***Predicting Injury Outcomes of Traffic Accidents in Orange County, Florida***

ISC4242 Final Report

Group 3: Rhea Nibert, Ava Nafisi, Jordan Allwood, Meghana Mukkoti, Shivani Ananth

**Problem Statement and Motivation**

According to the World Health Organization, around 1.19 million people die in traffic accidents each year, and between 20 and 50 million more people suffer injuries[[2]](https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries). This shows that traffic accidents are a persistent public safety concern. While some accidents result in minor damage and injuries, others can have devastating consequences, including serious injuries or fatalities. Understanding the underlying factors that influence the severity of injury outcomes in traffic accidents is essential for designing effective road safety interventions, shaping transportation policy, and informing risk assessment models for insurers and law enforcement.

Our project aims to build machine learning models that predict the severity of injuries resulting from traffic accidents using real-world data. Specifically, we aim to classify accidents into one of four categories: no injuries, nonfatal injuries, serious injuries, or fatalities. Rather than treating this as a simple binary classification task of injury/no injury, we attempt to address a multi-class classification problem that captures the nuanced spectrum of injury outcomes. Our approach emphasizes feature selection, data preprocessing, and insights generation to support evidence-based decision-making.

Our motivation for this work comes from the need to enhance traffic safety measures and reduce injury rates by identifying the most influential factors contributing to accident outcomes. Insights derived from this project can help public safety officials prioritize high-risk scenarios, guide driver education efforts, and improve road design. Moreover, insurance companies and emergency responders can use such models to optimize resource allocation and response strategies. By leveraging a comprehensive traffic accident dataset, we aim to bridge the gap between raw data and practical road safety improvements.

**Introduction and Description of Data**

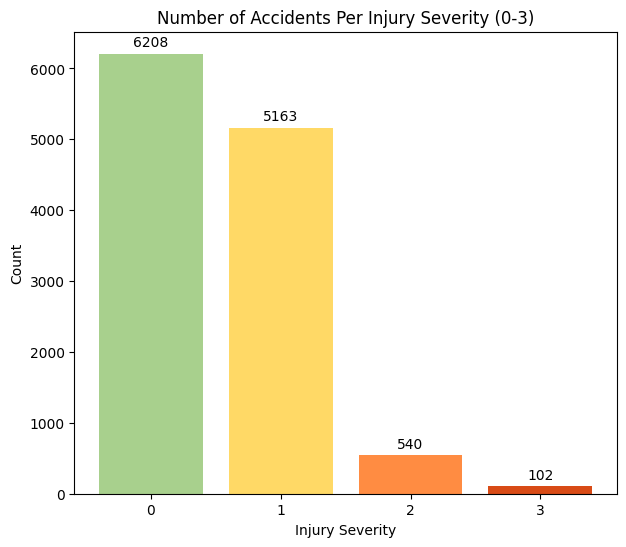
We obtained that dataset for this project from the Florida Department of Transportation (FDOT), specifically from the State Safety Office Geographic Information System (SSOGis) query tool[[1]](https://gis.fdot.gov/ssogis/). We scoped our data collection to Orange County in the year 2020 to maintain a manageable and focused dataset, while still capturing a diverse range of accident types and circumstances. Due to download limitations imposed by the SSOGis platform, we segmented our queries by month and merged the resulting CSV files into a single dataset using Python’s Pandas library.

The final compiled dataset includes 12,038 traffic accident records with 126 original features describing various conditions surrounding each incident. These features span a wide array of categories, including road conditions, driver actions, vehicle information, environmental factors, and time-of-day specifics.

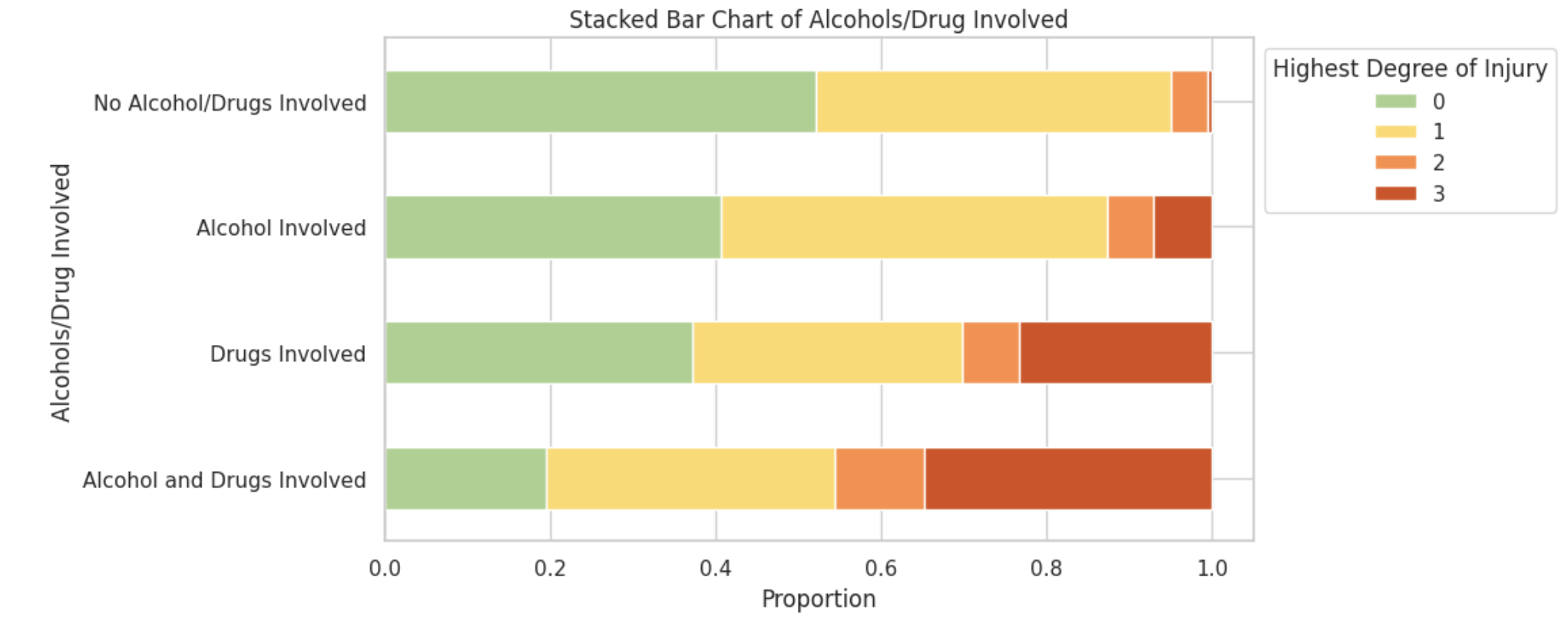
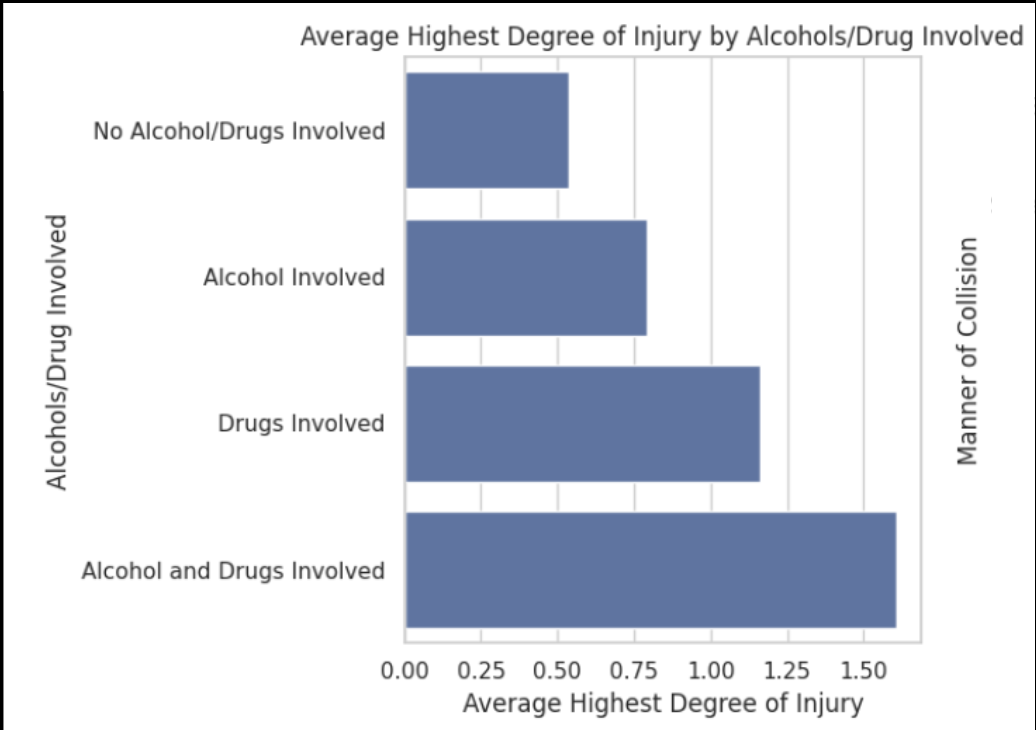
Since not all of these features are relevant or clearly defined, we undertook significant preprocessing steps. This included:

* Dropping ambiguous or redundant columns
* Consolidating overlapping variables (e.g., merging various driver behavior indicators into a single “Driver Action” field)
* Standardizing categorical variables by merging rare or overly specific categories into broader, more interpretable groups
* Creating a new response variable that categorizes the severity of each accident based on the most serious injury reported:
* **0**: No Injuries
* **1**: Nonfatal Injuries (Minor)
* **2**: Severe Injuries (Major)
* **3**: Fatalities

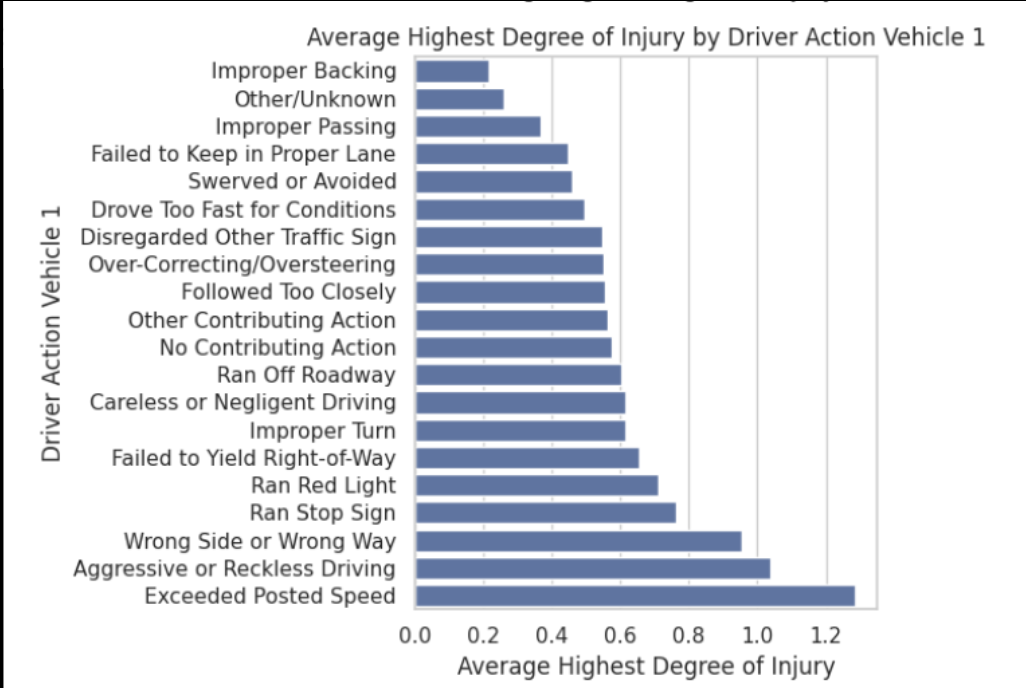
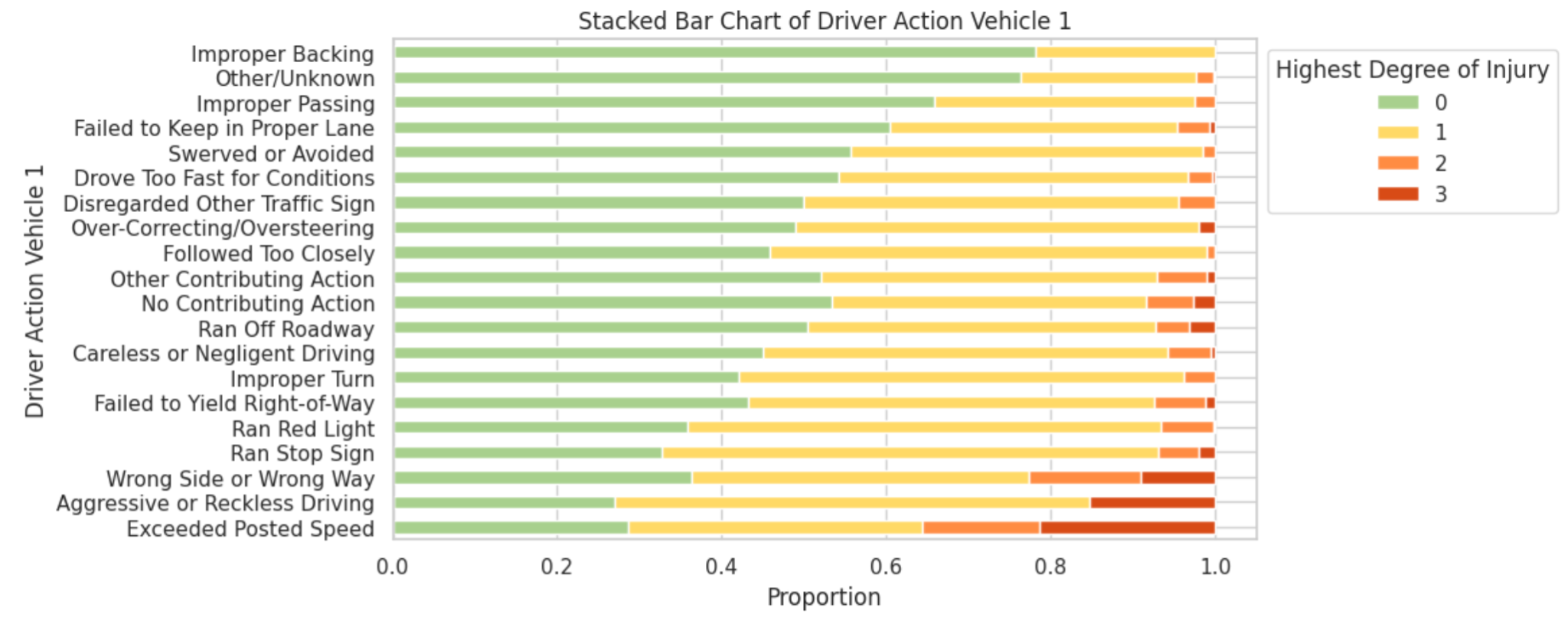
We reduced the dataset to 50 meaningful features and handled missing data based on the data type of the column. For example, we imputed missing categorical values with “Other/Unknown” and filled numerical gaps using either the mean or median, depending on the variable type. In cases where columns had excessive missing data (such as GPS coordinates), we decided to remove them rather than relying too much on imputation.

After cleaning and processing the data, our next step was to create visualizations to explore the relationship between the data features and injury severity. We aimed to use insights from this process to identify key features to use in our models. We began by examining the class imbalance with a bar chart, which revealed a sharp drop from nonfatal injuries to severe injuries and fatalities. This imbalance is important and will guide how we approach both the data and the modeling process moving forward. One consequence of this imbalance is that scatter-based visualizations become less effective; the majority classes overshadow the minority ones, making it hard to identify clear patterns. For this reason, we focused on proportional visualizations, such as bar charts showing the average injury severity and stacked bar charts that better capture the overall distribution.

In creating visualizations to explore our data, we focused on a few key discrete and categorical variables, as these are well-suited for bar chart representations. For each of these variables, we plotted the average injury severity by category level and created stacked bar charts to show the proportion of each injury class within each level. When the proportions vary noticeably across levels, it suggests that the variable could be a strong predictor, as different levels appear to correspond to different injury severities. One particularly insightful set of visualizations compared alcohol and drug involvement with injury severity. As shown in the plots, drivers who were impaired at the time of the accident were involved in crashes that tended to result in more severe injuries.

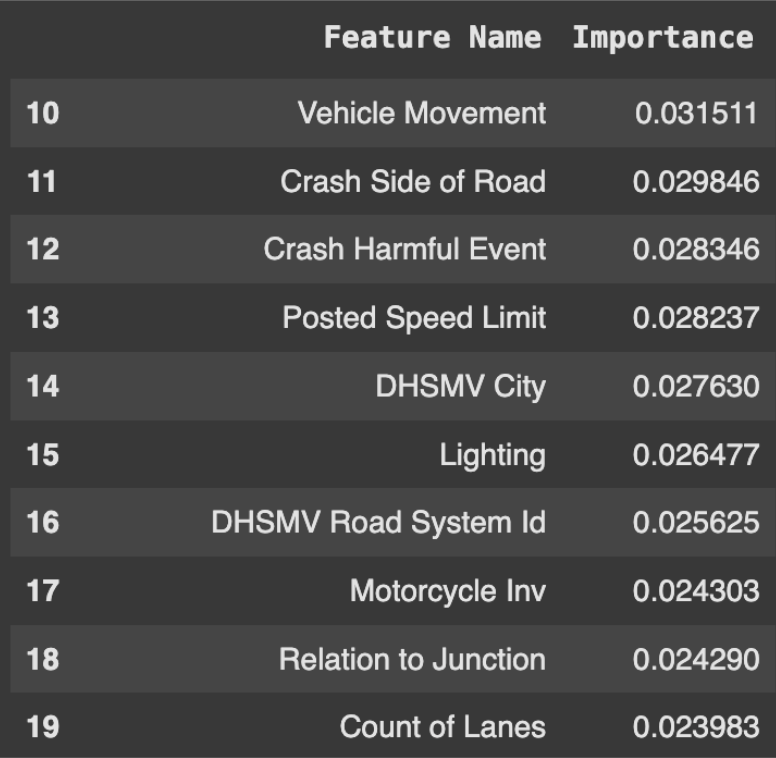
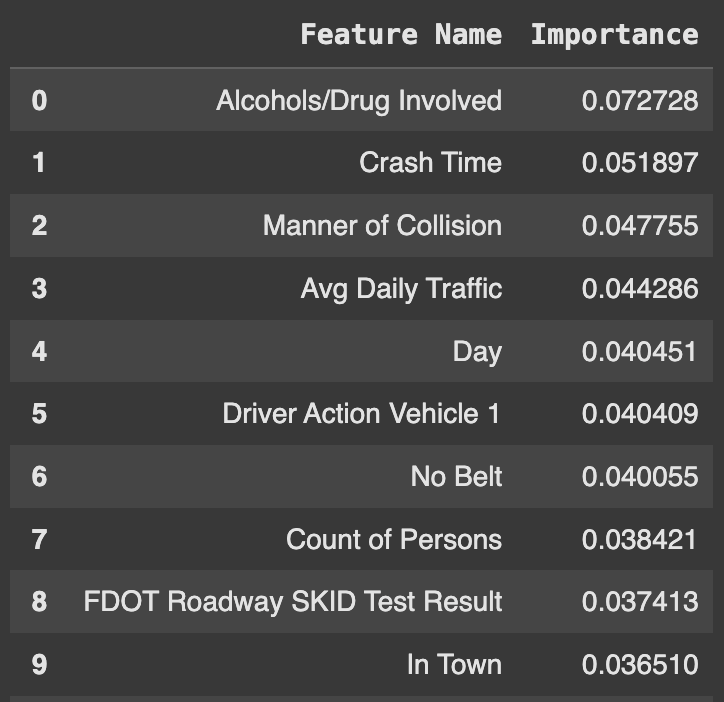


Another meaningful chart was for the driver action variable, which records which action the driver engaged in to cause the accident. The chart reveals that driving on the wrong side of the road, speeding, and reckless driving are behaviors that are most likely to result in a higher degree of injury.



**Modeling Approach**

Now that we’ve explored the structure of our data and identified potentially important features, we began our modeling process by further narrowing down the feature list. Our dataset currently has 50 recorded features about each accident, so our goal was to narrow this down to 20, since this is a more manageable number for machine learning models. We used a Random Forest Classifier due to its built-in feature importance metrics and efficiency with large feature sets. To prepare the data, we scaled numeric variables (e.g., average daily traffic) to have a mean of 0 and a standard deviation of 1, ensuring consistent comparison across features. Categorical variables were target encoded, replacing each level with the average degree of injury for that level, similar to our earlier bar chart analysis. To address class imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples, allowing the Random Forest to better assess feature importance across all injury levels. The top 20 features are shown in the table below:



Next, we spot-checked several machine learning models to identify which ones to prioritize moving forward. The models we tested included logistic regression, K-nearest neighbors, support vector machines, and boosting algorithms. We emphasized F1-score over overall accuracy, since accuracy can be misleading in the presence of class imbalance. The F1-score, which is the harmonic mean of precision and recall, provides a better measure of a model’s ability to correctly identify minority classes while minimizing false positives and false negatives (i.e., predicting an accident resulted in a fatality when it didn’t, or misclassifying fatalities as lower degrees of injury). This focus ensures our models are more reliable in detecting the most severe outcomes.

For all our models, we applied SMOTE to balance the class distribution and used a train-test split with 30% of the data held out for testing. We continued the scaling and target encoding steps detailed previously, taking care to apply target encoding only to the training set to avoid data leakage; that is, preventing the test data from influencing the encoded values for categorical levels.

Our goal is to identify the models that show the most promise in classifying accidents by their highest degree of injury and fine-tune their parameters to achieve optimal performance. If multiple models perform well, or if certain models demonstrate complementary strengths and weaknesses, we plan to ensemble them, combining their predictions to enhance overall accuracy and robustness.

**Project Trajectory, Results, and Interpretation**

In this section, we’ll highlight the results of our modeling process and changes we had to make to combat challenges with our imbalanced data.

We first explored a baseline **logistic regression** model without applying SMOTE. Logistic regression is a statistical model that estimates the probability of an outcome based on a linear relationship between the independent variables and the log-odds of the outcome. Initially, the model achieved moderate performance, with an accuracy of 61.3% and a weighted F1-score of 0.60. However, a clear class imbalance was evident: the model performed well on the majority classes (0 and 1) but poorly on the minority classes (2 and 3), as reflected by low F1-scores and recall values. After applying SMOTE to oversample the minority classes, the model’s accuracy dropped to 49.99%, but there was a notable improvement in recall for classes 2 and 3, indicating increased sensitivity to minority classes. This improvement, however, came at the cost of precision, which declined significantly for the minority classes, suggesting a higher rate of false positives. Overall, the macro-average recall increased from 0.33 to 0.54, highlighting better balance across classes, though the decrease in precision reflects the typical trade-off between improving recall and maintaining precision.

To explore a model less constrained by such assumptions, we next tested a nonparametric approach: **K-Nearest Neighbors (KNN).** KNN is a simple algorithm that classifies data points based on the majority class of their k-nearest neighbors, where the number of neighbors (k) is a key parameter. For the KNN model, we selected the number of neighbors by plotting model performance up to 100 neighbors and identifying the optimal value using the elbow method. Our initial KNN model performed moderately, achieving an accuracy of 60.2%; however, it failed to correctly identify any instances of classes 2 and 3, resulting in a poor overall F1-score of 0.31.

To address this, we trained a second KNN model using stratified data to ensure the train-test split better represented the minority classes. This slightly lowered the accuracy to 57.9%, but the model successfully classified some instances of Class 2, representing a modest improvement. We also attempted KNN with Principal Component Analysis (PCA), reasoning that the high dimensionality of the data might be hindering performance. Unfortunately, the KNN model with PCA performed the worst, again failing to classify any instances of classes 2 and 3. Condensing the data into two principal components likely caused the loss of subtle patterns associated with the minority classes. Overall, KNN’s performance was middling, suggesting it is not the most suitable model for this dataset.

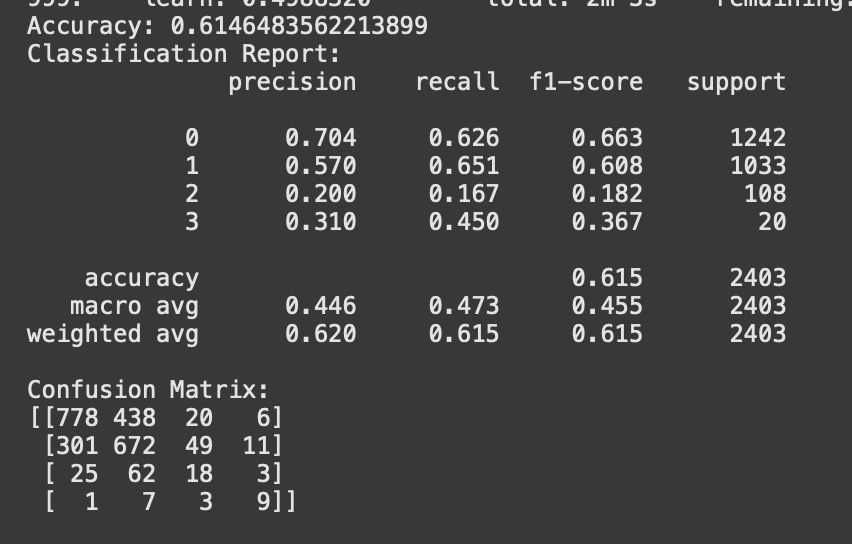
We next tested **Support Vector Machines (SVM)** to address some of the limitations observed with KNN. SVM is a supervised learning method that identifies the hyperplane that best separates classes by maximizing the margin between them. This approach is effective for both linear and non-linear classification problems. For our project, we trained an initial SVM model using an RBF kernel and applied SMOTE to help manage the class imbalance. This baseline model achieved an accuracy of 58% and a weighted F1-score of 0.56, but, similar to previous models, it struggled to correctly classify the minority classes representing severe injuries and fatalities.

To improve sensitivity to these minority classes, we trained a Class-Weighted SVM, which assigns greater penalties to errors involving underrepresented categories. By adjusting the class weights, the model is encouraged to treat misclassifications of severe injuries and fatalities as more costly, thereby improving recall for these outcomes. The class-weighted SVM achieved a slightly higher accuracy of 59% and a weighted F1-score of 0.58, indicating modest improvement over the standard SVM. However, despite better balance across classes, the model continued to face difficulties in accurately distinguishing between injury severities due to the minimal feature separation within the dataset.

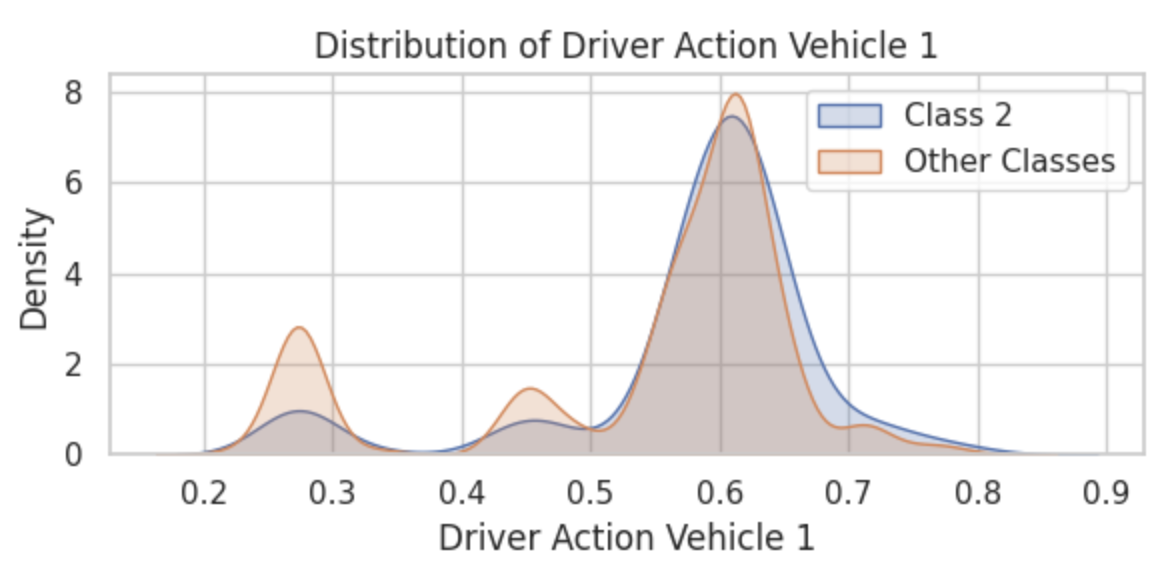
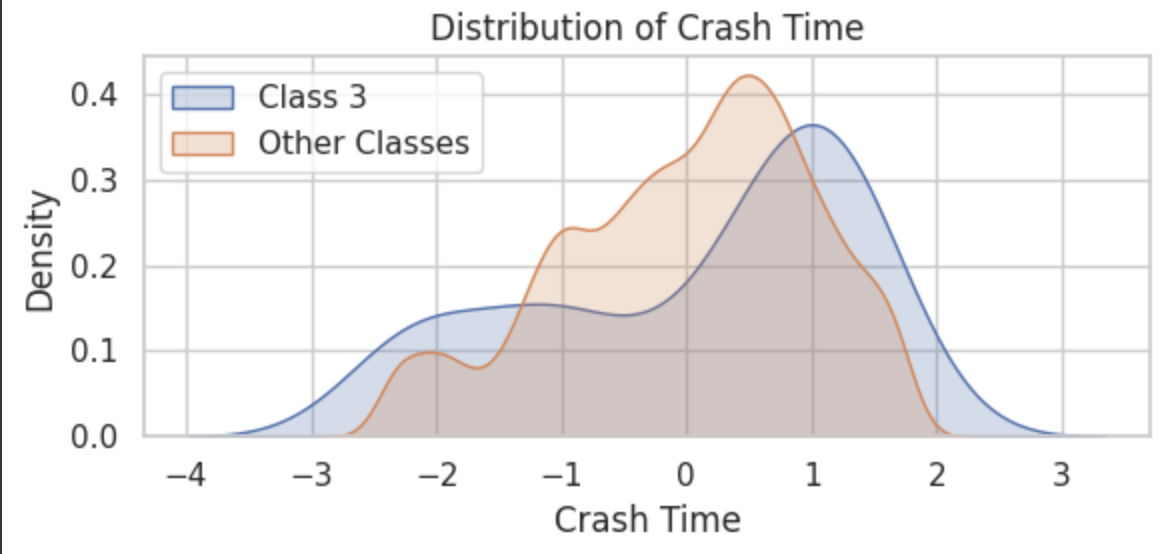
Due to ongoing challenges in accurately classifying severe injuries and fatalities, we shifted our focus to a class of algorithms that are inherently more robust to class imbalance: **boosting**. Boosting is an ensemble method that builds a sequence of weak learners, typically decision trees, where each new model attempts to correct the errors made by its predecessors. This sequential learning process enables the algorithm to steadily improve its performance, particularly on difficult-to-classify cases like minority classes.

Given that most of our features are categorical, we initially concentrated on **Categorical Boosting (CatBoost)**, which is specifically designed to handle categorical variables effectively. Our baseline CatBoost model achieved the strongest results to date, with an overall accuracy of 61.1% and a weighted F1-score of 0.605.

To further improve model performance, we conducted a grid search to tune key hyperparameters, including tree depth, regularization strength, number of iterations, and learning rate. For each grid search iteration, the model was trained using three-fold cross-validation, where it was trained on two-thirds of the data and validated on the remaining third. The optimal hyperparameters, based on the highest cross-validation accuracy, were found to be a tree depth of 8, 1000 iterations, a learning rate of 0.01, and a regularization parameter of 3. With these tuned settings, the model achieved a slight improvement, reaching an overall accuracy of 61.5% and a weighted F1-score of 0.615.

To explore whether ensembling multiple boosting algorithms could further boost performance, we briefly tested XGBoost and LightGBM. However, both algorithms performed worse in predicting Class 2 (severe injuries) and Class 3 (fatalities) compared to CatBoost. As a result, we determined that ensembling with these models would likely reduce the overall performance rather than enhance it, and we proceeded with CatBoost as our final model.

We’ve included the classification report (image to the right) for this best-performing model. As shown in the confusion matrix, Class 3 (fatalities) is predicted with moderate accuracy and recall. However, Class 2 (severe injuries) still suffers from a high number of false positives, resulting in lower precision. This indicates that the model continues to struggle in distinguishing severe injuries from other classes, suggesting a need for additional or more discriminative features.

To better understand how to distinguish Class 2 from the other classes, we created kernel density plots comparing the distributions of our predictors across different injury types. These plots revealed a critical limitation of our dataset: there is very little meaningful separation between major injuries (Class 2) and the other classes. For example, when examining the density plot for Driver Action, the distribution for Class 2 closely mirrors the distributions of the other classes, showing no significant distinction. This pattern was consistent across nearly all features we analyzed.

We repeated the same analysis for Class 3 to investigate why the model performed somewhat better in predicting fatalities. In contrast to Class 2, some features exhibited clearer differences for Class 3. For instance, the density plot for Crash Time showed that incidents resulting in fatalities tended to occur slightly later in the day compared to other classes. This subtle but observable pattern suggests that, while imperfect, there is at least some feature-level distinction for Class 3 that the model can leverage during training.

**Conclusion and Future Work**

Our results highlight that due to the clear class imbalance and minimal separation between features, our machine learning models struggled to predict specific injury outcomes with high confidence. While our goal was to differentiate between varying degrees of injury severity (minor, major, fatal), the limited feature-level distinctions made this task unreliable. Even after applying SMOTE to balance the dataset, although the model became more sensitive to minority classes, precision declined significantly, leading to inconclusive results. Among the models we tested, boosting methods—particularly CatBoost—performed the best, achieving the highest accuracy and weighted F1-score. However, even CatBoost continued to struggle with distinguishing major injuries (Class 2) from other outcomes, mainly due to the lack of meaningful feature separation. Influential features such as alcohol or drug involvement and driver behavior did emerge as important contributors to injury severity; however, the dataset lacked the granularity needed to drive stronger predictive performance.

In future work, simplifying the problem into a binary classification task (injury vs. no injury) may be more appropriate, given the minimal separation between the injury severity classes. Additionally, gathering richer data, such as detailed driver demographics, vehicle make and model, or more nuanced behavioral factors, could provide the stronger feature differentiation needed to better predict detailed injury outcomes. Using alternative or enhanced datasets could ultimately improve both model accuracy and reliability, especially for distinguishing between major and fatal injuries.

**References**

1. *FDOT SSOGis Traffic Accident Query Tool*. FDOT State Safety Office GIS. (n.d.). https://gis.fdot.gov/ssogis/
2. World Health Organization. (2023, December 13). *Road traffic injuries*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries