Classification of Accident Severity

Seattle collisions analisys

Almerindo Uazela

10/14/20

IBM Data Science

Contents

1. Introduction

One of the current concerns of the Seattle Department of Transportation (SDT) is to find solutions that can minimize the number of traffic accidents, as well as deaths, injuries and damage from traffic accidents, in this context, all relevant information about the occurrence of accidents are registered and maintained by the department for access by all researchers. This data is necessary for the planning and implementation of countermeasures, operational control and for evaluating road safety programs and improvements.

1.1. Problem

Accidents have different severities and the implementation of countermeasures must be able to take this aspect into account in order to prioritize their projects, therefore, the process of classifying the severity of accidents must be carried out as accurately as possible. This project intends to analyze all collisions registered since 2004, in order to build a Machine Learning Model to predict the classification of the severity of new accidents based on their characteristics with 100% accuracy.

1.2. Interest

The department may have more accurate and automatic accident classification and, consequently, programs that receive grant funding in the Seattle region have benefited from this information to define their action plan. City citizens may be interested in knowing under what circumstances they are most vulnerable to being involved in car crashes.

2. Data Description

2.1. Data Source

The data set used in this project is available in a comma-separated values (CSV) file format and has been downloaded from Seattle Open GeoData Portal and includes all types of collisions since 2004 to Present. The data set contains 221738 redcords and 40 fields. The dataset contains columns with 3 diferent type of values, float64, object and Int64.

The dataset metadata was found at Department of Transportation Seattle.

2.2. Data Cleaning

The intention was to keep as many variables as possible after cleaning. In this sense, the following steps were performed:

a) a) Dataset is part of an information system of the Seattle government and relates to other tables, with this, some columns represent secondary keys of these relationships were all dropped as well as the primary key and other references columns.

b) As columns with Boolean behavior values where ***Y*** corresponds to ***TRUE*** and ***N*** or null value field (NAN) corresponds to ***FALSE*** were standardized for numeric values ***0*** and **1**, where ***1*** = ***TRUE*** and ***0*** and ***NAN*** = ***FALSE***.

c) Day of the week and Month was extracted from the column containing the date of occurrence and the original column was dropped.

d) The target variable appears represented by two columns, SEVERITYCODE which has the severity code with the values (1,2,3,2b, 0) has been changed to the type int64 with the values (1,2,3,4, None) the other column is SEVERITYDESC which includes the description and severity maintained in the data set to give more emphasis on the interpretation of the graphs.

e) Some Columns use the value ***Unknow*** and ***other*** to represent lack of Data, therefore THESE Values were replaced by **NAN**.

f) All records containing at least one NAN value were dropped from the Dataset.

With data cleaning, 221738 records generated 148,171 (68%) and the 40 columns were transformed into 23 features (58%).

3. Methodology

3.1. Target

The target is the classification of accident severity represented by two columns in dataset, one containing the severity code and the other description. Both columns were maintained, however, the first was converted to numerical values to allow the application of mathematical operations and the second will be used for data visualization purposes.

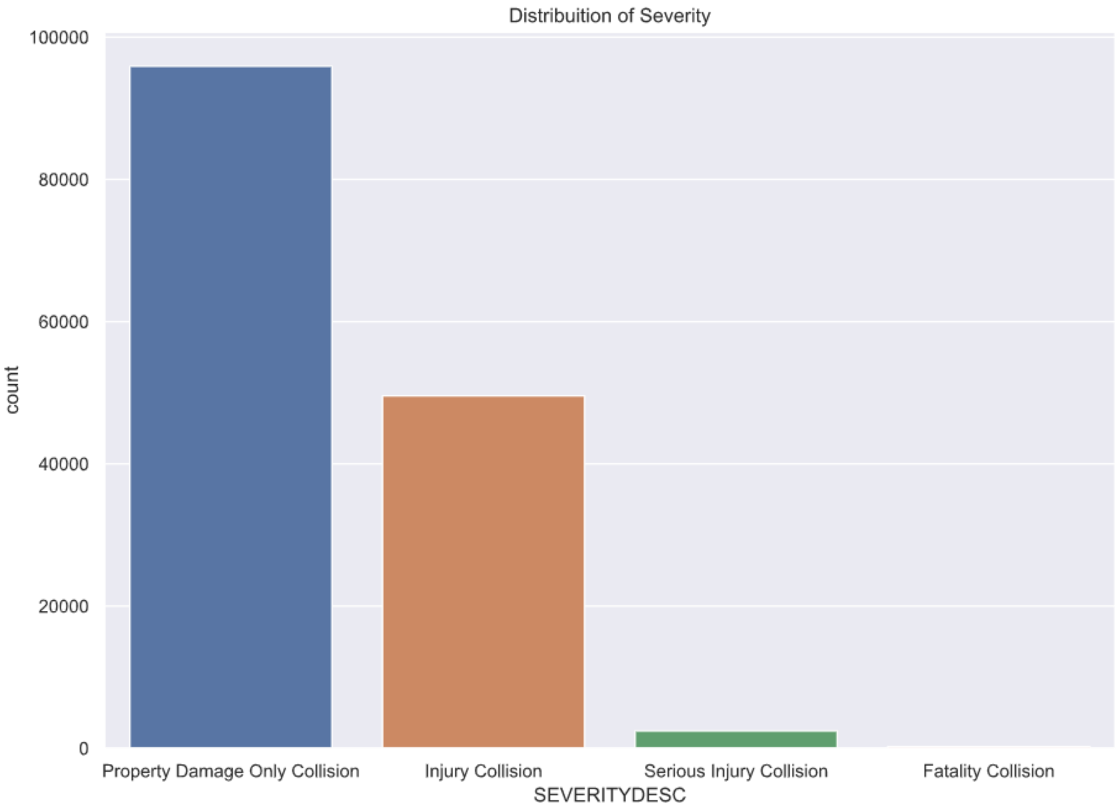


Figure 1. Severity distribuition by occorrences

3.2. Exploratory Data Analysis

After cleaning, we proceeded to the selection of variables that have an important impact on the classification of the severity of an accident for this, the variables were separated into two groups, namely, Continuous Features and Categorical Features.

3.2.1. Continuous Variables

Continuous variables Include all independent variables of the numeric type float64 and int64. Here the mean of each variable was analyzed in relation to the severity distribution in order to identify the classification trend.

Geographic distribution of accident severity

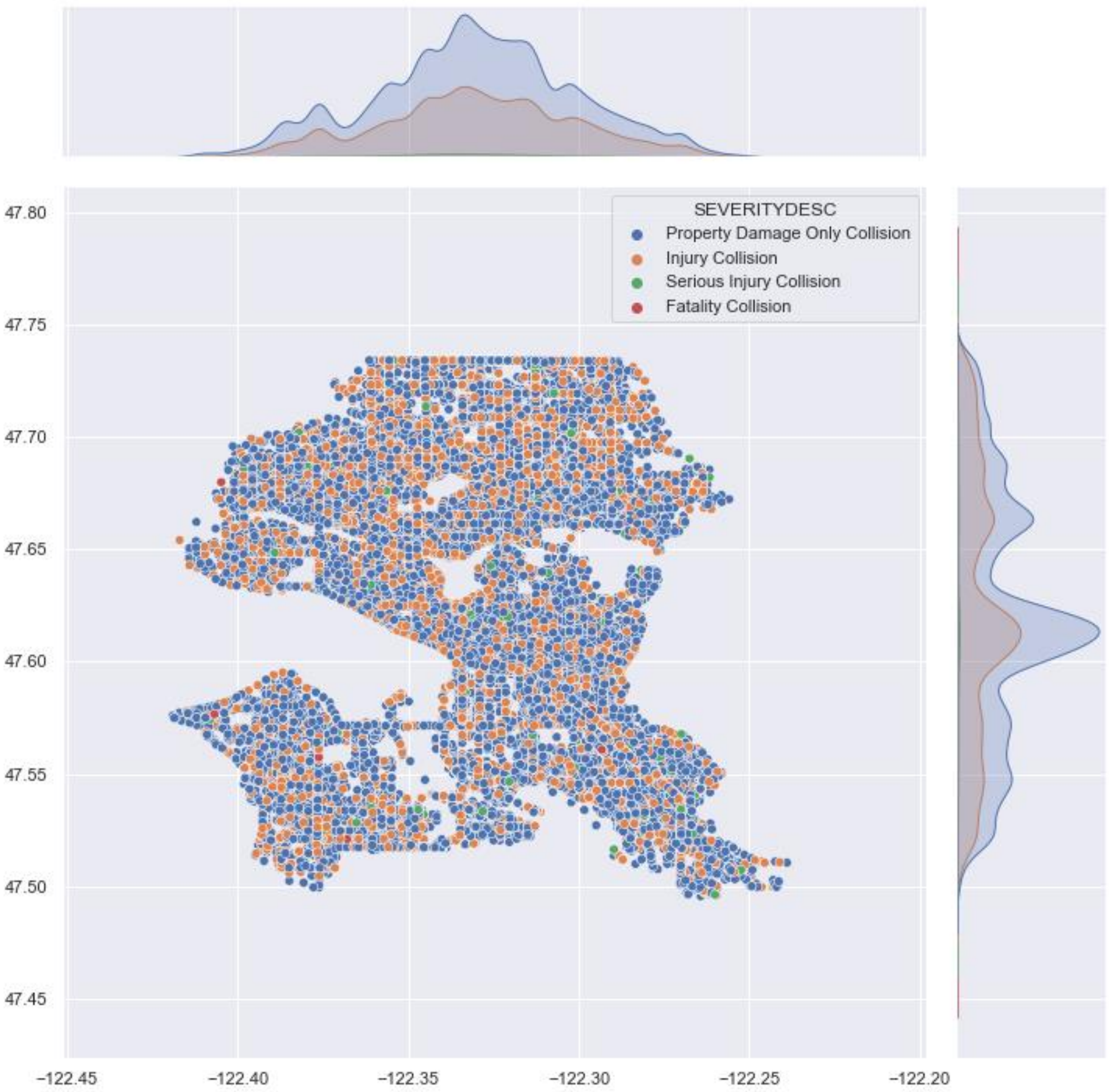
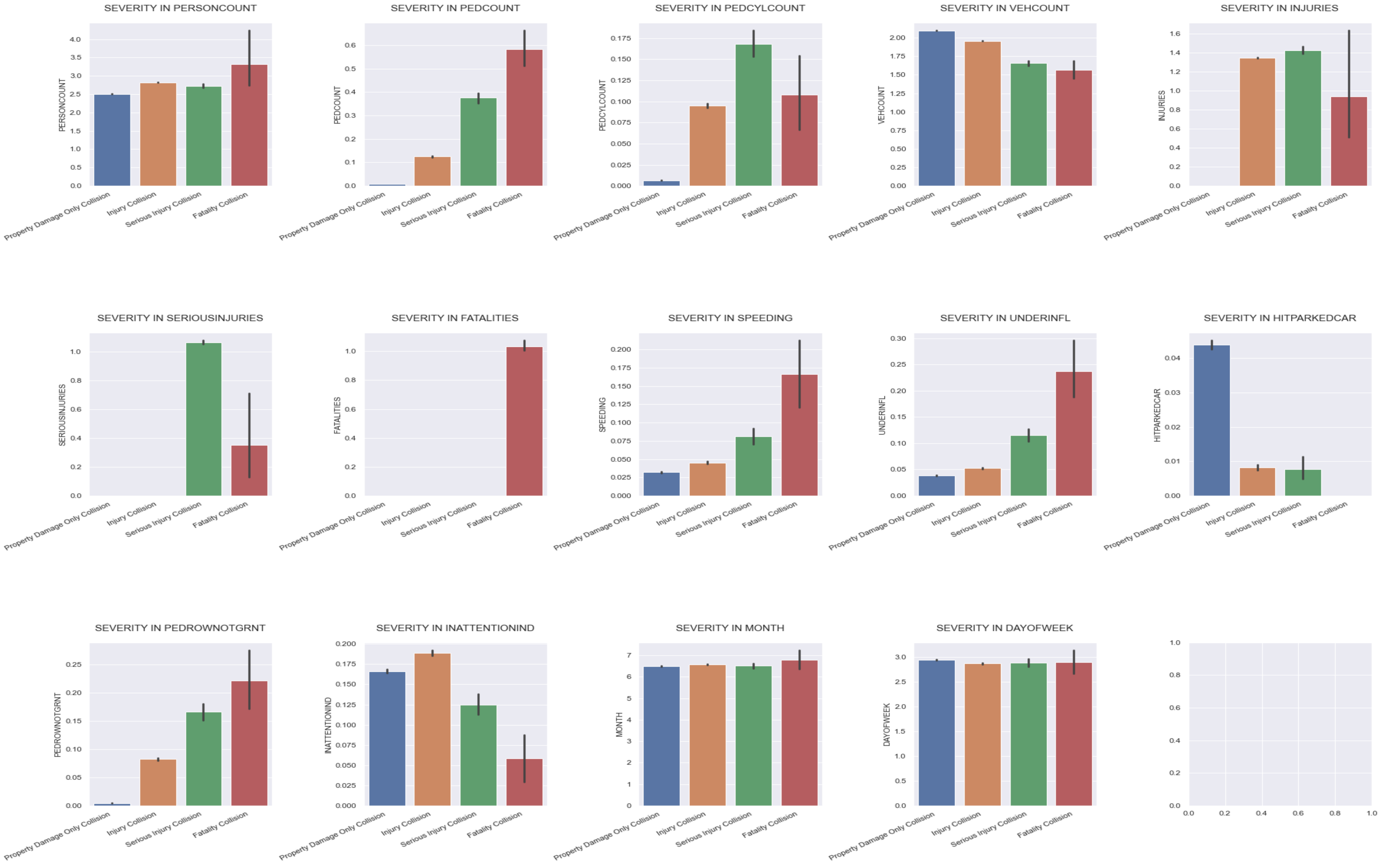


Figure 2. Geodistribuition of accidents severity

Observing the map, it can be seen that the severity of accidents is proportionally distributed, making it difficult to classify an accident only by the place of occurrence.

Accident severity distribution vs numerical variables


SEVERITY IN PERSONCOUNT

SEVERITY IN PEDCOUNT

SEVERITY IN PEDCYLCOUNT

SEVERITY IN VEHCOUNT

SEVERITY IN INJURIES

SEVERITY IN SERIOUSINJURIES

SEVERITY IN FATALITIES

SEVERITY IN SPEEDING

SEVERITY IN UNDERINFL

SEVERITY IN HITPARKEDCAR

SEVERITY IN PEDROWNOTGRNT

SEVERITY IN INATTENTIONIND

SEVERITY IN MONTH

SEVERITY IN DAYOFWEEK

Figure 3. Severity distribution by continuos features

Looking at Figure 4, it can be seen that each characteristic influences the classification of severity accident, in the relationship *SEVERITY IN FATALITIES,* for example, it is clear that any accident involving fatality tends to be classified as *Fatality collision,* in the relationship, *SEVERITY IN HIPARKEDCAR* it is noticed that accidents involving a high number of parked cars tends to be classified as *Property Damage Only Collision* and if the number of cars envolver is relatively low then it tends to be classified as *Seriuos Collision* or *Serious Injury Collision.*

In relationships like *SEVERITY IN MONTH* and *SEVERITY IN DAYOFWEEK* is very difficult to classify the accident because the distribution of severity is almost equal.

3.2.2. Categorical Variables

Categorical Variables Include all independent variables of the numeric Object type. Here, the severity distribution for each value that the variable assumes was analyzed in order to identify the values that have influence in classifcation of accident.

3.2.2.1. Distribuição de Severidade do acidente em função das variaveis Categoricas independents

Figure 6 explores the relationship between the severity of the accident and the categorical variables, we can see that some values of the variables do not contain enough information to classify the severity of the accident. These values will be dropped from the model to be created because it does not show a trend towards the target.

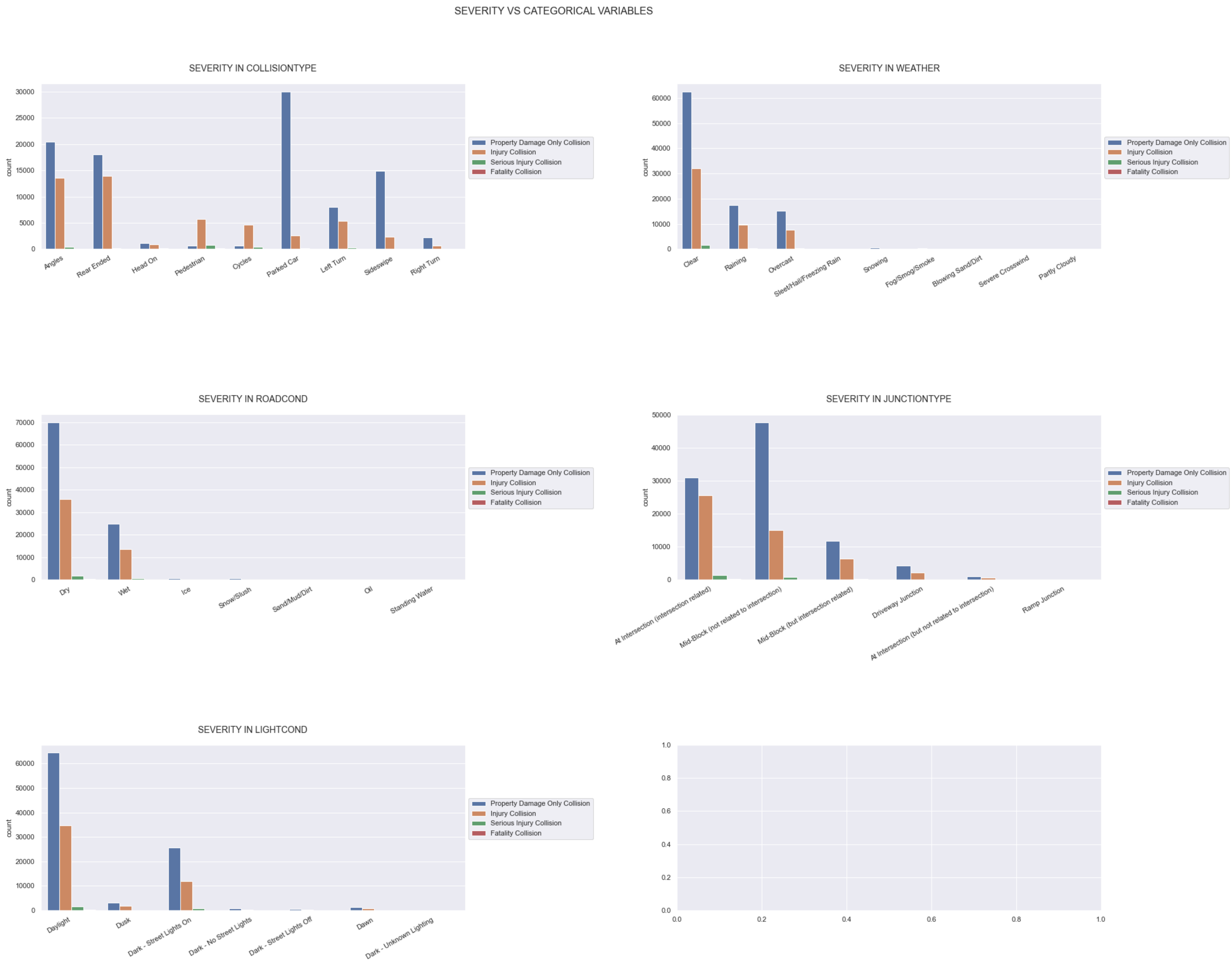


Figure 4. Severity distribiution by categorical features

3.2.2.2. Correlation

After identifying the relevant numerical as well as categorical variables, both variables were merged and then the values of categorical variables were transformed into dummy variables, making it possible to apply a correlation.

As we observed before in the analysis of categorical variables, some values with insufficient data have no impact on the classification of an accident and therefore should be dropped, therefore, all variables were correlated and all variables with very low (correlation greater than **- *0.03*** and lower ***0.03***) coefficient were dropped***.***

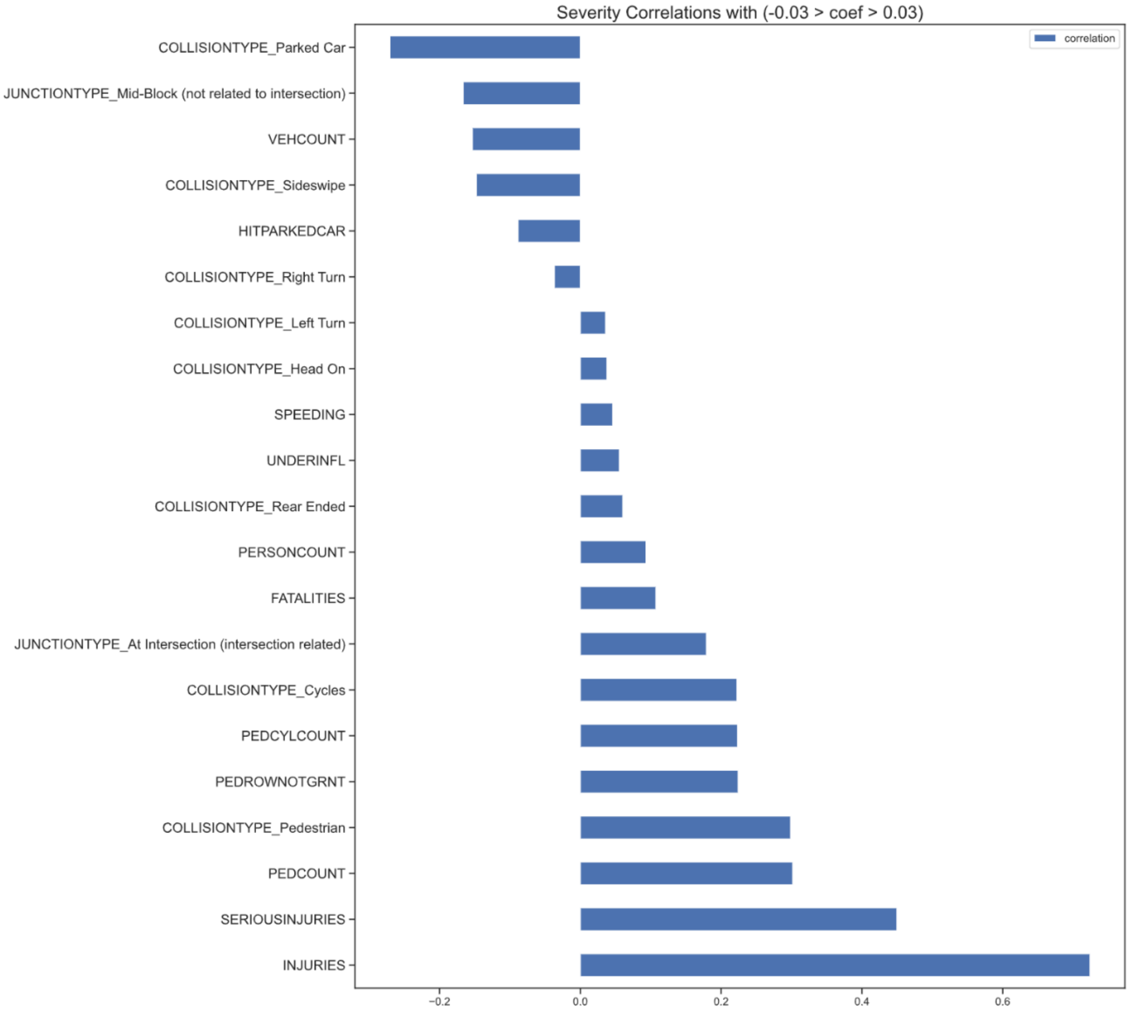


Figure 5. Severity accident correlactions

4. Modeling

The Records on the dataset contains all collisions already classified and the model must be able to learn from this data and then determine the classification of new accidents not yet classified. To achieve this, use used the approach *Train test split* using 80% of the dataset for the *Training Set* and the remaining 30% for the *Testing Set*. Since the problem is a case of supervised classification, the model used was a Classification Model.

4.1. Classification Models

The models were built using the Support Vector Machine (SVM), Decision Tree Classification and K-Nearest Neighbors (KNN) algorithms. All algorithms have an approximate 100% average accuracy using different metrics such as *f1score*, and *accuracy classification score.*

Algorithm	**F1 Score**	**Classification Accuracy**
Support Vector Machine	0.999629	0.999626
Decision Tree Classification	1.000000	1.000000
K-Nearest Neighbors	0.998583	0.998583

Table 1. Model acurracy

5. Discussion

The data set used in the reference project contained 221,738 records corresponding to collision occurrences so far, however, after data cleaning only 148,171 records were used as a sample, this is mainly due to the lack of information in some fields. 32% of records have at least 1 missing data.

Field Name	Count
X	7478
Y	7478
SEVERITY CODE	1
COLLISION TYPE	26451
WEATHER	26641
ROAD COND	26560
JUNCTION TYPE	11979
LIGHT COND	26730
INATTENTION ND	1

Table 2. Fields with missing data

Table 1 is mostly represented by categorical variables, which during our analysis were noted there was insufficiency of data to analyze some values of the categorical variables and consequently they ended up being dropped because the accident cannot be classified through the variables.

It cannot be categorically stated that the missing data are the dropped values, but it is undeniable that a data set without missing data enables the construction of richer models and it is important to standardize the data collection process.

6. Conclusion

Purpose of this project was to analyze Seattle collisions data and build a machine learning model in order to predict the classification of accident severity by its characteristics. By splitting the variables in categorical and continuous we identified and select as features the independent variables that have significative impact in accident classification. Several classification models were built and tested obtaining an average result of 100% accuracy using different accuracy metrics.

The implementation of this model can help Seattle Department of Transportation to classify the severity of accident more accurately and automatically with the data from the accident record. The model can be adjusted to include new variables if necessary.