifier is also used. This model, which leverages an ensemble of decision trees, is capable of capturing complex, non-linear interactions between features, potentially improving predictive performance.

Both models are trained using the engineered GSV derived features, ensuring consistency with the regression analysis. Their performance is evaluated through cross-validation, with precision, recall and F1-score performance.

### 3.9.2 Municipal Administrative Level (Statistical Models)

At the municipal level, statistical regression models are employed instead of ML models due to the relatively small number of data locations (213 municipalities). Unlike the ward-level analysis, which relied solely on features derived from GSV imagery, the municipal-level models incorporate additional datasets. These include crime data, land-use data, and financial data, providing a broader perspective on the factors influencing voting patterns. By integrating these diverse data sources, the analysis aims to capture the complex socioeconomic and infrastructural conditions at the municipal level, allowing for a more comprehensive understanding of voter behaviour.

*Statistical Model Pipeline (Municipal Administrative Level)*

The statistical modelling workflow begins by standardising all numerical features using Z-score normalisation to ensure that variables with different scales do not disproportionately influence the model. Following this, variance inflation factor (VIF) analysis is conducted to identify and remove highly collinear features [103]. Instead of simply eliminating features with high VIF values, an responsive selection process is applied, where the feature most correlated with the target variable is retained while redundant features are discarded. This helps preserve valuable information while mitigating multicollinearity.

After filtering features based on VIF, three different regression techniques are applied to further refine feature selection and improve model performance: Lasso Regression, Ridge Regression, and Elastic Net Regression. Lasso Regression [48] (L1 regularisation) is used to perform automatic feature selection by shrinking the coefficients of less important variables to zero, effectively eliminating them from the model. Ridge Regression [49] (L2 regularisation) penalises large coefficients, reducing the risk of overfitting while retaining all features. Elastic Net Regression [50] combines both L1 and L2 penalties, striking a balance between feature selection and regularisation, which is particularly useful when dealing with highly correlated variables. The outputs of these models are then generated to asses the perfromance of the various models.

Once the regression models have been executed, an analysis is performed to determine the most influential features contributing to voting patterns at the municipal level. To achieve this, the top 15 most important features are extracted from each regression method — Lasso, Ridge, and Elastic Net. These selected features represent the variables with the strongest relationships to the target outcomes and provide valuable insights into the factors influencing voting behaviour.

Importance plots are then generated from the 15 most important features acting as a explainability tool, illustrating how each feature influences voting patterns at the municipal level. By visualising feature importance, these plots provide insights into the relative impact of different predictors, helping to interpret the relationships uncovered by the regression models. Each regression method produces its own independent importance plot, allowing for a comparative analysis of how different regularisation techniques prioritise features. This approach ensures that the selection of significant predictors is not reliant on a single method but rather on a consensus among the models.

The importance plots also highlight the direction of the relationships between features and voting outcomes. Features with positive coefficients indicate a direct association, where an increase in that feature corresponds to an increase in the predicted voting percentage for a particular party. Conversely, negative coefficients suggest an inverse relationship, where higher values of a feature correspond to lower voting percentages. The standardisation of all numerical features prior to modelling ensures that these comparisons remain meaningful, as all coefficients are measured on the same scale.

# Chapter 4

# Results and Model Performance

This study aimed to model voting patterns at both the ward and municipal levels using a combination of Google StreetView (GSV) imagery, satellite imagery, and additional datasets. The objective is to predict the vote percentage of the four dominant political parties, applying machine learning (ML) techniques at the ward level and statistical regression methods at the municipal level. By integrating multiple data sources, the study identified spatial and socioeconomic factors that influence voting behaviour.

This chapter presents the results obtained from each stage of the analysis. It first evaluates the effectiveness of the coverage and quality of the gathered imagery. Next, it presents the performance of the object detection models applied to GSV imagery, outlining their ability to extract relevant urban and environmental features.

The chapter then shifts to reporting the predictive performance of the ML and statistical regression models in modelling voting patterns. It presents the results of the ward-level ML models, followed the municipal-level statistical models.

i. **Image collection and computer vision:** Results of the image collection process and analysis of the inferred features in those images using computer vision (see 4.1)).

ii. **Modelling voting patterns at the ward level:** Performance of the ML models (see 4.2).

iii. **Modelling voting patterns at the municipal level:** Performance of the regression models (see 4.3).

## 4.1 Image Collection and Computer Vision

A total of 157 338 images were collected, exceeding the minimum required sample size of 82 390 by 74 948 images. These images were distributed across municipalities following the sampling methodology outlined in Section 3.5, ensuring proportional representation based on road networks and population proxies.

### 4.1.1 Potholes

A total of 34 536 potholes were detected across the 157 338 collected images. Figure 4.1 shows the distribution of confidence scores assigned to pothole detections by the pothole object detection model. The majority of detections fall within mid-range confidence intervals (0.3–0.8), with fewer detections assigned very high confidence scores (above 0.9). This distribution suggests variability in the model's confidence, which may be influenced by factors such as the distance of the pothole from the camera, as illustrated in the top half of Figure 4.2. Additionally, lower confidence scores may occur when potholes are smaller in size or deviate from the typical appearance of potholes in the training dataset, as shown in the bottom half of Figure 4.2.

Upon visual inspection, the significant majority of classified potholes were indeed actual potholes, indicating that the model performs reliably despite variations in confidence scores.

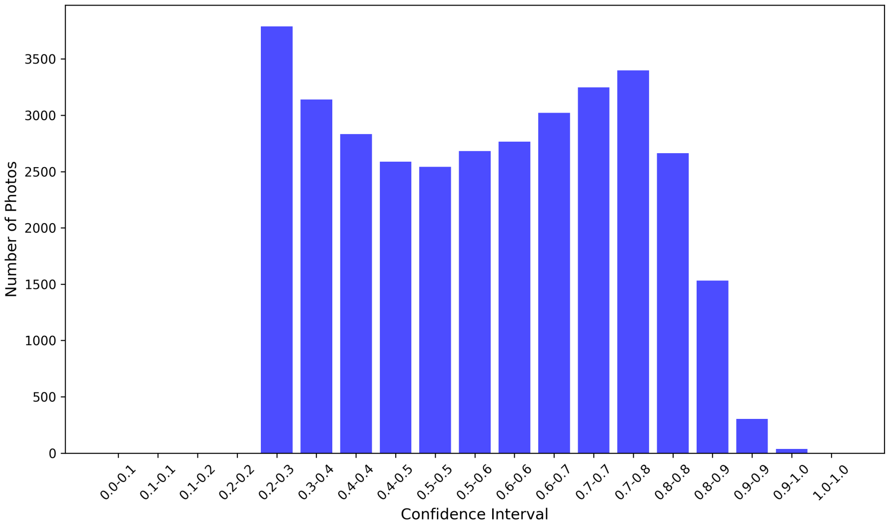


FIGURE 4.1: Distribution of pothole detection confidence scores across collected GSV images in South Africa


FIGURE 4.2: Images showing variability in confidence scores due to pothole non-conformity and distance from the camera

Figure 4.3 presents chloropleth plots depicting the distribution of pothole ratio density across three metropolitan municipalities in South Africa at the ward administrative level: Ekurhuleni (Gauteng), the City of Cape Town (Western Cape), and eThekwini (KwaZulu-Natal). Each plot illustrates spatial variations in pothole density, with darker regions indicating lower pothole ratios and lighter regions representing higher densities of pothole ratios.

While a nationwide visualisation of pothole density is possible, much of the detail would be lost due to scale. Instead, these municipal-level plots provide a clearer visualisation of localised variations in road conditions. The observed differences in pothole density across these municipalities may reflect variations in infrastructure maintenance, road usage patterns, and environmental factors.

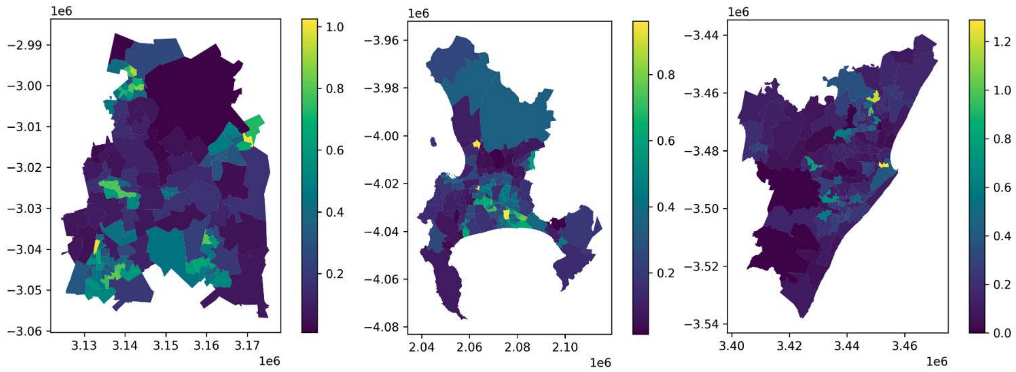


FIGURE 4.3: Pothole ratio density across three major municipalities in South Africa at the ward level, City of Tshwane (left), City of Cape Town (middle), ethekwini Metropolitan Municipality (right)

### 4.1.2 People

A total of 25 456 people were detected across the 157 338 collected images. Figure 4.4 presents the distribution of confidence scores, with most detections falling within the lower confidence ranges and few exceeding 0.7. This reflects the inherent variability in detecting people within GSV imagery due to factors like occlusions, varying poses, and image perspectives.

Despite lower confidence scores, visual inspection confirms that most detections were correct. As shown in Figure 4.5, even low-confidence detections correspond to actual people rather than misclassifications, suggesting the model maintains high precision.

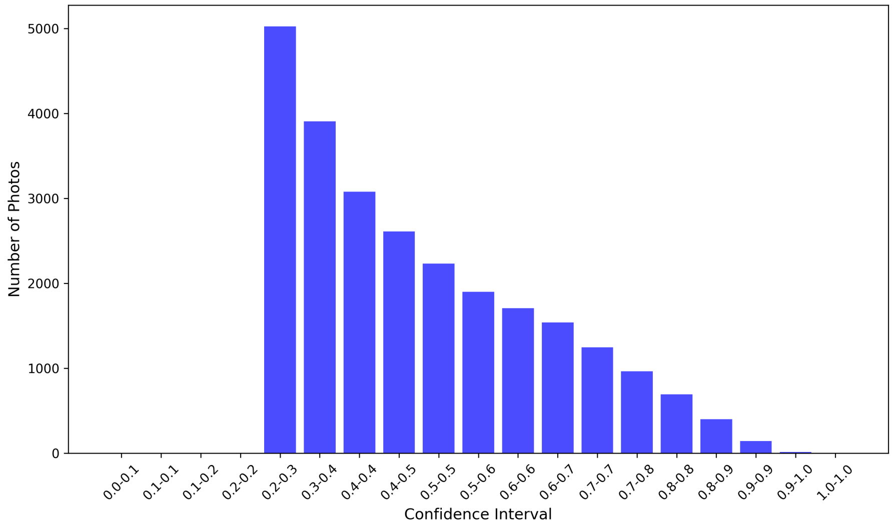


FIGURE 4.4: Distribution of people detection confidence scores across collected GSV images in South Africa


FIGURE 4.5: Images showing variability in confidence scores due to people non-conformity and distance from the camera

A similar to Figure 4.3 can be generated for the distribution of people ratio, highlighting variations in pedestrian presence across different municipalities. However, the areas with high people ratios differ from those with high pothole ratios, reflecting differences in urban density, commercial activity, and pedestrian infrastructure. This choropleth is shown in Appendix 4.1.2.

### 4.1.3 Vehicles

A total of 108 513 vehicles were detected across the 157 338 collected images. For this analysis, vehicles, including cars, motorbikes, buses, and trucks, have been grouped into a single category. Figure 4.6 presents the distribution of confidence scores assigned to these detections. Unlike previous categories, vehicle detections are skewed towards higher confidence scores, with a notable increase in detections above 0.8.

This distribution suggests that the model exhibits greater certainty in detecting vehicles compared to the other models, likely due to their relatively uniform structure and distinct visual characteristics in GSV imagery. However, variations in detection confidence may still arise due to factors such as occlusion, vehicle size, or distance from the camera. Despite these variations, visual inspection confirms that the majority of detected objects were indeed vehicles, with minimal instances of misclassification. Figure 4.7 provides examples of detected vehicles across different confidence levels.

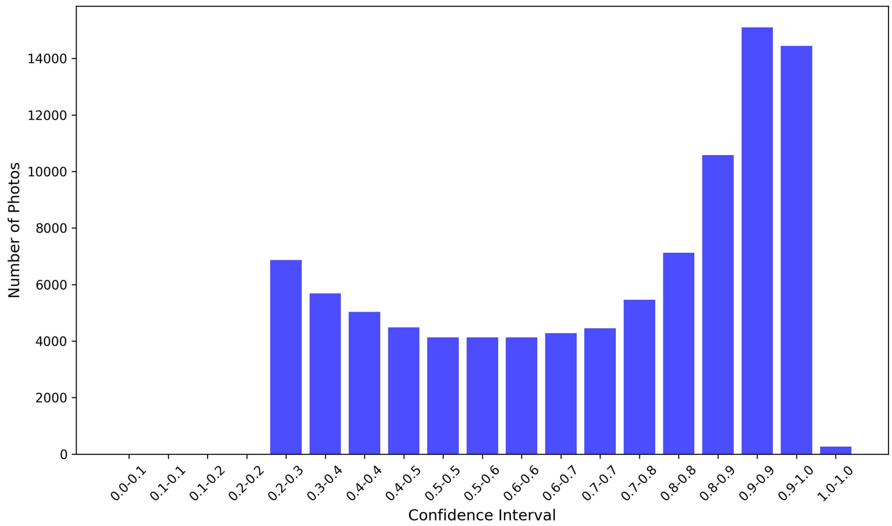


FIGURE 4.6: Distribution of vehicle detection confidence scores across collected GSV images in South Africa


FIGURE 4.7: Images showing high vehicle detection confidence regardless of image variation

Again a similar plot to Figure 4.3 can be generated for vehicle ratio distribution, illustrating spatial variations in vehicle presence across municipalities. However, the spatial distribution differs from that of potholes and people, reflecting differences in road density, traffic flow, and urban planning. This choropleth is shown in Appendix 4.1.3.

### 4.1.4 Illegal Dumping

A total of 10 039 illegal dumping sites were detected across the 157 338 collected images. Figure 4.8 presents the distribution of confidence scores assigned to these detections. Compared to other detected features (vehicles, people, and potholes), confidence scores for illegal dumping sites tend to be lower, with most detections falling below 0.5.

This lower confidence is likely due to the significant variation in the appearance of illegal dumping sites. Unlike more structured objects like vehicles and potholes, illegal dumping can take many different forms, ranging from scattered debris to large waste piles, making it more challenging for the model to generalise detections with high certainty.

Visual inspection indicates that while the model generally classifies illegal dumping sites correctly, the rate of false positives is higher compared the other models. As shown in Figure 4.9, some detections with lower confidence scores include objects or areas that resemble illegal dumping but are not actually waste sites, such as construction debris or natural ground cover.

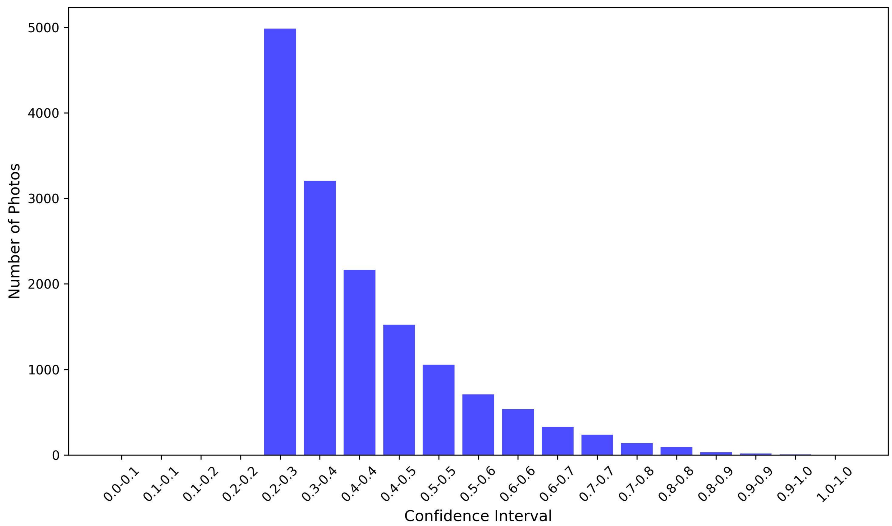


FIGURE 4.8: Distribution of illegal dumping detection confidence scores across collected GSV images in South Africa


Figure 4.9: Images showing variability in confidence scores due to variation in illegal dumping sites

A similar spatial distribution plot to Figure 4.3 could be generated for illegal dumping, illustrating its prevalence across municipalities. However, the detected locations differ significantly from those of potholes, vehicles and people, reflecting differences in waste disposal practices and service delivery across urban and rural environments. This choropleth plot in shown in Appendix 4.1.4.

## 4.2 Modelling Voting Patterns at The Ward Level

At the ward level, ML models were used to predict the vote percentage of the four dominant political parties (African National Congress (ANC), Democratic Alliance (DA), uMkhonto we Sizwe (MK), and Economic Freedom Fighters (EFF), as well as the non vote percentage). The analysis relied on features derived from GSV imagery, capturing urban and environmental characteristics such as the presence of potholes, illegal dumping, vehicles, and pedestrians. These high-resolution features provided a spatial perspective into the built environment, while additional datasets, such as municipal financial and crime data, were only available at the municipal level.

Among the tested models, linear regression and random forest demonstrated the strongest performance. linear regression required feature reduction to improve interpretability and prevent overfitting, while random forest handled high-dimensional data without the need for explicit feature selection. Model evaluation was conducted using cross-validation, measuring R-2 scores, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to assess predictive accuracy. Furthermore, plots comparing the predicted vs. actual vote percentages are presented, to further enhance the analysis of the model predictions.

To complement the regression analysis, a classification approach was also implemented to determine the dominant political party in each ward. Instead of predicting vote percentages, this method classified wards based on the party with the highest vote share. By framing the problem as a categorical prediction task, the classification models offered an alternative way to examine voting patterns.

Logistic regression and random forest classifiers were used to model this classification task. Logistic regression provided a probabilistic framework for understanding the likelihood of a ward being assigned to a specific party, while random forest captured non-linear feature interactions that may influence party dominance. These classification models were evaluated using precision, recall, and F1-scores.

*A Note On How The Results Are Presented*

The results of the ML models are presented across two distinct subsets of data. The first subset includes all available wards across South Africa (3 636 wards), providing a comprehensive analysis of voting patterns at a national level. The second subset focuses specifically on the top 23 municipalities with the highest population density (1 258 wards).

These high-density municipalities have a greater number of collected images and represent urban environments that may differ from the more rural municipalities. Figure 4.10 illustrates the spatial distribution of these two groups, highlighting the distinction between densely populated urban areas and the remaining, predominantly rural areas.

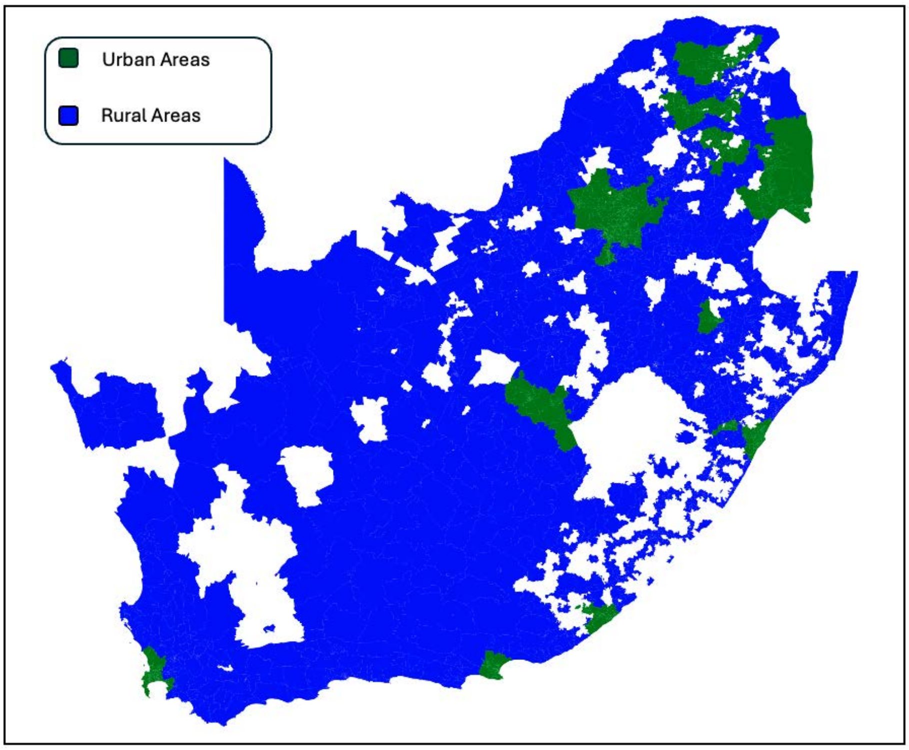


FIGURE 4.10: Spatial distribution of high-density urban and rural municipalities in South Africa

### 4.2.1 ML Regression Model Performance

Table 4.1 presents the R-2 values for both the linear regression (linear regression (LR)) and random forest (Random Forest (RF)) models on the train and test sets. These values correspond to the models predicting the vote percentage of the four dominant political parties (ANC, DA, EFF, and MK), as well as the non-vote percentage, across all wards (3 636 wards). Table 4.2 presents the same values for the models trained on the urban subset (1 258 wards), which includes the 23 most densely populated municipalities.

Tables 4.3 and 4.5 provide the MAE, MSE, and RMSE for the linear regression and random forest models, respectively, evaluated on the test set across all wards. Similarly, Tables 4.4 and 4.6 present these error metrics for the models trained on the urban wards.

TABLE 4.1: $R^2$ values of the linear regression (LR) and random forest (RF) models using cross-validation with 10 folds on all wards on the train and test set

| Target (Label) | LR Train $R^2$ | LR Test $R^2$ | RF Train $R^2$ | RF Test $R^2$ |
|---|---|---|---|---|
| Non Vote % | 0.3683 | 0.3461 | 0.4617 | 0.4136 |
| ANC Vote % | 0.6767 | 0.6720 | 0.7561 | 0.7004 |
| DA Vote % | 0.5828 | 0.5402 | 0.6981 | 0.6428 |
| MK Vote % | 0.7849 | 0.7769 | 0.8212 | 0.8256 |
| EFF Vote % | 0.3861 | 0.3788 | 0.4161 | 0.3942 |

TABLE 4.2: $R^2$ values of the linear regression (LR) and random forest (RF) models using cross-validation with 10 folds on urban wards on the train and test set

| Target (Label) | LR Train $R^2$ | LR Test $R^2$ | RF Train $R^2$ | RF Test $R^2$ |
|---|---|---|---|---|
| Non Vote % | 0.5226 | 0.3480 | 0.5700 | 0.3658 |
| ANC Vote % | 0.7300 | 0.7056 | 0.8078 | 0.6887 |
| DA Vote % | 0.6679 | 0.5839 | 0.7394 | 0.5839 |
| MK Vote % | 0.7423 | 0.7876 | 0.8416 | 0.7735 |
| EFF Vote % | 0.3349 | 0.2649 | 0.3957 | 0.2743 |

Across both datasets, the random forest model consistently outperformed the linear regression model, achieving higher R-2 values on both the training and test sets. This suggests that the random forest model captures non-linear relationships in the data more effectively, leading to improved predictive performance.

Notably, the difference between the train and test R-2 values is relatively small, particularly in the national dataset, where the test R-2 values are close to the train R-2 values for all models. This indicates that the models generalise well, meaning they do not suffer significantly from overfitting and can be applied to new data. The urban dataset shows slightly larger gaps between train and test R-2 performance, particularly for the random forest model, which suggests a minor overfitting effect in denser urban environments where the data distribution may be more complex.

The models performed particularly well in predicting the vote percentages for the ANC, DA, and MK, with R-2 values above 0.67 across both datasets. In contrast, the EFF and non-vote percentages yielded lower R-2 values, though they remain within a range that suggests statistical significance. The lower predictive power for these categories may reflect greater variability in voting patterns for these groups, potentially due to differences in voter behaviour or regional disparities that are harder to capture with the available features.

Table 4.3: MAE, MSE, RMSE of the linear regression (LR) model using cross-validation with 10 folds on all wards on the test set

| Target (Label) | LR MAE | LR MSE | LR RMSE |
|---|---|---|---|
| Non Vote % | 4.6769 | 37.6348 | 6.1347 |
| ANC Vote % | 11.3474 | 205.9050 | 14.3494 |
| DA Vote % | 11.2060 | 205.1869 | 14.3243 |
| MK Vote % | 5.3521 | 72.9212 | 8.5394 |
| EFF Vote % | 4.2589 | 33.0997 | 5.7532 |

Table 4.4: MAE, MSE, RMSE of the linear regression (LR) model using cross-validation with 10 folds on urban wards on the test set

| Target (Label) | LR MAE | LR MSE | LR RMSE |
|---|---|---|---|
| Non Vote % | 4.5249 | 32.0464 | 5.6610 |
| ANC Vote % | 10.7686 | 180.1923 | 13.4236 |
| DA Vote % | 11.5192 | 200.8454 | 14.1720 |
| MK Vote % | 6.4227 | 84.9906 | 9.2190 |
| EFF Vote % | 4.4302 | 35.7597 | 5.9799 |

Table 4.5: MAE, MSE, RMSE of the random forest (RF) model using cross-validation with 10 folds on all wards on the test set

| Target (Label) | RF MAE | RF MSE | RF RMSE |
|---|---|---|---|
| Non Vote % | 4.2997 | 31.8082 | 5.6399 |
| ANC Vote % | 9.6274 | 168.4916 | 12.9804 |
| DA Vote % | 8.3559 | 148.2863 | 12.1773 |
| MK Vote % | 3.9696 | 52.6565 | 7.2565 |
| EFF Vote % | 4.1701 | 32.7503 | 5.7228 |

Table 4.6: MAE, MSE, RMSE of the random forest (RF) model using cross-validation with 10 folds on urban wards on the test set

| Target (Label) | RF MAE | RF MSE | RF RMSE |
|---|---|---|---|
| Non Vote % | 4.1548 | 27.6450 | 5.2579 |
| ANC Vote % | 8.5987 | 143.8259 | 11.9927 |
| DA Vote % | 9.2557 | 174.0401 | 13.1924 |
| MK Vote % | 4.6433 | 61.4011 | 7.8359 |
| EFF Vote % | 4.3708 | 36.5264 | 6.0437 |

Among the five predicted targets, the ANC vote percentage exhibits the largest error values. However, this is expected given the broader distribution of ANC votes across wards. In contrast, the MK vote percentage shows lower error values, which aligns with its right-skewed distribution, where the majority of wards have low MK vote shares with a few high-density concentrations. As a result, predictions for the MK vote percentage are inherently less variable, leading to lower error metrics.

Figures 4.12 and 4.11 further illustrate this difference, showing the actual and predicted distributions of the ANC and MK vote percentages on the national test dataset using the final random forest model where the x axis shows the vote % buckets and the y axis shows the number of wards that full into each bucket. The ANC vote percentage has a more dispersed distribution, while the MK vote percentage is more concentrated, reinforcing the relationship between the distribution shape and prediction error magnitude.

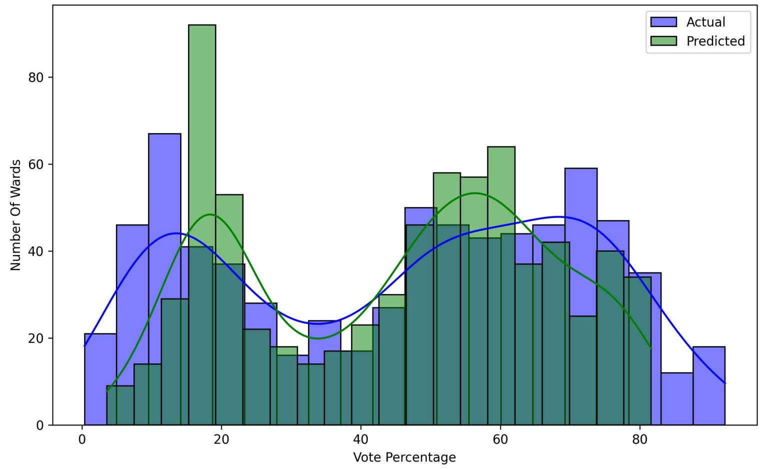


FIGURE 4.11: Actual and predicted distribution of ANC vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category

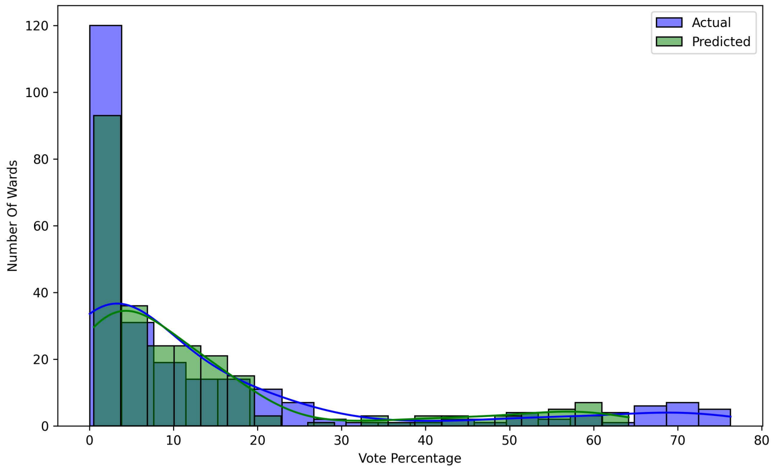


FIGURE 4.12: Actual and predicted distribution of MK vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category

These findings indicate that model performance is influenced not only by feature importance and predictive power but also by the underlying distribution of the target variables, which affects how errors propagate across different voting patterns. The distribution plots for the additional targets (DA, MK, EFF, Non Vote) are shown in the Appendix 4.2.1.

### 4.2.2 Visualising Model Performance

To further assess the predictive accuracy of the ML models, this section presents visualisations comparing actual and predicted vote percentages, as well as the distribution of residual errors. These plots provide insights into how well the models generalise across different wards and political parties. For this analysis, we focus exclusively on the national-level dataset, due to its better performance and generalisability, and the random forest models which showed better results as compared to the linear regression models, allowing for a comprehensive evaluation of model.

Two types of visualisations are presented for each political party:

i. **Actual vs Predicted Scatter Plots:** These plots illustrate the predicted vote percentage against the actual vote percentage across all wards. Ideally, a well-performing model should exhibit a strong alignment along the diagonal line, indicating minimal error. Deviations from this trend may suggest systematic biases or challenges in capturing extreme cases.

ii. **Residual Distribution Histograms:** These histograms depict the distribution of errors, calculated as the difference between actual and predicted vote percentages. A symmetric, bell-shaped distribution centred around zero suggests a well-calibrated model, whereas skewness or heavy tails may indicate areas where the model systematically under- or overpredicts.

*Non-Vote Percentage*

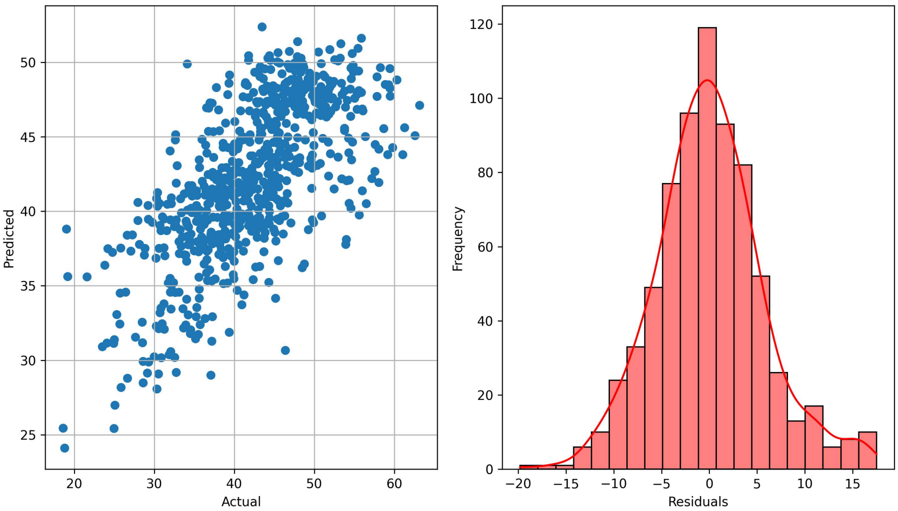


FIGURE 4.13: Actual vs predicted scatter plot (left) and residual distribution (right) for non-vote percentage

Figure 4.13 presents the actual vs predicted scatter plot and the residual distribution for the non-vote percentage.

i. The scatter plot indicates a strong positive correlation between actual and predicted non-vote percentages. The data points generally cluster along the diagonal, demonstrating that the model captures broad trends in non-voting behaviour. However, some degree of dispersion remains.

ii. The residual distribution is approximately normal and centred around zero, suggesting that the model does not systematically over- or under-predict non-voting rates. However, the slight asymmetry suggests that the model is overestimating non vote %.

*ANC Vote Percentage*

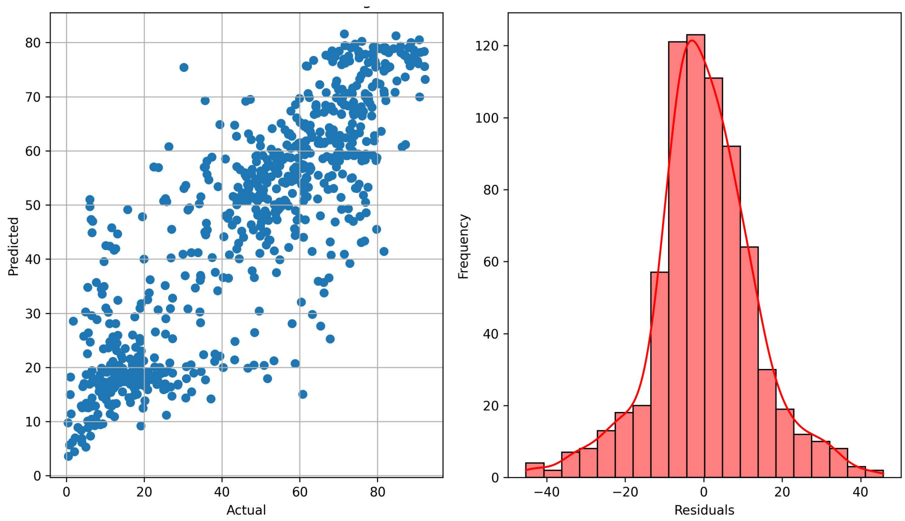


FIGURE 4.14: Actual vs predicted scatter plot (left) and residual distribution (right) for ANC vote percentage using the final random forest model

Figure 4.14 presents the actual vs predicted scatter plot and the residual distribution for the ANC vote percentage.

i. The scatter plot shows a strong correlation between actual and predicted values, with data points following a diagonal trend. However, some dispersion is evident, particularly at higher vote shares, where the model appears to struggle more with extreme cases.

ii. The residual distribution is approximately symmetric and centred around zero, indicating no strong bias in over- or under-predicting ANC vote share. However, the presence of larger residuals in the tails suggests that certain wards deviate significantly from the predicted values.

*DA Vote Percentage*

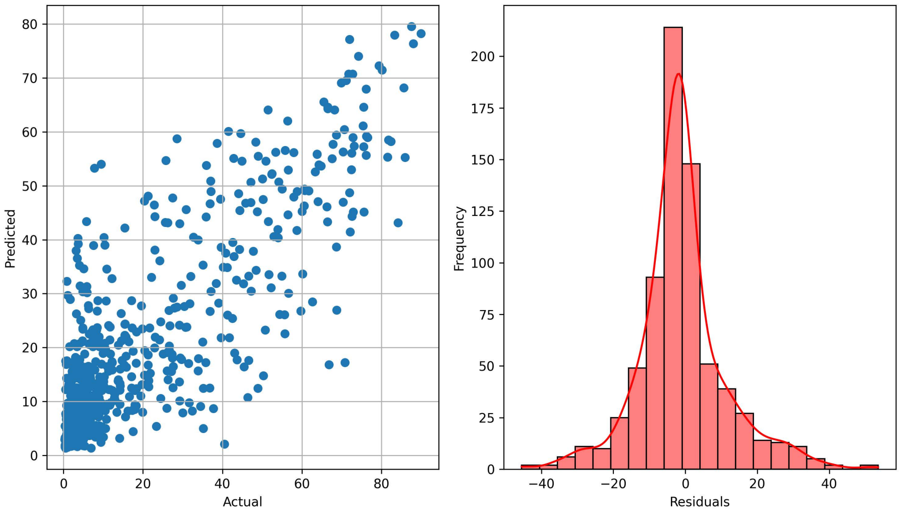


FIGURE 4.15: Actual vs predicted scatter plot (left) and residual distribution (right) for DA vote percentage using the final random forest model

Figure 4.15 presents the actual vs predicted scatter plot and the residual distribution for the DA vote percentage.

i. The scatter plot shows a general alignment along the diagonal trend, indicating that the model captures DA vote share reasonably well. However, a denser concentration of points near the lower vote percentages suggests that the model may struggle to differentiate among wards with low DA support.

ii. The residual distribution is approximately symmetric and centred around zero, indicating an overall balanced prediction. However, the slightly wider tails suggest that the model produces larger errors in some wards, potentially in areas where DA support is low.

*MK Vote Percentage*

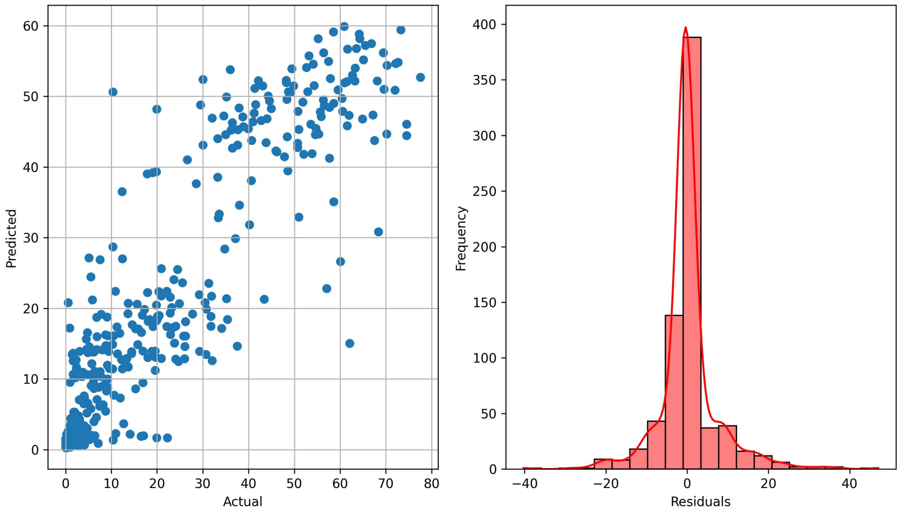


FIGURE 4.16: Actual vs predicted scatter plot (left) and residual distribution (right) for MK vote percentage using the final random forest model

Figure 4.16 presents the actual vs predicted scatter plot and the residual distribution for the MK vote percentage.

i. The scatter plot shows a clear alignment along the diagonal, indicating that the model captures the MK vote percentage well. The spread of predictions is relatively narrow, suggesting that the model is confident in its estimates. However, variations can be seen at the mid ranges.

ii. The residual distribution is tightly centred around zero, with fewer extreme residuals compared to other parties. This suggests that the model makes relatively consistent predictions with minimal large errors. The slight asymmetry in the residuals may indicate some overestimations, potentially in wards where MK support is high.

*EFF Vote Percentage*

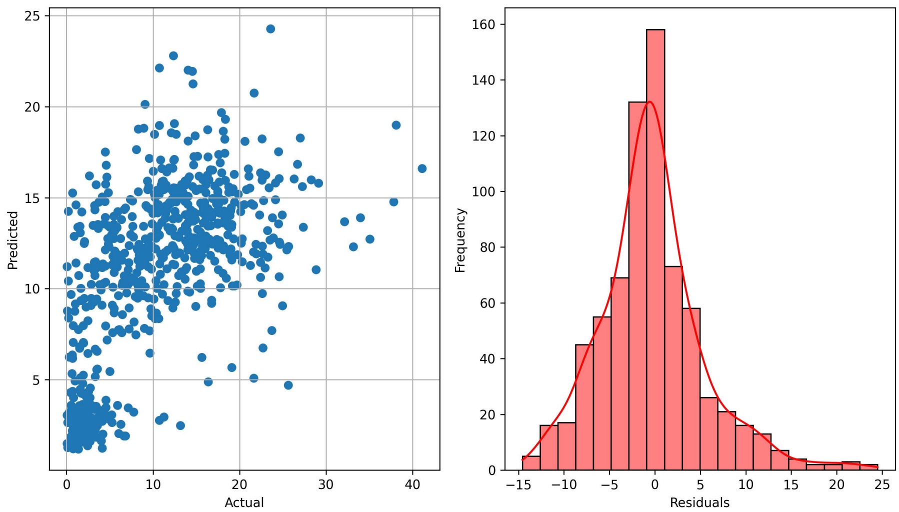


FIGURE 4.17: Actual vs predicted scatter plot (left) and residual distribution (right) for EFF vote percentage using the final random forest model

Figure 4.17 presents the actual vs predicted scatter plot and the residual distribution for the EFF vote percentage.

i. The scatter plot shows a poor alignment along the diagonal, indicating that the model captures some underlying patterns in the EFF vote percentage. However, there is notable dispersion, particularly at lower vote shares, suggesting higher variability in predictions for areas where EFF support is weaker.

ii. The residual distribution is asymmetric compared to the other models. The large right skewed tail reflects the overestimation of EFF vote percentage seen in the scatter plot.

*Summary Of Visualisation*

The visualisations indicate that the models predict vote percentages well, with ANC, DA, and MK showing strong correlations, while EFF and non-vote percentages have lower accuracy but remain statistically significant. Residual distributions are centred around zero, suggesting no major bias, though heavy tails indicate occasional over- or under-predictions. Overall, the models capture broad voting trends but may struggle with extreme cases. The next section explores feature importance to better understand these predictions.

### 4.2.3 ML Classification Model Performance

In addition to the regression analysis, a classification approach was employed to identify the dominant political party in each ward. Rather than predicting vote percentages, this method categorised wards based on the party with the highest vote share, providing a different perspective on electoral trends.

Logistic regression and random forest classifiers were used to model this task, offering both a probabilistic interpretation and the ability to capture complex, non-linear relationships. Performance was assessed using key metrics including precision, recall, and F1-score. This classification framework helps to further understand the patterns that drive party dominance across wards.

The classification was performed on the full national dataset (3 636 wards), ensuring a comprehensive analysis of voting patterns across different geographic and socioeconomic contexts.

Figure 4.18 presents the distribution of wards where each political party secured the highest vote share, along with the number of wards won by the summation of the remaining parties. The ANC dominates the classification dataset, winning the majority of wards, followed by the DA and MK. Furthermore, figure 4.18 shows that the number of wards won by the EFF and the combined set of remaining parties is notably small. Due to this limited representation, the EFF and remaining parties are excluded from further analysis to ensure model stability and meaningful classification performance as shown in figure 4.19.

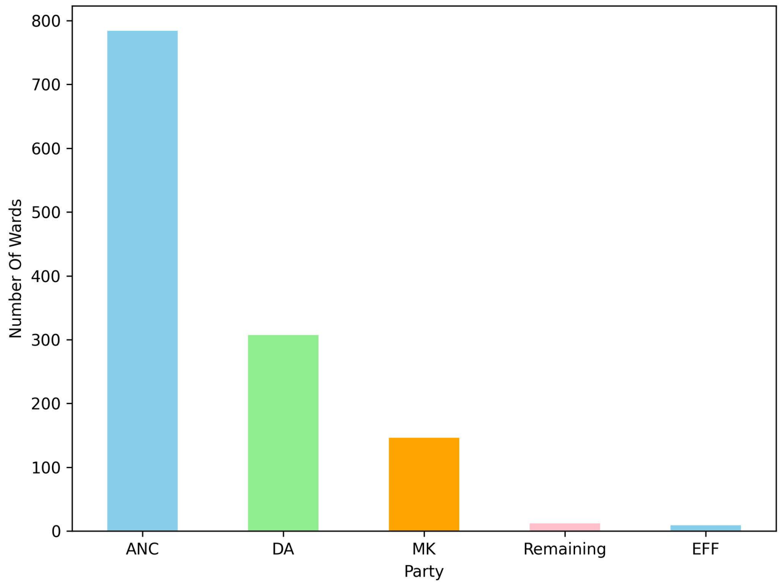


FIGURE 4.18: Distribution of number of wards won by the dominant parties and the summation of the remaining parties in the national dataset

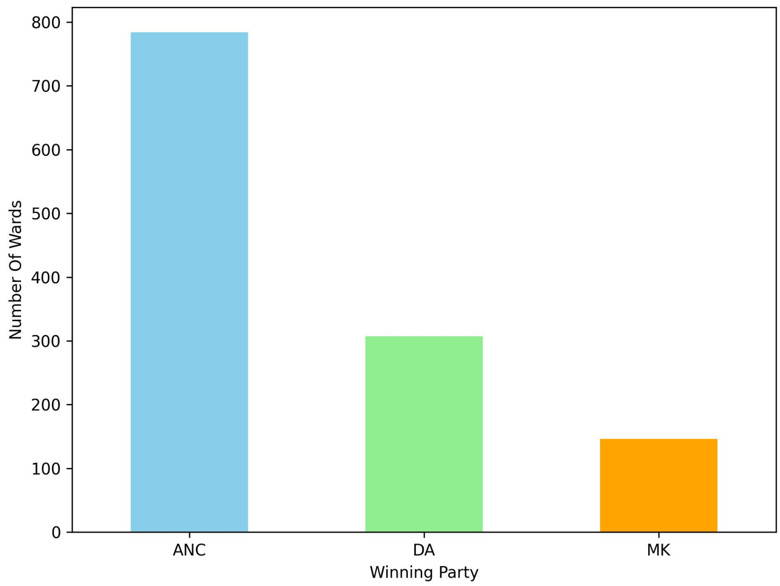


FIGURE 4.19: Distribution of number of wards won by the ANC, DA, MK in the national test dataset

Tables 4.7 and 4.8 present the performance metrics for the logistic regression and random forest classifiers, respectively, while Figures 4.20 and 4.21 provide the corresponding confusion matrices. Both models demonstrate strong performance in classifying the majority vote for each ward, with high precision, recall, and F1-scores across all three dominant political parties: ANC, DA, and MK.

For both classifiers, the ANC class achieves the highest recall, indicating that most ANC-majority wards are correctly identified. Precision and recall for the DA is slightly lower than those for the ANC, reflecting a higher degree of misclassification between DA and ANC-majority wards. The MK class also achieves high precision, meaning that when the model predicts an MK-majority ward, it is usually correct. However, the recall for MK is slightly lower in the random forest model compared to logistic regression, indicating that some MK-majority wards are misclassified.

The confusion matrices further illustrate the classification performance, showing that the most frequent misclassification occurs between the DA and ANC, which aligns with their close competition in some areas. Misclassification between ANC and MK is minimal, likely due to their distinct geographic strongholds.

Overall, both models perform well, with only marginal differences in performance. Logistic regression offers slightly higher recall for MK, while the random forest model provides better overall stability across the three classes. Given these results, both classifiers provide valuable insights into the dominant political party in each ward, reinforcing the patterns observed in the regression analysis.

Table 4.7: Performance metrics of the logistic regression classifier on classifying majority vote on the test national dataset

| Target (Label) | Precision | Recall | F1-score |
|---|---|---|---|
| ANC | 0.96 | 0.97 | 0.96 |
| DA | 0.89 | 0.89 | 0.89 |
| MK | 0.96 | 0.90 | 0.93 |

Table 4.8: Performance metrics of the random forest classifier on classifying majority vote on the test national dataset

| Target (Label) | Precision | Recall | F1-score |
|---|---|---|---|
| ANC | 0.94 | 0.96 | 0.95 |
| DA | 0.84 | 0.85 | 0.85 |
| MK | 0.96 | 0.83 | 0.89 |

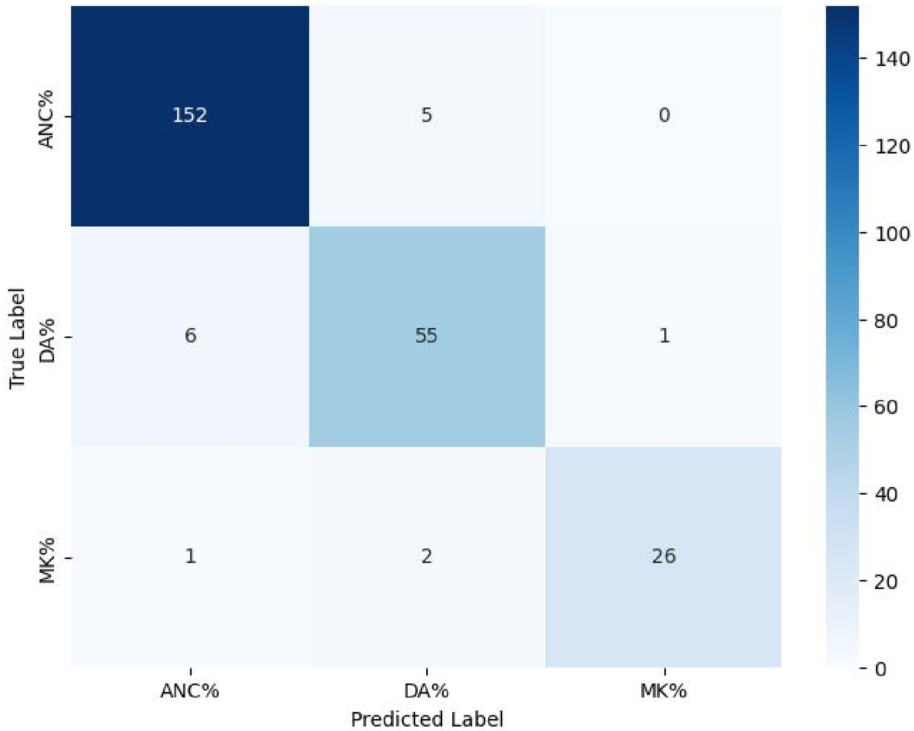


Figure 4.20: Confusion matrix of the logistic regression model in predicting majority vote on national test dataset

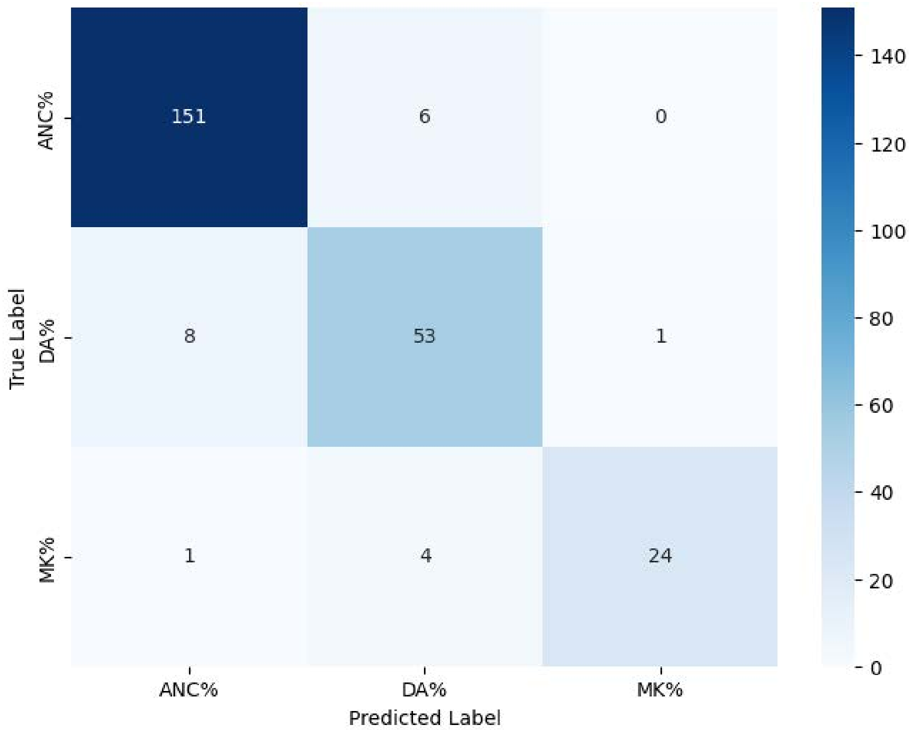


FIGURE 4.21: Confusion matrix of the random forest model in predicting majority vote on national test dataset

*Effect of Removing Provincial One-Hot Encoding*

To assess the extent to which regional factors drive classification performance, an additional experiment was conducted by removing the one-hot encoded provincial features from the dataset. This allows us to determine whether the models rely on geographic location as a key predictor of voting patterns or whether other socio-economic and infrastructural features are sufficient for accurate classification.

The results, presented in Tables 4.9 and 4.10, show a significant drop in classification performance when the provincial one hot encoded features are dropped, particularly for the MK. The recall for the MK drops to near zero for both models, indicating that without regional data, the models struggle to correctly identify MK-dominated wards. This suggests that the MK vote is heavily driven by location, reinforcing the observation that its support is highly concentrated in KwaZulu-Natal.

Additionally, the performance for the ANC also declines, albeit to a lesser extent, which could imply that ANC and MK share similar characteristics in the feature space. The DA, on the other hand, maintains relatively stable performance, suggesting that its vote share is less dependent on regional factors and more associated with broader socio-economic or infrastructure-related features.

The confusion matrices in Figures 4.22 and 4.23 further illustrate these effects. The MK is frequently misclassified, primarily as ANC, reinforcing the notion that these two parties share similar feature characteristics outside of location. Again, the ANC also experiences a decline in predictive performance, while the DA remains largely unaffected.

These findings highlight the extent to which regional context influences voting behaviour, particularly for the MK and ANC. The stark decline in MK classification performance without provincial data suggests that geographic information is one of the strongest determinants of its voter base.

TABLE 4.9: Performance metrics of logistic regression classifier on classifying majority vote on the test national dataset without including provinces

| Target (Label) | Precision | Recall | F1-score |
|---|---|---|---|
| ANC | 0.83 | 0.97 | 0.89 |
| DA | 0.89 | 0.87 | 0.88 |
| MK | 0.75 | 0.10 | 0.19 |

TABLE 4.10: Performance metrics of random forest classifier on classifying majority vote on the test national dataset without including provinces

| Target (Label) | Precision | Recall | F1-score |
|---|---|---|---|
| ANC | 0.82 | 0.96 | 0.88 |
| DA | 0.84 | 0.84 | 0.84 |
| MK | 1.0 | 0.03 | 0.07 |

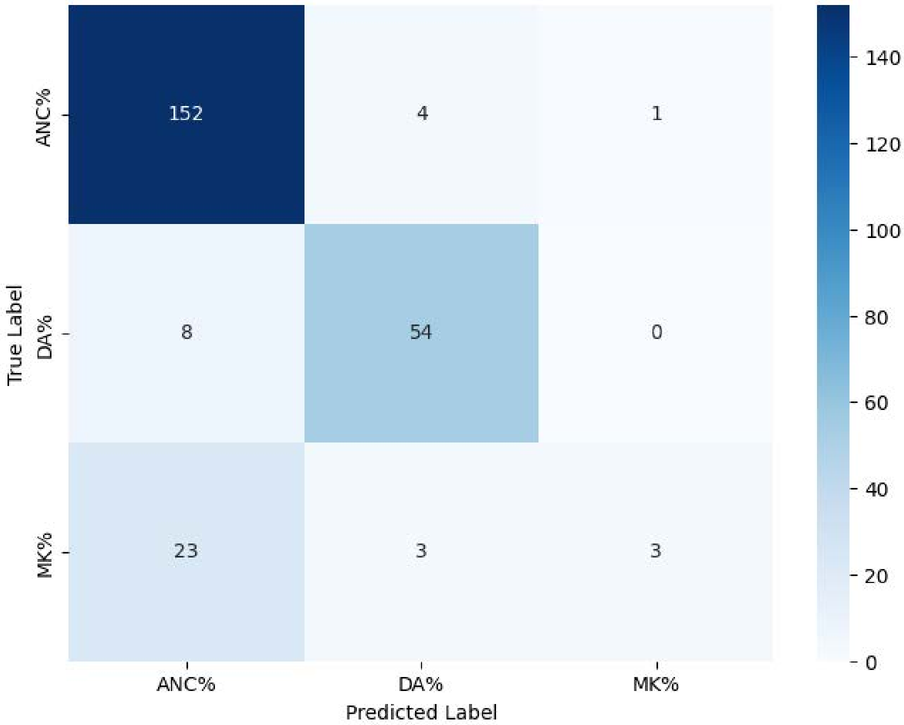


FIGURE 4.22: Confusion matrix of logistic regression model in predicting majority vote on national test dataset without including provinces

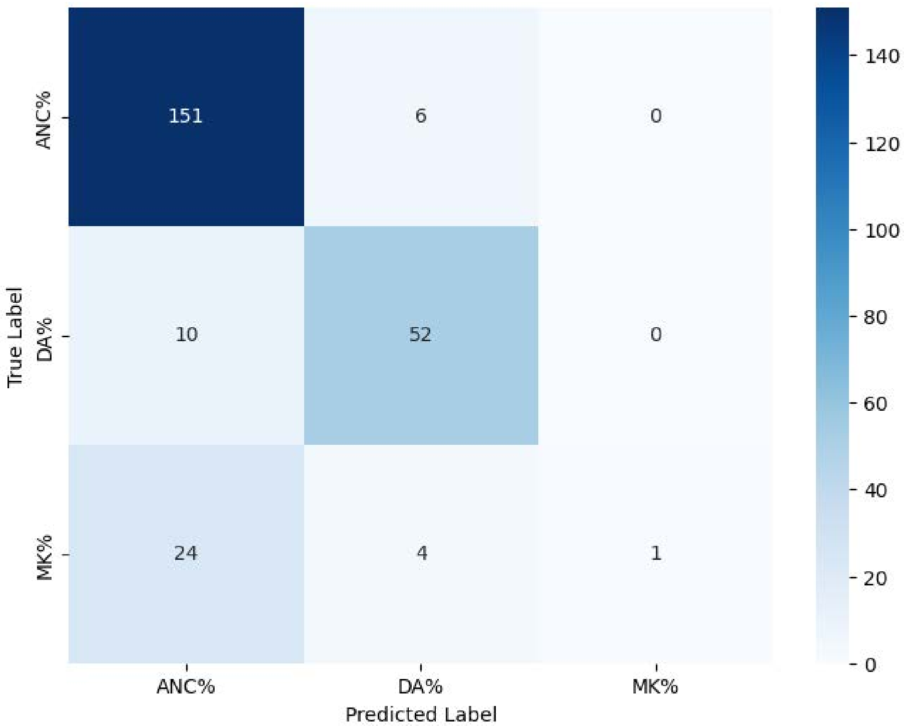


FIGURE 4.23: Confusion matrix of random forest model in predicting majority vote on national test dataset without including provinces

## 4.3 Modelling Voting Patterns at The Municipal Level

At the municipal level, the analysis shifts from wards to municipalities, which represent broader administrative regions encompassing multiple wards. This transition allows for the integration of additional data sources that provide a more comprehensive view of the socioeconomic and environmental factors influencing voting patterns.

Statistical regression models were used instead of ML due to the smaller number of municipalities (213). The modelling process involved standardising numerical features using Z-score normalisation, performing variance inflation factor (VIF) analysis to reduce multicollinearity, and applying three regression techniques independently: Lasso, Ridge, and Elastic Net. These models help identify the most influential predictors of voting behaviour by selecting features that exhibit strong statistical relationships with voting outcomes.

Unlike the ward-level analysis, which primarily relied on GSV-derived features, the municipal-level study incorporates additional datasets, including satellite imagery, crime data, and financial data. The inclusion of these variables enables a more holistic understanding of voting behaviour by capturing broader economic and infrastructure conditions.

At this resolution, we removed the one-hot encoded provincial features. During initial tests, province-based categorical data dominated model performance, significantly reducing the impact of other relevant socioeconomic and environmental factors.

The model outputs from the regressions are provided to evaluate the performance of the models. These outputs include key performance metrics such as R-squared values, coefficient estimates, confidence intervals, and statistical significance values, which help assess how well the models explain variation in voting outcomes.

### 4.3.1 Model Performance Post Feature Reduction

As shown in tables (4.11,4.12,4.13,4.14,4.15), all three regression techniques (Lasso, Ridge, and Elastic Net) yield fairly consistent results within each target, as evidenced by similar R-squared and Adjusted R-squared values. This consistency suggests that the choice of regularisation method does not drastically alter predictive accuracy for these datasets, at least in terms of the standard Ordinary Least Squares (OLS) metrics shown. However, each method might still select or weight predictors differently due to the nature of their penalties—Lasso encourages sparsity by zeroing out weaker predictors, Ridge shrinks coefficients more uniformly, and Elastic Net combines elements of both.

The Non-Vote category sees R-squared values around 0.45–0.46, indicating that these predictors explain less than half of the outcome's variability. While lower than the others, it may still be meaningful if the Non-Vote behaviour is driven by complex or unobserved factors not fully captured by the chosen predictors. The Condition Number remains moderate, so collinearity is not a severe concern, but the lower R-squared and Adjusted R-squared highlight a potentially more challenging modelling problem.

For the ANC, the R-squared values hover around 0.70 across all three models, with an Adjusted R-squared close to 0.68. These figures indicate that roughly 70% of the variance in the ANC target variable is explained by the chosen predictors, which is a respectable level of explanatory power. The Condition Number for the ANC is between 4.5 and 5.2, suggesting that collinearity among features is not extreme, although it should still be monitored.

By contrast, the DA models achieve higher R-squared values, in the range of 0.85 to 0.86, which implies that the models capture more of the variability for the DA target variable than they do for the ANC. The Adjusted R-squared, at around 0.84–0.85, confirms that these models remain robust even after accounting for the number of predictors. The F-statistic is also notably higher than for the ANC, reflecting stronger overall model significance. Condition Numbers for the DA models are generally below 5.6, indicating a moderate level of multicollinearity.

In the MK case, the R-squared is around 0.74, while the Adjusted R-squared typically hovers near 0.72. This points to a solid explanatory model, although not quite as high as the DA. The Condition Number remains under 5, suggesting that while there may be some correlation among predictors, it is not alarmingly high.

Finally, the EFF shows slightly lower R-squared values, around 0.62, and an Adjusted R-squared near 0.59, which means the model explains a bit over half of the variance in the target. Although the Condition Number is similar to the MK's, the drop in R-squared suggests that the EFF's data might be less straightforward to model with linear predictors or that additional variables could improve fit.

Across all tables, the F-statistics are generally significant, indicating that at least one of the regression coefficients is non-zero in each model. The Condition Numbers—mostly between 3 and 6—suggest moderate correlations among predictors, but not so severe as to invalidate the regression assumptions. In practice, one might choose among Lasso, Ridge, or Elastic Net based on secondary considerations such as interpretability (Lasso often yields a sparser model), resistance to over-fitting (Ridge), or a balance of both (Elastic Net). Given the similar R-squared values, all three methods are viable. However, going forward the features from the Elastic Net Model are used to present the feature importance.

Table 4.11: OLS model performance summary for non-vote

| **Metric** | **Lasso Values** | **Ridge Values** | **Elastic Net Values** |
|---|---:|---:|---:|
| R-squared | 0.457 | 0.464 | 0.451 |
| Adjusted R-squared | 0.415 | 0.423 | 0.409 |
| F-statistic | 10.93 | 11.27 | 10.70 |
| Condition Number | 3.74 | 4.39 | 4.04 |

Table 4.12: OLS model performance summary for the ANC

| **Metric** | **Lasso Values** | **Ridge Values** | **Elastic Net Values** |
|---|---:|---:|---:|
| R-squared | 0.708 | 0.704 | 0.704 |
| Adjusted R-squared | 0.685 | 0.681 | 0.681 |
| F-statistic | 31.48 | 30.90 | 30.90 |
| Condition Number | 4.25 | 5.17 | 5.17 |

Table 4.13: OLS model performance summary for the DA

| **Metric** | **Lasso Values** | **Ridge Values** | **Elastic Net Values** |
|---|---|---|---|
| R-squared | 0.858 | 0.856 | 0.860 |
| Adjusted R-squared | 0.847 | 0.845 | 0.850 |
| F-statistic | 78.55 | 77.23 | 80.14 |
| Condition Number | 4.79 | 5.56 | 4.69 |

Table 4.14: OLS model performance summary for the MK

| **Metric** | **Lasso Values** | **Ridge Values** | **Elastic Net Values** |
|---|---|---|---|
| R-squared | 0.747 | 0.739 | 0.742 |
| Adjusted R-squared | 0.727 | 0.719 | 0.722 |
| F-statistic | 38.31 | 36.89 | 37.29 |
| Condition Number | 4.63 | 4.78 | 4.93 |

Table 4.15: OLS model performance summary for the EFF

| **Metric** | **Lasso Values** | **Ridge Values** | **Elastic Net Values** |
|---|---|---|---|
| R-squared | 0.620 | 0.616 | 0.620 |
| Adjusted R-squared | 0.591 | 0.586 | 0.591 |
| F-statistic | 21.25 | 20.82 | 21.25 |
| Condition Number | 4.45 | 4.45 | 4.39 |

# Chapter 5

# Findings: Key Features Affecting Voting Patterns

## 5.1 ML Correlations and Feature Analysis At The Ward Level

This chapter explores the most influential features associated with voting patterns at the ward level, using a combination of correlation analysis and feature importance scores derived from the trained random forest models. For each voting target, Non-Vote, African National Congress (ANC), Democratic Alliance (DA), uMkhonto we Sizwe (MK), and Economic Freedom Fighters (EFF), correlation tables present the top Pearson and Spearman coefficients, while feature importance plots illustrate the variables that contributed most significantly to each random forest model's predictions. This multi-pronged approach provides insight into both linear and non-linear relationships between built environment features and electoral outcomes.

By comparing these patterns across parties, the analysis sheds light on how lived physical conditions, captured through indicators such as potholes, illegal dumping, and Google StreetView (GSV) coverage, interact with regional, infrastructural, and socioeconomic contexts to shape voting behaviour. These findings are interpreted in light of the theoretical frameworks introduced in the literature review (Chapter 2), including rational choice theory, clientelism, and social structure determinants. In doing so, this chapter helps unpack the complex and often localised dynamics that influence political participation and party support in South Africa's evolving democracy.

### 5.1.1 Non Vote Percentage Feature Analysis

TABLE 5.1: Top 10 most important Pearson and Spearman correlation scores for the non-vote percentage

| Feature | Pearson Score |
|---|---|
| Pothole ratio | 0.126402 |
| People to car ratio | 0.125813 |
| Pothole lag difference | 0.114530 |
| Motorbike ratio | 0.064970 |
| People ratio | 0.059543 |
| Dumping lag difference | 0.032715 |
| Car lag difference | 0.027114 |
| Car ratio | -0.039446 |
| Road lag difference | -0.066431 |
| Road ratio | -0.346078 |

| Feature | Spearman Score |
|---|---|
| People to car ratio | 0.185008 |
| Pothole lag difference | 0.111472 |
| People lag difference | -0.033908 |
| Road lag difference | -0.065281 |
| People ratio | -0.035279 |
| Motorbike ratio | -0.069987 |
| Truck ratio | -0.108054 |
| Car ratio | -0.139190 |
| Bus ratio | -0.140755 |
| Road ratio | -0.361594 |

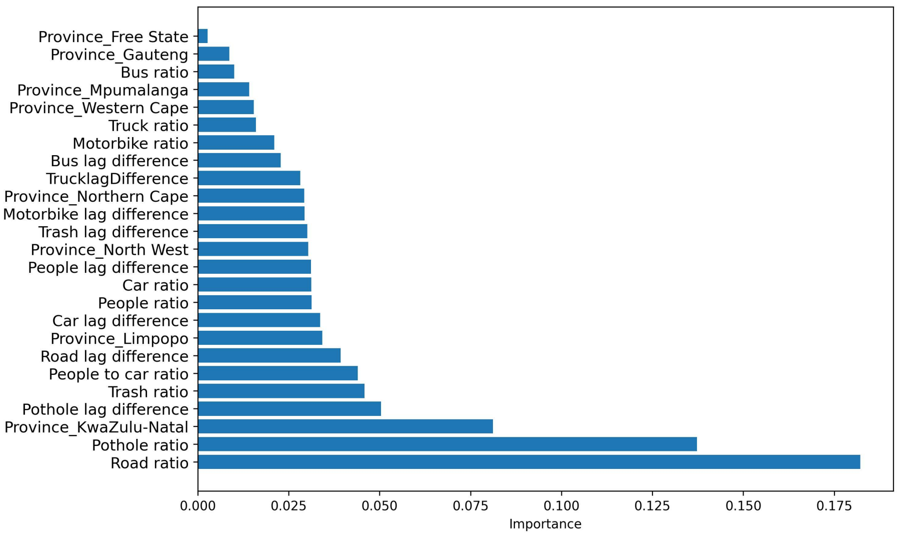


FIGURE 5.1: Feature importances of final random forest model in predicting non-vote percentage

Table 5.1 and Figure 5.1 reveal a consistent relationship between physical infrastructure, spatial representation, and voter abstention. Among the strongest positive correlations in Table 5.1 are the pothole ratio and pothole lag difference, suggesting that areas with poorer road conditions, either in absolute terms or relative to their surroundings, tend to exhibit higher rates of non-voting. This supports prior research linking poor service delivery to political disengagement [13, 19, 20]. Similarly, the people-to-car ratio shows a positive association with non-voting, implying that areas with limited car access, potentially a proxy for lower income and mobility, are less likely to participate in elections. These findings align with rational choice perspectives, in which citizens weigh the perceived benefits of voting against the effort or cost of participation [19, 7, 5].

The negative correlation between road ratio, a GSV coverage feature, and non-vote percentage introduces a spatial dimension. Areas with sparse GSV imagery are more likely to exhibit high abstention rates. As shown in the literature, GSV coverage tends to prioritise commercially important, well-serviced, and safe urban areas [26, 28, 29], while neglecting regions with poorer infrastructure, lower economic value, or safety concerns. Therefore, limited GSV coverage may signal deeper patterns of spatial exclusion and underdevelopment, both of which are likely to compound barriers to democratic participation.

Notably, province-level features, apart from KwaZulu-Natal, rank among the least important in the random forest model, suggesting that non-voting is not concentrated in particular geographic regions. Rather, abstention appears to be shaped by broader structural inequalities, transcending provincial boundaries. This insight is particularly important in the South African context, where voter turnout has declined steadily across the country despite the presence of regional political dynamics [16].

Taken together, these findings suggest that non-voting at the ward level is not merely a function of apathy or political indifference, but is rooted in material conditions and spatial inequality. Areas that are physically neglected, poorly mapped, or socially marginalised appear systematically less likely to engage electorally.

### 5.1.2 ANC Percentage Feature Correlations and Importance Plots

TABLE 5.2: Top 10 most important Pearson and Spearman correlation scores for the ANC vote percentage

| Feature | Pearson Score |
|---|---|
| People to car ratio | 0.190288 |
| Pothole lag difference | 0.088292 |
| Pothole ratio | 0.085806 |
| People ratio | 0.085176 |
| Motorbike ratio | 0.040984 |
| Dumping lag difference | 0.031513 |
| Dumping Ratio | 0.028536 |
| Road lag difference | -0.051285 |
| Car ratio | -0.166062 |
| Road ratio | -0.447259 |

| Feature | Spearman Score |
|---|---|
| People to car ratio | 0.302442 |
| Pothole lag difference | 0.092580 |
| People lag difference | -0.052192 |
| Road lag difference | -0.058099 |
| Motorbike ratio | -0.120574 |
| Dumping ratio | -0.128534 |
| Truck ratio | -0.199804 |
| Bus ratio | -0.211260 |
| Car ratio | -0.229042 |
| Road ratio | -0.473124 |

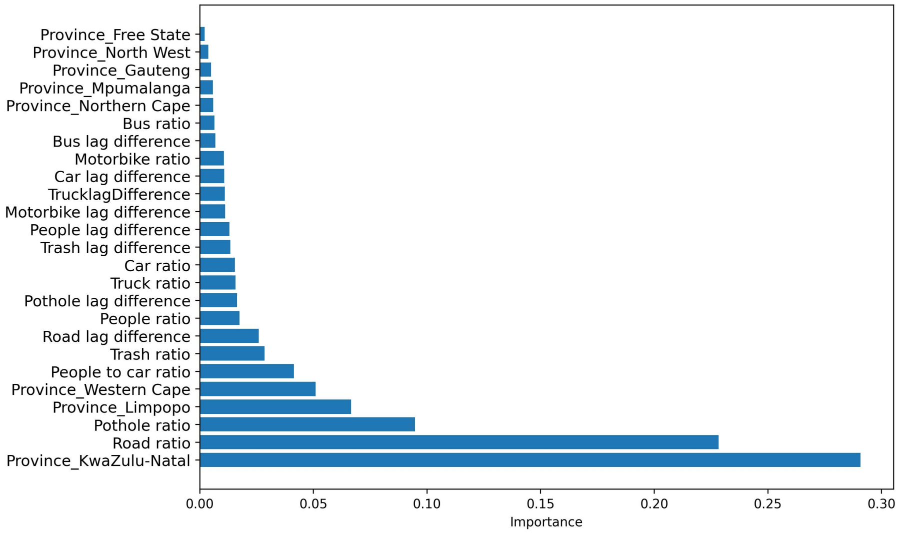


FIGURE 5.2: Feature importances of final random forest model in predicting ANC percentage

Table 5.2 and Figure 5.2 reveal several insights about ANC voting patterns at the ward level. Across both Pearson and Spearman measures, the People to Car Ratio exhibits the strongest positive correlation with ANC vote share. This suggests that areas with fewer cars per capita, potentially linked to lower-income communities, are more likely to support the ANC. This aligns with existing research showing that economically disadvantaged communities, particularly those historically excluded under apartheid, tend to favour the ANC, either due to material dependence on state resources (as suggested by clientelism theory) or enduring psychological ties to the party's liberation legacy [35, 32, 43].

Notably, GSV image coverage, captured through the road ratio, shows the strongest negative correlations with ANC support. Road ratio also ranks among the most important in the random forest model. According to the literature, GSV coverage is uneven and tends to prioritise economically significant, navigable, and safe areas, resulting in rural, poorer, or marginalised communities being underrepresented [26, 28, 29]. The strong negative relationship between GSV coverage and ANC vote share therefore offers a compelling signal: areas excluded from digital visibility are also those where ANC support remains concentrated. This suggests that GSV underrepresentation may serve as a proxy for spatial, infrastructural, and economic marginalisation; conditions which, as the literature shows, continue to shape patterns of political loyalty and state dependence.

While the Pearson and Spearman correlations for pothole ratio, and pothole lad difference as well as illegal dumping ratio, and illegal dumping lag difference are modest, their consistent appearance in both correlation tables and model importance rankings indicates a persistent, albeit subtle, relationship between poor infrastructure and ANC support. One possible explanation is that in areas facing infrastructural decline, voters may still support the incumbent party due to a lack of viable alternatives, long-standing local party networks, or psychological identification with the ANC as a historical liberator [44, 36]. This illustrates the complexity of voter behaviour, where dissatisfaction with services does not always translate into electoral punishment.

The feature importance plot further highlights KwaZulu-Natal and Limpopo as the two most influential provincial features for predicting ANC support. These provinces have historically served as ANC strongholds [40] and reflect the enduring relevance of regional identity and political history in shaping voter behaviour. This geographic persistence echoes broader findings in the literature on the intersection of race, region, and political affiliation in South Africa [37, 39].

Together, these results reinforce the notion that ANC support is deeply embedded in South Africa's socio-political geography. Support remains high in structurally disadvantaged, under-mapped, and poorly serviced areas. This suggests that social structure determinants, clientelism, and affective party loyalty continue to play a central role in driving voter preferences, even in the face of infrastructure decay and democratic dissatisfaction.

### 5.1.3 DA Percentage Feature Correlations and Importance Plots

Table 5.3: Top 10 most important Pearson and Spearman correlation scores for the DA vote percentage

| Feature | Pearson Score |
|---|---|
| Road ratio | 0.519156 |
| Car ratio | 0.172240 |
| Road lag difference | 0.085813 |
| Trash lag difference | -0.055363 |
| Dumping lag difference | -0.055363 |
| Dumping ratio | -0.055847 |
| Motorbike ratio | -0.067959 |
| Pothole lag difference | -0.118204 |
| Pothole ratio | -0.159776 |
| People to car ratio | -0.172658 |

| Feature | Spearman Score |
|---|---|
| Road ratio | 0.551136 |
| Car ratio | 0.236176 |
| Bus ratio | 0.235145 |
| Truck ratio | 0.221728 |
| People ratio | 0.163417 |
| Dumping ratio | 0.150044 |
| Motorbike ratio | 0.144519 |
| Pothole ratio | 0.012409 |
| Road lag difference | 0.078975 |
| People to car ratio | -0.237005 |

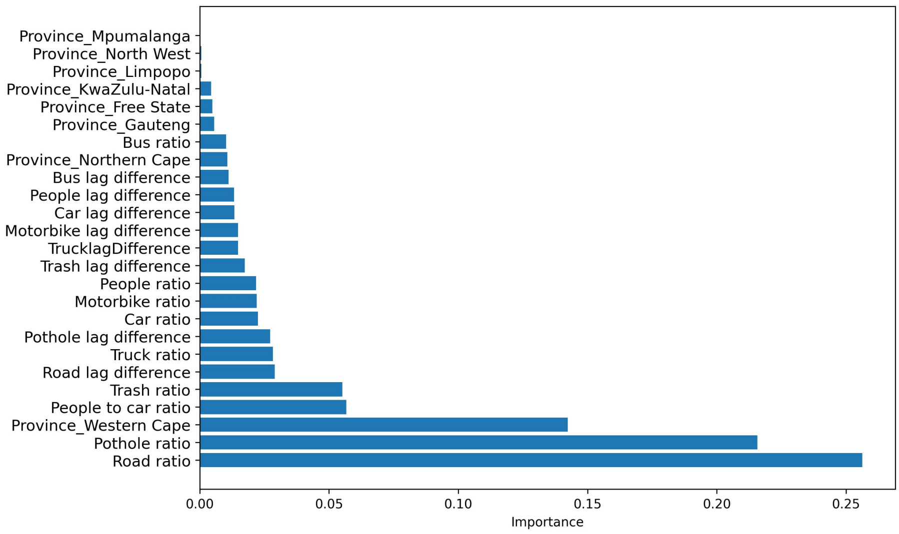


FIGURE 5.3: Feature importances of final random forest model in predicting DA percentage

Table 5.3 and Figure 5.3 reveal a strong association between urban infrastructure indicators and DA support. Most notably, the road ratio, a measure of GSV coverage, show the highest positive correlations with DA vote share across both Pearson and Spearman metrics. These two features also dominate the random forest feature importance plot, reinforcing their predictive power. Given GSV's selective spatial coverage — driven by commercial priorities, infrastructure accessibility, and safety perceptions, these results suggest that DA support is concentrated in areas that are digitally visible, navigable, and deemed commercially or infrastructurally significant.

This pattern aligns with social structure theories of voting, which posit that economic class and urban geography shape political alignment. The DA has historically found support among middle-class, urban, and predominantly non-black constituencies, groups more likely to reside in well-serviced and well-mapped environments [39]. The negative correlation with people-to-car ratio further strengthens this interpretation. A lower ratio implies greater car ownership and therefore higher household wealth and mobility, reinforcing the DA's connection to economically advantaged communities. Similarly, positive correlations with car, bus, and truck ratios suggest that DA-voting areas tend to exhibit higher vehicle presence, which may reflect both affluence and transport infrastructure.

Interestingly, the correlations with indicators of poor service delivery, such as pothole ratio and illegal dumping ratio, are generally negative, implying that the DA performs worse in areas with visible signs of infrastructural neglect. However, some of these correlations are weak or even slightly positive in the Spearman metrics (e.g., pothole ratio), suggesting the possibility of localised variation. This may reflect cases where the DA is seen as a viable alternative in areas experiencing service delivery failure, or where opposition support exists despite persistent infrastructure challenges.

From a geographic perspective, provincial features contribute relatively little to the model as compared to the other parties, with the Western Cape, the DA's traditional stronghold [40], being the only region with notable importance. This contrasts with the ANC and MK, where provinces like KwaZulu-Natal and Limpopo are significant predictors. The DA's relatively diffuse spatial distribution supports the idea that its base is shaped less by identity and historical allegiance and more by material living conditions and infrastructure quality.

Overall, these findings reinforce the link between DA support and digitally visible, infrastructurally serviced, and economically stable environments. The DA appears to represent a unique electoral geography, one rooted in physical accessibility, digital representation, and perceived governance efficacy. These insights, to some degree, support both rational choice and social structure theories, illustrating how physical and digital infrastructures interact to influence political outcomes.

### 5.1.4 MK Percentage feature correlations and importance plots

TABLE 5.4: Top 10 most important pearson and spearman correlation scores for the MK vote percentage

| Feature | Pearson Score |
|---|---|
| Pothole ratio | 0.040936 |
| Dumping ratio | 0.040741 |
| Pothole lag difference | 0.025725 |
| Dumping lag difference | 0.025350 |
| Motorbike ratio | 0.024351 |
| Truck lag difference | 0.017738 |
| People to car ratio | -0.036403 |
| Road lag difference | -0.050294 |
| People ratio | -0.058248 |
| Road ratio | -0.060057 |

| Feature | Spearman Score |
|---|---|
| Pothole ratio | 0.103315 |
| Motorbike ratio | 0.068273 |
| Dumping ratio | 0.062684 |
| Car ratio | 0.035876 |
| Road ratio | 0.033443 |
| Pothole lag difference | 0.030691 |
| Bus ratio | 0.025854 |
| Dumping lag difference | 0.019143 |
| People lag difference | 0.014323 |
| People to car ratio | -0.043965 |

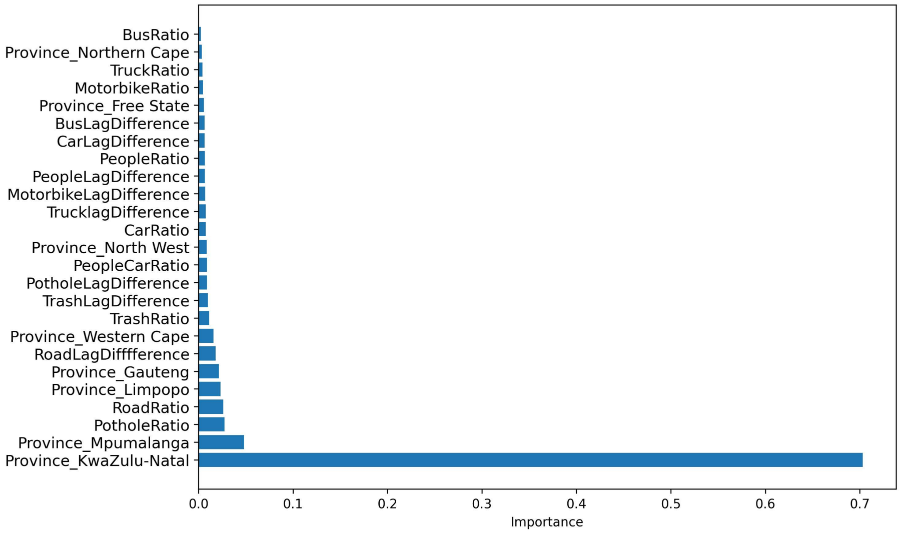


FIGURE 5.4: Feature importances of final random forest model in predicting MK percentage

Table 5.4 and Figure 5.4 reveal a strikingly regionalised and identity-driven dynamic. As shown in Table 5.4, the features most strongly correlated with MK vote share, such as pothole ratio, illegal dumping ratio, and motorbike ratio, exhibit only weak positive associations. Meanwhile, the road ratio display small but consistent negative correlations.

This suggests that MK support is concentrated in areas with low digital visibility and reduced infrastructural mapping; regions that are often rural, less accessible, and deemed commercially unimportant by platforms like GSV.

The presence of weak but positive correlations with service delivery challenges, such as potholes and illegal dumping, align with rational choice theory, where voters dissatisfied with the status quo may seek alternatives. In this case, MK may attract support from citizens who perceive a failure in governance, particularly from the ANC [40].

However, the dominant factor in predicting MK vote share is not infrastructure or service delivery, but regional identity. As Figure 5.4 demonstrates, KwaZulu-Natal overwhelmingly outstrips all other features in importance. This aligns with the 2024 general election results, where MK's rise was heavily concentrated in this province. Such localisation underscores the relevance of social structure and psychological theories of voting, which emphasise group affiliation, emotional loyalty, and charismatic leadership over rational cost-benefit calculations. MK's appeal appears to stem not merely from material conditions but from a broader sense of political and cultural identity tied to the legacy of its leader and its symbolic opposition to the ANC.

The prominence of KwaZulu-Natal in the feature importance plot also suggests that MK's electoral surge is less about national trends and more about a highly specific regional mobilisation. The limited importance of other provinces reinforces the party's localised character. In this way, MK mirrors the ANC's historical base in rural strongholds but reflects a fragmentation of that base along lines of dissatisfaction, political rivalry, and identity realignment.

Overall, MK's support profile illustrates a hybrid pattern. On the one hand, it shares some predictors with the ANC, such as weaker infrastructure and low digital mapping, but on the other, its support is far more geographically concentrated. This duality highlights the importance of integrating both infrastructural and identity-based explanations in interpreting voter behaviour, particularly in contexts of party splintering and political realignment.

### 5.1.5 EFF Percentage Feature Analysis

TABLE 5.5: Top 10 most important pearson and spearman correlation scores for the EFF vote percentage

| Feature | Pearson Score |
|---|---|
| Pothole ratio | 0.138998 |
| People ratio | 0.081855 |
| Car ratio | 0.067123 |
| People to car ratio | 0.063973 |
| Dumping ratio | 0.048620 |
| Pothole lag difference | 0.046217 |
| Car lag difference | 0.041299 |
| Motorbike lag difference | 0.039952 |
| Bus ratio | 0.033236 |
| Road ratio | -0.053801 |

| Feature | Spearman Score |
|---|---|
| Pothole ratio | 0.168161 |
| People ratio | 0.123710 |
| People to car ratio | 0.099904 |
| Pothole lag difference | 0.068340 |
| Motorbike ratio | 0.033154 |
| Dumping ratio | 0.023388 |
| Car lag difference | 0.020895 |
| Bus ratio | -0.025602 |
| Truck ratio | -0.040330 |
| Road ratio | -0.045979 |

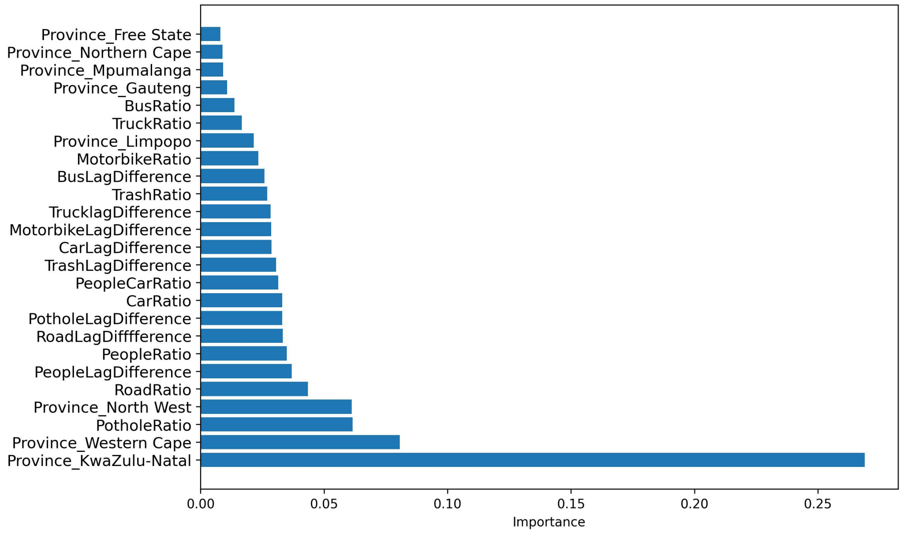


FIGURE 5.5: Feature importances of final random forest model in predicting EFF percentage

Table 5.5 and Figure 5.5 highlight an association between EFF support and indicators of infrastructural neglect and digital invisibility. Among the top features, the pothole ratio, people ratio, and car ratio all show positive correlations with EFF vote share across both Pearson and Spearman metrics, suggesting that the party finds support in areas with weaker service delivery, lower levels of private vehicle ownership, and denser, urban populations. These findings are consistent with rational choice theory, where voters in underserved areas evaluate their lived experiences and may turn to opposition parties in response to perceived state failure [8, 9].

The negative correlation between EFF support and the road ratio suggests that EFF voters are more likely to reside in areas with limited digital representation, possibly due to commercial disinterest, poor infrastructure, or safety concerns that deter comprehensive GSV coverage [26, 28, 29].

The prominence of geographic location, especially KwaZulu-Natal and the Western Cape, while likely a negative impact, suggests regional variation in EFF support, possibly shaped by localised contestation with other populist parties such as the MK, or by identity-based factors discussed under social structure determinants [39]. While province-level predictors were less prominent for the DA and NonVote, they are central to the EFF, highlighting how geography interacts with structural inequalities to shape electoral support. [23, 10, 20].

## 5.2 Feature Analysis at The Municipal Level

At the municipal level, statistical modelling complements the machine learning (ML) analysis at the ward level by offering more interpretable insights into the direction and magnitude of feature relationships. Elastic Net Regression is employed here for its ability to manage multicollinearity while performing variable selection and shrinkage, allowing for a balanced examination of correlated and high-dimensional features. Unlike ensemble-based models, which can be difficult to interpret directly, the linear nature of the Elastic Net model enables a clearer understanding of how specific features relate to voting outcomes.

The following importance plots visualise the top 15 most influential features for each outcome variable. These plots not only highlight the strongest associations between variables and voting behaviour but also distinguish between positive and negative influences. While the relationships observed should be interpreted with caution, they offer a valuable lens through which to understand how infrastructural, environmental, and social factors correlate with voter support and abstention at the municipal scale.

### 5.2.1 Non Vote Importance Plots

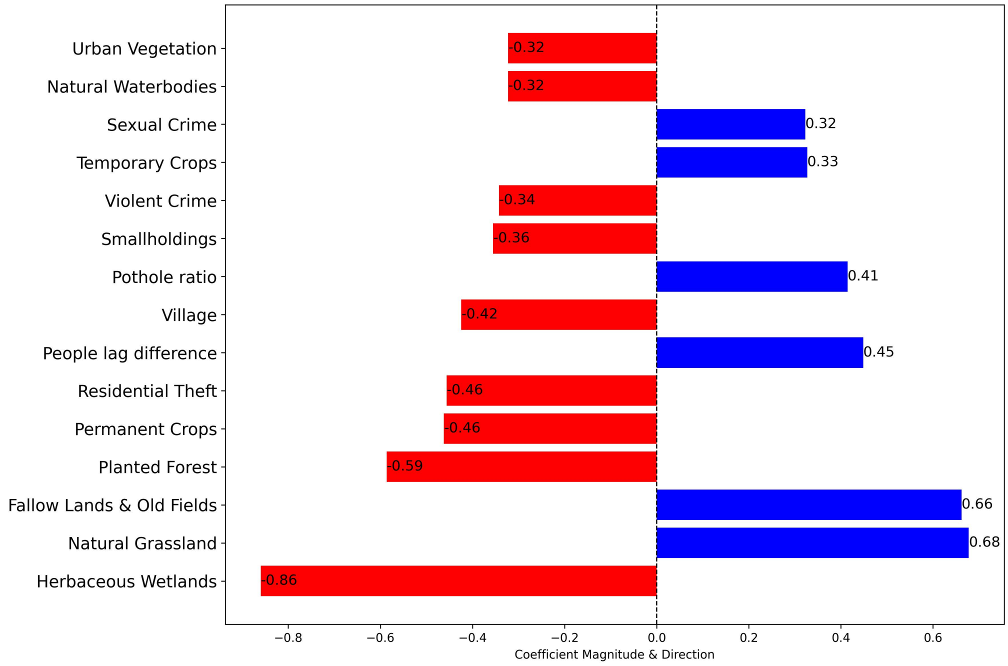


FIGURE 5.6: Feature importance plot for the non vote elastic net regression model

Figure 5.6 reveals a positive association between pothole prevalence and non-vote percentage, suggesting that inadequate infrastructure may contribute to political disengagement. This aligns with rational choice theory, where citizens withdraw participation when they perceive government failure and see no viable alternative worth supporting [13, 7].

Land use features captured via satellite imagery reveal a more nuanced picture of rural disengagement. Areas dominated by fallow land, natural grassland, and smallholdings are positively associated with abstention, whereas villages and agricultural zones, such as permanent and temporary crop areas, are negatively associated. This suggests that not all rural areas behave uniformly - more established or economically active rural areas may foster greater civic engagement, while more isolated or less developed zones face higher barriers to participation. These findings support the notion that in some communities, especially those where the costs of voting, such as transport or time, outweigh the perceived benefits, and as a result citizens may rationally choose not to vote [8, 19].

Finally, the negative correlation between non-voting and residential theft offers further insight. Areas with higher levels of residential crime appear to have greater political engagement, possibly because such crime is more common in wealthier urban areas where civic participation tends to be higher. Conversely, areas with lower residential theft, often poorer, more rural communities, show higher non-voting rates. Together, these results reinforce the broader finding that abstention is more prevalent in structurally marginalised, less-developed regions where the material and psychological costs of voting are highest.

### 5.2.2 ANC Importance Plots

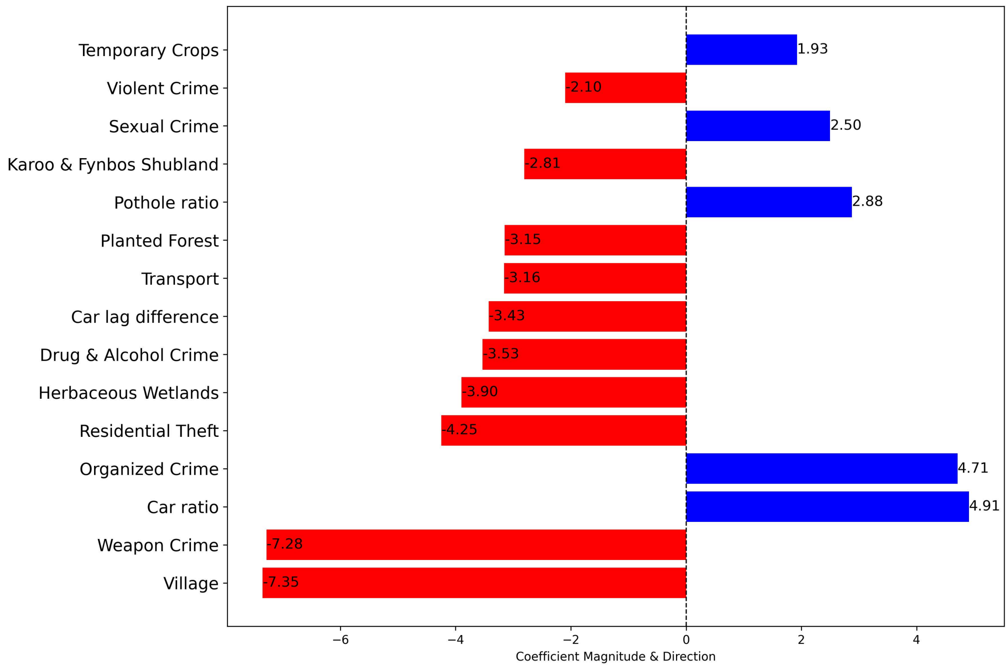


Figure 5.7: Feature importance plot for the ANC elastic net regression model

Figure 5.7 show that a higher pothole ratio is positively associated with ANC vote share, which could reinforce findings at the ward level that suggest the party retains support in areas experiencing visible infrastructure decline. While a higher absolute number of cars in an area seems to correspond with increased ANC support, a higher car lag difference, indicating more cars relative to neighbouring areas, appears to be negatively associated. This may indicate that ANC support is stronger in relatively low-mobility areas, aligning with clientelism and social structure theories that propose lower-income communities with limited transport access might continue to support the ruling party due to historical loyalty, dependence on state support, or a lack of viable alternatives [35, 36].

Crime variables also seem to follow a distinct pattern. Residential theft and violent and weapon-related crime, which are both features of residential theft [82, 104], are negatively associated with ANC support, which may indicate that the party struggles to retain backing in middle-income or urban residential zones characterised by stability and better services. Conversely, the positive associations observed for organised crime and sexual crime might suggest an alternative dynamic in these contexts.

Finally, land use patterns add further nuance. The strong negative correlation between ANC support and villages might challenge the common assumption that rural areas uniformly favour the party. Instead, the output shows that ANC support is positively associated with more ambiguous or undeveloped land-use types, such as planted forests and shrublands, and negatively associated with clearly defined rural settlements like villages. This may indicate potential shifts in the ANC's rural support base that warrant further investigation.

### 5.2.3 DA Importance Plots

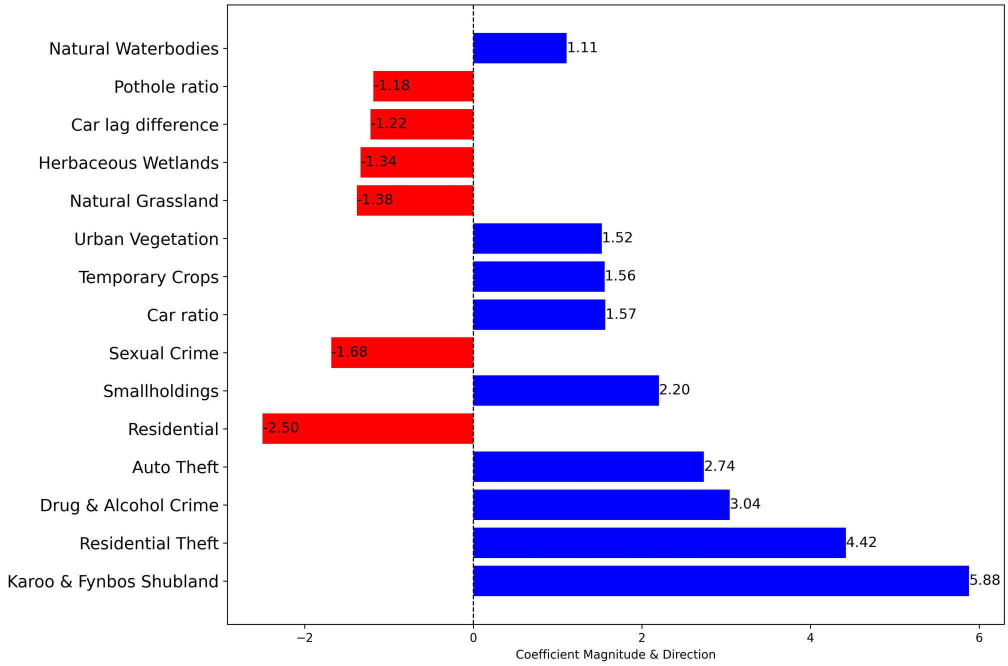


FIGURE 5.8: Feature importance plot for the DA elastic net regression model

The DA's municipal-level regression results, presented in Figure 5.8, reveal several notable patterns that differ from those observed in the ANC model, particularly in relation to infrastructure, land use, and crime.

Starting with infrastructure, the pothole ratio is negatively associated with DA support, suggesting that areas with more deteriorated road conditions are less likely to vote for the DA. This is consistent with the party's stronger performance in urban, well-maintained municipalities, where service delivery tends to be more consistent. These findings may support the notion that the DA's electoral base is concentrated in areas with relatively better infrastructure and access to services.

In terms of land use, the feature Karoo and Fynbos shrubland shows a positive association with DA vote share. This aligns with geographic expectations, as these vegetation types are largely located in the Western Cape, historically the DA's strongest region. This suggests that local geography may play a role in shaping political preferences, potentially through its overlap with administrative boundaries and regional identities.

The crime-related features display trends that appear to contrast with those in the ANC. Both drug and alcohol crime and residential theft are positively associated with DA support. One possible interpretation is that these crimes are more frequently reported or recorded in relatively affluent, urban areas — contexts in which the DA performs well. Conversely, sexual crime is negatively associated with DA support, differing from its positive association in the ANC model. These differing patterns may reflect broader socio-environmental or reporting differences between DA-leaning and ANC-leaning municipalities, though this remains speculative.

Overall, the results suggest that DA support is associated with areas that are more urbanised, have better infrastructure, and may differ in how crime is experienced or recorded.

### 5.2.4 MK Importance Plots

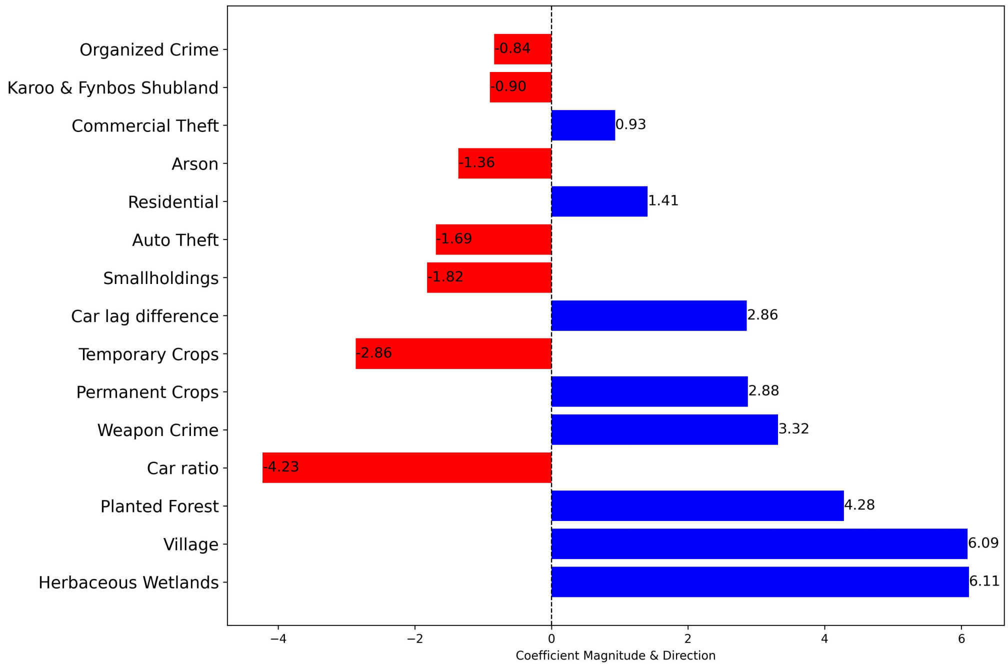


FIGURE 5.9: Feature importance plot for the MK elastic net regression model

The feature importance plot for MK, shown in Figure 5.9, reveals notable distinctions from both the ANC and DA. A negative association between car ratio and MK support suggests that the party tends to draw votes from areas with fewer cars — potentially reflecting lower levels of affluence or urbanisation. This contrasts with the ANC, where car ratio is positively associated with support. Additionally, villages are strongly positively correlated with MK support, in contrast to the ANC's negative association. This pattern may indicate that the ANC's traditional rural base is shifting, from more developed rural settlements like villages toward less developed or more remote areas, while the MK increasingly captures support in these formerly ANC-aligned regions. This apparent redistribution of rural backing between the two parties may again warrant further investigation to better understand the changing dynamics of rural voter behaviour.

Crime-related factors also differ significantly. Unlike the ANC, which is positively associated with organised and sexual crime, MK shows a negative association with both organised crime and sexual crime. This inversion may indicate that MK-leaning areas experience different socio-criminal dynamics compared to ANC strongholds.

Another notable distinction is the nature of theft-related crimes. While DA-supporting areas are positively associated with residential theft, MK-leaning areas are positively correlated with commercial theft. This suggests that theft in MK-supporting regions is more concentrated around shopping and commercial centres rather than residential areas.

### 5.2.5 EFF Importance Plots

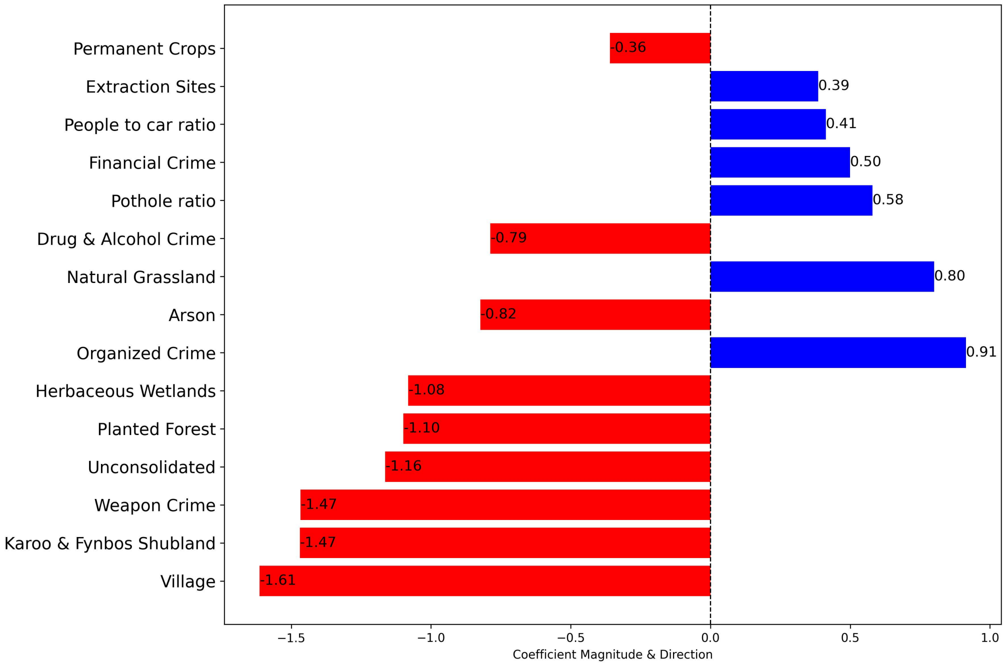


Figure 5.10: Feature importance plot for the EFF elastic net regression model

The feature importance plot for the EFF, shown in Figure 5.10, reveals several patterns that differentiate the party's support base from that of both the ANC and MK. Notably, villages, natural grasslands, and permanent crop areas are all negatively associated with EFF support. This pattern contrasts with the MK, which sees strong positive associations with villages, and instead suggests that the EFF may be drawing from a similar base as the ANC, particularly in less structured rural environments. In this context, rational choice theory may provide a useful lens, as the shift toward opposition parties like the EFF in areas traditionally aligned with the ANC could reflect disillusionment with the ruling party's performance and a desire for political alternatives.

However, this interpretation is complicated by the negative correlation between EFF support and planted forests, an area where the ANC shows positive association. This divergence may suggest that although the two parties share some overlapping rural constituencies, there remain important variations in the characteristics and geography of their respective support bases.

Crime-related features present further distinctions. Financial crime is positively associated with EFF support, a pattern not observed in any of the other models. Given that financial crime in the South African context often includes corruption and bribery [105], this finding may point to unique dynamics. One possible interpretation is that support for the EFF emerges in areas where citizens or local actors perceive opportunities to exert influence over state processes through informal or non-transparent means. In this view, clientelism may take on an alternative form, where instead of state resources being exchanged for loyalty in the traditional top-down sense, citizens anticipate that once the EFF is in power, political sway can be gained through networks of influence or bribery. This interpretation is tentatively supported by the positive association between EFF support and both financial crime and organised crime — categories that, in the South African context, often involve corruption and illicit governance arrangements [106, 105].

Alternatively, the findings may reflect a more conventional rational choice explanation. Citizens experiencing poor service delivery and institutional neglect, signalled by the positive association between EFF support and pothole prevalence, may rationally choose to withdraw support from the ruling party in favour of an opposition alternative. In this case, the EFF could be viewed not as a vehicle for informal power, but as a protest vote or a strategic attempt to pressure the state into more responsive governance.

While speculative, these interpretations underscore the complexity of political realignment in contexts marked by economic marginalisation and weak local governance. They also highlight the need for further research into how different forms of dissatisfaction, whether driven by pragmatic calculations or perceptions of corruption, shape emerging patterns of opposition support in South Africa over time.

# Chapter 6

# Discussion

This study demonstrates the feasibility and value of using computer vision and geospatial analysis to model voting patterns in South Africa. Through the optimisation of parallel programming techniques and location generation algorithms, the system was able to collect a total of 157 338 Google StreetView (GSV) images across both the ward and municipal levels, almost double the required number of images. These images, combined with satellite data, enabled the construction of a rich and spatially granular dataset to explore the intersection of physical environments and electoral outcomes.

From the GSV imagery, computer vision techniques were used to extract features such as potholes, people, vehicles, and instances of illegal dumping. These features served as proxies for local living conditions, offering alternative and often underutilised insights into how infrastructure, mobility, and visible disorder may relate to political behaviour. Satellite-derived land use data complemented these GSV features, providing a broader spatial and ecological context to each administrative unit.

At the municipal level, additional data, including crime and financial metrics, were incorporated. Interestingly, while crime variables emerged as influential in all regression models, financial data did not. These variables were removed during the feature reduction process, potentially due to the challenge of normalising financial figures in a meaningful spatial context. Unlike land use or GSV features, financial metrics could not be scaled as ratios of administrative area, limiting their utility in the modelling process.

The performance of the models across both administrative levels was strong. At the ward level, R-2 values ranged from 0.3658 to 0.8078, while at the municipal level, they spanned from 0.451 to 0.860. Classification performance at the ward level was particularly noteworthy, with F1 scores reaching as high as 0.93. These results offer compelling evidence that living conditions are deeply intertwined with political behaviour — suggesting a triangular relationship between living conditions, voting considerations, and voting outcomes. This lends support to rational choice theory, which posits that voters weigh costs and benefits (voting considerations) about their circumstance (living condiditions) when deciding whether and how to participate in elections (voting patterns).

However, the story is more nuanced than a single theory can capture. For instance, non-voting was positively associated with indicators of poor service delivery, such as pothole density and illegal dumping, which may reflect a sense of disillusionment or protest. In rural areas, abstention may be driven by structural barriers, where the cost of reaching a polling station outweighs the perceived benefit of voting. These findings are consistent with rational choice models, particularly in contexts where government performance is visible, and alternative options appear absent.

Yet, not all cases conform to this logic. For the uMkhonto we Sizwe (MK) party, geographical and identity, based variables — particularly features related to KwaZulu-Natal, far outweighed infrastructural factors like potholes. When geographic location was removed from the MK classification model, its F1 score plummeted from 0.93 to 0.07, highlighting how spatial and social identity determinants may override material considerations. This suggests that psychological theories of voting and social structure determinants, those grounded in place, community, and historical affiliations, play a dominant role in shaping support for emerging parties.

While the analysis in this study is grounded in the 2024 South African general elections, the methodological framework and approach has potential broader applicability.

These findings point to a broader conclusion: no single theoretical framework can fully account for the complexity of South Africa's political landscape. While rational choice theory helps explain participation in contexts of service delivery and infrastructure, clientelism offers insight into the role of state dependence and material exchange. Social structure theories capture the enduring influence of race, geography, and community, while psychological models speak to identity, history, and emotion. Each framework contributes meaningfully, yet none is sufficient on its own.

South African voting behaviour emerges not from a uniform logic, but from overlapping and sometimes contradictory forces. Voters respond to lived realities - some shaped by material deprivation, others by collective memory, regional affiliation, or disillusionment. In this context, electoral choices are not solely strategic calculations or expressions of loyalty, but multifaceted responses to a society in transition. Understanding these patterns requires embracing theoretical plurality and acknowledging that political behaviour is as diverse and dynamic as the country itself.

# Chapter 7

# Conclusion

## 7.1 Introduction

This study set out to examine whether voting patterns in South Africa can be modelled using open-source data, satellite imagery, and Google StreetView (GSV) images. With declining voter turnout and growing disillusionment with traditional politics, the research aimed to find new ways of understanding how people vote, focusing not only on attitudes or demographics, but also on the physical environments people live in.

The main research question asked: "To what extent can voting patterns be modelled using open-source, satellite, and GSV data?" This was supported by sub-questions on what voting patterns are most relevant, how to collect and process imagery, how to extract features using computer vision, and how to connect these features to existing theories of voting behaviour.

The study successfully collected over 15 000 GSV images, along with satellite data, at both the ward and municipal levels. Using computer vision, it extracted features like potholes, people, cars, and illegal dumping, offering new ways to measure living conditions. These were combined with open source data on crime, land use, and financial data to create detailed models of voter behaviour.

The results showed strong predictive performance, with models explaining a large share of voting outcomes and achieving high classification accuracy. More importantly, the findings highlighted how different types of political behaviour are linked to distinct environmental and social conditions. This shows that combining imagery, spatial data, and machine learning offers a powerful new approach for understanding voting in South Africa — and adds a valuable dimension to existing theories of political behaviour.

## 7.2 Literature Review

This study was motivated by a critical gap in the literature: the lack of detailed analysis linking physical living conditions to voting patterns in South Africa. While previous research has focused on factors like race, income, and age, it often relies on high-level census and survey data, which overlook the localised and material realities that shape how people experience democracy. In particular, service delivery and infrastructure, key concerns for many communities, have been underexplored as direct influences on political participation. Yet, research has shown that people's lived environments deeply shape their engagement with the state, and thus, their voting decisions.

To make sense of these relationships, the literature draws on several theoretical frameworks. Rational choice theory suggests that people vote based on whether they believe political participation will improve their circumstances. Clientelism, by contrast, highlights how material benefit can be exchanged for political loyalty, especially in poorer communities. Social structure and psychological theories further explain how identity, memory, and community dynamics shape political behaviour. However, while each theory captures part of the picture, none fully addresses how daily realities, like potholes, illegal dumping, or visible signs of neglect, can influence both voter turnout and party preference.

Recent advances in machine learning and computer vision have begun to close this gap internationally, showing that features visible in the built environment can be strong predictors of political outcomes. Studies in the United States, for example, have used street-level imagery to estimate party support with surprising accuracy. However, this approach remains largely unexplored in the South African context. This study contributes to the literature by using imagery to directly observe living conditions and test their relationship to electoral behaviour — offering a novel lens through which to understand the everyday, material drivers of political engagement.

## 7.3 Findings

This study found a consistent and measurable link between living conditions and voting behaviour at both ward and municipal levels. Across levels, infrastructure indicators such as pothole prevalence and illegal dumping showed strong associations with voter turnout and party support. In particular, areas with poor service delivery tended to have higher non-voting rates, supporting the notion that visible neglect can lead to political disengagement. Meanwhile, features like car ownership, road coverage, and satellite-observed land use offered insight into the socio-economic and spatial dimensions of voter behaviour, distinguishing between urban and rural dynamics and between different party support bases.

However, the findings also revealed that no single theory of voting behaviour can fully explain these patterns. While rational choice theory helps interpret abstention and shifts in support to opposition parties like the Economic Freedom Fighters (EFF), parties such as uMkhonto we Sizwe (MK) show strong geographic and identity-based patterns better captured by Social Structure and Psychological Theories. These results suggest that voting in South Africa is shaped by a complex interplay of lived experience, historical loyalty, and material conditions — requiring a multidimensional approach to fully understand how and why people choose to vote, or not.

## 7.4 Limitations and Future Work

This study faced several limitations, particularly regarding data availability and spatial coverage. One of the most significant challenges was the limited extent of GSV imagery across South Africa. The coverage of GSV is skewed towards towards urban and commercially important areas, resulting in underrepresentation of rural and marginalised regions.

This imbalance introduced potential bias into the analysis, particularly at the ward level where alternative data sources were not available. As a result, voting patterns in less-mapped areas may not be as accurately captured.

Satellite imagery was used consistently across both the ward and municipal levels, and providing broader spatial coverage. However, the features derived from this imagery were sourced from an existing dataset rather than custom-processed imagery. This limited the flexibility to extract more tailored, locally relevant features that could reflect context-specific voting influences in South Africa. With access to raw satellite imagery, future work could design and train models to detect infrastructure and land-use features more directly linked to electoral dynamics from satellite imagery.

Looking forward, an important avenue for future research involves investigating how voting patterns shift over time in response to changes in lived conditions. For instance, the evolution of support between the African National Congress (ANC) and MK, or changing levels of voter abstention, could be analysed in relation to changes in infrastructure or service delivery. This would provide a dynamic perspective on political behaviour beyond single-election snapshots. Additionally, more features could be extracted from street-level imagery, such as the condition of sidewalks, walls, signage, or vehicle quality, offering deeper insights into "environmental cues" of inequality or governance. Furthermore, combining such image-based features with demographic or survey data would enable more robust, multi-dimensional models of voter behaviour and engagement.

Finally, the relationship between the features used to generate the models could be investigated. This will give a better understanding of the relationships between the target variables, namely the ANC, Democratic Alliance (DA), MK, and NonVote. It is also important to note that the classification tasks excluded the EFF and other smaller parties due to low representation. While this simplification helped to reduce model complexity, it may understate political diversity and further imbalance the dataset, thereby limiting generalisability and masking meaningful variation in voter preferences outside the dominant parties. Future research could therefore consider alternative approaches such as SMOTE, class balancing, or hierarchical classification to better capture the full spectrum of electoral behaviour.

## 7.5 Summary of Discussion and Findings

The models developed in this study performed well, demonstrating that physical living conditions are meaningfully linked to both voter turnout and party support. Across ward and municipal levels, infrastructure indicators such as pothole density and illegal dumping consistently correlated with patterns of political engagement. High model accuracy and strong performance metrics, particularly an F1 score of 0.93 in some classification tasks, offer compelling evidence that environmental conditions are deeply entwined with electoral behaviour. These results reinforce the idea that voting is not just a matter of ideology or identity but is shaped by the realities people see and experience in their everyday surroundings.

However, the study also makes clear that no single theory, whether rational choice, clientelism, social structure, or psychological, can fully explain South African voting patterns. While material conditions influence voter engagement, identity-based and geographic factors often override them, especially for parties like the MK. This suggests that voting behaviour in South Africa is best understood through a combination of theories, each offering partial insights into a broader, more complex picture. In this light, the vote becomes not just a strategic decision or inherited allegiance, but a layered response to history, place, and lived experience.

# Bibliography

[1] A. Kiani and H. Sartipi, "Functions of election in a democratic system," *International Research Journal of Interdisciplinary & Multidisciplinary Studies (IRJIMS)*, vol. II, no. IX, pp. 20–35, 2016.

[2] W. Wojtasik, "Functions of elections in democratic systems," *Political Preferences*, vol. 4, pp. 25–38, 2016.

[3] Electoral Commission of South Africa, "NPE Results Dashboard 2024." `https://results.elections.org.za/dashboards/npe/`, 2024. [Accessed: 26-Feb-2025].

[4] C. Schulz Herzenberg, "The 2021 municipal elections: Rise of the volatile voter or disaffected citizen?," *Politikon*, vol. 49, no. 4, pp. 318–336, 2022.

[5] K. Nweke and E. Etido-Inyang, "National elections and political apathy: A comparative study of voter turnout in the 2019 general elections of Nigeria and South Africa," *Advances in Social Sciences Research Journal*, vol. 5, no. 12, pp. 643–654, 2018.

[6] W. Kulachai, U. Lerdtomornsakul, and P. Homyamyen, "Factors influencing voting decision: A comprehensive literature review," *Social Sciences*, vol. 12, no. 9, p. 469, 2023.

[7] M. Chandler, "Models of voting behavior in survey research," *Synthese*, vol. 76, pp. 25–48, 1988.

[8] D. de Kadt and E. S. Lieberman, "Nuanced accountability: Voter responses to service delivery in southern africa," *British Journal of Political Science*, pp. 1–31, 2017.

[9] T. Matsubayashi and S. Sakaiya, "Income inequality and income bias in voter turnout," *European Journal of Political Economy*, vol. 66, p. 101966, 2021.

[10] A. De-Juan and E. Wegner, "Social inequality, state-centered grievances, and protest: Evidence from South Africa," *Sage*, 2017.

[11] V. Graham, "South africa's democracy: The quality of political participation over 25 years," *Journal of African Elections*, vol. 19, no. 1, pp. 28–51, 2020.

[12] L. Lues, "Citizen participation as a contributor to sustainable democracy in south africa," *International Review of Administrative Sciences*, vol. 80, no. 4, pp. 789–807, 2014.

[13] S. Booysen, "Power through the ballot and the brick," in *The African National Congress and the Regeneration of Political Power* (S. Booysen, ed.), pp. 136–164, Johannesburg: Wits University Press, 2011.

[14] D. Berliner and J. Wehner, "Audits for accountability: Evidence from municipal by-elections in south africa," *The Journal of Politics*, vol. 84, no. 3, pp. 1581–1594, 2022.

[15] C. Schulz-Herzenberg, "Elections and accountability in south africa," Tech. Rep. 188, Institute for Security Studies (ISS), 2009.

[16] C. Schulz Herzenberg, "The South African non-voter: An analysis," *The Midpoint Paper Series*, no. 2, 2020.

[17] L. Tracey, *Do you want my vote? Understanding the factors that influence voting among young South Africans.* No. 193 in ISS Monograph, Institute for Security Studies, 2016.

[18] C. Schulz-Herzenberg, "The influence of the social context on south african voters," *Journal of Southern African Studies*, vol. 40, no. 4, pp. 839–859, 2014.

[19] M. Bekker, C. Runciman, and B. Roberts, "Beyond the binary: Examining dynamic youth voter behaviour in south africa," *Politikon*, 2022.

[20] T. Dooms, "Reclaiming people's power through electoral reforms," *Policy Insights*, no. 158, 2024.

[21] P. Pillay, "Public trust and good governance: A comparative study of brazil and south africa," *African Journal of Public Affairs*, vol. 9, pp. 31–47, September 2017.

[22] A. Morar, M. Venter, and T. Chuchu, "To vote or not to vote: marketing factors influencing the voting intention of university students in johannesburg," *Journal of Economics and Behavioral Studies*, vol. 7, no. 6, pp. 81–93, 2015.

[23] T. Fransman, "Voting and protest tendencies associated with changes in service delivery," Working Paper WP08/2021, Department of Economics, University of Stellenbosch, 2021.

[24] T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. Aiden, and L. Fei-Fei, "Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the United States," *Proceedings of the National Academy of Sciences*, vol. 114, no. 50, pp. 13108–13113, 2017.

[25] B. Mukiibi and P. Obiezu, "Putting africa on the map," *Nature Africa*, 2023. Accessed: 23 March 2025.

[26] M. D. M. Bader, S. J. Mooney, B. Bennett, and A. G. Rundle, "The promise, practicalities, and perils of virtually auditing neighborhoods using google street view," *The Annals of the American Academy of Political and Social Science*, vol. 669, pp. 18–40, 2017.

[27] Google, "How Google Street View Works," 2024. Accessed: 23 March 2025.

[28] F. Biljecki and K. Ito, "Street view imagery in urban analytics and gis: A review," *Landscape and Urban Planning*, vol. 215, p. 104217, 2021.

[29] M. J. Power, P. Neville, E. Devereux, A. Haynes, and C. Barnes, "'why bother seeing the world for real?': Google street view and the representation of a stigmatised neighbourhood," *New Media & Society*, vol. 15, no. 7, pp. 1022–1040, 2013.

[30] F. Umar, A. Elsey, D. Chamberlain, and A. Nelson, "On the potential of google street view for environmental waste quantification in urban africa: An assessment of bias in spatial coverage," *Sustainable Environment*, vol. 9, no. 1, 2023.

[31] W. H. Riker, "The political psychology of rational choice theory," *Political Psychology*, vol. 16, no. 1, pp. 23–44, 1995.

[32] Y. Sadie and L. Patel, "Zuma versus Ramaphosa: Factors influencing party choice of South Africans in the run-up to the 2019 elections," *Journal of African Elections*, pp. 1–27, 06 2020.

[33] T. J. Feddersen, "Rational choice theory and the paradox of not voting," *Journal of Economic Perspectives*, vol. 18, no. 1, pp. 99–112, 2004.

[34] N. Van de Walle, "Presidentialism and clientelism in Africa's emerging party systems," *The Journal of Modern African Studies*, vol. 41, no. 2, pp. 297–321, 2003.

[35] M. Pellicer, E. Wegner, L. Benstead, and E. Lust, "Poor people's beliefs and the dynamics of clientelism," *Journal of Theoretical Politics*, vol. 33, no. 3, pp. 300–332, 2021.

[36] A. Hicken, "Clientelism," *Annual Review of Political Science*, vol. 14, pp. 289–310, June 2011.

[37] P. Jöst and E. Lust, "Receiving more, expecting less? social ties, clientelism and the poor's expectations of future service provision," *World Development*, vol. 158, p. 106008, 2022.

[38] M. J. Masenya, "Neo-patrimonialism, corruption and governance in south africa," *African Journal of Public Affairs*, vol. 9, no. 9, pp. 146–156, 2017.

[39] M. Paret and C. Runciman, "Voting decisions and racialized fluidity in South Africa's metropolitan municipalities," *African Affairs*, vol. 122, no. 487, pp. 269–298, 2023.

[40] "IEC Women's Dialogue on Gender Parity & Gender-Based Political Violence." Event.

[41] Y. Sadie, L. Patel, and K. Baldry, "A comparative case study of the voting behaviour of poor people in three selected south african communities," *Journal of African Elections*, vol. 15, no. 1, 2016. Published Online: 1 June 2016.

[42] A. Campbell, P. E. Converse, W. E. Miller, and D. E. Stokes, *The American Voter*. New York: John Wiley & Sons, 1960.

[43] A. M. Mbete, C. O. Mongale, and V. Ojakorotu, "Social discontent and voting behaviour in south africa's 2024 general elections: The transition from one-party dominance to a multi-party coalition," *E-Journal of Humanities, Arts and Social Sciences*, vol. 5, no. 13, pp. 2413–2425, 2024.

[44] L. Patel, Y. Sadie, and J. de Klerk, "Factors determining voter choice in south africa's 2024 national general elections," research brief, Centre for Social Development in Africa (CSDA), University of Johannesburg, Johannesburg, South Africa, May 2024. Supported by the Department of Science and Technology (DST), National Research Foundation (NRF), and CSDA, UJ.

[45] R. Davis, "Mk party trades on identity politics as zuma invokes zulu nationalism, traditional leadership," 2024. Accessed: 26 March 2025.

[46] M. Braun, "South african elections: Research explores how disillusioned anc supporters might use their vote," May 2024. Accessed: 22 March 2025.

[47] A. Spanos, "Where do statistical models come from? revisiting the problem of specification," in *IMS Lecture Notes–Monograph Series: 2nd Lehmann Symposium–Optimality*, vol. 49, pp. 98–119, Institute of Mathematical Statistics, 2006.

[48] IBM, "Lasso Regression - IBM Think." `https://www.ibm.com/think/topics/lasso-regression`, 2025. [Accessed: 3-Mar-2025].

[49] IBM, "Ridge Regression - IBM Think." `https://www.ibm.com/think/topics/ridge-regression`, 2025. [Accessed: 3-Mar-2025].

[50] Corporate Finance Institute (CFI), "Elastic Net - Understanding Hybrid Regularization." `https://corporatefinanceinstitute.com/resources/data-science/elastic-net`, 2025. [Accessed: 3-Mar-2025].

[51] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: An overview,” in *Journal of Physics: Conference Series*, vol. 1142, p. 012012, IOP Publishing, 2018.

[52] L. Alkhawaldeh, “Support vector machines (svm).” ResearchGate, 2023.

[53] L. Breiman, “Random forests,” tech. rep., Statistics Department, University of California, Berkeley, 2001. Technical Report.

[54] I. H. Sarker, “Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions,” *SN Computer Science*, vol. 2, no. 420, 2021.

[55] R. Fairstein, A. Lauz, R. Meir, and K. Gal, “Modeling people’s voting behavior with poll information,” in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, (Montreal, Canada), pp. 1–9, International Foundation for Autonomous Agents and MultiAgent Systems (IFAAMAS), 2019.

[56] T. Phan, G. Navas, C. Ignace, C. Kribs, C. Castillo-Garsow, and V. Moreno, “Modeling voting dynamics in a two-party system: person–to–person interactions and media effect,” *Preprint*, 2016. Available via the Mathematical and Theoretical Biology Institute (MTBI), Arizona State University.

[57] R. Fairstein, A. Lauz, K. Gal, and R. Meir, “Predicting strategic voting behavior with poll information,” *arXiv preprint arXiv:1805.07606*, 2024. Presented as part of research on heuristic and rational models of voting behavior under uncertainty.

[58] T. Kyriazos and M. Poga, “Application of machine learning models in social sciences: Managing nonlinear relationships,” *Encyclopedia*, vol. 4, pp. 1790–1805, 2024.

[59] M. Bahrami, Y. Findik, B. Bozkaya, and S. Balcisoy, “Twitter reveals: Using twitter analytics to predict public protests,” *Preprint*, 2021. Available upon request or from author.

[60] H. Papageorgiou, J. Baggott, and M. Bekker, “Protest events have a ‘twitter signature’: Evidence from south africa’s feesmustfall,” *SA Crime Quarterly*, vol. 73, pp. 2–24, 2024.

[61] S. Ghosh and D. A. Linzer, "Measuring swing voters with a supervised machine learning ensemble," *Political Analysis*, vol. 29, no. 3, pp. 318–330, 2021.

[62] R. D. Pollard, S. M. Pollard, and S. E. Streit, "Predicting propensity to vote with machine learning," 2020. Unpublished manuscript.

[63] A. M. Pineda, P. Kent, C. Connaughton, and F. A. Rodrigues, "Machine learning-based prediction of q-voter model in complex networks," *arXiv preprint arXiv:2310.09131*, 2023.

[64] S. Mancilla, G. Wences, E. Hernández-López, and I. Cohen, "Sub-spatial prediction of votes integrating socioeconomic, educational, and age strata with machine learning and topological data analysis," *Journal of Big Data*, vol. 12, pp. 1–28, 2025.

[65] M. Ledwaba and V. Marivate, "Semi-supervised learning approaches for predicting south african political sentiment for local government elections," *Manuscript submitted to ACM*, 2022. Available at: `https://arxiv.org/abs/2205.02223`.

[66] S. A. Papert, "The summer vision project," Technical Report AIM-100, Massachusetts Institute of Technology, Artificial Intelligence Laboratory, July 1966.

[67] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, pp. 1097–1105, Curran Associates Inc., 2012.

[68] Ultralytics, "YOLOv8 - Real-Time Object Detection and Segmentation." `https://yolov8.com/`, 2025. [Accessed: 3-Mar-2025].

[69] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[70] Q. C. Nguyen, J. M. Keralis, P. Dwivedi, A. E. Ng, M. Javanmardi, S. Khanna, Y. Huang, K. D. Brunisholz, A. Kumar, and T. Tasdizen, "Leveraging 31 million google street view images to characterize built environments and examine county health outcomes," *Public Health Reports*, vol. 136, no. 2, pp. 201–211, 2021.

[71] J. L. Abitbol and M. Karsai, "Socioeconomic correlations of urban patterns inferred from aerial images: interpreting activation maps of convolutional neural networks," *arXiv preprint arXiv:2004.04907*, 2020.

[72] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke, "Using publicly available satellite imagery and deep learning to understand economic well-being in africa," *Nature Communications*, vol. 11, no. 1, p. 2583, 2020.

[73] Google Developers, "Google Maps Street View API Overview." `https://developers.google.com/maps/documentation/streetview/overview`, 2025. [Accessed: 28-Feb-2025].

[74] OpenStreetMap Contributors, "OpenStreetMap." `https://www.openstreetmap.org/#map=6/-28.68/24.68`, 2025. [Accessed: 28-Feb-2025].

[75] GeoPandas Developers, "Geometric Manipulations — GeoPandas Documentation." `https://geopandas.org/en/stable/docs/user_guide/geometric_manipulations.html`, 2025. [Accessed: 3-Mar-2025].

[76] Python Software Foundation, "concurrent.futures — Launching Parallel Tasks." `https://docs.python.org/3/library/concurrent.futures.html`, 2025. [Accessed: 28-Feb-2025].

[77] KEP UPNVJ, "Common Mistake Series: Sample Size Estimation." `https://kep.upnvj.ac.id/home/post/common-mistake-series-sample-size-estimation`, 2025. [Accessed: 28-Feb-2025].

[78] A. Althubaiti, "Sample size determination: A practical guide for health researchers," *Journal of General and Family Medicine*, vol. 24, no. 2, pp. 72–78, 2022.

[79] South Africa Info, "The nine provinces of south africa." `https://southafrica-info.com/land/nine-provinces-south-africa/`. Accessed: 26 March 2025.

[80] S. K. Ahmed, "How to choose a sampling technique and determine sample size for research: A simplified guide for researchers," *Oral Oncology Reports*, vol. 12, 2024. [Accessed: 3-Mar-2025].

[81] Cuemath, "Sample Size Formula - Definition, Examples, and Uses." https://www.cuemath.com/sample-size-formula/, 2025. [Accessed: 3-Mar-2025].

[82] Statistics South Africa, "Crime in south africa up in 2022/23," 2024. Accessed: 23 March 2025.

[83] Esri ArcGIS, "ArcGIS Online Item: Spatial Data Resource." https://www.arcgis.com/home/item.html?id=6394e587066e4d358f9f95707bbc02fa, 2025. [Accessed: 3-Mar-2025].

[84] National Treasury, South Africa, "Municipal Money - South African Municipal Data Portal." https://municipaldata.treasury.gov.za/, 2025. [Accessed: 3-Mar-2025].

[85] Planet Labs PBC, "Planet Science - Advancing Earth Data for Research." https://www.planet.com/science/, 2025. [Accessed: 3-Mar-2025].

[86] Esri ArcGIS, "Albers Equal-Area Conic Projection - ArcGIS Documentation." https://desktop.arcgis.com/en/arcmap/latest/map/projections/albers.htm, 2025. [Accessed: 3-Mar-2025].

[87] Department of Forestry, Fisheries and the Environment (DFFE), South Africa, "South African National Land Cover Datasets." https://egis.environment.gov.za/sa_national_land_cover_datasets, 2025. [Accessed: 3-Mar-2025].

[88] QGIS Development Team, "QGIS - A Free and Open Source Geographic Information System." https://qgis.org/, 2025. [Accessed: 3-Mar-2025].

[89] T. Bhonde, H. Temare, and Y. S. Dadwhal, "Enhanced object detection using yolov8: Identifying vehicles and pedestrians in urban environments," in *2024 IEEE Pune Section International Conference (PuneCon)*, (Pune, India), pp. 1–4, IEEE, 2024.

[90] Encord, "Object Detection: Algorithms, Applications, and Future Trends." https://encord.com/blog/object-detection/, 2025. [Accessed: 3-Mar-2025].

[91] Roboflow, "Roboflow Universe - Discover and Share Computer Vision Datasets." https://universe.roboflow.com/, 2025. [Accessed: 3-Mar-2025].

[92] OpenCV and Intel, "CVAT - Computer Vision Annotation Tool." `https://www.cvat.ai/`, 2025. [Accessed: 3-Mar-2025].

[93] Ultralytics, "YOLOv5 Training Tips - Best Practices for Optimal Results." `https://docs.ultralytics.com/yolov5/tutorials/tips_for_best_training_results/`, 2025. [Accessed: 3-Mar-2025].

[94] R. V. Niyobuhungiro and C. Schenck, "Exploring community perceptions of illegal dumping in fisantekraal using participatory action research," *South African Journal of Science*, vol. 118, no. 5/6, pp. 1–8, 2022.

[95] H. Smolic, "How much data do you need for machine learning," 2024. [Accessed: 16-Mar-2025].

[96] Scikit-Learn Developers, "sklearn.neighbors.BallTree — Scikit-Learn Documentation." `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.BallTree.html`, 2025. [Accessed: 3-Mar-2025].

[97] Jason Brownlee, "Skewness Be Gone: Transformative Tricks for Data Scientists." `https://machinelearningmastery.com/skewness-be-gone-transformative-tricks-for-data-scientists/`, 2025. [Accessed: 3-Mar-2025].

[98] scikit-learn developers, *sklearn.preprocessing.StandardScaler — scikit-learn documentation*, 2024. Accessed: 26 March 2025.

[99] S.-L. Developers, "sklearn.linear_model.linearregression — scikit-learn documentation." `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html`, 2025. [Accessed: 3-Mar-2025].

[100] S.-L. Developers, "sklearn.ensemble.randomforestclassifier — scikit-learn documentation." `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`, 2025. [Accessed: 3-Mar-2025].

[101] S. Chowdhury, Y. Lin, B. Liaw, and L. Kerby, "Evaluation of tree based regression over multiple linear regression for non-normally distributed data in battery performance," *arXiv*, 2021.

[102] G. Leckie, "Multilevel models for continuous outcomes," *arXiv*, 2019.

[103] Investopedia, "Variance inflation factor (vif)," 2024. Accessed: 26 March 2025.

[104] Statistics South Africa, "Crime statistics continue to raise concern," 2021. Accessed: 23 March 2025.

[105] K. K. Arthur, S. A. Asongu, P. Darko, M. O. Ansah, S. Adom, and O. Hlortu, "Financial crimes in africa and economic growth: Implications for achieving sustainable development goals (sdgs)," Working Paper 07-24, EDWRG Working Paper Series, 2024. University of Energy and Natural Resources; Peadato Group; University of Johannesburg.

[106] BBC News, "South africa opposition in turmoil as malema's deputy jumps ship," 2024. Accessed: 23 March 2025.

[107] J. Pardede and A. S. Purohita, "The advantage of transfer learning with pre-trained model in cnn towards ct-scan classification," *Khazanah Informatika*, vol. 9, no. 2, pp. 155–161, 2023. [Accessed: 3-Mar-2025].

[108] IBM, "IBM API - Understanding and Utilizing APIs." `https://www.ibm.com/think/topics/api`, 2025. [Accessed: 3-Mar-2025].

[109] Cloudflare, "Hypertext Transfer Protocol (HTTP) - Cloudflare Learning Center." `https://www.cloudflare.com/learning/ddos/glossary/hypertext-transfer-protocol-http/`, 2025. [Accessed: 3-Mar-2025].

[110] I. S. Kim, S. Kalyanaraman, M. Raghavan, S. Sacco, and Y. Velez, "Deep learning for social science: Applications and opportunities," *Patterns*, vol. 4, no. 12, p. 100878, 2023.

[111] K. F. Mathibedi and M. M. Tshwarelo, "Deterioration of anc support: Conceptual analysis of rural-urban voters patterns in south african elections," *International Journal of Social Science Research and Review*, vol. 6, no. 8, pp. 207–219, 2023.

# Note on appendix structuring

The appendix is structured to mirror the chapter layout of the dissertation, with each appendix section corresponding numerically to its associated chapter. This structure allows readers to easily locate supplementary material.

# Appendix 3.8

# Computer Vision Model Performance

## 3.8.1 Potholes

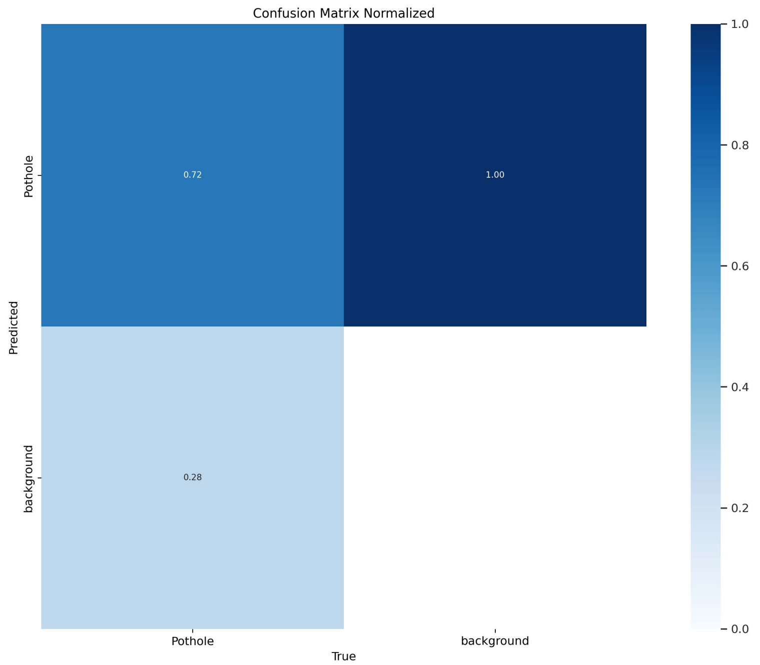


FIGURE 3.8.1: Normalised confusion matrix of final pothole object detection on test set

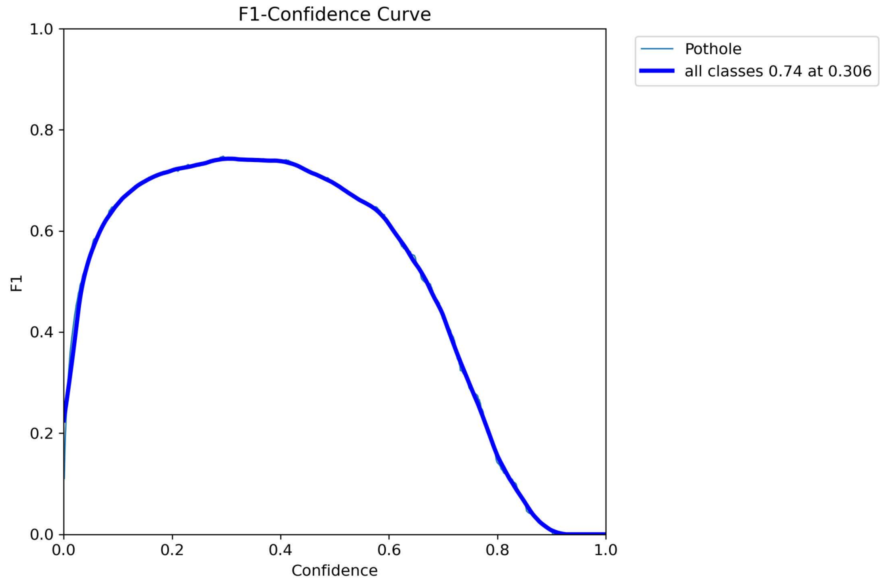


FIGURE 3.8.2: F1 curve of pothole object detection model

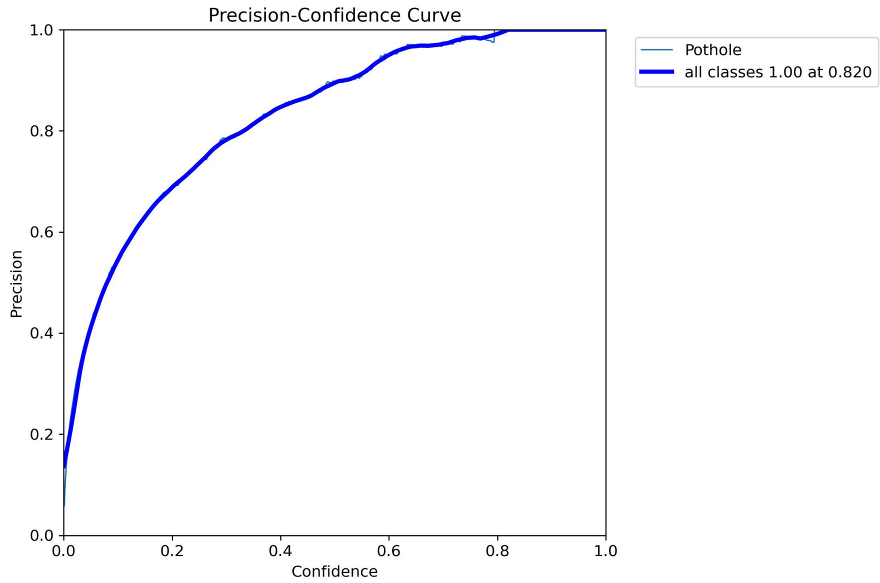


FIGURE 3.8.3: Precision-confidence curve of final pothole detection model

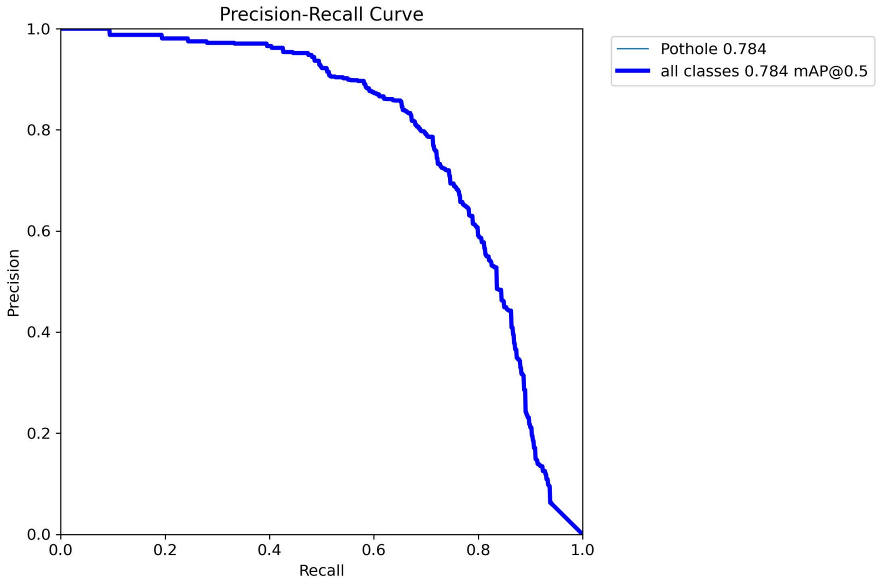


FIGURE 3.8.4: Precision-recall curve of final pothole detection model

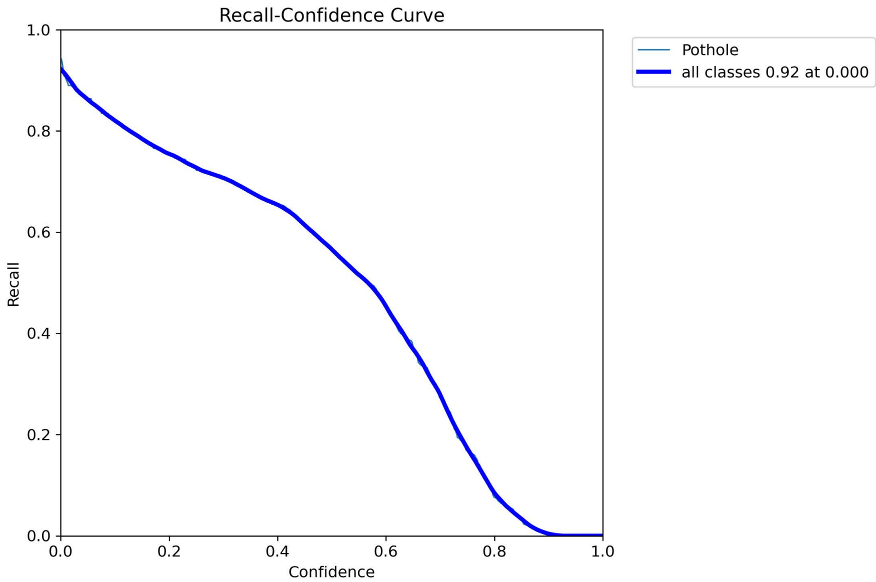


FIGURE 3.8.5: Recall-confidence curve of final pothole detection model

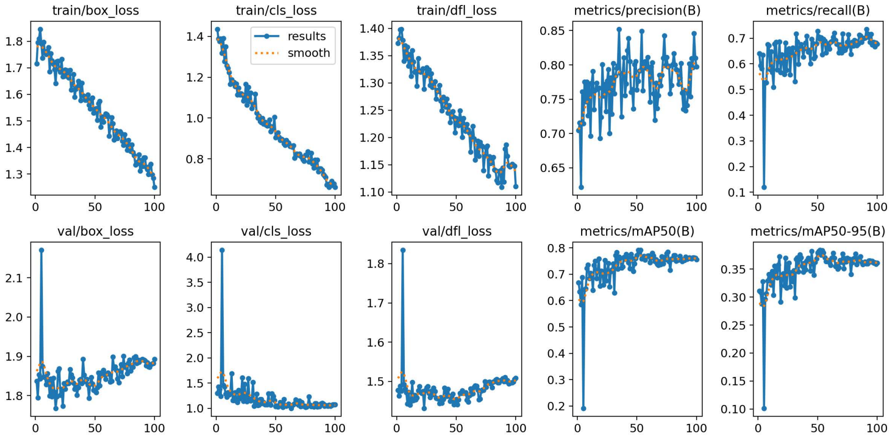


FIGURE 3.8.6: Training results of final pothole detection model

## 3.8.2 People

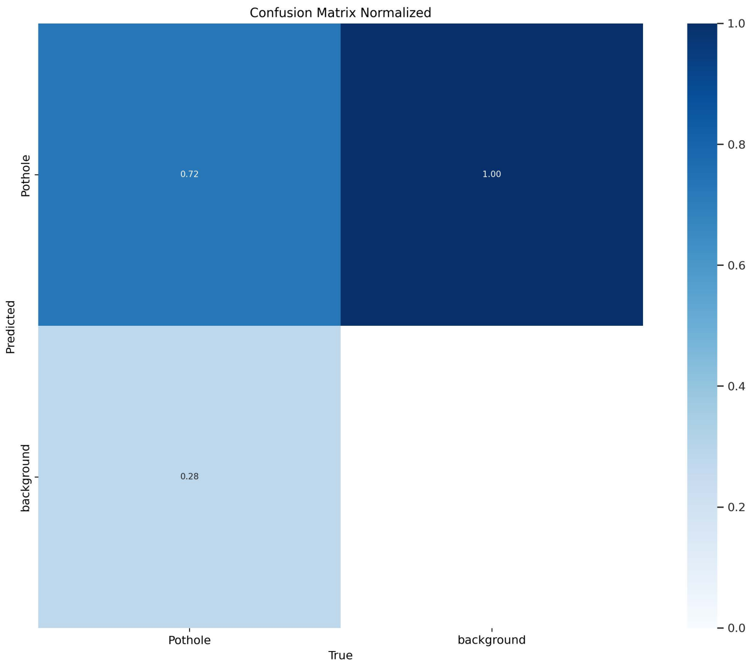


FIGURE 3.8.7: Normalised confusion matrix of final pothole object detection on test set

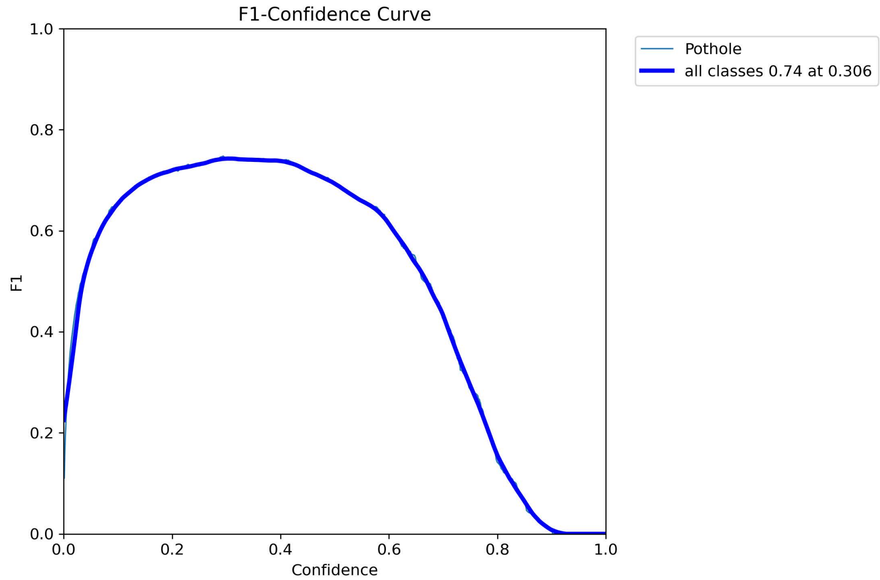


FIGURE 3.8.8: F1 curve of pothole object detection model

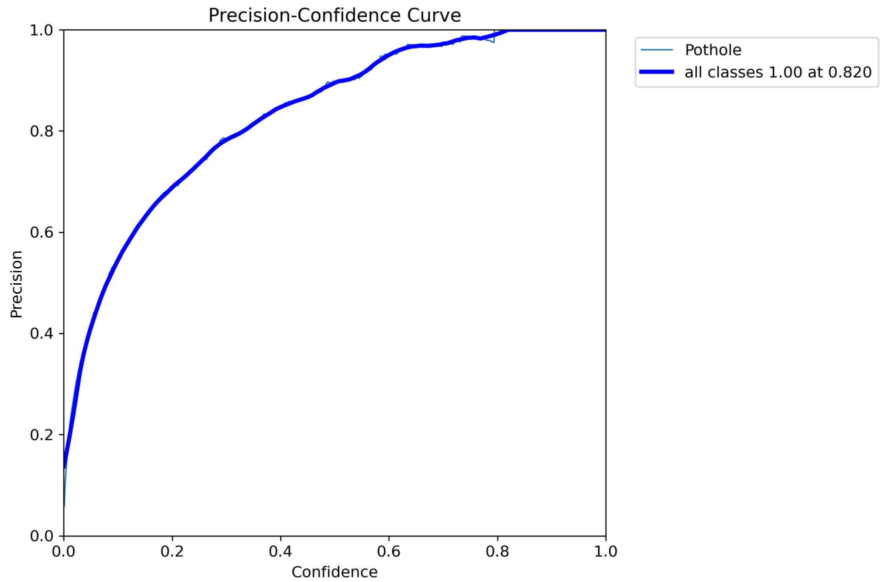


FIGURE 3.8.9: Precision-confidence curve of final pothole detection model

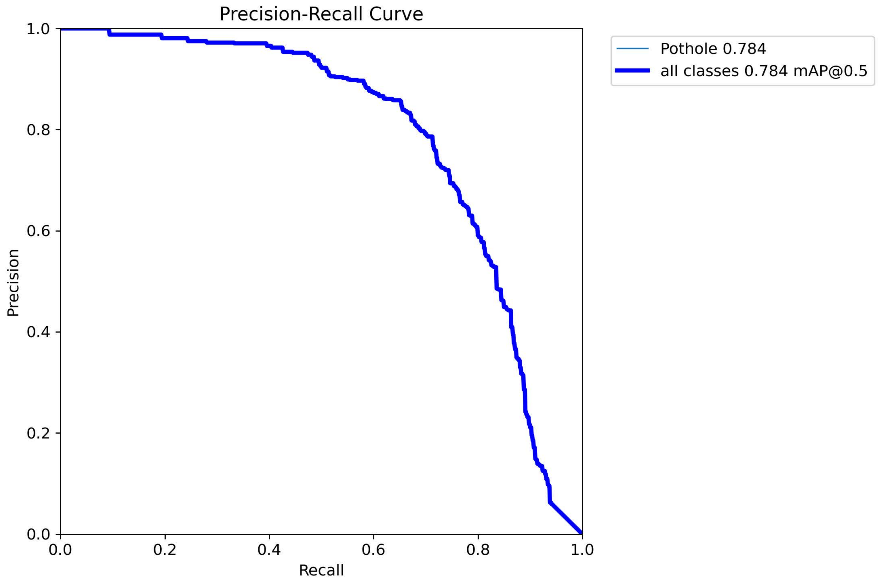


FIGURE 3.8.10: Precision-recall curve of final pothole detection model

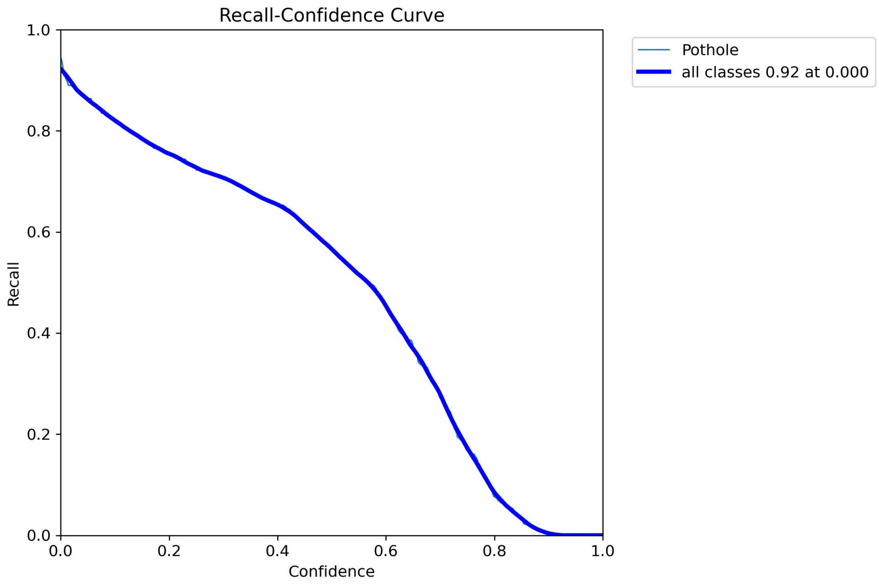


FIGURE 3.8.11: Recall-confidence curve of final pothole detection model

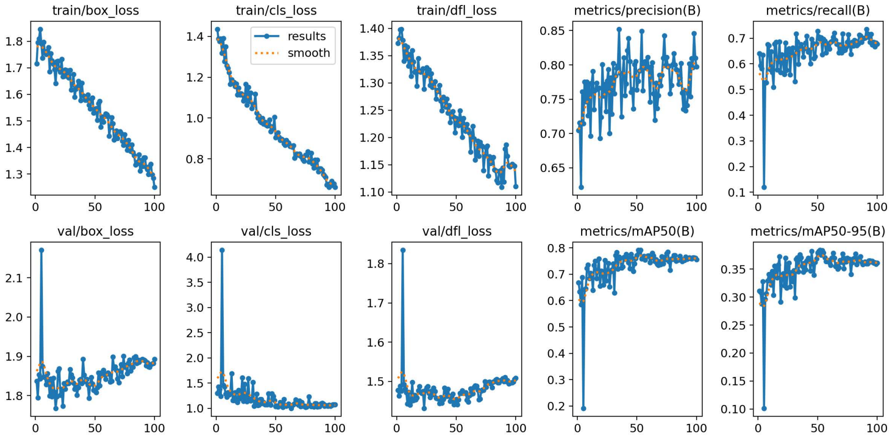


FIGURE 3.8.12: Training results of final pothole detection model

### 3.8.3 Vehicles

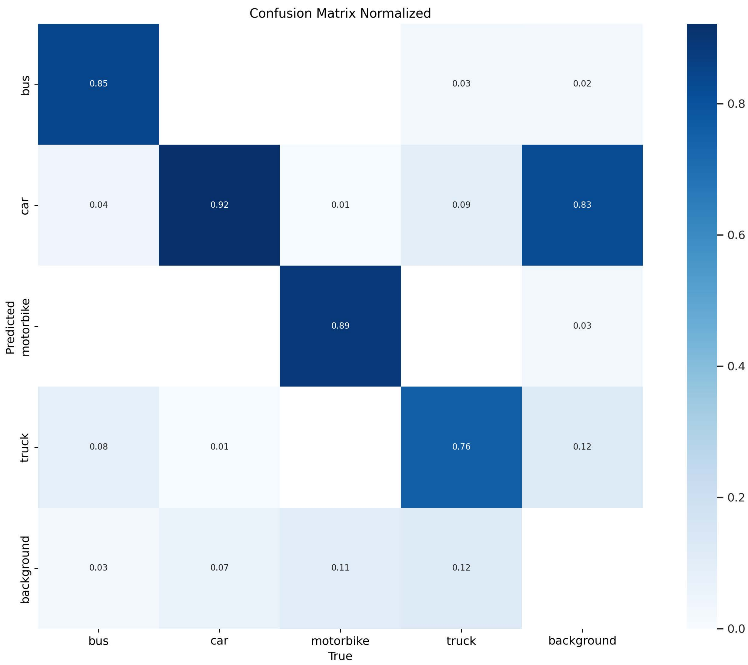


FIGURE 3.8.13: Normalised confusion matrix of final car object detection on test set

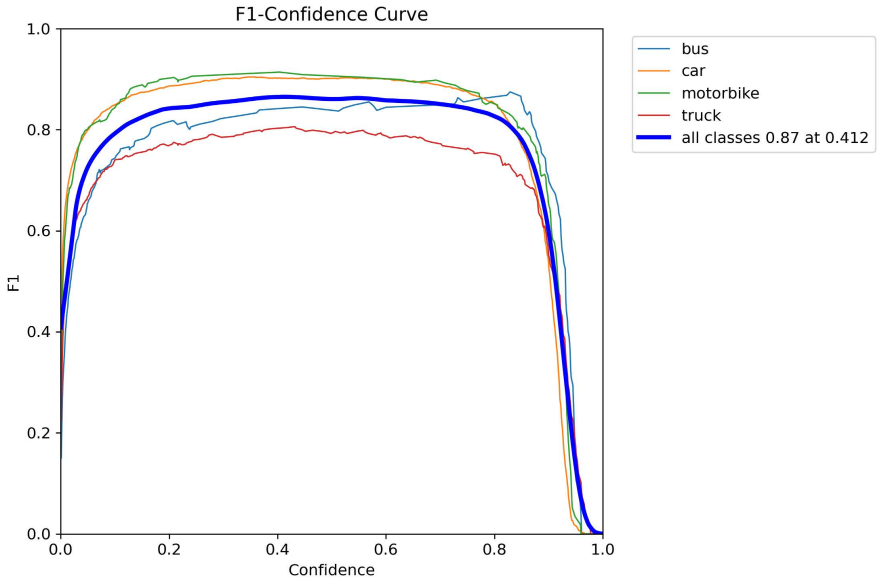


FIGURE 3.8.14: F1 curve of car object detection model

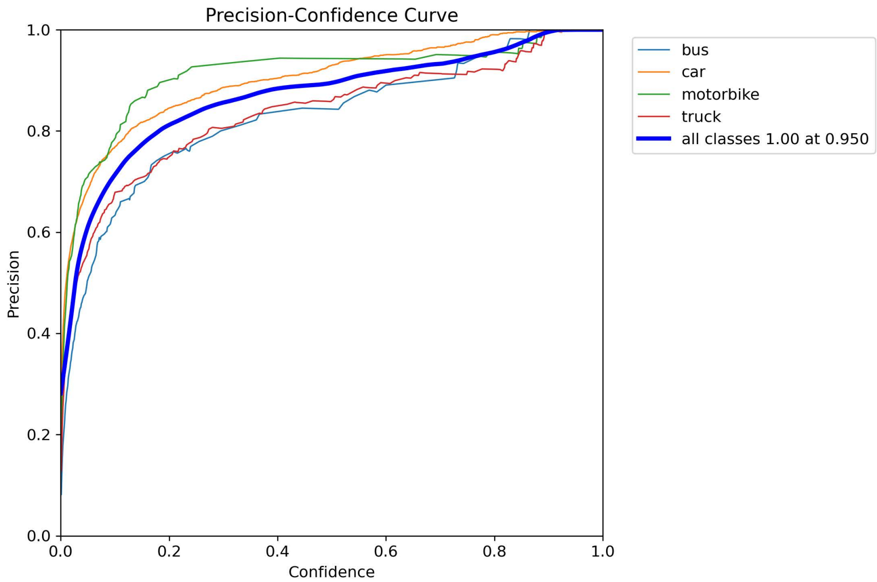


FIGURE 3.8.15: Precision-confidence curve of final car detection model

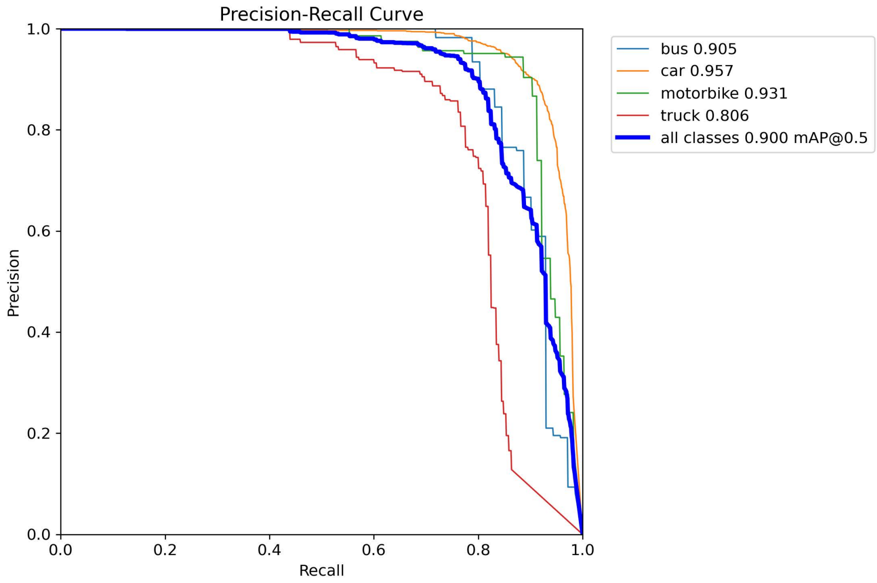


FIGURE 3.8.16: Precision-recall curve of final car detection model

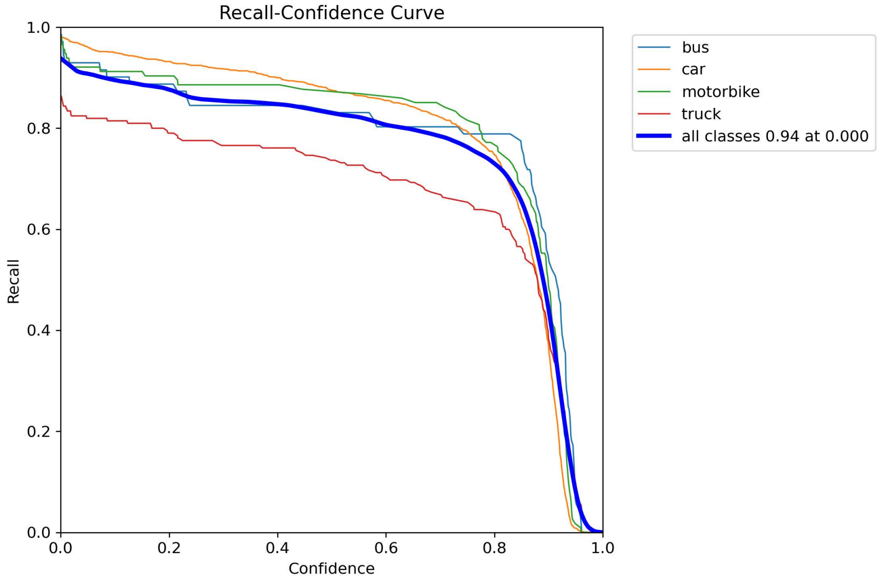


FIGURE 3.8.17: Recall-confidence curve of final car detection model

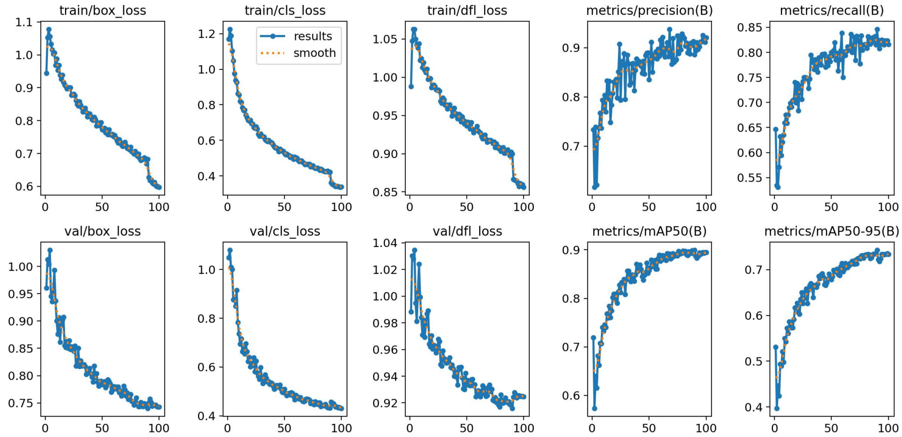


FIGURE 3.8.18: Training results of final car detection model

### 3.8.5 Illegal Dumping

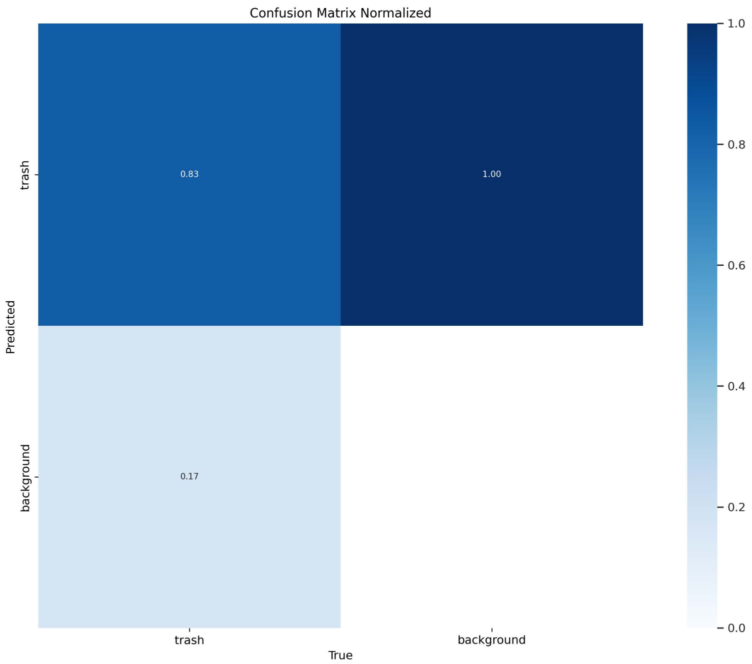


FIGURE 3.8.19: Normalised confusion matrix of final illegal dumping object detection on test set

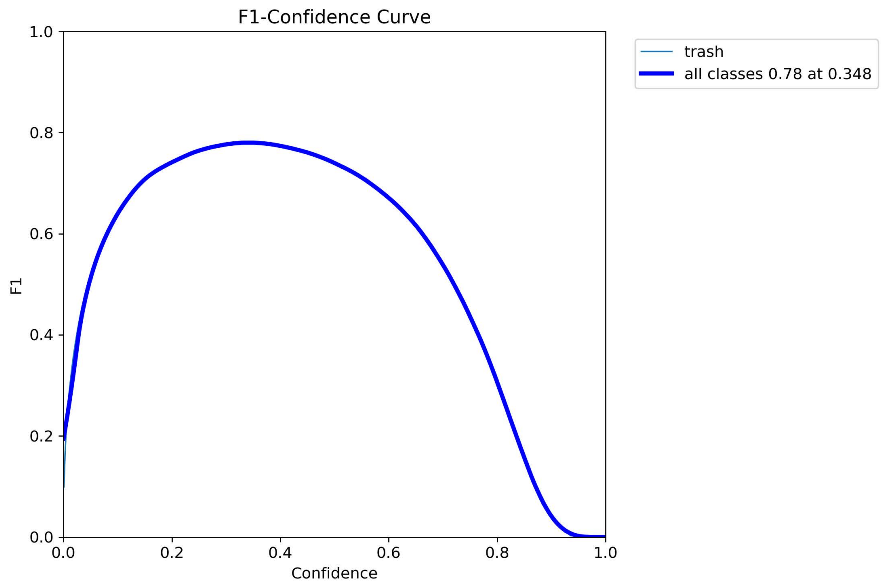


FIGURE 3.8.20: F1 curve of illegal dumping object detection model

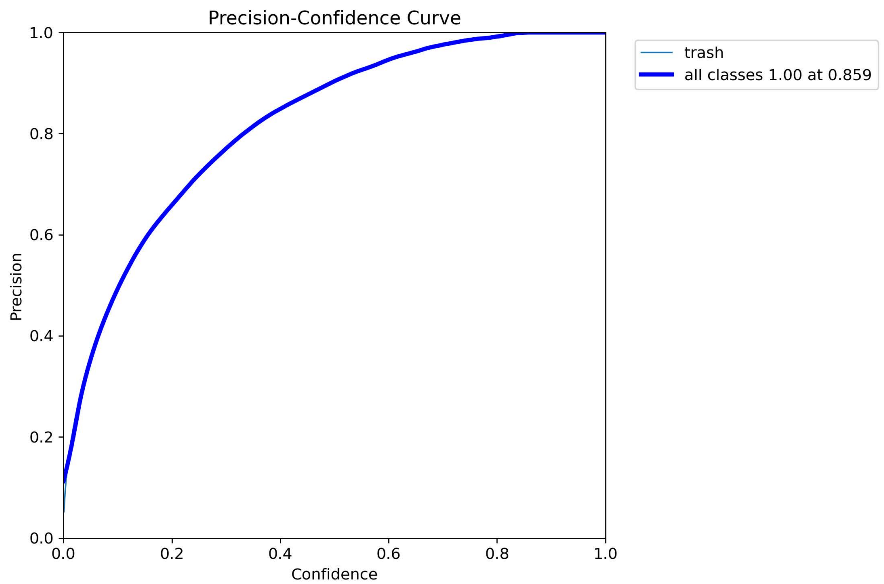


FIGURE 3.8.21: Precision-confidence curve of final illegal dumping detection model

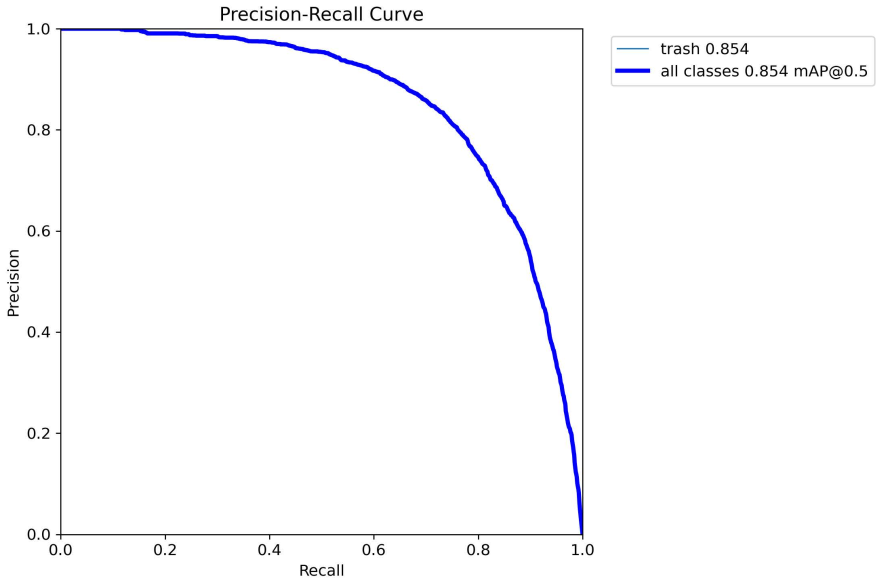


FIGURE 3.8.22: Precision-recall curve of final illegal dumping detection model

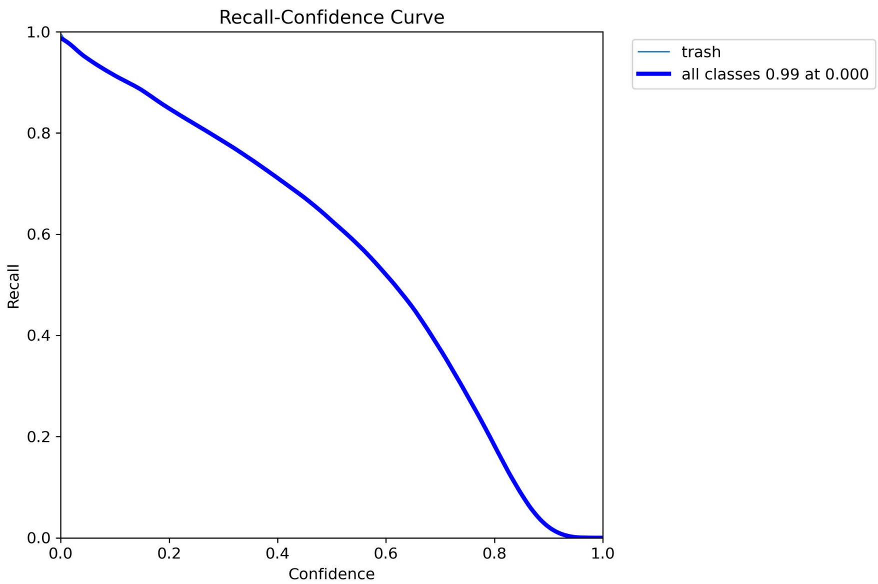


FIGURE 3.8.23: Recall-confidence curve of final illegal dumping detection model

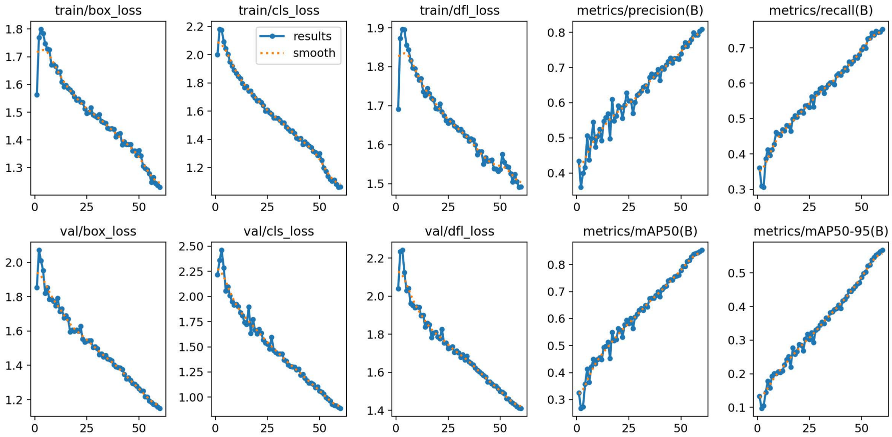


FIGURE 3.8.24: Training results of final illegal dumping detection model

# Appendix 4.1

# Choropleth Outputs

### 4.1.2 People Choropleth

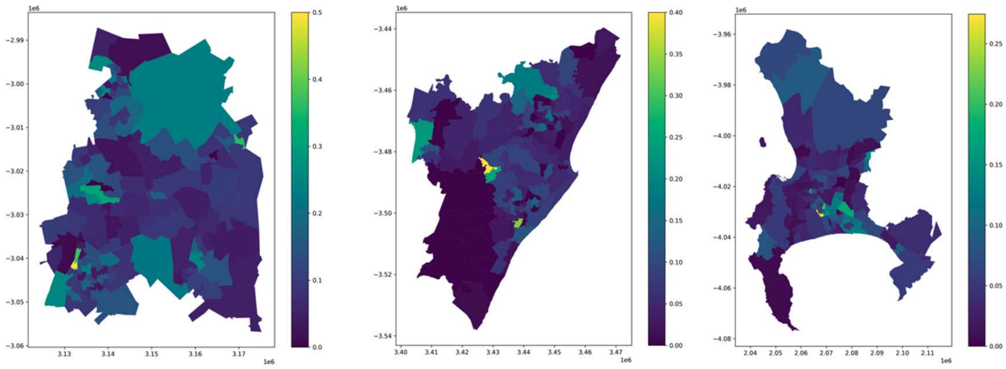


FIGURE 4.1.1: People ratio density across three major municipalities in South Africa at the ward level

### 4.1.3 Vehicle Choropleth

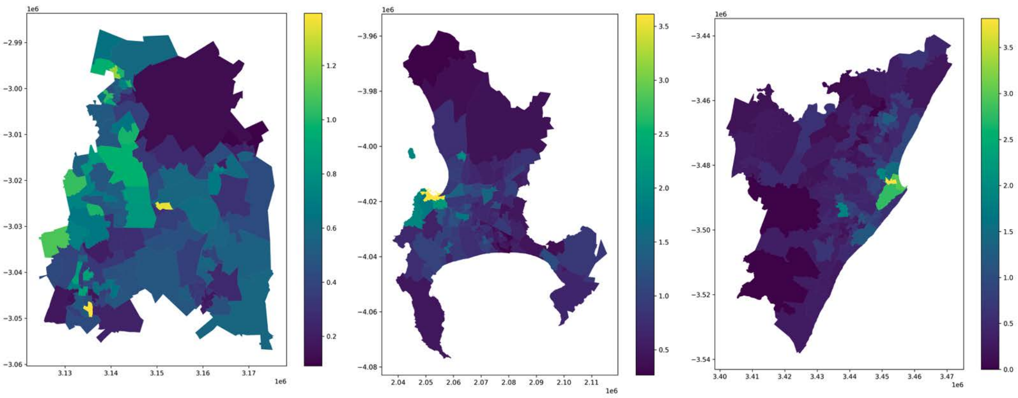

FIGURE 4.1.2: Vehicle ratio density across three major municipalities in South Africa at the ward level

### 4.1.4 Illegal Dumping Choropleth

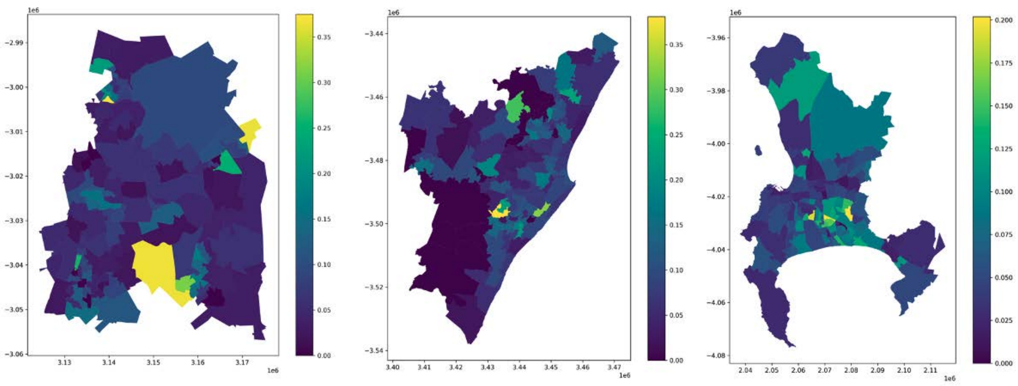

FIGURE 4.1.3: Vehicle ratio density across three major municipalities in South Africa at the ward level

# Appendix 4.2.1

# Actual and Predicted Distributions

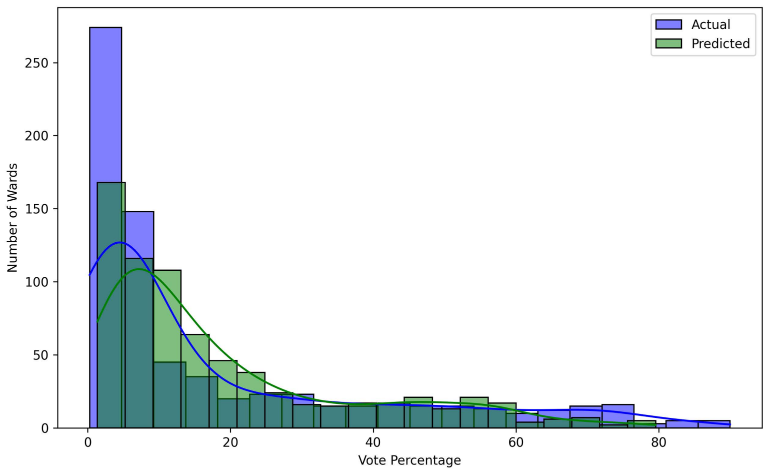


FIGURE 4.2.1.1: Actual and predicted distribution of Democratic Alliance (DA) vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category

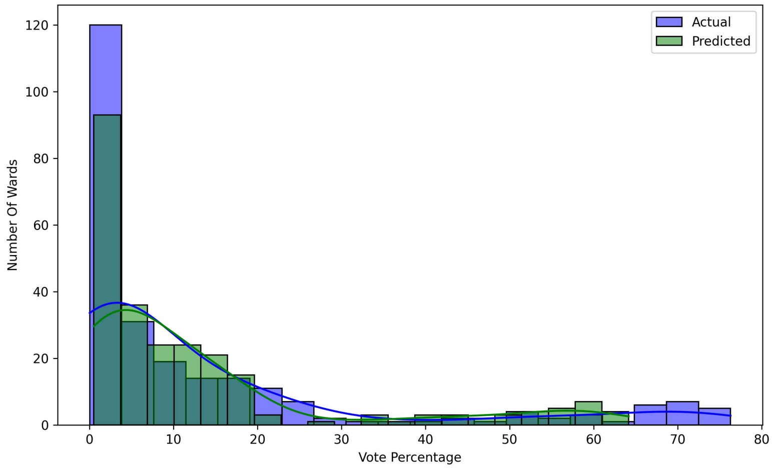


FIGURE 4.2.1.2: Actual and predicted distribution of uMkhonto we Sizwe (MK) vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category

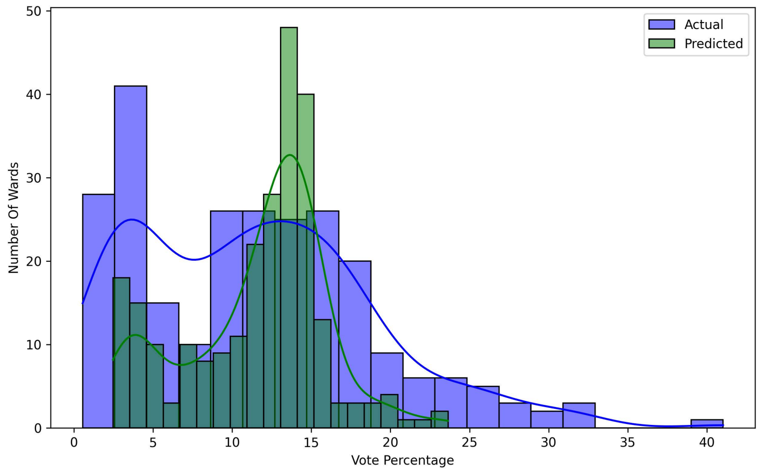


FIGURE 4.2.1.3: Actual and predicted distribution of Economic Freedom Fighters (EFF) vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category

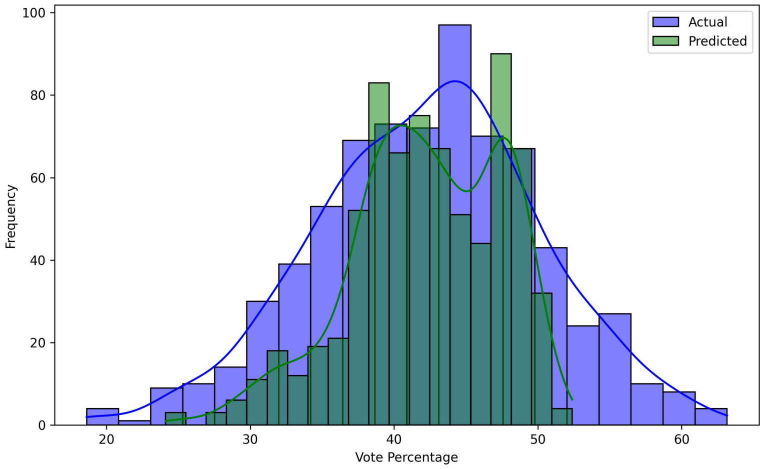


FIGURE 4.2.1.4: Actual and predicted distribution of Non Vote percentage on the national test dataset, showing the number of wards that fall into each vote percentage bucket for each category