Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2021, Article ID 5537902, 14 pages
https://doi.org/10.1155/2021/5537902

Research Article

South Africa Crime Visualization, Trends Analysis, and Prediction Using Machine Learning Linear Regression Technique

Ibidun Christiana Obagbuwa[1] **and Ademola P. Abidoye**[2]

[1]*Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley, South Africa*
[2]*Department of Information Technology, Cape Peninsula University of Technology, Cape Town, South Africa*

Correspondence should be addressed to Ibidun Christiana Obagbuwa; ibidun.obagbuwa@spu.ac.za

Received 25 February 2021; Accepted 13 May 2021; Published 9 June 2021

Academic Editor: Babak Daneshvar Rouyendegh (B. Erdebilli); babek.erdebilli2015@gmail.com

South Africa has been classified as one of the most homicidal, violent, and dangerous places across the globe. However, the two elements that pushed South Africa high in the crime rank are the rates of social violence and homicide. It was reported by Business Insider that South Africa is among the most top 15 ferocious nations on earth. By 1995, South Africa was rated the second highest in terms of murder. However, the crime rate has reduced for some years and suddenly rose again in recent years. Due to social violence and crime rates in South Africa, foreign investors are no longer interested in continuing or starting a business with the nation, and hence, its economy is declining. South Africa's government is looking for solutions to the crime issue and to redeem the image of the country in terms of high crime ranking and boost the confidence of the investors. Many traditional approaches to data analysis in crime-related studies have been done in South Africa, but the machine learning approach has not been adequately considered. The police station and many other agencies that deal with crime hold a lot of databases that can be used to predict or analyze criminal happenings across the provinces of South Africa. This research work aimed at offering a solution to the problem by building a model that can predict crime. The machine learning approach shall be used to extract useful information from South Africa's nine provinces' crime data. A crime prediction system that can analyze and predict crime is proposed. To accomplish this, South Africa crime data on 27 crime categories were obtained from the popular data repository "Kaggle." Diverse data analytics steps were applied to preprocess the datasets, and a machine learning algorithm (linear regression) was used to build a predictive model to analyze data and predict future crime. The appropriate authorities and security agencies in South Africa can have insight into the crime trends and alleviate them to encourage the foreign stakeholders to continue their businesses.

1. Introduction

The causes of high crime rates in South Africa were linked to factors including the low standard of education, alcohol abuse, a lack of social and vocational skills, poor housing and living conditions, and a lack of parenting skills [1]. Social violence crime and homicide are increasing faster than any other crime in South Africa [2, 3]. The report from the INFOGRAPHICS on Crime stats 2020: What you need to know, July 31, 2020, stated that South Africa's violent crime continued to increase (https://www.news24.com/news24/southafrica/news/infographics-crime-stats-2020-what-you-need-to-know-20200731). According to the South Africa annual crime report, 2019, the fight between different groups involving in certain activities such as taxi-related, illegal mining, political motives, and hostel-related violence are the major causes of homicide in South Africa [4]. South Africa yet suffers extreme crime after two decades of postapartheid. As stated by George Otieno et al., and Lindegaard, there is a high rate of homicide-like crime in the typical rural population of South Africa, which requires timely attention to avoid loss of life [5, 6].

With a significant increase in crime across the nations, it has become necessary to analyze crime data to reduce the crime rate. This helps the police, other security agencies, and citizens to take required actions and unravel the crimes faster. Yearly, enormous data is generated by the police and other law enforcement organizations, and analyzing these

data to execute the decision to prevent future crime is the main issue. Performing the analyses of the data will facilitate the recognition of the features responsible for the increase in crime and important steps to curb the crimes. Data mining processes involve evaluating and examining large data such as South Africa crime datasets at Kaggle [7]. New information is generated from the existing data which may be crucial for the prevention of crime in the nation. The extraction of new information will be predicted from the existing datasets. A lot of researchers reported the application of machine learning algorithms to crime analysis in the literature. Hossain et al. used the supervised learning technique (decision tree and k-nearest neighbor) to predict crimes with the San Francisco city dataset of twelve years [8]. Ramasubbareddy et al. built a crime prediction system using a decision tree and Naïve Bayesian classifier [9]. Kim et al. investigated crime prediction in Vancouver using a machine learning approach. They build predictive models with K-nearest neighbor and boosted decision trees [10]. Ahishakiye et al. developed a crime detection prototype model using the decision tree J48 machine learning algorithm which predicted 94.2528% accuracy, and the system is good for predicting future crime [11]. Saltos and Haig investigated instance-based learning, regression, and decision trees machine learning algorithms for predicting crimes by LSOA code (Lower Layer Super Output Areas: an administrative system of areas used by the UK police) and the frequency of antisocial behavior crimes. The results of their experiments show that the decision tree model is the most efficient among the three models [12]. Isha investigated crime analysis and prediction in San Francisco using the data available at the San Francisco Police Department. Prediction models were built with K-nearest neighbor, multiclass logistic regression, decision tree, random forest, and Naïve Bayes machine learning algorithms, and the model predicts the type of crime that will occur in each district of the city and finds applications in resource allocation of law enforcement in a Smart City [13]. Kumar et al. modeled a crime prediction system using the K-nearest neighboring machine learning algorithm [14]. Linear regression machine learning predictive technique is very efficient for building predictive models [15–19].

There is a need for an innovative system and new crime analytics methods for protecting South African communities from crime. By using data mining methods shown in Figure 1, several patterns were revealed and used to predict the amount of crime that is likely to occur in the future, and hence, the police and all security agencies can effectively safely guide the communities in all provinces across the nation. The proposed linear regression predictive model was built based on South African's crime data [7] with the 27 crime categories shown in Table 1, population data [20], and province area (km2) depicted in Table 2 and density that was computed in this work.

2. Methodology

Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology is considered for this work. CRISP-DM is very efficient and suitable for data mining projects. It has been widely used for data mining research in the literature. CRISP-DM steps are described below and depicted in Figure 1.

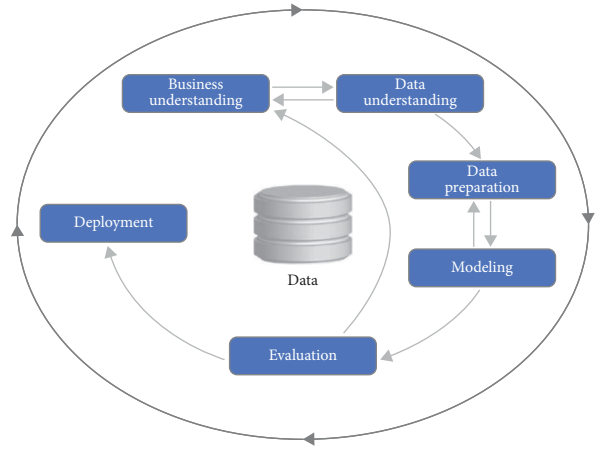


Figure 1: Cross-industry standard process for data mining (CRISP-DM).

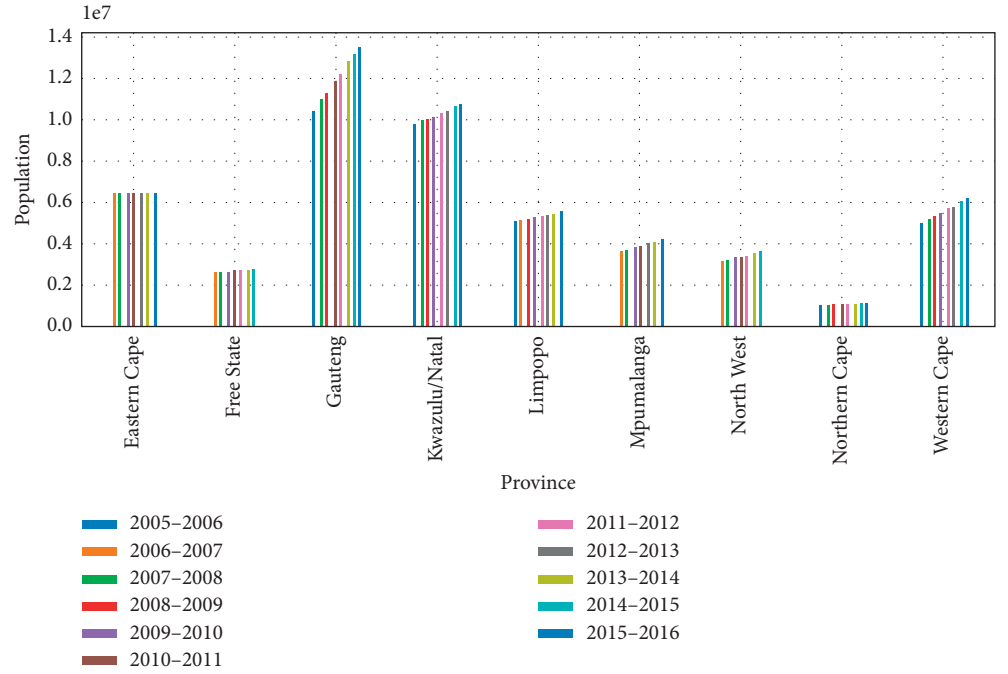
Table 1: Crime category.

Number	Category
1	All theft not mentioned elsewhere
2	Arson
3	Assault with the intent to inflict grievous bodily harm
4	Attempted murder
5	Bank robbery
6	Burglary at nonresidential premises
7	Burglary at residential premises
8	Carjacking
9	Commercial crime
10	Common assault
11	Common robbery
12	Driving under the influence of alcohol
13	Drug-related crime
14	Illegal possession of firearms and ammunition
15	Malicious damage to property
16	Murder
17	Robbery at nonresidential premises
18	Robbery at residential premises
19	Robbery in cash transit
20	Robbery with aggravating circumstances
21	Sexual offenses
22	Sexual offenses as a result of police action
23	Shoplifting
24	Stock-theft
25	Theft of motor vehicle and motorcycle
26	Theft out of or from motor vehicle
27	Truck hijacking

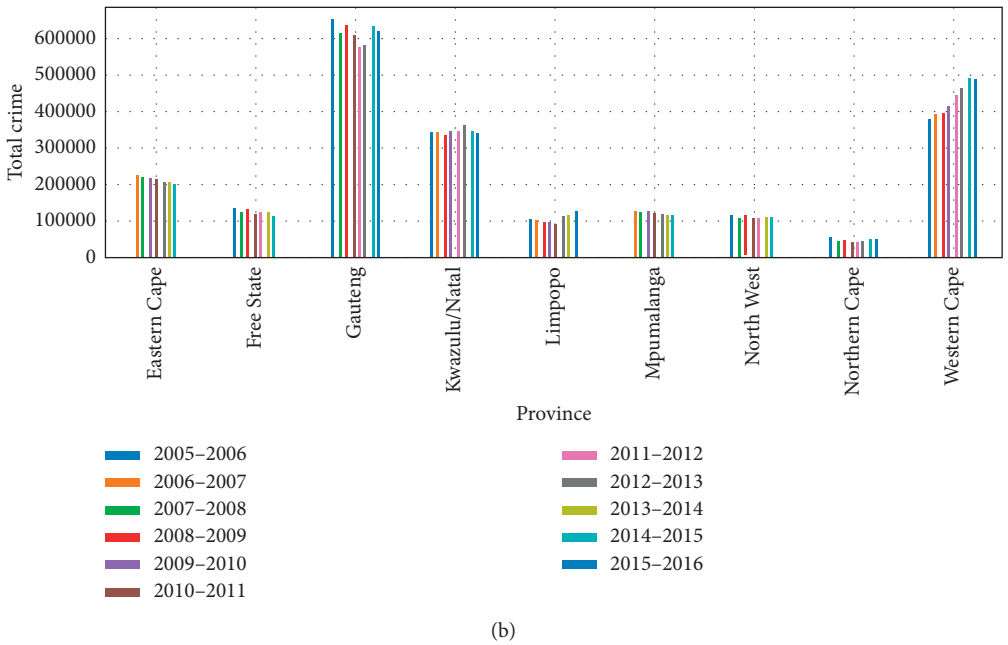
2.1. Business Understanding. This study aims to build a predictive model that can analyze the existing South Africa crime data, detect hidden patterns, and generate useful information that can be communicated to the

Table 2: Province area (km^2) [21].

Rank	Province	Area (km^2)	Percentage
1	Northern Cape	372,889	30.5
2	Eastern Cape	168,966	13.8
3	Free State	129,825	10.6
4	Western Cape	129,462	10.6
5	Limpopo	125,755	10.2
6	North West	104,882	8.6
7	KwaZulu-Natal	94,361	7.7
8	Mpumalanga	76,495	6.3
9	Gauteng	18,178	1.5
Total	South Africa	1220813	100.0



(a)



(b)

Figure 2: 2005–2016 South Africa population and crime statistics. (a) Population statistic. (b) Crime statistic.

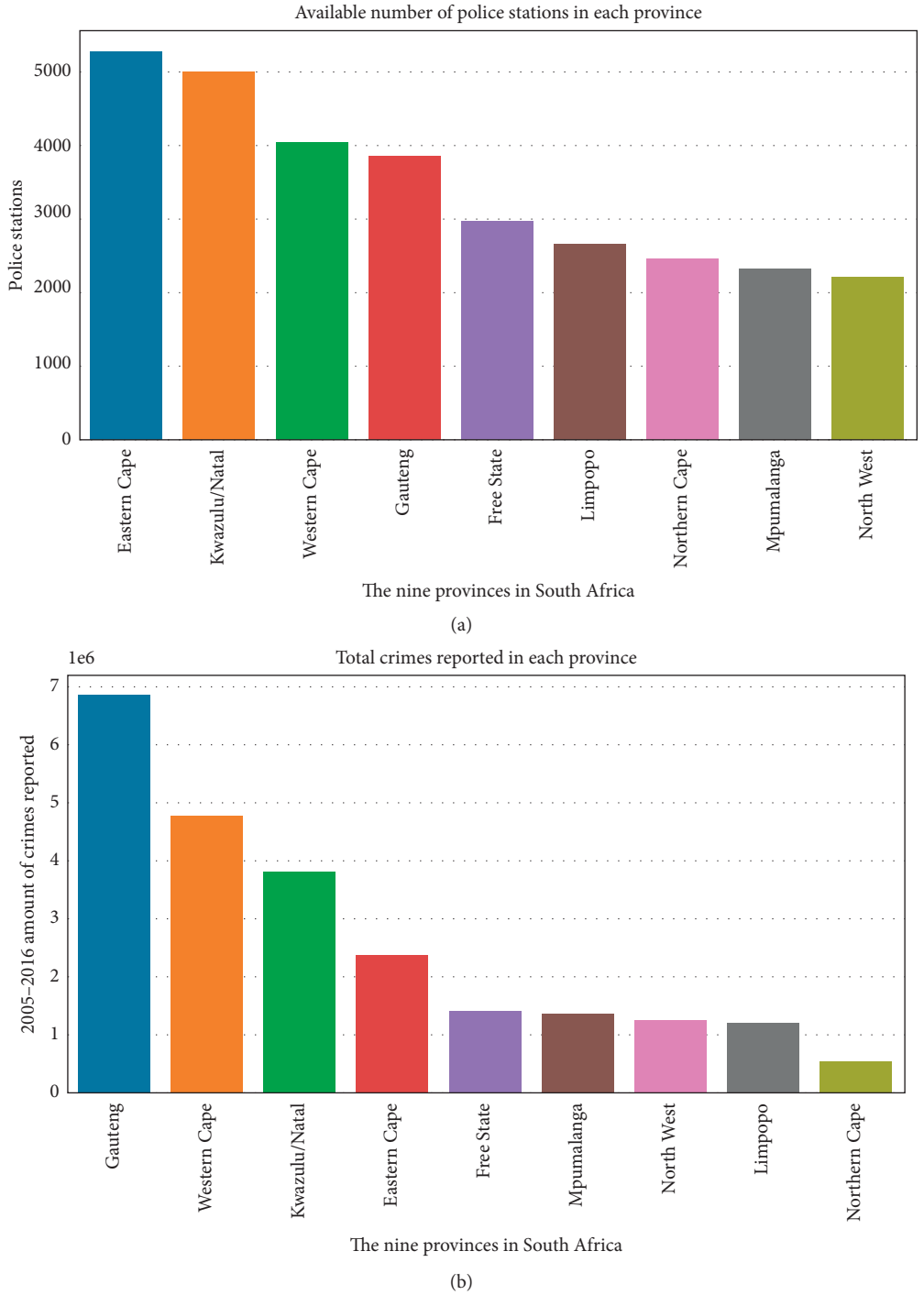


Figure 3: Number of police stations and number of reported crimes in South African Provinces (2005–2016). (a) Number of police stations per province. (b) Total crimes per province.

government and/or security agencies to make timely decisions on how to curb crime in the country.

2.2. Data Understanding. South Africa crime data obtained from the Kaggle repository is used for this work. The activities carried out at this stage include data description, data exploration, and verification of data quality.

2.3. Data Preparation. The crime dataset was organized and makes ready for data analytics. Data selection, data

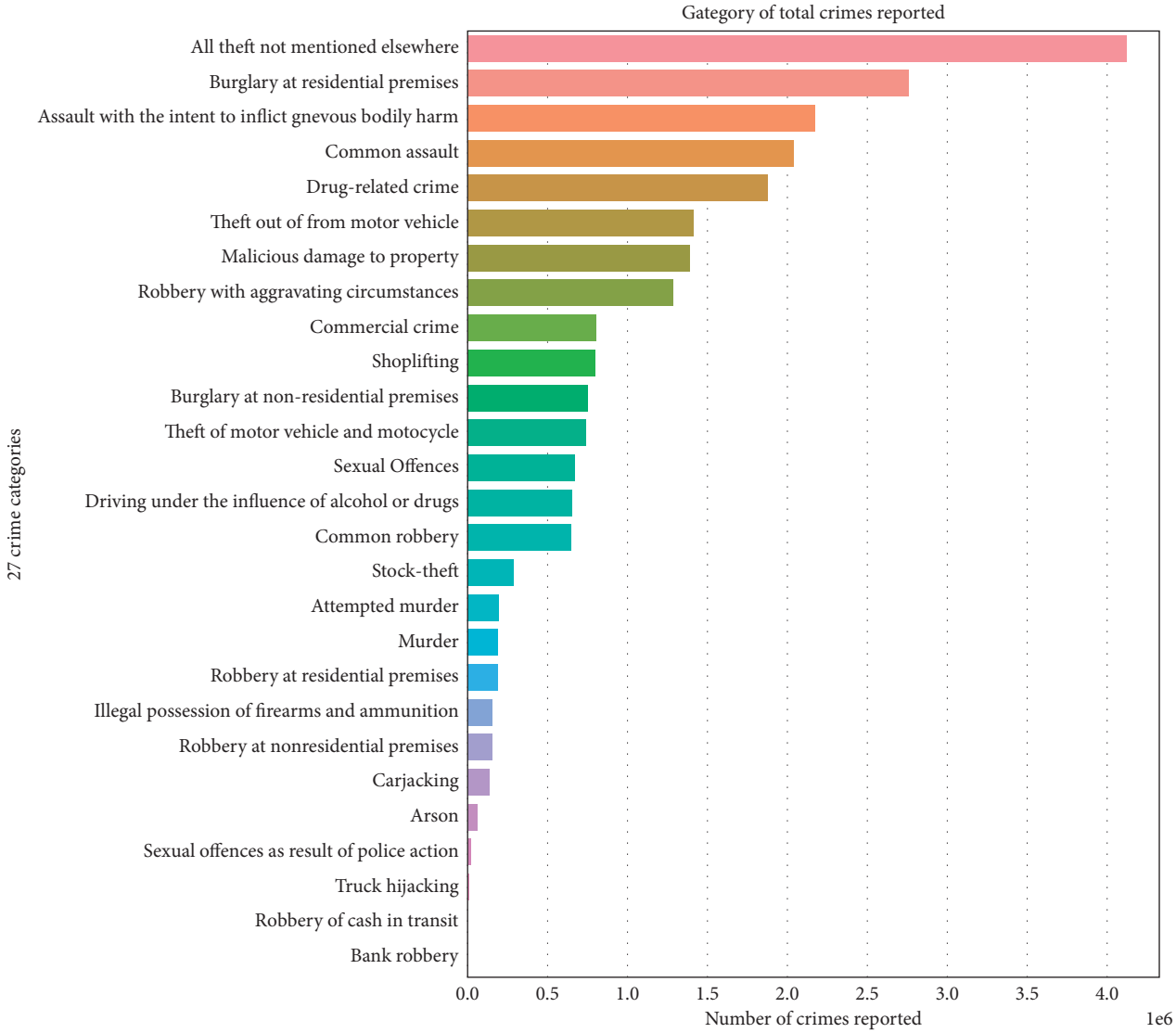


Figure 4: Crime category reported in South African provinces (2005–2016).

cleansing, data construction, and data integration were performed using Python library Scikit-learn (sklearn). Some attributes in the comma-separated values (CSV) files contain string values as well as numeric values.

2.4. Modeling. This is a very crucial stage of the data mining process where machine learning algorithms are applied to the prepared data to analyze the data and create predictive models to make predictions into the future using the useful information generated from the hidden patterns of the data. The activities of this stage include select suitable modeling techniques, that is, the appropriate machine learning algorithm to build the predictive models; generate test design to test the model quality and validity; build a model and run the model tool on the prepared dataset to create one or more models; interpret the model according to domain knowledge, success criteria, and the desired test designs; and ensure the accuracy and generality of the model.

In the execution of the linear regression of some dependent variable y on the set of independent variables $x = x_1,...,x_r$, where r is the number of predictors, a linear relationship between y and x is expressed in the following equation:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_1 x_1 + \varepsilon. \quad (1)$$

This equation is the regression equation where $\beta_0, \beta_1, \ldots, \beta_r$ are the regression coefficients and ε is the random error [22]. Linear regression computes the estimators of the regression coefficients or the predicted weights, denoted with $b_0, b_1, \ldots, b_r$. They define the estimated regression function as shown in the following equation:

$$f(x) = b_0 + b_1 x_1 + \ldots + b_1 x_1. \quad (2)$$

This function detects the inputs and output dependencies. The estimated or predicted response, $f(x_1)$, for each observation $i = 1, \ldots, n$, should be as close as possible to the

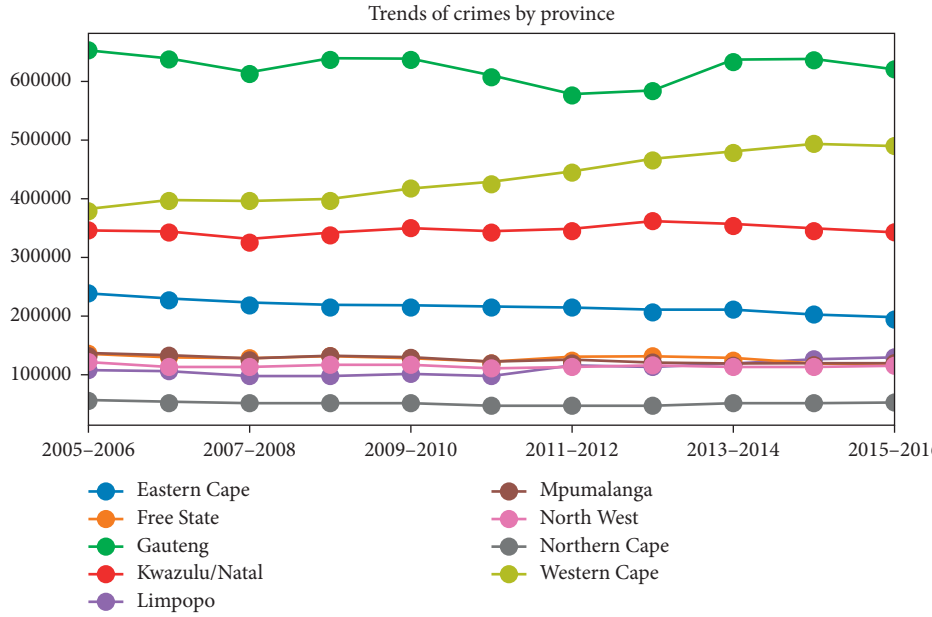


Figure 5: All province total crime trends.

corresponding actual response y_1. The differences $y_i - f(x_i)$ for all observations $i = 1, \ldots, n$ are called the residuals. Regression is about determining the best-predicted weights, that is, the weights corresponding to the smallest residuals. To get the best weights, the sum of squared residuals (SSR) for all observations $i = 1, \ldots, n$: $\mathrm{SSR} = \sum_1 (y_1 - f(x_1))^2$ is minimized [22].

In this work, a linear regression which is one of the machine learning algorithms was considered for building crime predictive models. Linear regression is a predictive modeling method where the target variable to be estimated is continuous. The technological tool is used for implementing the linear regression model in Python using Scikit Learn Modules. This is an efficient data mining tool built on NumPy, SciPy, and matplotlib modules of Python. Sklearn Linear Regression in Scikit allows studying relationships between two continuous (quantitative) variables: one variable, denoted by X, is referred to as the predictor—population, density, and so forth. The other variable denoted by y is regarded as the target—crime variable. A linear regression line has an equation of the form shown in the following equation:

$$y = a + bX, \tag{3}$$

where X is the predictor variable and y is the dependent variable. Hence, the classifier syntax can thus be illustrated as follows:

```
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
regr = linear_model.LinearRegression ()
regr.fit (X, y).
```

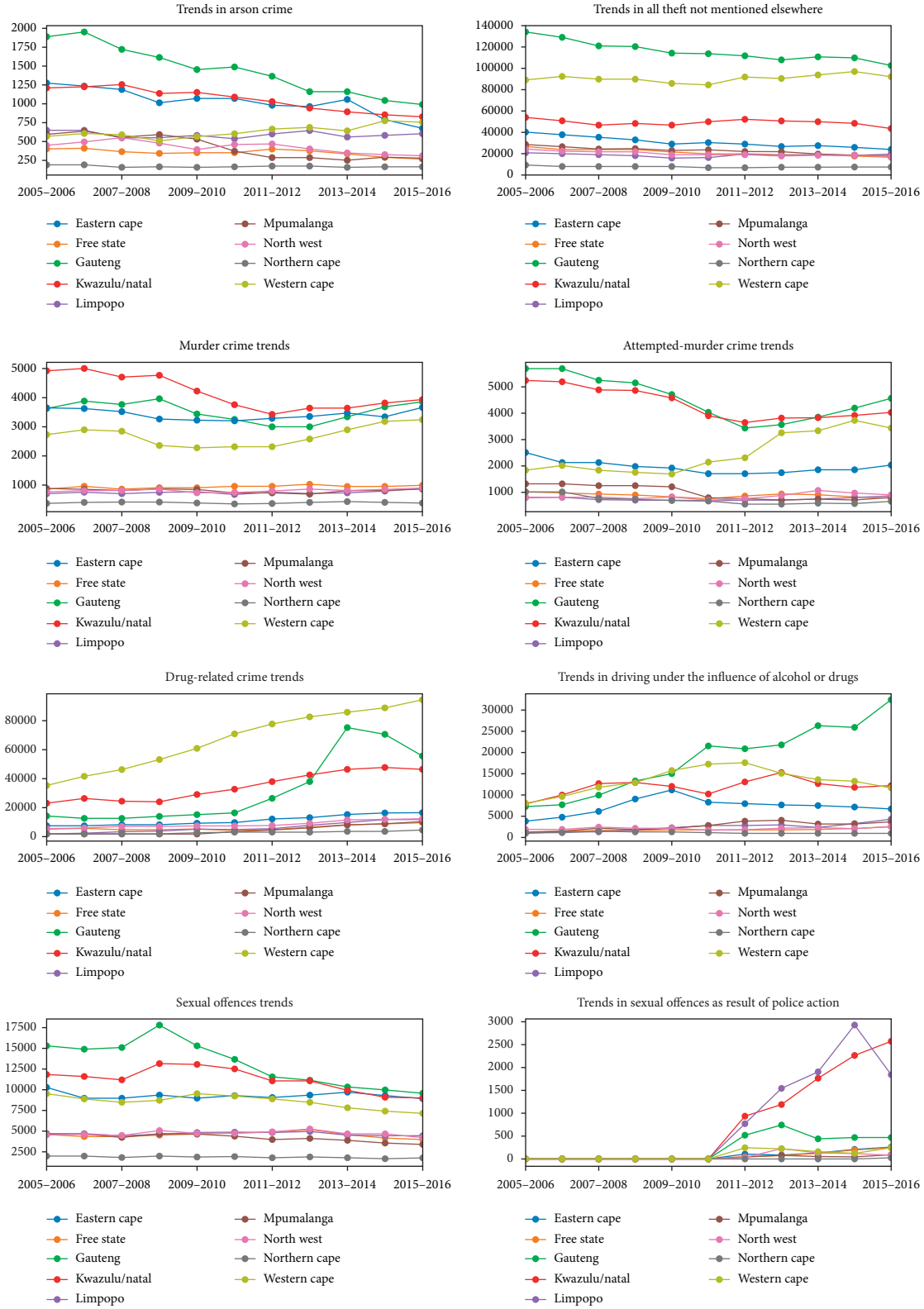
The efficiency of a linear predictive regression model has evaluated the squares of the errors average or deviations (i.e., the difference between the estimator, "features," and what is estimated, "Target Variable"). The difference of actual responses y_i, $i = 1,\ldots,n$, occurs partially due to the dependence on the predictors x_i. Nevertheless, there is also an additional intrinsic variance of the output. The coefficient of determination denoted as R^2 shows which amount of variation in y can be described by the dependence on x using the regression model. Larger R^2 indicates a better fit and means that the model can better describe the difference of the output with different inputs. The value $R^2 = 1$ corresponds to $\mathrm{SSR} = 0$, which is the perfect fit since the values of the predicted and actual responses fit completely to each other.

2.5. Evaluation. The degree to which the model meets the project objective is assessed at this stage. After assessing the models, the generated model that meet the objective of the project is considered.

2.6. Deployment. Strategies to establish the evaluation results will be determined at this stage which includes the final report.

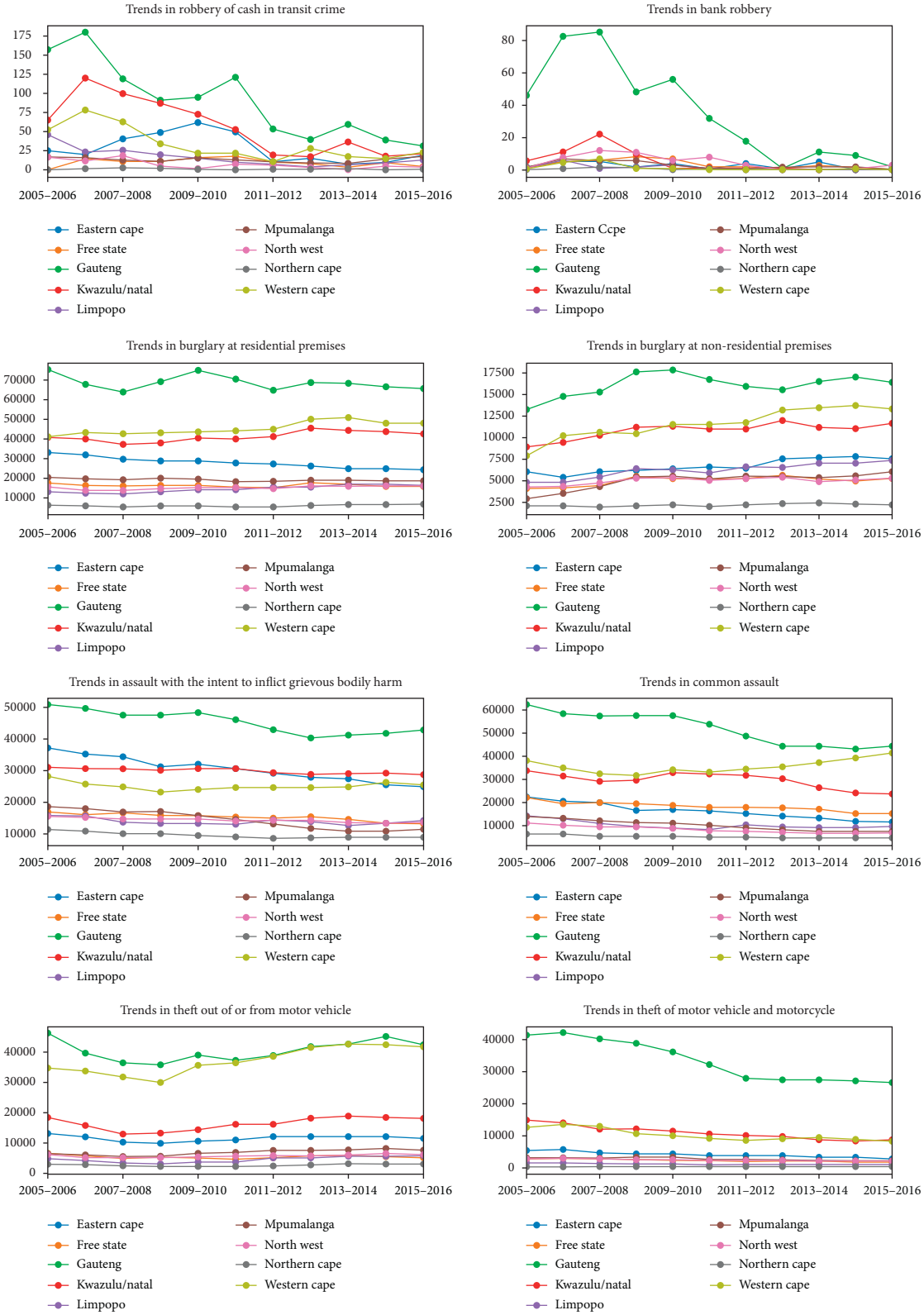
3. Experiments, Results, and Discussion

The visualization of 27 crime categories, trends in crimes, and all the results from the linear regression implementation are provided in this section. Linear regression machine learning predictive technique has been widely used in the literature for building predictive models [15–19]. Using equation (3), which detects the relationship in the crime dataset; with the trend line, the corresponding y-value for a future x-value can be determined. Python Libraries for data visualization were used for visualizing the south Africa crime dataset 2005–2016. Figure 2 depicts South Africa's crimes and population statistics 2005 to 2016; there is a correlation between population rates and crime rates; the higher the population, the higher the number of crimes. The four



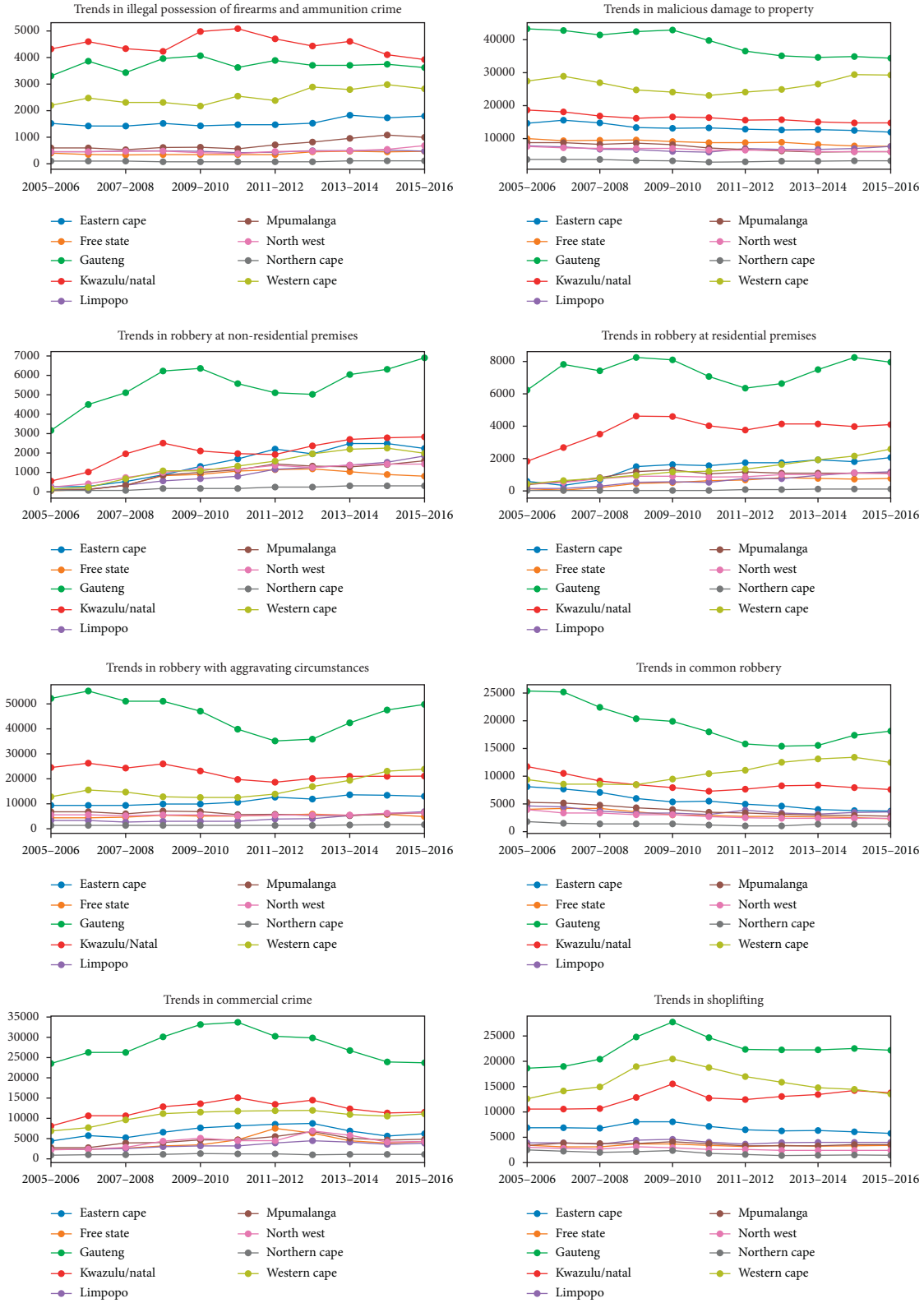
(a)

Figure 6: Continued.



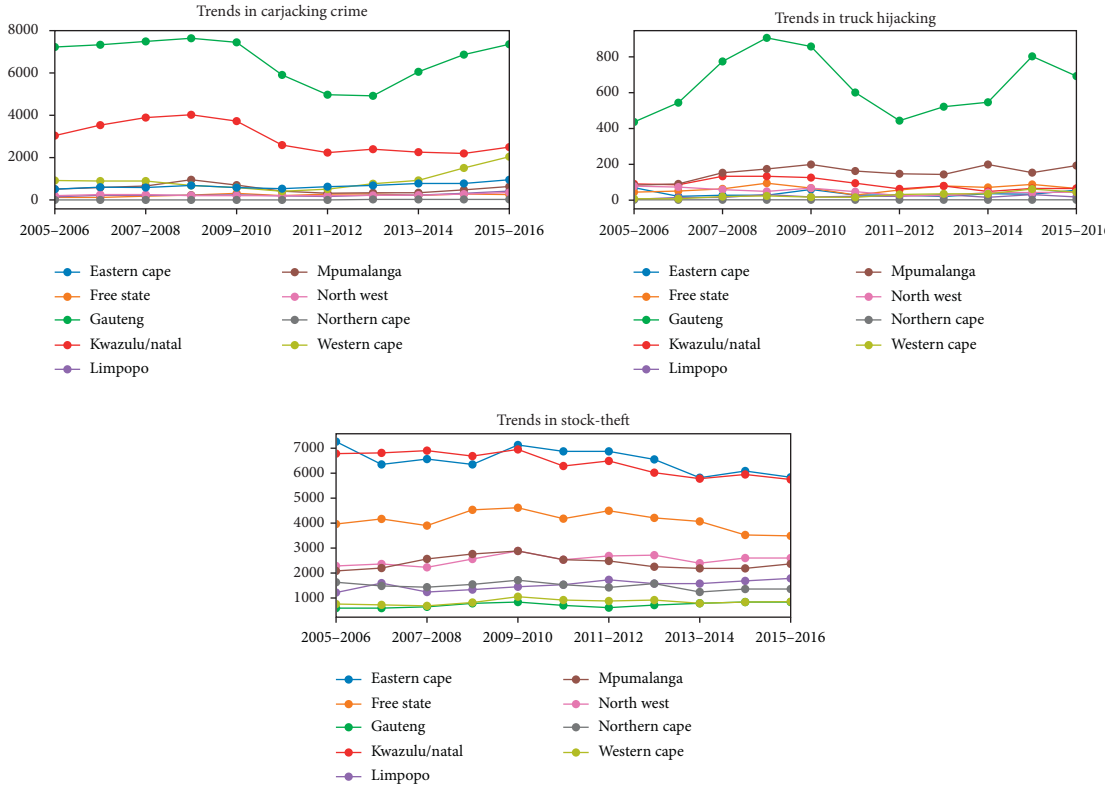
(b)

Figure 6: Continued.



(c)

Figure 6: Continued.



(d)

Figure 6: Trends in the 27 crime categories per province in 2005–2016.

(a) (b) (c) (d)

Figure 7: Word cloud of the crime category in (a) Gauteng, (b) KwaZulu-Natal, (c) Western Cape, and (d) Eastern Cape.

provinces (Gauteng, Kwazulu-Natal, Western Cape, and Eastern Cape) with high population statistics also have the highest number of crimes. Figure 3 illustrates the number of police stations in each of the nine provinces of South Africa and the number of crimes reported, respectively. There is no correlation between the number of police stations in a province and the number of crimes committed. For instance, Gauteng is shown to have the highest total crime rate, but it does not have the highest number of police stations as shown in Figures 3(a) and 3(b). The South Africa dataset considered for this work [7] has 27 crime categories as shown in Table 1, and Figure 4 illustrates the 27 crime categories and the number of crimes for each category. The categories (all theft not mentioned elsewhere, burglary at residential premises, and assault with the intent to inflict grievous bodily harm) have the highest crime occurrences while the categories (bank robbery, robbery of cash in transit, and truck hijacking) have the least number of crime occurrences (Figure 4). Figure 5 depicts the trend of total crime in each province from 2005 to 2016; Gauteng Province has the

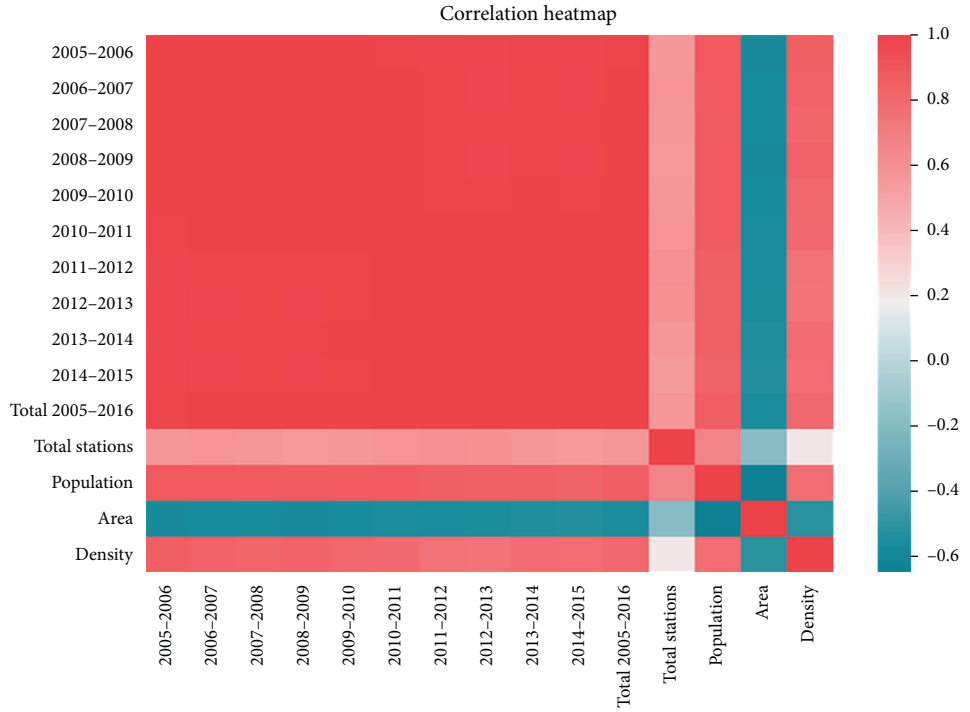


Figure 8: Correlation heatmap.

	2005–2006	2006–2007	2007–2008	2008–2009	2009–2010	2010–2011	2011–2012	2012–2013	2013–2014	2014–2015	Total 2005–2016	Total stations	Population	Area	Density
2005–2006	1	1	1	1	1	0.99	0.98	0.98	0.98	0.98	0.99	0.57	0.89	−0.57	0.86
2006–2007	1	1	1	1	1	1	0.99	0.98	0.99	0.99	1	0.57	0.88	−0.57	0.84
2007–2008	1	1	1	1	1	1	0.99	0.99	0.98	0.99	1	0.57	0.87	−0.57	0.83
2008–2009	1	1	1	1	1	1	0.99	0.99	0.99	0.99	1	0.56	0.88	−0.57	0.84
2009–2010	1	1	1	1	1	1	0.99	0.99	0.99	0.99	1	0.57	0.87	−0.56	0.82
2010–2011	0.99	1	1	1	1	1	1	1	1	0.99	1	0.58	0.86	−0.56	0.8
2011–2012	0.98	0.99	0.99	0.99	0.99	1	1	1	1	1	1	0.6	0.86	−0.55	0.76
2012–2013	0.98	0.98	0.99	0.99	0.99	1	1	1	1	1	1	0.59	0.85	−0.55	0.74
2013–2014	0.98	0.99	0.99	0.99	0.99	1	1	1	1	1	1	0.57	0.85	−0.54	0.78
2014–2015	0.98	0.99	0.99	0.99	0.99	0.99	1	1	1	1	1	0.56	0.83	−0.54	0.77
Total 2005–2016	0.99	1	1	1	1	1	1	1	1	1	1	0.57	0.86	−0.56	0.8
Total stations	0.57	0.57	0.57	0.56	0.57	0.58	0.6	0.59	0.57	0.55	0.57	1	0.67	−0.19	0.21
Population	0.89	0.88	0.87	0.88	0.87	0.86	0.86	0.85	0.85	0.83	0.86	0.67	1	−0.65	0.77
Area	−0.57	−0.57	−0.57	−0.57	−0.56	−0.56	−0.56	−0.55	−0.54	−0.54	−0.56	−0.19	−0.65	1	−0.52
Density	0.86	0.84	0.83	0.84	0.82	0.8	0.76	0.74	0.78	0.77	0.8	0.21	0.77	−0.52	1

Figure 9: Correlation gradient.

highest number of crimes. The top 3 crime-prone provinces are Gauteng, Western Cape, and KwaZulu-Natal, respectively. Moreover, the three provinces with fewer crimes are Northern Cape, Limpopo, and North West, respectively (Figure 5).

The trends of the 27 crime categories per province in 2005–2016 are depicted in Figure 6. A clear illustration of provinces that are prone to any of the 27 categories of crime is given in Figure 6; for instance, Gauteng province has the highest number of crime in 23 crime categories out of 27; KwaZulu-Natal province has the highest number of crime in the categories of murder and illegal possession of firearms and ammunition; for the category of stock theft, KwaZulu-Natal and Eastern Cape take the lead; and for the category of drug-related crime, Western Cape has the highest crime rate.

A data visualization technique known as the Word Cloud (Figure 7) depicts the crime category. The size of each word specifies its frequency or importance. Observing the most prominent crime category in the respective provinces shown in Figure 7 is made easy and quickly at a glance.

Intercept:
38825.39788764125
Coefficients:
[2.79062433e–02 5.52800083e + 02]
Predicted crime:
[230033.17129904]

OLS regression results

Dep. variable:	Crime	R-squared:	0.847
Model:	OLS	Adj. R-squared:	0.795
Method:	Least Squares	F-statistic:	16.56
Date:	Fri, 04 Dec 2020	Prob (F-statistic):	0.00361
Time:	10:08:44	Log-Likelihood:	−113.23
No. observations:	9	AIC:	232.5
Df residuals:	6	BIC:	233.1
Df model:	2		
Covariance type:	nonrobust		

	coef	std err	t	P > |t|	[0.025	0.975]
const	3.883e+04	6.44e+04	0.603	0.569	−1.19e+05	1.96e+05
Population	0.0279	0.014	2.038	0.088	−0.006	0.061
Density	552.8001	241.073	2.293	0.062	−37.085	1142.685

Omnibus:	12.958	Durbin-Watson:	1.081
Prob (omnibus):	0.002	Jarque-Bera (JB):	5.346
Skew:	1.545	Prob (JB):	0.0690
Kurtosis:	5.170	Cond. No.	1.36e + 07

Figure 10: Linear regression results.

A machine learning model was built using linear regression (with the existing data on crime, population, area, and density) to predict future crime occurrence. Multicollinearity among features can be identified by doing Feature-Feature correlation analysis. In linear regression, the input variables should not be multicollinear, that is, dependent on each other. The heatmap shown in Figure 8 depicts a positive correlation between total crime, population, and density. Density is the number of people per square kilometer. Figure 8 also shows a positive correlation between the number of police stations and the total crimes in a province; this may not necessarily mean that the more the police stations, the more the crimes, but it is an indication that there are more police stations to handle a higher number of crimes. However, Figure 8 depicts a negative correlation (no relationship) between the size of a province (area) and the number of crimes that occurred in a province. Figure 9 illustrates further these relationships with a correlation gradient.

Series of experiments were carried out and the regression results are shown in Figure 10. Moreover, the future crime in South Africa can be predicted using the linear regression model built in this work. Figure 11 depicts the sample prediction output. Linear regression fits a straight line to the data using two continuous variables: predictor variables and the response variable. A linear regression line has an equation of the form $Y = a + bX$, where X is the explanatory variable and Y is the dependent variable.

Target = $a + b$ (features). For instance:

Crime_Number = $a + b$ (population).

Crime_Number = $a + b$ (density).

Linear regression reduces the sum of squares of the variables predicted by linear approximation.

From the illustration of linear regression results in Figure 10, as P values of population and density are low, the crime rate is closely related to population and density R-squared value = 0.847 means 84.7% variability of crime rate is described by the population and density features. R-squared is an estimate of the strength of the relationship between your model and the response variable that lies between 0 (worst) and 1 (best), the R-squared value of 0.847 of this model indicates a strong relationship. F-Stat is a statistical test that compares the fit of the intercept-only model with our model; the P value for the F-Stat of 0.00361 less than 0.005 (significant level) indicates that our model is a very good one. Akaike information criterion (AIC) estimates the relative information lost by a given model; the less information a model loses, the higher the quality of the model. Hence, the lower the AIC, the better. AIC and BIC (Bayesian information criterion) of values 232.5 and 233.1 represent the good quality of the model. The actual crime rate and the predicted crime rate plot are linear. Hence, the crime rate prediction is almost the same as the actual crime rates, and therefore, the linear model is working correctly (Figure 10). The Omnibus/Prob (Omnibus) is a test of the skewness and kurtosis of the residual characteristic. The Prob (Omnibus) shown in Figure 10 performs a statistical test, indicating the probability that the residuals are normally distributed. The Prob (Omnibus) value of 0.002 that is close to zero indicates the normalcy of the data. Skew is a measure of data symmetry and its value drives Omnibus; here, the small value of the skew (1.545) indicates residual distribution is normal. However, the Kurtosis is the measure of peakedness or curvature of the data; the higher Kurtosis of the 5.170 value shows tighter clustering of residuals around zero, meaning a better model with few outliers. For the Durbin–Watson tests

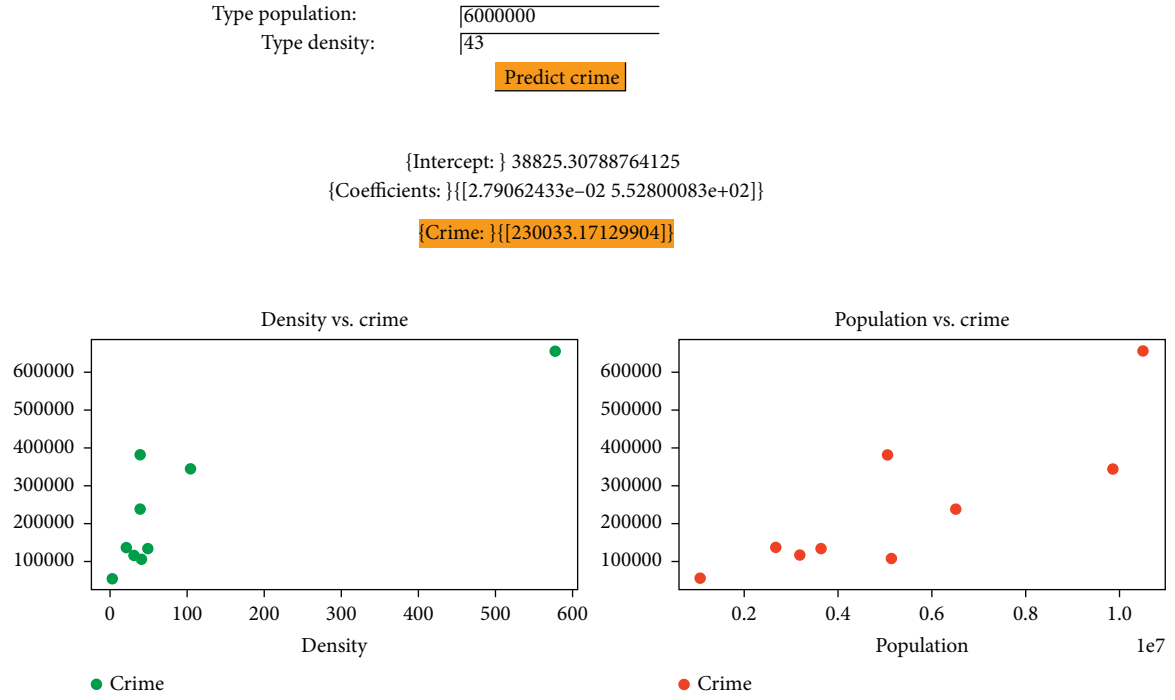


Figure 11: Crime prediction with linear regression approach.

for homoscedasticity characteristic, the values must be between 1 and 2; the Durbin–Watson output of 1.081 indicates that the data is within the limits. Jarque-Bera (JB)/Prob (JB) is like the Omnibus test tests both skew and kurtosis; the Prob (JB) value 0.0690 that is very close to 0.002 of Prob (Omnibus) is a confirmation of the Omnibus test (Figure 10). The overall results depicted in Figure 10 show the linear regression model is an efficient model for predicting the crime rates in South Africa. It can be applied to predict crime in any of the nine provinces of the country. The prediction system has been trained with the crime data and takes inputs (population and density) to predict crime occurrence (Figure 11).

4. Conclusion

Machine learning technique can effectively detect the hidden patterns in crime data that are valuable, give good visualization for crime prediction, and thus provide support for crime prevention in South Africa. Crime data analytics can extract unknown, vital information from raw data and, thus, assist the government to speed up the procedures of resolving crime. It would enable appropriate authorities in the government to gain a better understanding of crime trends and mitigate against them. When the crime is prevented and the environment is peaceful, foreign investors are happy to continue with their businesses in South Africa, and hence economic growth is sustained. This work presents a predictive model trained with crime data and can take population and density as inputs to predict the total crime of any province of South Africa. The extension of this work shall seek information on crime other factors from the South Africa Police authority and build a predictive model considering those factors.

Data Availability

South Africa crime data are obtained from Kaggle, https://www.kaggle.com/slwessels/crime-statistics-for-south-africa.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

Ibidun Christiana Obagbuwa and Ademola P. Abidoye contributed equally to the manuscript. Ibidun Christiana Obagbuwa contributed to generation of ideas, design, implementation, literature review, and writing of the paper. Ademola P. Abidoye contributed to implementation, literature review, and paper writing.

Acknowledgments

The authors would like to appreciate the support of Sol Plaatje University for this research. Furthermore, the authors are thankful to Kaggle for the availability of South Africa crime data (2005–2016) and the notebooks.

References

[1] R. McCafferty and U. C. Action, "Murder in South Africa: a comparison of past and present," *United Christian Action*, pp. 1–24, 2003.

[2] Business Insider South Africa, *South Africa Climbed the Ranks to Become One of the Most Murderous Countries in the World*, Business Insider South Africa, Cape town, South Africa, 2019, https://www.businessinsider.co.za/south-africa-crime-2019-11.

[3] Statistics South Africa, *South Africa's Crime Statistics for 2018/19*, https://citizen.co.za/news/south-frica/crime/2178462/factsheet-south-africas-crime-statistics-for-2018-19/2019.

[4] Statistics South Africa, *South African Annual Crime Report 2017/2018*, RP NUMBER: 299/2018, ISBN NUMBER: 978-0-621-46552-5, Crime Registrar, 2019.

[5] G. Otieno, E. Marinda, T. Bärnighausen, and F. Tanser, "High rates of homicide in a rural South African population (2000-2008): findings from a population-based cohort study," *Population Health Metrics*, vol. 13, no. 1, p. 20, 2015.

[6] M. R. Lindegaard, *Homicide in South Africa, the Handbook of Homicide*, F. Brookman, E. R. Maguire, and M. Maguire, Eds., John Wiley & Sons, New Jersey, NJ, USA, First edition, 2017.

[7] Statistics South Africa, *Crime Statistics for South Africa*, https://www.kaggle.com/slwessels/crime-statistics-for-south-africa. Accessed December 2020.

[8] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," 2020, https://arxiv.org/abs/2003.09322.

[9] S. Ramasubbareddy, T. Aditya Sai Srinivas, K. Govinda, and S. S. Manivannan, "Crime prediction system," in *Innovations in Computer Science and Engineering*, H. Saini, R. Sayal, R. Buyya, and G. Aliseri, Eds., vol. 103, Singapore, Springer, 2020, Lecture Notes in Networks and Systems.

[10] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime analysis through machine learning," in *Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, BC, Canada, November 2018.

[11] E. Ahishakiye, E. O. Omulo, D. Taremwa, and I. Niyonzima, "Crime prediction using decision tree (J48) classification algorithm," *International Journal of Computer, and Information Technology*, vol. 6, no. 3, ISSN: 2279 – 0764, 2017.

[12] G. Saltos and M. Cocea, "An exploration of crime prediction using data Mining on open data," *International Journal of Information Technology & Decision Making*, vol. 16, no. 5, pp. 1155–1181, 2017.

[13] P. Isha, *Exploratory Data Analysis and Crime Prediction in San Francisco*, Master's Projects. 642, https://scholarworks.sjsu.edu/etd_projects/642, 2018.

[14] A. Kumar, A. Verma, G. Shinde, Y. Sukhdeve, and N. Lal, "Crime prediction using K-nearest neighboring algorithm," in *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, pp. 1–4, Vellore, India, February 2020.

[15] K. Sukhija, S. N. Singh, and M. Kumar, "Using Linear Regression to investigate parameters associated with Rape crime in Haryana," in *Proceedings of the 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 107–111, Noida, India, January 2020.

[16] J. J. Gonzalez and A. Leboulluec, "Crime prediction and socio-demographic factors: a comparative study of machine learning regression-based algorithms," *Journal of Applied Computer Science and Mathematics*, vol. 13, no. 27, 2019.

[17] M. A. Awal, J. Rabbi, S. I. Hossain, and M. Hashem, "Using linear regression to forecast future trends in the crime of Bangladesh," in *Proceedings of the 2016 International Conference on Informatics, Electronics, and Vision (ICIEV)*, pp. 333–338, Dhaka, Bangladesh, May 2016.

[18] P. Yerpude and V. Gudur, "Predictive modelling of crime dataset using data mining," *International Journal of Data Mining and Knowledge Management Process*, vol. 7, no. 4, pp. 43–58, 2017.

[19] B. Cavadas, P. Branco, and S. Pereira, "Crime prediction using regression and resources optimization," in *Progress in Artificial Intelligence Portuguese Conference on Artificial Intelligence*, pp. 513–524, Springer, Cham, Switzerland, 2015.

[20] Statistics South Africa, *Statistical Release (P0318), General Household Survey*, http://www.statssa.gov.za/publications/P0318/P03182018.pdf, 2018.

[21] Statistics South Africa, *Stats in Brief*, vol. 3, Statistics South Africa, Pretoria, South Africa, 2009, 978-0-621-38774-2 http://www.statssa.gov.za/publications/StatsInBrief/StatsInBrief2009.pdf.

[22] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer New York Heidelberg Dordrecht London, Library of Congress Control Number: 2013936251 © Springer Science+Business Media New York, (Corrected at the printing 2017), 2013.