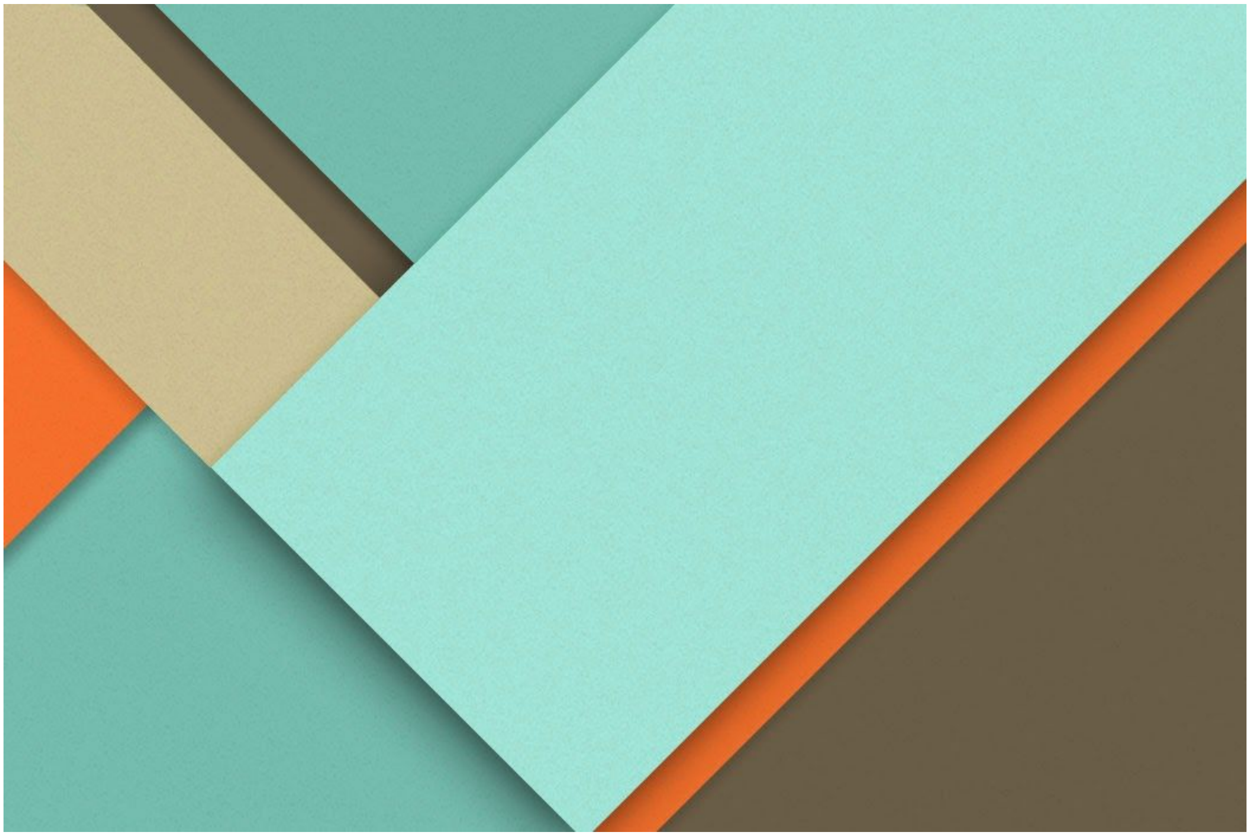


Project Predicting Car Accident Severity in Seattle

29-08-2020

Aditya Raj Singh Rathiore

Overview

Background For any traffic participant, pedestrian, cyclist, or motorist, an accident is an unexpected and undesired thing to experience. Property damage, personal injury or death as consequences of collisions not only impact the lives of individuals but also society and economy at large. When cars were first introduced in the US at the beginning of the 20th century, they were few numbers but the fatalities they caused were many. car manufacturers have introduced seatbelts and airbags; civil engineers have paved roads and added reflecting guard-rails to highways; moral campaigners have lobbied against drunk driving; and last but not least government has promulgated and enforced many traffic (safety) laws. For comparison, in 2015 almost 5,000 pedestrians died in traffic accidents, whereas in 1937 15,000 pedestrians were killed, when the US had far fewer cars and two fifths of its current population. Although motor vehicle accident deaths and pedestrian deaths in the US are at an all-time low, each death or injury that still occurs is a tragedy and thus remain a public health concern.

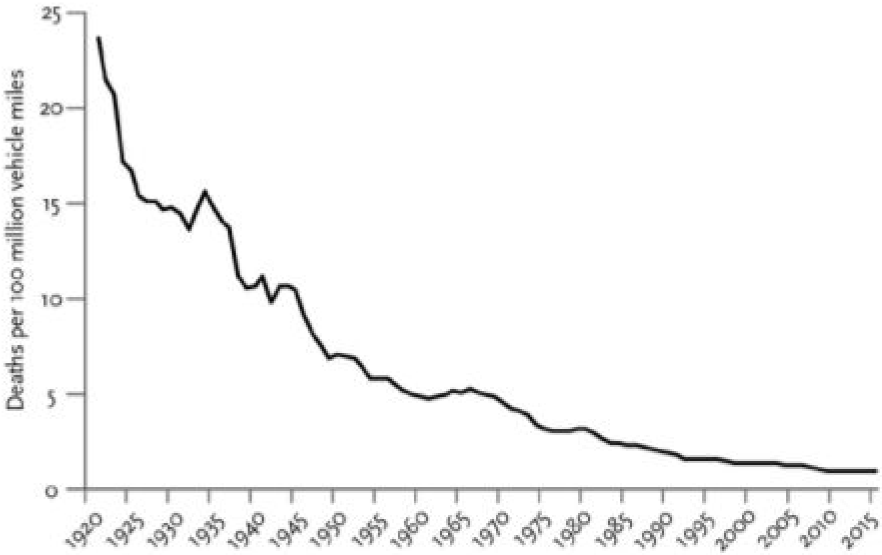


Figure 1: Motor Vehicle Accident Deaths, US, 1921-2015

Source: Pinker, Steven. *Enlightenment now*. Penguin, 2018. http://www.informedforlife.org/demos/FCKeditor/UserFiles/File/TRAFFICFATALITIES(1899-2005).pdf. http://www-fars.nhtsa.dot.gov/Main/index.aspx. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812384.

[1] Both figures are taken from Pinker, Steven. *Enlightenment now*. Penguin, 2018, p. 222-225.

[2] The WHO defines it as public health issue that causes approximately 1.35 million deaths around the world each year and leave between 20 and 50 million people with non-fatal injuries. https://www.who.int/health-topics/road-safety.

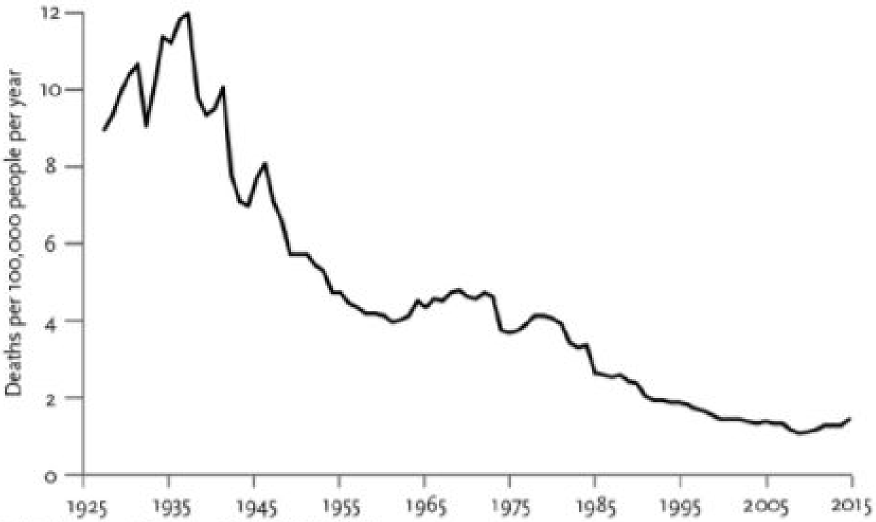


Figure 2: Pedestrian Deaths, US, 1927-2015

Sources: Pinker, Steven. *Enlightenment now*. Penguin, 2018.
For 1927–1984: Federal Highway Administration 2003.
For 1985–1995: National Center for Statistics and Analysis 1995.
For 1995–2005: National Center for Statistics and Analysis 2006.
For 2005–2014: National Center for Statistics and Analysis 2016.
For 2015: National Center for Statistics and Analysis 2017.

Goal

1. Before AI-driven cars will become ubiquitous and hopefully reduce traffic accidents to near zero, this report argues that Machine Learning Algorithms are the next important step to reduce traffic accidents. Machine Learning Algorithms can analyze historical data on traffic collisions and determine what features can best predict the occurrence and severity of accidents. The immediate application is that traffic authorities can use the algorithm's output in combination with electronic traffic signs to warn motorists (and other traffic participants) about potentially dangerous traffic conditions. The future application is that such an algorithm can update AI-driven cars on traffic condition in the area (in addition to the data that the car collects and analyzes itself) to improve its driving and the safety for all traffic participants. As a case study, this report will use the Collision Data collected by SDOT Traffic Management Division, Traffic Records Group (from 2004 to present) for the city of Seattle to build a classification model (a supervised machine learning algorithm) for predicting the severity of collisions.3 Following best practice and to ensure reproducible results, this report uses the Cross-Industry Standard for Data Mining methodology (CRISP-DM, see Figure 3). 4 In the second chapter, this report will investigate the Collision Dataset to determine the potential features for a machine learning model by performing an exploratory analysis on the attributes. The third chapter

uses the insights from the second to pre-process all relevant feature variables and the target variable putting them into final dataset ready for machine 3 See https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0 . 4 See https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining . learning. Chapters four and five will serve to build various type of classification machine learning models and to evaluate their performance in predicting collision severity. Given the limitations of this report, the final chapter will discussion deployment options for potential stakeholders (i.e. public traffic authorities, emergency services, or car manufacturers).

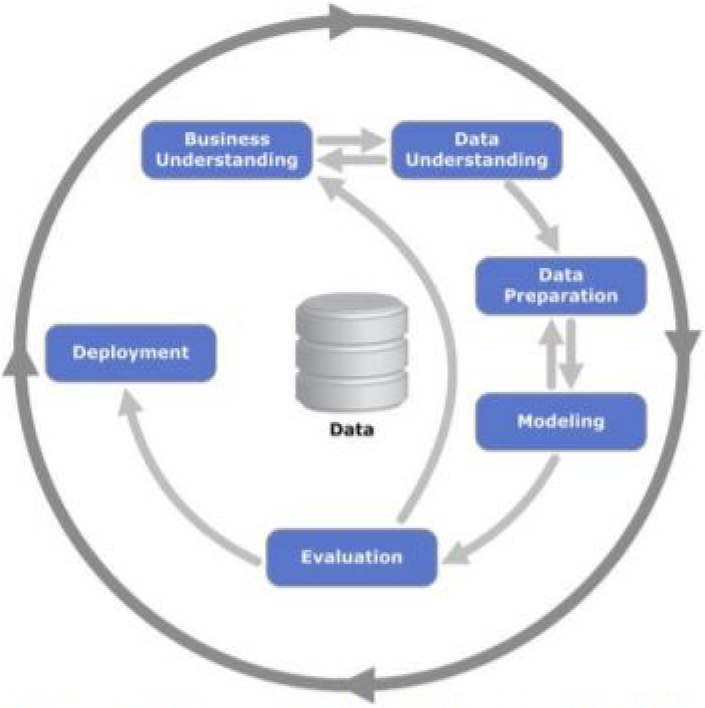


Figure 3: Cross-industry standard process for data mining, known as CRISP-DM

Sources: https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

Data acquisition and cleaning

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

This is the only place of data acquisition provided by IBM.

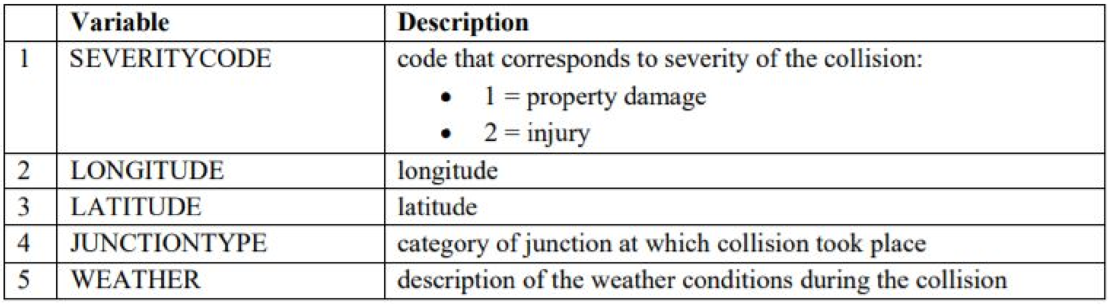
For cleaning purpose we will replace al the nan values with values depending on the attribute.

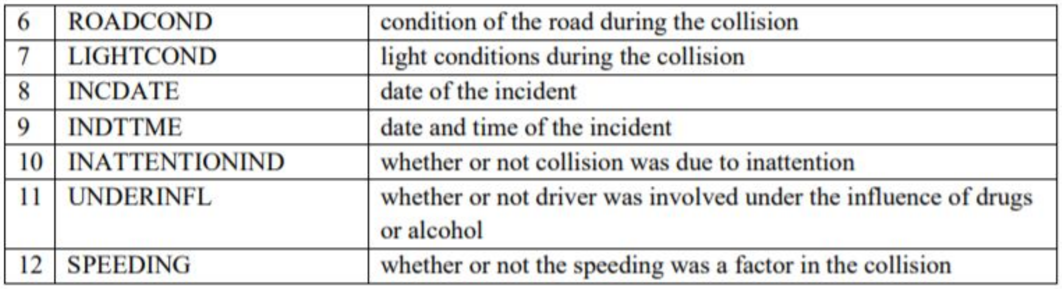
Data Preparation

We will create whole dataset of int values.

Use one hot encoding and create dummies.

Table 1: Pre-Selected Features





	Variable	**Description**
1	SEVERITYCODE	code that corresponds to severity of the collision: • 1 = property damage • 2 = injury
2	LONGITUDE	longitude
3	LATITUDE	latitude
4	JUNCTIONTYPE	category of junction at which collision took place
5	WEATHER	description of the weather conditions during the collision
6	ROADCOND	condition of the road during the collision
7	LIGHTCOND	light conditions during the collision
8	INCDATE	date of the incident
9	INDTTME	date and time of the incident
10	INATTENTIONIND	whether or not collision was due to inattention
11	UNDERINFL	whether or not driver was involved under the influence of drugs or alcohol
12	SPEEDING	whether or not the speeding was a factor in the collision

These are the columns we are gonna use out of 38 attributes as they are of significance.

And apply necessary replacing nan by mode,median values

Thereafter converting the whole dataset to int number like 0,1,2,3 etc.

We are also gonna drop location and date time variables as they are of lesser significance or no significance on our final model.

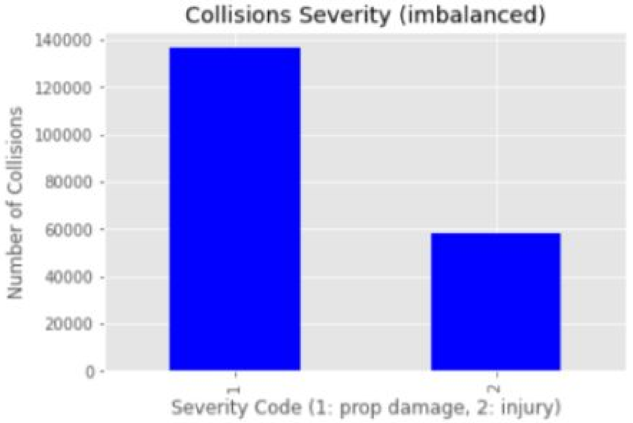


Figure 3: Collision Severity (imbalanced)

[1] The dataset was obtained here: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv.

An updated version of the dataset is available on the Seattle City website: https://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0?geometry=-123.374%2C47.452%2C-121.288%2C47.776.

[2] The metadata was obtained here: https://www.seattle.gov/Documents/Departments/SDOT/GIS/Collisions_OD.pdf.

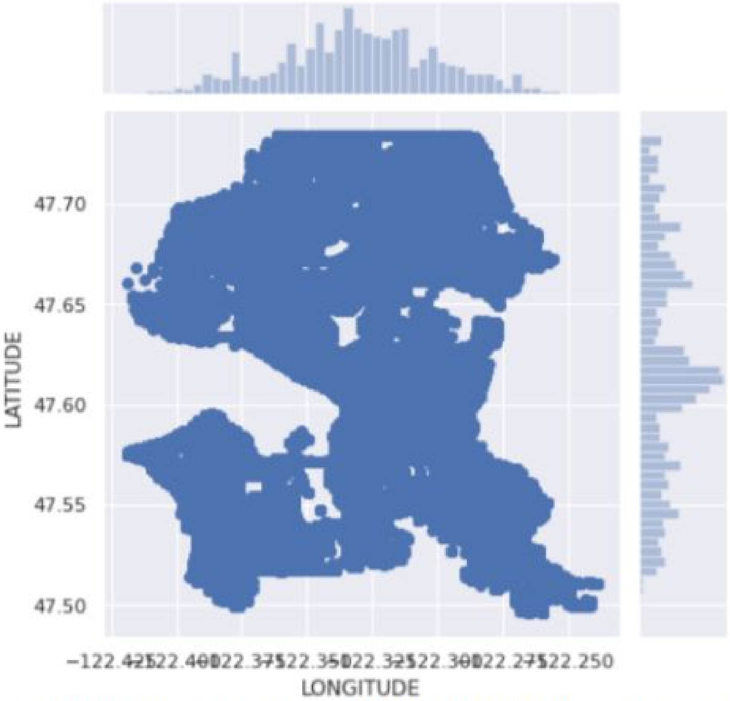


Figure 4: Location and Distribution of Collisions (by LONGITUDE and LATITUDE)

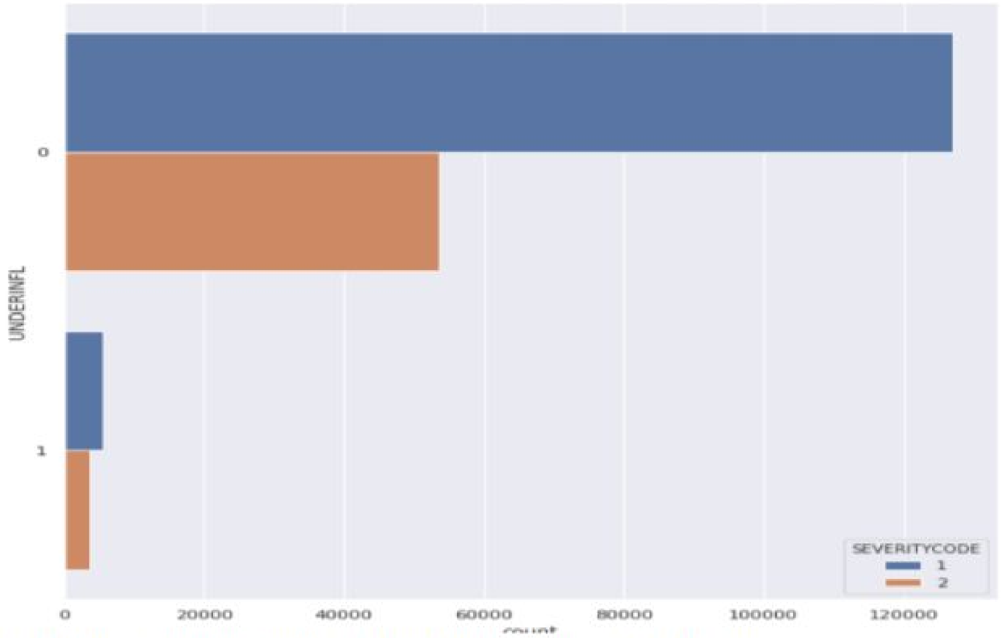


Figure 5: Influence of Drugs or Alcohol on the Frequency of Collisions

Modelling

Using differentclassification models .

KNN, LOGISTIC REGRESSION,DECISON TREE

*Classification models are used as problem are based on classification.

Training the model

```
x = test_condition
y = df_data_1['SEVERITYCODE'].values.astype(str)
x = preprocessing.StandardScaler().fit(x).transform(x)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=1234)

# obtaining data dimensions
print("Training set: ", x_train.shape, y_train.shape)
print("Testing set: ", x_test.shape, y_test.shape)
```

```
Training set: (155738, 6) (155738,)
```

```
Testing set: (38935, 6) (38935,)
```

Selecting the methods: Tree model, Logistic Regression and KNN methodology

```
#Tree model
Tree_model = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
Tree_model.fit(x_train, y_train)
predicted = Tree_model.predict(x_test)
Tree_f1 = f1_score(y_test, predicted, average='weighted')
Tree_acc = accuracy_score(y_test, predicted)

#Logistic Regression
LR_model = LogisticRegression(C=0.01, solver='liblinear').fit(x_train, y_train)
predicted = LR_model.predict(x_test)
LR_f1 = f1_score(y_test, predicted, average='weighted')
LR_acc = accuracy_score(y_test, predicted)


#KNN methodology
KNN_model = KNeighborsClassifier(n_neighbors = 4).fit(x_train, y_train)
predicted = KNN_model.predict(x_test)
KNN_f1 = f1_score(y_test, predicted, average='weighted')
KNN_acc = accuracy_score(y_test, predicted)
```

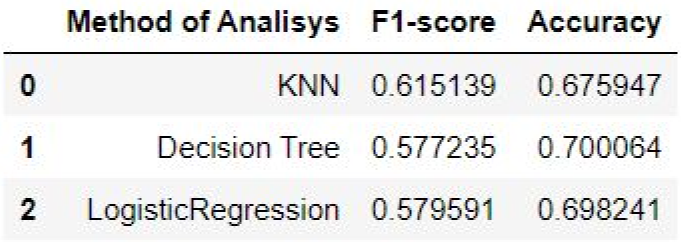
Here our model training part is completed and we have used classification model instead of other models as the problem is clearly of classification.

5.Evaluation

```
results = {
   "Method of Analisys": ["KNN", "Decision Tree", "LogisticRegression"],
   "F1-score": [KNN_f1, Tree_f1, LR_f1],
   "Accuracy": [KNN_acc, Tree_acc, LR_acc]
}
```

```
results = pd.DataFrame(results)
results
```

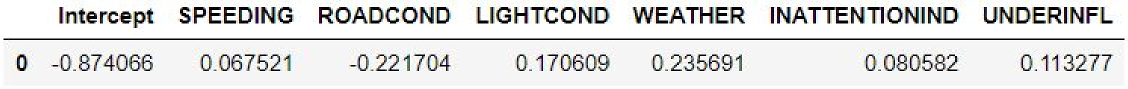
Here we are comparing F1-score, accuracy of different models
You can also use jaccard index,log value etc.



	Method of Analisys	F1-score	Accuracy
0	KNN	0.615139	0.675947
1	Decision Tree	0.577235	0.700064
2	LogisticRegression	0.579591	0.698241

```
results = {
    "Intercept": LR_model.intercept_,
    "SPEEDING ": LR_model.coef_[:,0],
    "ROADCOND ": LR_model.coef_[:,1],
    "LIGHTCOND ": LR_model.coef_[:,2],
    "WEATHER ": LR_model.coef_[:,3],
    "INATTENTIONIND": LR_model.coef_[:,4],
    "UNDERINFL": LR_model.coef_[:,5],
}

results = pd.DataFrame(results)
results
```



	Intercept	SPEEDING	ROADCOND	LIGHTCOND	WEATHER	INATTENTIONIND	UNDERINFL
0	-0.874066	0.067521	-0.221704	0.170609	0.235691	0.080582	0.113277

Looking at the results obtained in the comparison, it is understood that speed,road conditions, weather,lighting conditions,attention on driving and influence of others influence the severity of traffic accidents.