**Car Accident Severity Capstone Project:**

**Business Understanding**

The goal of this project is to predict severity of an accident and allow the Seattle government to possibly prevent/reduce car accidents that depend on factors like weather and road condition(wet/dry), traffic situation, light condition etc. About 100,000 crashes happen in Seattle every year since 2010 as per WSDOT ten-year summary report([WSDOT 10 Year Summary Report](https://remoteapps.wsdot.wa.gov/HighwaySafety/Collision/Data/Portal/Public/PublicReport/TenYearSummary)).

The model developed can be used to warn drivers, health authorities, government and police about the possibility of a car accident and its severity so that everyone can be more careful and/or change travel plans (if able to) in such critical situations.

The solution/model developed via this project can be utilized by the end user(car drivers),local government, car insurance companies and hospitals who are considered the target audience for this business problem to use the model and make required decisions to help prevent accidents in the city.

**Data Understanding**

The data used in this project is about “accident severity” and was provided by SPD (Seattle Police Department) and recorded by Traffic Records department from 2004 to present. This includes all types of collisions. This problem is seen as supervised machine learning as it has labeled data to train and validate the model.

In total, there are 37 attributes(columns) and 1 dependent variable (labelled data) which is “SEVERITYCODE” in the data and 194,673 rows. “SEVERITYCODE” is the code that corresponds the severity of the collision (as indicated in meta data):

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

There are numerical and categorical types of data which are needed to be converted into numerical data. and some attributes have missing data therefore the data requires some **preprocessing and preparation** to construct the final dataset to be fed into the model. Not all attributes are useful/relevant for my model, so I have decided to **select some features** such as SEVERITYCODE (target-variable), SEVERITYDESC, INATTENTIONIND, UNDERINFL, WEATHER, SPEEDINGROADCOND, LIGHTCOND as effects of these variables are considered significant. It was also noticed that data set is unbalanced as distribution of target variable is in favor of property damage so some data balancing will be required. There were only instances of 2 severity types (property damage and injury collisions) in the data set.

I plan to use different machine learning models to solve the problem and evaluate the model.