Car Accident Severity Prediction

Riccardo Bellio

October 8, 2020

1. Introduction

1.1. Background

Various types of transportation are available to move around. Evidence shows that vehicles like cars, vans and trucks represent the main types of transportation in our society and that the vast majority of people within each country use a car on a daily basis. Thus, car safety is an important issue, for drivers, cyclists and pedestrians. Car collisions can be severe and end or badly impact people's lives, or less serious and cost the people involved a lot of money.

Data released by WHO (World Health Organisation) on the 7th Feb 2020 reveals some alarming trends. On the globe, every year:

- **1.35 million** people die as a result of a car accident.
- **20 to 50 million** people suffer non-fatal injuries.
- Minor injuries cause considerable economic losses to individuals and countries.
- Car accidents cost most countries app. 3% of their entire GDP.

These numbers clearly indicate that a final solution to this problem has not been found yet. In order to address this topic, it is essential to collect relevant data and explore the various factors that can impact people's safety while driving (unsafe road infrastructure, speed, distraction, driving under the influence of alcohol or other substances, as well as the light and weather condition).

1.2. Goals

This research had two main goals:

1. Creating a **predictive model** for car accident severity by analysing and working with a dataset released by the city of Seattle, U.S.A. This set contained data relating to external factors (such as road conditions, weather ...) as well as factors associated with specific drivers behaviour (speeding, driving under the influence of alcohol and substances,...). The accuracy of various algorithm (such as: Decision Tree, Logistic Regression, Key Nearest Neighbours and Support Vector Machine) was tested out to find the best performing one.
2. Analysing the data contained in this dataset to explore the factors that impacted accident severity with the intent to **share the collected information** with the general population, in order to improve their sense of awareness of the risks involved while driving.

1.3. Interest

This project was directed to:

- The general public, who is interested in the topic to improve their safety while driving.
- Governments and private companies: these entities could be interested in further improving the project content and develop mobile applications to inform drivers and positively impact their behaviour on the road. An alert system could inform drivers on their level of risk of being involved in an accident given their specific circumstances. As a response, they could choose to drive more carefully or change route, thus reducing the overall risk of causing/being involved in a car collision.

2. Data Acquisition & Cleaning

2.1. Dataset Overview

The dataset used in this project was the shared dataset provided by IBM on the course IBM-Data-Science and it contained detailed data relating to accident severity provided by the State Police Department of Seattle, U.S.A. The collected information was recorded on a weekly basis from 2004 until today. The version provided by the course was slightly modified, but the overall content was similar.

The dataset provided on the course was also available on the open data portal of the city of Seattle (Dataset). It contained:

- **39 input attributes**: they included detailed information about the recorded accidents. Some features were used to train our models and others disregarded, based on their importance for model creation.
- **1 output attribute**: the car accident severity evaluation, which included all the types of collisions displayed at the intersection or mid-block of a segment, which were divided into five different categories.

A brief description of these attributes is shown below:

OUPUT VARIABLE	DESCRIPTION
SEVERITY CODE	This included the codes that correspond to the severity of a recorded collision: • **3** = Fatality • **2b** = Serious Injury • **2** = Injury • **1** = Property Damage • **0** = Unknown

INPUT VARIABLES	DESCRIPTION
X	Location, longitude.
Y	Location, latitude.
OBJECTID	ESRI unique identifier.
INCKEY	A unique key for the accident.
COLDETKEY	A secondary key for the accident.
REPORTNO	The number of the reported accident.
STATUS	Not Specified.
ADDRTYPE	The collision address type (Alley, Block, Intersection).
INTKEY	A key that corresponds to the intersection associated with a collision.
LOCATION	The description of the general location of the collision.
EXCEPTRSNCODE	Not specified.
EXCEPTRSNDESC	Not specified.
SEVERITYDESC	Detailed description of the severity of the collision.
COLLISIONTYPE	Collision type.

PERSONCOUNT	The number of people involved in the collision.
PEDCOUNT	The number of pedestrians involved in the collision.
PEDCYLCOUNT	The number of bicycles involved in the collision.
VEHCOUNT	The number of vehicles involved in the collision.
INJURIES	The number of injuries in the collision.
SERIOUSINJURIES	The number of serious injuries in the collision.
FATALITIES	The number of fatalities in the collision.
INCDATE	The date of the accident.
INCDTTM	The date and time of the accident.
JUNCTIONTYPE	Category of junction at which collision took place.
SDOT_COLCODE	A code given to the collision by Seattle Department Of Transportation (SDOT).
SDOT_COLDESC	A description of the collision corresponding to the collision code.
INATTENTIONIND	Whether or not collision was due to inattention (Y/N).
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	A description of the weather conditions at the time of the collision.
ROADCOND	The condition of the road during the collision.
LIGHTCOND	The light conditions during the collision.
PEDROWNOTGRNT	Whether/not pedestrian right of way was not granted (Y/N).
SDOTCOLNUM	A number given to the collision by SDOT.
SPEEDING	Whether or not speeding was a factor in the collision (Y/N).
ST_COLCODE	A code provided by the state that describes the collision.
ST_COLDESC	A description that corresponds to the state's coding designation.
SEGLANEKEY	A key for the lane segment in which the collision occurred.
CROSSWALKKEY	A key for the crosswalk at which the collision occurred.
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N).

2.2. Data cleaning

The dataset provided in .csv format was downloaded and ingested into a Pandas dataframe. It contained some non-relevant information for the purpose of this project, thus some features were discarded as indicated in the table below:

The remaining column headers were renamed for clarity.

ATTRIBUTE	DESCRIPTION	DROPPING REASON
OBJECTID	ESRI unique **Identifier**.	Irrelevant.
INCKEY	Unique accident **key**.	Irrelevant.
COLDETKEY	Secondary accident **key**.	Irrelevant.
INTKEY	**Key** corresponding to the intersection for the collision.	Irrelevant.
SEGLANEKEY	**Key** corresponding to the lane segment for the collision.	Irrelevant.
CROSSWALKKEY	**Key** corresponding to the crosswalk for the collision.	Irrelevant.
EXCEPTRSNCODE	Unknown **keys**, not specified.	Irrelevant.
REPORTNO	Unknown **keys**, not specified.	Irrelevant.
STATUS	Unknown **keys**, not specified.	Irrelevant.
SDOTCOLNUM	Unknown **keys**, not specified.	Irrelevant.
SDOT_COLDESC	A **description** of the collision.	Irrelevant.
ST_COLDESC	A **description** of the state's coding designation.	Irrelevant.
EXCEPTRSNDESC	A non-specified **description**.	Irrelevant.
LOCATION	A **description** of the location of the accident.	Irrelevant.
SDOT_COLCODE	Collision **code** (SDOT), already in ST_COLCODE.	Double entry.
ST_COLCODE	A code of the collision.	Irrelevant.
INCDATE	Accident **date**, already contained in INCDTTM.	Double entry.
INJURIES	Information contained in the label data.	Irrelevant.
SERIOUSINJURIES	Information contained in the label data.	Irrelevant.
FATALITIES	Information contained in the label data.	Irrelevant.

Time, **day of the week**, **month** and **year** of the recorded accidents were extracted from the column `*INCDTTM*´ and displayed in separate columns to explore them individually and try to identify meaningful patterns. The new feature **Time** was converted in **parts of the day** for simplicity (0.00 - 6.00 into **night**, 6.00 - 12.00 into **morning**, 12.00 - 18.00 into **afternoon** and 18 .00- 24.00 into **evening**).

Nan values were explored column by column: some features presented lots of these values, and since the dataset was quite big, they were dropped when not necessary:

- The presence of **others** and **unknown** values was identified together with **Nan**. These values were dropped because the model needed to predict specific values to be meaningful.
- With binary features, **nan** values were considered legitimate entries as they referred to **0** values, as opposed to 1 values. In these circumstances they were kept and converted into **0** values.
- In other circumstances, **Y** values both referred to **Y** and **1** values, and **N** values to **0** and **Nan** values. Thus, the feature was binarised to **0** and **1** values for **N** and **Y** entries.
- The labeled data included an **unknown** class with **21636** entries that was entirely dropped.

Then, the original values scale for the labeled output was changed for clarity, as shown in the table on the right.

OLD VALUE	NEW VALUE	DESCRIPTION
1	1	Property Damage Only
2	2	Injury
2b	3	Serious Injury
3	4	Fatality

The labeled data showed that the dataset was **extremely imbalanced**: the vast majority of data referred to **property damage only** and **injury collisions**. This result was expected because

collisions involving **fatalities** and **serious injuries** are obviously and luckily less frequent.
This imbalance in data distribution was a problem because algorithms used in classification problems are badly affected by imbalanced data. As a solution, data was balanced out after exploration to improve model quality.

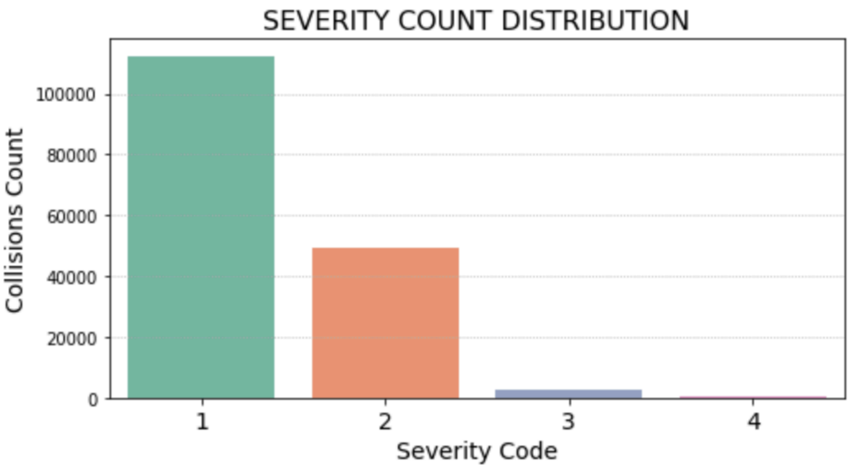


Fig 1: Extremely imbalanced labeled data distribution.

2.3. Feature selection

After the initial feature reduction due to their irrelevance, other features were removed. **Person count**, **pedestrian count**, **bicycle count**, **vehicle count** and **hit parked car** were dropped. These features possessed interesting information for data exploration, but were not relevant for predicting the severity of a car accident. They were considered as consequences of car collisions, not factors that could have been monitored to inform drivers about their level of risk of incurring into an accident. The correlation of independent variables was inspected and did not show any significant correlations between features:

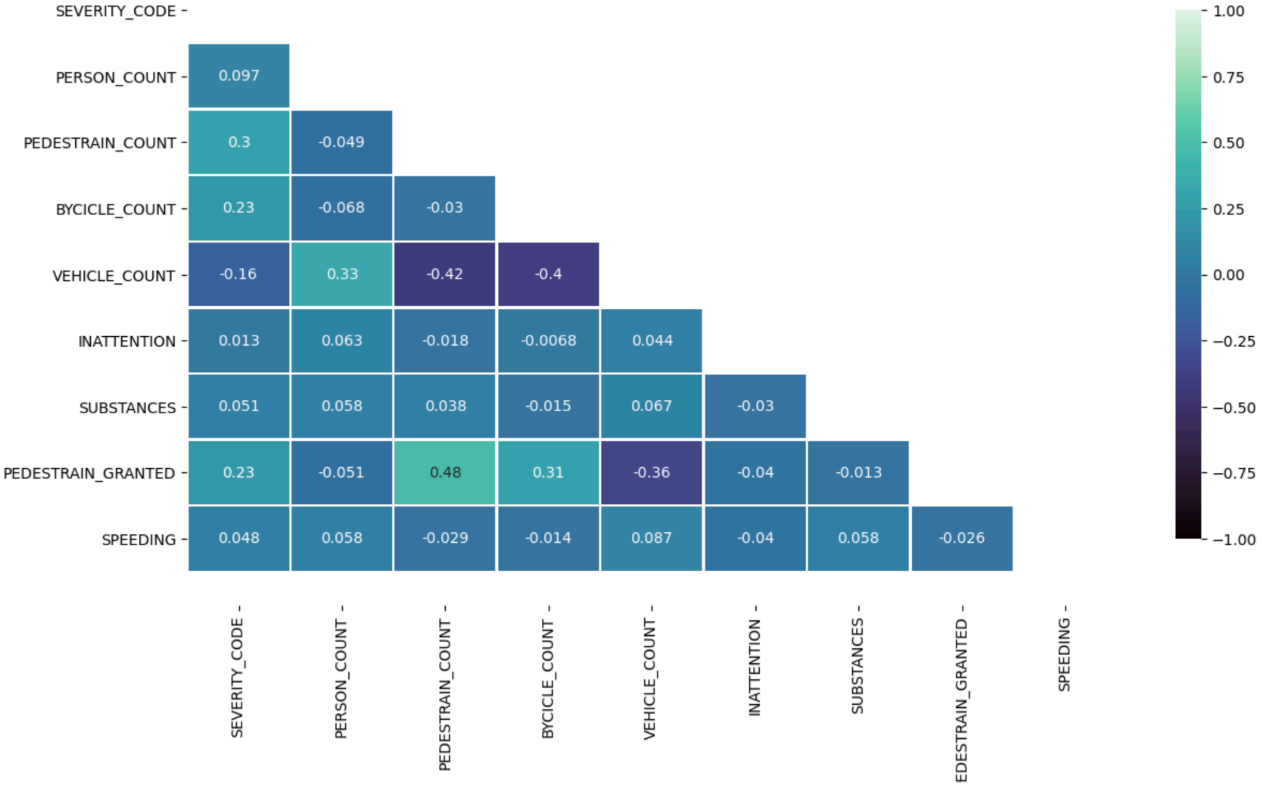


Figure 2: Heat-map showing the poor statistical correlation of the selected features for modelling.

Only **pedestrian_count** and **pedestrian_granted** were positively correlated, but their correlation was not strong enough to be considered significant.
The presence of outliers was visualised with **scatterplots**: they were identified and considered as legitimate entries because the dataset was imbalanced and these values were equivalent to the extremities of our dataset (labelled values 3 and 4).
The final features chosen for module creation were:

- Collision time, day and month.
- Type of accident location and accident type.
- Road and light conditions.
- Weather conditions.
- Inattention while driving.
- Driving under the influence of alcohol or drugs.
- Speeding.
- Whether pedestrians were granted or not.

3. Exploratory Data Analysis

3.1. Collision severity (CS) in relation to time factors

Collision severity was analysed in relation to time factors and revealed:

- A similar distribution on a **daily**, **weekly** and **monthly** basis.
- Time of the day had an impact on it, with a higher rate of accident in the **afternoon** (maybe because there are more people on the road at that time of the day: after work, children school pick-ups, …).
- The day of the week impacted accident severity: **Friday** was the day with the highest rate.
- **October**, **November**, **January** and **June** were the worst months.

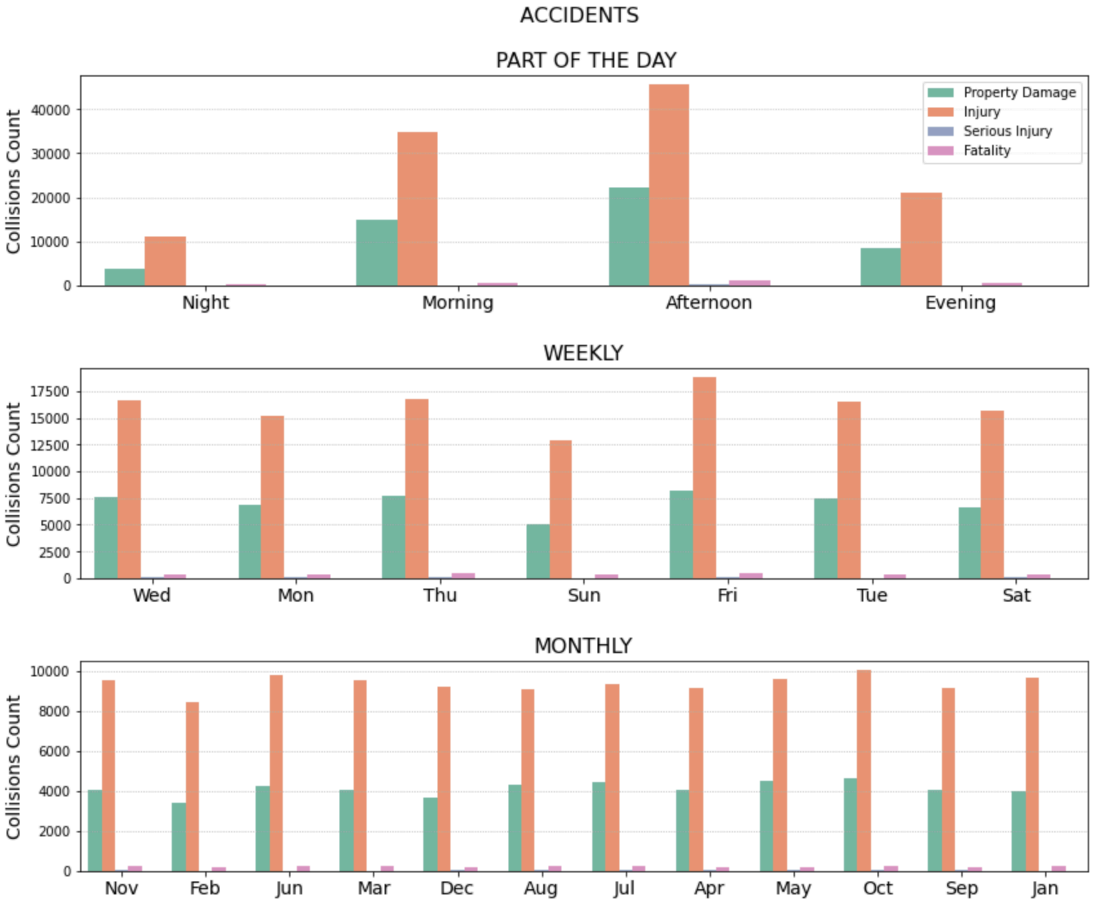


Fig 3: Collision Severity distribution. Part of the day, day and month.

3.2. CS in relation to location type & accident type

Accident location showed that **property damage only** collisions peaked at blocks and **fatalities** did not take place at alleys.

Accident types indicated that **property damage only** collisions peaked when a parked car was involved, **injuries** were the product of different types of collisions (mainly **rear ended** and **angles**) and **fatalities** peaked when accidents involve pedestrians (129).

3.3. CS in relation to road & light conditions

Daylight accidents were the most frequent and the absence of light **did not affect** collision severity distribution.

3.4. CS in relation to weather conditions

Weather conditions revealed an interesting result: the majority of the accidents took place with **clear weather conditions** (64.9%).

Raining and **overcast conditions** together represent 34.3% of the total.

This result could be due to many factors: either drivers were more cautious when driving in bad weather or the weather in Seattle is generally nice.

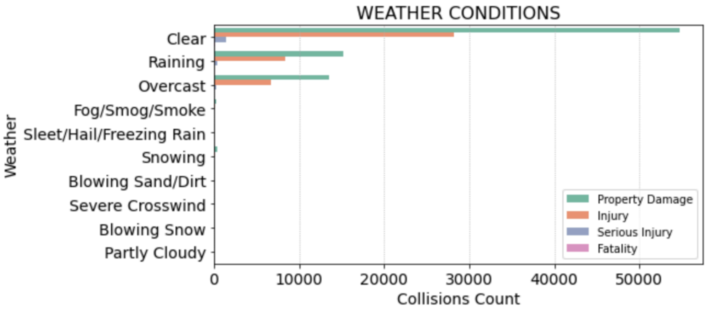


Fig 4: Collision severity in relation to weather conditions.

3.5. CS in relation to inattention

Most frequent accidents were not caused by distracted drivers. This information **deserved further investigation**. In general, drivers are expected to be reluctant to admit their inattention. For example, if they were not paying attention because they were using their mobiles while driving, the insurance company would not cover their costs. It was questioned if this information was reliable.

3.6. CS in relation to driving under the influence

The distribution of collision severity when driving under the influence of alcohol and drugs or not was very different. In fact, driving under the influence dramatically increased accident severity:

- **Fatalities** increased from 0.1% to 0.9%.
- **Serious injuries** increased from 1.5% to 4.3%.
- **Injuries** increased from 33.1% to 39.2%.

Interestingly, substances consumption varied during the day: they increased throughout the day and peaked in the **evening**.

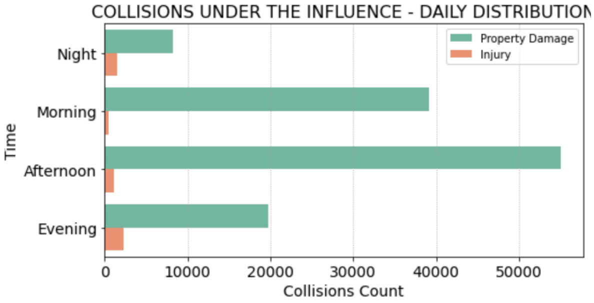


Fig 5: Daily variation of accident severity.

3.7. CS in relation to speeding

Like driving under the influence, speeding increased the accident severity in various ways::

- **Fatalities** from 0.1% to 0.7%.
- **Serious injuries** from 1.6% to 3.5%.

- **Injuries** from 33.1% to 40.3%.

3.8. Creating an Interactive Map

The **latitude** and **longitude** contained in the data after feature selection was used to create an interactive **Folium map** to explore the street of the city of Seattle and visualise the distribution of the collisions. It revealed that the majority of accidents took place in the city centre.

3.9. Convert Categorical Data into Int64 Data

To be ready for modelling, **object** and **categorical data types** needed to be converted into **int64** data types. The method **pd.get_dummies** for one hot encoding was used so that the column names of category was preserved.

4. Predictive Modelling

4.1. Preparing the imbalanced multi-class data set for modelling

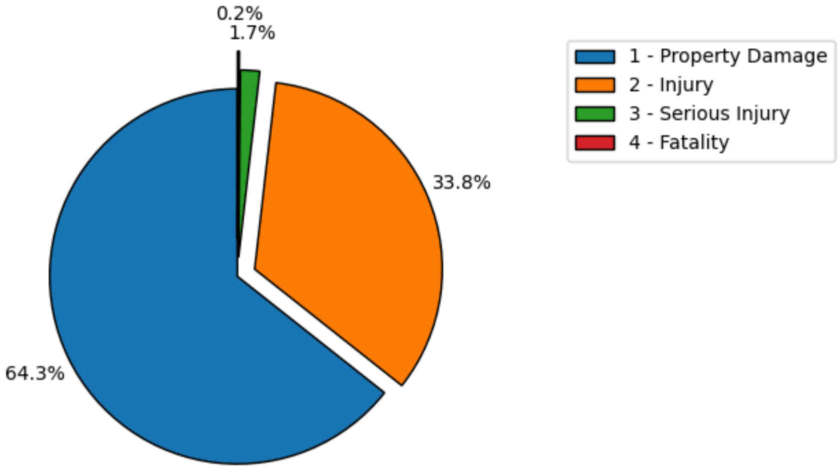


Fig 6: Four-class labeled data, imbalanced distribution.

The cleaned dataset had a total of **127505 entries** and the four classes of our labeled data were not represented equally, as shown on the pie-chart on the right.

Creating a classification model with this **multiple-class imbalanced dataset** presented some challenges. If the predictive model was trained with this type of imbalanced data, since data was heavily biased towards **Class 1 (property damage)**, the model would have over-fitted on this class label and ended up predicting it in most of the cases and **not effectively predicting the minority classes**.

If the accuracy of such a model had been tested, the result would have been a great score. However, this would have been a misleading result: the model **may have never classified Class 3 (serious injuries**) or **class 4 (fatalities)**. In fact, with imbalanced classes, it is easy to get a high accuracy without actually making useful predictions. Accuracy as an evaluation metric makes sense only if the class labels are uniformly distributed.

To improve model prediction, the dataset was balanced:

- The majority class/classes (the over-represented classes) was **under-sampled** (or removed).
- The minority class/classes (the under-represented classes) was **over-sampled** (or added). It is important to notice that over-sampling did not introduce new information in the dataset, it only shifted it around so as to increase the "numerical stability" of the resulting models.
- A slight imbalance was maintained to reflect the original dataset distribution.

CLASS	ACTION
1 - Property Damage	Randomly under-sampled to 35000
2 - Injury	Randomly under-sampled to 30000
3 - Serious Injury	Over-sampled to 25000
4 - Fatality	Over-sampled to 25000

After balancing the dataset, another problem aroused. When machine learning algorithms are applied to multiple-class classification problems with balanced datasets, they normally do not perform well. There are different possible strategies to improve the model performance and the solution adopted in this project was **binarising the labeled data**. The four classes in the labeled output were turned into two classes: property damage, injury and serious injury were merged into a new class **non-fatality**, while fatality was preserved:

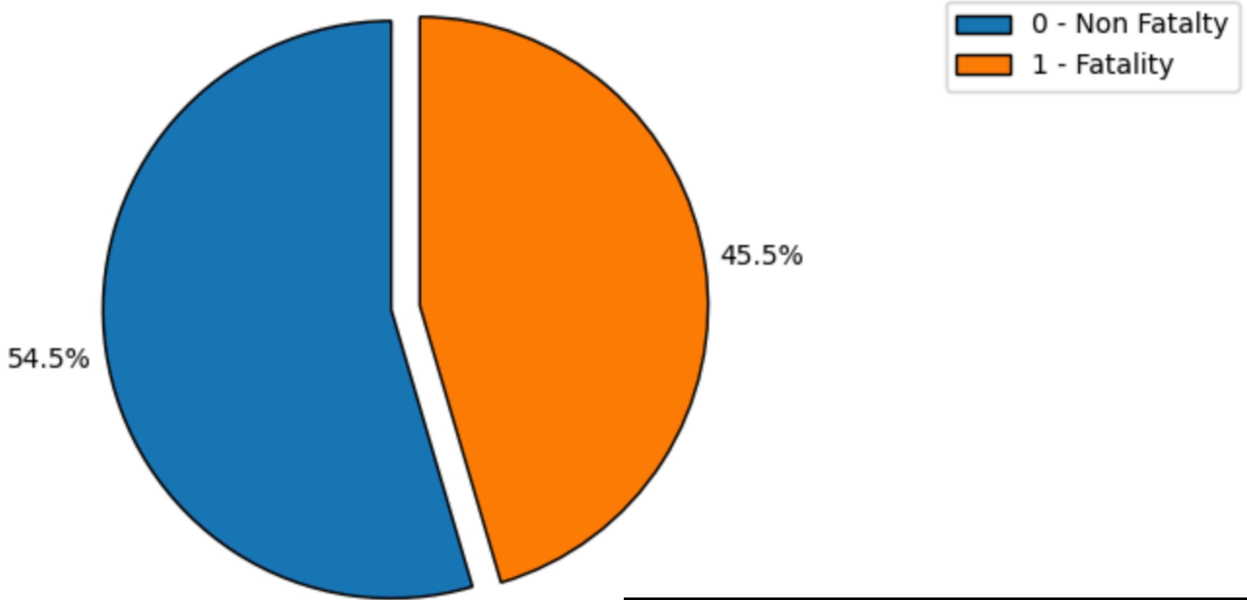


Fig 6: Binarised labeled data.

OLD CLASS	NEW CLASS
1 - Property Damage	**1 - Non_Fatality**
2 - Injury	**1 - Non_Fatality**
3 - Serious Injury	**1 - Non_Fatality**
4 - Fatality	**2 - Fatality**

Four different algorithms were used and their performances were tested:

- K Nearest Neighbours.
- Decision Trees.
- Logistic Regressions.
- Support Vector Machine.

The process followed for model creation was the same for each algorithm:

1. Data was **scaled**: machine learning models require each feature value to be close to zero or that all features vary on comparable scales. Data was put in a format that works better for algorithms.
2. The dataset was split data into **Train** (67% of the total) & **Test** (33 %) **sets**.
3. The specific algorithms **hyper-parameters** were optimised.

The metrics used to evaluate model performance were:

- **Accuracy**, **Precision** and **Recall**.
- **Confusion-Matrix** (to visualise more accurately what the models predict).

4.2. Specific algorithm adjustment

K Nearest Neighbours is an algorithm that can be both used for regression and classification predictive problems. It uses feature similarity and considers the **k** nearest neighbours (or points) to predict the classification of the test point.
The value for k as a great impact on the model accuracy: firstly k was randomly set at **2**, then the model accuracy was tested by looping the value of k from 1 to 15. The value that returned the best accuracy score was **3**.

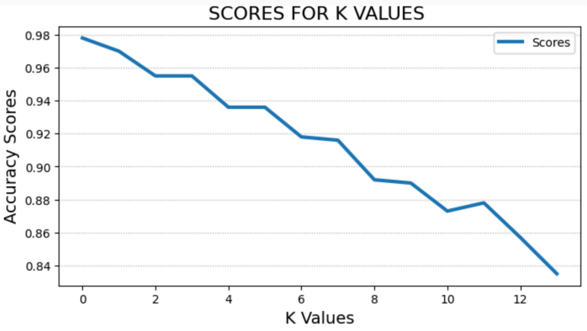


Fig 7: K value optimisation

Logistic regression is used to predict the probability of a target variable. In this project, this variable was binomial because it could have only two values (Fatality, Non_fatality). The model could calculate this probability using different types of numerical optimisers including Newton-cg, lbfgs, liblinear, Sag and Saga solvers. After testing these parameters, the **liblinear** solver was chosen as it produced the best results.

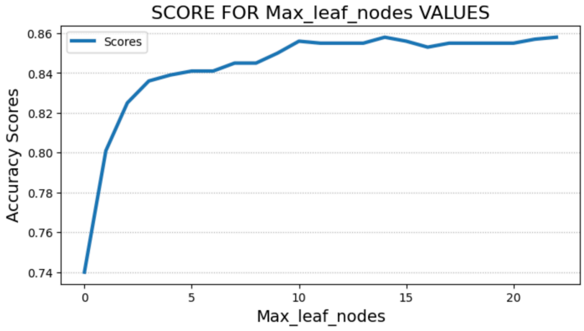


Fig 8: Max_leaf_nodes value optimisation.

Decision trees is a predictive model tool that can be applied across many areas. The dataset is split in different ways based on different conditions creating a nodes/leaves structure. The **max_leaf_nodes** (maximum number of leaf nodes) has an impact on the model accuracy: this value was randomly set at 10**,** then the model accuracy was tested while looping the value of max_leaf_nodes from 2 to 25. The value that returned the best accuracy score (while avoiding over-fitting) was **11**.

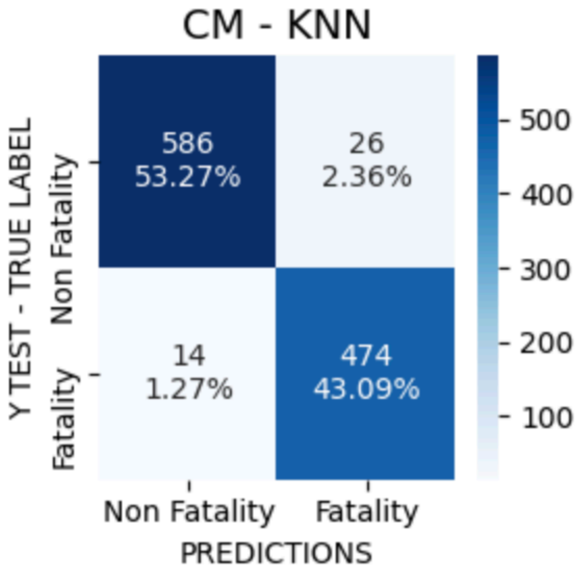
Support vector machine (SVM) is a very versatile algorithm that can handle multiple continuous and categorical variables. It is a representation of different classes in a hyperplane in a multidimensional space. The hyperplane is generated in an iterative manner by the algorithm so that error can be minimised. The model divides the dataset into classes and found the maximum marginal hyperplane. The mathematical function **radial basis** (rbf) was used to map data into a higher dimensional space as it was the one that was performing better.

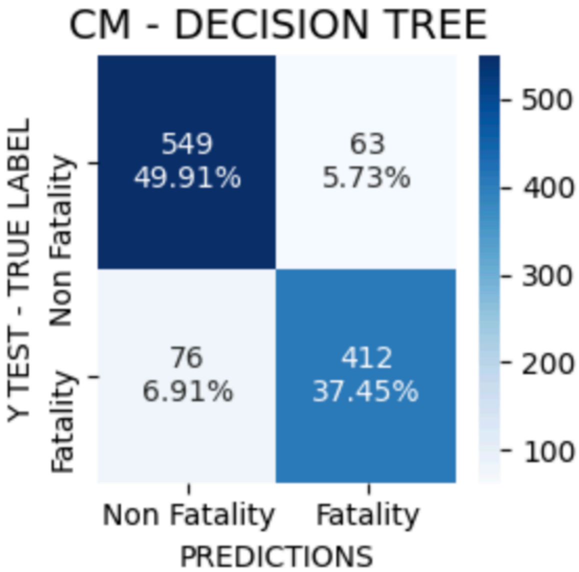
4.3 Model performances

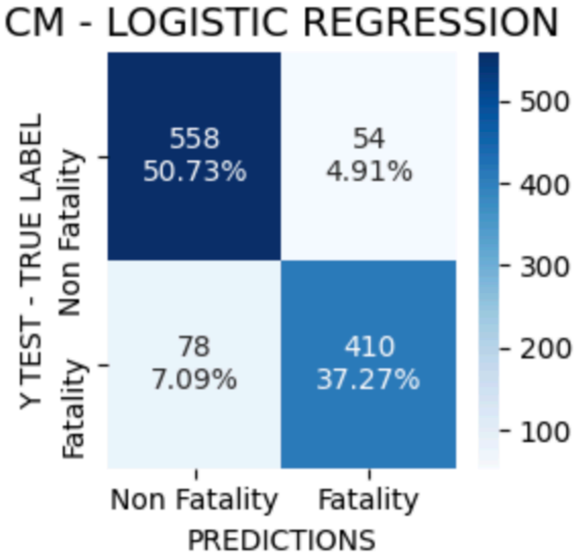
The four different classification model performed extremely well and the doubt that they were overfitting was strong. The high model accuracies was confirmed by their **confusion matrix**, thus eliminating the suspect that the model was not equally predicting **true values**. As it is possible to see in the following table, the **K nearest neighbours** model as the one that scored the best performances in terms of precision, recall and F1-score:

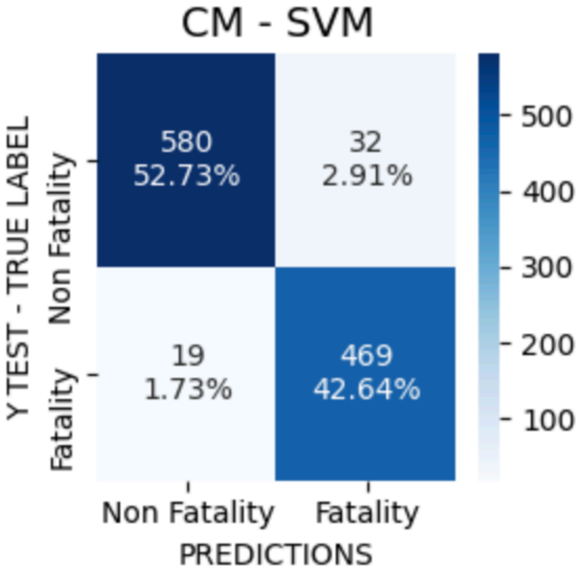
MODEL	VALUE	PRECISION	RECALL	F1-SCORE	SUPPORT
KNN	0	**0.98**	**0.96**	**0.97**	612
	1	**0.95**	**0.97**	**0.96**	488
DECISION TREE	0	0.88	0.90	0.89	612
	1	0.87	0.84	0.86	488
LOGISTIC REGRESSION	0	0.88	0.91	0.89	612
	1	0.88	0.84	0.86	488
SVM	0	0.97	0.95	0.96	612
	1	0.94	0.96	0.95	488

The confusion matrices of the predictive models show that the distribution of True/False Negative and True/False Positive is even:









This positive result was probably due to the fact that the **under-represented data was oversampled 10 times**. This meant that the dataset did not posses a great variety and that part of the **test set** already contained most of the data included in the **test set**.

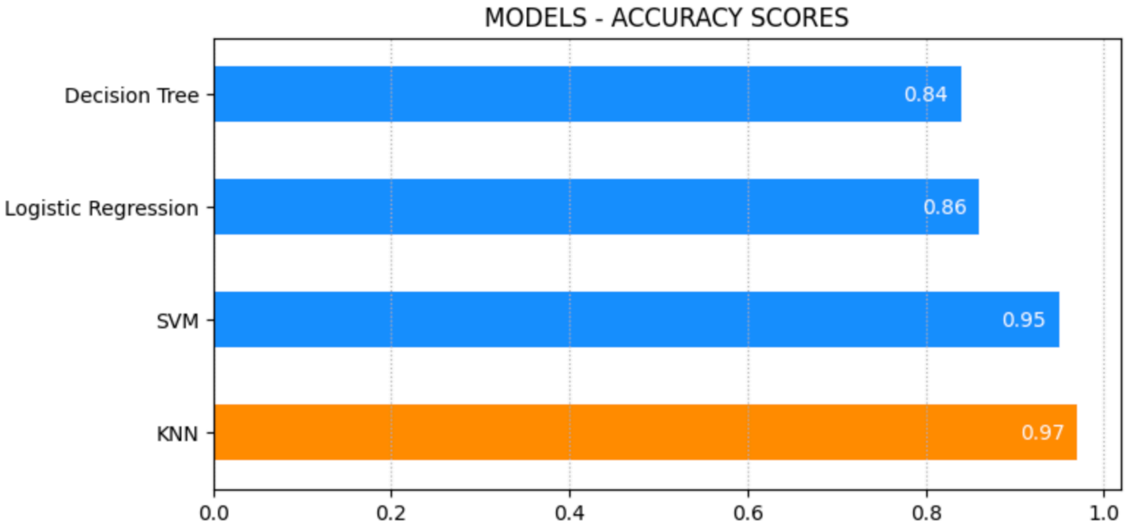


Fig 9: Final models accuracy scores.

5. Conclusions

In this project, the relationship between **accident severity** and data relating to the recorded accidents was analysed. The number of features contained in the dataset was reduced in relation to their significance for model creation. Co-dependencies amongst features were researched during the exploratory analysis phase.

The **labeled data** was imbalanced and to prepare the data for modelling it was **balanced out** and then **binarised**. The predictions of four different classification models were compared and the best one was chosen: the best model accuracy was very high (**97%** achieved using KNN algorithm). This high accuracy was confirmed by the **confusion matrix**, thus eliminating the suspect that the model was not equally predicting **true values**. This result is probably due to the fact that under-represented data was oversampled 10 times. This meant that the dataset did not posses a great variety and that part of the **test set** already contained the data included in the **test set**.

The method used to collect some features deserves further investigation: it would be interesting to know how the information relating to **inattention** was obtained. Drivers were expected to be reluctant to admit their inattention. For example, if they were not paying attention because they were using their mobiles while driving, the insurance company would not have covered their costs.

Data exploration revealed some interesting information that should be shared with the general population to increase their sense of awareness on the topic:

- The collision yearly distribution showed a **downtrend** from 2006 to 2010 followed by an uptrend and then another downtrend from 2015 to 2019. This means that there is still a lot to do to address this issue.
- The highest rate of collisions took place in the **afternoon**, on **Fridays** in **October**, **November**, **January** and **June**.
- Collisions were more frequent at blocks and when a parked car is involved.
- Injuries were mainly the product of collisions at rear ended and angles.
- Fatalities peaked when accidents involved **pedestrians**.
- Unexpectedly, the majority of collisions took place with **clear weather conditions** (64.9%). **Raining and overacted conditions** together represented 34.3% of the total.

- The majority of the accidents took place when the road condition was **dry**, followed by **wet** conditions.
- **Daylight accidents** were the most frequent. The absence of light did not affect the severity distribution.
- Collisions caused by drivers driving under the influence were less frequent, but this factor increased the collisions severity: **fatalities** +0.8%, **serious injuries** +2.8%, **injuries** +6.1%.
- Speeding increased the accident severity: **fatalities** +0.6%, **serious injuries** +1.9%, **injuries** + 7.2%.

6. Future Directions

A further research would have to take into account of other information. It seems that the majority of the accidents took place when the **weather condition was clear**. It is not clear if this was due to the fact that drivers are generally more careful with bad weather or simply because the overall weather in Seattle is nice. Data relating to weather conditions will have to be compared to the one collected in order to give a meaning tho this outcome.

A different way of dealing with a **classification problem with imbalanced data** could have been maintaining the original four classes and use more **complex predictive models**. The ones used in this project were tested on the cleaned multi-class dataset and never returned an accuracy score higher than 50%. This type of solution will be implemented in future projects.

Different datasets relating to this topic from different parts of the world could be tested to establish common trend and points of differentiation.