# Introduction

Traffic accidents as one of the most important and frequent factors in everyday societies have always been discussed and discussed. These accidents, in addition to the physical and financial damages they cause to people and machines, affect people's lives in all aspects. Therefore, predicting traffic accidents and reducing the probability of their occurrence can be one of the effective solutions to increase safety and improve driving conditions.

In this project, our goal is to provide a complete method for predicting traffic accidents using two methods of statistical analysis and machine learning. In the statistical analysis method, we use historical data related to traffic accidents to examine the patterns and statistical relationships between different factors and the occurrence of accidents. This method can help us to predict the factors affecting the probability of an accident or...

In the machine learning method, we use advanced algorithms and artificial intelligence methods such as neural networks and regression. By using historical data and complex algorithms, a model is trained to be able to predict issues such as the severity of injuries in accidents by inputting the information of new conditions. This method can increase the accuracy and efficiency of predictions and be used for decisions related to driving safety.

# A review of traffic accident literature

In the field of traffic accidents, it is very important to study and analyze the data related to accidents and the factors affecting them. In specialized literature, a large number of studies and researches have been done in the field of traffic accidents. In some of these studies, statistical analysis methods such as regression and principal component analysis have been used to investigate the relationship between various factors and the occurrence of accidents. These methods can identify patterns and statistical relationships between factors.

In contrast, machine learning methods such as neural networks are able to use historical data to train models. According to the patterns and relationships in the data, these models are able to predict the probability of traffic accidents in new conditions. Considering the complexity and high

volume of driving data, the use of machine learning methods can lead to higher accuracy and efficiency in predicting accidents.

# Description of analysis methods:

The statistical and machine learning methods and methods used to analyze traffic accident data are as follows:

Statistical Methods:

- Import the necessary libraries
- Reading CSV file
- Display a summary of the descriptive statistics of the data
- Drawing a diagram of the distribution of accidents by year
- Analysis of injury severity column
- Drawing a diagram of the distribution of types of injuries
- Examining the relationship between the severity of the injury and the age of the victim
- Drawing the accident distribution chart based on age group
- Calculation of correlation matrix of numerical variables

machine learning:

- Logistic regression to predict injury severity
- Random forest for damage severity classification
- Multilayer perceptron (MLP) neural network for classification
- XGBoost for classification and AUC calculation
- Calculating the importance of features in algorithms

- Drawing a learning curve to evaluate the performance of models

# Results and analysis

## Analysis of statistical plots

### # Plot the distribution of accidents by year

All accidents in 2022 have fallen

### # Analyze the 'casualty_severity' column

This chart has two bars representing two levels of damage severity: "3.0" and "2.0".

The values in the "label" column are:

- **2**: Serious

- **3**: Slight

The bar corresponding to injury severity "3.0" (slight) is much higher than "2.0" (serious), indicating that the number of injuries with severity "3.0" is much higher than that of severity "2.0". In other words, there were too many accidents with "serious" injury severity.

### # Plot the distribution of casualty type values

It is a chart titled "Distribution of Damage Type Values". In this graph, the horizontal axis (x) represents the "type of damage" and the vertical axis (y) represents the "number".

- Damage type "9" or  has a very high number, reaching about 25,000, which indicates the frequent occurrence of this type of damage in accidents.

- Other damage types have lower numbers, most of them below 5,000.

Overall, this chart shows that injury type "9" is much more common among accidents than other injury types. This information can be useful for driving safety planning and accident prevention.

It is a chart titled "Distribution of Damage Type Values". In this graph, the horizontal axis (x) represents the "type of damage" and the vertical axis (y) represents the "number".

- Damage type "9" (passenger of the car) has a very high number, reaching about 25,000, which indicates the high occurrence of this type of damage in accidents.

- Damage type "0" (pedestrian) and "1" (cyclist) are slightly higher than the number of 5000.

- Other damage types have lower numbers, most of them below 5,000.

Overall, this chart shows that injury type "9" is much more common among accidents than other injury types. This information can be useful for driving safety planning and accident prevention.

The different types of `casualty_type` are as follows:

0. Sidewalk

1. Cyclist

2. Motorcycle under 50 cc driver or passenger

3. Motorcycle above 125 cc and up to 500 cc driver or passenger

4. Motorcycles above 125 cc and up to 500 cc passenger

5. Motorcycle over 500 cc driver or passenger

6. Taxi/rental car

7. The car is sitting

.....

## # Analyze the relationship between 'casualty_severity' and 'age_of_casualty'

The graph shows the age distribution of the victims according to the severity of the injury. Two categories with damage severity 2 (Serious) and 3 (Slight) are shown. In both categories, the age of the victims varies between about 20 and 80 years. Also, there is no big difference in the age distribution of the victims between the two categories. The average age for both levels of injury

severity appears to be around 40 to 60 years. This graph can be useful for examining the relationship between the age of the victims and the severity of their injuries.

## # Plot the distribution of accidents across age bands

The graph shows the distribution of accidents according to the age group of the victims. The horizontal axis of the graph shows the age groups of the victims and the vertical axis shows the number of accidents. The highest number of accidents occurred in the age group of 6 (26-35), then a significant decrease was observed in the age group of 7 (36-45). Age groups 5 to 8 (21-55) have the largest number, with age 6 having the largest number at over 12,000. This graph can be useful for examining the relationship between the age group of victims and the number of accidents.

## #Plot casualty severity by gender

The graph shows the frequency of casualty severity by gender. The frequency of casualties varies depending on the gender of the victim. For example, the frequency of casualty severity for men is higher than the frequency of casualty severity for women.

Analysis:

Distribution: The frequency distribution of casualty severity is almost the same for both sexes.

Average: The average severity of casualties is slightly higher for men than for women.

Standard deviation: The standard deviation of casualty severity is almost the same for both sexes.

Interpretation:

Men are more likely to suffer severe casualties: This chart shows that men are more likely to suffer severe casualties than women. This can be due to various factors including biological differences, social roles and risky behaviors.

Gender difference in casualty severity is visible at different levels of casualty severity: This difference is visible at different casualty severity levels. For example, at the 2.0(Serious) casualty severity level, the frequency for males is almost twice that for females.

Data on unspecified gender is significant: data on unspecified gender accounts for 10% of all data. This shows that in many cases, the gender of the victim is not known.

## # Plot the relationship between casualty severity and casualty class

Horizontal axis: the severity of the injured, which is either serious or mild

Vertical axis: number of casualties

In the severity of low injuries, the number of drivers is more, followed by the number of passengers, and then the number of pedestrians. Passengers are placed.

## # Plot the relationship between casualty severity and casualty's home area type

The chart shows "damage severity by type of local area". In this graph, the severity of injuries is displayed on the horizontal axis and the number of injuries on the vertical axis. Four types of local areas are shown.

Based on injury severity 2.0:

- Rural areas have the highest number of injuries.
- Urban and semi-urban areas do not show any damage.

Based on damage severity 3.0:

- Urban areas have the highest number of injuries, which is much more than any other type of area.
- Semi-urban and unknown areas show few damages.
- Rural areas do not show any damage.

## # Plot the distribution of accidents by gender

A graph shows the "distribution of accidents by gender". In this graph, gender is displayed on the horizontal axis and the number of accidents on the vertical axis. There are three genders with different colors.

- For women: the number of accidents is much higher than that of men and reaches close to 30,000.

- For men: the number of accidents is about half of that of women.

- For unknown: no data is displayed.

## # Visualize the distribution of 'age_of_casualty'

A graph shows the "age distribution of accidents". In this graph, the age of accidents is displayed on the horizontal axis and the frequency on the vertical axis.

- The graph in blue shows that more frequency of accidents occurred in the age group of 20 to 40, with the highest frequency around the age of 20 to 25.

- The green line shows the average age of accidents, which is about 40 years.

- The red line shows the average age, which is around 25 years old.

The green line in the graph shows the average age of casualties. The red line shows the average age of the entire population.

The average age of casualties is lower than the average age of the entire population. This shows that young people are more at risk of death.

## # Plot a boxplot for 'age_of_casualty'

The graph is a boxplot that shows the "age of victims". In this graph, the age of the victims is displayed on the horizontal axis and the boxes and lines are displayed on the vertical axis.

- The blue box indicates the interval between the first and third quartiles (IQR) of the data, which indicates that the middle 50% of the victims' ages are between about 20 and 60 years old.

- The vertical line inside the box shows the median age of the victims, which is about 40 years.

- Two lines drawn on both sides of the box show fluctuations outside the first and third quartiles; Therefore, they include ages outside the middle 50% but within about 1.5 IQRs on either side of the IQR.

- Several small circles on the right side of the dashed line represent the outlier data above about 80 years.

Focus:

The age of the injured is directly related to the age of the participants.

Most of the injured are between 20 and 60 years old.

Few of the injured are under 20 years old or over 60 years old.

## # Plot scatter plots using plotly

The graph is a scatter plot, with "victims' age" on the vertical axis and "victims' gender" on the horizontal axis. One gender is male and two gender is female.

- The data are mostly scattered at the end of the horizontal axis, near values 1 and 2.

- The color bar on the right indicates the class of the victim, which ranges from 1 (dark purple, meaning the driver) to 3 (pedestrians).

- Most of the data are concentrated in the lower end of the victim class scale, i.e. drivers, and are from the age of 20 onwards.

The dispersion rate of pedestrian victims is higher in women.

## # Plot scatter plots using plotly

This graph is a scatter plot showing the relationship between "bus_or_coach_passenger" and "car_passenger". Also, the data color changes based on "sex_of_casualty".

In this chart, there is one notable data point:

- one at point (0,1) indicating that there is a bus or caravan passenger with "sex_of_casualty" equal to 1, i.e. male, who is not a car passenger.

## # Plot the distribution of pedestrian location values

This chart is a bar chart titled "Distribution of Pedestrian Locations". The x-axis labeled "Pedestrian Location" shows values from 0 to 10, and the y-axis labeled "Number" shows the number of pedestrians, which at location 0 is greater than 40,000. This point means no sidewalk.

- Other locations (1 to 10) have significantly fewer pedestrians, with the small red bars at locations 5 meaning on the freeway, crossing, and then the other locations have less.

## # Plot the distribution of Casualty_IMD_Decile values

According to the graph, it can be seen that the number of injured has decreased in areas with more deprivation (with higher IMD values). Regions 1 to 3, which are 10-20%, 20-30% and 10% more deprived respectively, have the highest number of injured. This suggests that greater deprivation

is associated with a greater number of injuries. As the value of IMD increases, i.e. the deprivation decreases, the number of injured decreases. This could indicate that less deprived areas have more resources for accident prevention or public health care. This graph can be useful for examining the relationship between deprivation and the number of injuries in surgeries.

## # Plot the distribution of pedestrian road maintenance worker values

A diagram shows the distribution of pedestrian road maintenance workers. According to this chart, a large number of workers (around 50,000 people) are in category 0 (No / Not applicable). It seems that no worker is included in categories 1 (yes) and 2 (unknown).

- Category 0 (No / Not applicable): The largest number of workers (about 50,000 people) are in this category. This means that for the largest number of workers there is no case for footpath maintenance or it is not applicable to them.

- Category 1 (Yes): No worker has been included in this category. This means that there are no footpath maintenance workers currently engaged in footpath maintenance.

- Category 2 (Not known): No worker has been included in this category. This means that the maintenance status of the footpath is not uncertain for any of the workers.

This chart can indicate a need for resource planning or more training for workers.

## # Plot vehicle_reference vs casualty_reference

The chart is a line chart that shows the number of cars on the vertical axis and the car reference on the horizontal axis.

The vertical axis is graded from 0 to 20 and the horizontal axis is graded from 0 to 60.

The graph shows that the number of cars increases as the car reference increases.

Specifically, the chart shows that:

* About 5% of people who have a car reference of 0 are willing to pay for the car.

* About 10% of people who have a car reference of 10 are willing to pay for the car.

* About 15% of people who have a car reference of 20 are willing to pay for the car.

* About 20 percent of people who have a car reference of 30 are willing to pay for the car.

* About 25% of people who have a car reference of 40 are willing to pay for the car.

* About 30% of people who have a car reference of 50 are willing to pay for the car.

* About 35% of people who have a car reference of 60 are willing to pay for the car.

There are several possible interpretations for these findings.

One interpretation is that people with higher car references are more likely to be economically prosperous and therefore able to afford a car.

Another interpretation is that people with higher car references are more likely to need a car and therefore willing to pay for it.

For example, people living in rural areas may need a car to get to work or school.

Finally, it is also possible that people with higher car references are more likely to be interested in driving and therefore willing to pay for a car.

There are insufficient data to determine the correct interpretation of these findings.

However, this graph shows that there is a positive relationship between car reference and willingness to pay for a car.

## # Calculate the correlation matrix for numeric variables

This graph is a correlation matrix between numerical variables that appear to be related to traffic accidents. Each box represents the correlation between two variables, and the colors change from yellow (strong positive correlation) to purple (strong negative correlation) and green (low or no correlation).

Variables such as "accident_index", "accident_year", "vehicle_reference" and others are listed in both rows and columns to show the correlation between each pair. Each cell contains the numerical value of the correlation coefficient between the relevant variables.

For example, there is a strong positive correlation between pedestrian movement and pedestrian position, and changes in one will directly affect the other.

# Analysis of machine learning plots

## LogisticRegression

The first model is LogisticRegression, which has a precision of 0.8179539099411821 and other evaluation parameters Precision: 0.7601719897782432 and Recall: 0.8179539099411821 and F1 Score: 0.736413216060629.

## RandomForestClassifier

The second model is RandomForestClassifier, which has accuracy results of 0.8175682190724135 and evaluation parameters of Precision: 0.7527464917762274, Recall: 0.8175682190724135, and F1 Score: 0.7394492535147685.

This model's analytic chart sorts the features by their importance, with the more important features at the top of the y-axis.

There are 15 bars, each representing a feature. The length of each bar corresponds to the relative importance of that feature in the model. Feature number 12 and then 5 are significant because their bar is stretched much more than other bars in the x-axis. 5 is related to age_of_casualty and 12 is related to casualty_type.

Both the training score and the cross-validation score are plotted against the size of the training set.

- As the size of the training set increases, there is a visible trend where both curves seem to converge but with significant fluctuations.
- The shaded area represents the variance or standard deviation of scores at any point in time; This helps in understanding the uniformity of our model performance.

This learning graph can help us understand whether our model improves with more training data. Also, it can show whether our model is overfitting or underfitting. If the two curves converge and end with a high score, this indicates that the model will improve with more training data. If the training curve starts with a high score and then decreases, and the test curve starts with a low score and then increases, this indicates that the model is overfitting. If both curves start with a low score and then slowly increase, this indicates that the model is underfitting.

This diagram shows the learning curve of a system during the training process. The horizontal axis shows the size of the training set, which is expressed in terms of the number of samples. The vertical axis is also divided into two separate sections to display the "Training Accuracy" and "Cross-Validation Accuracy" scores.

Training accuracy score: As the size of the training set increases, the training accuracy score increases steadily. This shows that the system learns better by considering more examples.

Cross-validation accuracy score: While the training accuracy score is increasing, the cross-validation accuracy score increases initially, but starts to decrease after reaching a peak.

## MLPClassifier

The next model was MLPClassifier, which had a value of 0.8143683702989393 for Score on training data and a value of 0.81785748722399 for Score on test data.

## XGBClassifier

The next model is XGBClassifier, which has 0.8384794893058958 for Score on training data, and its analytical graph has 0.8201119057771881 for AUCN.

Its solubility diagram is the ROC (Receiver Operating Characteristic) curve. The horizontal axis shows the False Positive Rate and the vertical axis shows the True Positive Rate.

The ROC curve is not a smooth curve and is drawn diagonally from the lower left corner to the upper right corner of the chart. The closer this curve is to the upper right corner, the better the system's performance in detecting positive cases (e.g. patients).

AUC is a numerical index to evaluate system performance. An AUC of 1 indicates perfect performance and an AUC of 0.5 indicates random performance.

The AUC of this graph is 0.85, which shows that the system performs well in detecting positive cases.

The cutoff point is the point on the ROC curve where the false positive rate and the true positive rate are equal. This point represents a balance between sensitivity and specificity of the system.

This graph shows that the system performs well in detecting positive cases. However, system performance can be optimized to increase sensitivity or specificity by adjusting the cut-off point.

In this graph, the ROC curve is not a straight line, which indicates the optimal performance of the system.

AUC of 0.85 indicates good system performance, but this performance can be improved to some extent by using various techniques.
The optimal cutting point depends on the type of system application.
Compared to the previous chart, this chart provides more information about system performance.
AUC 0.85 in this chart indicates better performance of the system compared to the previous chart.

# Discussion and conclusion

Statistical analyzes in this project show that the severity of injuries in traffic accidents is divided between two levels of "serious" and "low". The number of "minor" severity injuries far exceeds the number of "serious" severity injuries, indicating that most crashes result in "minor" severity injuries. Also, "type 9" injuries, which is the passenger of the car, has the highest number and is the most common in accidents.

Analysis of the relationship between the severity of injuries and other factors has also been done. Examining this relationship shows that factors such as the driver's age may affect the severity of injuries in accidents. In this way, the examination of the relationship between the severity of the injury and age showed that the age range of the victims in both categories of serious and mild injuries is between 20 and 80 years. Also, the analysis of the gender distribution of the victims showed that women had more accidents than men. Also, the severity of male injuries was higher. The distribution chart between the age and gender of the victims showed that most of the male victims are in the age group of 20 to 40 years. The analysis of the victim's residence area also determined that rural areas had the highest number of injuries. These analyzes can be effective in identifying risk factors and high-risk groups and planning related to driving safety.

In the use of machine learning methods, various algorithms such as logistic regression, random forest, multilayer perceptron neural network and XGBoost were used to predict and classify the severity of injuries of accident victims, which had acceptable accuracy. Second, the learning curve of the algorithms showed that the performance of the models improves with the increase in the amount of training data. Also, the algorithms are not subject to overfitting or underfitting. Thirdly, the calculation of the importance of the features in the random forest algorithm showed that the features of age and type of injury have the greatest effect on the output of the model. Finally, the

ROC plot of XGBoost algorithm showed that this algorithm has a very good performance in true positive detection.

# Suggestions to reduce traffic accidents

Based on the analysis, it can be concluded that in order to increase driving safety and reduce the number of accidents, more focus should be placed on preventing accidents with "low" intensity and "type 9" injuries, which is the passenger of the car. This information can help relevant authorities and relevant organizations in implementing appropriate programs and policies to increase driving safety and reduce the number of accidents. Also, knowledge of these results can help drivers and the general public to have a better understanding of factors contributing to accidents and to take appropriate safety measures.

**In general, the following are recommended:**

1. More focus on preventing accidents with "low" injury severity and "9" injury type, which is related to the passenger of the car. This information can help relevant authorities and related organizations in implementing appropriate programs and policies to increase driving safety.

2. Informing drivers and the general public about factors affecting the occurrence of accidents and taking appropriate safety measures by them.

3. Focusing on high-risk groups such as young male drivers (20-40 years old) and implementing driving and safety training programs for them.

4. More focus on rural areas that had the highest number of injured.

5. Using machine learning algorithms to predict and classify the severity of injuries caused by accidents.

6. Improving transportation infrastructure and monitoring vehicles to comply with safety regulations.

7. Implementing educational programs and promoting safety culture among the general public.

8. Increasing investment and improving infrastructure in more deprived areas that have had more injured people. This could lead to more accident prevention or better public health care.

9. Focusing on preventive measures for pedestrians, especially women, who had a higher dispersion rate.

10. Implementing special training programs for sidewalk maintenance workers, many of whom were in the "not applicable/inapplicable" category.

11. More detailed analysis of the relationship between car reference and willingness to pay for a car to better understand patterns and make more appropriate decisions.

12. Using more advanced algorithms such as neural networks for more accuracy and efficiency in predicting accidents.