Accident Severity Prediction Project Report

Elahe Talaie

October 2020

1. Introduction

Accident analysis and prediction have become a significant field of research considering its impact on public safety. There were 1.35 million road traffic deaths globally in 2016, according to WHO.[1] Understanding the most relevant factors contributing to road accidents is of utmost importance to improve road safety. Machine-learning modeling is a powerful and popular approach to accident analysis and prediction. The objective of this project is to develop classification models to accurately predict the severity of an accident based on given conditions.

2. Data Description

The dataset used in this project is offered by the Seattle Department of Transportation (SDOT) Traffic Management Division, Traffic Records Group. The .csv file is obtained from the website of Coursera. The dataset contains data on more than 194000 accidents recorded in Seattle, and each accident has up to 38 attributes, such as weather conditions, light condition, collision time and date, collision type, whether the driver was under influence of drugs or alcohol, etc. One attribute is accident severity code, 1 for property damage, and 2 for injuries. The accident severity is the target of prediction in this project. 70% of the data have a severity code of 1.

3. Methodology

Programming Language and Libraries

All implementations in this project are in Python. Libraries, such as Pandas, Matplotlib, scikit-learn were used.

Pre-processing

More than 20 irrelevant or redundant features, such as different codes defined by SDOT and given to each accident, accident geographical coordination, description notes on the accidents were dropped. Faceting histograms by subsets of data were plotted to visualize how each feature affects the severity code. Some of these plots are shown in the results section. The correlation matrix of different features was plotted as a heatmap and some dependent features were dropped from the features set. For example, the plot showed a strong correlation between rainy weather condition and wet road condition, as expected. One of the features for each pair with a strong correlation was dropped.

[1] https://www.who.int/gho/road_safety/mortality/en/

The categorical variables were converted to binary variables using One-Hot Encoding. The data was split to the train test (70%) and test set (30%).

Negative Sampling

Because the dataset has more frequency of accidents with severity code 1 than those with severity code 2, negative sampling was implemented on the training set to balance the frequency of samples and investigate the effect of the data balance on the performance of classification algorithms used in this study.

Classification Algorithms

In this project, four different classification algorithms were used, including K-Nearest Neighbors (KNN), Decision Tree (DT), Support Machine Vector (SVM), and Logistic Regression (LR). The optimum number of nearest neighbors for KNN was found using K-fold cross-validation.

Evaluation

The performance of prediction for each algorithm was evaluated by calculating its confusion matrix and F1_score on the test set.

4. Results and Discussion

Data Exploring

The list of initial features selected to explore are as follows: address-type, collision type, number of people, bicycles, pedestrians, and vehicle involved in the collision, accident time, whether or not the collision was due to inattention, whether or not a driver involved was under the influence of drugs or alcohol, weather condition, road condition, light condition during the collision, whether or not the pedestrian right of way was not granted, whether or not speeding was a factor in the collision, and whether or not the collision involved hitting a parked car.

Figures 1-4 show the distribution of accident severity over some of the above factors. The histograms in Figure 1 and Figure 2 clearly show that when a bicycle or pedestrian was involved in the collision, or when the driver was under the influence of drugs or alcohol, or when speeding was a factor in the collision, the ratio of accidents with injuries was higher, as intuitively expected. Figure 3 and Figure 4 show that in some specific collision types such as angles, rear-ended, and left turn, and in collisions happened at an intersection the ratio of accidents resulting in injuries was higher.

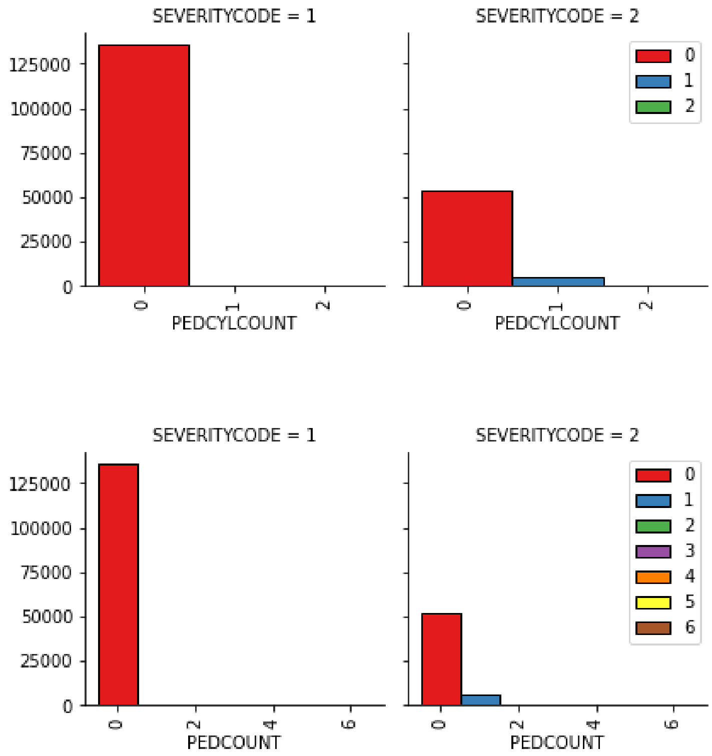


Figure 1. Distribution of accident severity over the number of bicycles (upper graph) and pedestrians (lower graph) involved in the collision.

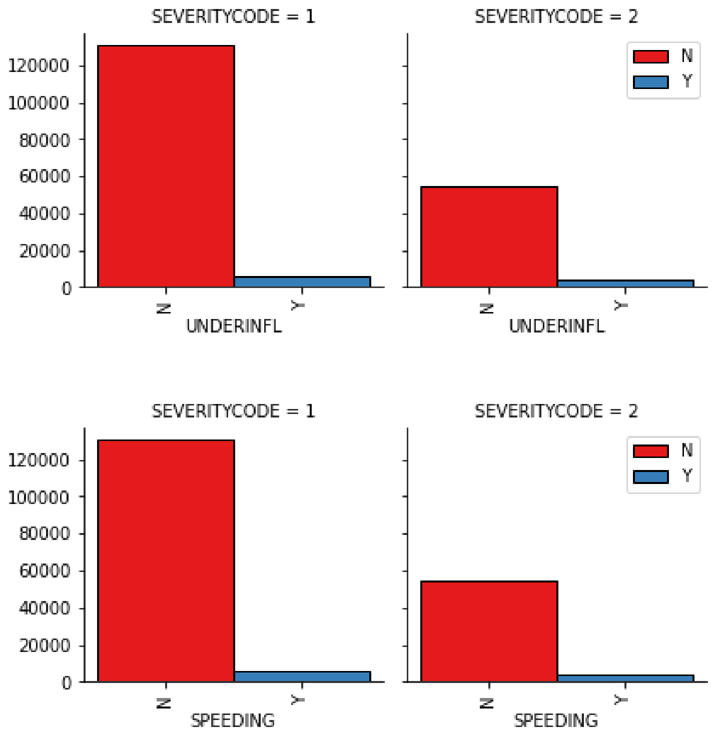


Figure 2. Distribution of accident severity over the conditions that if a driver involved was under the influence of drugs or alcohol (upper graph) and if speeding was a factor in the collision (lower graph).

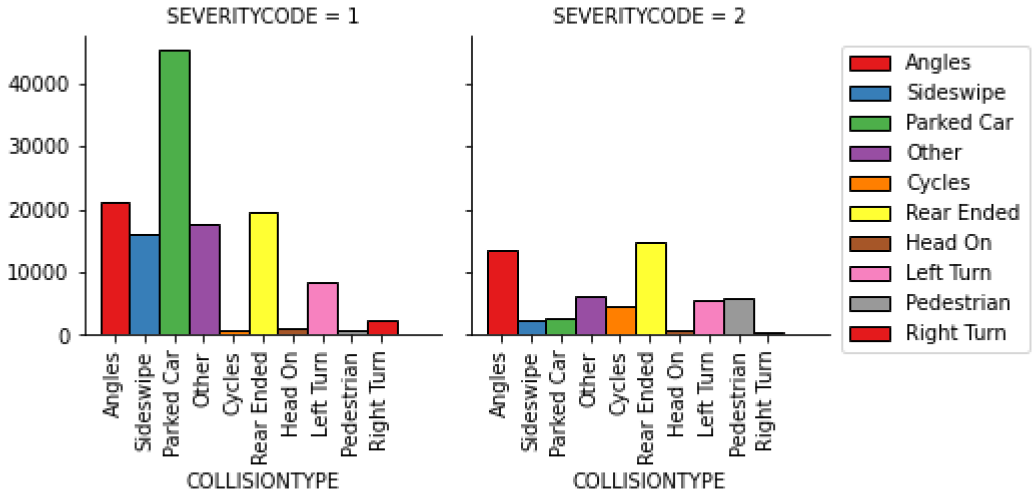


Figure 3. Distribution of accident severity over the collision type.

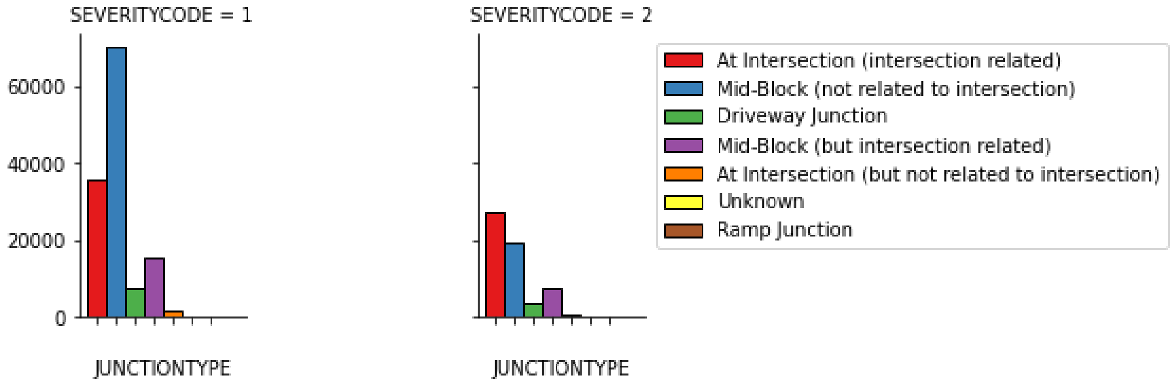


Figure 4. Distribution of accident severity over the junction type.

Figure 5 shows that the distribution of the severity ratio (defined here as the number of accidents with injuries over the number of accidents with property damage) over different hours of the day are not similar. Therefore, the time of the accident was mapped into 3 different bins. However, the day of the week on which the accident occurred was dropped from the feature set considering the similar severity ratio over all of them (see Figure 5, lower graph).

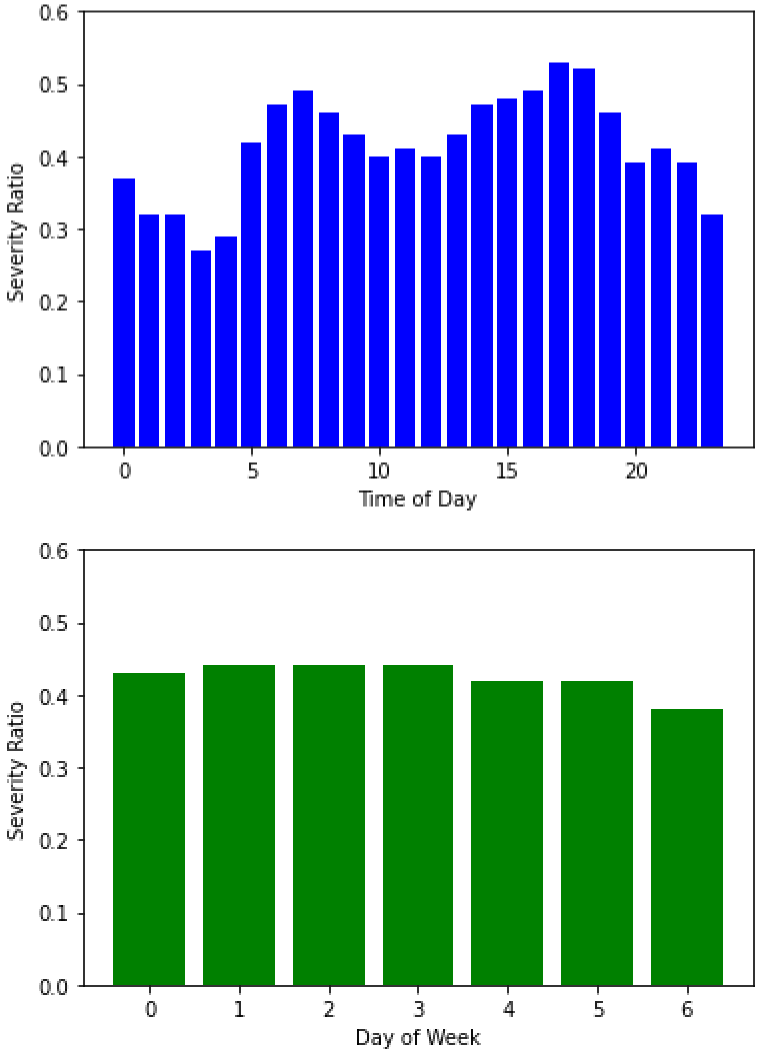


Figure 5. Distribution of severity ratio (defined here as number of accidents with injuries over the number of accidents with property damage) over the time of day (upper graph) and the day of week (lower graph) that the accident happed.

The heatmap in Figure 6 demonstrates the correlation between different features related to weather conditions, road conditions, and light conditions. Features ‘Dry’, ‘Wet’, 'Snow/Slush', and ' Dark - Street Lights On' were dropped because of strong correlations with some other features.

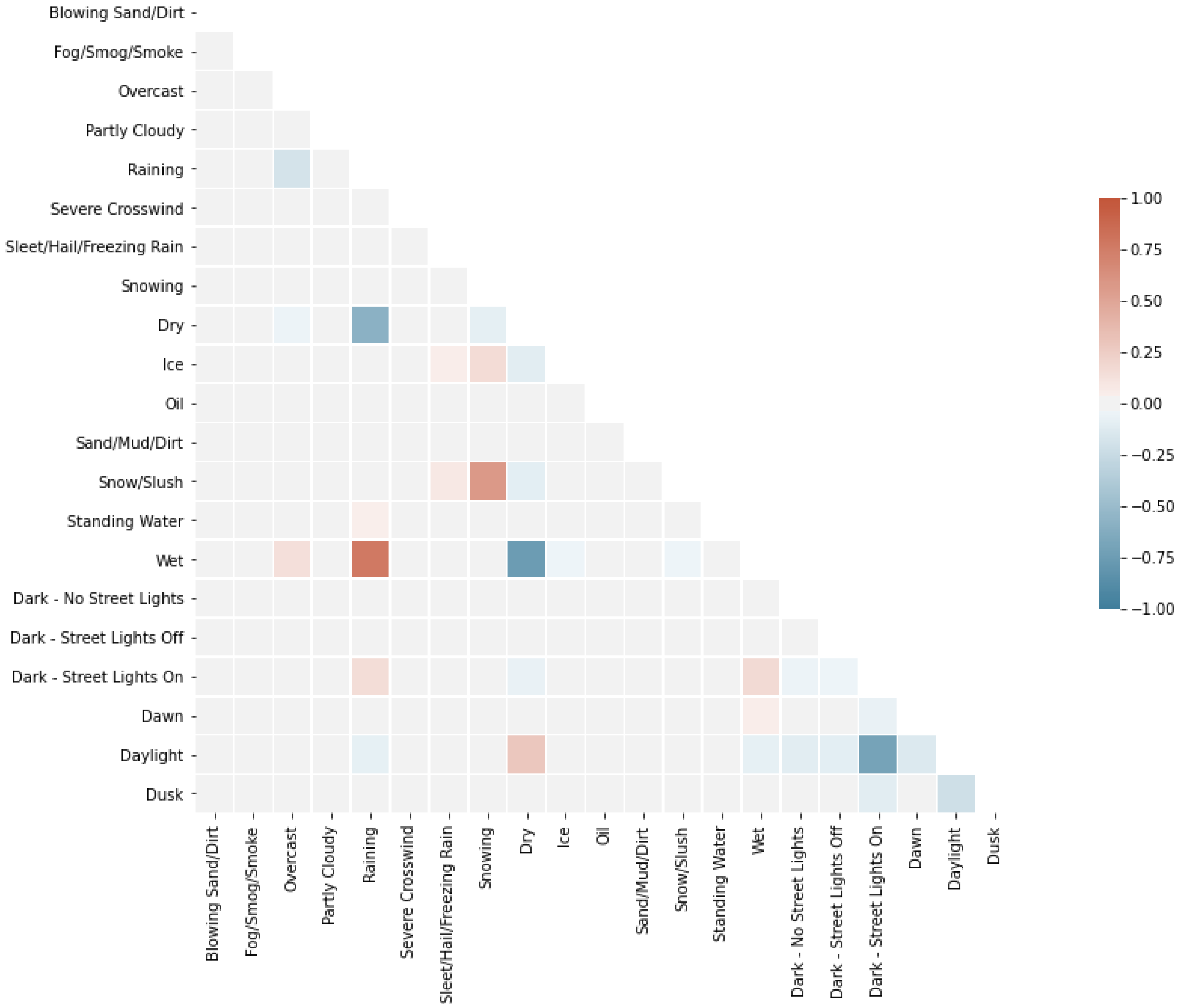
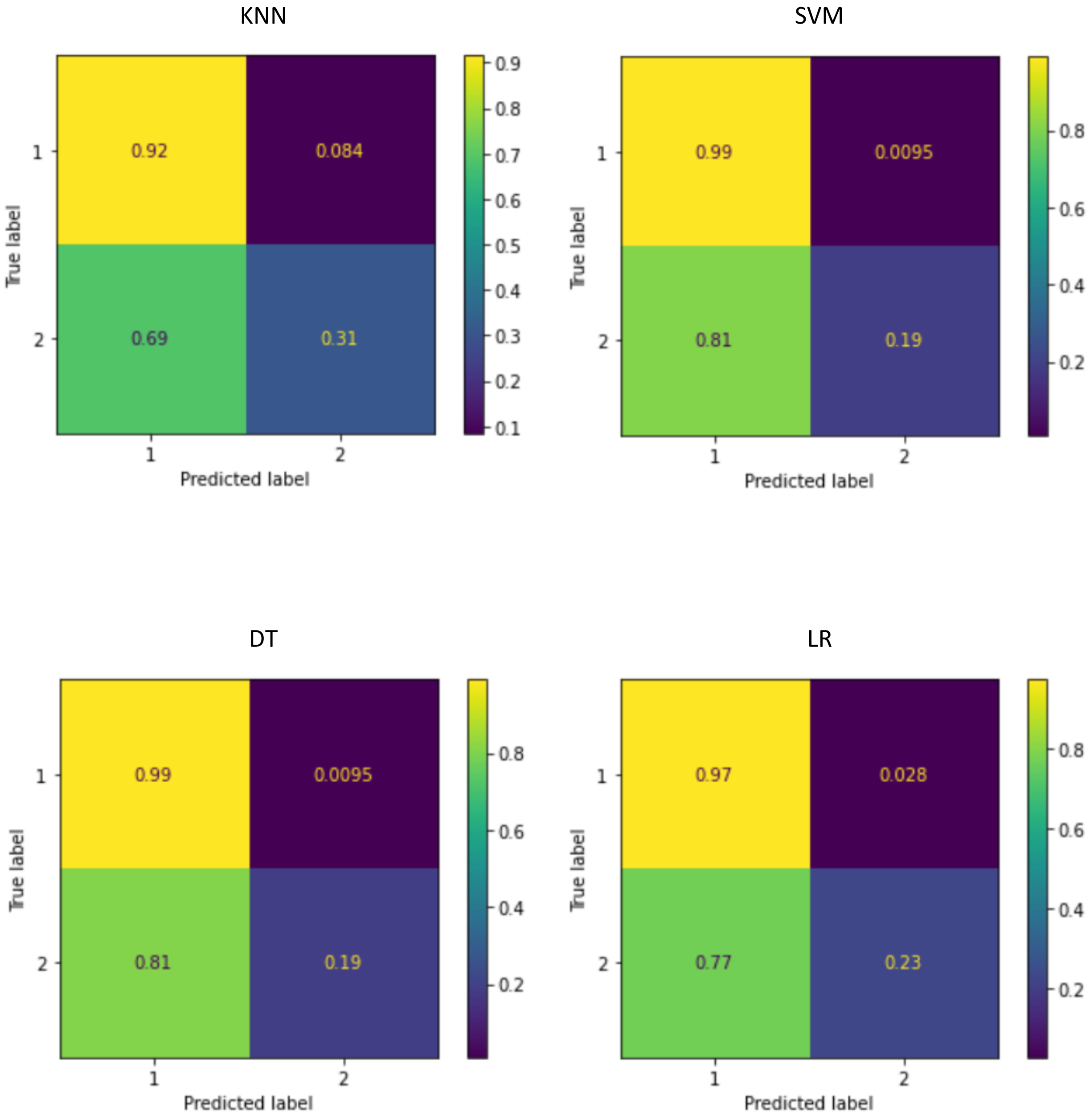


Figure 6. Heatmap demonstrating the correlation of various features related to the weather condition, road condition, and light condition during the collision.

Classification

Figure 7 and Figure 8 show the Confusion matrix, normalized to true test values, of the four classifiers trained on the imbalanced (original) training set and the balanced training set, respectively. Table 1 shows the regarding F1-scores for the class of Code 1 and the class of Code 2 and weighed average.

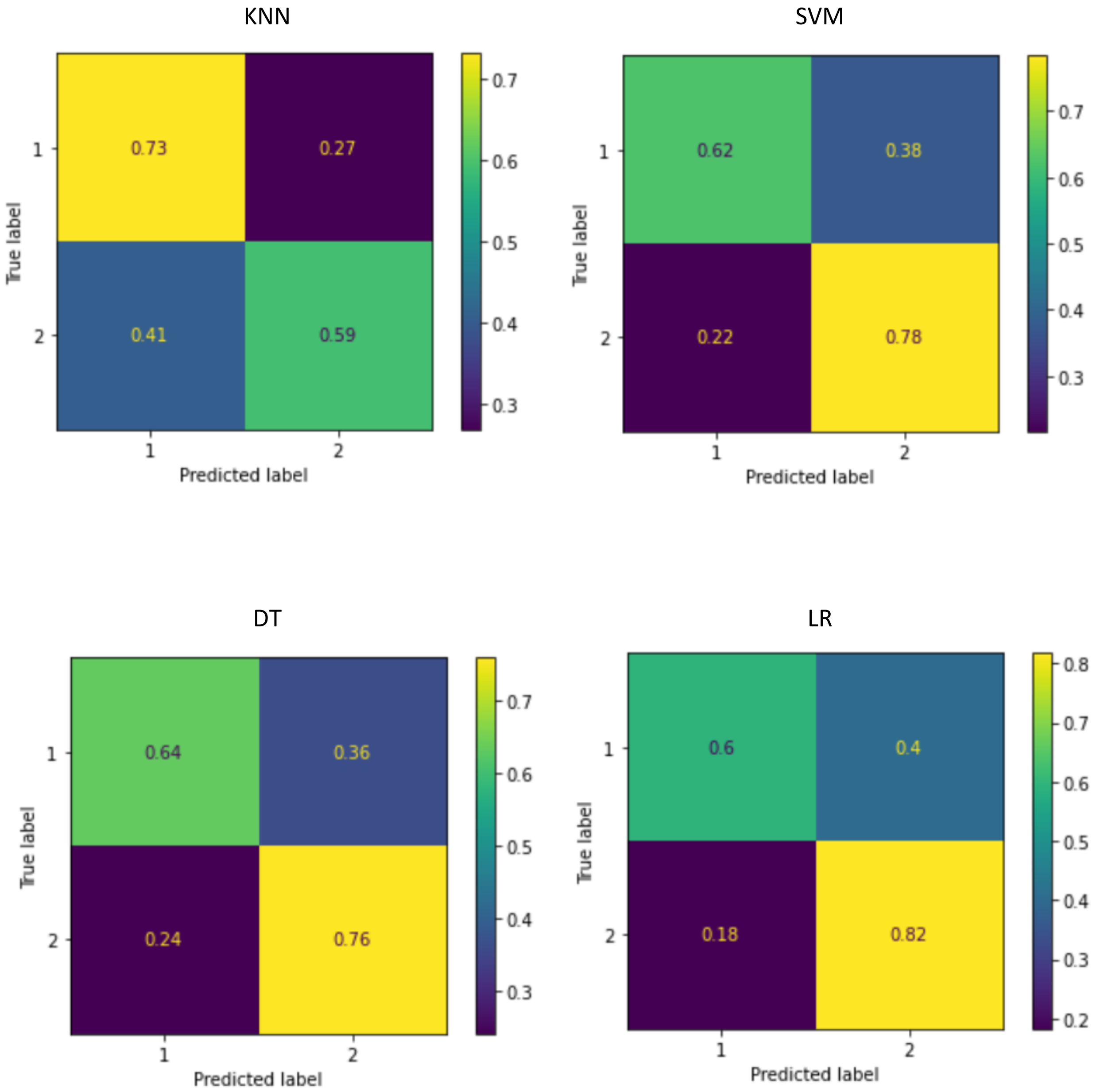

KNN

SVM

DT

LR

Figure 7. Confusion matrix, normalized to true test values, of the four classifiers trained on the **imbalanced** (original) training set.



KNN

SVM

DT

LR

Figure 8. Confusion matrix, normalized to true test values, of the four classifiers trained on the **balanced** training set.

Table 1. Prediction performance evaluation based on F1-score for the class of Code 1, Code 2, and weighted average (W-avg).

	KNN			**SVM**			**DT**			**LR**		
	Code 1	**Code 2**	**W-Avg**	**Code 1**	**Code 2**	**W-Avg**	**Code 1**	**Code 2**	**W-Avg**	**Code 1**	**Code 2**	**W-Avg**
Trained on Original Data	0.83	0.41	0.70	0.85	0.31	0.69	0.85	0.31	0.69	0.84	0.36	0.70
Trained on Balanced Data	0.77	0.53	0.70	0.73	0.59	0.69	0.73	0.58	0.69	0.72	0.59	0.68

The results show that when the classifiers were trained on the imbalanced training test, their performances on the prediction of the majority class (Code 1) were excellent; the support vector machine and decision tree algorithms classified 99% of the majority class instances into the correct one. However, the performances of all four classifiers were poor predicting the minority class. When trained on the balanced training set, their performance on predicting the minority class improved significantly, but their scores on the majority class were decreased.

5. Conclusions

The objective of this project was to develop classification models to accurately predict the severity of an accident based on given conditions. Accident analysis and prediction are critical for improving road safety. The various factors affecting the accident severity were explored through the data. Four different classifiers were examined. In this dataset, the number of accident instances with property damage was more than twice higher as the number of accidents with injuries. The classifiers were trained on both the imbalanced (original) training set and a balanced training set gained through negative sampling. The results show that the performance of the classifier on the prediction of the minority class into the correct one improved significantly when trained on the balanced training set.