CMPE 255 Data Mining

Team: Crazy Spartans (Group 14)

Students: Shengtao Li, Zixin Li, Yi Lai

Team Project Proposal

1. **Title**

   UK road safety data analysis and accident prevention

2. **Project Description**

   Road traffic accidents are large problems that hurt public health and safety. Accident prevention is a critical challenge for not only the government but also the whole society. In our team project, we will use data mining algorithms and techniques to study the UK road safety dataset in order to find useful patterns for better accident prevention.

3. **Data Source Description**

   Our training data set and testing data set come from an open online platform provided by the United Kingdom government. From this website, we can get road safety data in recent 10 years. For each year: the dataset comprises three csv files:

   - Accidents(year): record all the reported features in each accident.
   - Vehicles(year): record all the reported features about the vehicle involving in an accident.
   - Casualties(year): record all the reported features about the casualties in each accident.

   We can combine these three csv files by the unique accident_Id.

4. **Proposed Methodologies:**

a. Data Preprocessing

   i. Accidents, vehicles, and casualties can be related by the unique accident_Id

b. Main Research Areas

   i. Car Safety Evaluation based on the Seriousness statistics

1. Combining the seriousness information from the accidents.cvs and car model information from the vehicles.csv to rank the safety of different car model. The year of cars is also being included in this analysis.

ii. Dangerous vehicle maneuvers analysis

1. Combining the Junction information from accidents.csv and maneuvers information from vehicles.csv to rank the dangerousness of each maneuver in certain junction type.

iii. High-risk driver population analysis

1. Combining the accident location and severity information from accidents.csv and age and sex of driver from vehicles.csv to study the correlation between age and sex to the type of accidents.

iv. Time interval analysis

1. Using the day of week and time information from accidents.csv to find out the correlation between the percentage of accidents and time intervals in the week.

v. Road condition analysis

1. Using road surface condition and light condition from accidents.csv to find out the correlation between the percentage of accidents and road conditions(dry & daylight, wet/damp & daylight, dry & darkness, wet/damp & darkness).

c. Data Processing

i. The Technologies we may use in our project include Data Cleaning, Dimensionality Reduction, Decision Tree Classification, Association Analysis, Cluster Analysis.

d. Testing

i. We would test the model from data processing using cross-validation method to assess the efficiency and accuracy of the model.