JKAU: Comp. IT. Sci., **Vol**. **11** No. 2, pp: 53 – 66 (1444 A.H. / 2022 A.D.)
DOI: 10.4197/Comp.11-2.5

Using data analytics and machine learning for road traffic accident prediction and causal factor analysis

Abdullah Hassan Asiri
Department of Information Systems, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia
ahussainasiri@stu.kau.edu.sa

Abstract. Machine learning and data analysis together represent a powerful force in producing analysis and forecasting models which in turn support the growth of different fields. In terms of safety and people, road traffic accidents have a negative impact on individuals, governments, and companies. Therefore, predicting traffic accidents on the road has become very important in the decision-making process that will lead to the safety of others. It is also important to find the causes of these accidents. This study aims to predict road traffic accidents and their causative factors based on a data set from Kaggle [1] and to create a road vehicle accident prediction model using machine learning and decision tree algorithms. The results showed that one of the most important factors causing accidents on the road is the lack of distance between vehicles, and that weather and age play an important role. The model achieved an accuracy of 84.22% and a standard deviation of +/-1.08 % . The tools used in data analysis and machine learning and prediction were Python and RapidMiner an integrated software platform for data science that provides an integrated environment for machine learning, deep learning, and predictive analysis.

Keywords*—data analysis, algorithm, vehicle accidents, machine learning, prediction.***

I. INTRODUCTION

The world of cars is witnessing great developments, and although modern cars are equipped with modern technology and multiple safety features, the number of car accidents is increasing dramatically throughout the world. In a 2018 report, the World Health Organization reported that the number of traffic accident deaths was 1.35 million, of which a large proportion is disproportionately shared among pedestrians, motorcyclists, and cyclists [2].

To preserve human life, it is important to understand the factors causing traffic accidents so that the world will be able to avoid them in future. Among these factors are the following: weather, roads, age, alcohol, and driver behavior. Researchers have investigated the factors causing these accidents, such as: driver behavior, road design, and weather fluctuations [3]. Although these factors are considered major factors, this study did not delve deeply into some other factors, so some other reasons and factors are still not clear. Therefore, this paper aimed to create a model to predict vehicle accidents on the road and to analyze the factors causing these accidents. The data set was obtained from Kaggle [1], in an attempt to collect the largest possible number of factors which can cause accidents such as weather fluctuations, road conditions, driver age, time of day, etc.

In this paper, various techniques were used, starting with using the Python language to clean the data, through to using Rapid Miner for data analysis and using machine learning and the decision tree algorithm for making predictions. The aim was to discover effective predictions and detect the most important factors causing vehicle accidents.

II. RELATED WORKS

Data analytics is the science of analyzing raw data to make conclusions about that information [4]. Data analytics requires the automation of a set of techniques, processes, and some algorithms on raw data to produce information for human use.

Machine learning can be broadly defined as computational methods using experience to improve performance or to make accurate predictions [5]. Experience comes in the form of electronic data representing already existing

information that can be collected, cleaned, and analyzed.

Data analytics is used to avoid some problems and discover solutions or improvements to others. When reliable and clean data are used, along with the various machine learning algorithms, it is possible to obtain reliable information to make predictions. Data analytics and machine learning are used in many fields, but they are used effectively in predicting the severity of road accidents and determining the factors that cause them [6]. Specifically in Bangladesh, a study worked on examining the most important factors causing accidents and the extent of the severity of these accidents. An analysis was conducted using several tools, including Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes, and AdaBoost, to classify accidents and their severity into fatal, moderate, or minor injuries, and the best tool used in this study was AdaBoost, which gave the highest performance.

What we are witnessing in our time of population density, with the requirements of sprawling urban life, is that everyone is using the car to move around and accomplish tasks, so car accidents on the road have increased significantly. There are many factors that cause these accidents. Various studies have been conducted in a variety of ways to search for the most important of factors and their main causes.

Table I. Some studies that used different methods to study and analyze the factors causing road traffic accidents.

Objectives	Method	Results	Limitations	Study
Establishes models to select a set of influential factors and to build up a model for classifying the severity of injuries.	Machine learning algorithms, AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF)	RF algorithm has shown better performance with 75.5% accuracy than LR with 74.5%, NB with 73.1%, and AdaBoost with 74.5% Accuracy.	The study focused on only a certain group of factors that cause traffic accidents. We may need more than one group to study some other factors that may be more important than some of those factors selected. In other words, we need several models.	Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity [7].
To develop a method for novel Frequent Pattern tree (FP tree) based variable selection. The method works by identifying all the frequent patterns in the traffic accident dataset.	FP tree (Bayesian) and the random forest method.	Best model found was an FP tree-based Bayesian network model that can predict 61.11% of accidents while having a false alarm rate % of 38.16.	The method used in this study to predict road traffic accidents depends on the patterns of previous accidents, and the patterns of accidents cannot be counted, or there could be repeats in some patterns, so the model needed to be continuously updated.	A novel variable selection method based on frequent pattern tree for real-time traffic accident risk prediction [8].

Table II. Some studies that used different methods to study and analyze the factors causing road traffic accidents.

Objectives	Method	Results	Limitations	Study
To study the pattern of road traffic accidents, and antecedent factors influencing the road traffic accidents.	Cross-sectional study (Various parameters like age and sex distribution, time of occurrence, alcohol consumption, etc.)	There was a marked male preponderance (88.77%), Most of the accidents had taken place in the evening hours (6 pm - 12 midnight).	A cross-sectional study was used, and as we know that this method is used to collect data for a specific period and in a specific place, and therefore the data and numbers vary from month to month, from year to year, and from place to place. This study has been included to cover technical and non-technical studies and their use.	An epidemiological study of road traffic accident cases at a tertiary care hospital in rural Haryana [9].
This study aims to use a consensual neural network to predict with great accuracy the occurrence of traffic accidents.	Convolution neural network.	The prediction of (TAP-CNN) model accuracy comes to 78.5%, which is 7.7% higher than the model (TAP-BP).	The study focused on different specific sets of accident cases to train the neural network, but as also reported in this paper, it did not explicitly address the causative factors of accidents, including the road environment on which cars are traveling and so on. This method may succeed in predicting traffic accidents, but it will not tell us the causative factors to avoid them in the future	A model of traffic accident prediction based on convolutional neural network [10]

Table III. Some studies that used different methods to study and analyze the factors causing road traffic accidents.

Objectives	Method	Results	Limitations	Study
The aim of this study was the analysis of human behavior or the conditions which can lead to an accident or influenced the cause of an accident.	Qualitative analysis and Quantitative analysis (Pearson s chi-squared test).	The most common contributing factor (or accident cause) was inattention (regardless of the age or gender). The Pearson chi-squared test indicated statistically significant differences between the men and women and between age groups (young drivers under 25, middle age, and seniors up to 65).	This study focused on one factor, which is human behavior, and it was supported by age and gender. It was based on qualitative and quantitative analysis. Therefore, we need to analyze some other factors using better technical methods.	Human factors contributing to the occurrence of road traffic accidents [11].

A study references a major and very important factor, which is the age group, specifically the adolescent stage. This study stated that people between 18-24 years represent 23% of the number of deaths caused by traffic accidents [12]. Research needs to address the other age groups, especially the elderly and people with disabilities, so we need to detect, analyze, and study these other factors.

The main objective [13] is to analyze traffic accidents with the aim of developing models for accurate prediction of collision frequency in Anambra, Nigeria, using auto regressive integrated moving average (ARIMA) and auto regressive integrated moving average with explanatory variables (ARIMAX). The results showed that the ARIMAX model outperformed the ARIMA model. This study's findings reveal that incorporating human, vehicle, and environmental related factors in a time series analysis of the crash dataset produces a more robust predictive model than using the aggregated crash count.

Some of the classification models used in Mohanta's study [14], specifically Logistic Regression, Artificial Neural Network, Decision Tree, K-Nearest Neighbors, and Random Forest have been implemented to predict the severity of accidents. These models have been verified, and the first results prove that these classification models have attained considerable accuracy.

III. Method

In this study, a variety of techniques for analysis and prediction were used, starting with the Python language to clean the data because it is easy to use, accurate and highly flexible. Then Rapid Miner was used to analyze the data and create a prediction model using the decision tree. Rapid Miner was chosen because of its ease of use and fast creation of models and analysis. In this section we will see some more details.

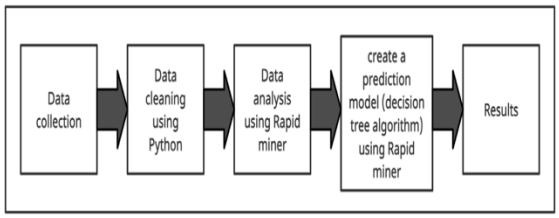

Fig 1. Workflow

A. Data source

The data for car accidents in Addis Ababa for the years 2017-2020 was collected by Shahane [15], obtained from Kaggle [1]. The data consisted of 32 columns and 12,316 rows. The dataset contains many factors that cause traffic accidents on the road, such as the age of the driver, the type of road, the weather condition, the driver's experience, the driver's gender, etc.

- Data description

The descriptive table shows us the count, uniqueness, top entry, and frequency for each column. For example, the time column contains 12,316 rows, the number of unique entries is 1074, the top entry is 15:30 with a frequency of 120, which means that 15:30 is repeated 120 times in this column. The other columns work in the same way. Another example is the driver gender column containing 12,316 rows, the number of unique entries is 3 (males, females and unknown), the top entry is male with a frequency of 11,437.

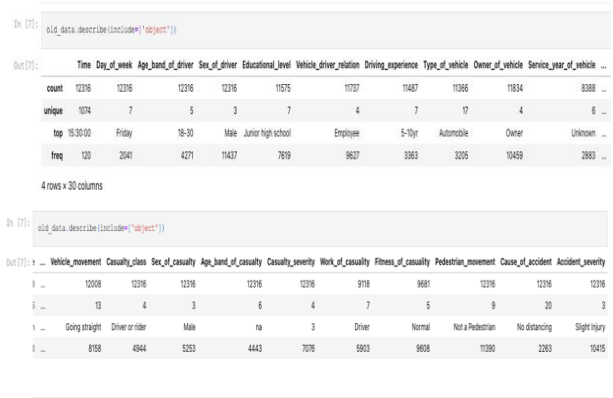

Fig 2. Descriptive table

B. Data cleaning

Data cleaning is a necessary part of the data analysis procedure. As we saw in the data description section, there are many missing data values in several columns and the total number of missing data values in the complete data set is 20,057, and there are also completely missing rows and useless columns. This empty data needs to be removed so that the missing data can become complete and reliable data. In this section, we will discuss in detail the data cleaning process.

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics, a multi-functional, maximally interpreted programming language with several advantages that are often used to streamline massive and complex data sets.

Jupyter notebooks provide an easy-to-use private data science environment that can be used in presentations or educational settings. Jupyter is often used with Python because of its flexibility and efficiency. Python and the Jupyter tool were used to understand and clean the data.

This code allows us to call the Panda library, which is one of the Python libraries that deals with data, and then read the data file via the file path.

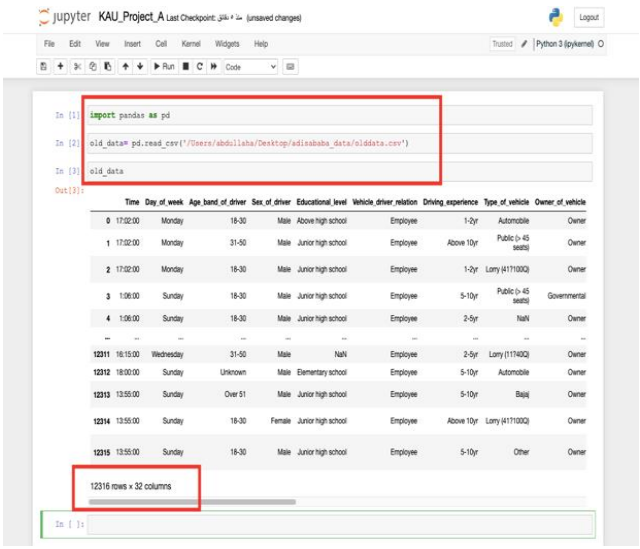

Fig 3. Data read

- Missing data

The missing data may be incomprehensible text or numeric data, and it may be corrupt data or empty fields that distort the data set.

This code shows us each column in the data with the number of missing values for each column.

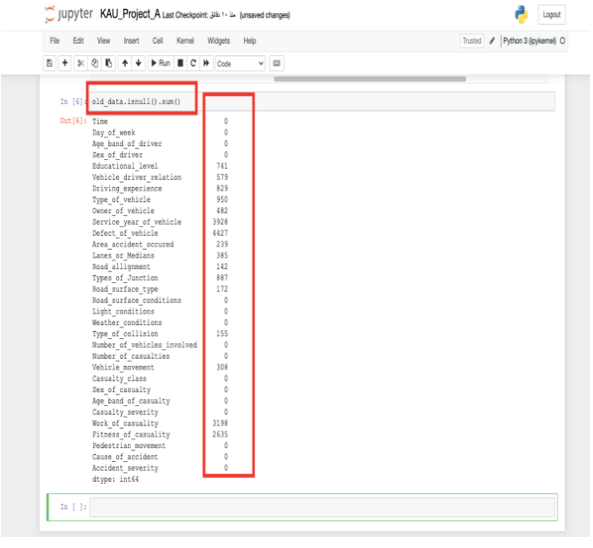

Fig 4. Number of missing data per column

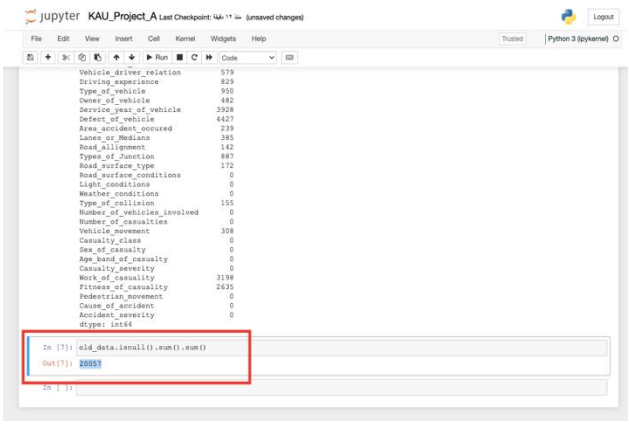

Fig 5. Missing values in all rows and columns.

Seven unwanted columns were removed until we were left with 25 out of 32 columns. The unwanted columns do not add any value to the analysis process and do not help in the search for the causes of accidents. For example, the Fitness_of_casuality column. This column talks about the fitness of casualty, but this does not help the process of analysis and detection of the causes of accidents. Another example, Type_of_vehicle, is a column that talks about the type of vehicle this also does not the help analysis process, etc.

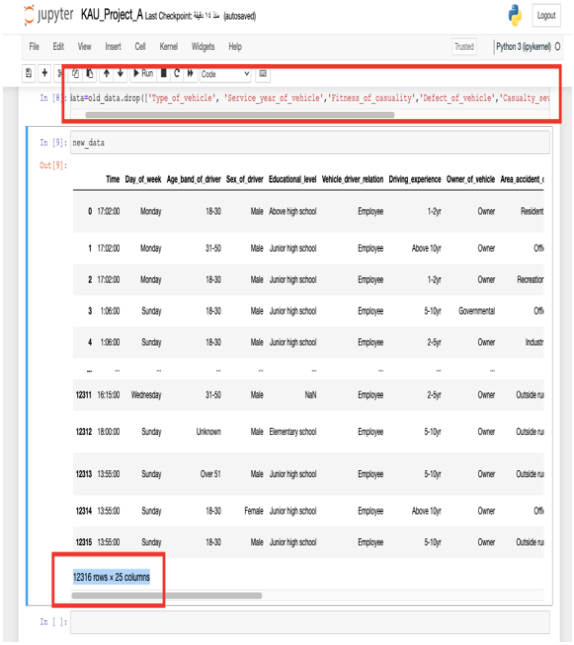

Fig 6. Deleted columns

This code deletes any row that contains at least one missing field.

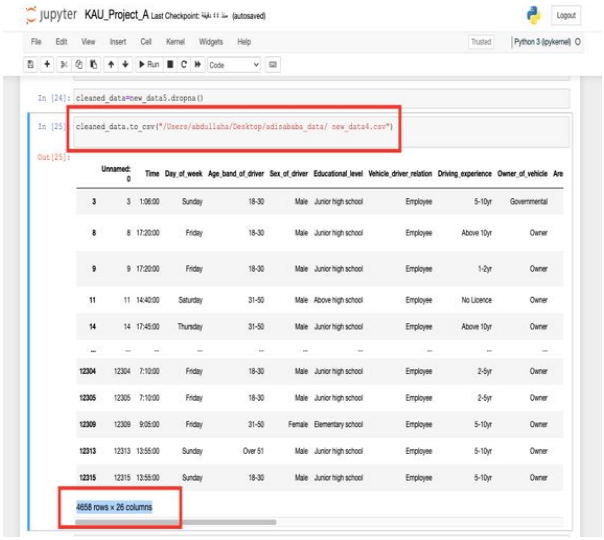

Fig 7. Remove missing values

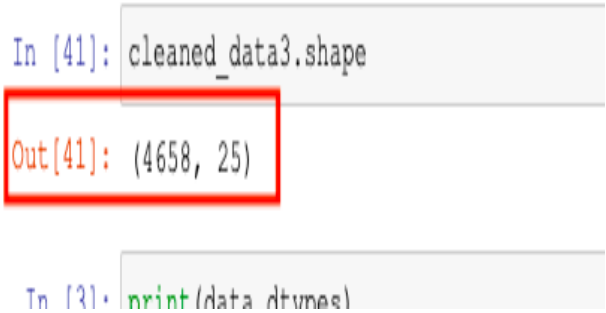

Fig 8. Final form of the data set

After the data cleaning process, we see in the descriptive table that the number of columns has shrunk to 25 out of 32 columns, and we see that the number of rows has become 4,658 rows instead of 12,316 rows. Also, the number of rows in the data after cleaning is equal in all columns, which is the opposite of what we had before cleaning. There was a discrepancy in the number of rows per column.

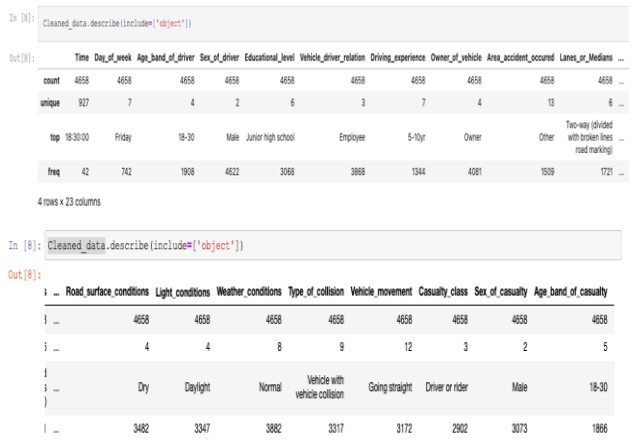

Fig 9. Descriptive table after cleaning

C. Data analysis

The data collector collected data from manual records of road accidents for the years 2017-20, and all sensitive information was excluded. The data

collector also expressed the belief that these are the most important causes of traffic accidents based on what he found in the manual records.

This chart contains the causes of accidents, and the biggest factor causing accidents is “No distancing,” as the number of accidents due to this factor reached 861, which is followed by “Changing the lane to the right” with 677 accidents. The least relevant factor causing accidents was “Improper parking” with 8 accidents, then “Drunk driving” with 10 accidents.

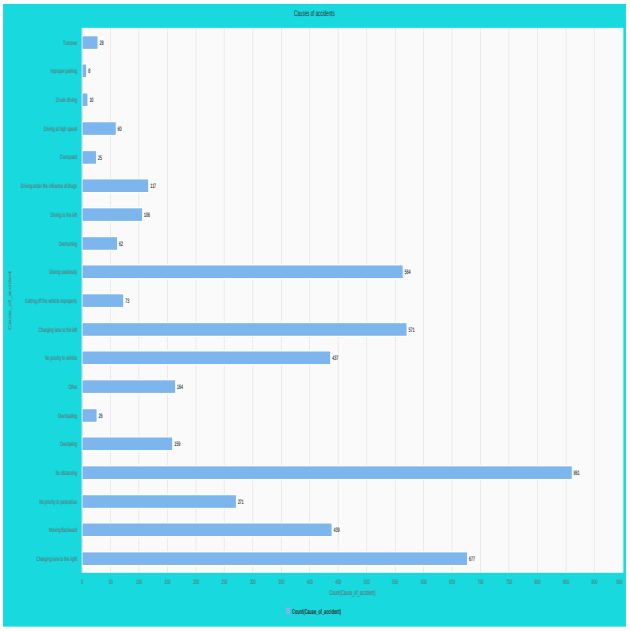

Fig 10. The causes of accidents

Here is the driver's gender column. This column shows us the gender of the drivers who caused accidents with any cause. The number of males who were one of the parties to an accident was 4,622, but there were only 36 females.

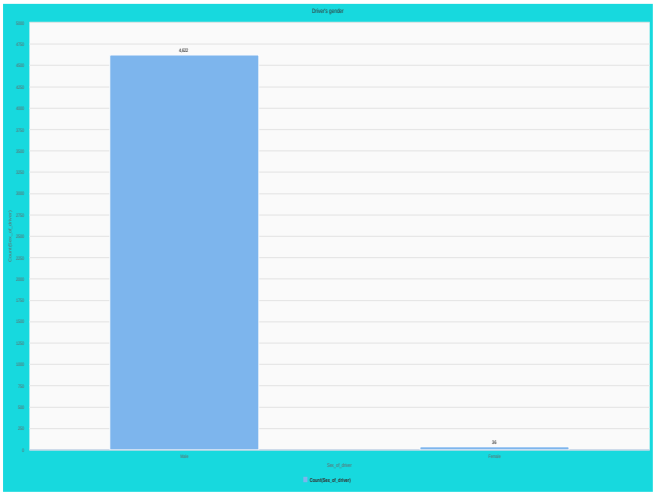

Fig 11. Driver's gender

This chart shows us the most vulnerable group to accidents, and they are the youth group in this case. The ages of the drivers who caused the accident were divided into four groups. 1,908 accidents involved drivers aged between 18-30 years, followed by drivers aged 31-50 years, where the number of accidents reached 1,731, then drivers over 51 years, where the number of accidents reached 662, and finally those under the age of 18 years, where number of accidents reached 357.

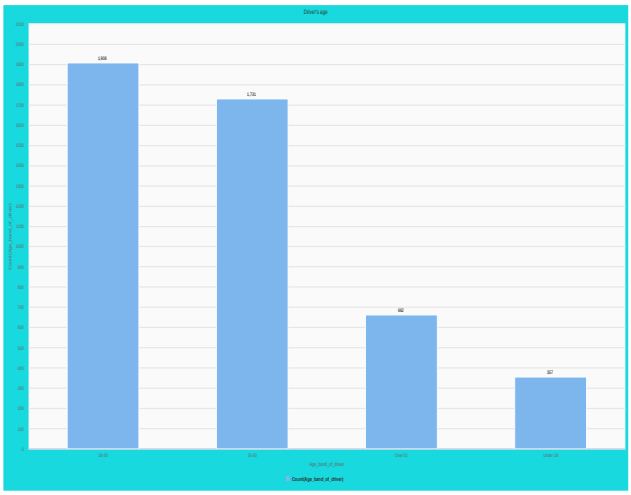

Fig 12. Ages of the drivers

Weather conditions are an important reason that may cause traffic accidents, depending on the condition: rain, fog, dust, etc. In the weather column, the number of accidents is shown by different weather conditions. As “normal weather” has the highest proportion with 3,882 accidents, then the “rainy” state with 516 accidents, then followed by “other weather” with 108 accidents, then the “cloudy” case with 57 accidents, then “winds” with 45 accidents, “snow” with 25 accidents, and finally “rain and wind” with 20 accidents.

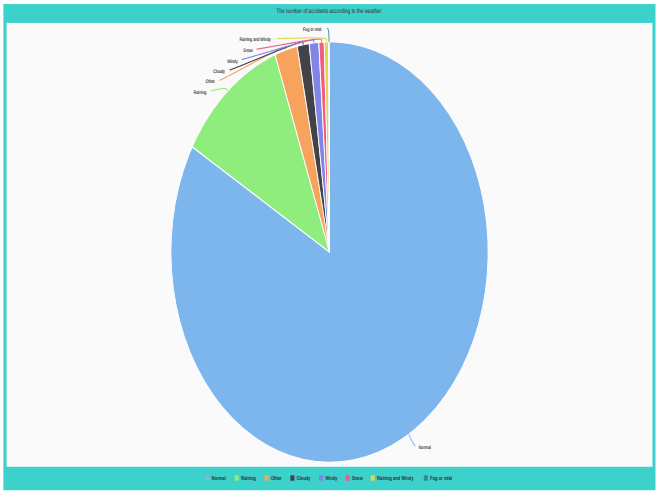

Fig 13. Weather conditions

The type of road the vehicle is traveling on is also an important factor that may cause accidents. There are several different circumstances such as dry, wet, snowy, and more. “Dry roads” got the highest rank with 3,482 accidents, then “wet roads” with 1,152 accidents, then “snowy roads” with 23 accidents, and “floods” with only one accident.

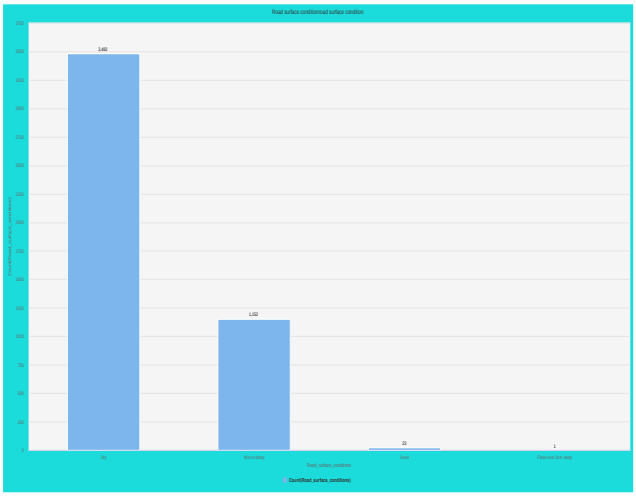

Fig 14. Road conditions

The lights on the road and on the vehicle are other important traffic safety factors for drivers on the road, and they reduce risk while driving. This chart shows us the accidents that occurred in different lighting conditions: the number of accidents in “daylight” reached 3,347, and the number of accidents in “darkness with lights lit” was 1,232, followed by “darkness without lighting” with 72 accidents, and finally “darkness with unlit lights” with 7 accidents.

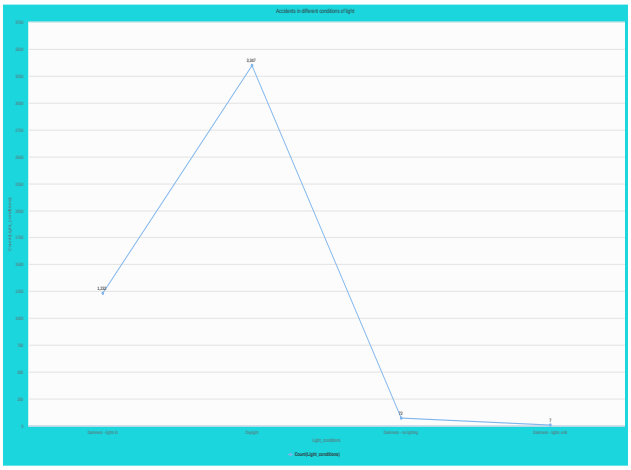

Fig 15. Lighting on the road

D. Machine learning

To create machine learning, the cleaned data was input, and a duplication filter was applied to remove any duplicate data. This operator removes duplicate examples from an ExampleSet by comparing all examples with each other based on the specified attributes. Two examples are considered duplicate if the selected attributes have the same values.

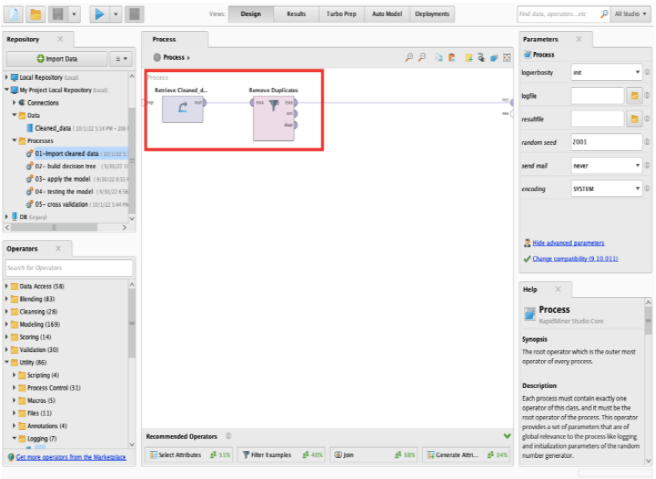

Fig 16. Cleaned data and a duplication filter

The search is a classification problem that we need for our predictions. A decision tree was chosen, and the decision tree algorithm was applied to the data that had been cleaned. This Operator generates a decision tree model, which can be used for classification and regression.

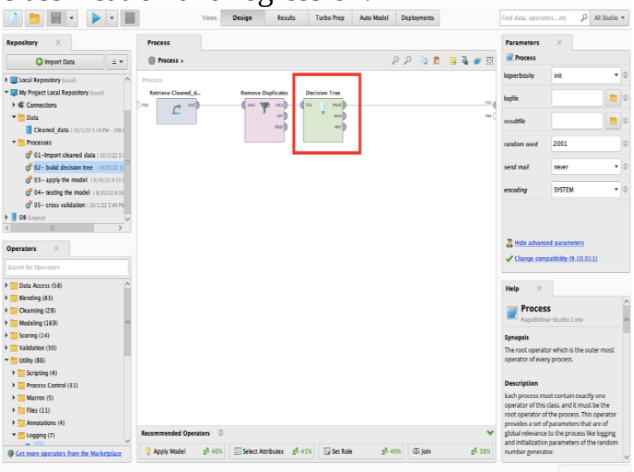

Fig 17. The decision tree algorithm

The huge decision tree could be used to support decision making manually, but this will be tedious, tiring and take a lot of time, so we will keep applying the model in the next steps until the decision support becomes automated after a short time.

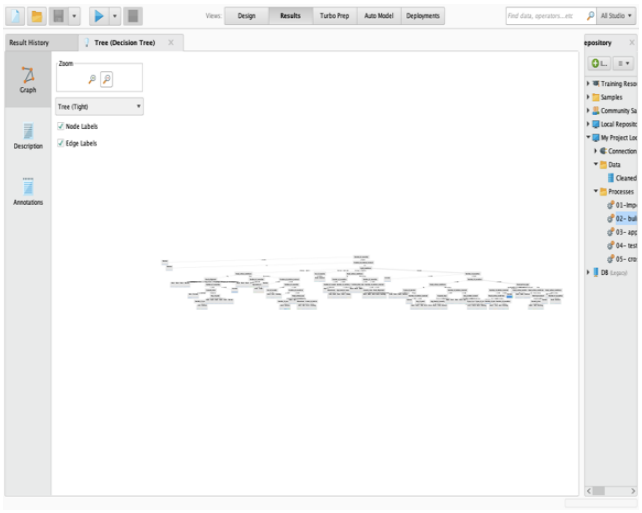

Fig 18. Huge decision tree

To apply the model, the data was divided into two parts. The first section contains 70% of the entries used for training the model. The training data is the initial dataset used to train the machine learning algorithms. The second part contains 30% of the entries used for testing the model. The test data is unseen data to test your model and evaluate its performance. The Split Data operator is used to partition data into subsets according to the specified relative sizes.

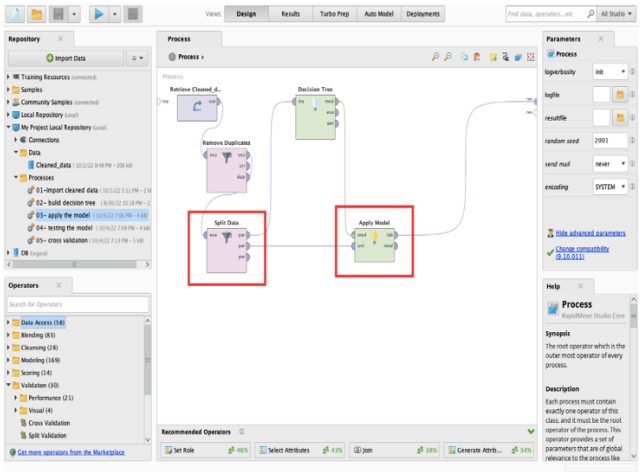

Fig 19. Applying the model

The weather column was chosen to predict accidents according to the weather conditions. There were several reasons behind choosing the weather column as the basis for prediction, including that the weather is one of the most important factors causing accidents. Also, the weather column is a comprehensive list of all different weather conditions. As we see in Figure 20, two columns are colored in green: the first column is the original data, and the second column is the prediction data.

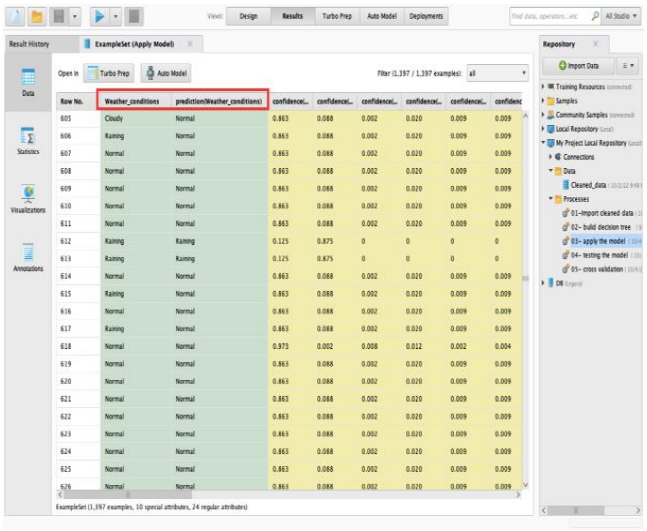

Fig 20. Original data and prediction data

In this step we need an operator to measure the accuracy of the model's performance for the training data, so we need a performance operator. This operator is used for performance evaluation. It delivers a list of performance criteria values.

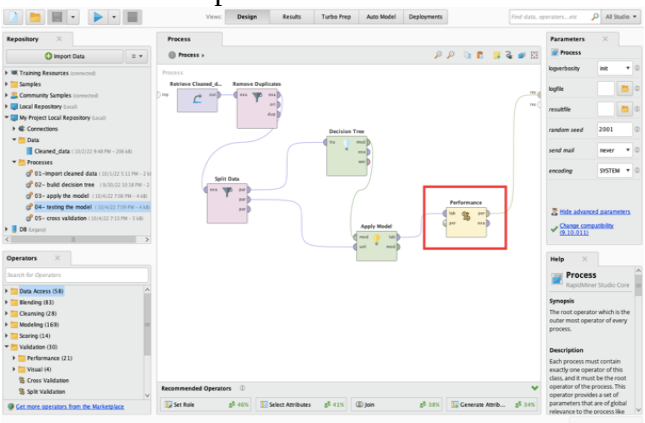

Fig 21. Performance

The accuracy of the model that we obtained was 84.75%. This is the accuracy of the model based on the training data, which totaled 70% of the entire data. Therefore, this should represent the accuracy of the model training.

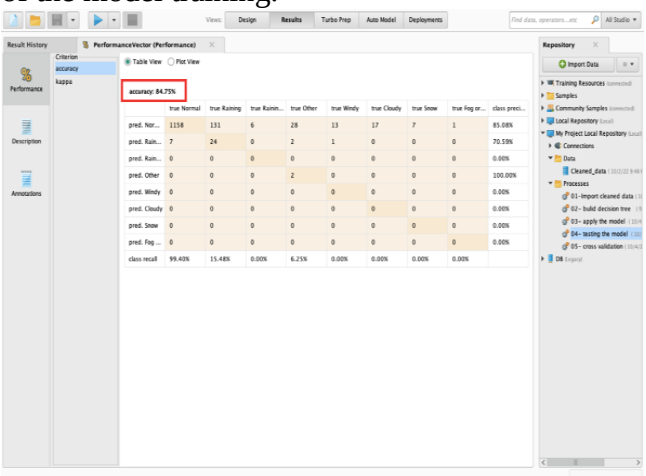

Fig 22. The accuracy of the model training

To obtain true model accuracy based on test data, a cross validation must be applied. The cross-validation operator performs a cross validation to estimate the statistical performance of a learning model.

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model.

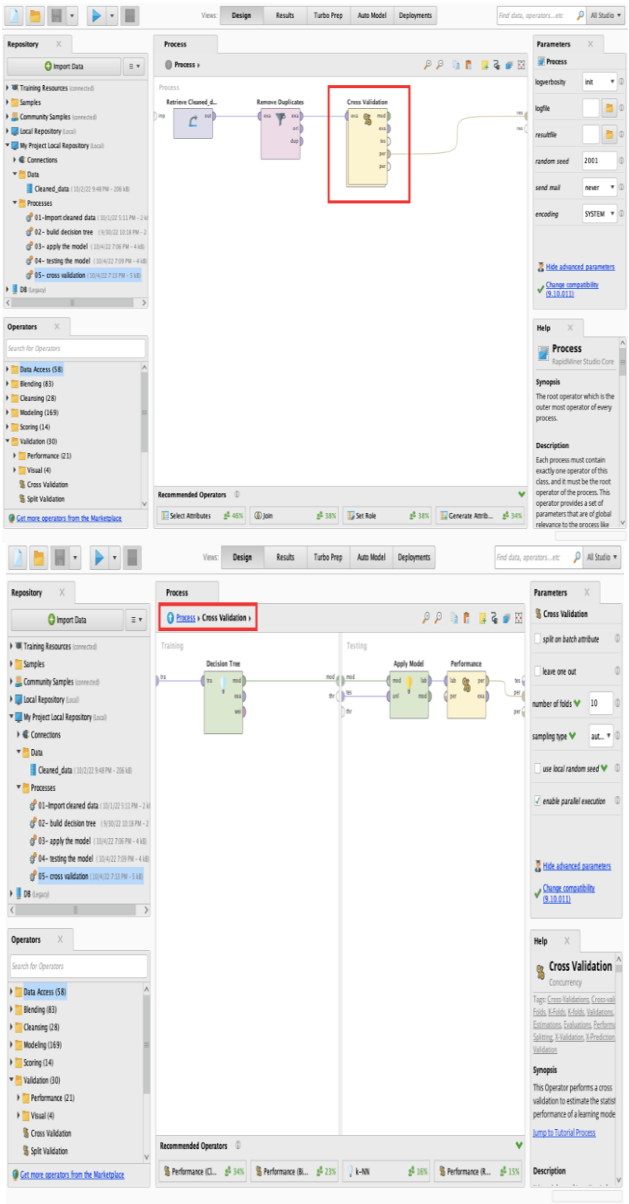

Fig 23. Cross-Validation

The model's accuracy is 84.22% and the standard deviation is +/- 1.08%. The lower the standard deviation, the more stable the model. The model's accuracy is not bad, but this needs to be proven. In the weather column, there are 8 different elements: “normal weather,” “Rain,” “snow,” “wind,” “rain and wind,” “cloudy,” “fog,” and “others”. In addition to that, there is a large amount of data. To be clearer, the driver's gender column contains “male” and “female” and the model's accuracy is 99.23% and the standard deviation is +/- 0.11%, but the driver's gender column is weak and ineffective for predicting accidents. Thus, we conclude that the greater the number of different elements in the target column, the lower the accuracy of the model and vice versa.

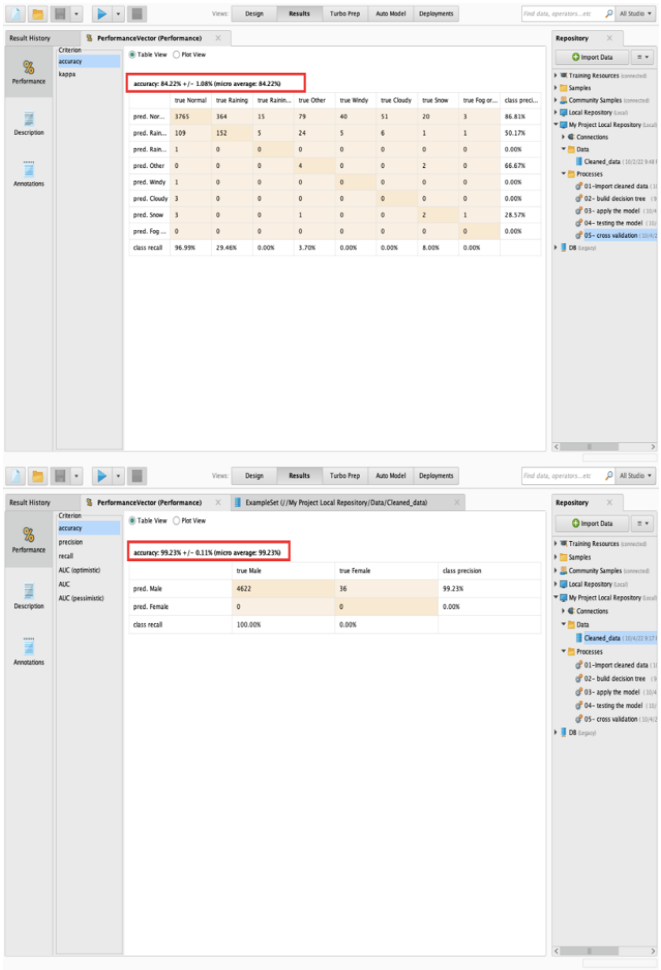

Fig 24. The model's accuracy

IV. Results and Discussion

The data analysis showed that there are multiple factors in the causes of accidents. In Figure 25, the results of all possible influential factors of accidents are presented. The most influential factors are “changing lane to right,” “moving backward,” “No distancing,” “No priority to vehicle,” “changing lane to left,” and “driving carelessly”. The order of the influential factors of accidents is shown in Figure 26. The order shows that “No distance between vehicles on the road” is the most common factor that causes traffic accidents on the road, with a rate of 861 accidents.

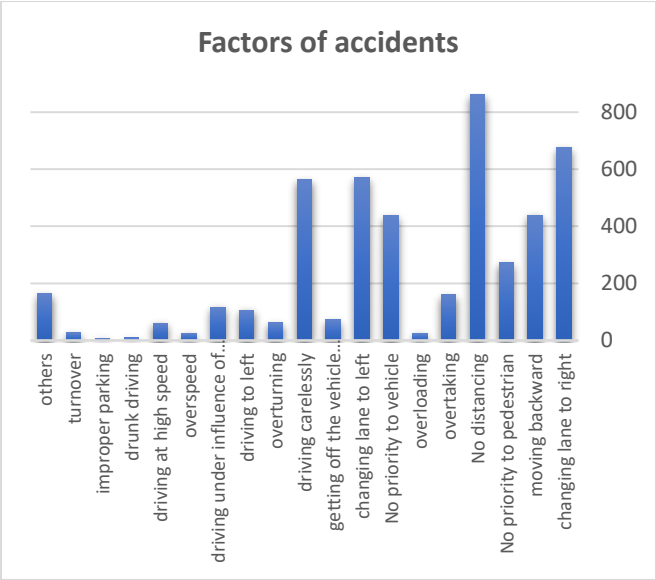

Fig 25. The results of all factors of accidents

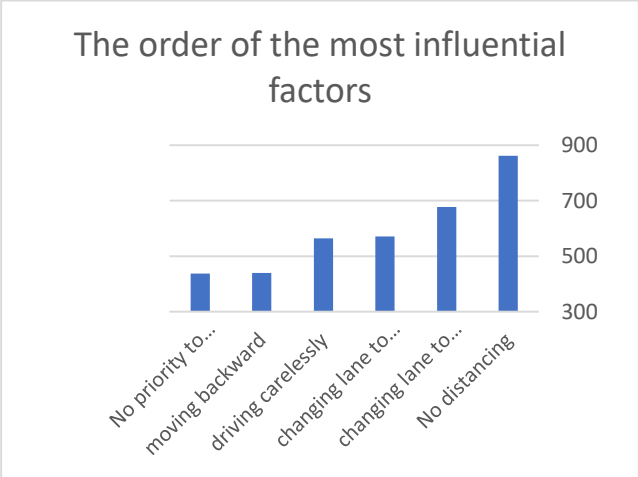

Fig 26. The order of the most influential factors of accidents

There are other factors that can be influential factors of accidents, which are drivers' gender and drivers' age. In Figure 27, the drivers' gender percentage shows that "male" had a high percentage, which is equal to 99%.

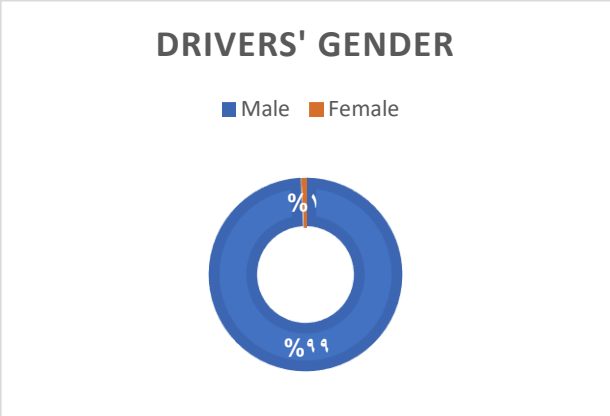

Fig 27. Percentage of drivers' gender

In addition, the drivers' age plays a major role. Figure 28 presents the percentages of drivers' ages. The age group that caused the most accidents was "between 18-30" years, which is equal to 41%, followed by "between 31-50" years.

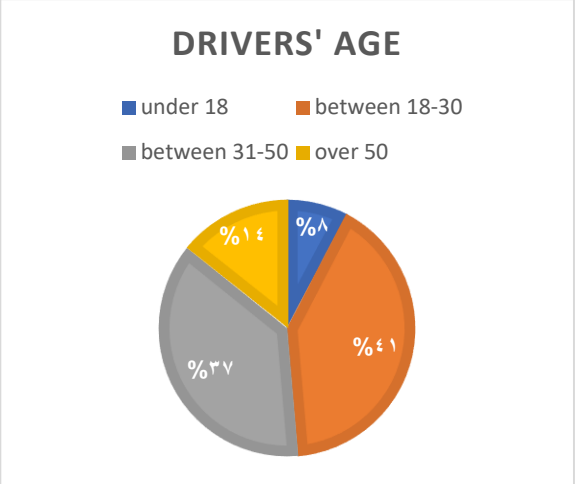

Fig 28. The percentages of drivers' ages

One of the unexpected results shown in Figure 29, is that dry roads are the most common road type that causes accidents. In addition, most accidents happen in normal weather and daylight as presented in Figure 30 and Figure 31. This gives us an indication that accidents do not depend on certain factors or special circumstances, as these accidents abound in normal conditions and factors.

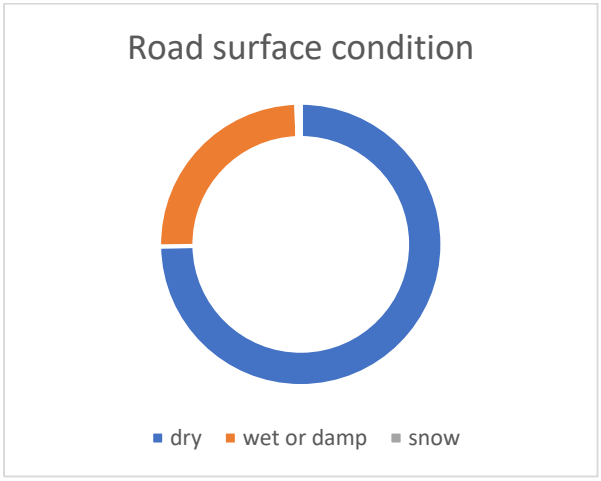

Fig 29. Percentages of road surface conditions

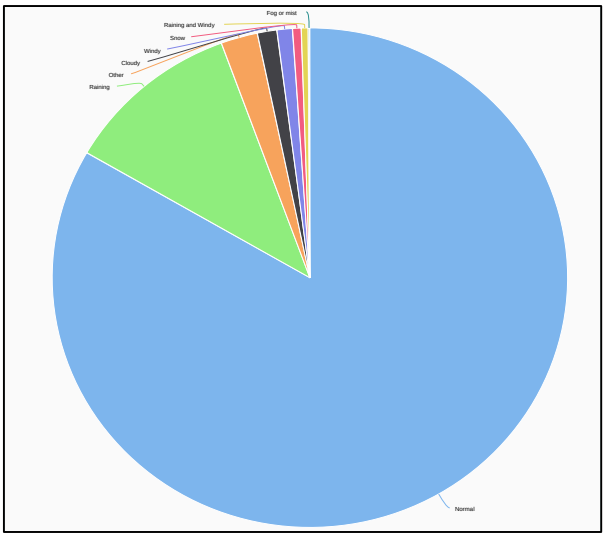

Fig 30. Percentages of weather conditions

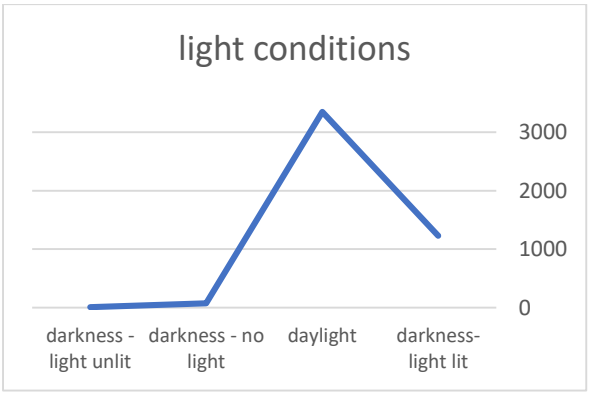

Fig 31. Curve of light conditions

The model used a decision tree algorithm based on the weather column, and weather is an important factor in predicting vehicle accidents on the road. This column contains 8 sub-factors which are the different weather conditions. This model achieved an accuracy rate of 84.22%, exceeding several previous studies by a difference of up to 10 degrees Additionally, this model is considered more stable based on the standard deviation achieved of +/- 1.08%. The closer this number is to zero, the more stable the model.

V. Conclusion and Future Work

It would be interesting to create more than one model using different machine learning algorithms and then compare the results of each algorithm with the other algorithms. It would also be worthwhile to create a dashboard to track data for a specific process, for example tracking the number of incidents on a quarterly basis.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Osama Rabie, who guided me throughout this project and spared no effort to help me. I would also like to thank Professor Mohammad Ramadan, who guided me at the beginning of this project. And thanks to my family who have supported me throughout this journey.

References

[1] Kaggle, "https://www.kaggle.com/datasets," 2010. [Online].

[2] World Health Organization. Available: https://www.who.int/publications/i/item/9789241565684 , 2018. [Online].

[3] Rolison, J. J., Regev. S., Moutari. S. & Feeney. A. "https://www.sciencedirect.com/science/article/pii/S0001457518300873," 2018.

[4] Runkler, T. A., "https://link.springer.com/book/10.1007/978-3-658-14075-5," 2016.

[5] Mohri, M., Rostamizadeh, A. & Talwalkar, A. "https://cs.nyu.edu/~mohri/mlbook/," 2018.

[6] Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K. & Nawrine, F. "https://ieeexplore.ieee.org/document/8843640," 2019.

[7] AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. "https://ieeexplore.ieee.org/abstract/document/8717393," 2019.

[8] Lin, L., Wang, Q. & Sadek, A. W. "https://www.sciencedirect.com/science/article/abs/pii/S0968090X15000947," 2015.

[9] Singh, A., Bhardwaj, A., Pathak, R. & Ahluwalia, S. "https://www.semanticscholar.org/paper/AN-EPIDEMIOLOGICAL-STUDY-OF-ROAD-TRAFFIC-ACCIDENT-A-Singh-Bhardwaj/a0a2fdd8c7bc9b0209d11b7b1d4df0b9d2f90adb," 2012.

[10] Wenqi, L., Dongyu, L. & Menghua, Y. "https://ieeexplore.ieee.org/document/8056908," 2017.

[11] Bucsuházya, K., Matuchová, E., Zůvalaa, R., Moravcová, P., Kostíková, M. & Mikuleca, R. "https://www.sciencedirect.com/science/article/pii/S2352146520302192," 2020.

[12] Gicquel, L., Ordonneau, P., Blot, E., Toillon, C., Ingrand & P. Romo, L. "https://pubmed.ncbi.nlm.nih.gov/28620324/, " 2017.

[13] Ihueze, C. & Onwurah, O., "https://www.sciencedirect.com/science/article/abs/pii/S0001457517304542," 2018.

[14] Mohanta, B. K., Jena, D., Mohapatra, N., Ramasubbareddy, S., Rawal, B.S., "https://content.iospress.com/articles/journal-of-intelligent-and-fuzzy-systems/ifs189743," 2022.

[15] Shahane, S. "Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents.

استخدام تحليلات البيانات والتعلم الآلي للتنبؤ بحوادث الطرق وتحليل العوامل السببية

عبدالله حسن العسيري

قسم نظم المعلومات، كلية الحاسبات وتقنية المعلومات،

جامعة الملك عبد العزيز، جدة، المملكة العربية السعودية

مستخلص. يمثل التعلم الآلي وتحليل البيانات معًا قوة قوية في إنتاج نماذج التحليل والتنبؤ والتي بدورها تدعم نمو المجالات المختلفة. من حيث السلامة والأشخاص، فإن حوادث المرور على الطرق لها تأثير سلبي على الأفراد والحكومات والشركات. لذلك، أصبح التنبؤ بالحوادث المرورية على الطريق أمرًا مهمًا للغاية في عملية اتخاذ القرار الذي سيؤدي إلى سلامة الآخرين. من المهم أيضًا معرفة أسباب هذه الحوادث. تهدف هذه الدراسة إلى التنبؤ بحوادث الطرق والعوامل المسببة لها بناءً على مجموعة بيانات من Kaggle [1] وإنشاء نموذج للتنبؤ بحوادث الطرق باستخدام التعلم الآلي وخوارزميات شجرة القرار. وأظهرت النتائج أن أحد أهم العوامل المسببة للحوادث على الطريق هو عدم وجود مسافة بين المركبات، وأن الطقس والعمر يلعبان دورًا مهمًا. حقق النموذج دقة ٨٤,٢٢٪ وانحراف معياري +/- ١,٠٨٪. كانت الأدوات المستخدمة في تحليل البيانات والتعلم الآلي والتنبؤ هي Python وRapidMiner وهي منصة برمجية متكاملة لعلوم البيانات توفر بيئة متكاملة للتعلم الآلي والتعلم العميق والتحليل التنبئي.

الكلمات المفتاحية: تحليل البيانات، الخوارزمية، حوادث المركبات، التعلم الآلي، التنبؤ