Predicting Car Accident Severity

Paulina Leiva

October 15, 2020

1. Introduction/Business Problem

1.1 Background

More than 3,000 people die every day in the world, as a result of traffic accidents.These are the leading cause of death in young people. The World Health Organization (WHO) qualifies the situation of traffic accidents as a priority public health problem throughout the world .Citizens in general are not aware of the magnitude of the problem of traffic accidents in the world, nor is it possible to quantify the number of people who, as a result of a traffic accident, are disabled for life and yet appear in the figures only as injured.In addition to the loss of human life, traffic accidents produce a huge economic impact, which, directly or indirectly, all citizens bear. It is estimated that on average the costs of traffic accidents reach 3% of the GDP of a country according to WHO figures, if this money were invested in education, housing, health, social assistance, imagine the benefit it would represent for our society.hese are the leading cause of death in young people.

1.2 Problem

In this proyect will build a model to predict the severity of an accident.The problem that this model seeks to solve is the ignorance that drivers have about the probabilities of an accident on a certain route and its severity, due to different variables such as the weather and road conditions, among others, that makes it impossible for drivers to make decisions in advance, such as driving more carefully or changing routes.

1.3 Interest

Not only drivers are interested in knowing the probabilities of a traffic accident to choose a better route, be careful when driving or postpone their trip. But it is also considered a global health problem, so it is in the interest of those responsible for transport in each country, to better understand the contributing factors and the relationships between them in order to introduce specific awareness campaigns and programs to reduce costs. road safety incidents, seeking as a major goal to reduce in this way the people affected by their health and the GDP expenditure of each country in traffic accidents.

2. Data acquisition and cleaning

2.1 Data sources

It was decided to use the data set called Data-Collisions.csv, which shows data provided for this project from the Seattle Department of Transportation for the period between 1st January 2004 and 20th May 2020. This data set conteint a total of 194,673 collision incidents with 37 attributes .

2.2 Data cleaning

There are some problems in the data set. First, it was identified that there are some variables that have many missing values, more than 50% of the rows have no values, for this reason it was decided to eliminate these variables: INTKEY, EXCEPTRSNCODE EXCEPTRSNDESC, INATTENTIONIND, PEDROWNOTGRNT, SDOTCOLNUM, SPEEDING. Besides the variable SEVERITYCODE is duplicated, one of the columns is eliminated.

Second, for the purpose of the model there are variables that are not useful to use, which is why they are also eliminated, this happens with the location variables X, Y and LOCATION, and the date and time variables, INCDATE,INCDTTM. Also, the HITPARKEDCAR variable is eliminated because most of its values are NO (N), STATUS eliminated because only had one value Unmatched, the rest are Matched and the SEGLANEKEY, CROSSWALKKEY variables are also eliminated because most of its values are 0.

In addition to eliminating variables, it is necessary to transform variables, in our case the UNDERINFL variable, a variable that may be important for the future model because it tells us whether or not a driver involved was under the influence of drugs or alcohol, has values Y (yes) and N (No), these values are transformed into 0 and 1 respectively.

Finally, it is also necessary to eliminate rows with missing values in some variables. There are several variables that can be considered as possible important variables for the model that have empty values, that is why those rows must be eliminated ROADCOND, UNDERINFL, ADDRTYPE, COLLISIONTYPE, JUNCTIONTYPE, WEATHER, LIGHTCOND. With the previous action, the data set went from having 194,673 collision incidents to having 182,895.

2.3 Feature selection

For the construction of the model, it is decided first not to use any description variable, for example, there is the variable SDOT_COLDESC that describes what the value of the variable SDOT_COLCODE means, this is repeated with other variable ST_COLDESC, none of these 2 variables will be selected.

On the other hand, there are variables that provide similar information, redundant variables. There are two variables that are collision codes SDOT_COLCODE (a code given to the collision by SDOT) and ST_COLCODE (a description that corresponds to the state's coding designation), it is decided to work with the latter because we have the description of what each value of the code. The same happens with the unique IDs for each collision, there are 3 OBJECTID, INCKEY and COLDETKEY, it is decided to only consider INCKEY, which is the unique key for the incident.

After ruling out redundant features, I inspected the correlation of independent variables, and not found pairs that were highly, you can see in Table 1. Finally, 15 features were selected.

	SEVERITYCODE	INCKEY	PERSONCOUNT	PEDCOUNT	VEHCOUNT	UNDERINFL
SEVERITYCODE	1.000000	0.034259	0.124545	0.245656	-0.081166	0.035763
INCKEY	0.034259	1.000000	-0.049568	0.032564	-0.014410	0.702275
PERSONCOUNT	0.124545	-0.049568	1.000000	-0.027211	0.399674	-0.028545
PEDCOUNT	0.245656	0.032564	-0.027211	1.000000	-0.317361	0.029523
VEHCOUNT	-0.081166	-0.014410	0.399674	-0.317361	1.000000	-0.014674
UNDERINFL	0.035763	0.702275	-0.028545	0.029523	-0.014674	1.000000

Table 1: Correlation of numerical variables

3. Exploratory Data Analysis

3.1 Target variable

To build a model that predicts the severity of a car accident with the dataset we have, the most logical thing is to use SEVERITYCODE as the target variable. This variable is a code that corresponds to the severity of the collision, the possible values are:

- 3 — fatality
- 2b — serious injury
- 2 — injury
- 1 — prop damage
- 0 — unknown

But within our dataset there are only values 1 and 2, and has no missing values. It can be seen in figure 1, number of accidents per collision severity, that approximately 2/3 of the accidents are "prop damage", in other words, most accidents are of low severity.

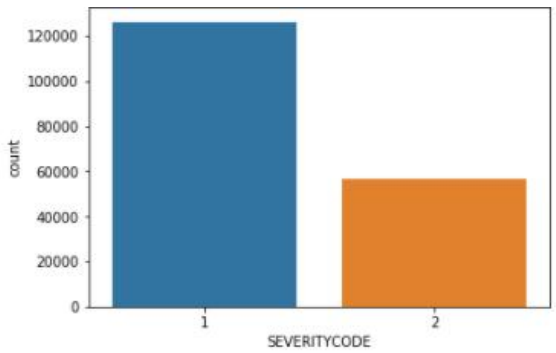


Figure 1: number of accidents per collision severity

3.2 Relationship between ADDRTYPE and SEVERITY of Accidents

In our models, the target variable is a binary categorical variable, it can only be worth 1 or 2, and most of the possible independent variables are also categorical, the catplot type of the seaborn library allows us to graph the relationship between two categorical variables.

The independent variable ADDRTYPE, collision address type, has 3 possible values:

- Alley
- Block
- Intersection

We can see in Figure 2, which seems to be related to our target variable, that many more accidents occur in "Intersection" and "Block" than those that occur in "Alley". Therefore, it is decided to consider the variable in the models to be built.

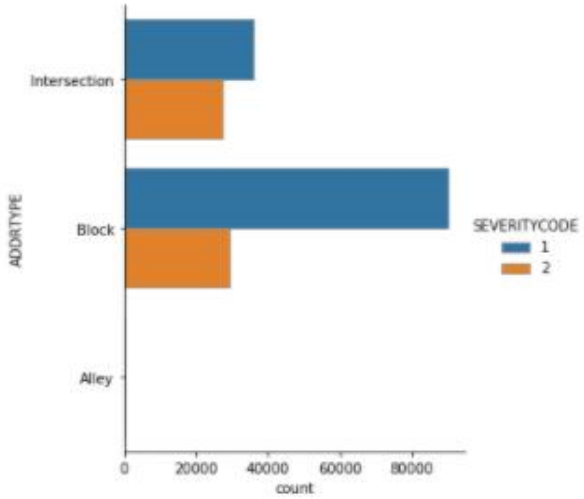


Figura 2: Relationship between ADDRTYPE and SEVERITY of Accidents

3.3 Relationship between COLLISIONTYPE and SEVERITY of Accidents

The independent variable COLLISIONTYPE shows us how its name says Collision type of the accident. We can see in figure 3, that there is not a prevalence of one type or another, and less a differentiation of gravity according to the type of collision, therefore, it is decided not to consider this variable in the models to be built.

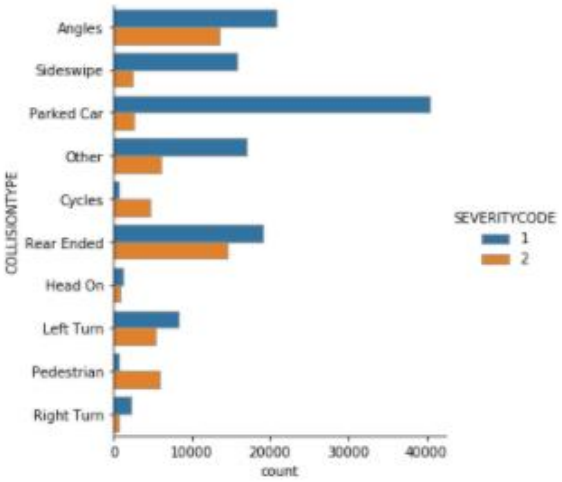


Figura 3: Relationship between COLLISIONTYPE and SEVERITY of Accidents

3.4 Relationship between VEHCOUNT and SEVERITY of Accidents

The independent variable VEHCOUNT shows the number of vehicles involved in the collision, this is entered by the state. We can see that accidents with one or two vehicles predominate, but the most interesting thing about this relationship is that accidents with one vehicle in general are more serious than those with two vehicles, these make us decide that it is an interesting variable to build the models.

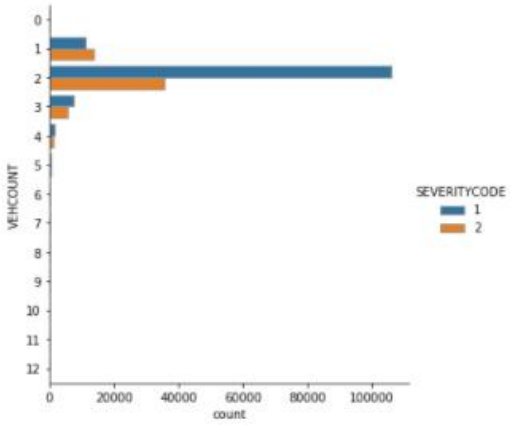


Figura 4: Relationship between VEHCOUNT and SEVERITY of Accidents

3.5 Relationship between PERSONCOUNT,VEHCOUNT and SEVERITY of Accidents

The independent variable PERSONCOUNT shows The total number of people involved in the collision. Intuition tells us that the number of people involved should be related to the number of vehicles, that is why through relplot the relationship between the number of vehicles, people involved and the severity of the accident is reviewed. Figure 5 shows the strong relationship

between the number of vehicles and the number of people involved, since the VEHCOUNT variable will already be considered in the models, the PERSONCOUNT variable is discarded.

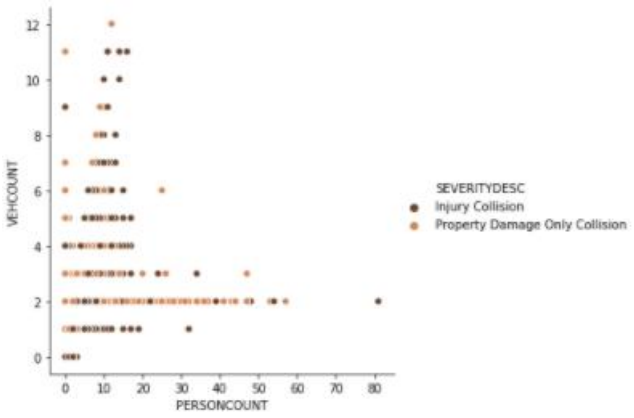


Figura 5: Relationship between PERSONCOUNT,VEHCOUNT and SEVERITY of Accidents

3.6 Relationship between PEDCOUNT,VEHCOUNT and SEVERITY of Accidents

The independent variable PEDCOUNT shows the number of pedestrians involved in thecollision, this is entered by the state. Although in this case the intuition is less clear regarding the relationship between the number of pedestrians, the number of vehicles and the severity of the accident, it is also decided to analyze them together. When graphing it, as seen in figure 6, it is noted that if there is a strong relationship between the number of vehicles and the number of passengers, which is why this variable is discarded as a possible variable for the models to be built.

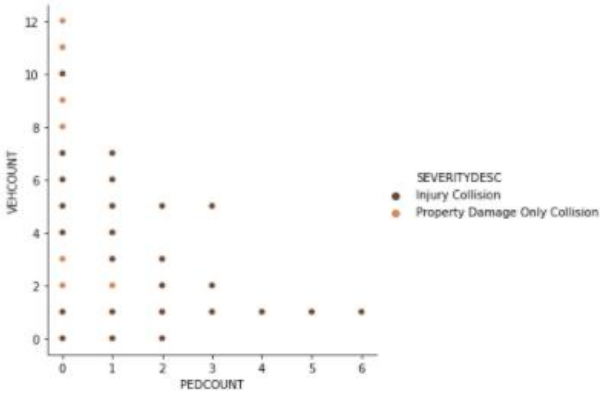


Figura 5: Relationship between PEDCOUNT,VEHCOUNT and SEVERITY of Accidents

3.7 Relationship between JUNCTIONTYPE and SEVERITY of Accidents

The JUNCTIONTYPE categorical variable shows the category of junction at which collision took place. Figure 7 shows the relationship between this variable and the number / severity of accidents, we can see that depending on the type of crossing there may be more or fewer accidents, that there are markedly 4 types of junction where most accidents occur. On the other

hand, it can be seen that depending on the type of junction, certain gravity is also more likely than another. For all the above, it is decided to consider the variable for future models.

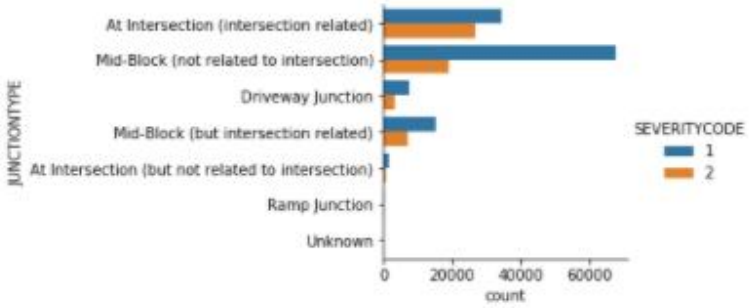


Figura 7: Relationship between JUNCTIONTYP**E** and SEVERITY of Accidents

3.8 Relationship between UNDERINFL and SEVERITY of Accidents

The UNDERINFL variable shows whether (0) or not (1) a driver involved was under theinfluence of drugs or alcohol. Intuition tells us that an accident is more likely to occur under the influence of drugs or alcohol, but it can be seen in image 8, where the relationship between the UNDERINFL variable and the objective variable is plotted, that there is practically no difference in quantity or severity of accidents. Therefore, it is decided not to use the variable in future models.

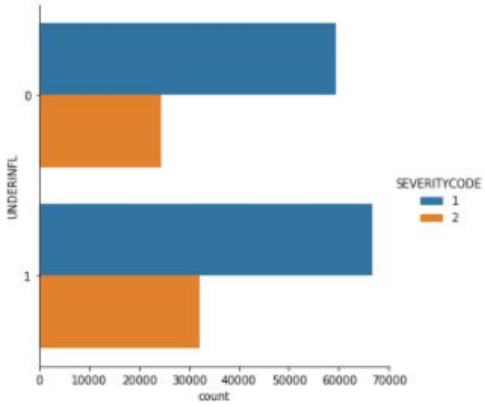


Figura 8: Relationship between UNDERINFL and SEVERITY of Accidents

3.9 Relationship between WEATHER and SEVERITY of Accidents

The WEATHER categorical variable is a description of the weather conditions during the time of the collision. Figure 9 clearly shows that the relationship between climates and accidents is strong, the vast majority of accidents occur in 3 weather conditions "overcat", "rainning" and "clear", and that within these conditions accidents occur with greater severity (2) when there is "clear". For all the above, the variable is considered for the models to be built.

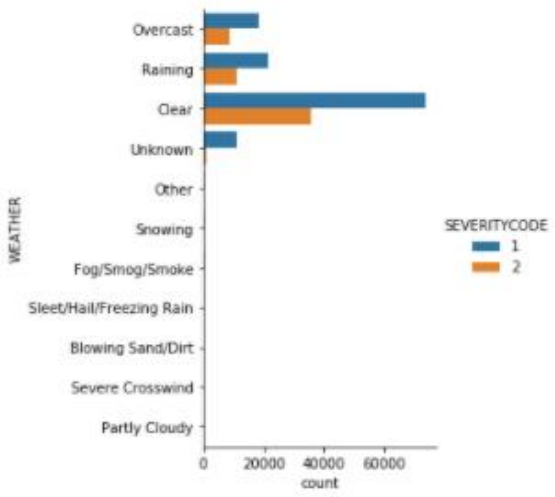


Figura 9: Relationship between WEATHER and SEVERITY of Accidents

3.10 Relationship between ROADCOND and SEVERITY of Accidents

The ROADCOND variable shows the condition of the road during the collision. Intuition tells us that it exists relationship between this variable and the objective variable, accident severity. It can be seen in figure 10, that this relationship exists and is strong, accidents occur markedly in 2 road conditions, "Wet" and "Dry", and serious accidents occur mostly in "Dry" conditions. All of the above makes it reasonable to consider this categorical variable in future modeling.

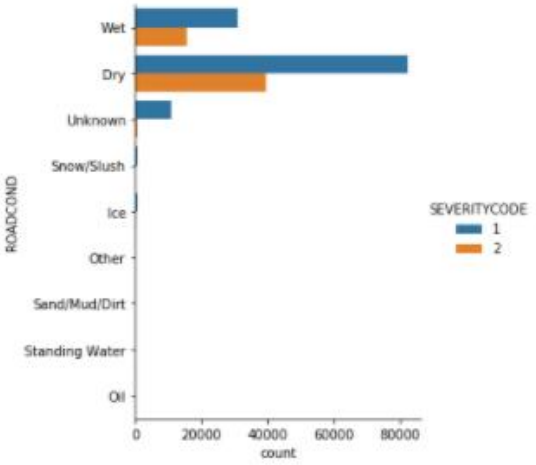


Figura 10: Relationship between ROADCOND and SEVERITY of Accidents

3.11 Relationship between LIGHTCOND and SEVERITY of Accidents

A diferencia de ROADCOND que muestra las condiciones del camino, LIGHTCOND muestra the light conditions during the collision. La intucion tambien nos dice que la cantidad de luz debe influir en la probabilidad de que ocurra un accidente, lo que se confirma con la figura 11, donde de aprecian que los accidentes ocurren con "daylight" y "dark-street light on", lo que va contra la intuic

ion de que es más probable que ocurra un accidente con menos luz pero que si confirma que puede ser influyente esta variable en la ocurrencia y gravedad de una colision, se decide por lo mismo utilizar la variable.

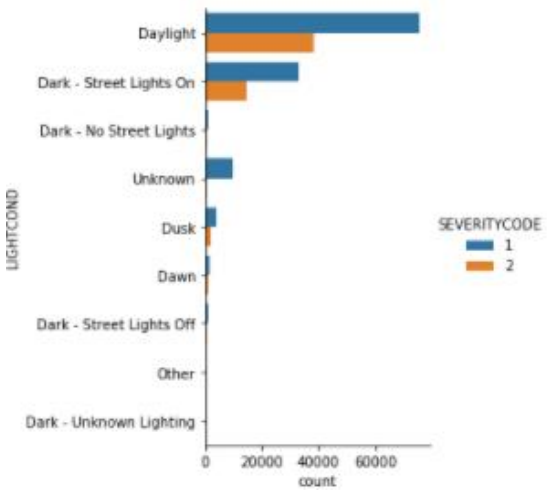


Figura 11: Relationship between LIGHTCOND and SEVERITY of Accidents

4. Machine Learning Models

4.1 Data Preparation for Models

As is clear from the figure1, SEVERITYCODE shown above, the vast majority of accidents involve "property damage only collision" , and only a 1/3 of accidents involve "injury collision". If we train a classification model on these data, the model will be biased. To fix this issue we need to resample the data. The new dataset contains the same number of collisions for each severity, each with 56625 collisions.

In our data set, the objective variable is categorical, for this reason the best models that we can use are machine learning models, in our case we will build 4 different models:

- KNN
- Decision tree
- SVM
- Logistic Regression

For implementing the ML-Predictive modeling, I have used Github as a repository and IBM Watson studio for running Jupyter Notebook to preprocess data and build Machine Learning models. Regarding coding, I have used Python and its popular packages such as Pandas, NumPy and Sklearn.

Based on exploratory data analysis, I can see that the categorical variables have key impact in predicting severity in accidents hence I have selected these most important features to predict the severity of accidents in Seattle. Among all the features, the following features have the most influence in the accuracy of the predictions:

- “WEATHER”,

- “ROADCOND”,
- “LIGHTCOND”
- “ADDRTYPE”
- “JUNCTIONTYPE”
- “VEHCOUNT”

Also, as I mentioned earlier, “SEVERITYCODE” is the target variable.

Features in this dataset are categorical ADDRTYPE, JUNCTIONTYPE, WEATHER, ROADCOND, LIGHTCOND. Sklearn KNN, Decision Trees, Logistic models do not handle categorical variables. For solution that, we convert these features to numerical values.

In order to develop a model for predicting accident severity, the re-sampled, cleaned dataset was split in to testing and training sub-samples (containing 20% and 80% of the samples, respectively) using the scikit learn “train_test_split” method. In total, four models were trained and evaluated.

	SEVERITYCODE	INCKEY	REPORTNO	ADDRTYPE	COLLISIONTYPE	SEVERITYDESC	PERSONCOUNT	PEDCOUNT	VEHCOUNT	JUNCTIONTYPE	UNDERINFL	WEATHER	ROADCOND	LIGHTCOND	
129091	1	168816	3581067	1	Sideswipe	Property Damage Only Collision	2	0	2	4	1	1	0	5	
175353	1	277820	3753800	2	Left Turn	Property Damage Only Collision	5	0	2	1	1	1	0	5	
110094	1	137477	3593833	1	Parked Car	Property Damage Only Collision	2	0	2	4	1	1	0	2	
46167	1	67135	2795005	2	Left Turn	Property Damage Only Collision	5	0	2	1	0	1	0	5	
38310	1	58699	2410680	1	Parked Car	Property Damage Only Collision	2	0	2	3	0	4	8	5	

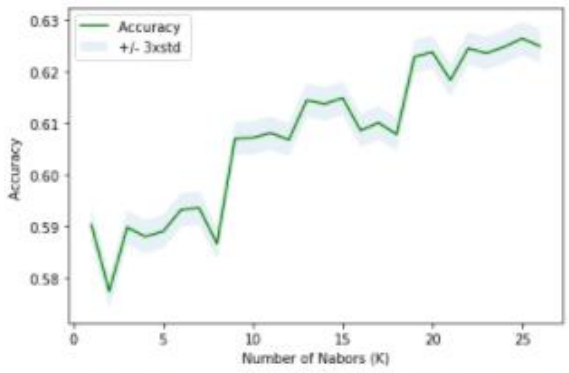
Table 2: table with variables converted from categorical to numerical

4.2 K Nearest Neighbor (KNN)

kNN models seek to categorise the outcome of an unknown data sample based on its proximity in the multi-dimensional hyperspace of the feature set to its “k” nearest neighbours, which have known outcomes.

Here while implementing KNN algorithm, I have taken it for set of 27 values the model will be trained on training set of data, and then predicting the values based on test data set, Atlast the model accuracy will be calculated by comparing actual values with predicted values of test data set.

While evaluating model, plotting all 27 accuracy values, so the highest value of is the best accuracy KNN Model. In this case it is for k=25 we have got maximum accuracy of 0.63




```
The best accuracy was with 0.6265342163355409 with k= 25
```

Figura 12: Accuracy for k between 1 and 27

4.3 Decision Tree

Decision tree models identify the key features on which the data can be partitioned (and the thresholds at which to partition the data) in the hope of arriving, after some iterations, at “leaves” which contain only accidents belonging to one target variable value (in this case, accident severity code). A decision tree model was trained on the data according to the “entropy” criterion.

As when implementing KNN, where the K that would give us the best fit through accuracy was sought, in this case the best number of levels(Depth) was sought, iterating from 3 to 10 depth nivel, where the model with the best indicators of f1_score and Jaccard returns. In this case, the best model is achieved with max_depth = 7, where jaccard = 0.64 and f1_score = 0.64. The accurrecy with max_depth is 0.64

```
from sklearn.tree import DecisionTreeClassifier

jaccard_array=[]
f1_score_array=[]

md_initial=3

for md in range (md_initial,10,1):
    dt=DecisionTreeClassifier(criterion='entropy', max_depth=md)
    dt.fit(X_train,y_train)
    dt_yhat=dt.predict(X_test)

    jaccard=jaccard_similarity_score(y_test,dt_yhat)
    f1=f1_score(y_test,dt_yhat,average='weighted')

    jaccard_array.append(jaccard)
    f1_score_array.append(f1)

print(f'Best value for max depth = {jaccard_array.index(max(jaccard_array))+md_initial}')
print(f"Evaluation Jaccard = {jaccard}")
print(f"Evaluation F1 score = {f1}")
```

```
Best value for max depth = 7
Evaluation Jaccard = 0.6381456953642384
Evaluation F1 score = 0.6378330080896863
```

Figura 13: Python code iteration to find better decision tree

4.4 Support Vector Machine

SVM models seek to separate data based on different values of the target variable by mapping the dataset to a higher-dimension space and identifying the support vectors which best-describe the hyper planes that most effectively partition the data. An SVM model was built using the scikit learn C-Support Vector Classification method (svm.svc), with a "rbf" mapping kernel employed in order that the model could return a list of the features with the most diagnostic power for determining accident severity.The built model has an accuracy of 0.64. The performance will be evaluated through f1_score and jaccard in the next section.

4.5 Logistic Regression

Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression. A Logistic Regression model was trained using an inverse-regularisation strength C=0.01, and tested on the testing subset.

Recall that the model accuracy is calculated by comparing actual values with predicted values of test data set, in this case the model accuracy is 0.60. The performance will be evaluated through f1_score and jaccard in the next section.

5. Models Evaluation

Car accident data for the city of Seattle between 2004–2019 have been used to train and evaluate machine learning models for predicting accident severity based on the circumstances of the accident. Four classes of models have been trained : (i) k-Nearest Neighbours, (ii) Decision Tree, (iii) Support Vector Machine ant (iv) • Logistic Regression. In this section evaluate the performance of each of these models, through their Jaccard indices, their F1-score index and the Logloss index for the case of logistic regression, all these indices measure the precision with which the models are predicting the severity of collisions. The Decision Tree model perform best, with an average F1-score of 0.64 and Jaccard of 0.63. KNN and SVM have good performance too, very close to Decision tree, both with Jaccard of 0.63 and f1-score of 0.62 and 0.63 respectively.The Logistic Regression model performs poorest, with an average F1 score of 0.60, f1-score of 0.59 and LogLoss 0.67.

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.63	0.62	NA
Decision Tree	0.64	0.63	NA
SVM	0.63	0.63	NA
LogisticRegression	0.60	0.59	0.67

Figura 14: Jaccard y F1-score ML-models

6. Conclusions

The information provided by Seattle Police Department is a first step to prove that a model can be generated to predict future accidents on the road and identify the type of information (independent variables) that can be used. i can concluid this because i have decent accuracy value for all classification algorithms. With the models evaluation, the best classifier of this problem are Decision Tree, SVM and KNN based on their accuracy value.

By revealing hidden patterns in predicting severity in accidents based on the features Weather, Road and Light conditions, addresstype, junctiontype and vehcount, have significant impact on decision whether to travel or not which often result in injury and property damage. And even bigger, it has an impact on those responsible for transport, who with this information can take action on these

variables seeking to reduce the number of people with health injuries due to accidents and the GDP spending of the countries in traffic accidents.

7. Future Directions

This work highlights that machine learning techniques can be used to probe historical data in order to make reliable predictions about the outcome of road traffic accidents, given information which is available at the time when an accident is reported.

Regarding the specific case of Seattle, the following model has a decent level of precision, but with a lot of opportunity for improvement, remember that the best models have Jaccard and F1-score indexes close to 1 and in our case they are approximately 0.64. Therefore, another project can start to collect more information from other sources or directly from the roads.