**Proposal Template**

**Team Name (Choose a name to represent your team)**

SIKD MARK

**Team Members**

201400944 산업경영공학과 김홍범

201803379 전자물리학과 정현희

201801271 전자물리학과 김효연

201803603 환경학과 최민영

**Project Objective**

The Clustreing aims to group accidents data in accordance to age of victim, number of deaths, road type, etc. and Algorithms such as K-means and DBSCAN will be used as analysis methods, and the dummy variable or preprocessing will be changed according to the algorithm.



**Project Scope**

Classification – To see the relationship between variables with the severity of the road accident, the casualty class in the data set is the target variable. The casualty class refers to the degree of injury to both the victim and the perpetrator. Available variables are the categorical data and need to choose some of them by using feature selection methods. Since the ages of victim and perpetrator are numerical values, they need to be discretized into intervals. After this preprocessing, I’ll use the information gain method for the feature selection and select 5 attributes that are best ranked. The purpose is to create a decision tree with the selected features and evaluate them to create an optimal decision tree.

Clustering - The main data set to be used to determine the relevance of traffic accidents is Number of Deaths, reported injures, Age\_victim, Age\_perpetrator. Depending on the direction of the analysis, we will combine the number of injured and deaths to create a new data variable. Based on this, we will use k-means, hierarch methods algorithm. Because both algorithms use numerical variables, we will use dummy variables when analyze variables such as road conditions and vehicle types. The purpose of this algorithm is to identify the insights of traffic accidents and similarities between accidents.

Dataset Explanation

**2018 CheoIn-gu road accidents**: This dataset is about the accidents from last year on the CheoIn-gu. It shows accident information about where and when was the accident happen, how was the road conditions, and who was injured. The details on the Column Details section.

* Since original the dataset was in Korean, we changed the data into English and some unnecessary data (deleted reference number, address, street) has been deleted. Accident Date is separated into two columns, Accident Month and Accident Time.

Column details

Accident Month - categorical data

Accident Time - categorical data

Day of the week – categorical data

Casualty Class - categorical data. Categorized by death injured accident, serious injured accident, slightly injured accident, injury reported accident. Injury reported accident is the accident that does not count for the slightly injured accident.

Number of Deaths – numerical data

Number of serious injuries - numerical data

Number of slightly injuries - numerical data

Number of reported injuries – numerical data

Accident type – categorical data. Data that explains how the accident happened. Ex) car vs car, car vs people accident

Violation of the Law - categorical data. Ex) Signal violation, safety operation default, safety distance not secured, pedestrian protection obligation violation, cross-passing method violation, central line invader, illegal U-turn, speeding), lane violation, straight right turn obstruction, other

Road Surface condition - categorical data categorical (Data that explains that road surface condition the accident happened. Ex) dry road, wet road…

Weather Conditions - categorical data. Ex) sunny, cloud, etc.

Road Type - categorical data. Ex) intersection, single road…

Type of Vehicle\_perpetrator - categorical data. Ex) van, truck, bicycle and so on

Sex\_perpetrator - categorical data. male and female

Age of Casualty\_perpetrator - numerical

Casualty Severity\_perpetrator - categorical and consists of no injury, serious injury, slightly injury, reported injury and death,

Type of Vehicle\_Victim - categorical and means the type of vehicle like van, truck, bicycle and so on

Sex \_Victim -categorical and consists of male and female.

Age of Casualty\_Victim - numerical

Casualty Severity\_Victim - categorical and consists of no injury, serious injury, slightly injury, reported injury and death

**Role of team members:**

|  |  |
| --- | --- |
| Students | Tasks |
| 201400944 김홍범 | Do the Clustering Analysis |
| 201803379 정현희 | Do the regression Analysis |
| 201801271 김효연 | Do the Classification Analysis |
| 201803603 최민영 | Do the Association Analysis |

**References**

VÁCLAV JIROVSKÝ(2015), CLASSIFICATION OF ROAD ACCIDENTS FROM THE PERSPECTIVE OF VEHICLE SAFETY SYSTEMS

김정민(2015), 교통사고 데이터의 마이닝을 위한 연관규칙 학습기법과 서브그룹 발견기법의 비교