PREDICTING MOTORCYCLE CRASH INJURY SVERITY IN UTAH: UTILIZING MACHINE LEARNING METHODS

By

Isaac Baah

Laramie, Wyoming

December 2023

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1.0 Introduction

1.1 Background

Motorcycle use in the United States has increased significantly over the years. From 2002 to 2021, the number of motorcycle registrations has almost doubled, from 4.3 million to 8.6 million (Teoh, n.d.). However, motorcyclists, a part of vulnerable road users are highly exposed and have less protection in comparison with occupants of enclosed vehicles (Yannis et al., 2020, Constant & Lagarde, 2010). Moreover, vulnerable road users in 2021 alone accounted for 20 percent of all those who died in motor vehicle crashes in the United States, a 13 percent increase from 2020 (FHWA, 2017). The crash fatality rate for motorcyclists is nearly 5 times higher than for passenger car occupants and 8 times higher than that of light-truck occupant vehicles according to the National Highway Traffic Safety Administration (National Highway Traffic Safety Administration & U.S. Department of Transportation, 2021).

This project uses 11-years of historical motorcycle crash data (2010 – 2021) in Utah, including variables like driver characteristics (age group, sex, etc.), environmental conditions, vehicle contributing events, and crash circumstances. It is worth mentioning that three datasets were used in this study, crash level, vehicle level, and person level datasets. Various machine learning algorithms (tree and non-tree-based models) were employed and compared to determine the most efficient model for predicting the severity of injuries.

This project aims to (1). Predict the injury severity of a motorcycle driver involved in a crash.

(2). Identify the crash characteristics of fatal and severe motorcycle crashes in Wyoming and Utah.

2.0 Data and Methods

2.1 Data Collection and Pre-Processing

This project used 11 years (2010 – 2021) of historical crash data in Utah. The crash level database contained crash information such as crash location, weather and roadway conditions, first harmful event (FHE), FHE location, etc. The vehicle level database contained information related to each vehicle such as vehicle contributing circumstances, etc. The person-level database contained information about each person involved in the crash, such as the sex of the driver, age, etc.

The crash injury severity levels reported are grouped into five (5) categories; fatal injury (K), suspected serious injury (A), suspected minor injury (B), possible injury (C), and no injury/property damage only (O). For this project, the crash injury severity levels were grouped into binary (1 or 0) response variables. Fatal (K) and suspected serious injuries (A) were represented as 1 – Fatal Injury (KA). All other crash injury severity levels were also aggregated into a single category - Non-Fatal Injury (BCO).

2.2 Feature Selection

As part of the preprocessing stage, the recursive feature elimination (RFE) with a cross-validation algorithm was implemented to select the independent features that are relevant in predicting motorcycle crash injury severity. Thirteen (13) features were selected for the crash-level dataset, nine (9) features were selected for the vehicle-level dataset, and seven (7) features were selected for the person-level dataset. The "*feature_selection.RFECV*" is available in Scikit-Learn. RFECV was instantiated and used to implement the RFE algorithm. The cross-validation method used in the process is "*StratifiedKFold*" and the algorithm that was also used is the Adaptive Gradient Boosting (AdaBoost).

2.3 Method

Several tree-based (RF - Random Forest, XGBoost - Extreme Gradient Boosting, AdaBoost - Adaptive Gradient Boosting, GradBoost - Gradient Boosting, and LightGBM - Light-Gradient Boosting Machines) and non-tree based (SVM - Support Vector Machines, LR – Logistic Regression) algorithms were applied and compared using classification metrics.

The datasets were then split into training and testing or hold-out datasets. The training dataset was then scaled, specifically*, StandardScaler()* was instantiated and *fit_transform* on the "X_train" data and transformed only on the "X_test" data. To train the model and make a prediction on the training and testing datasets, each algorithm was instantiated. For example, by making a prediction using the Random Forest, the *RandomForestClassifier()* was instantiated. A "*.fit*" was called on the training datasets for the dependent and independent variables. A "*.predict*" was then called on the test data to make a prediction.

When initially fitting the machine learning algorithms, the default hyper-parameters were used. To select the best combination of hyper-parameters to improve the model performance and to reduce the human effort required in optimization, a decision-theoretic approach (Randomized Search) for hyper-parameter optimization (HPO) was adopted, using a 10-fold cross-validation. Random Search (RS) hyper-parameter optimization is a method that combines hyper-parameters in the search space independently in a random manner (Yang & Shami, 2020).

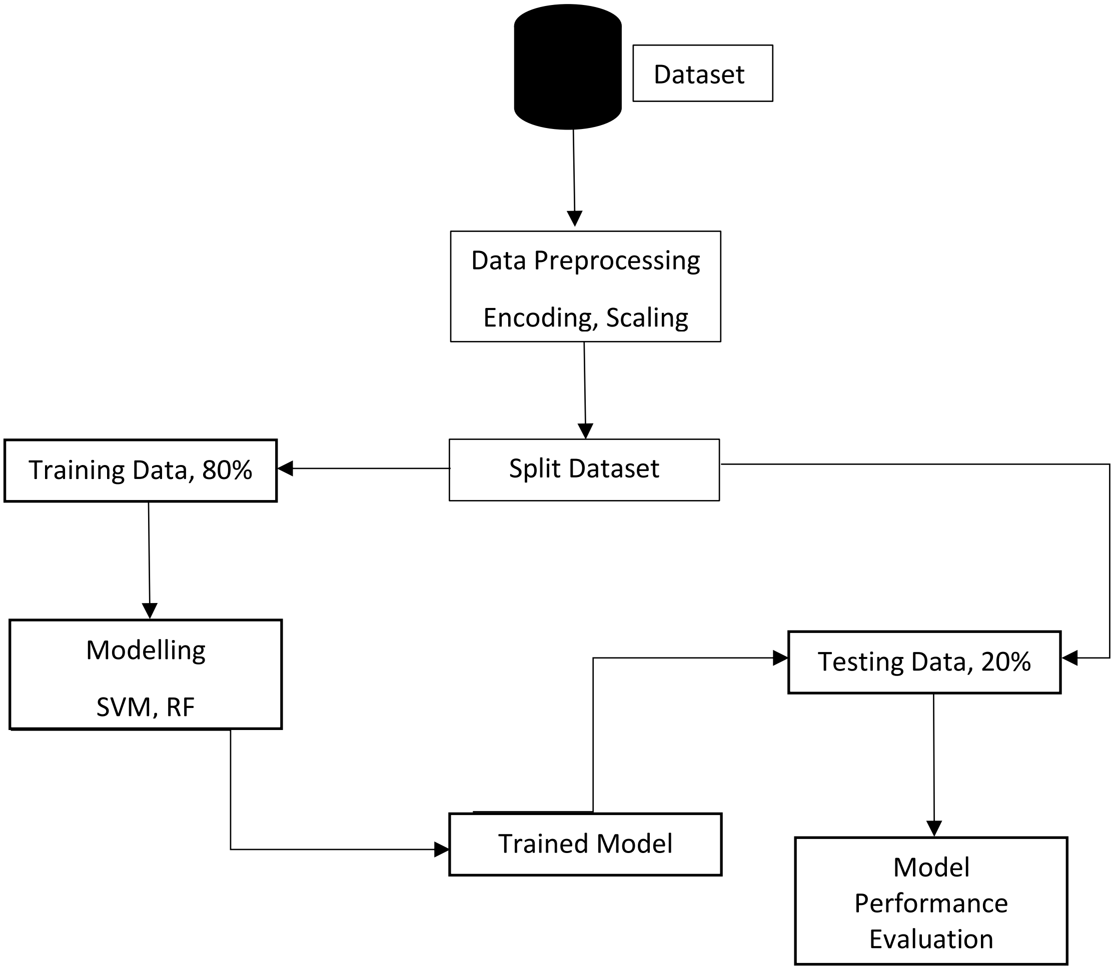


Figure 1. Machine Learning Process Adopted

2.4 Model Performance Evaluation

In this study, we evaluate the performances of these machine-learning models using accuracy, precision, and ROC-AUC.

Accuracy measures the ratio of correctly predicted instances to total instances.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Precision measures the ratio of correctly predicted positive class (Fatal) to all the predicted positives.

$$Precision = \frac{TP}{TP+FP}$$

The Receiver Operating Characteristic (ROC) curve is a non-threshold evaluation metric that visualizes the performance of a classifier by plotting the true positive rate against the false positive rate at various thresholds. The Area under the Curve (AUC) quantifies this performance, with a value of 1 representing a perfect classifier.

2.5 Results and Discussions

2.5.1 Classification Results for the ML Models on the Datasets

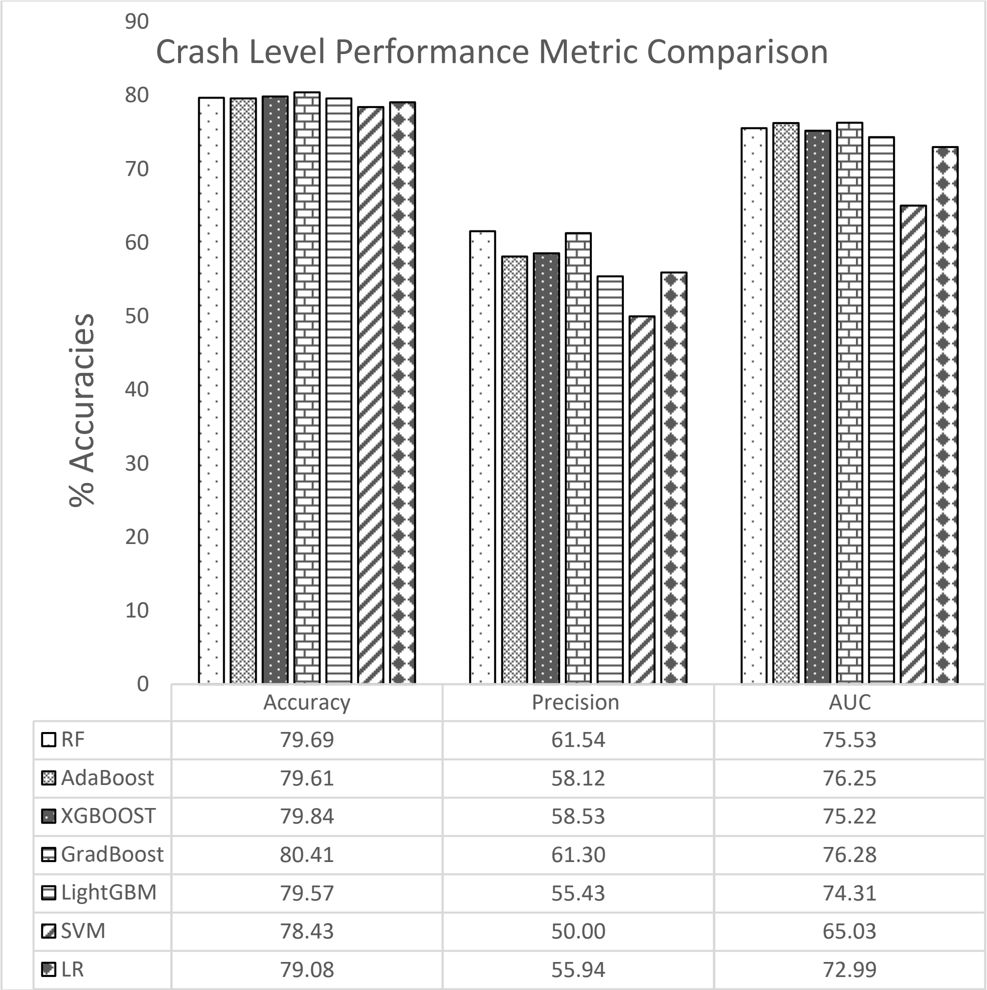
This section presents the ML performances on the three datasets. Table 1 below shows the summary performance of the seven (7) ML models.

Table 1. Performance of Models for Binary Class Injury Severity Prediction on Training and Test Datasets

Level	Model	Training Data			Test Data		
		Accuracy	Precision	AUC	Accuracy	Precision	AUC
Crash	RF	80.85	75.79	81.68	79.69	61.54	75.53
	AdaBoost	79.72	59.97	76.36	79.61	58.12	76.35
	XGBOOST	82.44	75.90	81.72	79.84	58.53	75.22
	GradBoost	80.73	64.74	78.73	80.41	61.30	76.28
	LightGBM	82.42	71.75	82.13	79.57	55.43	74.31
	SVM	78.81	89.87	78.73	78.43	50.00	65.03
	LR	78.97	57.19	72.24	79.08	55.94	72.99
Vehicle	RF	79.76	87.11	73.88	77.88	54.55	66.11
	AdaBoost	78.87	55.03	66.17	77.65	41.86	65.38
	XGBOOST	79.86	85.00	66.48	77.86	60.00	65.14
	GradBoost	79.43	80.00	68.19	77.60	41.18	66.04
	LightGBM	79.93	64.86	61.59	77.37	44.17	59.61
	SVM	80.08	86.56	66.12	77.51	37.74	56.97
	LR	78.69	33.33	61.36	77.69	33.33	60.73
Person	RF	78.95	78.36	69.52	78.96	66.67	69.43
	AdaBoost	78.39	62.67	67.71	79.24	67.31	68.88

XGBOOST	79.11	67.40	69.68	79.12	62.89	69.37
GradBoost	79.32	67.77	69.78	79.18	61.86	69.95
LightGBM	79.33	67.77	69.20	79.20	62.96	69.32
SVM	78.98	78.36	58.61	78.66	66.67	56.36
LR	78.21	64.29	63.52	78.70	66.19	63.83

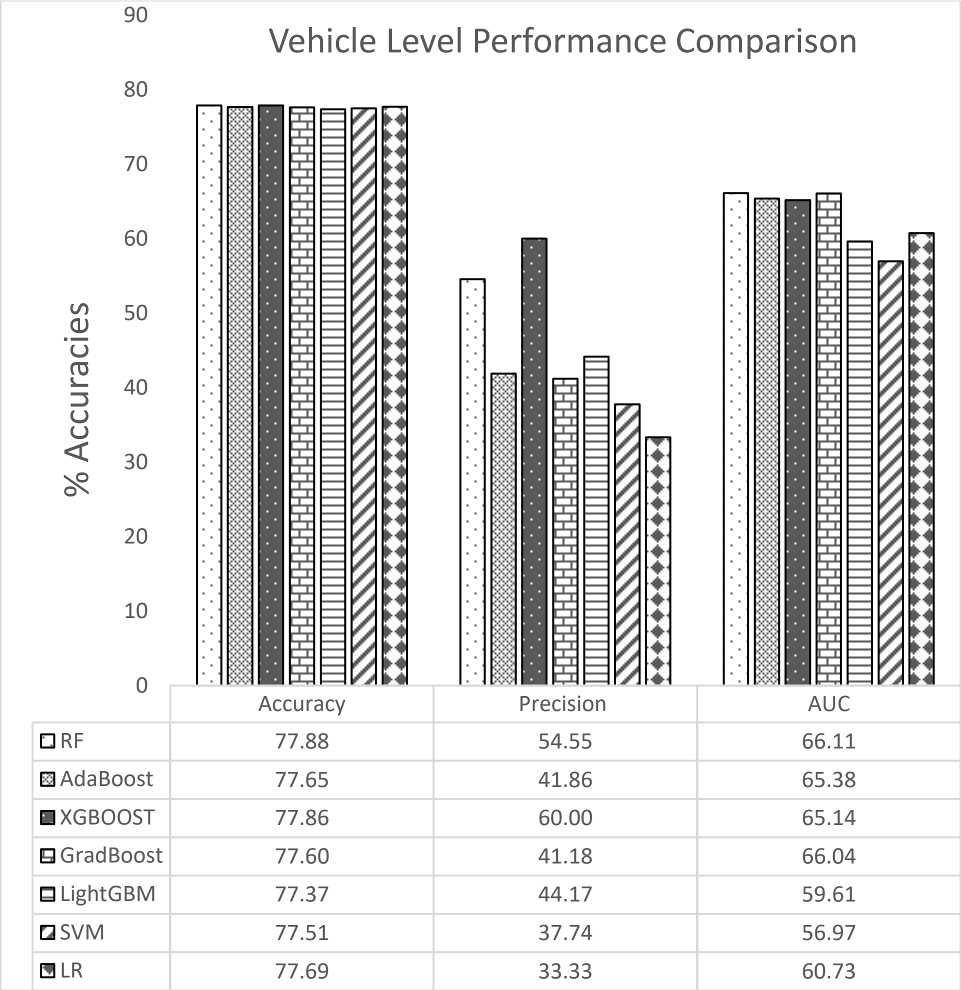
Figure 2 below shows the performance comparison for the test set using the three classification metrics on the crash dataset. The highest accuracy and area under the curve (AUC) were achieved by the gradient-boosting algorithm on the test dataset, 80.41% and 76.28%, respectively. The precision of the gradient boosting was 61.30%, however, the highest precision was achieved by the random forest model, 61.54%. Comparing the ensemble machine learning methods to the non-tree-based models, the ensemble methods performed slightly well for the test data when compared to the non-tree-based models on both accuracy and AUC. Apart from the light gradient boosting machine, the ensemble methods also performed better in terms of precision compared to SVM and LR.



	Accuracy	Precision	AUC
□ RF	79.69	61.54	75.53
▩ AdaBoost	79.61	58.12	76.25
■ XGBOOST	79.84	58.53	75.22
▤ GradBoost	80.41	61.30	76.28
▤ LightGBM	79.57	55.43	74.31
▨ SVM	78.43	50.00	65.03
▩ LR	79.08	55.94	72.99

Figure 2. Comparison of Different ML Algorithms – Crash Level

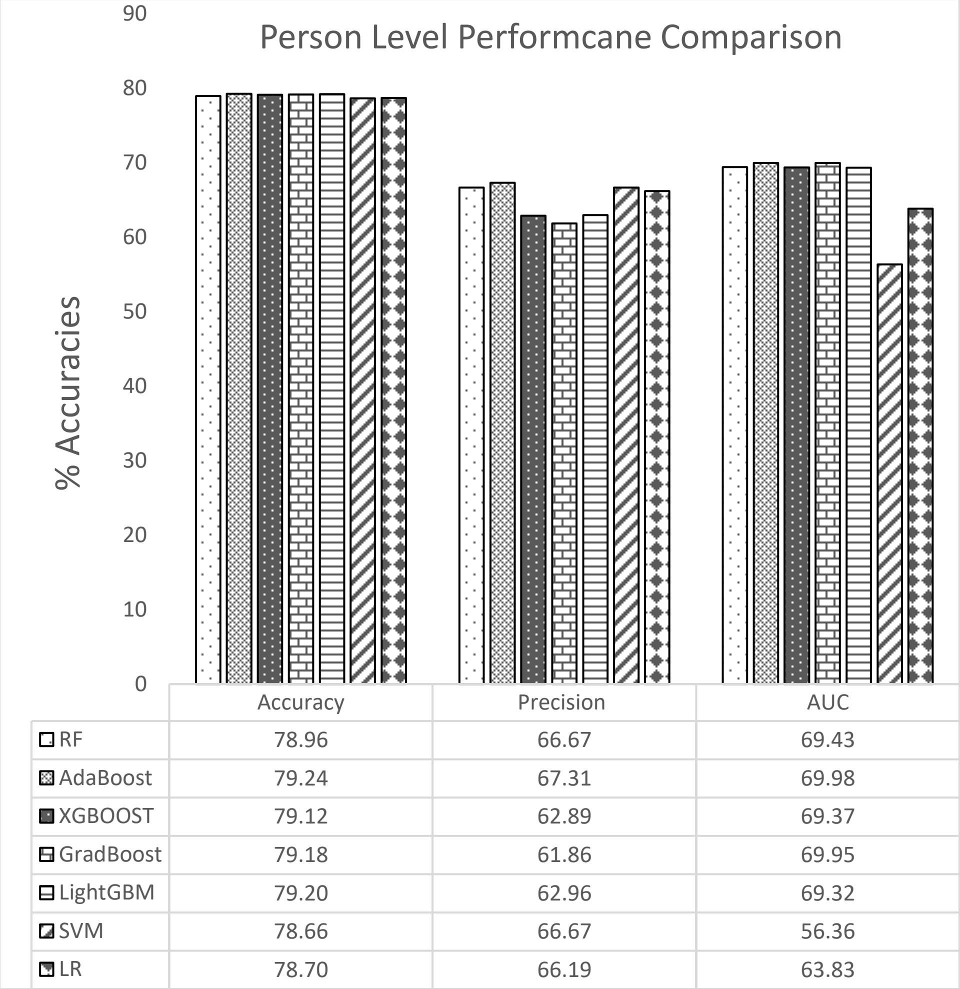
On the vehicle-level dataset, it was found that the RF classifier slightly outperformed all the other models in terms of accuracy and AUC. The highest precision was achieved by the extreme gradient boosting model, 60%, as seen in Figure 3 (XGBoost has the tallest bar). The random forest classifier on the other hand achieved a precision value of 54.55%. All the ensemble methods, except for the light gradient boosting machine algorithm were found to outperform the SVM and LR models. The lowest accuracy for the test set was found for the light gradient boosting machine and the lowest AUC was found for the support vector machine (SVM). Figure 3 below shows the performance comparison of the test set using the three classification metrics on the vehicle-level dataset.



	Accuracy	Precision	AUC
□ RF	77.88	54.55	66.11
▩ AdaBoost	77.65	41.86	65.38
■ XGBOOST	77.86	60.00	65.14
▤ GradBoost	77.60	41.18	66.04
▤ LightGBM	77.37	44.17	59.61
▨ SVM	77.51	37.74	56.97
◆ LR	77.69	33.33	60.73

Figure 3. Comparison of Different ML Algorithms – Vehicle Level

Figure 4 below shows the performance comparison for the test set using the three classification metrics on the person-level dataset. Similar to the result on the crash level, the gradient boosting classifier showed good performance for the test data. It had an accuracy of 79.18% and highest AUC value of 69.95%. However, the adaptive gradient boosting classifier was found to have the highest precision when compared to the other models. It was found that the RF classifier slightly outperformed all the other models in terms of accuracy and AUC. When comparing the tree-based and non-tree-based models, the tree-based models were found to be slightly superior to the non-tree-based models in terms of both accuracy and AUC. For precision, gradient boosting was found to have the lowest value. Figure 4 below shows the performance comparison of the test set using the three classification metrics on the vehicle-level dataset.



	Accuracy	Precision	AUC
□ RF	78.96	66.67	69.43
▩ AdaBoost	79.24	67.31	69.98
■ XGBOOST	79.12	62.89	69.37
⊟ GradBoost	79.18	61.86	69.95
⊟ LightGBM	79.20	62.96	69.32
◪ SVM	78.66	66.67	56.36
◩ LR	78.70	66.19	63.83

Figure 4. Comparison of Different ML Algorithms – Person Level

2.6 Discussions

2.6.1 Feature Importance for the Crash Level

Figure 5 below shows the relative feature importance ranking of the predictors based on the gradient-boosting classifier. As can be seen from the figure below, the driver injury area is the most dominant variable in predicting the crash injury severity, with a feature importance value of 0.47. The next important variable influencing crash injury prediction is the manner of collision having a feature importance value of 0.08, followed closely by whether the driver was riding under the influence (drugs and alcohol), having a feature importance value of 0.06. The maneuver of the vehicle, speeding, collision with a fixed object, and the location of the crash were some factors associated with motorcycle injury severity prediction.

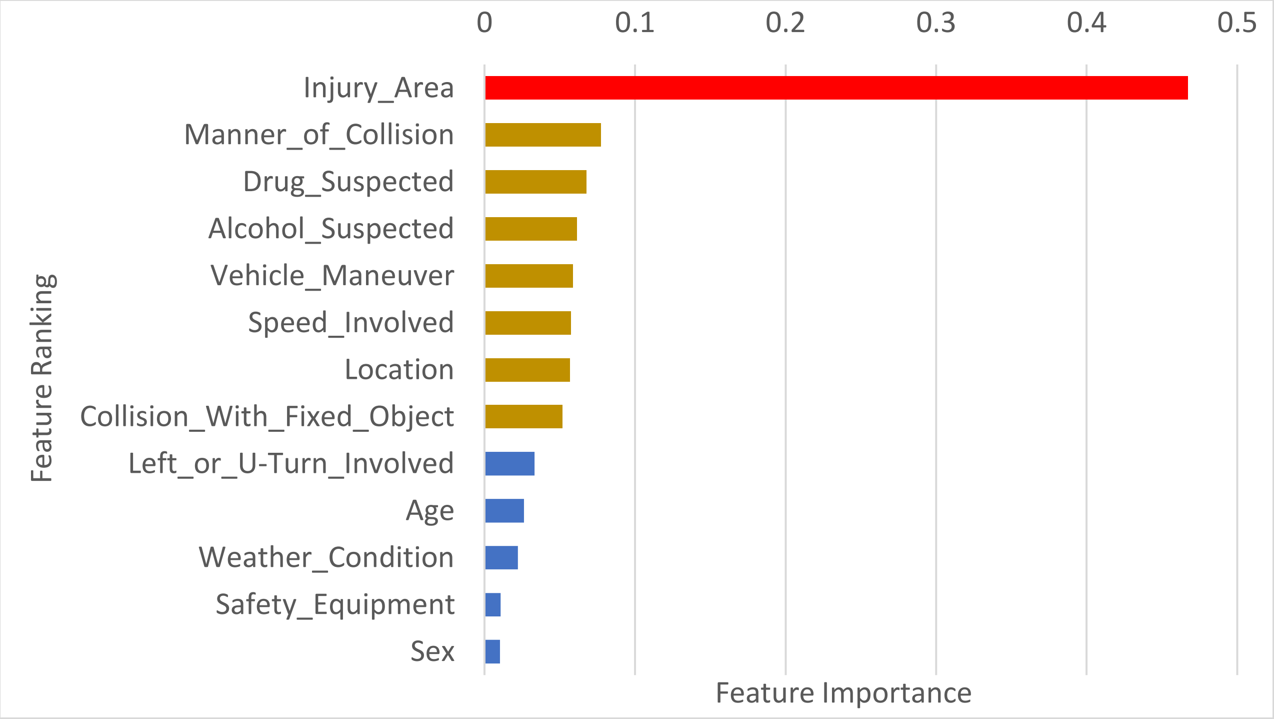


Figure 5. Variable Importance for the Crash Level

2.6.2 Feature Importance for the Vehicle Level

Figure 6 below shows the global relative feature importance ranking of the predictors based on the random forest classifier. From Figure 6 below, the variable that was found to have a high impact in determining the severity of motorcycle crashes is vehicle maneuvers. Vehicle maneuvers, the first harmful event, vertical alignment, and the direction of travel followed each other closely in predicting crash injury severity. The study also found that vehicle collisions with fixed objects, horizontal alignment, and whether overturn rollover was involved were found to contribute to crash injury severity prediction. The variable that was found to have minimal influence on motorcycle crash injury severity is the number of vehicles involved in the crash.

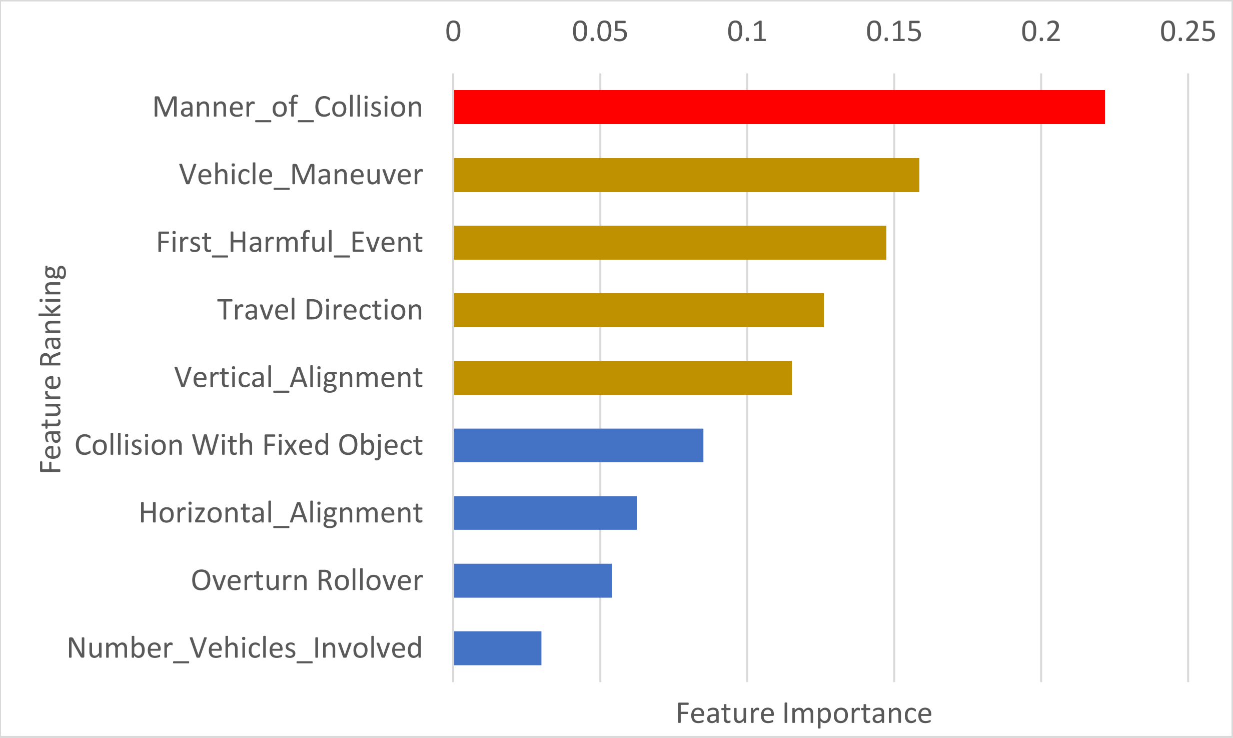


Figure 6. Variable Importance for the Vehicle Level

2.6.3 Feature Importance for the Person Level

An analysis of the result found that the injury area of the person involved in the crash can be important in crash injury severity prediction. The injury area can be any part of the body such as the head, neck, shoulder, lower extremities, etc. The driver's action was also found to contribute to the severity of motorcycle injury severity as shown in Figure 7. It is followed closely by driving under the influence of drugs or alcohol. The last three variables found to influence crash severity prediction for the person-level datasets are safety equipment use, age, and wrong-way driving.

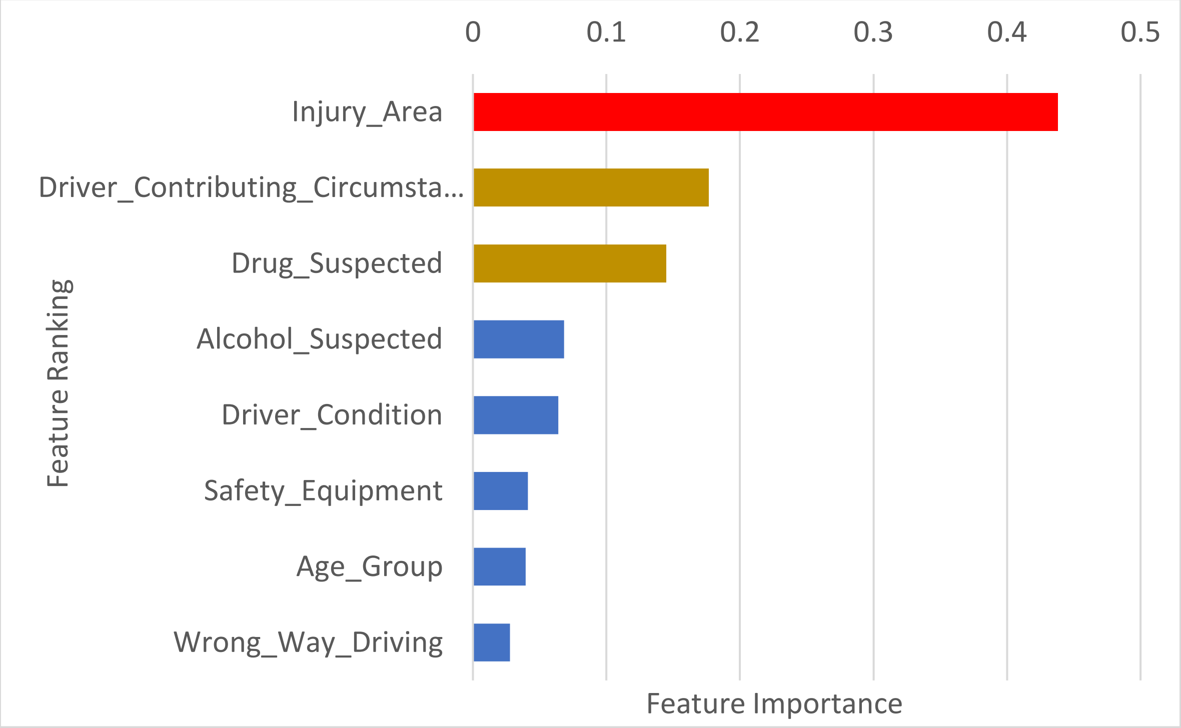


Figure 7. Variable Importance for the Person Level

2.7 Conclusion

Given that the percentage of all motorcycle deaths as a percent of all motor vehicle deaths in Utah is among the highest in the United States, we must identify and understand the contributing factors to crash injury severities. The objectives of this study are to predict motorcycle crash injury severity as well as to identify motorcycle crash contributing factors.

By applying five tree-based and two non-tree-based machine learning, the feature importance plot of each dataset (that is, crash, vehicle, and person levels) based on the best-performing model for each dataset was used to assess influential factors of motorcycle crash injury severity prediction. Feature importance analysis for the crash level data showed that the injury area of a driver, the manner of collision, driving under the influence of drugs and alcohol, vehicle maneuver, speeding, the location of the crash, and collision with a fixed object are the most sensitive features in predicting motorcycle crash injury severity outcome. The vehicle level variable importance analysis found that the manner of collision, vehicle maneuver, first harmful event, the travel direction of the driver, and the vertical alignment of travel were the top five contributing features of crash injury severity outcome prediction. The person-level feature importance analysis revealed that the injury area of the driver, driver actions, and driving under the influence of drugs are influential factors for predicting crash injury severity.

3.0 References

1. Constant, A., & Lagarde, E. (2010). Protecting vulnerable road users from injury. PLoS Medicine, 7(3), 1–4. https://doi.org/10.1371/journal.pmed.1000228.
2. Federal Highway Administration (FHWA). *State Motor-Vehicle Registrations – 2017*. https://www.fhwa.dot.gov/policyinformation/statistics/2017/mv1.cfm. [Accessed December 5, 2023].
3. National Highway Traffic Safety Administration, & U.S. Department of Transportation. (2021). Traffic Safety Facts 2021 Data: Motorcycles. 2023(June).
4. Teoh, E.R. (2021) Motorcycles Registered in the United States, 2002 -2021. IIHS. https://www.iihs.org/api/datastoredocument/bibliography/2225.
5. Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061.
6. Yannis, G., Nikolaou, D., Laiou, A., Stürmer, Y. A., Buttler, I., & Jankowska-Karpa, D. (2020). Vulnerable road users: Cross-cultural perspectives on performance and attitudes. *IATSS Research*, *44*(3), 220–229. https://doi.org/10.1016/j.iatssr.2020.08.006.