CAR CRASH SEVERITY ANALYSIS

SEATTLE, WASHINGTON

BACKGROUNG AND PROBLEM

- The cost of car accidents in terms of economy and society amounts to $871 billion per year
- >100 people die in the US every day
- This means a death every 14 min

→ GOAL OF THE STUDY: predict, based on selected factors, how the severity of car accidents could be reduced.

Data

- Collection:
  - The dataset contains 194673 observations of collisions in the city of Seattle
  - It includes all types of collisions
  - The dataset is provided by SPD and recorded by Traffic Records
  - The timeframe of the observations goes from 2004 to 2020
- The models' aim is to predict the severity of an accident, the data has been prepared accordingly
- The following 5 features were selected for this study along with the target variable: Severity Code.

Feature Variable	Data type, length	Description
INATTENTIONIND	Text,1	Whether or not collision was due to inattention. (Y/N)
UNDERINFL	Text, 10	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Text, 300	A description of the weather conditions during the time of the collision.
ROADCOND	Text, 300	The condition of the road during the collision.
LIGHTCOND	Text, 300	The light conditions during the collision
SPEEDING	Text, 1	Whether or not speeding was a factor in the collision. (Y/N)

Methodology

Logistic Regression: Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable

Decision Tree Analysis: The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

k-Nearest Neighbor: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance)

Results and Discussion

Algorithm	Avg. f-1 Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41
k-nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

Conclusions

- The f1-score is highest for k-Nearest Neighbor at 0.75 …
- But it also performs poorly in the precision of 1 at 0.08
- Decision Tree has a more balanced precision for 0 and 1
- The Logistic Regression is more balanced when it comes to recall of 0 and 1

→ Both Decision Tree and Logistig Regression models can be used side by side for the best performance