Accident Severity Prediction

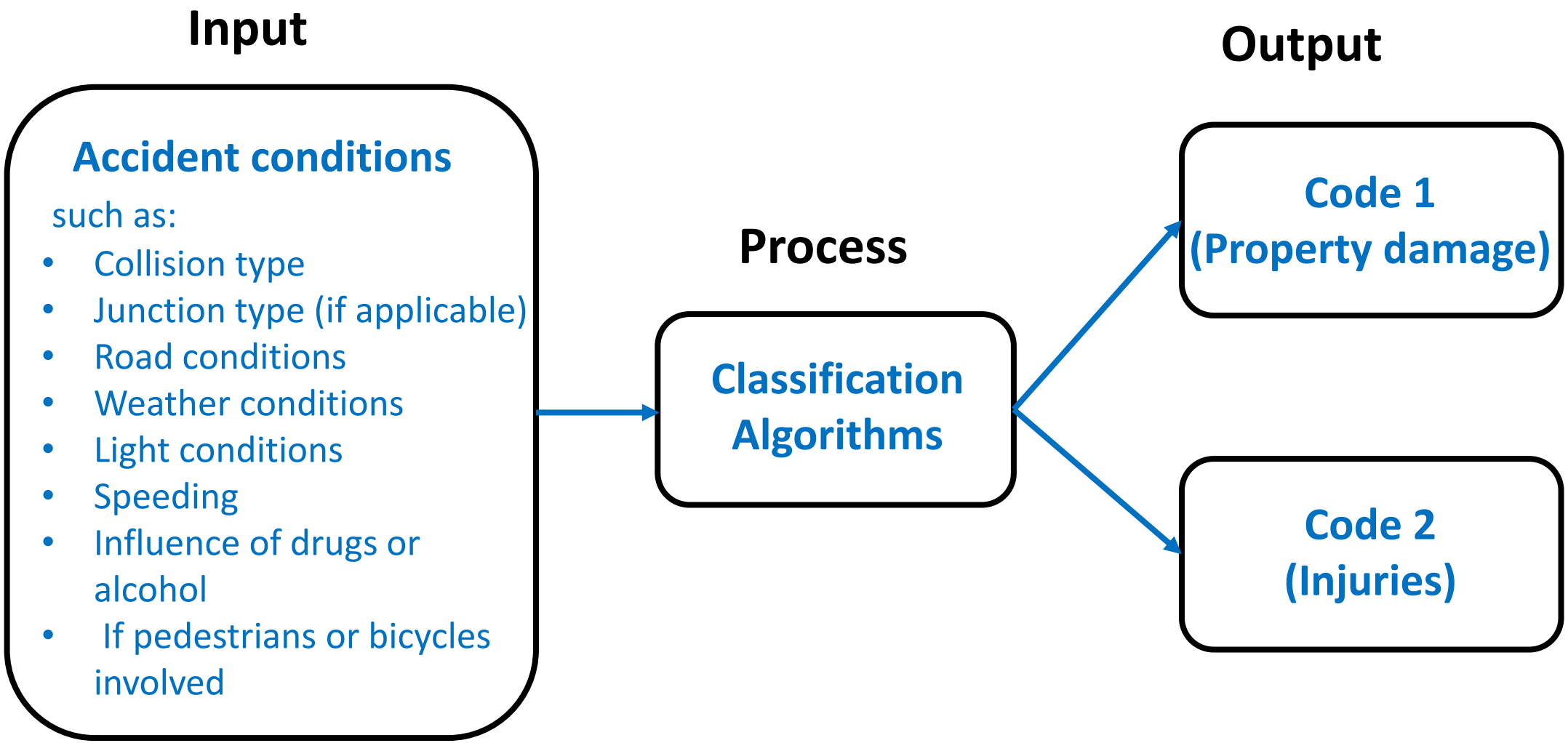
Elahe Talaie

October 2020

Introduction

- There were 1.35 million road traffic deaths globally in 2016, according to WHO.
- Understanding the most relevant factors contributing to road accidents is critical to improve road safety and save lives.
- Accident analysis and prediction is a significant field of research considering its impact on public safety.

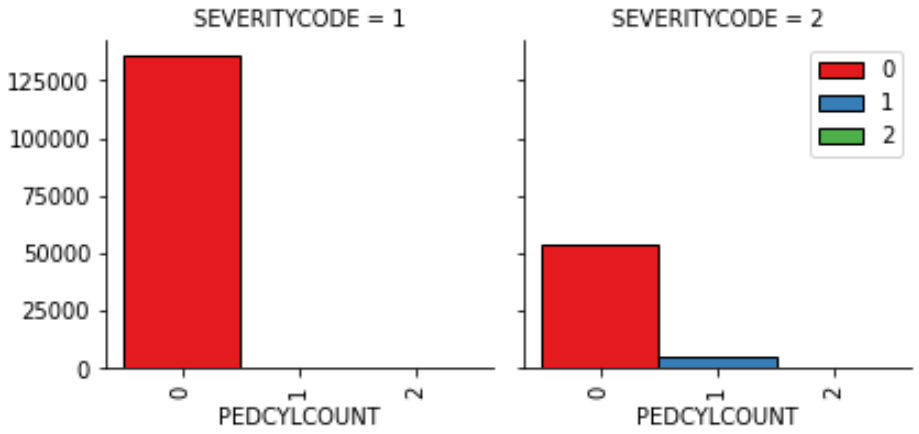
Objective: To Predict Accident Severity

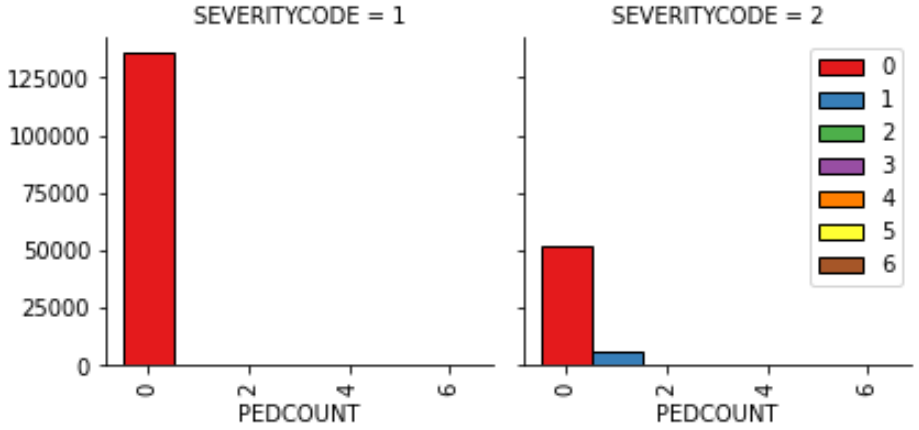


Data and Methods

- Data Source: the Seattle Department of Transportation (SDOT) Traffic Management Division, Traffic Records Group
- 194673 rows and 38 columns
- Target column: Severity Code (Code 1 and Code 2)
- One-Hot Encoding to convert categorical variables to binary variables
- 70% of data with Severity Code 1
- Negative sampling for data balancing
- Classification Algorithms: K-Nearest Neighbors (KNN), Decision Tree (DT), Support Machine Vector (SVM), and Logistic Regression (LR)

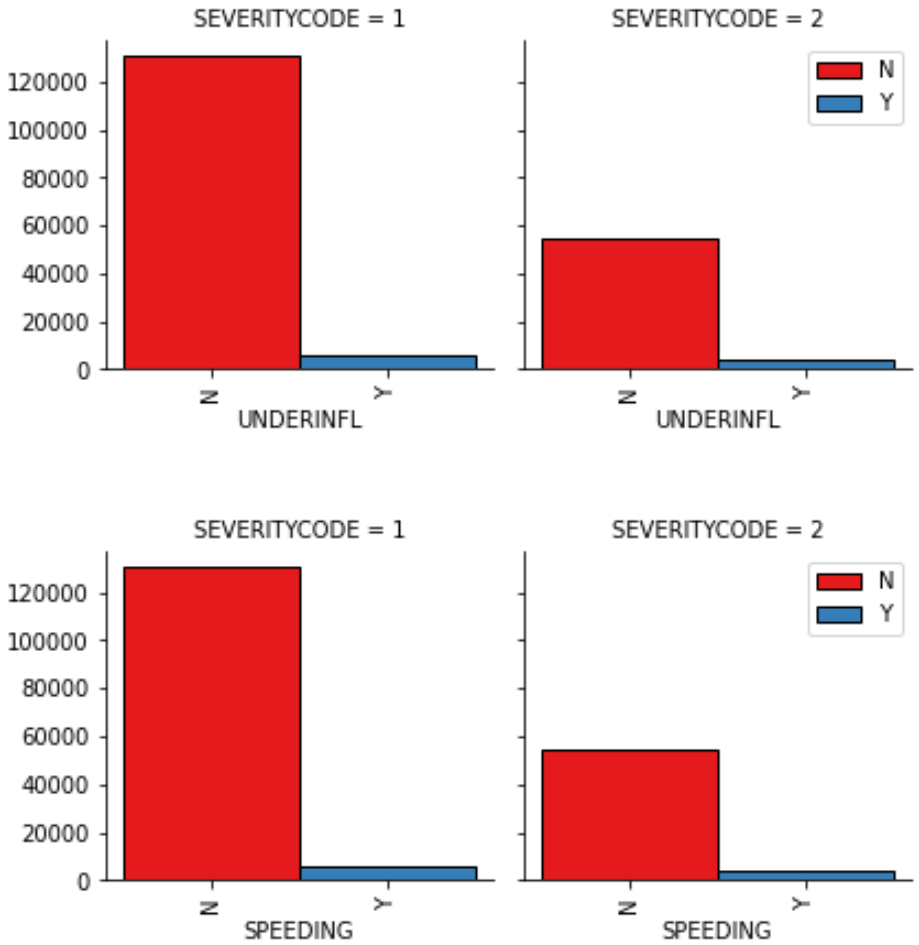
Data Exploring





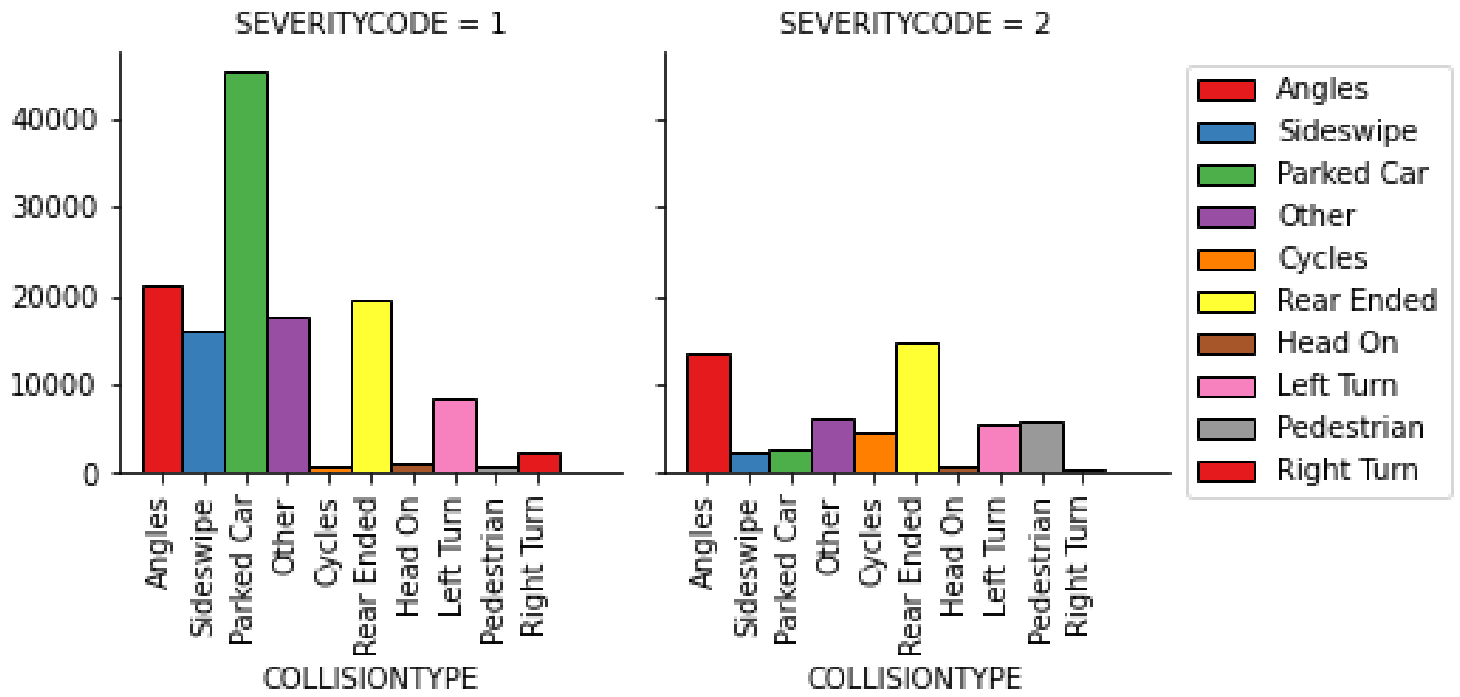
When pedestrians or bicycles were involved in the accident, the ratio of the accidents with injuries was increased.

Data Exploring



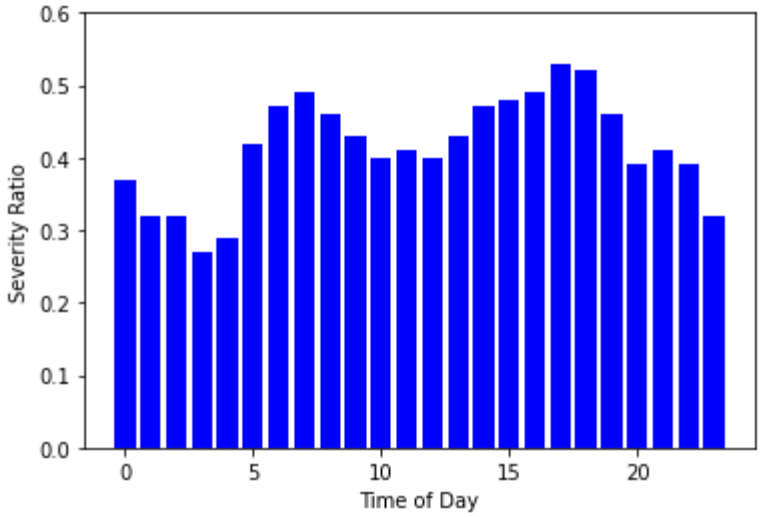
When the driver was under the influence of drugs or alcohol, or when speeding was a factor in the accident, the ratio of the accidents with injuries was increased.

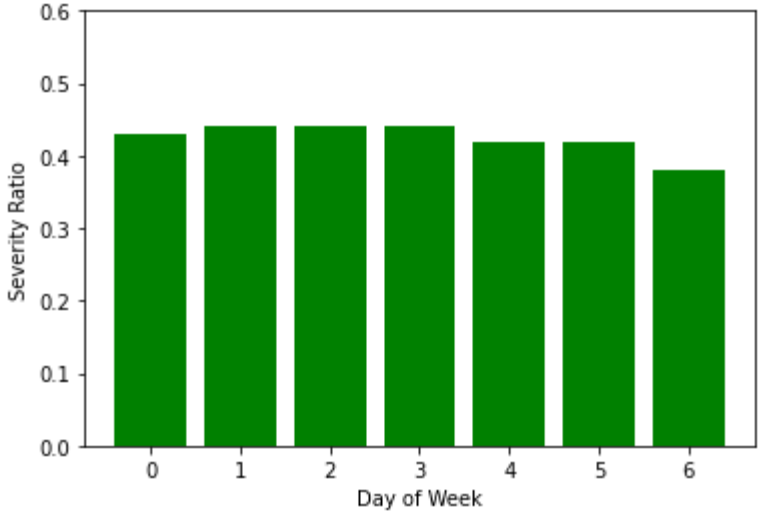
Data Exploring



Collision type was an significant factor affecting the severity of an accident.

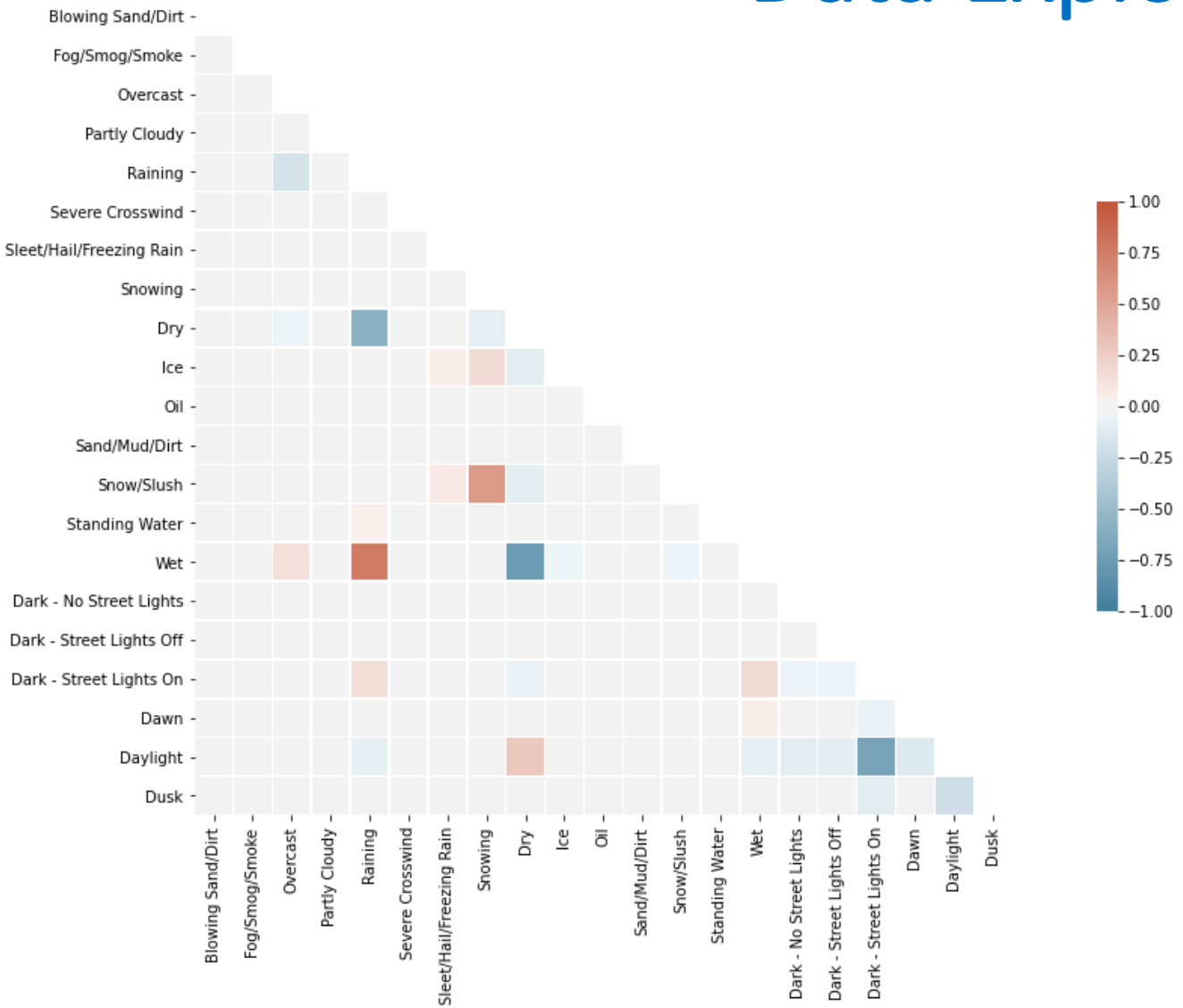
Data Exploring





The severity ratio (defined here as the number of accidents with injuries over the number of accidents with property damage) was higher during rush hours, but almost the same over different days of the week.

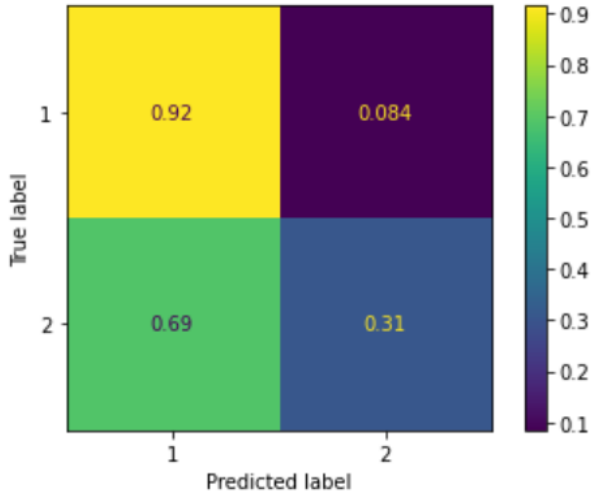
Data Exploring



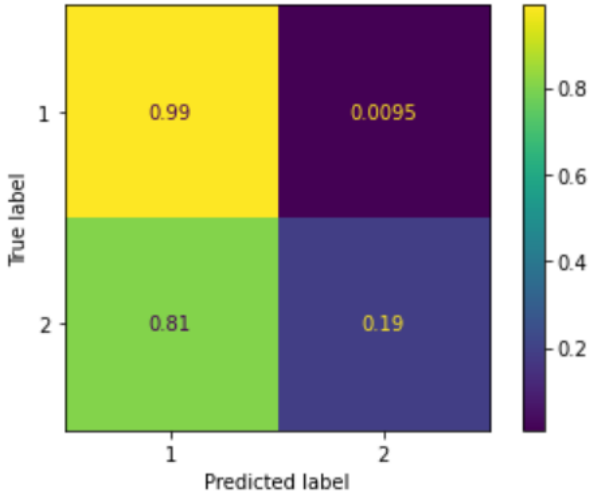
Heatmap showed that some features had strong correlations. Some of them were dropped from the feature set.

Confusion Matrix
(normalized to true test values)

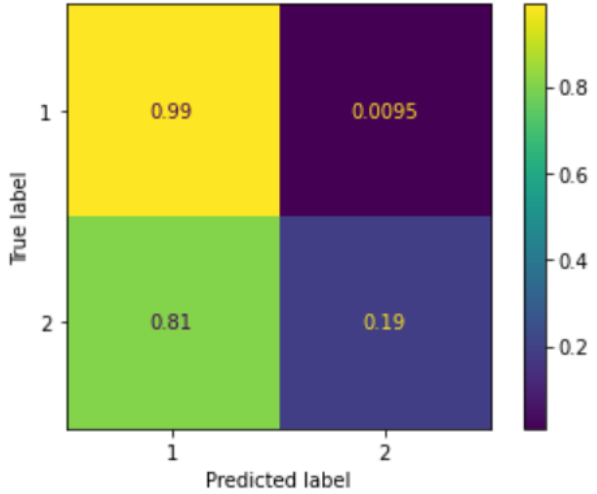
KNN



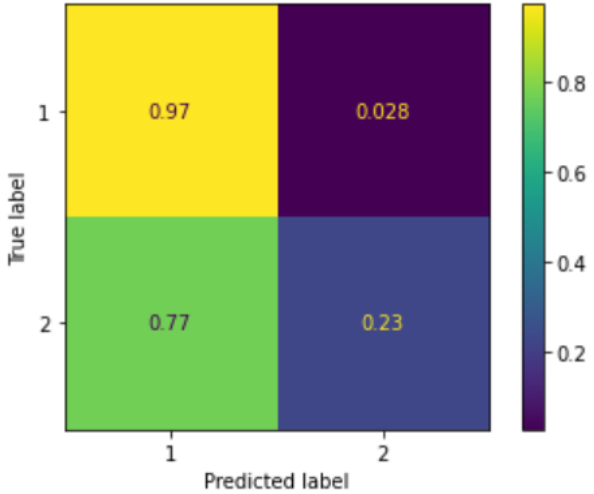
SVM



DT



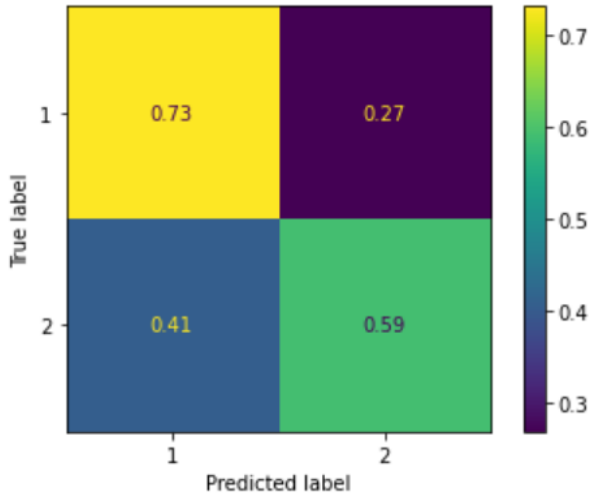
LR



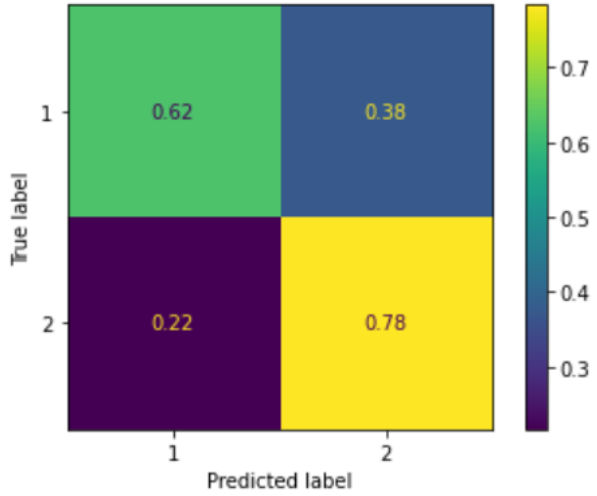
The classifiers were trained on **imbalanced** (original) test set.

Confusion Matrix
(normalized to true test values)

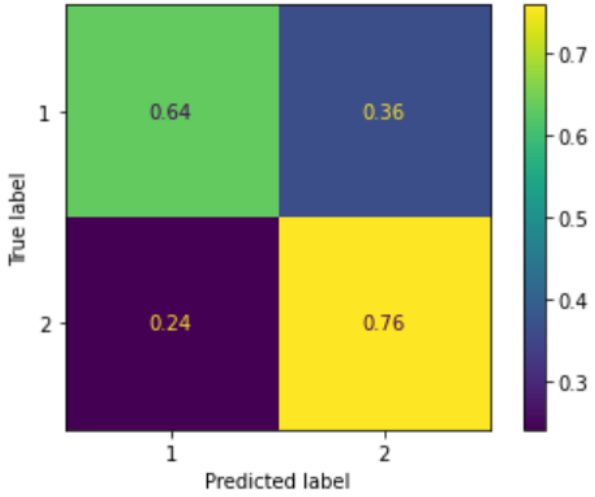
KNN



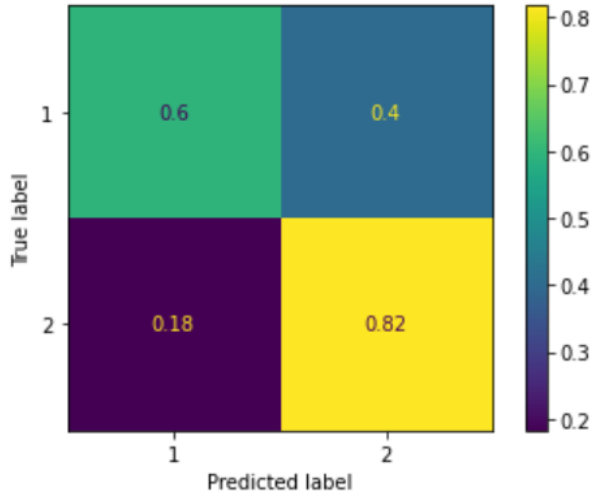
SVM



DT



LR



The classifiers were trained on **balanced** test set.

F1-Scores

	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** Training Set	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** Training Set	0.77	0.53	0.70	0.73	0.59	0.69	0.73	0.58	0.69	0.72	0.59	0.68

Conclusions

- 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.