Exploratory framework for analysing road traffic accident data with validation on Gauteng province data

Tebogo Makaba¹*, Wesley Doorsamy² and Babu Sena Paul²

Abstract: Exploratory data analysis (EDA) is often a necessary task in uncovering hidden patterns, detecting outliers, and identifying important variables and any anomalies in data. Furthermore, the approach can be used to gain insights by modelling the dataset through graphical representations. In this paper, we propose an exploratory framework for analysing a road traffic accidents real-life dataset using graphical representations and incorporating dimensionality reduction methods. Both Principal component and Linear discriminant analyses are performed on the dataset and the resulting performance metrics reveal some comprehensive insights of the road traffic accident patterns. The investigation also revealed which road traffic factors contribute more significantly to the events. Classification results were generated after applying the dimensionality reduction methods to the dataset and show that the application of Linear discriminant analysis dimensionality reduction together with Naïve Bayes classification performed better as compared to the other approaches for the dataset.

ABOUT THE AUTHOR
Ms Tebogo Makaba is currently a PhD candidate and working as a lecturer at the University of Johannesburg. Ms Makaba’s area of research is in computational intelligence, healthcare, Emerging Tech and Transportation. Prof Wesley Doorsamy completed the PhD from the University of Witwatersrand. He also completed the postgraduate diploma in education. Currently, he is an Associate Professor at the Institute for Intelligent Systems, University of Johannesburg. Prof Doorsamy’s area of research is in condition monitoring and diagnostics, computational intelligence, analytics and optimisation in different application areas including, inter alia, power and energy, banking, healthcare and manufacturing sectors. Prof Babu Sena Paul received his Ph.D. degree from the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati India. He is the currently serving as an Associate Professor and Director to the Institute for Intelligent Systems, University of Johannesburg. His research interests are in the area of Artificial Intelligence and machine learning.

PUBLIC INTEREST STATEMENT
Data analysis is of importance when dealing with large or different types of real-life datasets which can be used to understand further insights into the data. The study proposes an exploratory framework for analysis of road traffic accident data using some visualisation representations and consideration of feature reduction methods that can be used to analyse in detail the patterns of historical datasets. The authors extended the work by considering popular forecasting methods to access their performance. The proposed framework can be used widely to extract insights from data before it can be utilised for future data analytics. The study would mainly benefit transport planners, transportation researchers, policymakers and transport data scientists.
1. Introduction

Exploratory data analysis (EDA) has been widely used in research with literature thereof employing different graphical representations and statistical analyses, to perform preliminary investigations on datasets. EDA is well known as an approach that can be used to examine datasets to identify and uncover hidden patterns and answer some important questions (Martinez et al., 2010). The idea behind EDA is to obtain a background context of the dataset to be able to develop an appropriate prediction model. EDA approach can be employed to identify important variables, detect outliers and spot anomalies in the dataset (Martinez et al., 2010). Thus, EDA can be classified into four groups (Chambers, 2018; DuToit et al., 2012): the non-graphical univariate and multivariate methods mainly involve the calculation of summary statistics while the graphical univariate and multivariate methods use some graphical ways to summarise analyse and present the dataset. Furthermore, the univariate methods focus on two or more variables timely to discover their relation and the multivariate methods focuses only on two variables, or in some cases, it can expand to more than two variables. EDA is the best practice that can be applied in different domains such as anomaly detection, speech recognition, fraud detection, etc.

EDA can reveal hidden patterns in a dataset that can play an important role during the prediction phase. If EDA is not addressed during the early stages, this could negatively impact the performance of the model during the modelling stage. In machine learning (ML) EDA is significant as it helps to establish sound assumptions and answer questions thereby ensuring the best results are obtained during the model design phase.

Road traffic accidents (RTAs) are the major cause of the high number of fatalities and injuries globally. In addition, road traffic accidents are a major concern around the African continent, which is killing thousands of innocent people (World Health Organization, 2018). RTAs raw dataset is important as it can greatly help transport planners and researchers to uncover trends and hidden patterns. RTAs possess some varying and hidden characteristics which require EDA methods to gain more comprehensive insights and reveal the important characteristics. Graphical representation can be beneficial to the transport planners and the engineers to easily examine the raw dataset before any model construction is undertaken (Shbee & Awad, 2016). Numerous graphical representation tools have been employed by researchers such as boxplots, bar charts, histograms, scatter plots, among others (DuToit et al., 2012; Muguro et al., 2020; Wells-Parker et al., 2002). A considerable number of literature work is available which has addressed the importance of EDA across different research areas such as transport management, image processing, anomaly detection, speech recognition such as (Cuenca et al., 2018; Gazder et al., 2018; Lavrac et al., 2008; Michalaki et al., 2015 & Timmermans et al., 2019). Furthermore, other available research works by (Ahmadi et al., 2020; Abou Ellassad et al., 2020; De Andrade et al., 2014; Feng et al., 2017; Rende et al., 2013; Yong-dong et al., 2019) introduces the importance of applying PCA dimensionality reduction and classification methods in their work to reduce the scope of variables their dataset contained and further performed classification.

The main contributions of this paper are to propose an exploratory framework to analyse road traffic accidents dataset using graphical representation such as boxplot, bar chart and histogram. A comprehensive exploratory analysis of RTAs is common but in addition to this work, we have introduced dimensionality reduction techniques namely, principal component analysis (PCA) and
Linear discriminant analysis (LDA) techniques, which are some of the popular dimensional reduction techniques. We designed the exploratory framework using real-life road traffic accident data from the Gauteng province, South Africa (SA) with the use of dimensionality reduction techniques. Further on, we performed classification using Naive Bayes, Logistic regression and k-nearest neighbor. Post-processing is carried out on the processed data and model performance measures namely, accuracy and root-mean-square error (RMSE), are used to evaluate each classifier.

The presentation of the study is organized as follows: Section 2 focuses on the study’s methodology which comprises the experimental settings, dataset, key statistics of the road traffic accident data, measuring model performance and machine learning methods. Section 3 covers the results and discussion, with the conclusion of the paper covered in Section 4.

2. Study methodology
This section of the study focuses on the experimental setting, considered dataset, model performance measure methods, key statistics of the road traffic accident data and the machine learning methods used. During statistical analysis, the road traffic accident dataset was used to evaluate feature significance and correlations.

2.1. Experimental settings
The experiment aims to conduct data exploration and evaluate the impact of different types of data visualization methods and dimensionality reduction techniques on RTA analysis and classification modelling. A key objective of the experiments is to uncover hidden patterns from the dataset by proposing an exploratory framework. The implementation of the described methodology was done in Python environment. Experiments were conducted using 3 classifiers, 2 imputation methods, 2 dimensionality reduction methods and 2 evaluation methods as captured in Table 1. The classifiers and evaluation methods were selected due to their widespread usage and well-known advantages. The dimensionality reduction techniques are also widely used in many applications but are selected here to explore suitability and applicability for use on road traffic accident data.

2.1.1. Experimental framework
Figure 1 represents the experimental framework followed in this work to design the proposed exploratory framework. The framework contains three main stages with feature transformation, data visualisation and feature reduction being the sub-process involved during pre-processing stage 1 of the framework, stage 2 being the classification stage and lastly with performance evaluation stage which is mainly used to access the performance of the classifier. With stage 1 being the main focus of the study was data graphical representation and dimensionality reduction being the main concepts.

| Table 1. Methods applied during experiments |
|--------------------------------------------|
| **Methods**                                |
| Imputation methods                        | Mean                                      |
|                                            | k-nearest neighbor                        |
| Dimensionality reduction                  | Principal component analysis              |
|                                            | Linear discriminant analysis              |
| Classifiers                                | Naive Bayes                               |
|                                            | Logistic regression                       |
|                                            | k-nearest neighbor                        |
| Metrics                                   | Accuracy                                  |
|                                            | Root means square error                   |
2.2. **Road traffic accident dataset**

The study employs an actual dataset obtained from the Gauteng Department of Community Safety (GDCS) in South Africa (Gauteng Department: Community Safety, 2012–2020). The road accident dataset consists of the department’s recordings during the period from 2012 to 2019. The features of the dataset are shown in Table 2. The raw dataset was obtained in a Microsoft Excel Spreadsheet format from the department. Preparation of the dataset for the study included rectification of some inconsistencies such as incomplete string values, duplication of the same data fields in different forms and other errors, in the original raw data. The data were also converted to numeric values using the Statistical Package for the Social Sciences (SPSS) (Miller, 2017).

| #  | Features             | Description                                                                 |
|----|----------------------|-----------------------------------------------------------------------------|
| 1  | InvolvedVehicles     | Refers to the number of vehicles that were involved in the road accident.   |
| 2  | TotalNoVictims       | Refers to the number of people/individuals that lost their lives during a specific road accident. |
| 3  | Gender               | Refers to the different gender involved during a traffic accident such as F, M and Others. |
| 4  | Location             | Refers to the place where an incident/road accident occurred.               |
| 5  | Season               | Refers to one of the four seasons: Spring, Summer, Winter and Autumn.     |
| 6  | Month                | Refers to the different number of months in a year with peak and off-peak RTAs. |
| 7  | DayOfWeek            | Refers to the days of a week, this covers the number of RTAs for weekdays and weekends. |
| 8  | CrashDate            | Refers to the date on which an RTA occurred.                               |

Table 3 shows accident type classes for the RTAs dataset. The top three classes are *Pedestrians*, *Overturned* and *unknown* with a percentage of 32.57, 12.84 and 9.17, respectively. This means that from the dataset used for the study, the *Pedestrian* class is most commonly occurring as compared to the other classes. This means the event contributes more to the overall high number of incidents.

2.3. **Measuring model performance**

The study considered the accuracy and root mean square error (RMSE) measures to evaluate performance. These methods are commonly used to evaluate the performance of the constructed models (Chai & Draxler, 2014; Lu et al., 2019; Nguyen et al., 2017), and were chosen due to their popularity from the literature studies. The metrics are computed by (1) and (2).
Accuracy = \frac{TP}{TP + FN} \times 100\% # (1)

In (1), TP represents the true positive predictions of the constructed classifier, and TN the true negative, FN, the false negative, and FP the false-positive predictions.

\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (X_{\text{obs,}i} - X_{\text{model,}i})^2}{n}} (2)

In (2) X_{\text{model}} are the modelled values and X_{\text{obs}} refers to the observation values at sample i with n size as observed set.

2.4. Key statistics of road traffic accident data

In this section of the paper, a statistical analysis of the dataset is presented using different types of charts to review data patterns.

Figure 2 shows a summary of RTAs (formulated using data presented in Tables 2 and 3) in terms of the total number of road traffic accidents yearly. The pie chart shows that the high percentage of road traffic accidents was recorded in 2012, followed by RTAs recorded in 2017. A decline in road traffic accidents is observed during the period from 2015 to March 2019. The overall results indicated that the road traffic accidents recorded in 2012, 2017 and 2015 contributed more to the overall number of road traffic accidents in the study period.

Table 3. Dataset class value distribution

| # | Events          | Values | %   |
|---|-----------------|--------|-----|
| 1 | Pedestrian      | 71     | 32.57 |
| 2 | Overturned      | 28     | 12.84 |
| 3 | Unknown         | 20     | 9.17  |
| 4 | Fixed Object    | 19     | 8.72  |
| 5 | Head-On         | 16     | 7.34  |
| 6 | Head-Rear       | 15     | 6.88  |
| 7 | Collision       | 11     | 5.05  |
| 8 | Sideswipe       | 10     | 4.59  |
| 9 | Rear-End        | 9      | 4.13  |
| 10| Multiple vehicles| 7      | 3.21  |
| 11| Hit and run     | 6      | 2.75  |
| 12| Motorcycle      | 6      | 2.75  |
| Total|             | 218    | 100  |

Figure 2. Summary of RTAs yearly.
Figure 3 shows different summary statistics charts for the road traffic accidents dataset. In Figure 3(a) data distribution of Gender is represented by Other, Female and Male. From the chart, it was observed that the Male variable contributed more to the overall datasets when compared to the others. This means that men are most likely to die in road accidents when compared to women. Furthermore, the authorities can make mandatory the awareness campaigns for road safety to educate all the drivers and try mitigate the high number of road accidents which result in to people losing their lives.

Figure 3(b) shows RTA data distributed monthly, and it is observed that month September represents most of the road traffic accidents dataset. This means that during the September month most of RTAs were recorded as compared to other months. This analysis shows that month November follows as the second most likely month for RTAs with January having the least number of recorded incidents during the study period. Months September and November contribute more to the high number of accidents since is during rainy seasons, public holidays and unforeseen factors could have added to them having such high numbers of incidents. From this analysis authorities can be able to plan better with the knowledge that month September and November are the once with high number of road accidents.

Figure 3(c) Location data distribution shows Wierdabrug as the location with the highest number of RTAs recorded, followed by Midrand. The rest of the locations shows a lower number of RTAs recorded as compared to areas such as Wierdabrug and Midrand. This could be due to the areas being nearby busy roads in the province such as national routes N1 and N14. More so, the areas could be experiencing a high number of road traffic congestion which could be some of the contributing factors to the high number of road accidents. The raw data revealed that Wierdabrug has the highest number of incidents, authors recommend that high visibility of police offices, installation of speed cameras and around areas with high population authorities consider installation humps to prevent vehicles from speeding.

Figure 3(d) shows Season data distribution for all the four seasons represented by Autumn, Spring, Summer and Winter. From this graph, it is found that Spring contributes most to the RTAs in the dataset, that is, the highest number of RTAs that occurred during the Spring season across the study period. This is followed by Winter season and on the other hand, a relatively lower number of RTAs records during Autumn and Summer seasons. Overall, during Summer times the lowest number of road traffic accidents were recorded when compared to the other seasons for dataset obtained from 2012 to March 2019. In overall, authors recommend that visibility of police on the roads during such times can be of value.

![Figure 3. Data distribution (a) Gender, (b) Month, (c) Location and (d) Season.](image-url)
Figure 4 shows summary statistics of DayOfWeek and EVENT histograms. Figure 4(a) DayOfWeek is analogous to a normal distribution. The x-axis on the graph represents day of the week as follows: Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday. The graph shows that the highest number of road traffic incidents occurs on Sundays, followed by Saturday. This means a lot of road accidents happen during weekends and this could be due to drink and drive. A lower number of road accidents are only expected during the week on Monday, Tuesday, Wednesday, Thursday and Friday. This result is contrary to the intuitive expectation that a higher number of vehicles on the roads during the working days of the week are more likely to yield more RTAs. However, these data suggest that other factors during the weekends contribute to the higher number of road traffic accidents. The number of road accidents are less during weekdays compared to weekends, during the week most people are travelling to different work places. It can be seen that the number of road accidents are low during weekdays reason being the drivers are familiar with the roads and that plays an important part when compared to weekends when someone is travelling on an unknown road in high speed. More safety measure can be considered during weekends.

Furthermore, Figure 4(b) data distribution by EVENTS is represented by the x-axis values Collision, Fixed Object, Head-On, Head-Rear, Hit and run, Motorcycle, Multiple vehicles, Overturned, Pedestrian, Rear-End and Sideswipe. It is observed that the Pedestrian event significantly contributes to the number of road accidents. This high number of fatalities is due to people that are walking or crossing on unauthorised side of the routes and not wearing reflective materials/clothes. This is followed by Overturned road traffic incidents and for Fixed Object event, there are relatively lower numbers with Motorcycle events contributing the least. In addition, the authorities should consider spending more on pedestrian educational awareness campaigns and teaching drivers about pedestrian crossing signs. From our analysis it was observed that pedestrian safety is crucial.

Figure 5 shows the numerical feature data distribution of TotalNoVictims and InvolvedVehicles grouped by EVENT using box plots. In Figure 5, it is observed that the distribution in 5(a)
TotalNoVictims has a larger range on event Head-On with the bigger size when compared to the others with less total number of victims. EVENT Multiple Vehicles and followed by EVENT Sideswipe. In 5(b) is observed that Overturned event covers most of the Involvedvehicles followed by Head-On and Collision events. This means that Overturned, Head-On and Collision covered a large number of vehicles that were involved during the incident when compared to the other events. For example, with Overturned event involved vehicles went up to five during different dates, time and year.

2.5. Machine and statistical learning methods
The methods used in this investigation include Principal component analysis, Linear discriminant analysis, Naïve Bayes, Logistic regression, k-nearest neighbor. PCA and LDA enable dimensionality reduction by reducing the number of features. Authors (Bro & Smilde, 2014), principal component analysis was used to reduce the dimensions from the overall number of 14 features. Resulting principal components can then be explored to determine which components comprise the highest variance, indicating the significance of that principal component. Thus, in (Bro & Smilde, 2014), it was found that the first two principal components represent 70% of the variance and only these two components were kept. In this study, the first two principal components were considered to represent 34%, which could be one of the reasons why we obtained the lowest accuracy with PCA technique. Then, it was found that the first three principal components represent 48% of the dataset, which slightly changed the performance of the models, then further PC8. Researchers such as (Eleyan & Demirel, 2006; Jain & Salau, 2019) used LDA and PCA techniques, which was observed that LDA outperformed PCA in many performed tasks. In this study, LDA technique was applied to the RTAs dataset with number of discriminants set to 2, 3 and 8. The LDA technique proved to have greatly increased the model’s performance when LD was set as 2 with 51% accuracy when compared to PCA models.

The study, later on, employed classifiers like Naïve Bayes, Logistics regression and k-nearest neighbor with their default settings. The classifiers were considered due to the following capabilities: the NB (prior = True) which was estimated to work well with a small dataset during model construction (Mukherjee & Sharma, 2012), LR (solver = “lbfgs” and random_state = 0), which have demonstrated success in some different real-world applications such as in image processing and anomalies (Fouad et al., 2015) and the k-NN(k = 5 and p = 2), a classifier that can be regarded to be simple, with some costly prediction times and with the fact that it revisits the entire training dataset (Fouad et al., 2015; Gou et al., 2012; Kuang et al., 2019). This work was set as a baseline to observe the performance of the classifiers with default settings. In this study, the classifiers presented worst results. In addition, the application of the dimensionality reduction techniques presented some promising results. For Naïve Bayes, Logistic regression and k-nearest neighbor in particular further parameter adjustment could have improved the results for this study.

3. Results and discussion

3.1. Analysis of road traffic accident data using PCA
Principal component analysis (PCA) is a well-known statistical method that can be used to reduce dimensions of a dataset to improve understanding and enable the graphical representation of the dataset (James et al., 2013; Martin-Martin et al., 2017; Yong-dong et al., 2019). The variance results shown in explained below were obtained when the number of components was set to “None” to be able to observe which PC contributes more information.

The dataset shows principal components produced in the order of the contribution to the variance in the dataset. It was observed that PC1 covers the most variance, PC2 is the second most leading to the PC8 which captures the lowest variance. All of the PC's contribute some information to the dataset and features. If any PCs are left out that means, an amount of information gets lost. Overall, the results show that the first two to three PCs capture most of the dataset i.e. the MEAN imputation method: PC1-17%, PC2-17% and PC3-14% and k-NN imputation method: PC1-19%, PC2-15% and PC3-14% both datasets explain 48%.
Table 4 shows the most important original features in reducing the dataset. Gender is the most important feature when compared to others such as, Season feature leading to the last feature which has the same most important feature (James et al., 2013). The most important features are regarded as the ones that influence the components more when compared to the rest. This means the features in Table 4 are important to this problem and are captured according to their maximum information concerning the dataset.

3.2. Comparison of dimensionality reduction techniques

In this section, a comparative analysis of the results of the Principal component analysis and Linear discriminant analysis techniques are discussed. Both PCA and LDA plots were constructed using the MEAN and k-NN imputation methods. The number of PCs and LDs were set to 2 for both the 2D graphs in Figures 5-8. Original class labelling naming 1-Collision, 2-Fixed Object, 3-Head-On, 4-Head-Rear, 5-Hit and run, 6-Motorcycle, 7-Multiple vehicles, 8-OVERTURNED, 9-Pedestrian, 10-Rear-End and 11-Sideswipe.

Figures 6 and 7 show scatter plots for PC1 and PC2 principal components against the 11 classes. The resulting plots show the PCA distributions of the dataset. Figure 6 was captured using the MEAN imputation method dataset. In Figure 6(c), it is observed that the data for classes 7-Multiple vehicles, 8-OVERTURNED and 9-Pedestrian are distributed and orientated toward PC2, in which most of the variance thereof is captured. This means that most of the information of these classes in the dataset are contained in PC2 with the Pedestrian class contributing more when compared to the other events. In (b) and (d) less data distribution is observed with the distribution oriented towards PC1 in both plots.

In Figure 7, it is observed that data for classes 7, 8 and 9 are orientated toward the positive PC1 with few outliers from class 7-Multi vehicles. Figure 7(a,c) data are scattered among PC1. This figure was constructed using k-NN imputation method dataset. About (a) plot showing more data distribution when compared to (c).

Figures 8 and 9 show scatter plots for LD1 and LD2 discriminant functions against the 11 classes/dependent variables. Results show that LDA clusters all classes. Linear discriminant analysis is well known as a classification method for predicting categories and is mostly used as a dimensionality reduction technique in data science (James et al., 2013). In Figure 8(c), the plots for classes 7, 8 and 9 projects further away from LD2 to LD1, with a smaller proportion of class 7 data distributed in LD2. Overall, the discriminant function plots in Figure 8 shows that class 2 in (b) is separated from classes 1 and 3.
In Figure 9(c), the LDs show classes 8 and 9 clustered together in LD2, with class 7 scattered across negative LD1. Furthermore, Figure 9(b,d) exhibit distributed classes among LD1 and LD2, and in Figure 9(b), the data points are orientated towards negative-valued LDs with some outliers from class 5- Hit and run.
3.3. Model results

Table 5 and Figure 10 show the tabulated and graphical results obtained during model design using the RTAs dataset shown in Tables 2 and 3. The results show the performance of the applied classifiers based on the two imputation methods. The following is found from the results.

Overall obtained results were poor across different classifiers applied for the study, particularly when dimensionality reduction was not applied to the data. However, overall when LDA was applied to the dataset performed much better in terms of performance when the k-NN imputation method dataset was applied across the three classifiers. Also, LDA, k-NN imputation dataset on Naïve Bayes classifier performance was promising when compared to the other classifiers. The reason for LDA’s to perform well is that it uses both the features information and considers the dependent variables.
In terms of PCA overall did not perform well when compared to LDA, only promising results were obtained by LR for both RMSE and performance with the k-NN imputation method dataset and when the number of components was set to 3. The results revealed that LDA dimensional reduction techniques have a good influence on the dataset when compared to the PCA technique. The analysis was expanded by setting PCs and LDs number of components to 2 and 8, with regards to number of components set to 8 the overall results were poor across all results. Then, with regards to number of components set to 2, the overall results show that when dimensionality reduction LDA was applied promising marginal results were observed when compared to previous number of components 3 and 8. In general, an observation was made that the results vary depending on the classifier and the dimensionality reduction technique. More so, the method has proven is data-dependent and the analysis of the proportion of variance is essential in deciding the applicable number of components to utilise.

Table 5. Summary of the model results

| Models Default Settings | Missing value method: Mean | Missing value method: k-NN |
|-------------------------|---------------------------|---------------------------|
|                         | Accuracy1 % | RMSE | Accuracy2 | RMSE |
| #                       | Model/ metrics |         |         |         |
|                         |              |         |         |         |
| 1                       | NB           | 7       | 13.233  | 12      | 15.953 |
| 2                       | LR           | 37      | 15.628  | 42      | 14.465 |
| 3                       | k-NN         | 28      | 20.86   | 37      | 18.744 |

Principal Component Analysis

|                          | Missing value method: Mean | Missing value method: k-NN |
|-------------------------|---------------------------|---------------------------|
|                         | n = 8 | n = 3 | n = 2 |
| #                       | Model/ metrics | Accuracy % | RMSE | Accuracy % | RMSE | Accuracy % | RMSE |
|                         |        |          |      |            |      |            |      |
| 4                       | NB     | 37       | 13.977 | 37         | 15.442 | 37         | 13.419 |
| 5                       | LR     | 37       | 15.628 | 33         | 13.07  | 33         | 11.953 |
| 6                       | k-NN   | 28       | 20.860 | 26         | 18.395 | 23         | 21.558 |

Linear Discriminant Analysis

|                          | Missing value method: Mean | Missing value method: k-NN |
|-------------------------|---------------------------|---------------------------|
|                         | n = 8 | n = 3 | n = 2 |
| #                       | Model/ metrics | Accuracy % | RMSE | Accuracy % | RMSE | Accuracy % | RMSE |
|                         |        |          |      |            |      |            |      |
| 10                      | NB     | 33       | 17.163 | 40         | 16.233 | 37         | 16.581 |
| 11                      | LR     | 40       | 15.512 | 37         | 18.953 | 37         | 18.558 |
| 12                      | k-NN   | 30       | 21.674 | 30         | 24     | 28         | 24.884 |

|                          | Missing value method: Mean | Missing value method: k-NN |
|-------------------------|---------------------------|---------------------------|
|                         | n = 8 | n = 3 | n = 2 |
| #                       | Model/ metrics | Accuracy % | RMSE | Accuracy % | RMSE | Accuracy % | RMSE |
|                         |        |          |      |            |      |            |      |
| 13                      | NB     | 47       | 11.558 | 49         | 14.047 | 51         | 13.023 |
| 14                      | LR     | 42       | 14.465 | 49         | 10.674 | 49         | 10.977 |
| 15                      | k-NN   | 40       | 18.047 | 40         | 12.093 | 42         | 12.512 |
Further on, an investigation was conducted using One-Vs-Rest (OvR) and One-Vs-One (OvO) strategies, which are mostly used for multi-class classification problems. These strategies were used to split the multi-classes into binary classification per or for each pair of classes. The investigation with OvO and OvR did not show any significant difference.

4. Conclusion
The study aimed to propose an exploratory framework for analysis of road traffic accident data and incorporate dimensionality reduction techniques to reduce the scope of the real-life dataset from Gauteng province. We have observed that the application of EDA can assist in uncovering hidden patterns in datasets and how important is it in data science. Also, an observation of the importance of introducing techniques like PCA and LDA to the road traffic accident data.

However, the overall findings of the study revealed that the NB classifier performed marginally better across all the experiments when LDA dimensional reduction was applied to the k-NN imputation method dataset. Additionally, so, the LR classifier obtained low RMSE value when compared to the rest of the classifiers, which means fewer errors. The authors went further to explore One-vs-Rest (OvR) and One-vs-One (OvO) strategies. However, there were no improvements in the results. This study has demonstrated the following:

1) The proposed framework is beneficial in providing useful insight at the outset relating to patterns from the road traffic accidents dataset.

2) EDA proved to be useful in model selection for this specific dataset, and

3) EDA and dimensional reduction utilised together can provide significantly improved model performance.

In conclusion, this study puts forward an investigation of an extended approach for analysing road traffic accident data, the result of which provides potential usage on other road traffic accident datasets for regions and countries globally.
Acknowledgements
Authors would like to thank the Department of Applied Information Systems, the Institute for Intelligent Systems and the University of Johannesburg. We would also like to thank the Gauteng Department of Community Safety for making the road traffic accident data available.

Funding
The authors did not receive any funding.

Author details
Tebogo Makaba 1
E-mail: tmakaba@uj.ac.za
Wesley Doorsamy 2
E-mail: wdoorsamy@uj.ac.za
Babu Sena Paul
E-mail: bspaul@uj.ac.za

1 Department of Applied Information Systems, University of Johannesburg, Johannesburg, South Africa.
2 Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa.

Declaration of interest
Authors declare that they have no conflict of interest.

Citation information
Cite this article as: Exploratory framework for analysing road traffic accident data with validation on Gauteng province data, Tebogo Makaba, Wesley Doorsamy & Babu Sena Paul,Cogent Engineering (2020), 7: 1834659.

References
Abou Elssad, Z. E., Mousaniff, H., & Al Moattassime, H. (2020). A proactive decision support system for predicting traffic crash events: A critical analysis of imbalanced class distribution. Knowledge-Based Systems, 205, 106314. https://doi.org/10.1016/j.knosys.2020.106314

Ahmadi, A., Jahangiri, A., Berardi, V., & Machiani, S. G. (2020). Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods. Journal of Transportation Safety & Security, 12(4), 522–546. https://doi.org/10.1080/15472450.2018.1505793

Bro, R., & Smilde, A. K. (2014). Principal component analysis. Analytical Methods, 6(9), 2812–2831. https://doi.org/10.1039/C4AY01907J

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chambers, J. M. (2018). Graphical methods for data analysis. CRC Press.

Cuencas, L. G., Puertas, E., Allone, N., & Andres, J. F. (2018, September). Traffic accidents classification and injury severity prediction. In 2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE) (pp. 52–57). IEEE.

De Andrade, L., Vissoci, J. R. N., Rodrigues, C. G., Finato, K., Carvalho, E., Pietrobon, R., de Souza, E. M., Nihei, O. K., Lynch, C., & de Barros Carvalho, M. D. (2014). Brazilian road traffic fatalities: A spatial and environmental analysis. PloS One, 9(1), e87244. https://doi.org/10.1371/journal.pone.0087244

DuToit, S. H., Steyn, A. G. W., & Stumpf, R. H. (2012). Graphical exploratory data analysis. Springer Science & Business Media.

Eleyan, A., & Demirel, H. (2006 September). PCA and LDA based face recognition using feedforward neural network classifier. In International Workshop on Multimedia Content Representation, Classification and Security (pp. 199–206). Berlin, Heidelberg: Springer.

Feng, S., Ke, R., Wang, X., Zhang, Y., & Li, L. (2017). Traffic flow data compression considering burst components. IET Intelligent Transport Systems, 11(9), 572–580. https://doi.org/10.1049/iet-its.2016.0128

Foud, M., Ellatif, M. M. A., Haggag, M., & Akl, A. (2015). Prediction of long term living donor kidney graft outcome: Comparison between rule based decision tree and linear regression. The International Journal of Advanced Computer Research, 3, 185–192. https://www.researchgate.net/profile/Ahmed_Akl/publication/283501232_Prediction_of_Long_Term_Living_Donor_Kidney_Graft_Outcome_Comparison_Between_Rule_Based_Decision_Tree_and_Linear_Regression/links/563b6e09080ed551de8621.pdf

Gauteng Department: Community Safety. (2012–2020). Retrieved September 2, 2020, from www.gauteng.gov.za/government/departments/community-safety

Gazder, U., Ahmed, A., & Shahid, U. (2018, November). Application of artificial neural networks for exploratory analysis of small dataset. In 2018 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (SICT) (pp. 1–6). IEEE.

Gou, J., Du, L., Zhang, Y., & Xiong, T. (2012). A new distance-weighted k-nearest neighbor classifier. Journal of Information Computing Science, 9(6), 1429–1436. https://www.researchgate.net/profile/Jianping_Gou2/publication/26687328_A_New_Distance-weighted_k_nearest_Neighbor_Classifier/links/54510cdf0c2f864cb99af9/A-New-Distance-weighted-k-nearest-Neighbor-Classifier.pdf

Jain, S., & Salau, A. O. (2019). An image feature selection approach for dimensionality reduction based on kNN and SVM for AKT proteins. Cogent Engineering, 6(1), 1599537. https://doi.org/10.1080/23311916.2019.1599537

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). springer.

Kuang, L., Yan, H., Zhu, Y., Tu, S., & Fan, X. (2019). Predicting duration of traffic accidents based on cost-sensitive Bayesian network and weighted K-nearest neighbor. Journal of Intelligent Transportation Systems, 23(2), 161–174. https://doi.org/10.1007/s10544-018-15369-78

Lavrac, N., Jesenovec, D., Trdin, N., & Kostal, N. M. (2008). Mining spatio-temporal data of traffic accidents and spatial pattern visualization. Metodoloski zvezki, 31(1–4). https://ibimf.uni-lj.si/m3/2008/no-1/lavrac.pdf

Lu, H., Xu, Y., Ye, M., Yan, K., Gao, Z., & Jin, Q. (2019). Learning misclassification costs for imbalanced classification on gene expression data. BMC Bioinformatics, 20(25), 1–10. https://doi.org/10.1186/s12859-019-3255-x

Martinez, W. L., Martinez, A. R., Solko, J., & Martinez, A. (2010). Exploratory data analysis with MATLAB. Crc Press.

Martín-Marin, P., González-Briones, A., Villarrubia, G., & De Paz, J. F. (2017, June). Intelligent transport system through the recognition of elements in the environment. In International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 470–480), Cham: Springer.

Michalaki, P., Quddus, M. A., Pitfield, D., & Huetson, A. (2015). Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model. Journal of Safety Research, 55, 89–97. https://reader.elsevier.com/reader/sd/pii/S0022347715000836?token=CA9254AA2FB23CAFBB9AFE93E4903AA9A5CD892
Muguro, J. K., Sasaki, M., Matsushita, K., & Njeri, W. (2020). Trend analysis and fatality causes in Kenyan roads: A review of road traffic accident data between 2015 and 2020. Cogent Engineering, 7(1), 1797981. https://www.tandfonline.com/doi/pdf/10.1080/23311916.2020.1797981

Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119–128. https://doi.org/10.1016/j.protcy.2012.05.017

Nguyen, H., Cai, L., & Chen, F. (2017). Automatic classification of traffic incident’s severity using machine learning approaches. IET Intelligent Transport Systems, 11(10), 615–621. https://doi.org/10.1049/iet-its.2017.0051

Rende, Y., Zhang, Q., Zhang, X., & Huo, L. (2013). Traffic accidents forecasting based on neural network and principal component analysis. Research Journal of Applied Sciences Engineering & Technology, 6(6), 1065–1073. https://doi.org/10.19026/rjaset.6.4014

Shbeeb, L. I., & Awad, W. E. H. (2016). Road traffic safety perception in Jordan. Cogent Engineering, 3(1), 1127748. https://doi.org/10.1080/23311916.2015.1127748

Timmermans, C., Alhajyaseen, W., Al Mamun, A., Wakjira, T., Qasem, M., Almollah, M., & Younis, H. (2019). Analysis of road traffic crashes in the State of Qatar. International Journal of Injury Control and Safety Promotion, 26(3), 242–250. https://doi.org/10.1080/17457300.2019.1620289

Wells-Parker, E., Ceminsky, J., Hallberg, V., Snow, R. W., Dunaway, G., Guiling, S., Williams, M., & Anderson, B. (2002). An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers. Accident Analysis & Prevention, 34(3), 271–278. https://doi.org/10.1016/S0001-4575(01)00021-5

World Health Organization. (2018). Global status report on road safety 2018: Summary (No. WHO/NMH/NVI/18.20). Yong-dong, W., Dong-wei, X., Peng, P., Yi, L., Gui-jun, Z., & Xiao, X. (2019). Kernel PCA for road traffic data non-linear feature extraction. IET Intelligent Transport Systems, 13(8), 1291–1298. https://doi.org/10.1049/iet-its.2018.5215