**Final Report**

**Team Name**

SKID MARK

**Team Members**

201400944 산업경영공학과 김홍범

201803379 전자물리학과 정현희

201801271 전자물리학과 김효연

201803603 환경학과 최민영

**Project Objective**

The purpose of the analysis is to create a model that can predict the casualty severity through the analysis and examine key indicators. The data can be used to determine the ideal value of compensation according to the Injury Severity of the customer, and to prepare for emergency treatment. Also if we know the type of accident which has high severity, we can prevent additional accidents.

**Project Scope**

We will analyze road accident data that occurred from January 2017 to December 2018 in cheoin-gu. As a pre-processing process, fields that are not related will be removed, and other data such as road surface conditions and road types will be added. Some fields of the data will be grouped and categorized.

**Role of team members**

|  |  |
| --- | --- |
| Students | Tasks |
| 201400944 김홍범 | Do the clustering analysis + Presentation |
| 201803379 정현희 | Do the regression analysis |
| 201801271 김효연 | Do the classification analysis |
| 201803603 최민영 | Do the association analysis |

**Data Preprocessing**

Dataset : 2017, 2018 cheoin-gu road accidents

Since the original dataset was in Korean, we changed the data into English. Some unnecessary data (deleted reference number, address, street) has been deleted. The original 23 columns have been reduced to 21, and our team will try 4 analysis with this processed data.



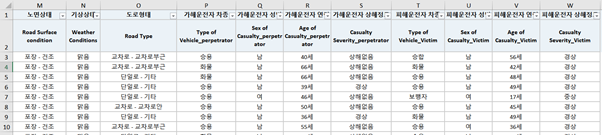
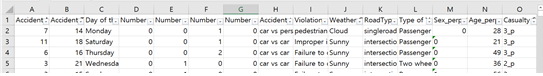


Fig 1, 2. Original Dataset(contains 1737 rows)



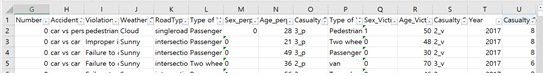


Fig 3,4. Preprocessed Dataset

h. Accident type: It has a large category and small category in original data, but we will use the large category only.

i. Violation of the Law: Since most of the obstruction of straight right-hand traffic is Improper intersection crossing, both of them are considered Improper intersection crossing. The rest are the same as the original data.

k. Road type: The details of a single road are grouped into a single road. Parking lots and etc. are grouped into etc. The rest are the same as the original data.

l,p. Type of vehicle\_perpetrator & victim: We convert the data to two wheel vehicle for motorcycle, motor and two-wheeled machine and to etc. for special vehicles, PM, other unknown and agricultural machine.

m,q. Sex\_perpetrator & victim: Sex is converted to 0 for males and 1 for females

o,s.casualty severity perpetrator & victim : The original data showed four categories in this section, but we changed it into three categories. 1 means fatal accident, 2 means seriously injured or slightly injured, and 3 means injury reported or no injury.

u. casualty class: Target variable is the casualty class, which means the degree of both victims and perpetrators injury. We've quantified it with three types of data, and divided it into 3 by 3 matrices.

t. Year: Since we added the 2017 data, it is necessary to identify which data is 2017 data and 2018 data. This data is only for convenience, it will be the meta data.

a,b,c, d, e, f, g, j.n,r Same as the original data.

**Data Visualization**

It is necessary to visualize the preprocess data with Orange tools. For the missing data, the entire row was deleted to improve accuracy, and it was 1% of the total data.

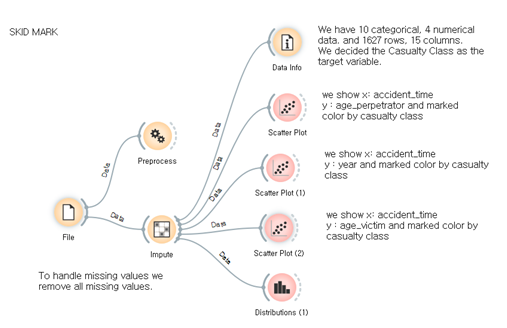


Fig 5, Data visualization

The number of cheoIn-gu road accidents is about 700cases in 2017 and about 900cases in 2018.

There were about 260 accidents in december and about 140 in other months.

There were less han 60 cases from 00:00 to 7:00, about 110 cases in 8:00, about 140 cases in 18:00, The rest of the time range is about 80 to 100 cases, except for 00:00 to 7:00, except 8:00 and 18:00.

On Friday and Wednesday, the accident rate is the highest, with about 270 accidents, about 240 accidents on Mondays and Thursdays, and about 160 accidents on Sundays.

There were about 800 Seriously injured accidents, 650 Slightly injured accidents, and about 100 Injury reported accidents and 30 fatal accidents.

There were 1,300 Car vs Car accidents, an overwhelming number, and about 300 Car vs Person accidents.

Failure to drive safely was the most Violation of the Law with about 1000 cases, and the frequency decreased as Improper intersection crossing, Disobedience of signals, Center line crossing, Pedestrian protection obligation violation, Failure to secure safe distance, Line violation, and Illegal U-turn, Speeding.

Weather Condition had the largest number of cases with about 1,500 in Sunny, and the frequency decreased as the rain cloud snow foggy.

There are single road, Intersection (within), Intersection (near), and Intersection (crosswalk) in road types. It occurs 750 accidents in a single road, 500 accidents in Intersection (within), 250 accidents in Intersection (near), and 50 accidents in Intersection (crosswalk).

Perpetrator drove more than 1000 passenger cars, 300 trucks, about 100 two-wheel vehicles, and about 50 bicycles. 1300 of perpetrators are males and 300 of perpetrators are females.

There are less than 20 perpetrators in teenagers, from 30 to 60 perpetrators in 20s, from 40 to 60 perpetrators in 30s, from 50 to 90 perpetrators in 40s, from 70 to 80 perpetrators in 50s, from 30 to 60 perpetrators in 60s, about 20 perpetrators in 70s, less than 10 perpetrators in 80s, and about 1 perpetrator in 90s. When the perpetrator is younger or older , the lower the accident rate.

Victim drove more than 800 passenger cars, 300 pedestrians, 200 trucks, 150 two-wheel vehicles, and 50 vans, 100 constructions and 100 bicycles. 1200 of Victims are males and 450 of Victims are females.

There are less than 10 Victims in under the age of ten, about 10 Victims in teenagers, from 20 to 40 Victims in 20s, from 50 to 80 Victims in 30s, from 60 to 90 perpetrators in 40s, from 70 to 80 Victims in 50s, from 20 to 40Victims in 60s, from 10 to 20 Victims in 70s, less than 20 Victims in 80s. The younger or older you are, the lower the accident rate.

**Data Analysis**

* Clustering

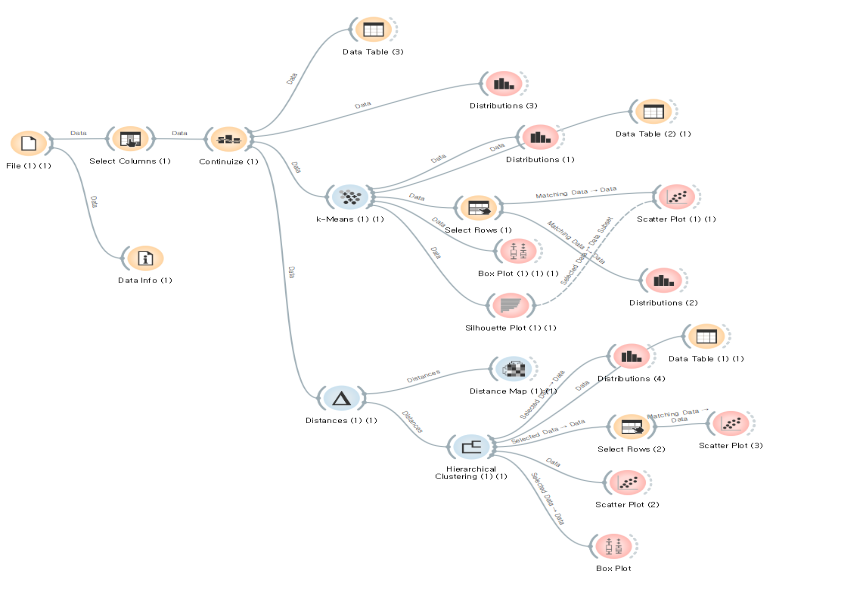
The aim of clustering is to find similarities between the severity of accidents and other features. The picture on the below shows the Orange file. For analysis, k-means and Hierarchical clustering were used. In addition, categorical data was converted to numeric during preprocessing. Also, the target variable is Casualty Class.

Fig 6, Clustering

The number of K -means clusters was determined to be 9 using the silhouette scores. Also, we practiced Hierarchical and pruning the clusters to nine. The picture shows the result of two clustering by using scatter plots. Based on the age of the victim and the perpetrator, Hierarchical was better clustered.

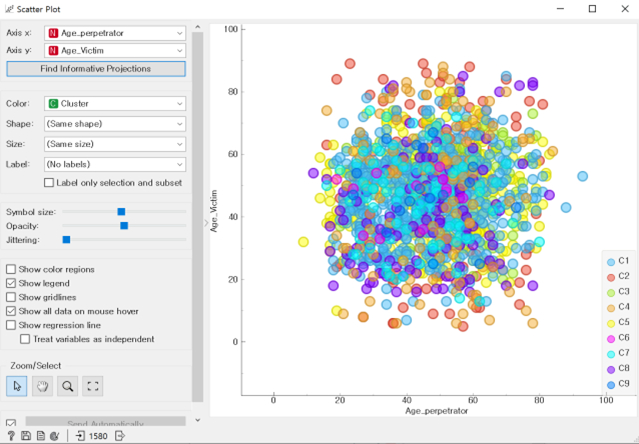
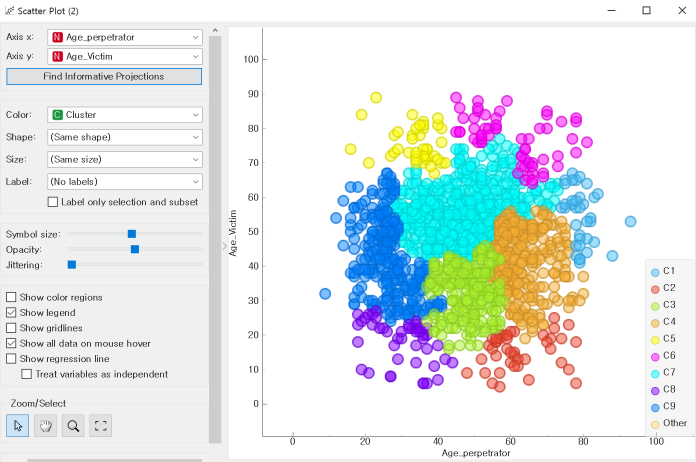


Fig 7, Scatter plot

We looked at two algorithms clustered based on the age of the perpetrator. As a result, while k-means do not make much difference, the Hierarchical algorithm shows that the C1 group is older than the other clusters.

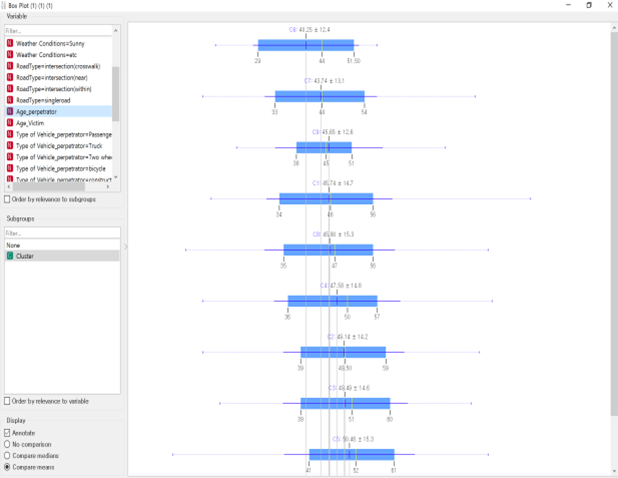
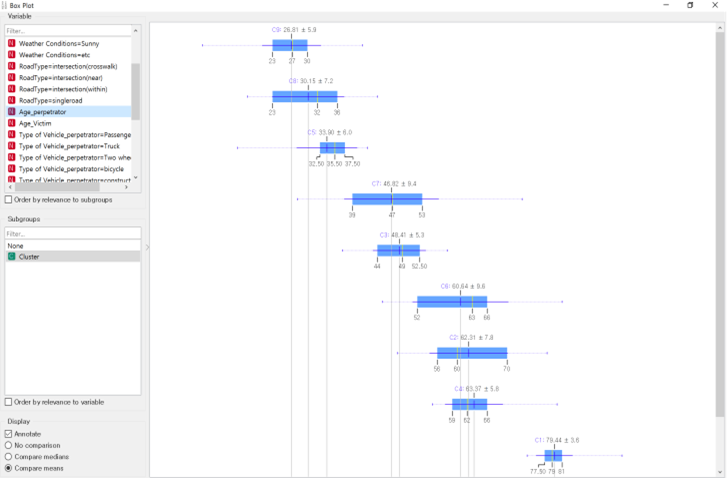


Fig 8, Box plot

* Association

In association rules, we tried to identify the rules of the conditions that occurred at the same time and to predict the severity of the accident depending on the condition.

Since the association rule is suitable for categorical data and the association's target implies the degree of injury of the victim and the perpetrator, categorizing and removing the victim and the perpetrator’s casualty class conditions is required. So we made additional preprocessing before starting the association analysis, and we chose the appropriate minsup and minconfi. When minsup is 0.1, minconf is 0.9, there are zero rules. And when minup is 0.1 and minconf is 0.8, there are 27 rules. However, all confidence of these rules did not exceed 0.9, so it was relatively small and there is only Casualty Class 8 in LHS. So we analyzed additional cases. When minsup is 0.01, minconf is 0.8, there are 721 rules. To reduce the number of rules, we turned the min values to minup 0.01, minconf 0.9 and found that there were 186 rules in the same way. but likewise there is only Casualty Class8 in LHS. when minsup is 0.01, minconf is 0.8, even if there were 721 rules but I can see the association with two classes. So I selected minsup 0.01, minconf 0.8.

When we run the orange program with this min it is difficult to see the result when there is only one condition in lhs. For example, I want to see the result of class6, so I put cas class6 in consequent, but It's recognized as a total casualty class. To make the analysis more clear and visualize the results, we selected R program

Before viewing the results in R. we limited the minsup, minconfi determined and limited RHS to leave only the severity of the injury which is a target.

At frequent itemsets, If you list the top6 in support order, The frequency was high in the order of clear weather, car-to-car accidents, male perpetrator, casualty class8, male victims, and perpaetraor’s passenger cars. (Fig1)

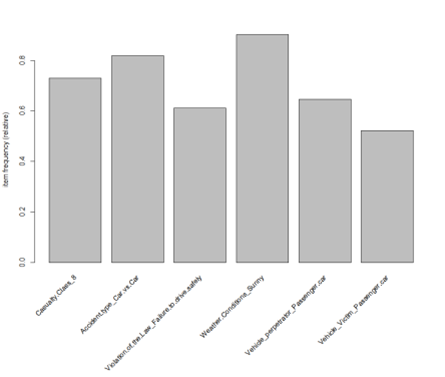


Fig 9,Frequent result

at association rules, the minsupport was set at 0.8 and min confidence at 0.01. As a result, 721 rules resulted and those rules had two kinds of lhs. There are only 3rules with casualty class 6 and about the class8, I listed them by lift and considered only the top 5 rules. As a result, When the severity of injury is Serious Perpetrator, There were Perpetrator Vehicles \_ bicycles & Car and Car, Perpetrator Vehicles \_ bicycles , Perpetrator Vehicles \_ bicycles & Sunny in lhs of rules.(Fig2) and When the severity of injury is Normal Perpetrator, Serious Victim, There were Autumn & Intersection/Crosswalk, Autumn & Violation of Laws, Monday & Victims in 70s, Violation of line crossing & Victims Vehicles \_ two wheeled, Monday Victims Vehicles \_ two wheeled in lhs of rules.(Fig3)



Fig 10,Association rules\_6



Fig 11,Association rules\_8

* Classification

The classification aims to identify patterns with the severity of the road accident and other categorized classes. The target variable is the Casualty Class, and the Casualty Severity will not be considered as features because they are the characteristic of the casualty class. After the preprocessing, 1554 instances were sampled with a fixed proportion, which is 80% of the total data, and were analyzed by five different algorithms. Those all can handle categorized data.

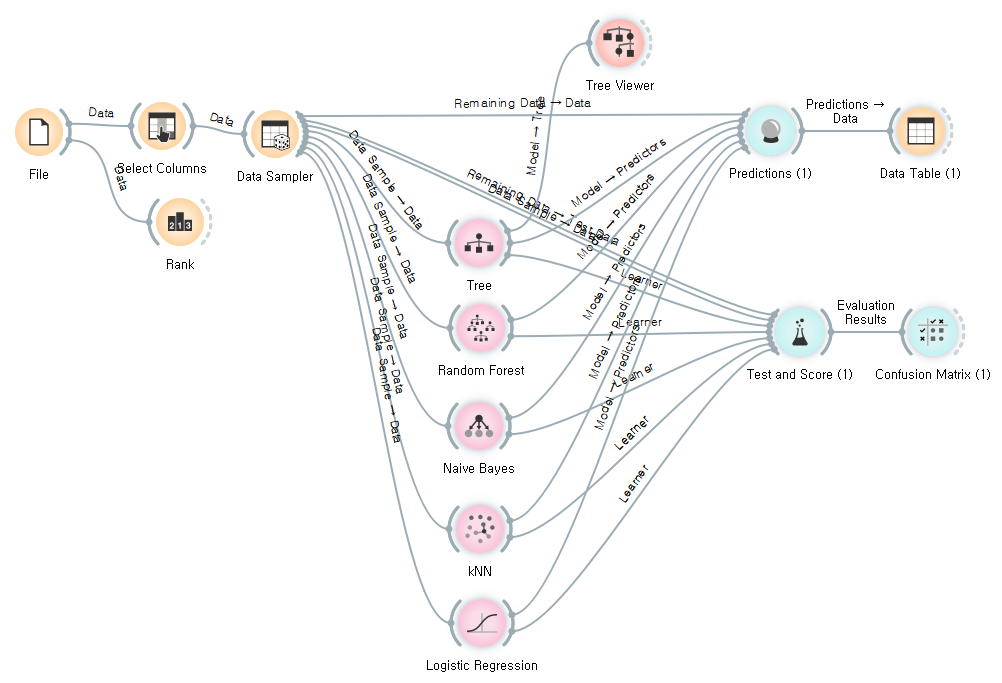


Fig 12. Evaluation of the five algorithms in Orange data mining

The picture above shows the screen of the Orange file. To determine which learner is the most suitable for modeling, the Predictions and the Test and Scores are used. In the prediction, shows models’ predictions on the data, the classification accuracy of the random forest was the highest, followed by the tree model. By using the remaining data, the accuracy of the model will be estimated. The train data and test data are evaluated in the test and score.

Next picture shows the test on train data.

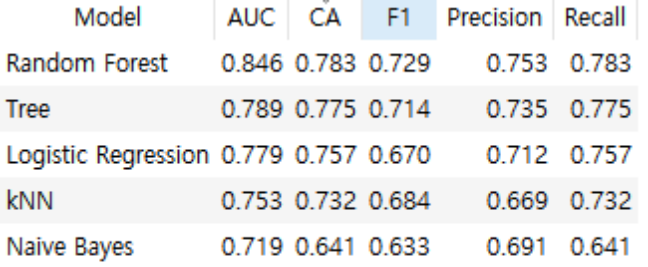


Fig 13. Model evaluation – Test on train data

On the other hand, the evaluation result of test data shows that logistic regression is the most accurate one.

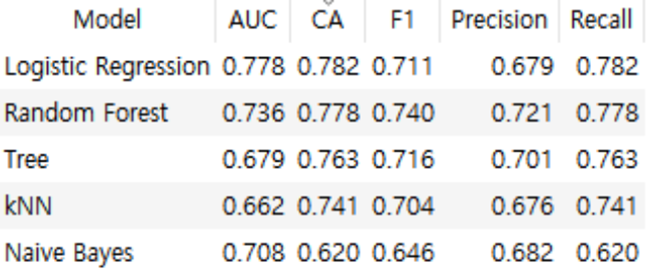


Fig 14.. Model evaluation - Test on test data

Since the original data is imbalanced, it is necessary to use the F1 score on evaluating the models. The most common technique to oversample a dataset is known as SMOTE. SMOTE-NC can handle a mix of categorical and continuous features, but it needs at least one continuous feature. In the original data, the only continuous value is the age column.

* Regression

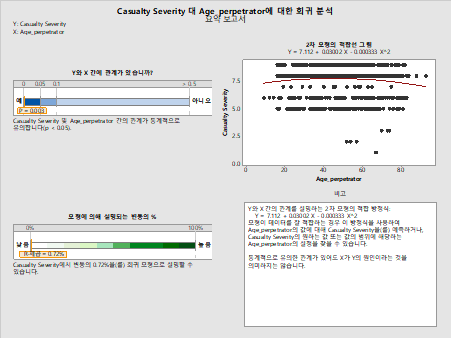


Fig 15, Summary report\_1

H0: 'Casualty Severity’ and ‘age\_perpetrator’ are irrelevant.

H1: 'Casualty Severity’ and ‘age\_perpetrator’ are related.

Y=7.112+0.03002X-0.000333X^2

The regression equation is a quadratic function. Therefore, as the ‘age\_perpetrator’ value increases, the 'Casualty Severity’ value increases or decreases

The P value is 0.003, which means that the slope (regression coefficient) 'Casualty Severity’ and ‘age\_perpetrator’ is significant. The R square is 0.72%, which is a little lower than the model fit.

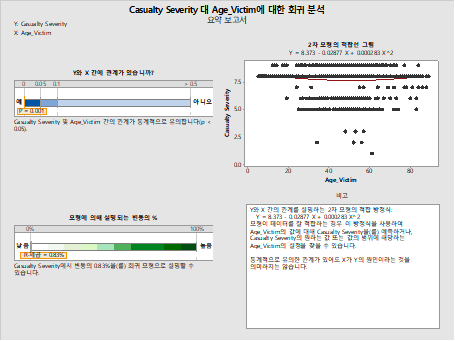


Fig 16, Summary report\_2

H0: 'Casualty Severity’ and ‘age\_victim’ are irrelevant.

H1: 'Casualty Severity’ and ‘age\_ victim’ are related.

Y=8.373+0.02877X-0.000283X^2

The regression equation is a quadratic function. Therefore, as the ‘age\_ victim’ value increases, the 'Casualty Severity’ value increases or decreases

The P value is 0.001, which means that the slope (regression coefficient) 'Casualty Severity’ and ‘age\_ victim’ is significant. The R square is 0.83%, which is a little lower than the model fit.

This regression analysis shows that the p value is less than 0.05, so statistically significant however the R2 value is very small, it means explanatory power is low. This means that the injury severity has little relation with the age of the victim and the perpetrator. This data has few numerical variables, making it difficult to analyze with other variables.

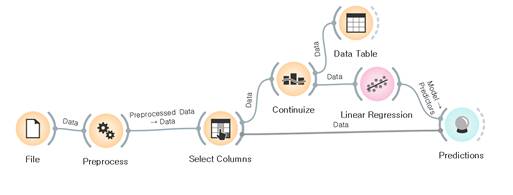


Fig 17, Orange

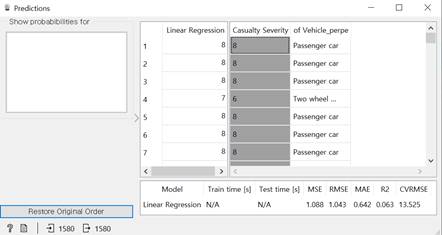
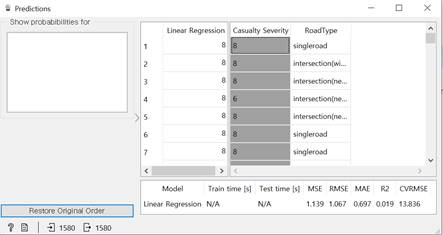
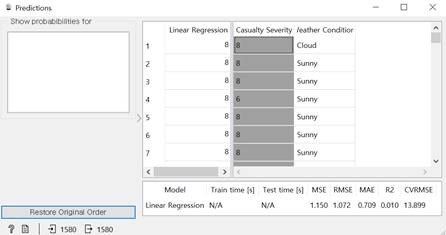
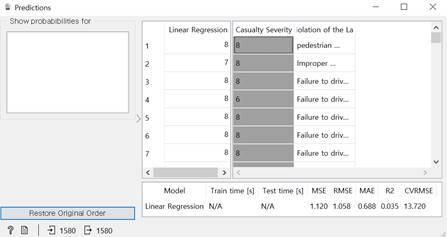
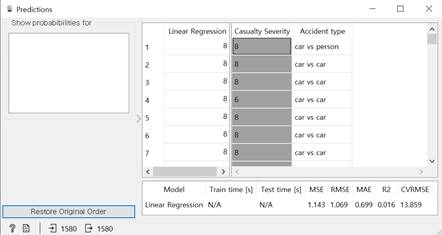
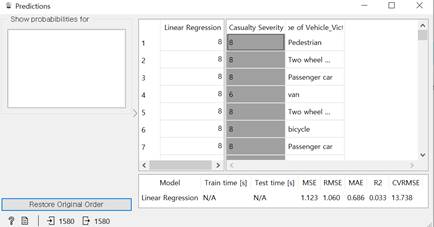
 

Fig 18,19,20,21,22,23 Prediction

This data has few numerical variables, so we changed categorical data to numerical values. The result of analysis, the R2 value is very small, it means explanatory power is low. This means that the injury severity has little relation with other variables.

Accident type’s R2 is 0.016. Violation of the Law’s R2 is 0.035.

Weather Conditions’ R2 is 0.010. Road Type’s R2 is 0.019.

Type of Vehicle\_perpetrator’s R2 is 0.063. Type of Vehicle\_Victim’s R2 is 0.033.

There are many factors in traffic accidents associated with the injury severity, a better analysis is expected if there are many numerical or categorical variables in other data sets.

**Conclusion**

Through clustering analysis, the hierarchical method was more effectively clustered than k-means, and we can identify the characteristics of the cluster. In association rules, the lift is the highest when the RHS is Perpetrator vehicle is bicycles and Cars and cars accidents, it is likely that the casualty class is 6. When the RHS is Autumn and the intersection is crosswalk, it is likely that the casualty class is 8. In classification, the most important feature was the type of perpetrator’s vehicle. Random forest model shows the smallest gap between test on test data and test on train data.

Since it is best to cluster the data into 9, it was an appropriate choice to set the target injury severity as 9 variables. This severity is highly influenced by perpetrator's vehicle, especially if perpetrator's vehicle is a cycle, analysis showed that perpetrator is more likely to be seriously injured.

Consequently, the analysis helping to predict the casualty severity can be used to determine the ideal value of compensation according to the Injury Severity of the customer, and to prepare for emergency treatment.

**Limits**

After looking at the analysis, we could find some limitations.

First, Because data is based on the target variable, classification is imbalanced. To supplement that, we can solve the data using python code. The most common technique to oversample a dataset is known as SMOTE. SMOTE-NC can handle a mix of categorical and continuous features, but it needs at least one continuous feature.

and about clustering, there are a lot of clusters. so it is hard to understand

Since there are nine variables in the target variable, the number of clusters was also set at nine and clustering was carried out. As a result, interpretation of the results was not easy, and there will be ways to supplement them, such as using domain knowledge.

About association, even if the rules contain some high frequency condition in RHS. they can not associate with casu class. because it can be just coincidence. The interpretation was often meaningless. For example, Sunny accounts for most of the weather. In other words, the rules often show that there is a Sunny condition in the RHS, which is a meaningless association. To supplement that, The rhs of the rule excluded the condition car & car, Sunny, and passenger car, which account for the majority of each condition. Also when the casu class in lhs, the max sup is 18.4%, so after constrainting(ex, min conf) the supports became smaller. the sup means the percent of total, so if the sup is too small. it can also considered as coincidence

**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