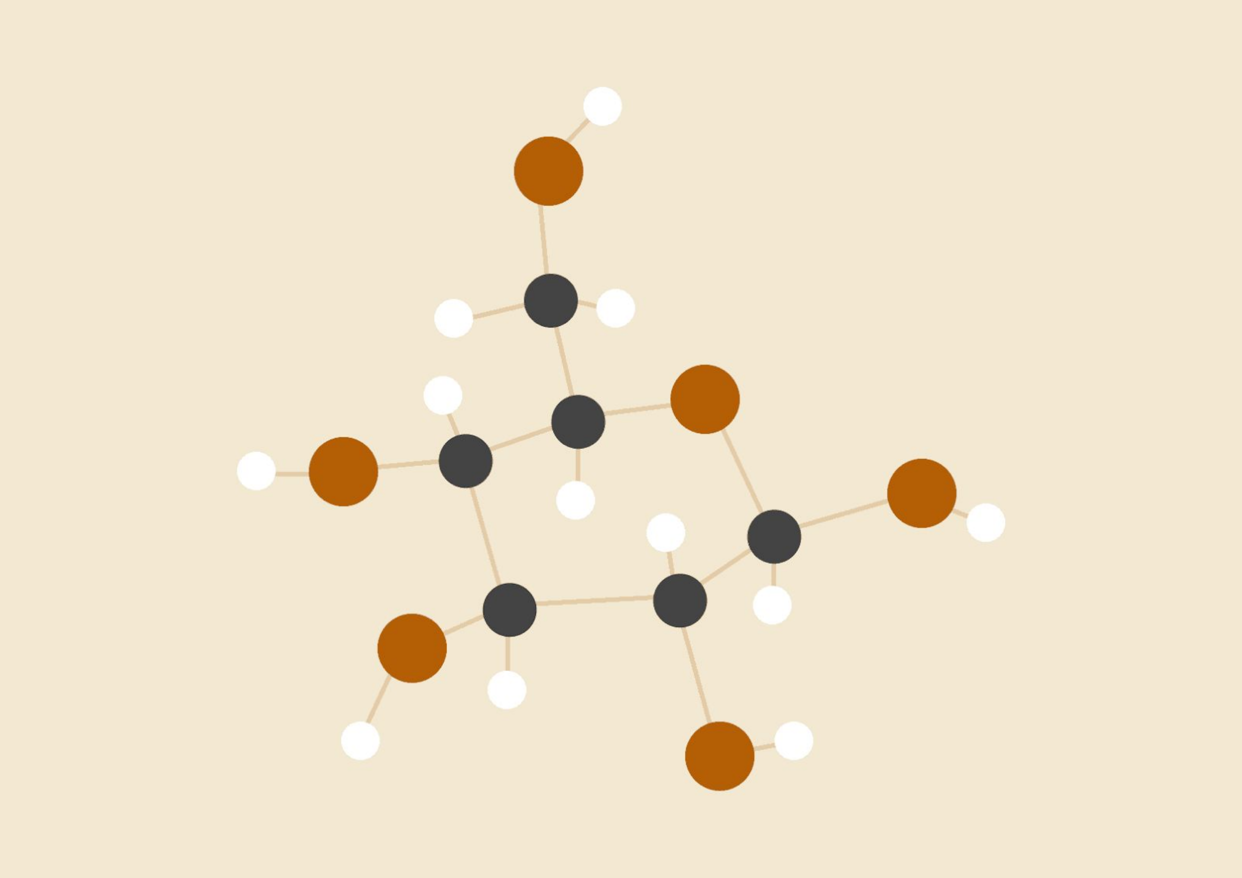
Data Science Final

IBM Course



Leoanrdo Inza

Introduction/Business Problem

The problem is that sometimes we are making a trip and in the middle of the road there was a collision and you have to wait two hours until all the accident area had taken care. So is there a way to predict this accident?

The main idea is to predict an accident by some data that include weather, road condition, location, severity of the accident, type of coallision, etc.

Hypothesis

We are going to make a deep learning model to see if we can predict traffic accidents, the idea is to see if the model can be at least 70 percent accuracy with data that doesn't know.

I will split data in training and test datasets and test data is the information that is new for the deep learning model.

Data

The Data atributes will be:

- SHAPE: x and y location

- ADDRTYPE: type of address can be Alley, Block or Intersection

- SEVERITYCODE: tells the severity of the collision can be: 3—fatality, 2b—serious injury, 2—injury, 1—prop damage, 0—unknown.

- COLLISIONTYPE: type of collision

- PERSONCOUNT: number of people that were involved in the accident.

- PEDCOUNT: number of pedestrians that were involved in the accident.

- PEDCYLCOUNT: number of bicycles that were involved in the accident.

- VEHCOUNT:number of vehicles that were involved in the accident.

- JUNCTIONTYPE: type of junction at the place of the accident.

- UNDERINFL: driver that was found involved was under the influence of drugs or alcohol?

- WEATHER: type of wather during accident.

- ROADCOND: condition of the road during accident.

- LIGHTCOND: condition of the light during accident.

- HITPARKEDCAR: in the accident a car that was parked was hit? (Y/N)

The idea of the data is to predict the severity of an accident, all variables except SEVERITYCODE are the independent variables and SEVERITYCODE y the dependent.

First stage is the cleaning data, then design a machine learning model, train and validate it, and after that will display a graph to explain the conclusion.

But the Data is not ready to use so we have to prepare it and the steps are:

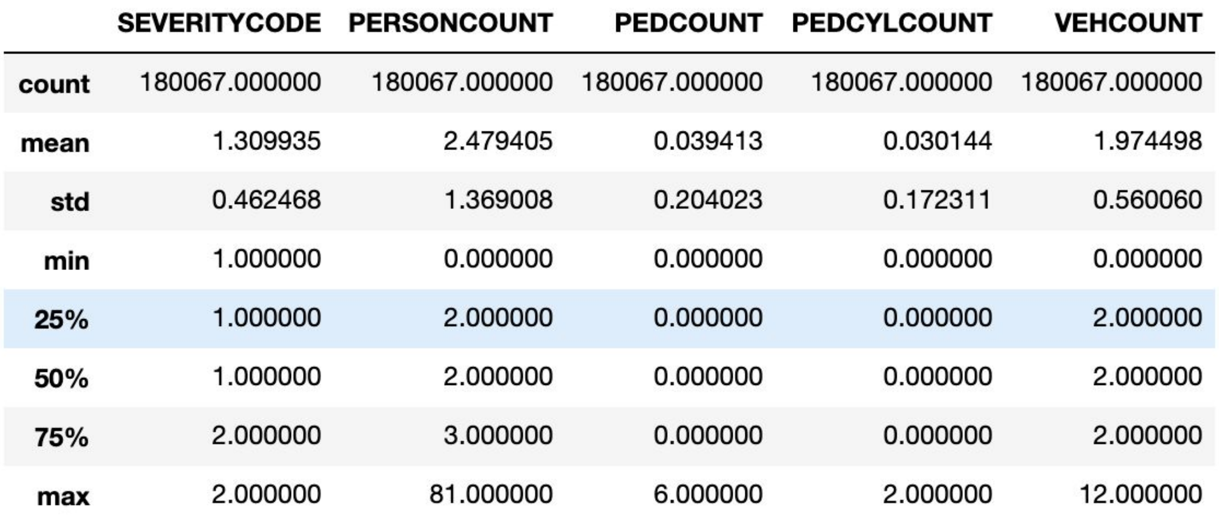
- Delete columns that will not be use
- Identify and delete rows nulls in X, COLLISIONTYPE, JUNCTIONTYPE, UNDERINFL, WEATHER, ROADCOND, LIGHTCOND
- Convert String to categorical type
- Split data to training and test
- Scale data

Methodology

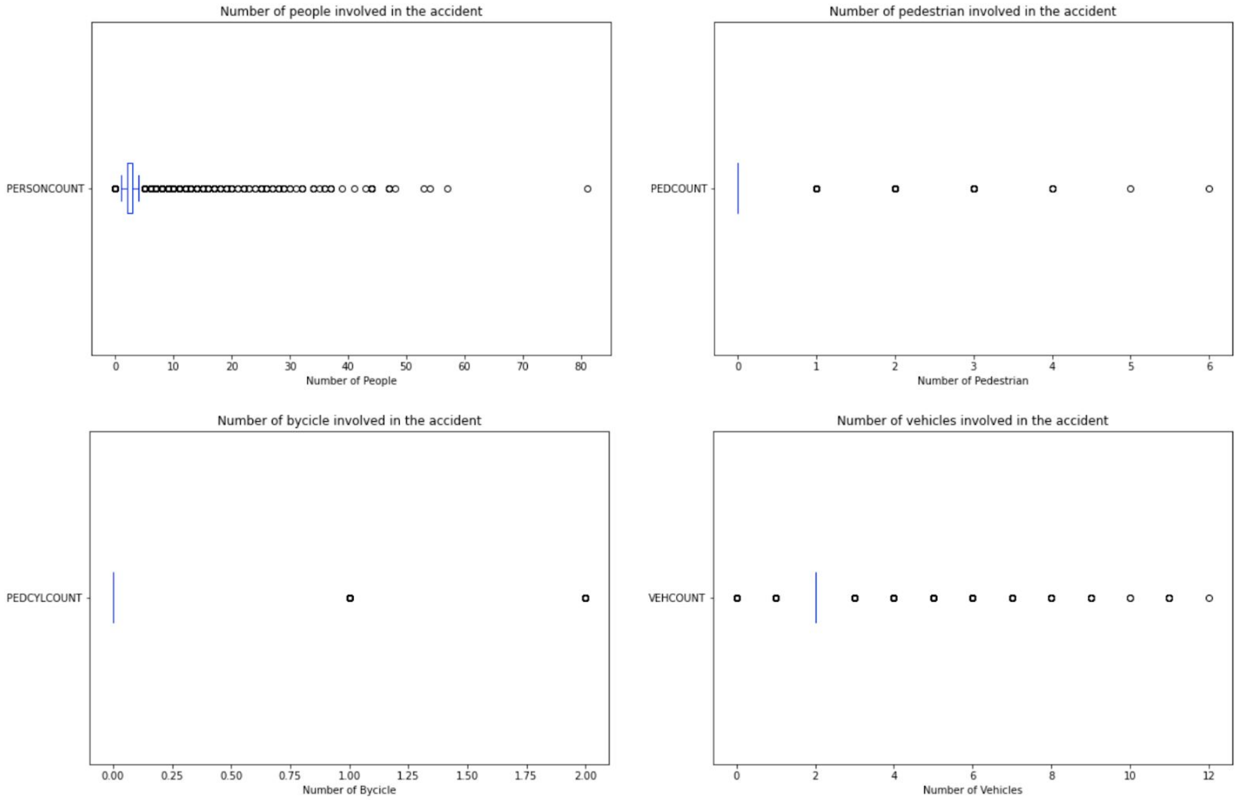
To make this proyect we are going to follow this steps:

- Data preparation: prepare data for our analysis
- Analysis: design machine learning and deep learning model and train it until the accuracy is more than 70%
- Results: explain the results of our analysis
- Conclusion: brief conclusion of our work

Before preparing our data we analize some of the variables to understand how they are, the results were the following:



	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT
count	180067.000000	180067.000000	180067.000000	180067.000000	180067.000000
mean	1.309935	2.479405	0.039413	0.030144	1.974498
std	0.462468	1.369008	0.204023	0.172311	0.560060
min	1.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	2.000000	0.000000	0.000000	2.000000
50%	1.000000	2.000000	0.000000	0.000000	2.000000
75%	2.000000	3.000000	0.000000	0.000000	2.000000
max	2.000000	81.000000	6.000000	2.000000	12.000000



ANN

After data was ready to be used we make the ANN and it contain:

- Input layer was 14 nodes
- Output layer was 1 node and predict values from 0 to 1
- 1 hidden layer

After design the artificial neural network was trained and finally we evaluate the model with a confusion matrix.

KNN

- We also did knn model and train with k from 1 to 10
- With each k numbers we calculated the accuracy with a confussion matrix
- Then we make a plot of the accuracy and saw that the best k was 10
- The data used was exactly the same used for the ANN model, so that we could comparate with the results.

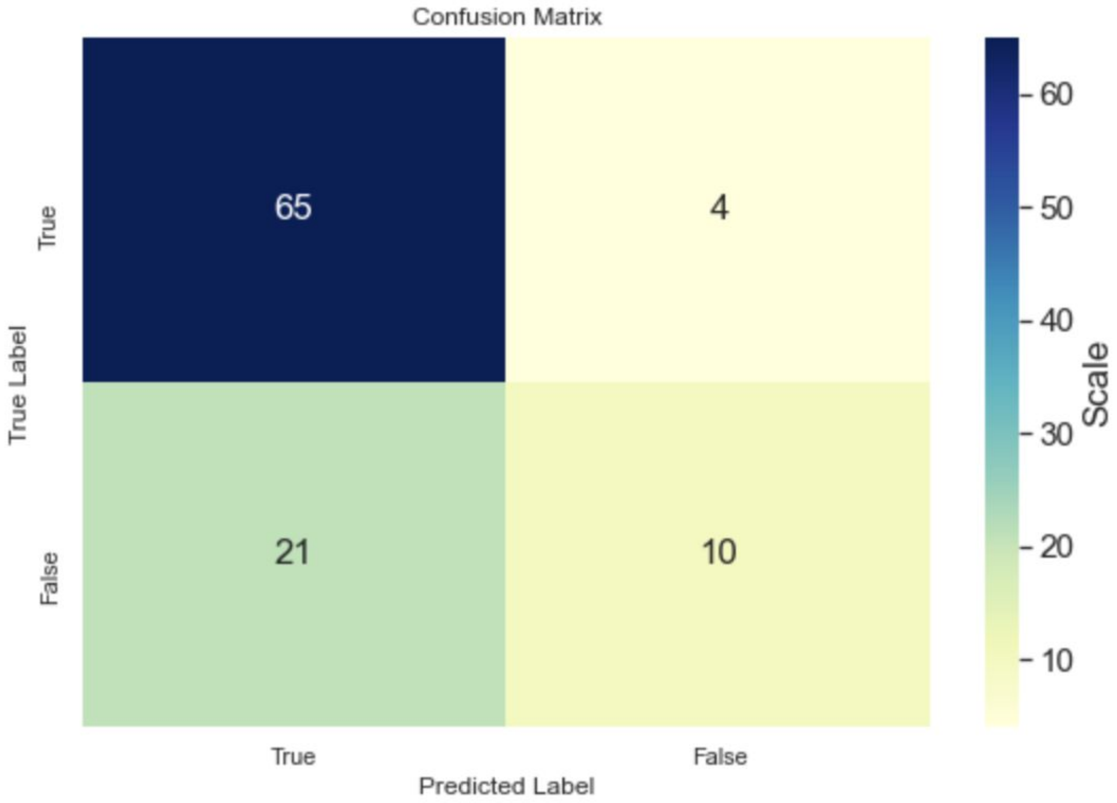
Decision Tree

- We also did decision tree model and train with number of trees from 1 to 10
- With each model with each numbers we calculat the accuracy with a confussion matrix
- Then we make a plot of the accuracies
- The data used was exactly the same used for the ANN and KNN model, so that we could comparate with the results.

Results and Discussion

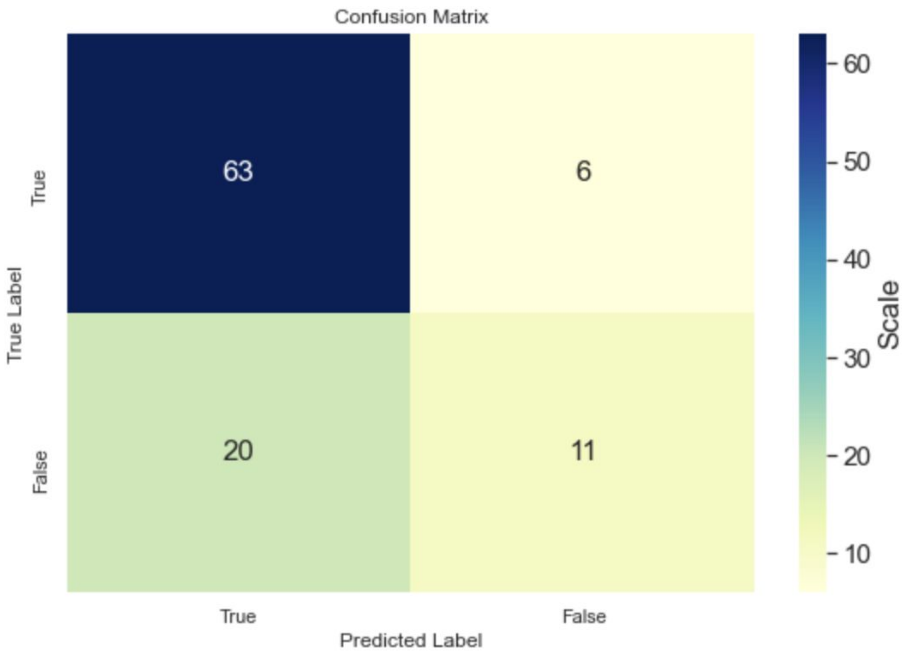
ANN

- We made a plot of the confussion matrix to explain the results.
- In the plot we can see the percentage of all posible options in the confussion matrix
- if we sum 65% with 10% we get 75% that the model predict one value and the reality was correct
- So we can say that our model have an accuracy of it
- If we sum 21% with 4% is 25% that the model predict one value and the reality was the other values.

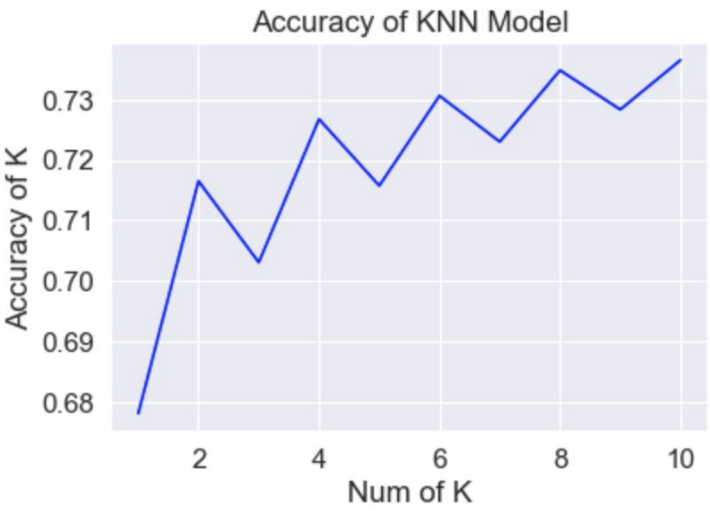


KNN

- We made a plot of the confussion matrix to explain the results.
- In the plot we can see the percentage of all posible options in the confussion matrix
- if we sum 63% with 11% we get 74% that the model predict one value and the reality was correct
- So we can say that our model have an accuracy of it
- If we sum 20% with 6% is 25% that the model predict one value and the reality was the other values.



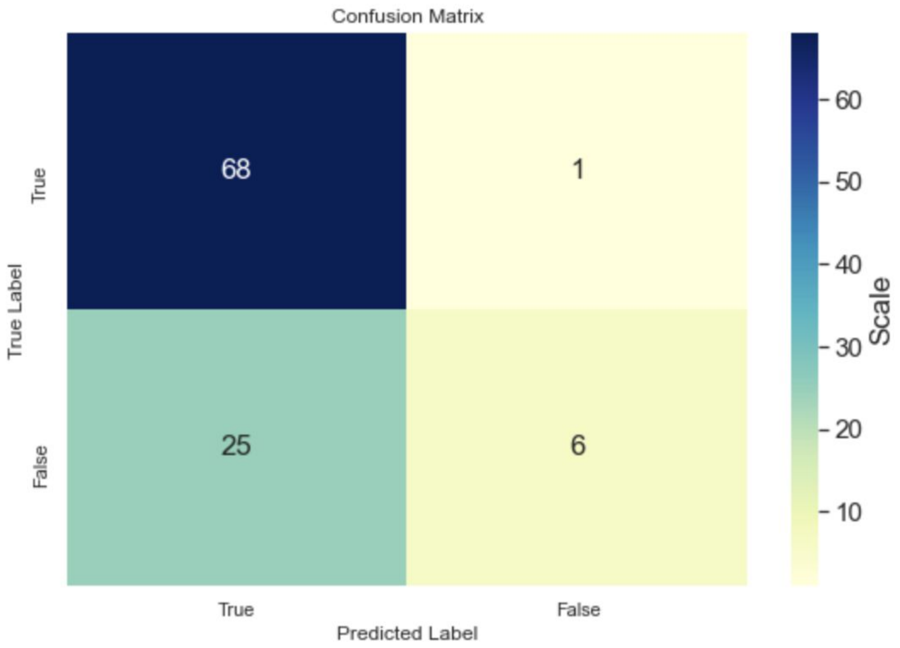
- Also we plot the accuracy values to see the best k number and value and the results was the following:



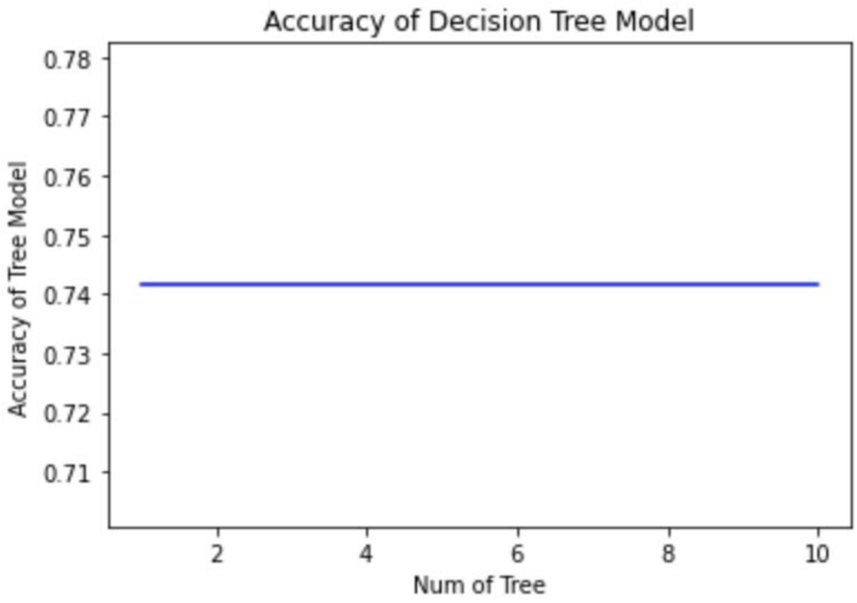
- So the best k is 10 with the accuracy value of 0.737

Desicion Tree

- We made a plot of the confussion matrix to explain the results.
- In the plot we can see the percentage of all posible options in the confussion matrix
- if we sum 65% with 10% we get 75% that the model predict one value and the reality was correct
- So we can say that our model have an accuracy of it
- If we sum 21% with 4% is 25% that the model predict one value and the reality was the other values.



- Also we plot the accuracy values to see the best number tree and value and the results was the following:



- So the best number tree is 1 to 10 with the accuracy value of 0.7416

Conclusion

In this proyecto we started with a data that wasn´t prepared for any machine learning model, so we had to prepare them, then we design a artificial neural network model, KNN and SVM model to predict future accidents, the results of all models were the following:

Model	Accuracy
ANN	0.751
KNN (k = 10)	0.737
Desicion Tree (num tree = 1 to 10)	0.7416

All models are very good and have a very good accuracy but the best one is the artificial neural network.

Our objective was to achieve a model that had 70% of accuracy and we get 75% so the results are greats, we end up deploying a model that can give us a trend of a possible accident, imagine of going on a trip and have a deep learning model that can help us avoid possible incidents, now we have it!