Predicting Car Accident Severity in Seattle, United States

Ibrahim Maassarani

October 12, 2020

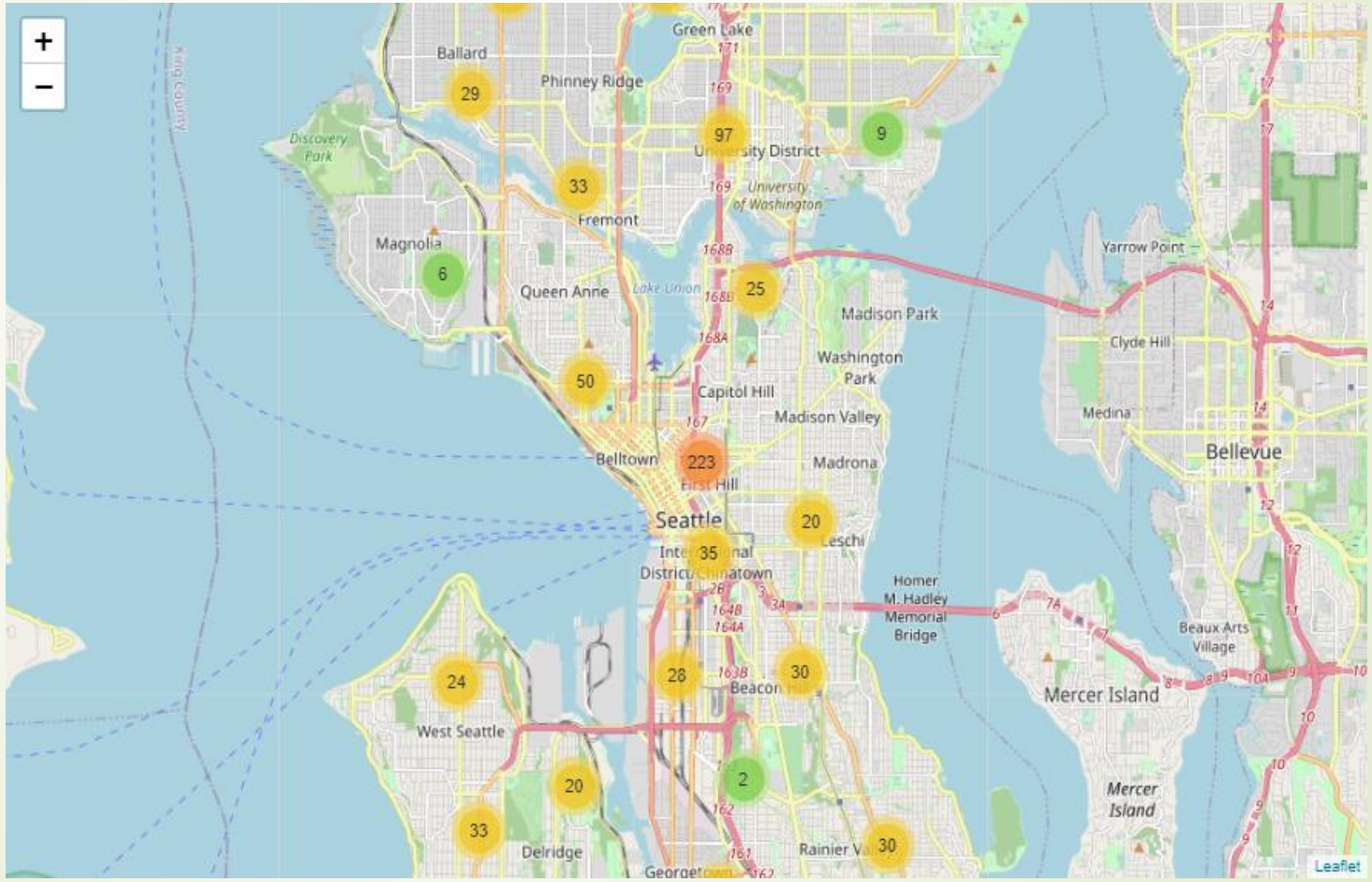
Predicting Car Accident Severity is of Great Importance

- Driving cars is a part of our daily lives in modern societies
- Car crashes is a major cause of death (a leading cause of death in the United States)
- Car collisions cause damages to roads, properties, etc…
- People lose their lives or get seriously injured and others would get hurt due to car accidents
- Predicting an accident's severity will offer insights on how to drive safely, reduce the number of crashes, as well as their severity

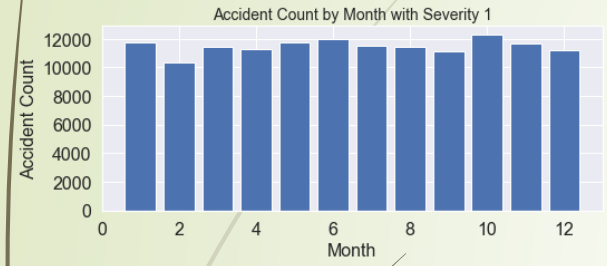
Data Acquisition and Cleaning

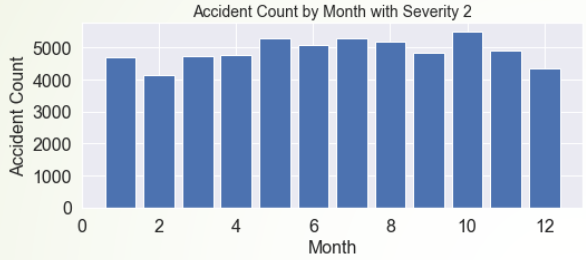
- Dataset that included all types of collisions in Seattle with a detailed description, from 2004 until 2018, found on Kaggle
- Seattle's coordinates were determined using geopy python package
- Raw data consisted of a total of 40 columns and 221738 records
- There were no duplicate values, but lots of problems such inconsistency in the data, missing values (some are meaningful but the majority is not), etc...
- Cleaned data consists of 15 columns and 171504 records

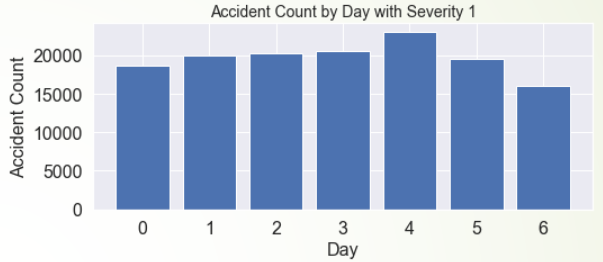
Location do Affect Crash Numbers and their Severity

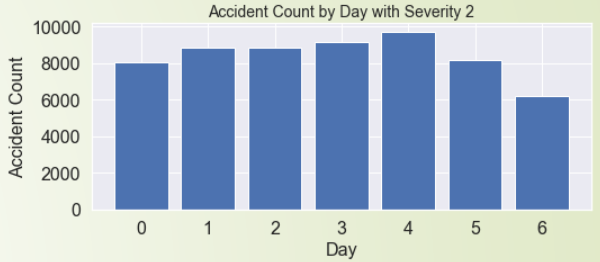


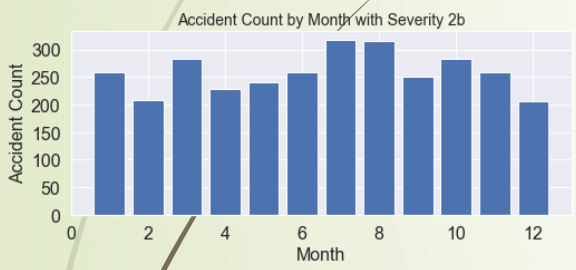
No Strong Correlation between Month or Day and Accident Severity

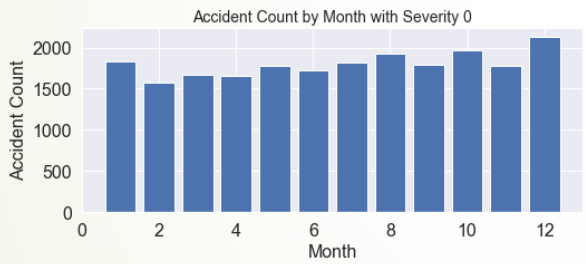


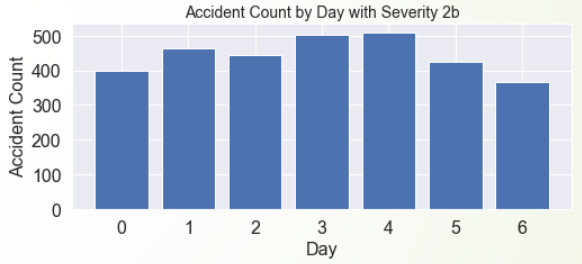


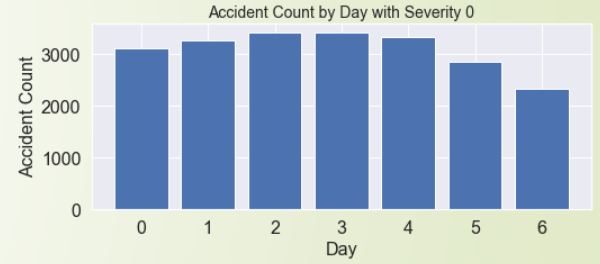


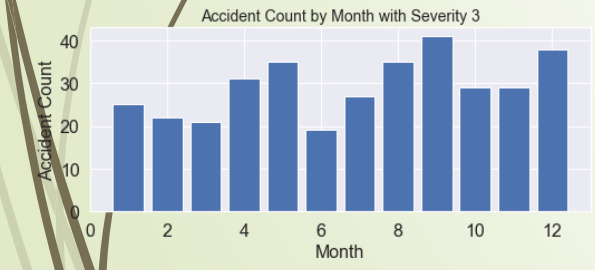


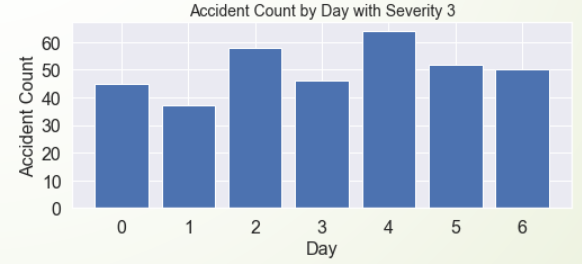




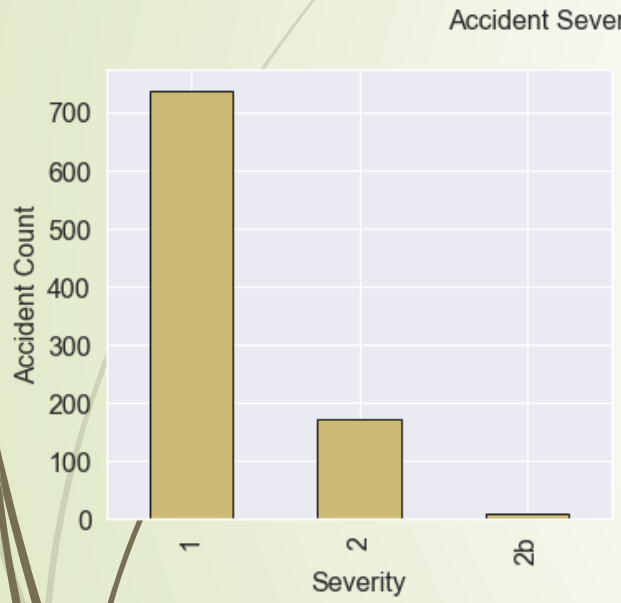


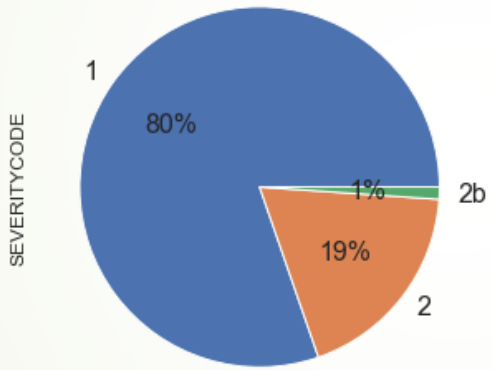


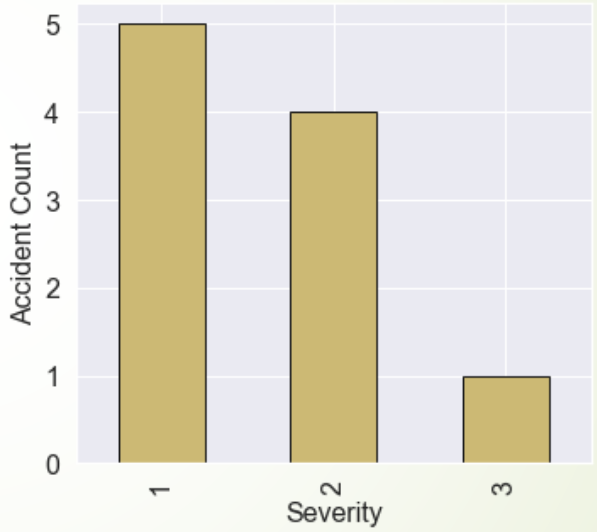


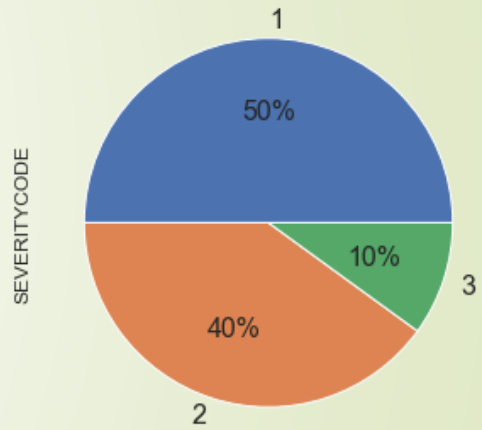


Special Weather Conditions Affect a Crash's Severity



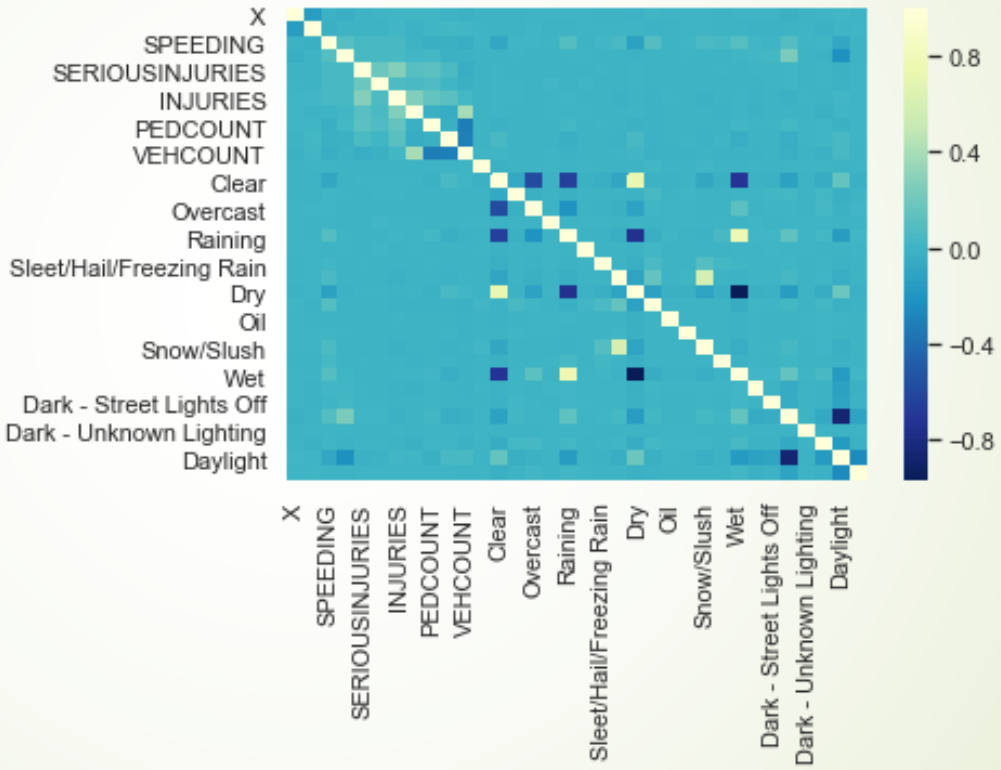






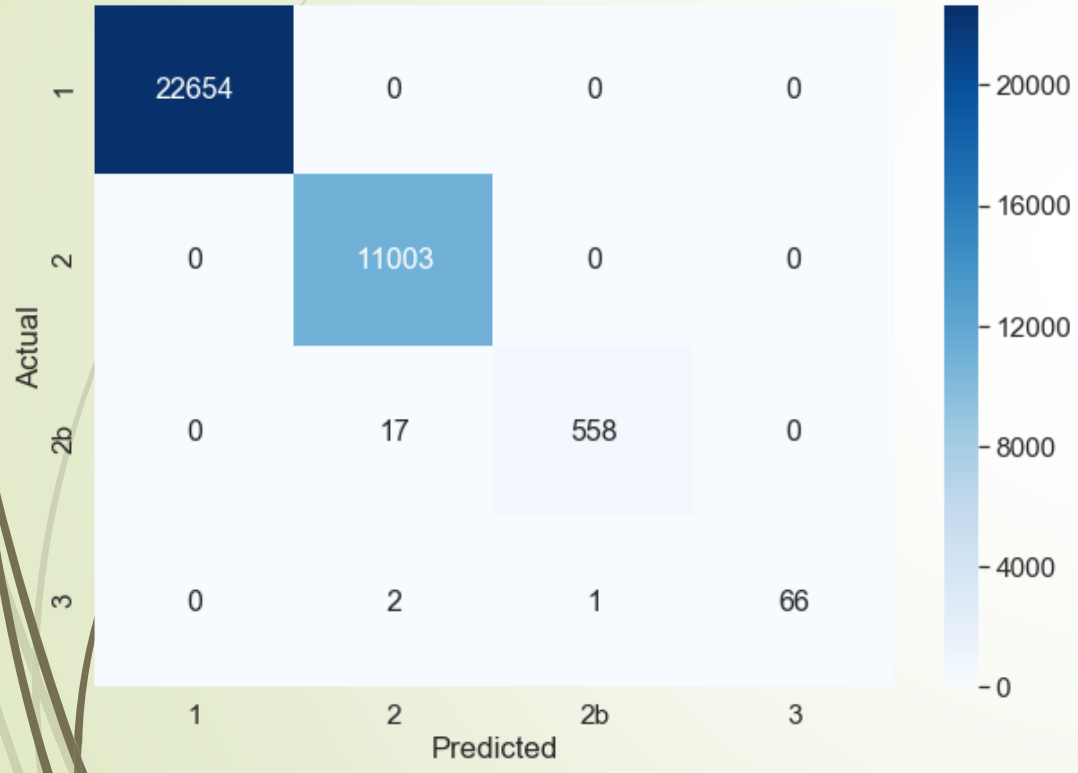
- We can see how the distribution of the severities, as well as the type of severity, change with certain weather conditions

Correlation between the Selected Features in the Clean Dataset



- It is obvious from the heat map above that certain features do have strong positive or negative correlation.

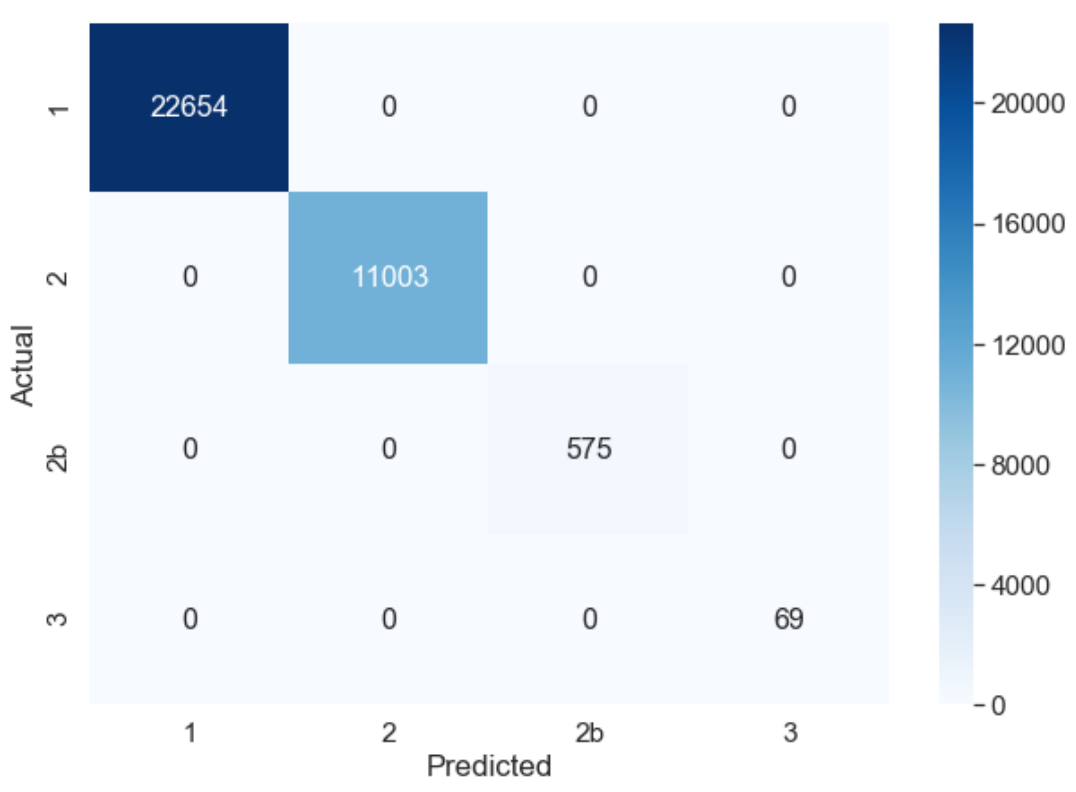
Logistic Regression Performance (Testing Phase)



	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	22654
2	1.00	1.00	1.00	11003
2b	1.00	0.97	0.98	575
3	1.00	0.96	0.98	69

- 19 crashes predicted as Severity 2 but 17 were Severity 2b and 2 were Severity 3
- 1 crash predicted as Severity 2b but was Severity 3
- This model misclassified 20 crashes
- 99.941% training accuracy and 99.958% testing accuracy

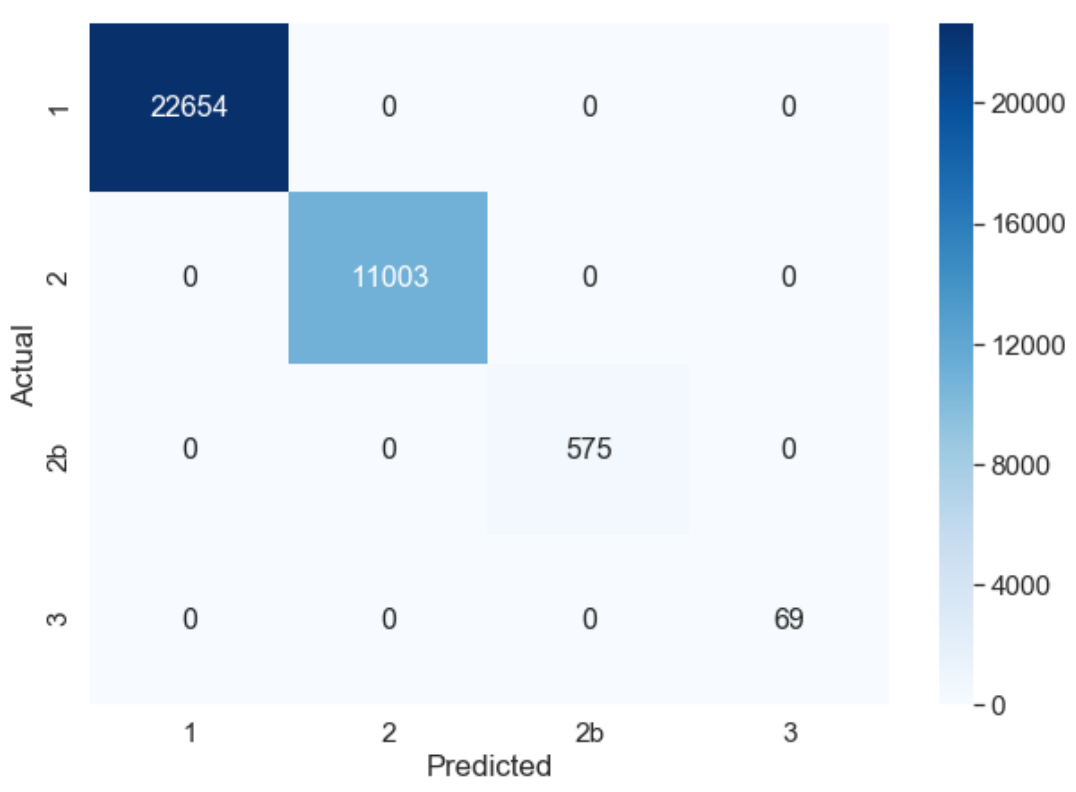
Support Vector Machine Performance



	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	22654
2	1.00	1.00	1.00	11003
2b	1.00	1.00	1.00	575
3	1.00	1.00	1.00	69

- The severity of every accident was predicted correctly
- 100% training accuracy and 100% testing accuracy

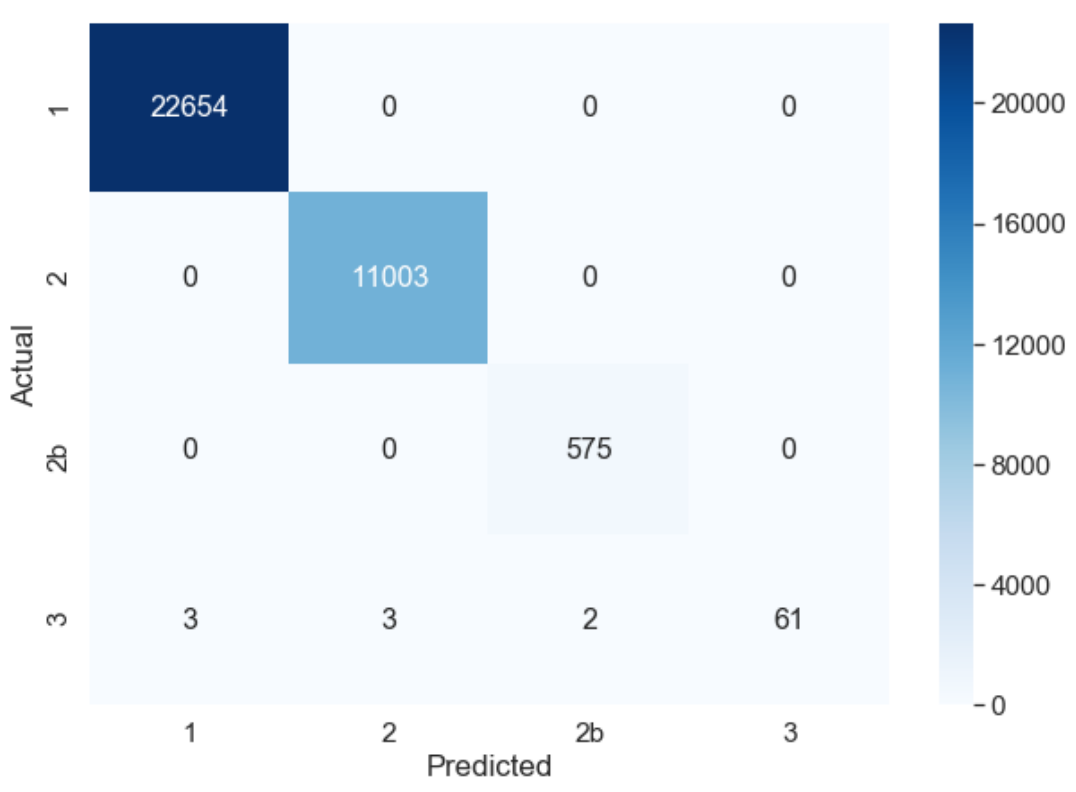
Decision Tree Performance



	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	22654
2	1.00	1.00	1.00	11003
2b	1.00	1.00	1.00	575
3	1.00	1.00	1.00	69

- The severity of every accident was also predicted correctly
- 100% training accuracy and 100% testing accuracy

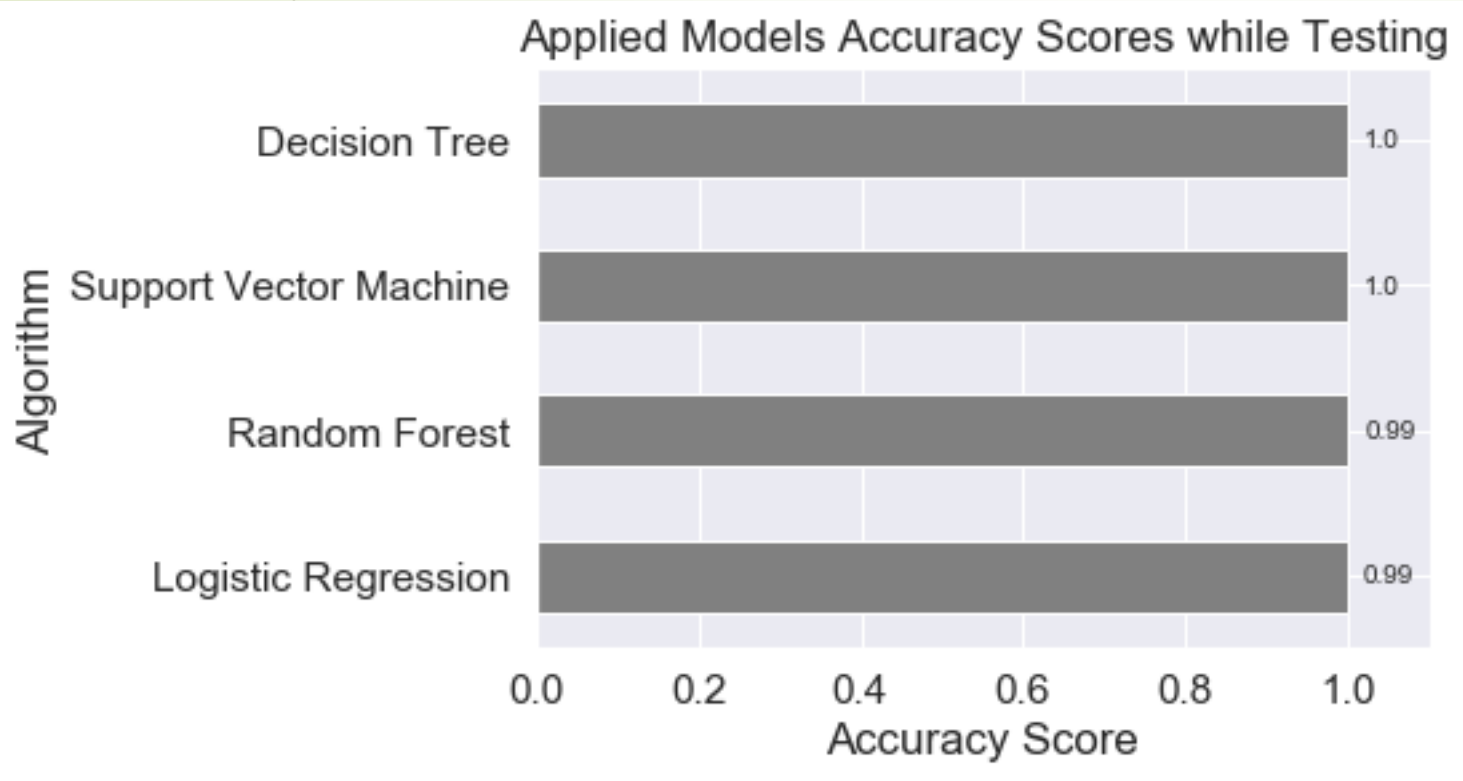
Random Forest Performance



	Precision	Recall	F1-score	Support
1	1.00	1.00	1.00	22654
2	1.00	1.00	1.00	11003
2b	1.00	1.00	1.00	575
3	1.00	0.88	0.94	69

- Predicted 3 of Severity 1, 3 of Severity 2, and 2 of Severity 2b, but all of the 8 were actually of Severity 3
- Recall and F1-score corresponding to Severity 3 from the table are lower due to the mentioned above
- 100% training accuracy and 99.976% testing accuracy

Results Summary



- As we can see, Support Vector Machine and Decision Tree performed best at an accuracy of 100% on the testing data. Still, Random Forest and Logistic Regression performed excellently with an accuracy above 99%.

Conclusion

- All 4 algorithms showed excellent accuracies, exceeding 99%, on both training and testing datasets
- Our models were well trained and fit to the training data
- Our models had excellent performance on the testing data
- We could state that our data was well handled and cleaned
- All 4 models can accurately predict a car accident's severity in the city of Seattle