Car accident severity in Canada

Applied Data Science Capstone

Introduction – the problem

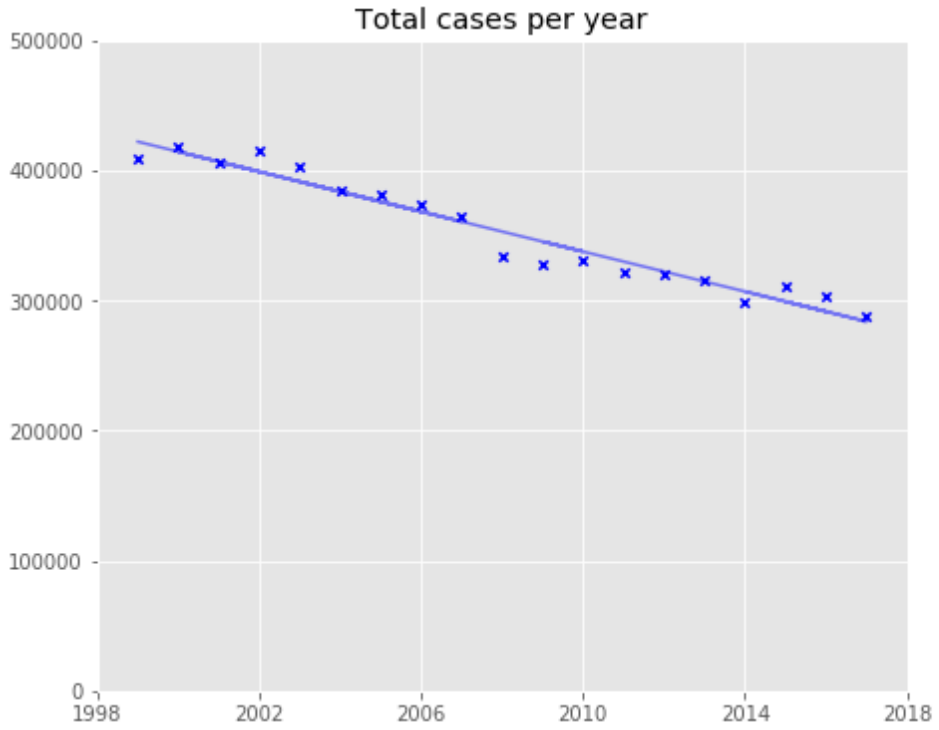
- On each day, thousands of people die at unnecessary car accidents
- Smart investments and clever decisions to reduce accidents require forecast models
- What are the main factors causing severe car accidents?
- Can we use **ML methods to predict the severity** of car accidents, based on factors such as weather or day time?

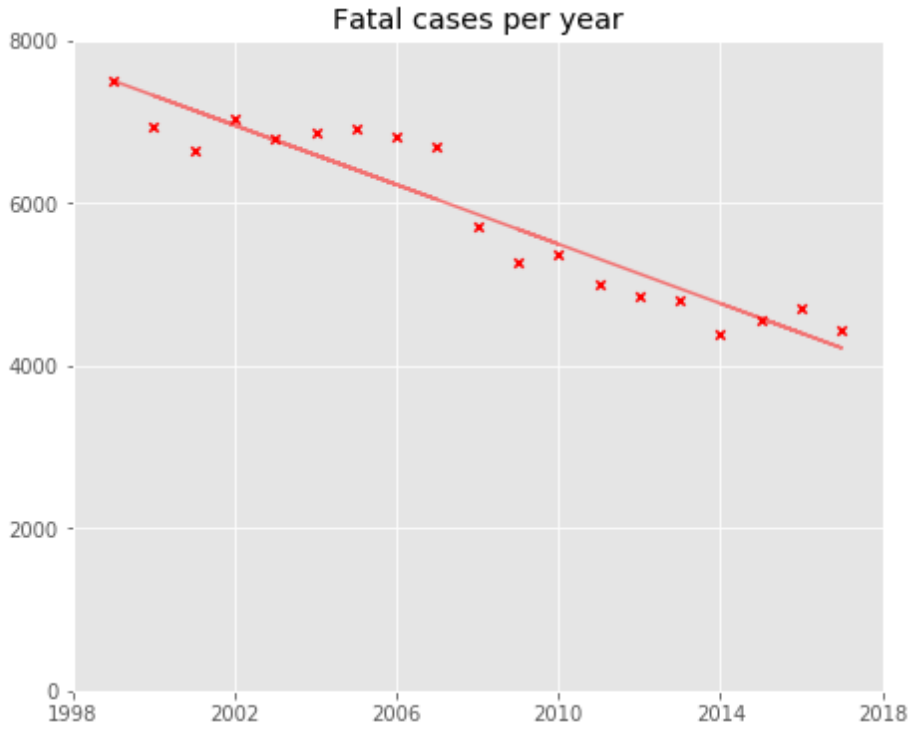
Data source

- National Collision Database (NCDB) of Canada, 1999-2017
- Almost 7 Million records
- Only 2 severity categories: fatal / not fatal
- Data set highly imbalanced: only 1.69 % of records are fatal

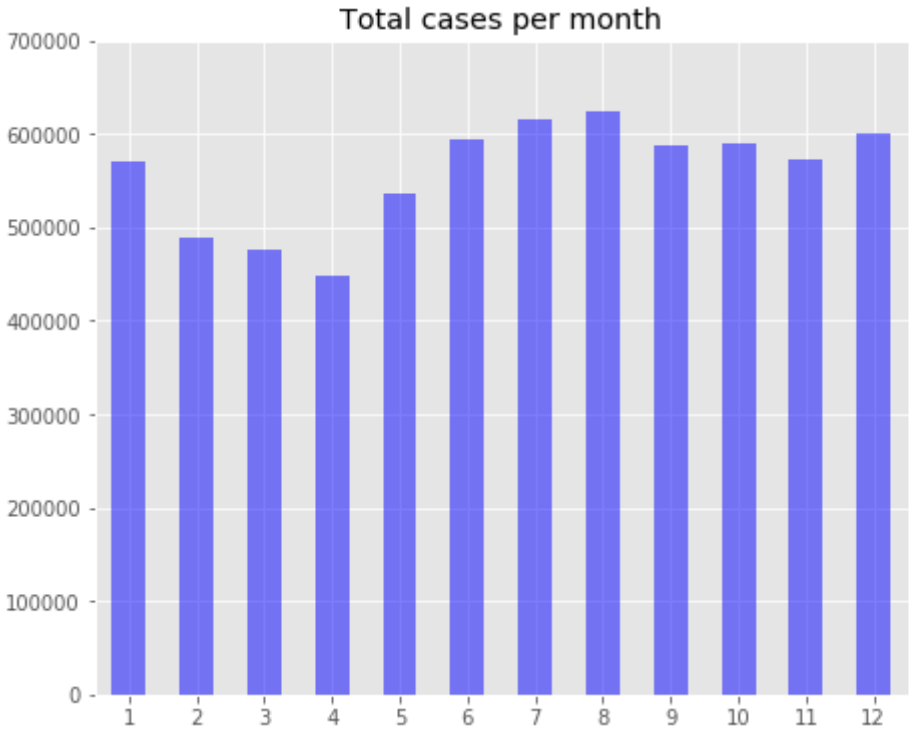
	C_YEAR	C_MNTH	C_WDAY	C_HOUR	C_SEV	C_VEHS	C_CONF	C_RCFG	C_WTHR	C_RSUR	C_RALN	C_TRAF	C_CASE
0	1999	1	1	20	2	02	34	UU	1	5	3	03	752
1	1999	1	1	20	2	02	34	UU	1	5	3	03	752
2	1999	1	1	20	2	02	34	UU	1	5	3	03	752
3	1999	1	1	08	2	01	01	UU	5	3	6	18	753
4	1999	1	1	08	2	01	01	UU	5	3	6	18	753
5	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
6	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
7	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
8	1999	1	1	17	2	03	QQ	QQ	1	2	1	01	820
9	1999	1	1	15	2	01	04	UU	1	5	U	UU	932

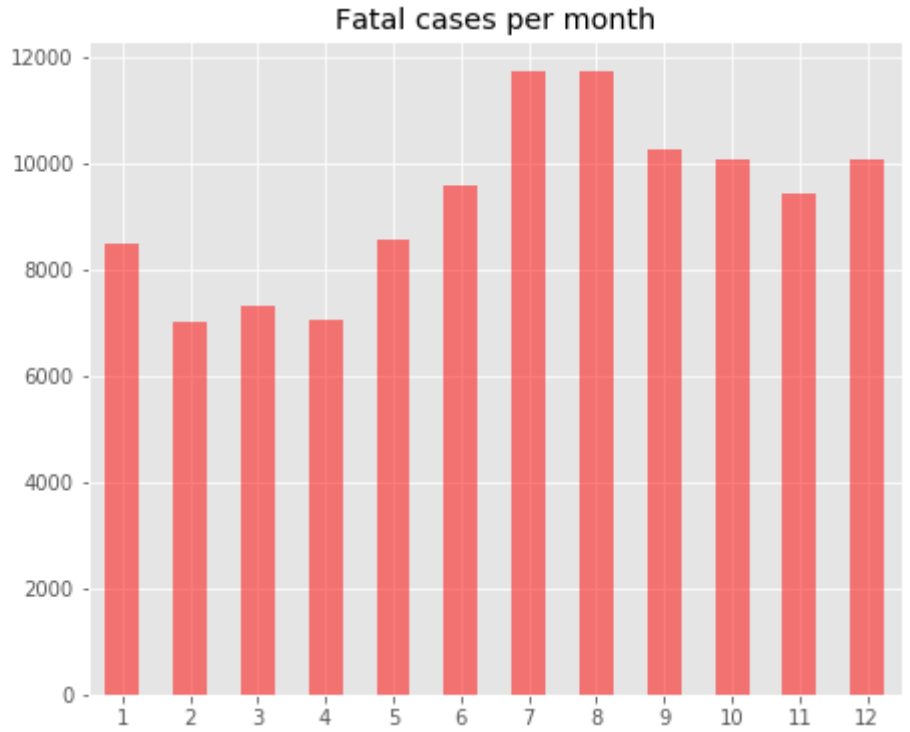
Car accidents per year



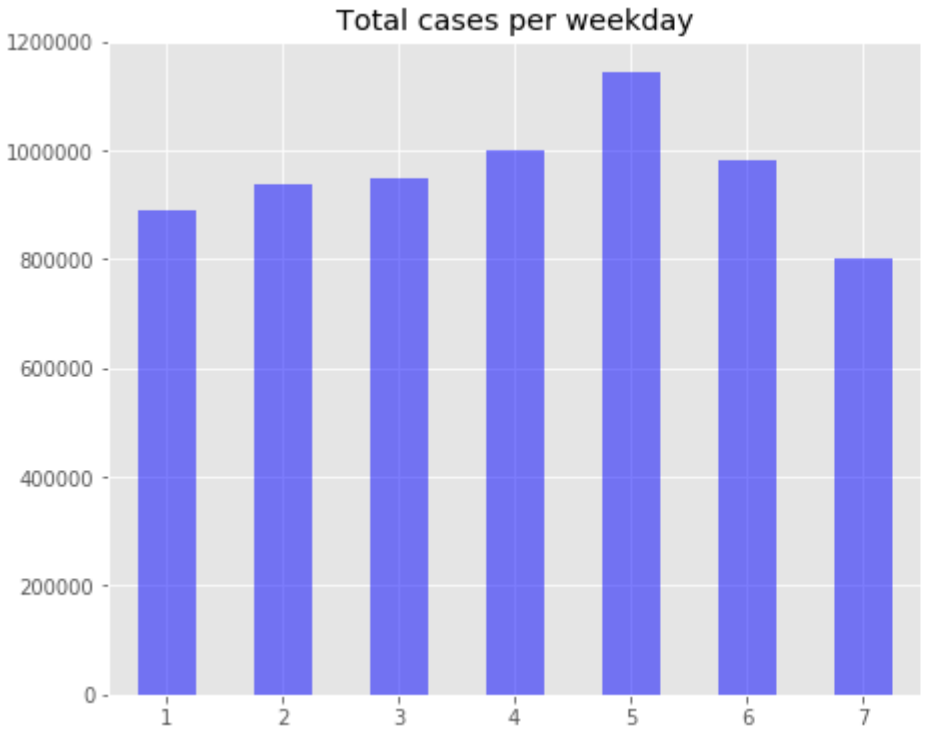


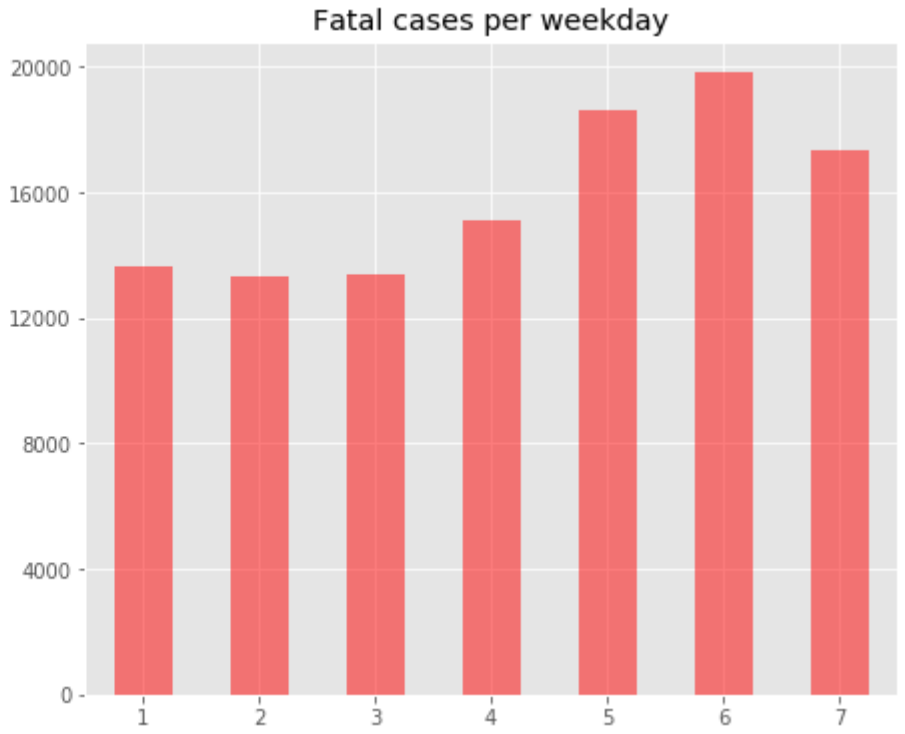
Car accidents per month



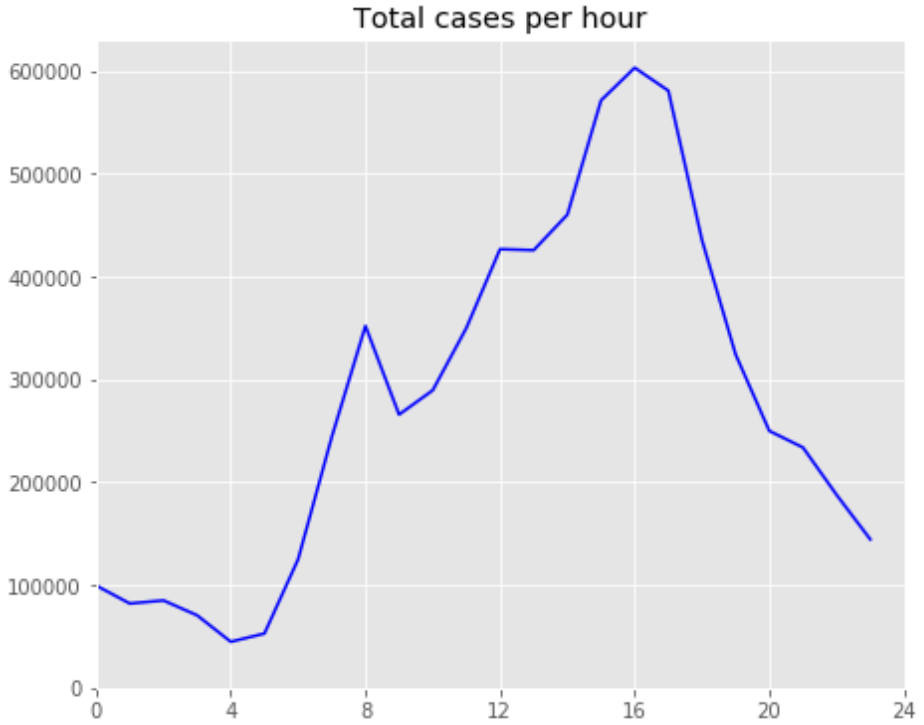


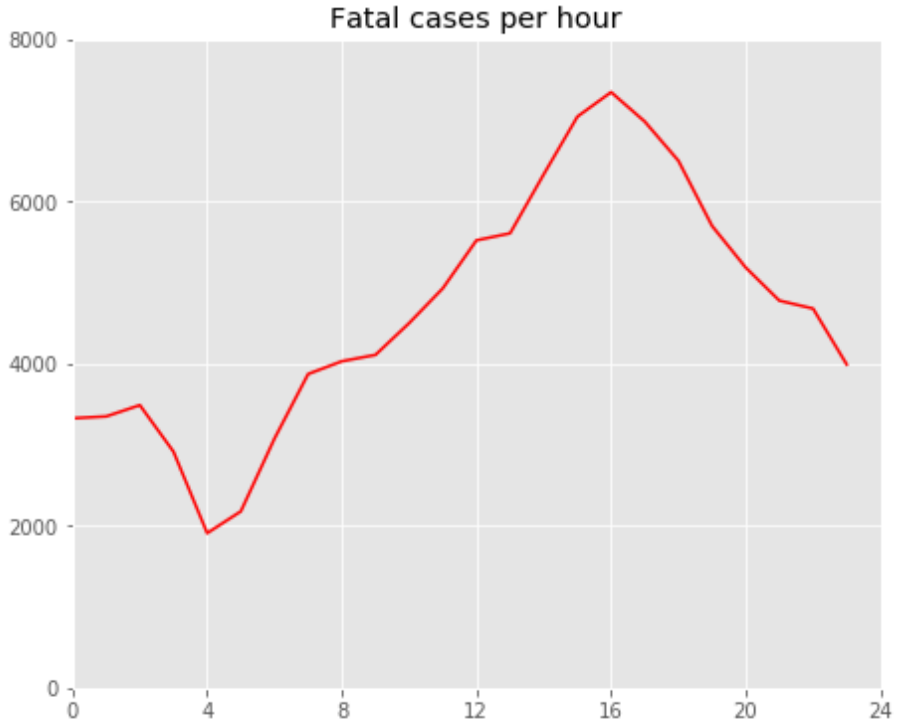
Car accidents per weekday



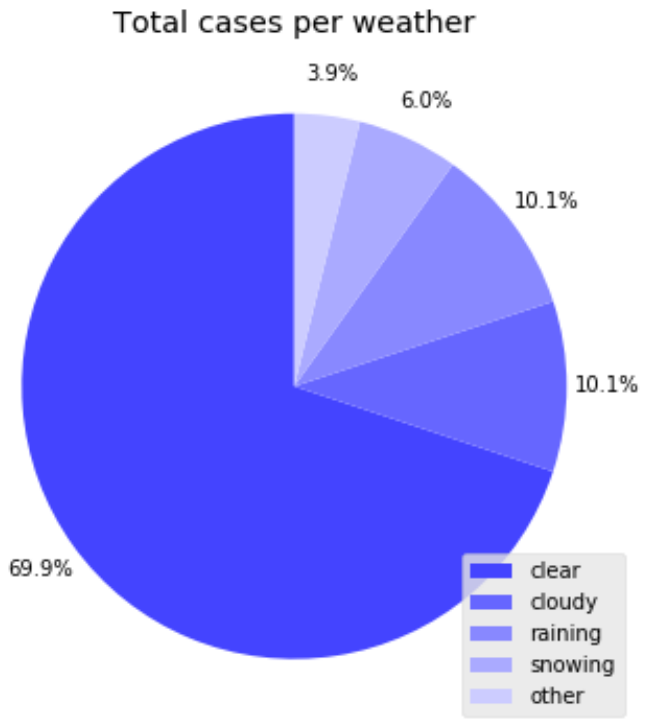


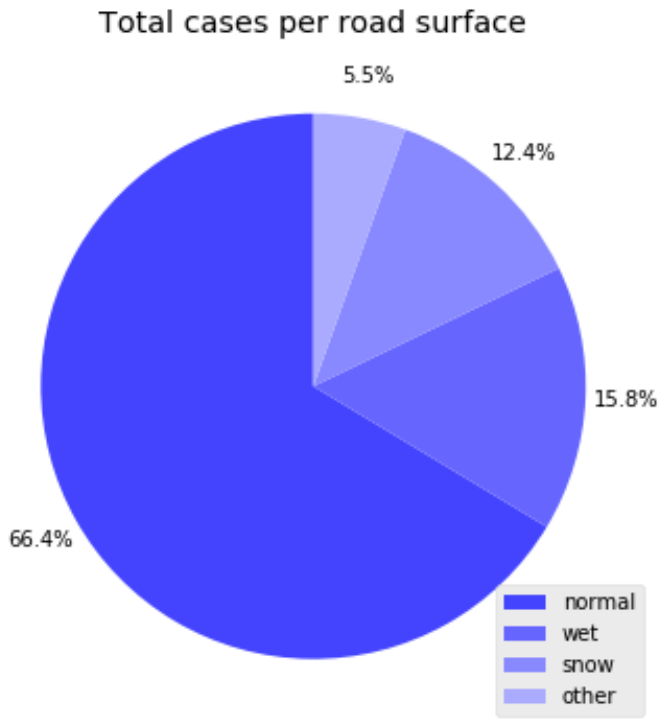
Car accidents per hour





Car accidents per weather and road surface





Data cleansing and preparation

- Removed many columns (e.g., personal data, traffic control)
- Removed rows with unknown date or time
- Cleaned data types and made column values consistent
- Simplified categories:
  - From 12 to 4 road surface categories
  - From 15 to 2 road configuration categories
- Applied **biased subsampling**:
  - Smaller data set to overcome memory issues
  - Achieved balanced data set

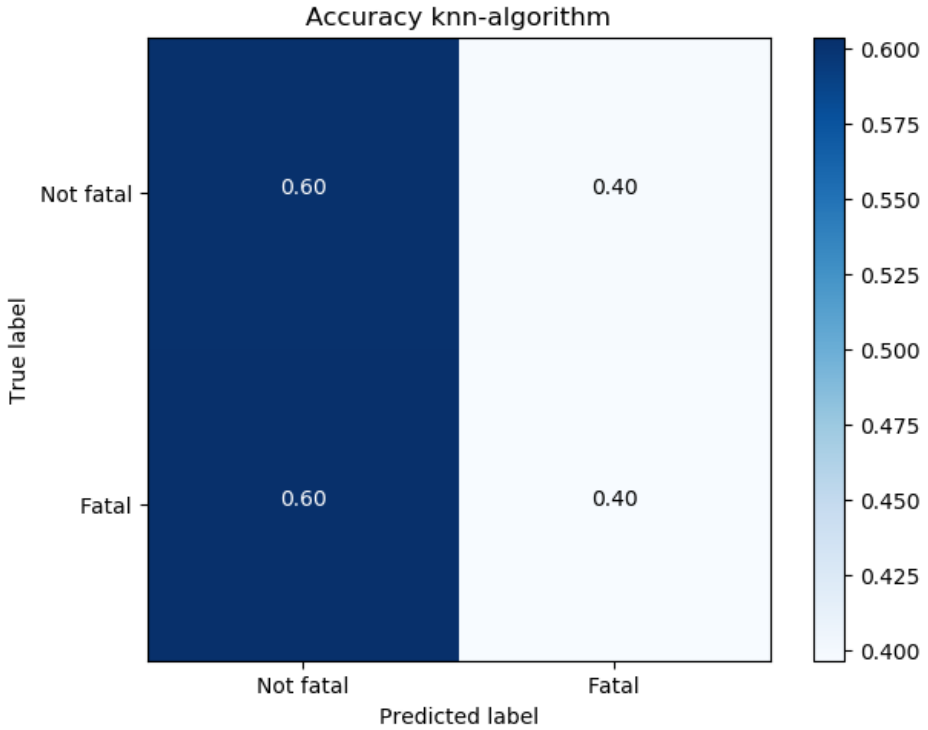
Cleaned data set

	Year	Month	Weekday	Hour	Fatal	Nbr Vehicles	Intersection	Weather	Road Surface
0	1999	1	1	20	0	2	0	Clear	snow
1	1999	1	1	20	0	2	0	Clear	snow
2	1999	1	1	20	0	2	0	Clear	snow
3	1999	1	1	8	0	1	0	Hail	snow
4	1999	1	1	8	0	1	0	Hail	snow
5	1999	1	1	17	0	3	0	Clear	wet
6	1999	1	1	17	0	3	0	Clear	wet
7	1999	1	1	17	0	3	0	Clear	wet
8	1999	1	1	17	0	3	0	Clear	wet
9	1999	1	1	15	0	1	0	Clear	snow

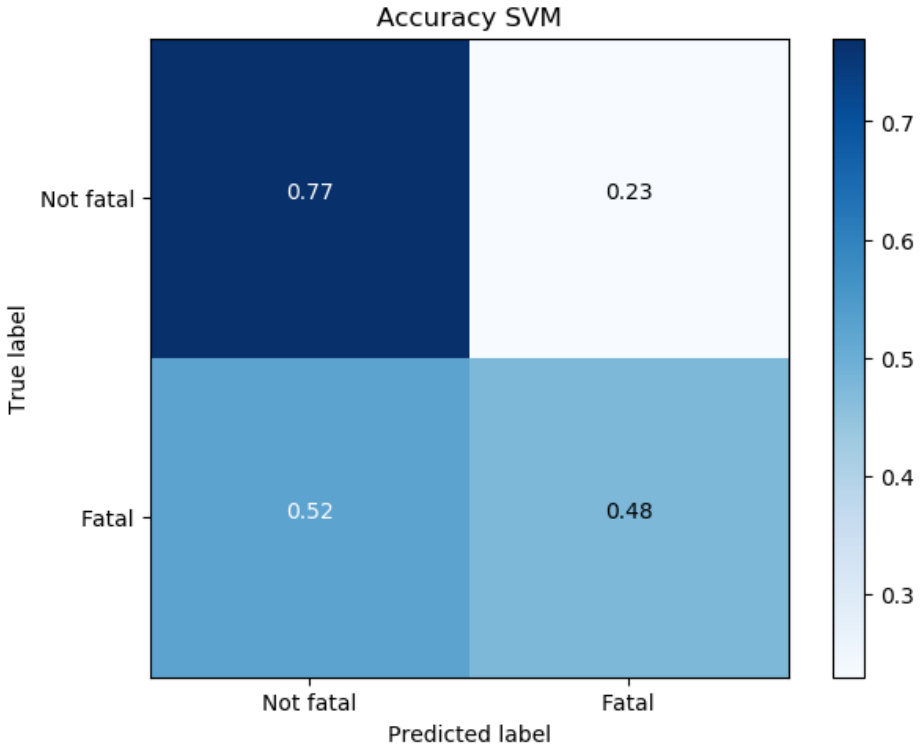
Methodology

- Used four different machine learning methods to predict severity of car accident
- 60% of data as training set
- 20% of data as evaluation set for tuning parameters
- 20% of data as test set

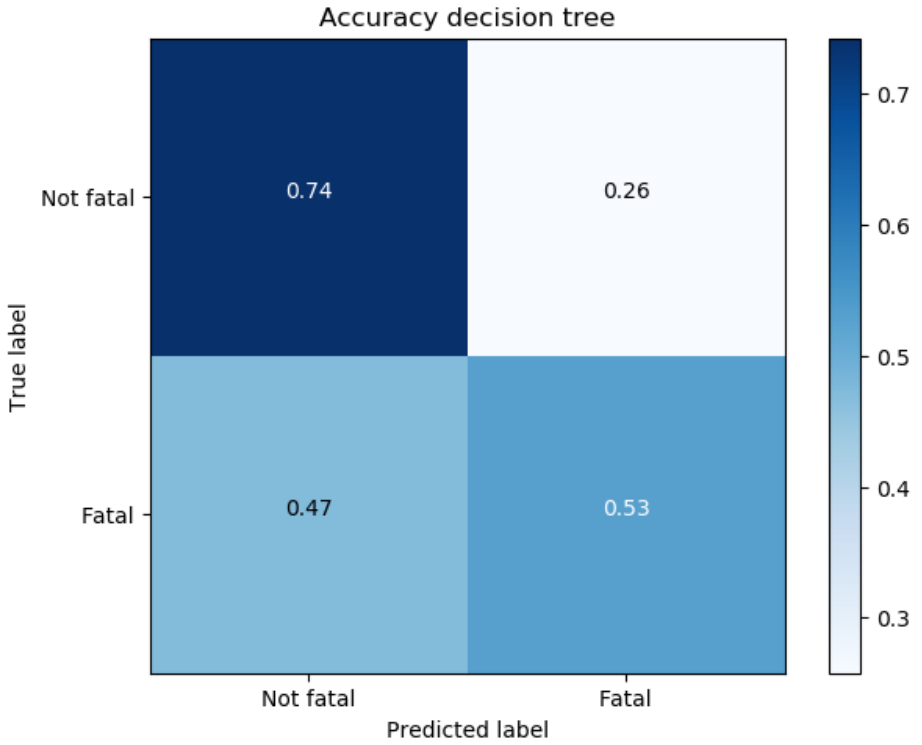
KNN-algorithm (k = 9)



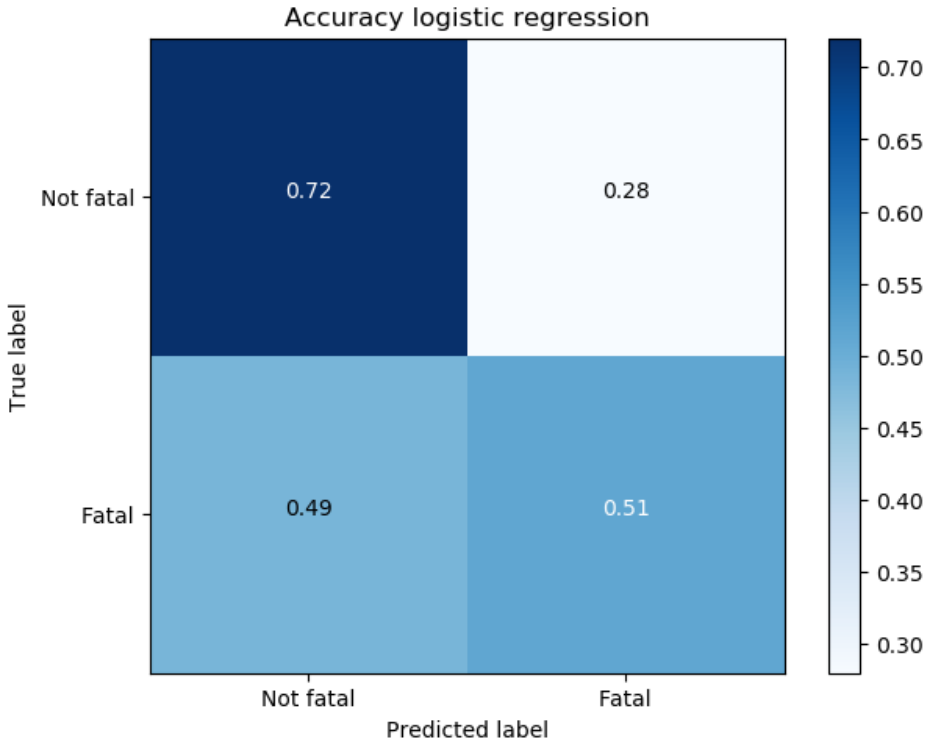
Support Vector Machines



Decision tree (depth 9)



Logistic regression



Discussion of results

- Poor accuracy of SVM and KNN
- Decision tree was the fastest and most accurate prediction
- Logistic regression achieved log-loss 0.64

- All 4 methods didn’t predict fatal cases convincingly

Interpretation

- There exist fatal and non-fatal cases with very similar conditions (weather, # cars, time, weekday)
  - For instance knn-algorithm can not distinguish
- Data seems to miss important features
  - Speed of vehicles
  - Consumed substances (alcohool)
  - Geographic location

Conclusion

- Decision tree is fastest and most accurate method
- Try out other models such as neuronal networks?
- Used data not sufficiently significant