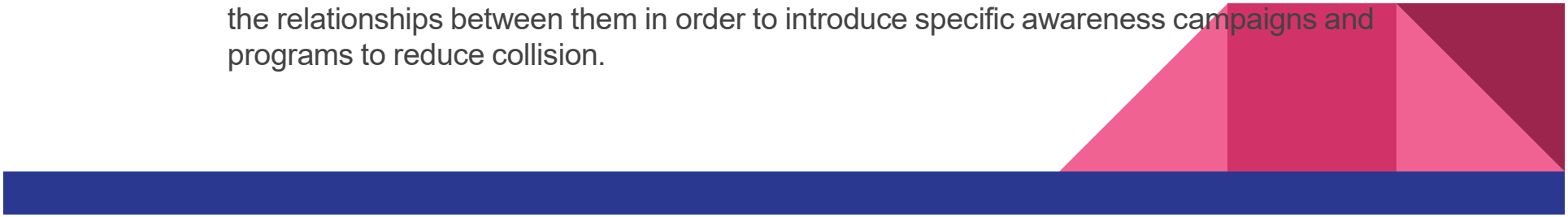
Predicting Car Accident Severity

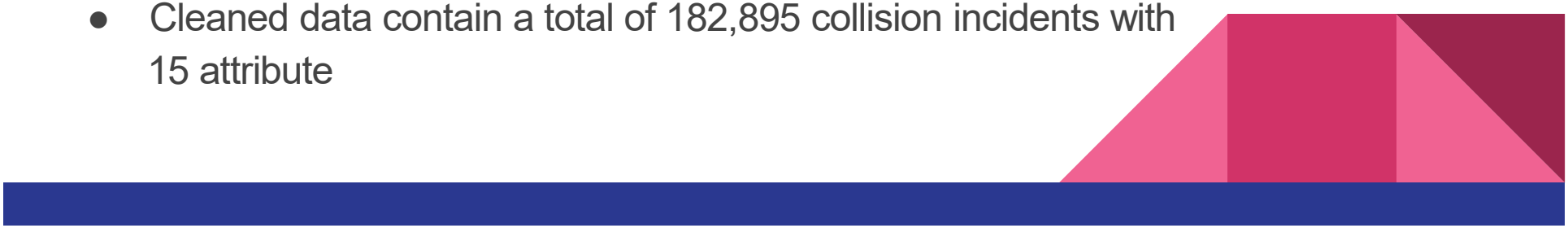
Introduction/Business

- More than 3,000 people die every day in the world, as a result of traffic accidents.
- The World Health Organization (WHO) qualifies the situation of traffic accidents as a priority public health problem throughout the world.
- In this proyect will build a model to predict the severity of an accident.
- Interested:
  - Drivers are interested in knowing the probabilities of a traffic accident to choose a better route, be careful when driving or postpone their trip.
  - Responsible for transport in each country, to better understand the contributing factors and the relationships between them in order to introduce specific awareness campaigns and programs to reduce collision.

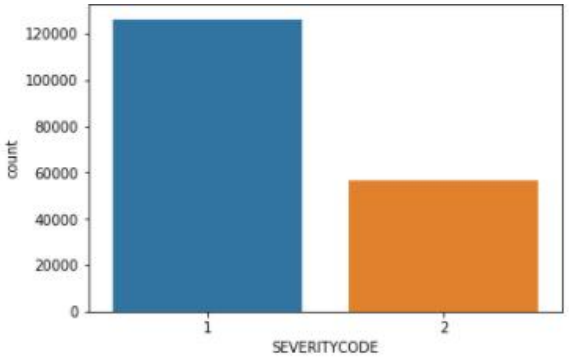


Data acquisition and cleaning

- Dataset called Data-Collisions.csv from the Seattle Department of Transportation for the period between 1st January 2004 and 20th May 2020.
- This data set contain a total of 194,673 collision incidents with 37 attribute.
- Dropped variables with many missing values and variables that do not add value
- Dropped row with missing values.
- Cleaned data contain a total of 182,895 collision incidents with 15 attribute



Exploratory Data Analysis

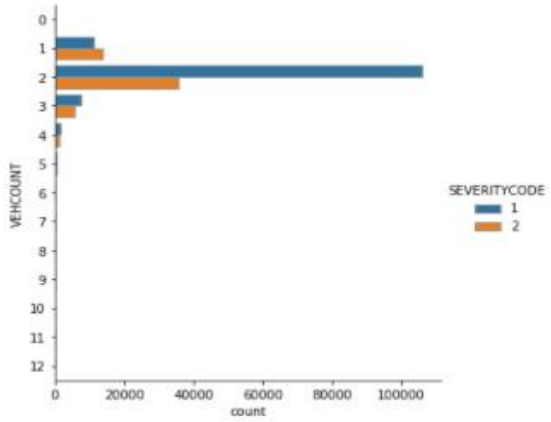


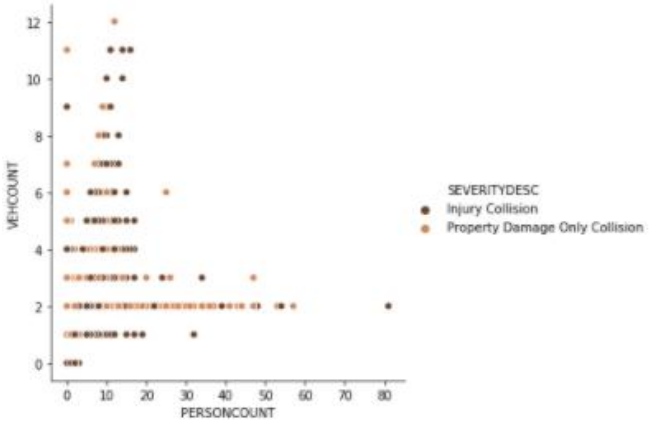
	SEVERITYCODE	INCKEY	PERSONCOUNT	PEDCOUNT	VEHCOUNT	UNDERINFL
SEVERITYCODE	1.000000	0.034259	0.124545	0.245656	-0.081166	0.035763
INCKEY	0.034259	1.000000	-0.049568	0.032564	-0.014410	0.702275
PERSONCOUNT	0.124545	-0.049568	1.000000	-0.027211	0.399674	-0.028545
PEDCOUNT	0.245656	0.032564	-0.027211	1.000000	-0.317361	0.029523
VEHCOUNT	-0.081166	-0.014410	0.399674	-0.317361	1.000000	-0.014674
UNDERINFL	0.035763	0.702275	-0.028545	0.029523	-0.014674	1.000000

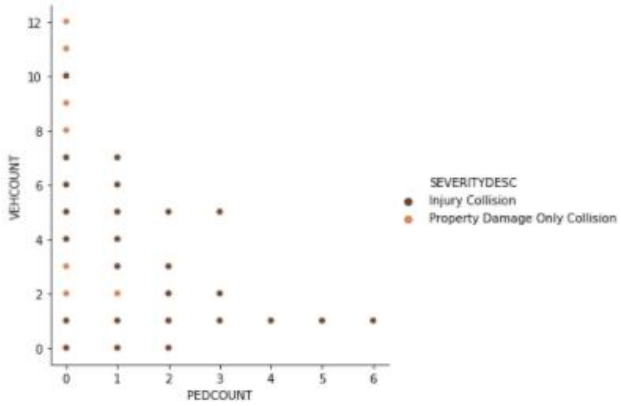
TARGET VARIABLE:
The vast majority of accidents involve "property damage only collision" (1), and only a 1/3 of accidents involve "injury collision“.

CORRELATION: not found correlation between variables numerical

Exploratory Data Analysis –Numerical variables



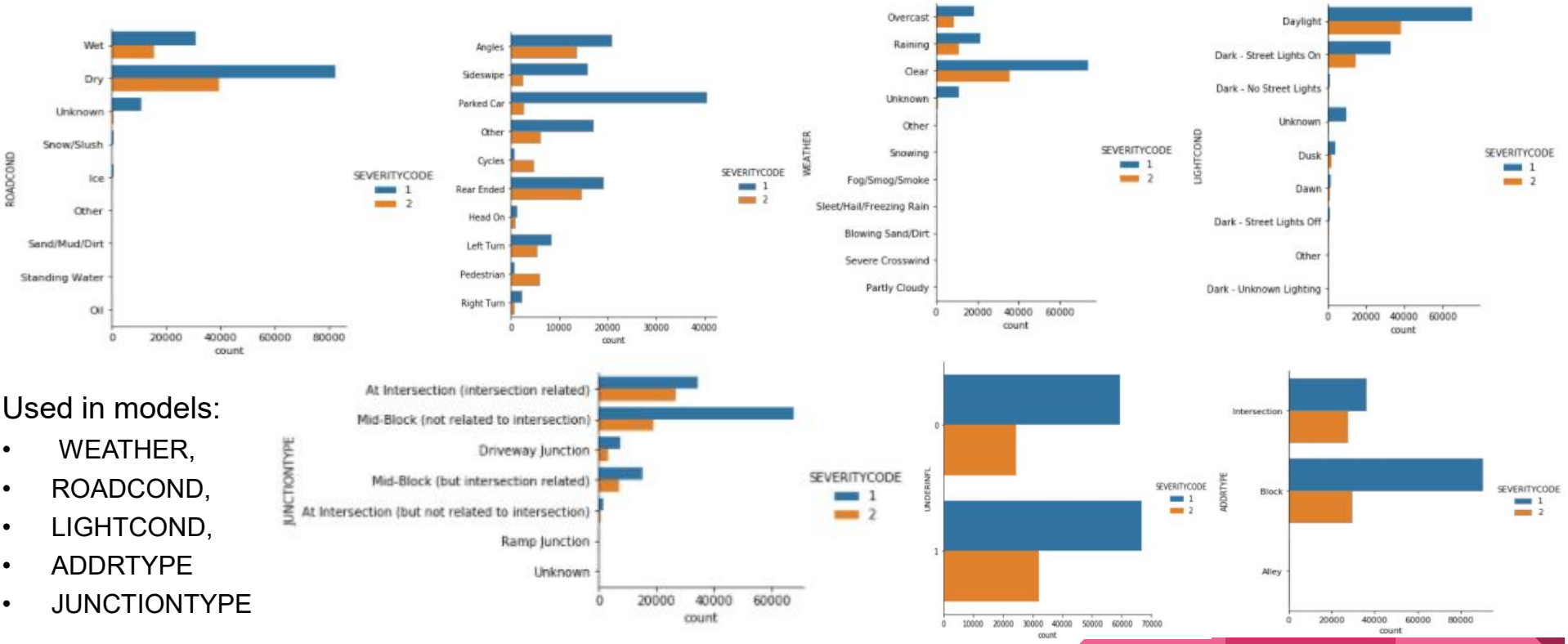




High relationship between the number of accidents and the number of cars, it was decided to use the variable VEHCOUNT. At the same time high relationship between:

- Number of cars and people in the car, PERSONCOUNT discarted
- Number of cars and pedestrians PEDCOUNT discarted

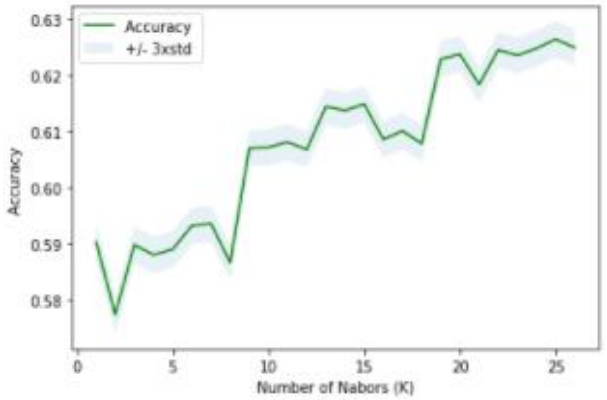
Exploratory Data Analysis –Categorical Variables



Used in models:

- WEATHER,
- ROADCOND,
- LIGHTCOND,
- ADDRTYPE
- JUNCTIONTYPE

Machine Learning Models



The best accuracy was with 0.6265342163355409 with k= 25

```
from sklearn.tree import DecisionTreeClassifier

jaccard_array=[]
f1_score_array=[]

md_initial=3

for md in range (md_initial,10,1):
    dt=DecisionTreeClassifier(criterion='entropy', max_depth=md)
    dt.fit(X_train,y_train)
    dt_yhat=dt.predict(X_test)

    jaccard=jaccard_similarity_score(y_test,dt_yhat)
    f1=f1_score(y_test,dt_yhat,average='weighted')

    jaccard_array.append(jaccard)
    f1_score_array.append(f1)

print(f'Best value for max depth = {jaccard_array.index(max(jaccard_array))+md_initial}')
print(f"Evaluation Jaccard = {jaccard}")
print(f"Evaluation F1 score = {f1}")
```

```
Best value for max depth = 7
Evaluation Jaccard = 0.6381456953642384
Evaluation F1 score = 0.6378330080896863
```

4 different models:

- KNN
- Decision tree
- SVM
- Logistic Regression

KNN and Decision Trees have the deficiency, that it is necessary to find the number of optimal neighbors k and the optimal depth, max_depth, in these cases they are k 25 and max_dept = 7

Models Evaluation

Algorithm	Jaccard	F1-score	LogLoss
KNN	0.63	0.62	NA
Decision Tree	0.64	0.63	NA
SVM	0.63	0.63	NA
LogisticRegression	0.60	0.59	0.67

- Used Jaccard, F1-score and LogLoss for evaluation:
  - Best model are Decision Tree.
  - KNN-SVM are Good models too.
  - The Logistic Regression model performs poorest.Team
  - In general, decent indices but with high possibility of improving models

Conclusion and future directions

- Built useful models to predict accident severity.
  - The Decision Tree model perform best, with an average F1-score of 0.64 and Jaccard of 0.63
- Accuracy of the models has room for improvement.
  - The best models have Jaccard and F1-score indexes close to 1 and in our case they are approximately 0.63
- Revealing hidden patterns in predicting severity in accidents based on the features Weather, Road and Light conditions, addresstype, junctiontype and vehcount.
- Another project can start to collect more information from other sources or directly from the roads.