Data Mining competition to understand causes for car incidents

Background

Real data of 50,000+ car incidents, covering traffic flow, demerits, registration information, criminal history.

Traffic police expect data scientist to present the causes of incidents, and also come up with suggestions of how to reduce incidents.

Proposal as a practise

Algorithm	
Clustering (K-means, SOM)	use heuristic methods to identify patterns that are previously unknown or unnoticeable
PCA	To reduce the dimensionality
Decision Tree/NN/SVM	Classify by the severity of incidents, to understand which factors are linked the casualty
Other Tools	**Function**
Neo4J	Network Analysis
Alteryx/SPSS/WEKA	Quick prototyping
H2O	Just to test the environment

Outcome

1. Understanding of the data
2. Explore auxiliary data: criminal history, weather, car type/safety rating, see how they can utilised (or not)
3. Suggest what data could be captured/acquired
4. Be wary of the difference between causal and correlate relationships
5. Use the analysed data to suggest how to reduce casualty/incident rates (e.g. defensive driving, regular servicing, warning signs @ high incident area, preventive reminders, alarm sensor for drink driving)

Future improvements

1. Algorithms
2. Collaboration
3. Optimising semi-labelled data mining