UNIVERSITI TEKNIKAL MALAYSIA MELAKA

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY (FTMK)

BACHELOR OF COMPUTER SCIENCE
(ARTIFICIAL INTELLIGENCE)

BITI 2513
INTRO TO DATA SCIENCE

PROJECT TITLE:
ANALYSIS AND PREDICTION PROJECTION OF ROAD ACCIDENTS ACCORDING TO ACCIDENT AND INJURY TYPE IN MALAYSIA

TEAM MEMBERS:

NO.	NAME	MATRIC NO.
1	NATASHA AMIERA BINTI AZMAN	B032120013
2	NURUL HAFIKA HAFINA BINTI MOHAMAD FADZLI	B032120014
3	NURFATIN NABILAH BINTI ABD AZIZ	B032120080

Analysis and prediction projection of road accidents according to accident and injury type in Malaysia

ABSTRACT

This paper presents an analysis of road accidents in Malaysia over a ten-year period, focusing on accident and injury types. Statistical models and machine learning techniques were used to identify trends and patterns in accident occurrence, severity, and injury type, and to predict future accident rates and injury types based on historical data. The study reveals that driver behaviour, road conditions, and vehicle safety are significant factors contributing to road accidents in Malaysia. The findings can be used to develop targeted interventions and policies to reduce the incidence of road accidents, save lives, and reduce the economic impact of road accidents in Malaysia.

Introduction

Road accidents continue to be a significant public health concern worldwide, causing loss of life and property, as well as serious injury and disability. Malaysia, like many other developing countries, faces the challenge of reducing the incidence of road accidents and improving road safety. To achieve this, it is crucial to have a better understanding of the factors that contribute to accidents and the types of injuries that result from them.

This paper presents an in-depth analysis of road accidents in Malaysia, focusing on the different types of accidents and injuries that occur. Using data released by Royal Malaysia Police (Polis Diraja Malaysia (PDRM)) gathered from data.gov.my which is an open data portal provided under the government of Malaysia, we have analysed accident statistics over a ten-year period to identify trends and patterns in accident occurrence, severity, and type of injury.

The study aims to identify the factors that contribute to road accidents in Malaysia, such as driver behaviour, road conditions, and vehicle safety, and to predict future accident rates based on these factors. We use statistical models and machine learning techniques to develop predictive models that can forecast accident rates and injury types based on historical data.

The findings of this study can be used to develop targeted interventions and policies to reduce the incidence of road accidents in Malaysia. By identifying the most significant risk factors for accidents and predicting future trends, we can develop effective prevention strategies that can help to save lives and reduce the economic impact of road accidents.

Data acquisition

The first step in acquiring the dataset is to identify the source. In this case, the dataset was released by the Royal Malaysia Police (Polis Diraja Malaysia) and is available on data.gov.my, a public data portal provided by the Malaysian government. The dataset was created by Mohd Hafizi bin Azmi on 28 April 2017 and updated on 8 September 2021. This dataset has accident statistics for ten years, from 2011 until 2021. Knowing the source of the data is important because it helps establish the credibility and reliability of the dataset.

Once the dataset is identified, the next step is to assess the source. This involves reviewing the dataset to ensure it contains the information needed to identify trends and patterns in accident occurrence, severity, and type of injury. Depending on the format of the dataset, has to clean and format the data to make it usable for analysis. For example, has to remove duplicate entries or missing data, convert data types, or create new columns. This step is important to ensure that the data is consistent and accurate.

With the dataset cleaned and formatted, import it into the analysis tool of choice, such as Google Colab. Once the dataset is imported, it can begin analysing trends and patterns in accident occurrence, severity, and type of injury.

Finally, this step is important to ensure that the insights gained from the analysis are used to make positive changes and improve road safety in the country.

Data Exploration

The dataset provides statistical information on road accidents and injuries in Malaysia, categorised by accident type and severity. It contains data for the year 2011 until 2021 and includes 10 columns of information with each column representing a different attribute of the dataset. However, data from the year 2016 cannot be retrieved since it is not available therefore it is not included in the analysis.

Table 1: First 10 rows from the data set used

						JUMLAH			
			JUMLAH	JUMLAH	JUMLAH	KMLG		JUMLAH	JUMLAH
		JUMLAH	KMLG	KMLG	KMLG	ROSAK	JUMLAH	CEDERA	CEDERA
TAHUN	NEGERI	KMLG	MAUT	PARAH	RINGAN	SAHAJA	KEMATIAN	PARAH	RINGAN
2011	PERLIS	1791	73	231	246	1241	79	259	288
2011	KEDAH	19699	506	499	1023	17671	515	608	1329
2011	PULAU PINANG	37158	375	160	320	36303	392	197	380
2011	PERAK	33506	739	700	1267	30800	811	898	1631
2011	SELANGOR	128876	1015	457	689	126715	1070	566	807
2011	KUALA LUMPUR	58795	230	83	497	57985	236	83	498
2011	NEGERI SEMBILAN	21157	343	444	945	19425	374	568	1221
2011	MELAKA	14720	224	147	382	13967	240	195	509
2011	JOHOR	59501	1001	366	1007	57127	1073	492	1254

The first two columns of the dataset include the year of the accident and the state in which the accident occurred. The remaining columns contain information about the type of accident, the number of accidents, and the number of injuries, categorised by accident type and severity.

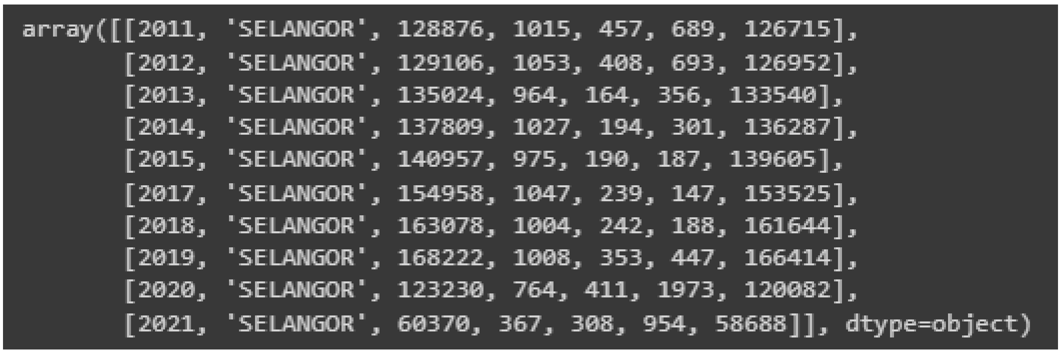
The severity of the injuries is classified into three categories: fatal, serious and minor injury. The dataset provides information on the number of fatalities and injuries for each accident type and severity category. Finally, the data reveals that the most common injury sustained in road accidents is minor injuries, followed by severe injuries and fatalities.

The dataset allows us to analyse road safety trends and patterns in Malaysia over the 10-year period covered by the data. By examining the data, we can see that the total number of accidents and injuries has increased over the years, with the highest number of accidents occurring in Selangor and the highest number of fatalities in Johor. The dataset also includes information on the number of accidents and injuries by year. This allows us to see if there are any seasonal patterns or trends in road accidents in Malaysia.

Overall, the dataset provides valuable information on road safety trends and challenges in Malaysia and can be used to inform policy decisions and initiatives aimed at reducing accidents and injuries on the roads.

Designing the predictive model

Our team decided on using ‘Google Colab’ for the training, testing and implementation of this model as Google Colab is a free to use online notebook environment similar to Jupyter Notebook. For the prediction model selection, we tested using three different algorithms which are support vector machines (SVM), or specifically support vector regression (SVR), neural network and linear regression. We chose these algorithms as they are algorithms useful for detecting patterns and predictions. During the training and testing of this model, we decided to use the columns ‘Year’, ‘Total Accident’, ‘Total death accident’, ‘Total critical accident’, ‘Total minor accident’ and ‘Total damage only accident’ for the independent variables and the dependent variable or item we chose to predict is the ‘Total death’ accidents. We also decided to narrow down the scope of data to only take the data for Selangor cases only to see the trend of cases. Figure 1 shows a snippet of the array produced when choosing the columns for the independent variable.



```
array([[2011, 'SELANGOR', 128876, 1015, 457, 689, 126715],
       [2012, 'SELANGOR', 129106, 1053, 408, 693, 126952],
       [2013, 'SELANGOR', 135024, 964, 164, 356, 133540],
       [2014, 'SELANGOR', 137809, 1027, 194, 301, 136287],
       [2015, 'SELANGOR', 140957, 975, 190, 187, 139605],
       [2017, 'SELANGOR', 154958, 1047, 239, 147, 153525],
       [2018, 'SELANGOR', 163078, 1004, 242, 188, 161644],
       [2019, 'SELANGOR', 168222, 1008, 353, 447, 166414],
       [2020, 'SELANGOR', 123230, 764, 411, 1973, 120082],
       [2021, 'SELANGOR', 60370, 367, 308, 954, 58688]], dtype=object)
```

Figure 1: Snippet of the array produced when choosing columns for independent variables

Modelling using Support Vector Machine Algorithm

This algorithm works based on a subset of data from our training batch as it does not take the training points that are outside the assigned margin into consideration. This means that there is a possibility of a high margin of error in the calculation when the wrong weight and margin is used. Figure _ shows the formula for the prediction algorithm when using SVR.

Figure 2 provides the template function that a prediction algorithm using SVR follows. Through the scikit-learn(sklearn) library, the algorithm can be accessed by importing svm and calling the SVR function.

$$\sum_{i \in SV} (\alpha_i - \alpha_i^*) K(x_i, x) + b$$

Figure 2: Template function of a support vector regression (SVR) algorithm

Based on the function, the αi - αi* holds the difference which can be acquired from the ‘dual_coef_’ attribute, K(xi, x) holds the support vectors which is acquired through the ‘support_vectors_’ attribute and b holds the interception point of each line to the y-axis which is accessed through the ‘intercept_’ variable.

Modelling using Neural Network

This algorithm works by using a multi-layer perceptron regressor (MLP) by using forward and backward propagation in order to find some hidden correlation between the data. Neural networks give a more in-depth view of the relationship of variables and find the more complex patterns within the data provided. Figure 3 defines the basic structure of a MLP neural network along with the structured view of the formula used.

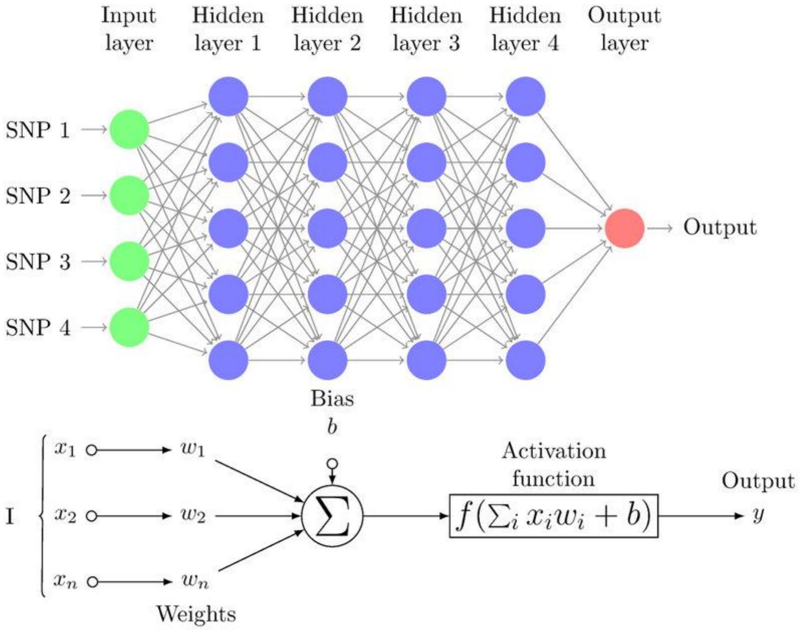


Figure 3: Basic structure of an MLP neural network and structured view of formula

Modelling using Linear Regression

This algorithm works by using a value taken from another variable to predict another variable. The predicted value is known as a dependent variable and the value used to predict the other variable is known as an independent variable. Linear regression reduces the differences between the predicted value and the actual value of a data by fitting a straight line between the values. A best-fit line is decided using the least squares method.

$$y = a_0 + a_1 x + \varepsilon$$

Figure 4: Formula of a linear regression algorithm

,where y represents the variable we wish to predict (dependent variable), x represents the independent variables, a0 is the interception of the line, a1 is the linear regression coefficient and ε represents the random error set for the model.

Result

This section focuses on discussing the results of the models by comparing the predicted value to the actual value and comparing the R-squared score, mean absolute error

(MAE) and root mean squared error (RMSE) of each model. The model with the highest accuracy rate and closest predicted value can be considered the best solution.

Figure 5 shows all the graphs produced after the predicted and actual values for each model is plotted into a graph. The graph on the upper-left side represents the comparison of predicted value to actual value of the result produced by the Linear Regression model, the upper-right graph represents the comparison of predicted value to actual value of the result produced by the neural network model, and the lower-left graph represents the comparison of predicted value to actual value of the result produced by the SVR model. We can clearly identify that the result produced by the linear regression model gives us the closest predicted value to actual value compared to the neural network model and SVR model. Table 1 shows us a tabulated view of the comparison of actual and predicted value, RMSE, MAE and R-squared score.

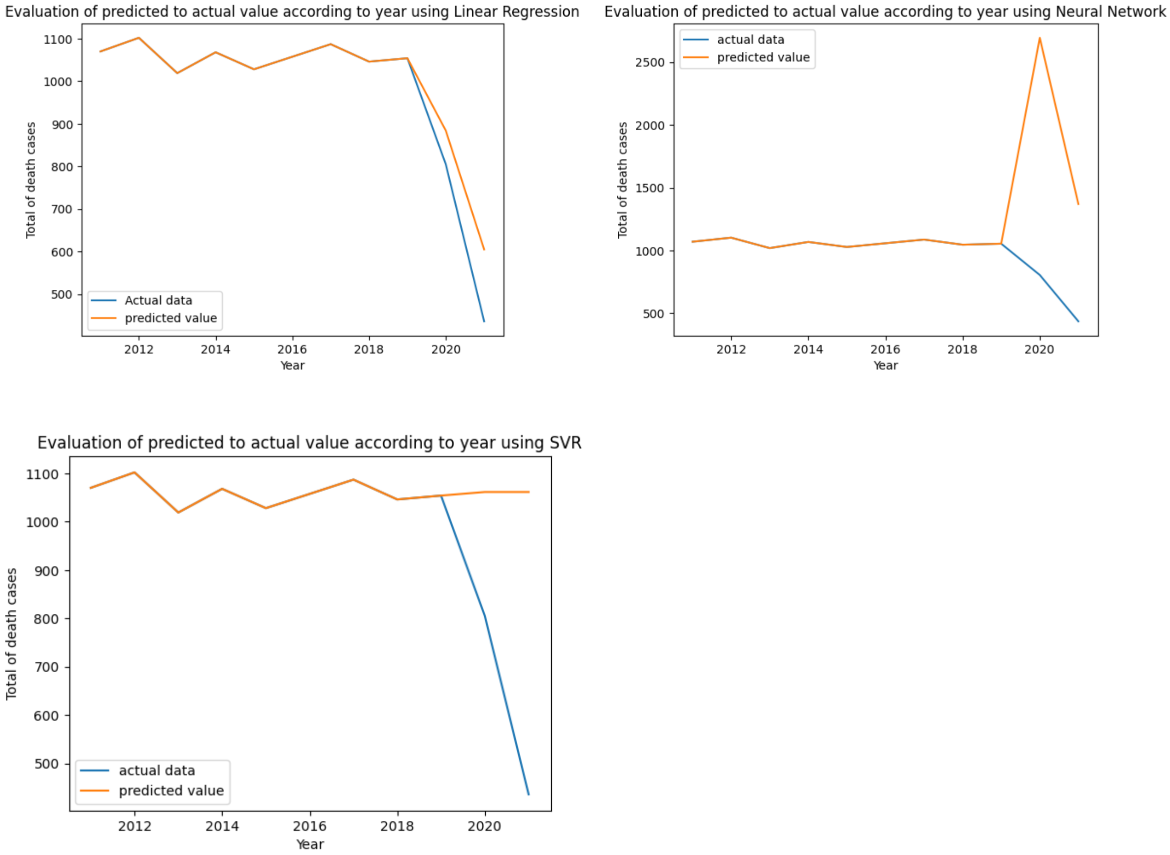


Figure 5: Graph of comparison on predicted value to actual value of each model

Table 2 shows us a tabulated view of the comparison of actual and predicted value, RMSE, MAE and R-squared score. According to table 1, we can see a more accurate representation of the actual and predicted value. As pertained by the tabulated data, we can see that the predicted value produced by the linear regression is much closer in comparison to the neural network and SVR model.

The R-squared score is calculated for each model produced to see the quality of each model that has been trained and tested. The table below shows a comparison of the resulting R-squared score for each model along with the predicted and actual value comparison.

Table 2: Comparison of R-squared score, RMSE, MAE and predicted value resulting from each model and the actual value.

Model	R-squared score (%)	RMSE	MAE	Value		
				Year	Actual	Predicted
Linear Regression	49.0%	124.02	11.14	2020	805	884.09976209
				2021	436	604.93121502
Support Vector Regression	-571.0%	440.92	21.0	2020	805	1061.42111082
				2021	436	1061.42111082
Neural Network	-6416.0%	1410.77	37.56	2020	805	2692.92807491
				2021	436	1369.60672805

As we can see the linear regression model is the only model to produce a positive r-squared score with a score of 49% compared to SVR which has a score of -571% and neural network with a score of -6416%. This means that while the accuracy is lower than 50%, linear regression algorithm is still the best choice for this prediction set. Mean Absolute Error (MAE) is used to see the precision of prediction. A large MAE value means that the model has poor prediction projection. All of the models produced give out quite a large MAE value. However, once again, compared to SVR which has a value of 440.92 and neural network with a value of 1410.77, the linear regression model still proves to give us a lower MAE value which stands at 124.02. The root mean squared error (RMSE) is a metric that is used to detect the error rate of result in the model and is compared to the MAE to see the margin of error. All the models created produce a higher RMSE compared to MAE which means the dataset used actually contains a rather large error. While the model produced does not give a proper conclusion that can tell us it can be used for predicting future accident cases. This research paper shows us that when proceeding with this research, the linear progression algorithm would most probably be the best solution into modelling a prediction model for road accidents in Malaysia.

Conclusion

In conclusion, this paper investigated the effectiveness of three different methods for predicting vehicle accidents: linear regression, neural network, and support vector regression. After careful analysis and comparison of the results, it can be concluded that linear regression is the best method of choice for predicting vehicle accidents.

The neural network and support vector regression showed quite inaccurate results possibly occurring due to the inconsistency of the data set used but the performance of linear regression outperformed them in terms of accuracy and computational efficiency. The linear regression model demonstrated the highest level of predictive ability, with a lower mean squared error and higher R-squared value compared to the other two methods.

Moreover, linear regression is a simple and easy-to-use method, making it an attractive option for predicting vehicle accidents in real-world applications. Overall, the findings of this study highlight the importance of selecting the appropriate method for accident prediction and suggest that linear regression is, while not the most accurate, the most reliable and effective approach for this purpose.

References

1. Azmi, M. H. (2021, September 8). *Statistik Kemalangan Jalan Raya Mengikut Jenis Kemalangan Dan Kecederaan*. STATISTIK KEMALANGAN JALAN RAYA MENGIKUT JENIS KEMALANGAN DAN KECEDERAAN. Retrieved May 7, 2023, from https://www.data.gov.my/data/ms_MY/dataset/statistik-kemalangan-jalan-raya-mengikut-jenis-kemalangan-dan-kecederaan
2. 1.4. Support Vector Machines. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,Effective%20in%20high%20dimensional%20spaces.
3. About Linear Regression | IBM. (n.d.). https://www.ibm.com/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.
4. csiplearninghub. (2023, March 12). AI Project Cycle Class 10 Important Notes. CS-IP-Learning-Hub. https://csiplearninghub.com/ai-project-cycle-class-10-important-notes/#:~:text=What%20is%20AI%20Project%20Cycle,us%20to%20achieve%20our%20goal.
5. GeeksforGeeks. (2022). Convert a NumPy array into a csv file. GeeksforGeeks. https://www.geeksforgeeks.org/convert-a-numpy-array-into-a-csv-file/
6. GeeksforGeeks. (2023). Adding new column to existing DataFrame in Pandas. GeeksforGeeks. https://www.geeksforgeeks.org/adding-new-column-to-existing-dataframe-in-pandas/
7. Google Colaboratory. (n.d.). https://colab.research.google.com/notebooks/charts.ipynb#scrollTo=08RTGn_xE3M
8. Malli. (2023). Pandas Add or Insert Row to DataFrame. Spark by {Examples}. https://sparkbyexamples.com/pandas/pandas-add-row-to-dataframe/#:~:text=By%20using%20append()%20function,with%20the%20newly%20added%20row.
9. Merge, join, concatenate and compare — pandas 2.0.1 documentation. (n.d.). https://pandas.pydata.org/docs/user_guide/merging.html

10. Ogunbiyi, I. A. (2022). Top Evaluation Metrics for Regression Problems in Machine Learning. freeCodeCamp.org. https://www.freecodecamp.org/news/evaluation-metrics-for-regression-problems-machine-learning/#:~:text=A%20regression%20model%20can%20only,residuals%20as%20being%20a%20distance.
11. Pérez-Enciso, & Zingaretti, Laura. (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. Genes. 10. 553. 10.3390/genes10070553.
12. P K, G. M. (2021, December 15). Machine Learning Basics: Support Vector Regression - Towards Data Science. Medium. https://towardsdatascience.com/machine-learning-basics-support-vector-regression-660306ac5226#:~:text=Support%20Vector%20Regression%20is%20similar,to%20plot%20the%20boundary%20line.
13. R, S. (2021). A Walk-through of Regression Analysis Using Artificial Neural Networks in Tensorflow. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/08/a-walk-through-of-regression-analysis-using-artificial-neural-networks-in-tensorflow/
14. Regression Analysis in Machine learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/regression-analysis-in-machine-learning
15. Sharp, T. (2023, April 3). An Introduction to Support Vector Regression (SVR) - Towards Data Science. Medium. https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2
16. sklearn.linear_model.LinearRegression. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
17. sklearn.neural_network.MLPRegressor. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html
18. Yulei. (2020, June 29). Yulei. https://yuleii.github.io/2020/06/27/data-manipulation-with-pandas.html