Enhancing Road Safety through Crash Data Analysis in San Jose: A Comparative Study of Feature Selection Methods & Classification Algorithms

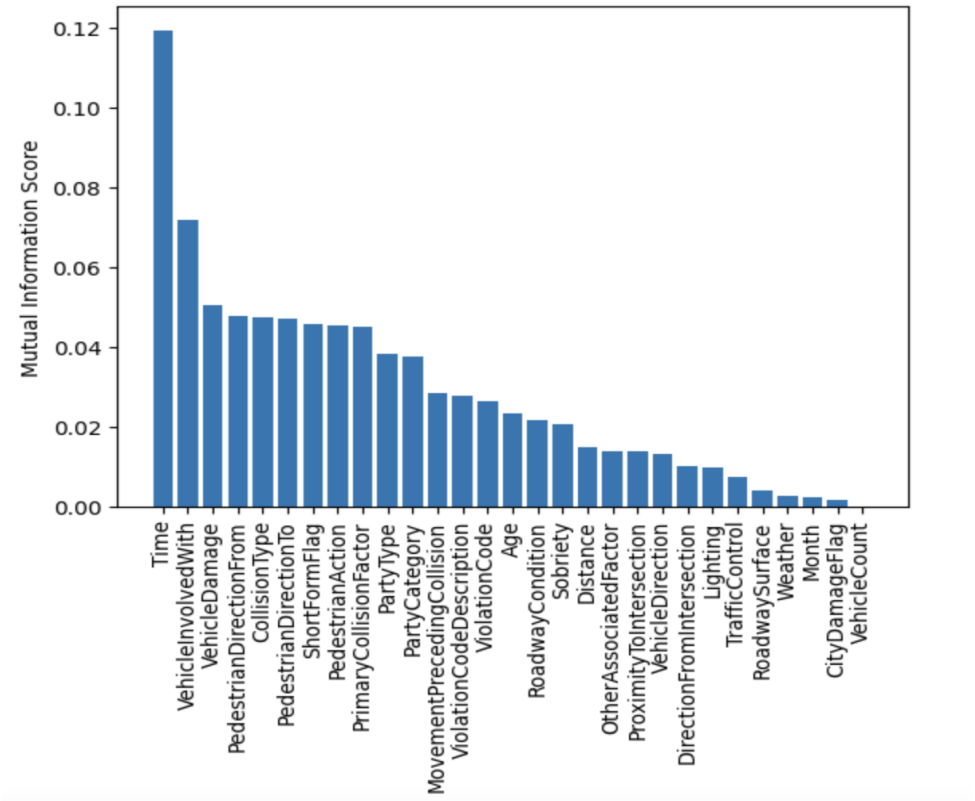
Feature Section Methods

For the purpose of this project, 8 feature selection methods were adopted to compare the reaction of various classification algorithms using the different generated feature sets. Each feature selection method will be discussed below.

Information Gain

Information Gain essentially measures the decrease in entropy and can be adapted likewise to assess important features. Utilizing this measure with respect to Accident Severity, i.e., the target feature, against each of the features we obtain importance scores. The mutual_info_classif function from the sklearn.feautre_selection library was used for this purpose. The importance scores for each feature was visualized using a bar graph as seen in figure 1 below.

Figure 1



It was observed that features like Time, VehicleInvolvedWith and VehicleDamage were recognised to be significant while features like Month and VehicleCount were not.

For the purpose of dropping columns that were not significant, statistics of the mutual information scores were analyzed. Figure 2 below shows the details for it.

Figure 2

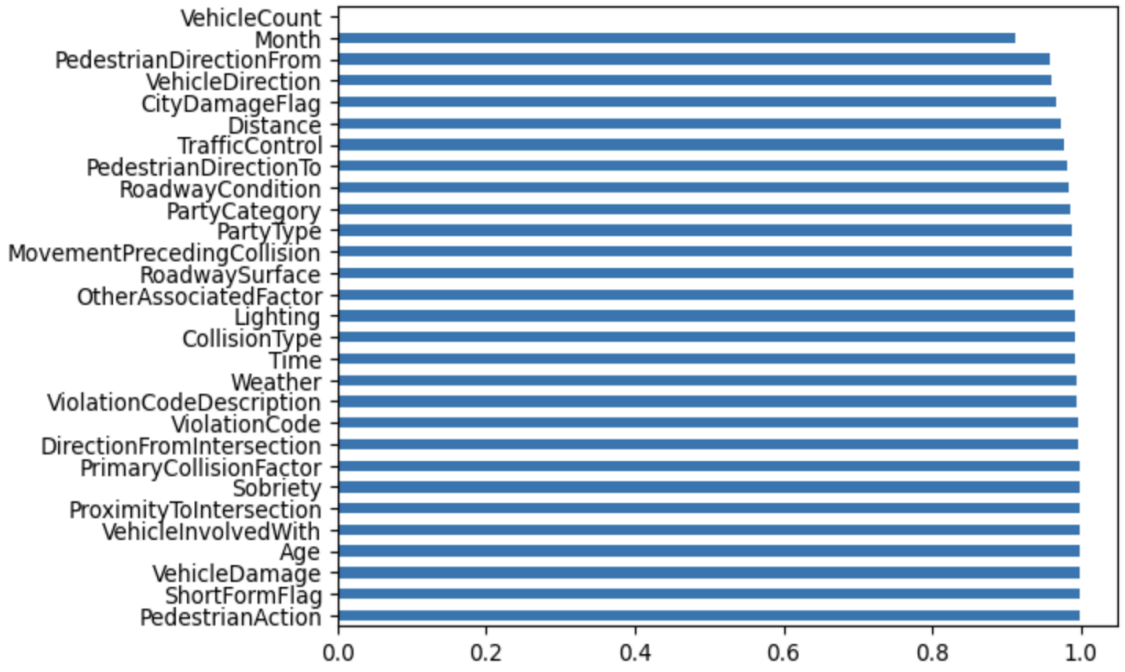
```
count    29.000000
mean      0.028914
std       0.025559
min       0.000000
25%       0.010327
50%       0.023324
75%       0.045461
max       0.119193
dtype: float64
```

On further comparison between the bar graph and the statistics of the score, median deemed to be the better threshold to obtain an optimal count of significant features. Also higher the score more the information. Hence, the threshold for dropping was set to median which was 0.023 where all values less than the median value were dropped. Finally, this method resulted in 15 important features.

Fisher Score

Fisher score gives the ratio of variance between classes to variance within classes. Hence, it can help gauge the degree of discrimination offered by a feature in terms of class separation. Thus, a higher value of fisher score is considered to be more informative. It can be calculated from F-value which measures the degree of variance between the sample mean of a feature and the sample mean of the target feature, Accident_Severity. The f_classif function from the sklearn.feautre_selection library was used for this purpose. Fisher score is calculated from f-value to determine feature ranking with respect to Accident Severity.

Figure 3



The fisher scores for each feature was visualized using a bar graph as seen in Figure 3 above. It can be observed that PedestrianAction, ShortFormFlag, VehicleDamage are highly important while Month and VehicleCount are least important.

For the purpose of dropping columns that were not significant, statistics of the fisher scores were analyzed. Figure 4 below shows the details for it.

Figure 4

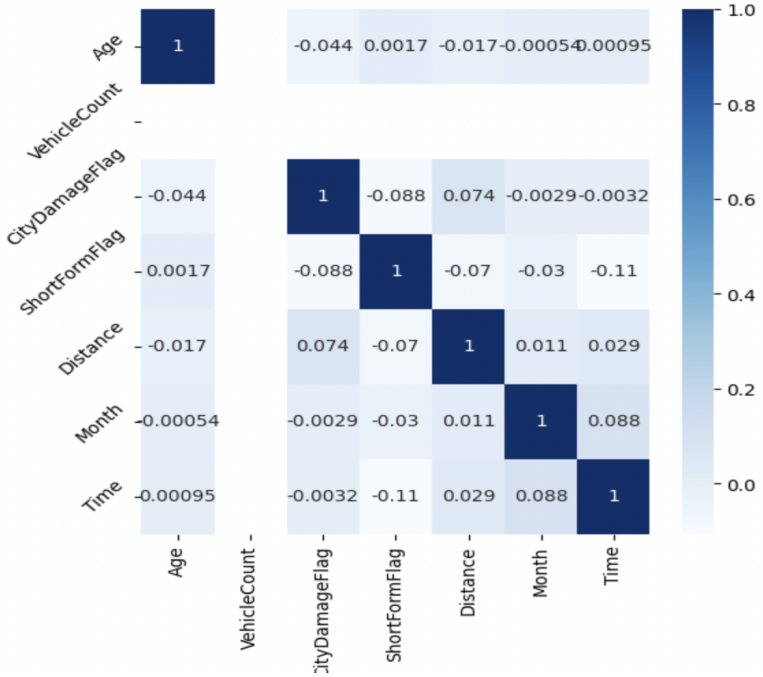
```
count    28.000000
mean      0.986137
std       0.017768
min       0.924589
25%       0.984098
50%       0.992742
75%       0.997844
max       0.999602
dtype: float64
```

On further comparison between the bar graph and the statistics of the fisher score, median deemed to be the better threshold to obtain an optimal count of significant features. Hence, the threshold for dropping was set to median which was 0.993 where all values less than the median value were dropped. Finally, this method resulted in 14 important features.

Correlation Coefficient

Correlation coefficient between all the numeric features of the dataset was determined. This was then used to plot a heatmap. The correlations between the features observed were weak as seen in Figure 5 below. Additionally the dataset is heavily populated with more categorical features hence this approach was not effective. Hence, the selection method was dropped and other ones were explored.

Figure 5



K-Means Method

K-Means was employed to assign data points to a predefined number of clusters. The cluster numbers were experimented with brute force for 4 and 3. Both values resulted in the same feature set. Hence, 3 was finalized as the final number of clusters for efficiency sake. Figure 6 shows the resulting columns for both 3 and 4 clusters. F-test Statistics was used to determine top 10 features. The F-test method in this context computes the difference between the means of the features against predicted clusters. This method identified some domain crucial features such as 'Age', 'ViolationCode' and 'Sobriety'

Figure 6

```
Index(['Age', 'VehicleDamage', 'PartyCategory', 'Sobriety', 'ViolationCode',
       'ShortFormFlag', 'RoadwaySurface', 'RoadwayCondition', 'Weather',
       'VehicleInvolvedWith'],
      dtype='object')
Index(['Age', 'VehicleDamage', 'PartyCategory', 'Sobriety', 'ViolationCode',
       'ShortFormFlag', 'RoadwaySurface', 'RoadwayCondition', 'Weather',
       'VehicleInvolvedWith'],
      dtype='object')
```

Chi Square Test

Chi-Square Test was enabled to find relevant features to crash severity. Chi-Square statistics were computed for the same. 'p-value' from the statistics was used to determine the top 10 features. Figure 7 shows the ordered p-values respective to each feature being evaluated.

This method identified some domain crucial features such as 'Age', 'DirectionFromIntersection', 'Weather' and distance from intersection.

Figure 7

	Feature	P-Value
0	Age	0.000000e+00
26	DirectionFromIntersection	0.000000e+00
23	VehicleInvolvedWith	0.000000e+00
20	Weather	0.000000e+00
13	Distance	0.000000e+00
12	ShortFormFlag	0.000000e+00
10	ViolationCodeDescription	0.000000e+00
9	ViolationCode	0.000000e+00
28	Time	0.000000e+00
1	VehicleDamage	0.000000e+00
7	OtherAssociatedFactor	0.000000e+00
6	PartyType	2.163381e-221
15	RoadwaySurface	1.170096e-215
14	PedestrianAction	5.225670e-179
21	CollisionType	1.013553e-144
22	ProximityToIntersection	1.523207e-129
5	MovementPrecedingCollision	1.181238e-124
3	Sobriety	4.874487e-117
19	TrafficControl	3.135964e-58
18	PrimaryCollisionFactor	3.993641e-55
17	Lighting	8.782146e-42
2	PartyCategory	1.308810e-24
4	VehicleDirection	3.685249e-23
11	CityDamageFlag	9.630456e-18
27	Month	7.760872e-17
25	PedestrianDirectionTo	2.339985e-07
16	RoadwayCondition	3.782564e-07
24	PedestrianDirectionFrom	3.381475e-02
8	VehicleCount	NaN

Lasso Regression

Lasso Regression with cross validation was adopted in order to autotune alpha. Alpha is the regularization parameter that governs the degree of sparsity in the model. Since autotuning is enabled for alpha, the model has the capacity to automatically deduce the suitable level of sparsity and thus choose the most significant set of features for prediction. Less important features have their coefficient shrunken to 0 and thereby removed. The LassoCV function in the sklearn.linear_model library was used for this purpose. Features selected were 'Age', 'VehicleDamage', 'PartyCategory', 'Sobriety', 'VehicleDirection', 'MovementPrecedingCollision', 'PartyType', 'OtherAssociatedFactor'. Similar to other methods this was also able to identify some domain critical features like ‘Age’, ‘Sobriety’ and ‘VehicleDirection’.

Random Forest Importance

Feature significance is deduced by measuring the average reduction in impurity or Gini importance that is attributed to each feature across all decision trees present in the Random Forest model. The feature_importances_ function of the RF model is used for this purpose. For the purpose of dropping columns that were not significant, statistics of the fisher scores were analyzed. Figure 8 below shows the details for it. The importance scores for each feature was visualized using a bar graph as seen in figure 9 below.

On further comparison between the bar graph and the statistics of the score, median deemed to be the better threshold to obtain an optimal count of significant features. Also higher the score more the information. Hence, the threshold for dropping was set to median which was 0.028 where all values less than the median value were dropped. Finally, this method resulted in 19 important features.

Figure 8

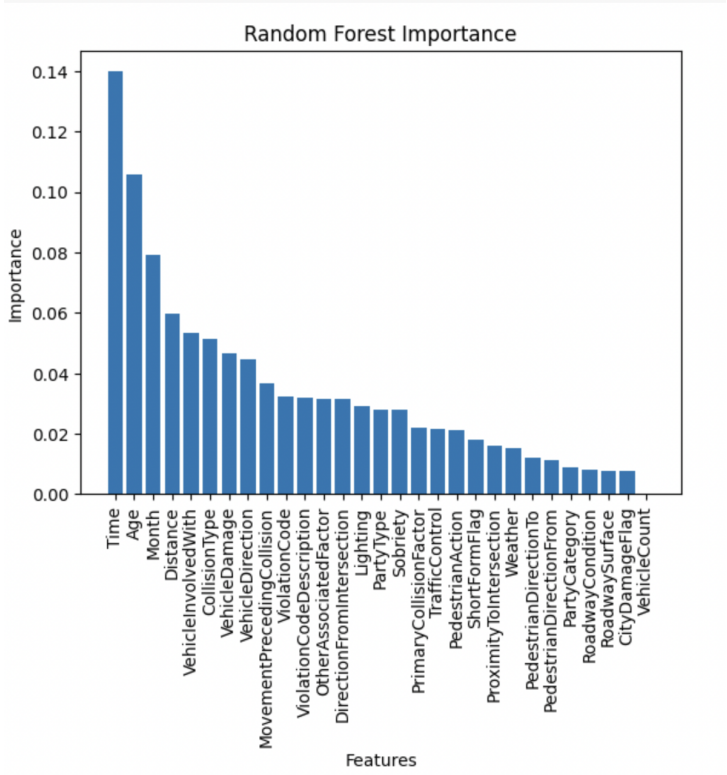


Figure 9

	Importance
count	29.000000
mean	0.034483
std	0.030665
min	0.000000
25%	0.015422
50%	0.028144
75%	0.044797
max	0.139711

Recursive Feature Elimination using RFE

Recursive Feature Elimination recursively removes the least important features from the dataset until a predetermined number of features is reached. It uses a model, in this case, logistic regression to determine the importance of each feature, and then eliminates the least important feature from the dataset. The values of max_iter were set to 1000, which is a higher value than the default as our dataset was quite large. Features selected were 'Age', 'PartyType', 'ViolationCodeDescription','CityDamageFlag', 'ShortFormFlag', 'PedestrianAction','Lighting','CollisionType','VehicleInvolvedWith'. The method proved quite effective in identifying some domain crucial features such as 'Lighting', ''ViolationCodeDescription', etc. We remained with 9 features.

Comparison between different methods for each of the 8 feature set

Decision Tree (DT)

Feature Set	Accuracy	F1 Score	Recall	Precision	AUC
Entire Feature Set	.77	.44	.46	.42	.65
Information Gain	.76	.40	.41	.40	.63
Fisher Score	.76	.44	.45	.43	.64
K-Means	.81	.36	.35	.38	.63
Chi-Square Test	.75	.41	.42	.41	.63
Lasso Regression	.70	.36	.39	.36	.60
Random Forest Importance	.75	.44	.46	.43	.64
RFE with Logistic Regression	.79	.47	.47	.47	.66

Random Forest (RF)

Feature Set	Accuracy	F1 Score	Recall	Precision	AUC
Entire Feature Set	85	.45	.39	.72	.88
Information Gain	.82	.40	.37	.50	.83
Fisher Score	.83	.43	.37	.56	.84
K-Means	.81	.36	.34	.40	.76
Chi-Square Test	.83	.41	.37	.51	.82
Lasso Regression	.76	.35	.33	.41	.70
Random Forest Importance	.84	.42	.38	.65	.87
RFE with Logistic Regression	.81	.41	.38	.50	.81

Naive Bayes (NB)

Feature Set	Accuracy	F1 Score	Recall	Precision	AUC
Entire Feature Set	.53	.23	.31	.28	.58
Information Gain	.60	.26	.30	.29	.61
Fisher Score	.60	.25	.30	.34	.62
K-Means	.63	.27	.30	.28	.66
Chi-Square Test	.52	.24	.32	.29	.59
Lasso Regression	.52	.24	.29	.28	.56
Random Forest Importance	.52	.23	.30	.28	.57
RFE with Logistic Regression	.60	.25	.30	.29	.62

Logistic Regression (LR)

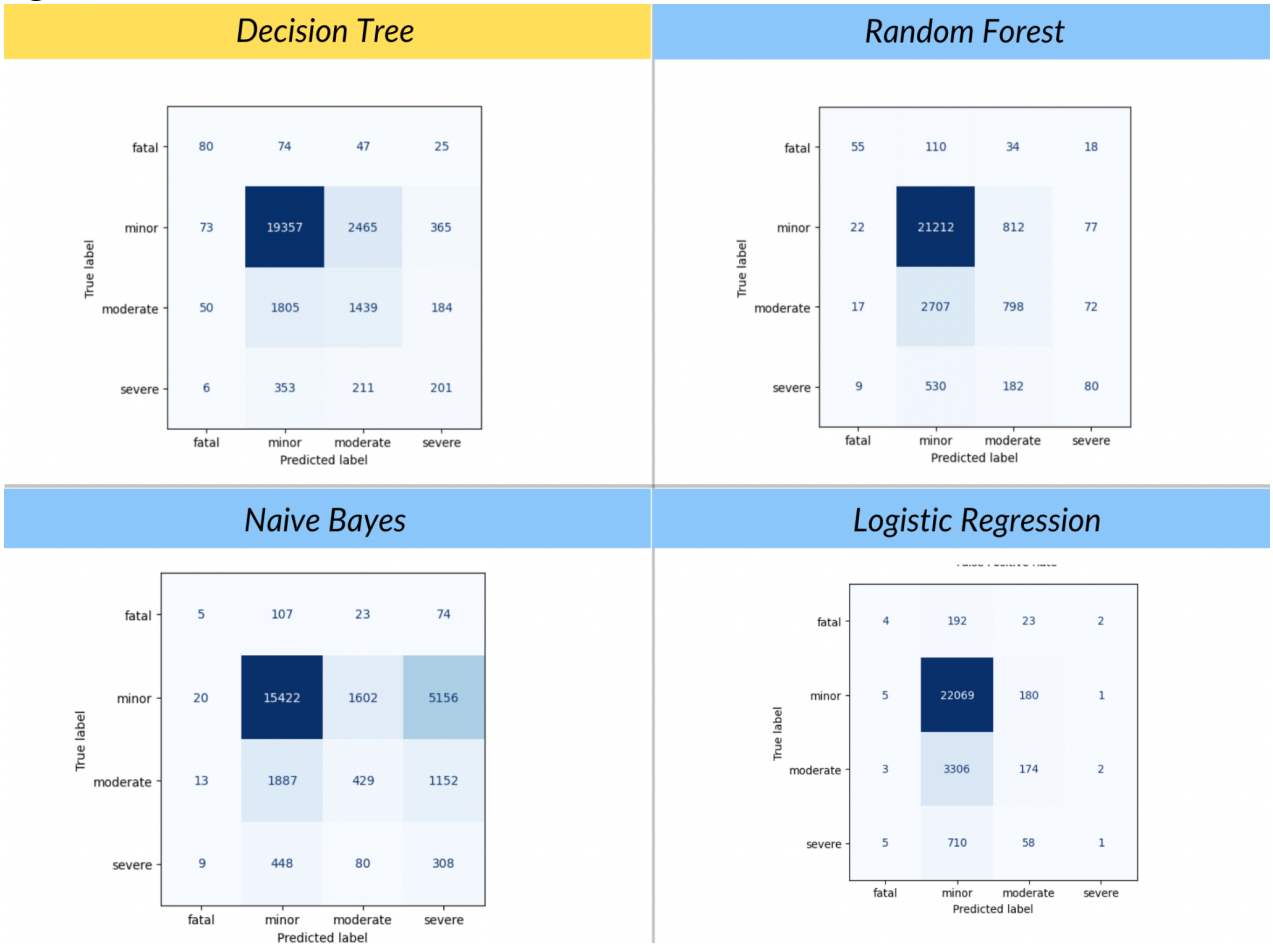
Feature Set	Accuracy	F1 Score	Recall	Precision	AUC
Entire Feature Set	.52	.61	.52	.81	.75
Information Gain	.53	.63	.53	.81	.74
Fisher Score	.50	.60	.50	.81	.75
K-Means	.44	.55	.46	.81	.71
Chi-Square Test	.42	.54	.42	.82	.71
Lasso Regression	.48	.57	.48	.78	.65
Random Forest Importance	.54	.63	.54	.79	.72
RFE with Logistic Regression	.48	.57	.48	.81	.65

Justification for Best Method - Why DT is better than rest?

DT Over RF : Each model has its own best features selected from feature selection methods. Please find below confusion matrices comparing the different best models as shown in below Figure 11. Looking at the confusion matrices we see, DT performed better than RF for predicting the Fatal class even though its overall accuracy is lower than RF. Fatal class is the predominant class in our dataset because it refers to crashes that result in death. As a result, these crashes are typically considered to be the most serious and have the greatest impact on society. Additionally, fatal crashes are often subject to more detailed investigation and analysis, as the goal is to identify the cause of the crash and prevent similar incidents from occurring in the future. This shows the ability of DT to correctly classify the Fatal class along with the minority class. RF also is able to predict all the minority, fatal and severe classes but isn't good enough when compared to DT as it still has a lot of false positives. Logistic Regression and Naive Bayes also are able to predict the minor class well but still cannot accurately predict the Fatal and severe classes and the false positives are more.

Overlap between Moderate and Minor Predictions: Another interesting factor to observe in the confusion matrices are there are many instances where minor and moderate classes are classified as the other. This overlap is expected primarily because the severity column was created based on the severity of injuries. If at least there existed one fatal injury, then crash was considered as fatal, following that if at least there was one severe injury and no fatal injuries then crash was considered as severe. Similarly moderate and minor were determined following this fashion. Moderate and minor injuries are more frequent than severe or fatal and the model classifying them as each other is quite expected.

Figure 11



Decision Tree Performance vs Other Models' Performance

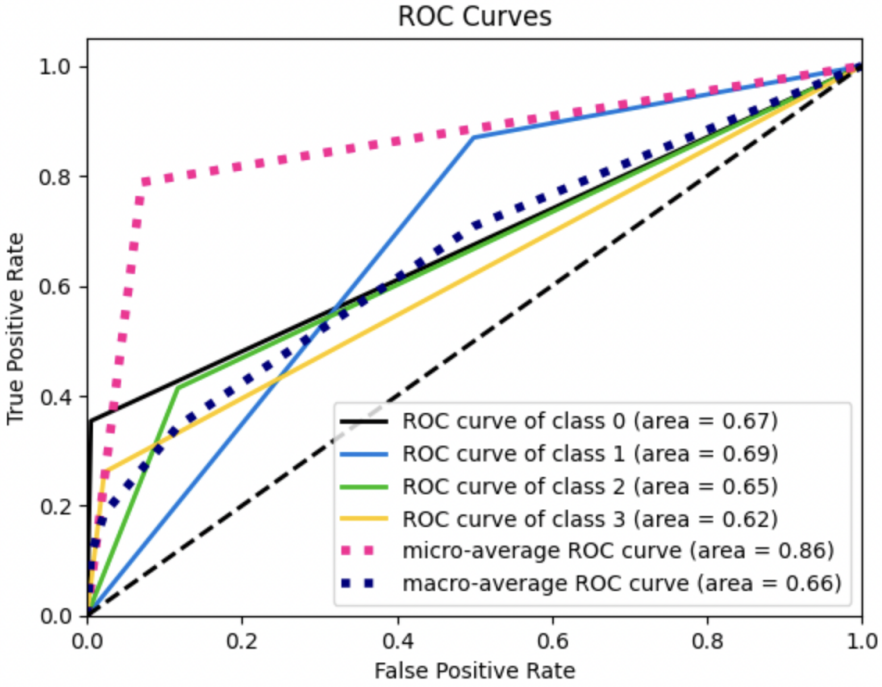
DT outperforms as it is a simpler model and it is easier to interpret. Moreover, it is less prone to overfitting than RF as they are less complex and have fewer parameters to tune and the relationships between the features are relatively simple. DT is known to be effective at handling imbalanced datasets as they divide the data based on the most informative features. By considering these selected features from RFE, the decision tree model captures important aspects related to age, party type, violation codes, pedestrian actions, lighting conditions, collision types, and other relevant factors that align with domain knowledge.Since DT performed well from the selected features from the RFE method. For the decision tree model, we can analyze why this result could be considered good:

- **Nonlinear Relationships**: Decision trees are capable of capturing nonlinear relationships between features and the target variable. In the context of crash severity prediction, there might be complex interactions and combinations of factors that contribute to the severity outcome. The decision tree algorithm can handle these nonlinear relationships effectively.
- **Feature Interactions**: Decision trees can identify and exploit feature interactions that affect the target variable. In the case of crash severity, different combinations of features might have a synergistic effect on the outcome. The decision tree algorithm can split the data based on these interactions, leading to accurate predictions.

- **Feature Importance**: Decision trees provide a measure of feature importance, allowing us to identify the most influential features in predicting the target variable. By selecting relevant features using the RFE method, we have likely included the most informative attributes related to crash severity. This feature selection process helps the decision tree model focus on the most significant factors, leading to improved performance.
- **Handling Categorical and Numerical Data**: Decision trees can handle both categorical and numerical data without requiring extensive preprocessing. In the given dataset, there are categorical variables such as party type, violation codes, and collision types, as well as numerical variables like age. The decision tree algorithm can naturally handle these different data types, making it suitable for this task.
- **Interpretability**: Decision trees offer interpretability, allowing us to understand the decision-making process. We can trace the tree's branches and nodes to interpret how the model arrives at its predictions. This interpretability is valuable in the context of crash severity prediction, as it helps identify the key factors and decision paths leading to different severity outcomes.
- **Robustness to Outliers and Missing Data**: Decision trees are generally robust to outliers and missing data. In real-world crash data, outliers and missing values are common. The decision tree algorithm can handle these data imperfections without significantly compromising performance.

Overall, the decision tree algorithm's ability to capture nonlinear relationships, identify feature interactions, handle different data types, provide interpretability, and handle outliers and missing data likely contributed to its good performance in predicting crash severity based on the selected features.

Figure 12 - ROC-AUC of DT



RF performed very well in terms of accuracy (Figure 11) as it is well-suited to datasets with large numbers of features and they are often used for classification tasks in which features have a complex relationship with target variables but the decision tree algorithm's advantage lies in its simplicity and interpretability compared to the ensemble-based random forest algorithm. Decision trees are easier to understand and visualize, allowing for more transparent decision-making. Therefore, if interpretability and simplicity are important factors, the decision tree algorithm might be preferred over random forest.The Recursive Feature Elimination (RFE) method, used for feature selection in the decision tree model, identified a subset of features that provided the best performance. This indicates that the selected features (such as Age, PartyType, ViolationCode, ViolationCodeDescription, etc.) contain relevant information for predicting crash severity. By focusing on these important features, the decision tree algorithm was able to capture the most discriminative patterns in the data and achieve better performance.The decision tree algorithm achieved comparable performance metrics (accuracy, F1 score, recall) to the random forest algorithm. This suggests that the decision tree was able to capture the underlying patterns in the data effectively, without the need for the more complex ensemble-based approach of random forests

NB and LR performed extremely poor amongst all models. This is due to factors such as imbalanced data, incompatible data, they are linear models which can't handle complex data structure.Naive Bayes struggled to effectively capture the relationships between the features and the target variable, likely due to its assumption of feature independence, which may not hold true in this dataset. Additionally, the imbalance in the class distribution might have affected the performance of Naive Bayes. With a low accuracy, F1 score, and precision, Naive Bayes struggled to correctly classify the instances, leading to lower overall performance.Logistic Regression, similarly performed relatively lower compared to other models. This could be due to the nature of the dataset and the relationship between the features and the target variable. Logistic Regression assumes a linear relationship between the features and the target, and if the relationship is non-linear or complex, the model may struggle to capture it effectively.

Moreover, NB is sensitive to feature selection, and the performance of the algorithm is affected by the quality of features selected, which is why it is the worst model overall.

Importance of Selected Features vs All Features

***All Features*:**

There are total 39 features which when pre-processed was reduced to 29 which are Sex(Text), Age(Numeric), Speed(Text), VehicleDamage(Text), PartyCategory(Text), Sobriety(Text), VehicleDirection(Text), MovementPrecedingCollision(Text), PartyType(Text), OtherAssociatedFactor(Text), VehicleCount(Numeric), ViolationCode(Text), ViolationCodeDescription(Text), TcrNumber(Text), CityDamageFlag(Text), ShortFormFlag(Text), Distance(Numeric), CrashDateTime(TimeStamp), PedestrianAction(Text), RoadwaySurface(Text), RoadwayCondition(Text), Lighting(Text), PrimaryCollisionFactor(Text), TrafficControl(Text), Weather(Text), CollisionType(Text), ProximityToIntersection(Text), VehicleInvolvedWith(Text), PedestrianDirectionFrom(Text), PedestrianDirectionTo(Text), DirectionFromIntersection(Text) and Comment(Text), Severity(Text).

Crashes dataset is a realtime dataset and contains a lot of features. It can be observed that there are a lot of features even after pre-processing to remove some generic irrelevant ones. This

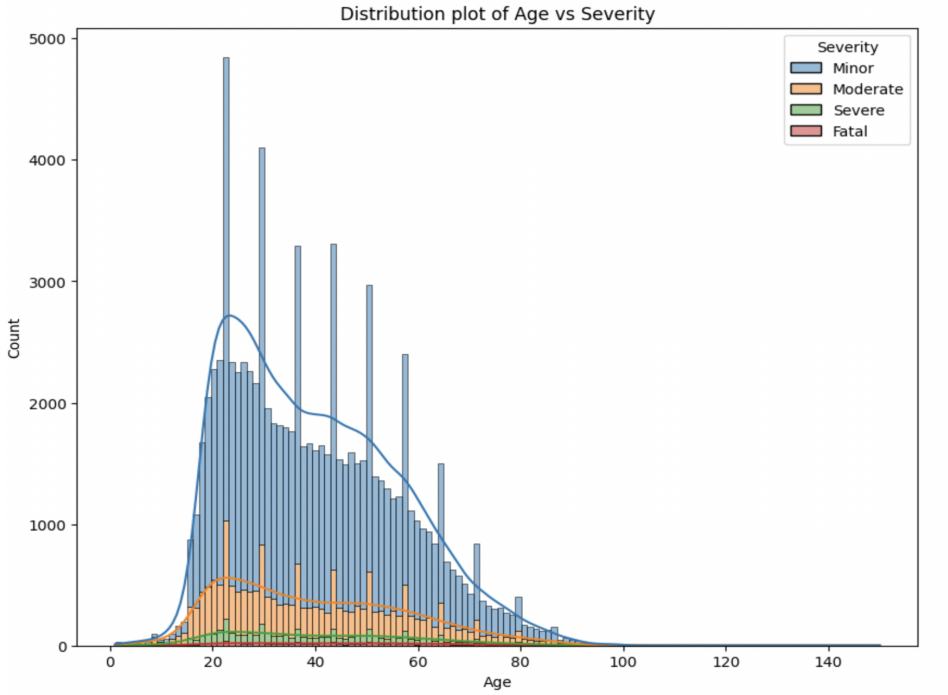
could affect performance as finding meaningful patterns can be hindered. Dimensionality reduction of this data is quite necessary for handling computational complexity. This can also improve accuracy by focusing on important features. Reducing overfitting is also another advantage. It is also crucial to find relevant features to establish safety protocols and improve safety conditions to justify project goals. Acclaimed feature selection methods were then applied.

***Selected Best Features with RFE for Decision Tree (Best Model)*:** Age, PartyType, ViolationCodeDescription, ShortFormFlag, PedestrianAction, Lighting, CollisionType, VehicleInvolvedWith, Time

The selected features are important because they provide valuable insights and information that can help in understanding severity of vehicle crashes. It was also interesting to observe certain features being picked up that stood out with insightful information at the time of EDA. Each selected feature is discussed below:

- **Age**: This can help identify if certain age groups are more prone to crashes or if there are specific trends based on age. Most of the feature selection methods gave Age as one of the important features and it is observed while performing EDA that ages between 20-40 are most likely to be involved in crashes as shown in Figure 13.

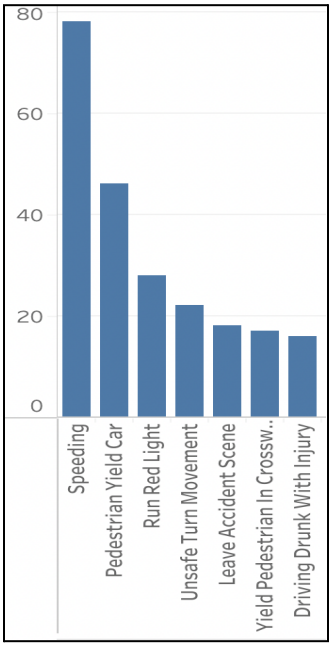
  Figure 13



- **PartyType**: This can help identify the types of parties involved in the crash, such as drivers, passengers, pedestrians, or cyclists. For instance, if the crash involves a pedestrian or cyclist, it is likely to be more severe than a crash that only involves drivers and passengers.
- **ViolationCodeDescription**: This can help identify if any traffic laws were violated leading up to the crash. The ViolationCodeDescription feature can be particularly useful in predicting the severity of crashes caused by reckless driving or DUI violations. These types of violations are more likely to result in severe crashes and injuries. Therefore, including the ViolationCodeDescription feature in the model can improve the accuracy of

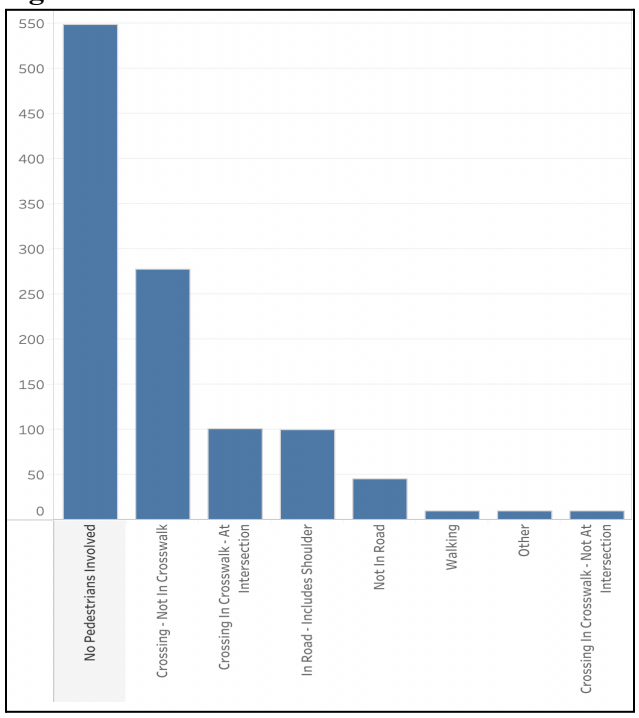
the prediction. From the given dataset, it can be seen that most of the fatalities occurred due to speeding followed by pedestrian yield cars and other violation codes as shown in below Figure 14.

Figure 14



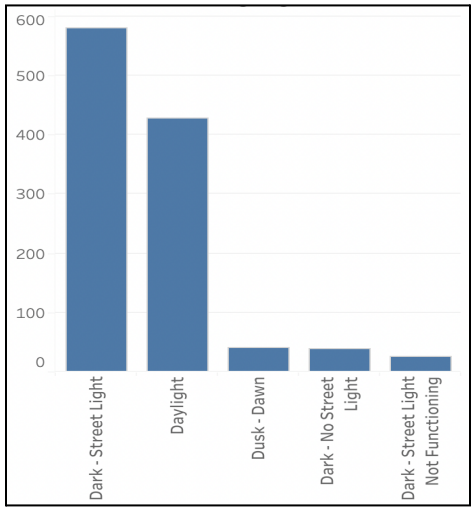
- **ShortFormFlag**: This can help identify if the crash report was a short form or a long form, which can provide information on the level of detail in the report. The ShortFormFlag feature is an indicator of whether the crash report is a short-form or long-form report. Short-form reports are mainly used for less severe crashes or minor crashes, while long-form reports are used for more severe crashes. By including this feature in the crash severity model, the algorithm can use it to help differentiate between less severe and more severe crashes, which can help emergency responders prioritize their response and allocate resources more effectively. In the given dataset, all crashes with no severity have ShortFormFlag as True and for crashes with a certain level of severity has ShortFormFlag as False.
- **PedestrianAction**: This can help identify if the crash involved a pedestrian and what action the pedestrian was taking at the time of the crash.The PedestrianAction feature is typically collected from video footage or other sources. This feature analyzes the pedestrian's actions during the moments leading up to the collision, such as walking, standing, or crossing the street. Based on this analysis, the PedestrianAction feature can provide information about the speed and direction of the pedestrian's movement, as well as their posture and body language. This information can then be used to determine the severity of the crash and the likelihood of injury to the pedestrian. For example, if the pedestrian was walking across the street at a slow pace and was hit by a car traveling at a low speed, the crash may be classified as less severe than if the pedestrian was running across the street and was hit by a car traveling at a high speed. It can be seen from the below Figure 15 that most of the fatalities have undertaken when there wasn't any pedestrian action followed by pedestrians walking not in the crosswalk.

Figure 15



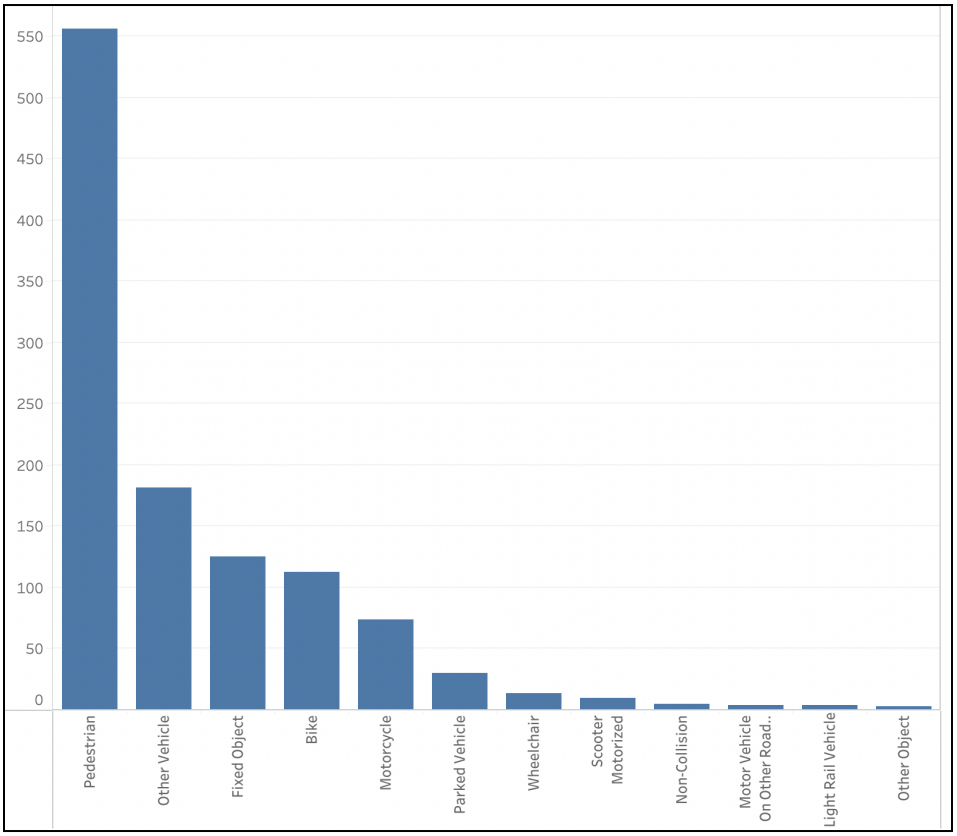
- **Lighting**: This can help identify if the crash occurred during daylight, dusk, or night time, which can provide information on visibility conditions. If a crash occurs in dark light conditions, it may be more difficult for drivers to see and avoid obstacles or hazards on the road, which can result in a more severe crash. Additionally, if the lighting conditions are poor, it may be more difficult for witnesses or investigators to accurately assess the scene of the crash and determine what happened. By taking into account the lighting conditions at the time of the crash, investigators can better understand the circumstances surrounding the crash and potentially identify factors that contributed to the severity of the crash. This information can be used to improve road safety and prevent future accidents. It can be verified from the EDA performed as shown in below Figure 16 that fatal accidents occur mostly when street light is dark.

Figure 16



- **CollisionType**: This can help identify the type of collision that occurred, such as rear-end, head-on, or sideswipe. By analyzing the CollisionType feature, accident investigators can gain insights into the factors that contributed to the accident and the severity of the crash. For example, head-on collisions are often more severe than other types of collisions, as they tend to involve higher speeds and greater forces. Similarly, rear-end collisions are typically less severe, but can still result in injuries or property damage.
- **VehicleInvolvedWith**: This can help identify the type of vehicle involved in the crash, such as a car, truck,pedestrian or motorcycle. This feature captures information about the type of object/individual that the vehicle collided with during the crash, such as another vehicle, a pedestrian, a fixed object, or an animal. For example, crashes involving collisions with pedestrians or fixed objects may be more likely to result in serious injuries or fatalities than crashes involving only other vehicles. It can be seen from below EDA Figure 17 that most of the fatal accidents have occurred when vehicles collided with pedestrians.

Figure 17



- **Time**: This can help identify the time of day or night the crash occurred, which can provide information on traffic patterns and congestion. By analyzing the time of day, day of the week, month, and year when the accident occurred, we can identify patterns and correlations with crash severity. For example, accidents that occur during peak traffic hours or on weekends may be more severe than those that occur during off-peak hours or on weekdays.

Hence, the total number of features reduced from 29 to 9 decreasing the training time of modeling and the model can focus on the most important features, leading to better performance. Feature selection can help prevent overfitting by reducing the complexity of the model .

Improving Safety Features and Protocols based on Our Results

In order to address the particular dangers and problems associated with a crash, policymakers and road safety practitioners can adjust interventions and tactics by knowing the distinct elements that contribute to crash severity in various circumstances.The following safety standards and measures are proposed based on the examination of the decision tree (DT) model, the features of pedestrian behavior, violation codes, and collision types, as well as the aspects that have been determined to be significant for enhancing safety, such as lighting conditions, sobriety, and time.

- **Pedestrian Protection:** Improve pedestrian infrastructure to make walking safer by upgrading crosswalks, sidewalks, and pedestrian signals. Increasing visibility for pedestrians To increase visibility and lower the risk of collisions, add enough illumination to busy pedestrian areas and pedestrian crossings. Organize educational campaigns: Inform pedestrians about safe practices like using designated crosswalks, adhering to traffic signals, and putting away cell phones when out walking.
- **Enforcement of Code Violations:** Boost enforcement actions: Increased traffic enforcement and surveillance, particularly for high-risk behaviors including speeding, careless driving, and running red signals. Put red-light cameras in place: Red-light cameras should be placed at intersections to increase compliance with traffic regulations and deter violators, which will lower the likelihood of collisions.
- **Manage collision types:** Redesign of intersections: To lower the frequency of particular crash types, analyze high-risk intersections indicated by the DT model and take into consideration reconstructing them with better turning lanes, more obvious signage, and improved signal timings.
- **Enhance road infrastructure:** To lessen the severity of collisions and reduce the likelihood of particular collision types, upgrade road infrastructure by installing rumble strips, crash barriers, and obvious road markers.
- **Lighting circumstances:** Upgrade the street lighting to ensure enough illumination on the roads, particularly in high-crash zones, to improve visibility for motorists, bicyclists, and pedestrians.
- **Perform routine maintenance**: Street lights should be routinely inspected and maintained to ensure that they are operating as intended; any malfunctioning or dim lights should be replaced right away.
- **Sobriety:** Enforce strong drunk driving regulations: Through sobriety checkpoints, random breath testing, and public awareness campaigns, increase the enforcement of rules against driving while intoxicated. Encouraging designated drivers to discourage driving while intoxicated, promote the use of designated drivers and offer accessible alternatives, such public transportation or ride-sharing services. Though Sobriety was not picked up by RFE it was identified in many other feature selection methods such as using Random Forest Importance which was best for RF model that did have highest accuracy among all models.
- **Time Management:**Optimize intersection traffic signal timings by analyzing traffic patterns in order to lessen congestion and the risk of accidents.

- **Utilize clever traffic management techniques:** Implement sophisticated traffic management systems that may modify signal timings in response to current traffic circumstances, thereby easing congestion and enhancing overall safety.

Interpretability based on domain knowledge and references

The goal of our project is to increase road safety by employing collision data analysis through classification. In order to understand the causes, contributing variables, and patterns of accidents, crash data analysis entails gathering and analyzing information about traffic accidents. Researchers and decision-makers can create effective strategies and interventions to prevent accidents and lessen their severity by analyzing this data.

The expected severity of a crash can be predicted by crash severity models, which can assist trauma centers in estimating the potential effects and offering appropriate and quick medical care. This is especially crucial when accidents take place in remote locations or close together. Rapid severity prediction allows trauma centers to send out emergency vehicles that are properly equipped to accidents and then direct them to hospitals or emergency facilities that can handle the patients quickly and efficiently. Our analysis reveals that certain features, such as Age, Pedestrian Actions, VehicleInvolvedWith and Lighting Conditions, exerted significant influence on the target class across various feature selection methods. These findings are consistent with existing research conducted on these factors within the context of the problem domain.

According to Dias et al. (2023), a number of elements, such as lighting, weather, infrastructure, human factors, accident types, vehicle category, and color, were taken into account to assess their impact on the severity of the incident. They were able to draw the following result from their research: inadequate road illumination had the biggest impact on crash severity. Street lighting and vehicle categorization were discovered to be the two most important factors in our research and analysis of the crash data in San Jose. These characteristics are highlighted as the top features by our best performing model, the decision tree model, using the RFE feature selection method, which supports the fact that poor lighting circumstances might impair driving visibility and judgment.

Also, based on the conclusions of Theofilatos et. al. (2012), young drivers, bicycles, intersections, and collisions with fixed objects are factors affecting road accident severity only in urban areas, while weather conditions, head-on collisions, and side collisions are factors affecting severity only outside urban areas. This contrast highlights the specific road users and traffic situations that should be the focus of road safety measures for the two different kinds of networks (inside and outside urban areas). In our analysis as well, we noticed that factors like age, collision type and vehicles involved play a crucial role in determining the severity of the crash.This supports the findings of Yannis et al. and highlights the significance of taking these factors into account when creating road safety measures.

Notably, the Age feature, which represents the age of the individuals involved, emerged as a strong predictor of severity in our study. Here’s prior research on this. The study titled "Age-Related Differences in Motor-Vehicle Crash Severity in California" examines various crash characteristics and provides recommendations to improve roadway safety for different age groups. The study was conducted by Adekunle Adebisi, Jiaqi Ma, Jaqueline Masaki, and John Sobanjo. The objective of the study is to analyze factors influencing motor-vehicle crash injury severity for young drivers (aged 16-25), middle-aged drivers (aged 26-64), and older drivers (above 64) in the state of California. The study found that there are close relationships between

severity determinants for young and middle-aged drivers. Young drivers are more likely to explore their driving skills due to their newness to driving, while middle-aged drivers tend to demonstrate less cautious behaviors, especially male drivers. On the other hand, older drivers tend to be most cautious among all age groups under all environmental and roadway conditions.

Our analysis indicates that the Lighting conditions feature exhibited substantial influence on the target class across multiple feature selection methods. This finding aligns with existing research conducted on lighting conditions within the problem domain. Jafari Anarkooli and Hadji Hosseinlou (2016) conducted a study to analyze the injury severity of crashes on two-lane rural roads under different lighting conditions. The results reveal that the Lighting Conditions feature played a critical role in determining the severity of the outcome. Its inclusion in the feature set led to improved performance across various feature selection methods, as evidenced by higher accuracy, F1 score, recall, precision, and AUC values.The influence of lighting conditions on accident severity is well-documented in the research.

Similarly, our analysis indicates that the VehicleInvolvedWith feature exhibited notable influence on the target class across multiple feature selection methods. This finding aligns with existing research conducted on factors related to the vehicles involved in accidents.
The results demonstrate that the VehicleInvolvedWith feature played a significant role in determining the severity of accidents. Its inclusion in the feature set resulted in improved performance across various feature selection methods, as evidenced by higher accuracy, F1 score, recall, precision, and AUC values.Research in the domain of accident analysis has highlighted the importance of considering the vehicles involved in incidents. The type of vehicle or vehicles involved in an accident can have a substantial impact on the severity of the outcome. Roudsari et al. (2004) conducted a study to examine the impact of vehicle type on pedestrian injuries, specifically comparing light truck vehicles (LTVs) with passenger vehicles.

References

Theofilatos, A., Graham, D. J., & Yannis, G. (2012). Factors Affecting Accident Severity Inside and Outside Urban Areas in Greece. *Traffic Injury Prevention*, *13*(5), 458–467. https://doi.org/10.1080/15389588.2012.661110

Dias, D., Silva, J. R., & Bernardino, A. (2023). The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal. *Informatics (Basel)*, *10*(1), 17. https://doi.org/10.3390/informatics10010017

Adebisi, A., Ma, J., Masaki, J., & Sobanjo, J. (2019). Age-Related Differences in Motor-Vehicle Crash Severity in California. Safety, 5(3), 48.
https://doi.org/10.3390/safety5030048

Jafari Anarkooli, A., & Hadji Hosseinlou, M. (2016). Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads. Journal of Safety Research, 57, 57-65.
https://doi.org/10.1016/j.jsr.2015.12.003

Roudsari, B. S., Mock, C. N., Kaufman, R., Grossman, D., Henary, B. Y., Crandall, J., ... & Dischinger, P. (2004). Pedestrian crashes: Higher injury severity and mortality rate for light truck vehicles compared with passenger vehicles. Injury Prevention, 10, 154-158.