Predicting Road Accidents Severity

Dipika Sharma

Bellevue University

Applied Data Science 680

Amirfarrokh Iranitalab

Project White Paper: Predict Road Accidents Severity

Introduction:

According to World Health Organization 1.19 M people dies every year because of road accidents. This number not only include adult but kids and young adults of age between 5-29. Road accidents is one of the major causes of unnatural deaths. To our surprise the rate for unnatural death is increasing every year.

As part of this project, I chose the dataset from Addis Ababa Sub-city police departments. Using this analysis, we will predict the road accidents severity.

Business Problem:

As we know the death cause by road accidents are increasing every year as per reports from WHO, what can be done to reduce the fatalities. Government is making sure to introduce rules and regulation and raise awareness among public so that they followed the rules, and the fatalities can be reduced. Machine learning playing an important role in all the fields and using machine learning we can predict the severity of the accidents so necessary precaution can be taken and this analysis can help government in reducing the road accidents rates.

Data Explanation:

I used the dataset from Kaggle website, this dataset is from Addis Ababa Sub-city police departments. This dataset contained the records of road accidents happened between 2017 to 2020.

The data can find at below location:

https://www.kaggle.com/datasets/kanuriviveknag/road-accidents-severity-dataset

This dataset is good for our purpose as it consists of target feature “Accident_

severity”.

The database has 12316 rows and 32 columns.

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_driver_relation	Driving_experience	Type_of_vehicle	Owner_of_vehicle	Service_
0	17:02:00	Monday	18-30	Male	Above high school	Employee	1-2yr	Automobile	Owner	
1	17:02:00	Monday	31-50	Male	Junior high school	Employee	Above 10yr	Public (> 45 seats)	Owner	
2	17:02:00	Monday	18-30	Male	Junior high school	Employee	1-2yr	Lorry (41?100Q)	Owner	
3	1:06:00	Sunday	18-30	Male	Junior high school	Employee	5-10yr	Public (> 45 seats)	Governmental	
4	1:06:00	Sunday	18-30	Male	Junior high school	Employee	2-5yr	NaN	Owner	

The detail of the columns is below.

Column Name	Column Description
Time	What time of a day accident happened?
Day_of_week	Day of a week when accident happened
Age_band_of_driver	Age group of the driver
Sex_of_driver	Gender of the driver
Educational_level	Education level of the driver
Vehicle_driver_relation	Relation of driver with vehicle
Driving_experience	driver's driving experience
Type_of_vehicle	Type of the vehicle
Owner_of_vehicle	The owner of the vehicle
Service_year_of_vehicle	When was the vehicle last service happened?
Defect_of_vehicle	Any defect in the vehicle?
Area_accident_occured	Area where the accident happened
Lanes_or_Medians	Lanes or Median at the area where accidents happened
Road_allignment	Any road alignment
Types_of_Junction	Types of junctions
Road_surface_type	Road surface type
Road_surface_conditions	Condition of the road surface
Light_conditions	what is the light condition at the accident site?
Weather_conditions	What is the weather condition?
Type_of_collision	Type of collision
Number_of_vehicles_involved	How many vehicles involved in accidents?

Number_of_casualties	Total number of causalities
Vehicle_movement	The movement of the vehicle before the accidents.
Casualty_class	Casualty class
Sex_of_casualty	Gender of the casualty
Age_band_of_casualty	Age group of the casualty
Casualty_severity	Severity of the casualty
Work_of_casuality	Work of the casualty
Fitness_of_casuality	Fitness of the casualty
Pedestrian_movement	Any pedestrian near the accidents?
Cause_of_accident	Cause of the accidents
Accident_severity	Severity of the accidents

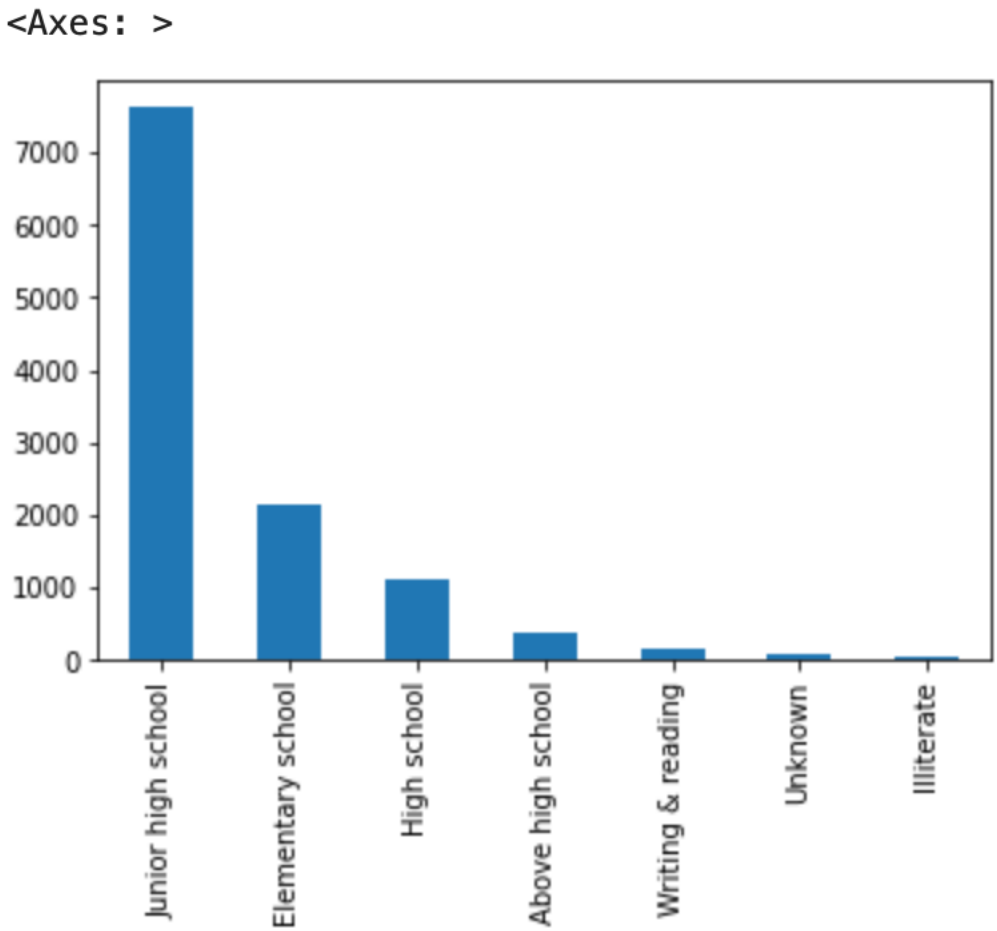
Project Methodology:

As mentioned, the objective of the project is to predict road accidents severity based on previously observed values. A classic approach is to perform the analysis on road accidents data with help of expertise, but a more modern approach includes machine learning and sentiment analysis. The model capable of this modern approach are supervised learning algorithm like decision tree and random forest algorithms. Both algorithms considered to be most widely used model when it comes to draw conclusions about a set of observation. This is appropriate models to deal with road accident data and most widely known for efficiency, reliability, and capability when we need to perform predictive modeling.

As part of this project both algorithms were performed, and Decision Tree model was selected over Random Forest algorithm because of its ease of application and interpretability.

Analysis:

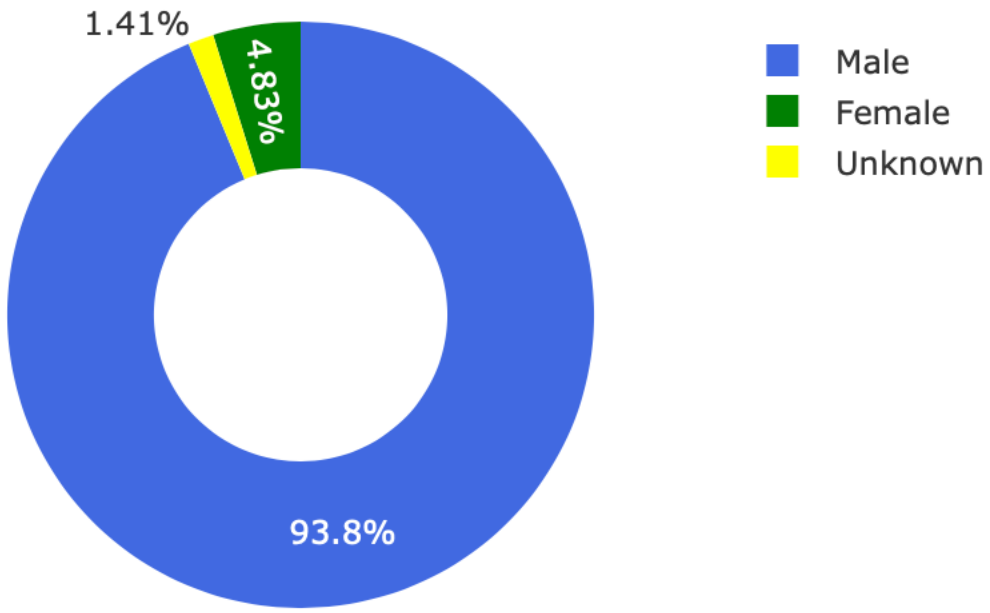
I started the project work with exploratory data analysis as this will help me understand the data. I used different visualization to draw some initial conclusion. As first visualization I check the education level of the driver.



The above result shows the driver involved with maximum accidents have passed the junior high school. And we see small fraction of drivers involved in accidents who have education above high school. This result show lower education can increase risk or number of road accidents. Using this information government can plan campaign to educate drivers about

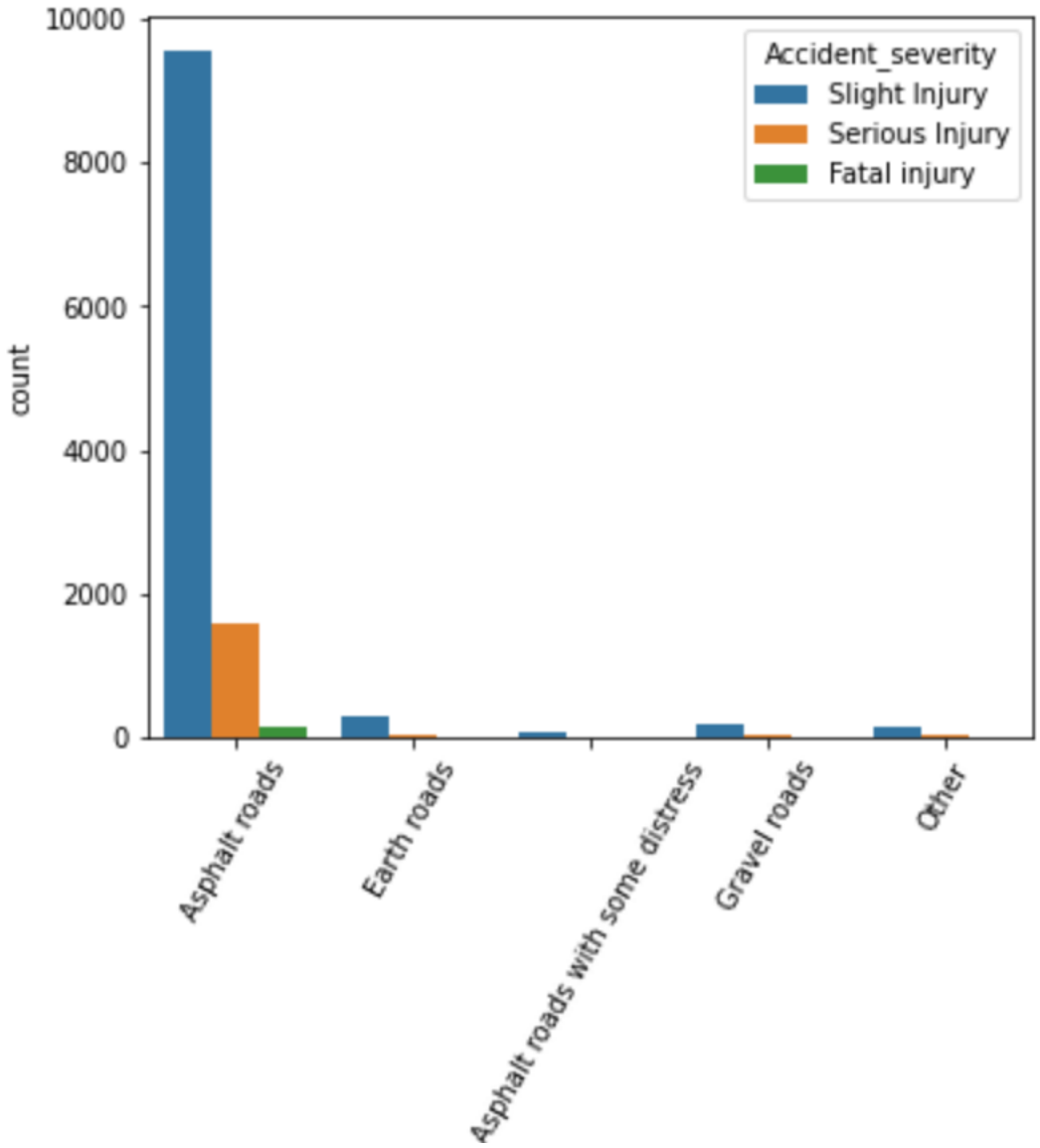
road sign and traffic laws. Also, we can introduce mandatory programs to pass the driving license to make sure all drivers understand the road rules and laws.

As next visualization I am showing the gender of drivers who were involved in road accidents.



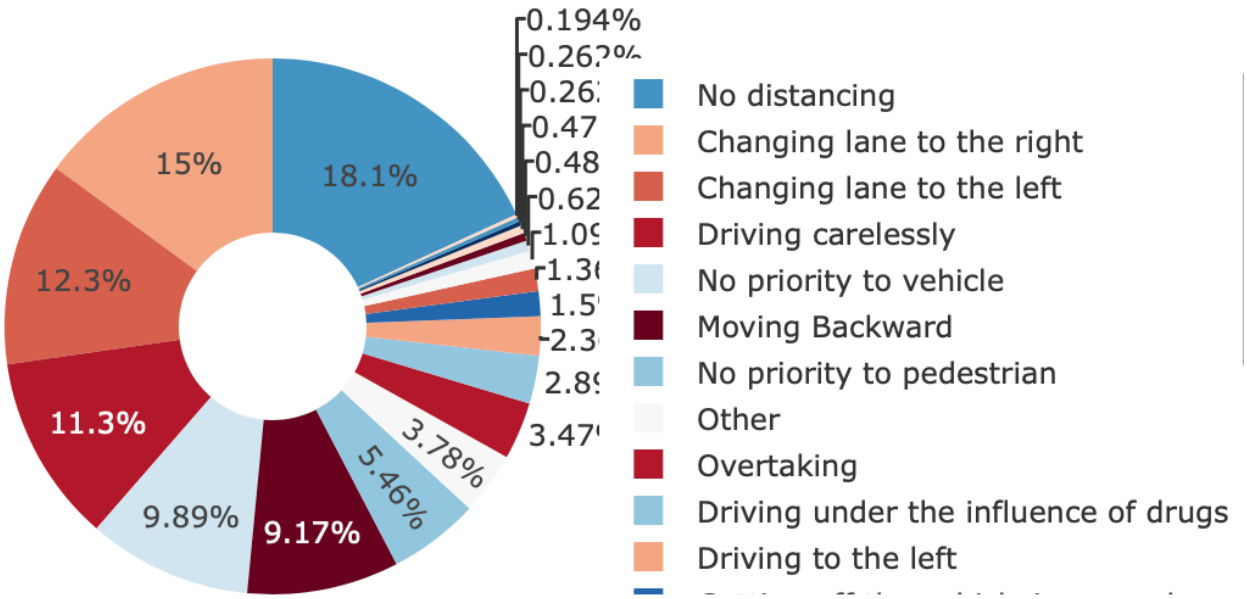
The above donut chart shows 93.8% of the drivers are male when the accidents happened.

Now let's see the road_surface_type and accident severity feature together



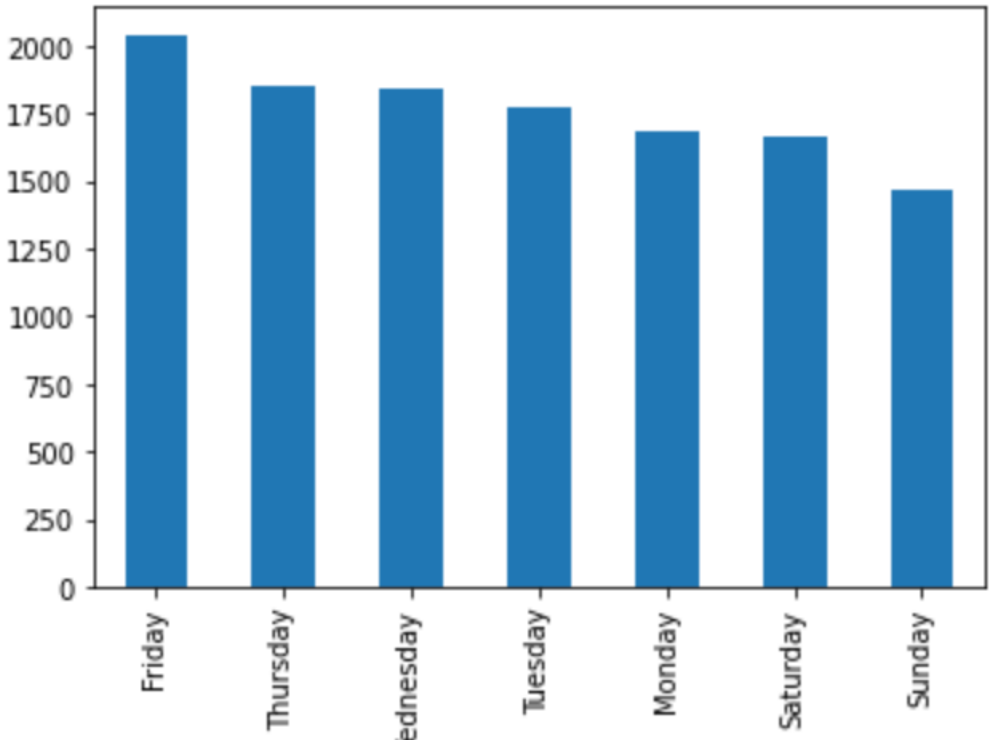
This chart show Asphalt roads have more accidents counts. May be government can perform some quality checks when they used the road surface as Asphalt.

As next I selected to show the cause of the incidents with help of donut chart.



The above donut chart shows 18% of the road accidents happened when drivers do not maintained distancing, 15% & 12.3% accidents happened when drivers trying to change lane to the right or to the left. The other reason shows when driver driving carelessly or moving backwards. May be government can impose some strict rules to reduce the number of road accidents.

With all the features we have with RTA dataset, I think it would be interesting to know which day of the week have most of the accidents.

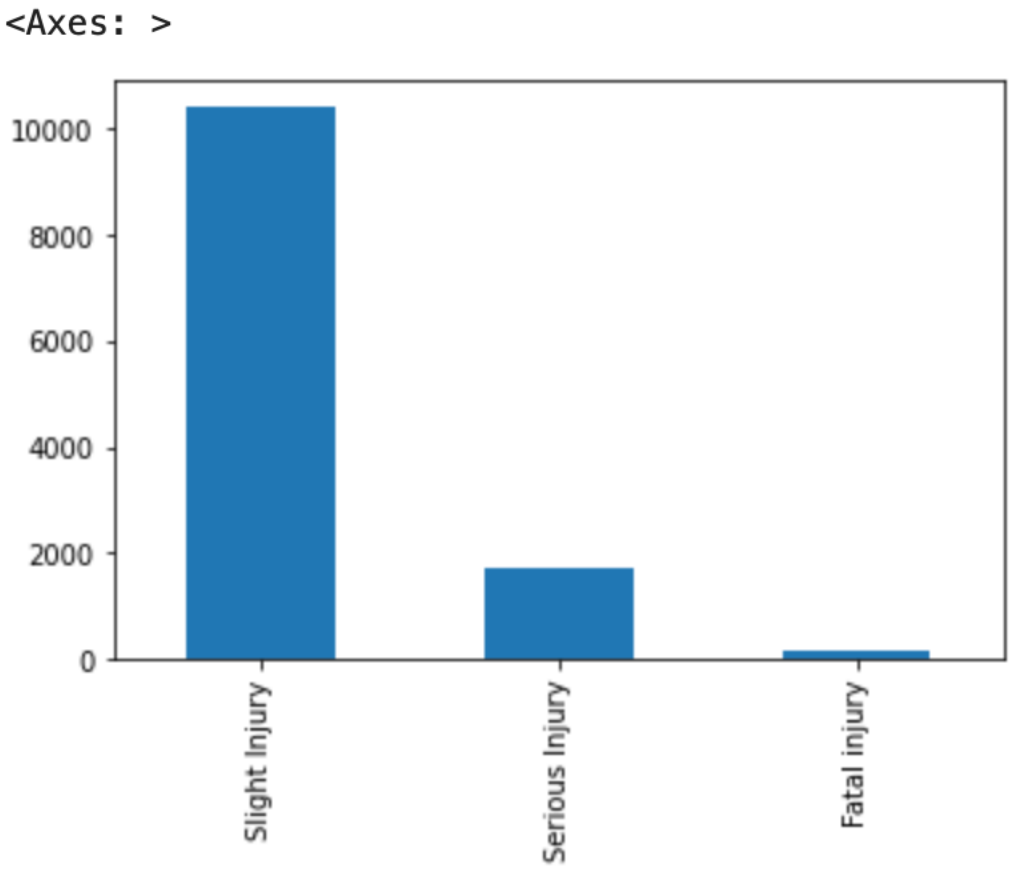


We can see above Friday has most of the accidents. Using this information, we can take some precaution on days which show more accidents. As most of the accidents happened during Friday. And less accidents happened on Sunday. Since averagely we see more accidents on weekdays compared to weakened which indicate drivers are in hurry and might results in more accidents during weekdays. Maybe we can implement some precaution plan on weekdays to reduce the road accidents.

As next step in cleaning process, I learned that some of the columns have missing values and all the columns with missing values are object type, so I planned to replace the missing values with “Unknown” or “Other”. The missing values can impact the result accuracy and reliability hence it is very important to handle them carefully in preprocessing step.

Since most of the features in our dataset is categorical columns so using the One-Hot encoding get_dummies() method to interpret categorical features more accurately for machine learning. This process will improve the prediction and avoid the biasness in system.

The project is about severity prediction, so my next visualization is on the target variable to see different severity distribution class



As we can see from above graph that most of the accidents involve slight injury compared to serious or fatal injury. The above outcome showing that the data is imbalanced.

As we can clearly see that the observation outcome of the dataset is unevenly distributed as the slight injuries is almost 5 times more than with accidents involves serious or fatal injuries. It is important to handle the imbalanced data to avoid building biased model and

inaccurate predictions. We used SMOTE to accurately unsampled the data. It is important as it avoid the biasness in model and improve the model accuracy.

See Appendix, Table 1: Target Variable class distribution after performing SMOTE

The next step is to split the data into training and testing datasets. The training dataset will be used to prepare the model and generate a prediction. The testing dataset, which is much smaller than the training dataset, will be the part of the data that we will try to predict.

Conclusion:

The two-model decision tree and random forest were trained to performed prediction, but Decision Tree show the accuracy of 98% and came out as best model in this project scenario.

The confusion matrix is created to see how the Decision Tree model performing:

```
[[8299   31    0]
 [ 169 8142    4]
 [  18  313 8020]]
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      8330
           1       0.96      0.98      0.97      8315
           2       1.00      0.96      0.98      8351

    accuracy                           0.98     24996
   macro avg       0.98      0.98      0.98     24996
weighted avg       0.98      0.98      0.98     24996
```

See Appendix, Table 2: Confusion matrix Random Forest Model

Assumptions:

All the analysis has been performed using the dataset available online assuming that the data is representative of the larger environment from Addis Ababa Sub-city police departments. This dataset contained the records of road accidents happened between 2017 to 2020. The biggest assumption is that this dataset is collected without any bias since we are not sure how this data is collected and the original purpose of collecting the dataset. Any bias in dataset can lead us to bias model and to inaccurate prediction of outcome.

Limitations:

One of the limitations is we using the historical dataset, we are not sure how old is this dataset. Also, this dataset has some predictors variable, but we are not sure if we are missing any important features that can changes the prediction of outcome then we build the model with certain limitation. In order to deal with this limitation, we need to try this model on different dataset and a proper review should happen on this model by expertise to understand if we are missing anything in this model.

Challenges and Risks:

With all the assumption and limitation of the dataset the biggest challenge is to make sure the model we build is accurate. Some of the potential risk we can see depending on how the missing data, misbalancing and scaling of the data is handled in the preprocessing steps as all these factors can lead us to inaccurate prediction and bias in model.

Future Uses/Additional Applications:

The model we build here using machine learning can help government in detecting severity in road accidents. Which in turns help them to introduce some rules and regulations which can reduce road accidents. Some preventative strategies can be implemented using the knowledge we gain from this analysis.

In order to advance our machine learning model, we can analyze more accidents data that is available to us ethically by states wise or county wise to identify the severity of the road accidents. If we can add the county or state data to our analysis we will be able to introduce rules and traffic laws as per the observation we made using machine learning model and it will add value to our model future use.

Recommendations:

Ensemble method is considered to be the essential when it comes to prediction in machine learning. We can use ensemble method to improve accuracy in predicting the severity of the road accidents. This method allows us to combine the outcome of multiple models effectively instead of relying of any single model status which results in producing the robust and reliable outcome in making prediction.

Implementation Plan:

The next step is to partner with different government sector, share the importance of the machine learning model with them, get access and collect more data so the model can be tested more on real time data.

Once the testing is done, we can implement the model live where model can provide support to government by sharing important information about predictor variable that can be used to reduce the road accidents and its severity. Once the model has all the predictor variables, it can identify severity of the road accidents and can alert the government to make preventive strategies to safe people life or reduce risk. Regular checks can be done to make sure of the consistency of the model and flow of the data.

Ethical Consequences:

Make sure all the ethical analysis guidelines will follow by me while preforming data analysis steps.

1. To ensure that legal and ethical ways are used to collect the data and make sure no personal information of driver or casualty get misused or disclosed without the consent.
2. Make sure that all the algorithm and models that are used in the project do not show any biases and discrimination towards any group.
3. There should be transparency in the data analysis methods so that everyone understands the analysis easily.
4. Will make sure the security of data so there will no unauthorized access to the data.
5. Finally, will ensure there will no negative impacts on population because of the project result.

References:

1. World health Organization: - Road Traffic Injuries https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries
2. Data Source : - https://www.kaggle.com/datasets/kanuriviveknag/road-accidents-severity-dataset/data

Appendix:

Table 1: Target Variable class distribution after SMOTE

```
2    10415
1    10415
0    10415
dtype: int64
```

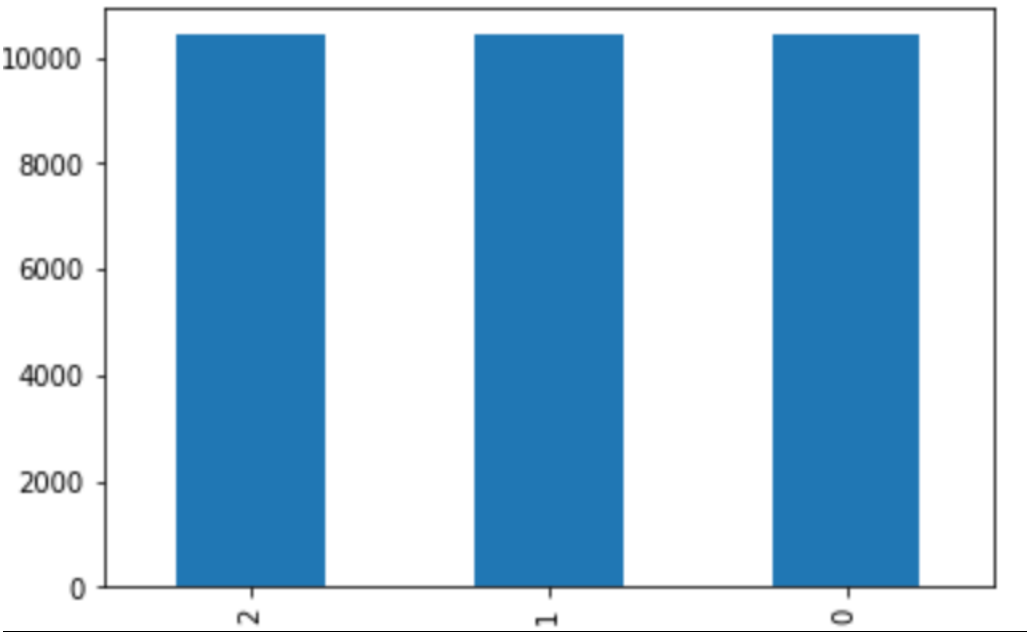


Table 2: Confusion Matrix Random Forest Model

```
[[1992   74   19]
 [  84 1747  269]
 [  21  262 1781]]
              precision    recall  f1-score   support

           0       0.95      0.96      0.95      2085
           1       0.84      0.83      0.84      2100
           2       0.86      0.86      0.86      2064

    accuracy                           0.88      6249
   macro avg       0.88      0.88      0.88      6249
weighted avg       0.88      0.88      0.88      6249
```