Introduction (Description of the Problem and a Discussion of the Background)

The number of motor vehicles have increased immensely over the years, as have the resulting car injuries and deaths per year rates. 1.35 million people die in road accidents worldwide every year that is 3,700 deaths a day. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities [1]. Also, fatal car wrecks and road accidents cost countries worldwide about 3% of their GDP [2].

There are several factors that influence the occurrence and severity of car accidents. These can be broadly categorized into behavioural, environmental and situational factors. Behavioural factors include careless driving, speeding and driving under the influence of drugs and alcohol. In fact, distracted driving is a leading cause of accidents, causing 25-50% of all crashes. Environmental factors are related to road, weather and visibility conditions. Situational factors include sudden malfunctioning of the vehicles.

Majority of accidents can be prevented as most result from poor and reckless driving and increase in specific environmental conditions. What if we can relate the occurrence and severity of car accidents to specific conditions such as road, weather and visibility conditions? Wouldn't we be able to alert drivers to be more careful and hence increase their attentive driving? Wouldn't such info also help healthcare institutions and governments such as the police and transportation departments by allowing them to maintain sufficient staff and alertness to adequately address any upcoming and unfortunate accidents?

Due to the availability of large amount of accident related data generated by different government bodies, it would be possible to build machine learning models to predict the occurrence and severity of car accidents given specific conditions. This would eventually help in both reducing the occurrence and severity of such incidents.

Many institutions including local governments, police departments, healthcare institutions, rescue groups, insurance companies may benefit from such predictive models.

References

[1] https://www.asirt.org/safe-travel/road-safety-facts/
[2] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Description of the Data

For this assignment, collision data collected by the Seattle Police Department (SPD) and recorded by the Traffic Records team will be used. The data was downloaded from Seattle's government site (link provided below). "Seattle City GIS Program" is the data owner.
https://data.seattle.gov/Land-Base/Collisions/9kas-rb8d

The data was last updated on September 22, 2020 and covers 214,050 collision records spanning from October 2003 to September 2020.

The data has 40 attributes including the dependent variable “SEVERITYCODE” which represents the severity of the collision represented by the following codes and their respective severity:

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

Following “Data Understanding”, only a subset of the attributes that may be related to the occurrence and severity of the collisions will be selected in this study and the rest non-relevant attributes or variables will be ignored. The following attributes of interest will initially be considered to train the model. As part of the “Data Preparation”, balancing unbalance data and cleaning the dataset will be performed if required.

Attribute	**Description**
ADDRTYPE	Collision address type: Alley, Block, or Intersection
LOCATION	Description of the general location of the collision
PERSONCOUNT	The total 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 total injuries in the collision
FATALITIES	The number of fatalities in the collision
INCDTTM	The date and time of the incident
JUNCTIONTYPE	Category of junction at which collision took place
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 during the time of the collision
ROADCOND	The condition of the road during the collision
LIGHTCOND	The light conditions during the collision
SPEEDING	Whether or not speeding was a factor in the collision (Y/N)
HITPARKEDCAR	Whether or not the collision involved hitting a parked car (Y/N)

The following machine learning classification algorithms will be employed to build the prediction model:

- K Nearest Neighbours (KNN)
- Decision Tree
- Logistic Regression
- Support Vector Machine (SVM)