Initial Thoughts on Modelling

Based on the EDA, the nature of data and the nature of variable we need to predict. We can say following:

1. Data has multiple factors on which severity can be determined on.
2. Based on bivariate analysis we could see that each factor brings in some data separation.
3. If we look at why accidents could happen then - some combination of factors in dataset could be the contributor.
4. i.e if a certain set of conditions are met we can expect a certain severity accident.

Based on above points **Decision Tree** model would be the first model that i would like to try out.

Reasons for choosing `Decision Tree` ?

1. At each level decision tree tries to make decision on which `variable` should the split happen.
2. On the basis of gini index / information gain it will split the node on that variable
3. We can see that there is already some level of separation of multiple variables when we try to look it from severity (variable to predict) perspective.
4. The predictions are explanable in nature.
5. It could be a simple baseline model to begin with.
6. It makes sense to use decision trees.

Further improvements of modelling

1. We could use ensemble models using decision trees like Random Forest or XgBoost