3. Problem solving:

We can use different supervised machine learning algorithms to predict car accidents severity such as Logistic Regression and Decision tree.

Machine Learning Algorithms:

Before we start using ML algorithms on our data, we need to do some normalizing on it

Normalizing Data:

It is the first step we take before we apply ML models to our data. The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. There are a few columns that we will standardize, so it would not affect negatively our machine learning algorithms. The age of the driver is from 18 to 88 in the dataset and we can normalize it. Also, the age of the vehicle is also from 0 to 100 and it can skew the performance of your machine learning algorithm and we will normalize this predictor too.

1- Logistic Regression:

We will take (Did_Police_Officer_Attend_Scene_of_Accident' , 'Age_of_Driver' ,'Vehicle_Type', 'Age_of_Vehicle','Engine_Capacity_(CC)','Day_of_Week' , 'Weather_Conditions' , 'Road_Surface_Conditions', , 'Light_Conditions', 'Sex_of_Driver' ,'Speed_limit') columns to use it as an input for our logistic regression model. In addition, I used (Accident_Severity) column as the output or prediction for the model.

We will train the model using 80% of the whole data that we have and the other reminded 20% we will use it for testing.

	precision	recall	f1-score	support
1	0.00	0.00	0.00	1507
2	0.00	0.00	0.00	23102
3	0.88	1.00	0.94	181109
Avg/total	0.78	0.88	0.82	205718

After we trained the model we will have the ability to predict the accidents severities but first we should test the accuracy for our model and I did that using two methods (jaccard index – Log Loss) and we got

Methods	Accuracy by percentage
jaccard index	88.0%
Log Loss	38.0%

So we can see that we could use logistic regression to predict car accidents severity but we will test another model to see if it perform better.

2- Decision Tree:

We will use the same input and output for our model that we used in logistic regression model.

Accuracy by percentage
88.0%

Conclusion:

As we saw, we could use both algorithms to predict car accidents severity in the UK based on the data that we have from 2005 – 2015.

Some Recommendations:

After we reviewed the data and analyzed it, there are a lot of factors that could help the authorities to reduce the number of car accidents such as the age of the driver, the day of the week, hours of the day,

road condition and specific areas or locations that have the most accidents rate.