FINAL PRESENTATION

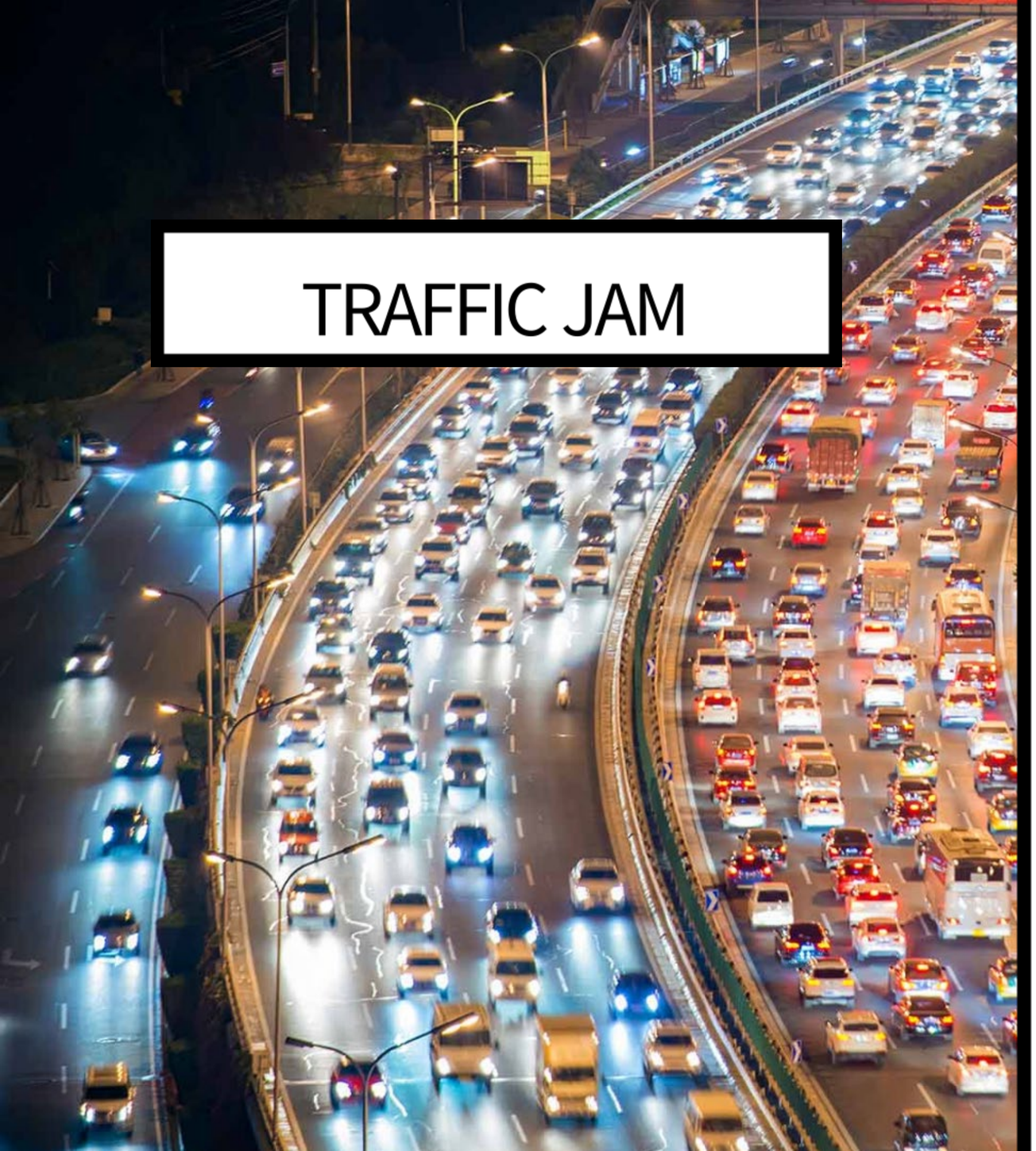
17102044 Jeewon Kim
18102093 Hwido Jung
20102119 Boyoung Lim

# DATA MINING

# CONTENTS

# BACKGROUND & PURPOSE

# 01




# BACKGROUND & PURPOSE

**1. TRAFFIC INCREASE**

Traffic levels are growing but the available road network capacity and the land capacity are limited.

**2. THE DIFFICULTY TO PREDICT TRAFFIC FLOW**

It is difficult to predict traffic flow because it involves data having a vast dimension.

=> If we can predict and manage the traffic congestion, we can give proper information to people

# OVERSEAS CASE | 02

# CASE OF INDIA

## THE TRAFFIC FLOW PREDICTION FOR INTELLIGEN T TRANSPORTATION SYSTEM

# DATASET

1. Collected in every 5 minutes from the application they developed with GPS technology

Location, direction, speed and start-end junction

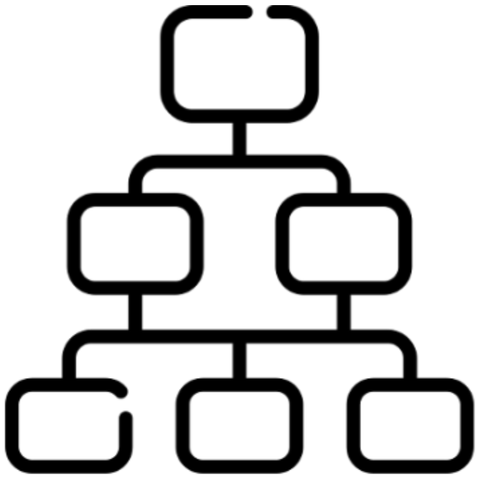
2. Grouping every 5 minutes interval with their corresponding data

3. With distance between each vehicle, the traffic congestion is identified.

- If distance < threshold, two vehicles are considered to be the neighbourhood.
- As the number of neighbourhood vehicles increases over time intervals, the traffic congestion is identified.

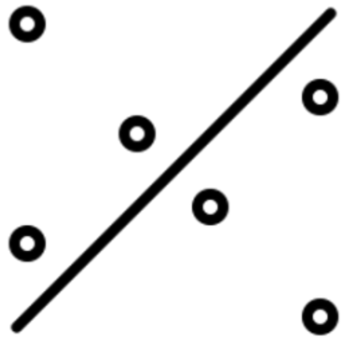

Fig. 2. Screen shot of the application

# METHODS



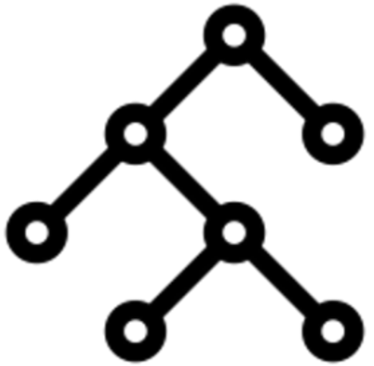
**Decision Tree**

- To predict the value of the target variables



**Support Vector Machine**

- To detect outliers
- Beneficial for high dimensional spaces



**Random Forest**

- To classify data

# METHODS

**Algorithm 1** For identifying the congested situation

1. Collect the traffic data in every 5 min with features:
   A. Location (Measured with GPS)
   B. Direction
   C. Speed
   D. Start-End Junction
2. Group every 5 min interval with their corresponding data.
3. Calculate the distance between each vehicle with all another vehicles within specified junction.

**if** the distance is less than the specific threshold between two vehicles **then**
  those vehicles are considered to be the neighbourhood vehicles
**else**
  Not considered as neighbour vehicles.
**end if**

**Algorithm 2** For classifying the congested situation

1. This will eventually give us the matrix A.
2. Now assign 1 to $A[i, j]$

**if** $A[i, j] < threshold$ **then**
  $A[i, j] = 1$
**else**
  $A[i, j] = 0$
**end if**

3. Count $A[i, j]$=1 and label $i, j$ as neighbourhood vehicles
4. Repeat above steps in every 5 min for 45 min
5. Plot the graph between neighbourhood vehicles and time interval.

**if** the neighbourhood vehicles shows an increasing graph **then**
  the traffic congestion is identified
**else**
  No traffic
**end if**

Steps involved in implementation

1. Created the application which can provide us the GPS coordinates.
2. Perform the proposed algorithm
3. Evaluate the matrix for the dataset
4. Divide the the dataset into training and testing.
5. Analyse different machine learning algorithms.
6. Predict the 45 min interval parameters through machine learning algorithm
7. Conclude about the traffic congestion

# RESULT

TABLE I
EVALUATION MATRIX FOR DIFFERENT MACHINE LEARNING ALGORITHMS

| Algorithm | Accuracy | Precision | Recall | Time |
|---|---|---|---|---|
| Decision Tree | 88% | 88.56% | 82% | 108.4sec |
| SVM | 88% | 87.88% | 80% | 94.1sec |
| Random Forest | 91% | 88.88% | 82% | 110.1sec |

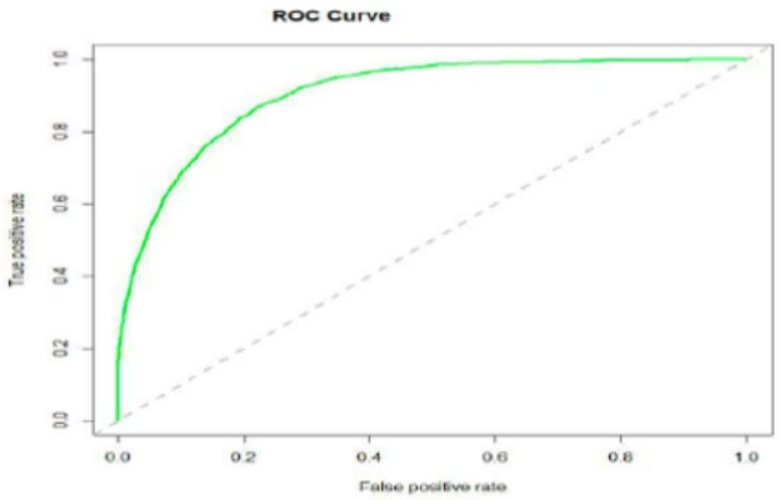



Fig. 3. ROC curve for Decision Tree Algorithm

The above steps we can implement this algorithm and can obtain the model which gives the higher accuracy of the machine learning model than the existing ones.

# LIMITATION

- The application’s measurement may not be accurate.

- Not enough features : Insufficient to apply in actual traffic environment. If they consider more features and apply the neural network, they could get a better prediction of traffic congestion.

# DOMESTIC CASE 1 | 03

# CASE OF KOREA

## Case 1 : TRAFFIC CONGESTION WITH TRAFFIC MANAGEMENT

# DATASET

1. VDS

- **Traffic and speed.** The number of congestion points can be determined based on the speed.

2. DSRC

- **Average travel time.** Converting the travel time to a speed allows the number of congested sections to be identified like VDS.

3. The change in DSRC traffic speed according to the change in VDS traffic volume aggregated every 5 minutes was analyzed.

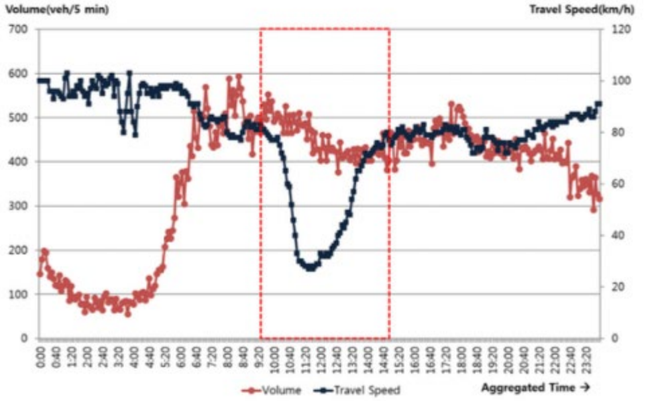



Fig. 1 Relationship of Volume and Travel Speed (Suwon IC~Giheung IC)

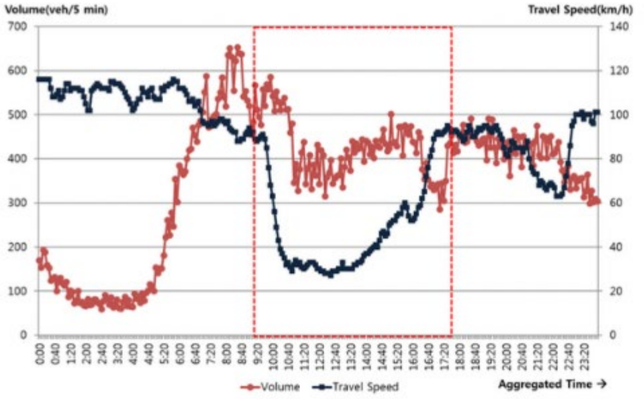
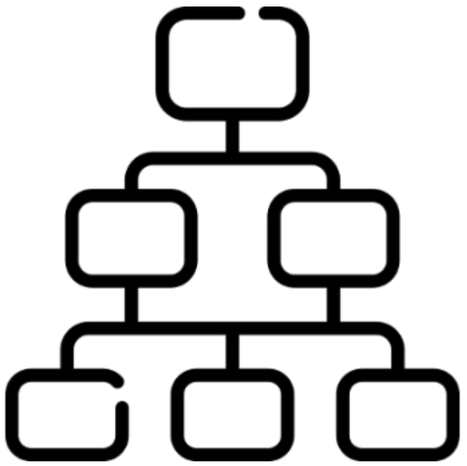



Fig. 2 Relationship of Volume and Travel Speed (Giheung IC~Osan IC)

# METHODS



## Decision Tree

Setting input variables in the decision tree m
odel by correlating changes in interval speed
with changes in congestion points and numb
er of intervals.

# METHODS

Table 1. Result of Correlation Analysis

| Conzone | # of VDS CP vs. Travel Speed | # of DSRC CL vs. Travel Speed | # of DSRC CL vs. # of VDS CP |
|---|---|---|---|
| Seoul TG ~Shingal JC | −0.69 | −0.74 | 0.89 |
| Shingal JC ~Suwon IC | −0.68 | −0.75 | 0.88 |
| Suwon IC ~Giheung IC | −0.70 | −0.91 | 0.78 |
| Giheung IC ~Osan IC | −0.77 | −0.90 | 0.82 |
| Osan IC ~Anseong JC | −0.64 | −0.71 | 0.75 |
| Anseong JC ~Anseong IC | −0.85 | −0.92 | 0.91 |

Table 2. Input Variable by Applied Decision Tree

| Classification | Input Variable |
|---|---|
| VDS | - VDS Point Congested Status<br>- Change of VDS Congested Points<br>- Change of VDS Section Speed |
| DSRC | - DSRC Link Congested Status<br>- Change of DSRC Congested Links<br>- Change of DSRC Section Speed |

## Setting input variables

Dependent variable: DSRC traffic speed based on departure time

Independent variables: congestion status, 15-minute interval speed change, 15-minute congestion point (interval) change

## Setting up decision tree

2 decision trees along to classification

# METHODS

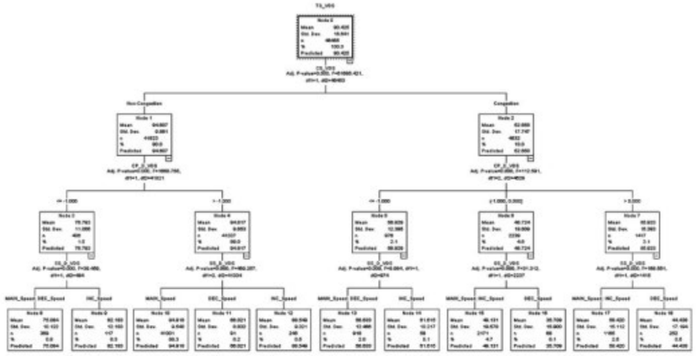


Fig. 5 Decision Tree Model using VDS Data

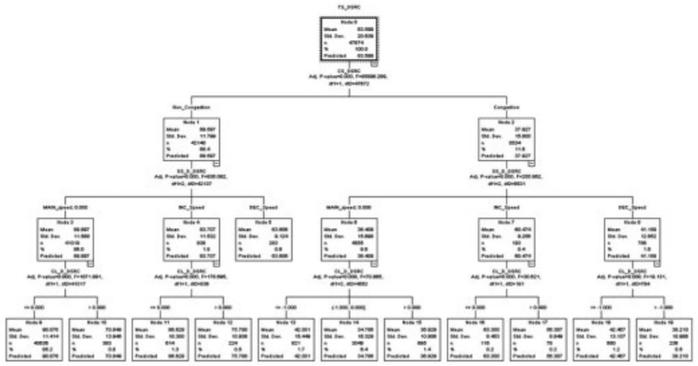


Fig. 6 Decision Tree Model using DSRC Data

As shown in Fig. 5, it was classified into a total of 11 situations in th e order of congestion, change in congestion index, and change in 1 5-minute interval speed.

The results of classifying the types of congestion situations based on DSRC collection data were also 11 situations, similar to those ba sed on VDS collection data.
However, as shown in Figure 6, there are differences classified in th e order of congestion status, change in the 15-minute interval spee d, and change in the number of congestion intervals.

# RESULT

Table 3. Accuracy by Applied Decision Tree Models

| Conzone | Congestion Status | VDS | DSRC |
|---|---|---|---|
| Seoul TG ~Shingal JC | Increase | 60.6% | 99.2% |
| | Maintain | 96.6% | 97.8% |
| | Decrease | 21.0% | 99.3% |
| Shingal JC ~Suwon IC | Increase | 47.6% | 100.0% |
| | Maintain | 95.1% | 97.7% |
| | Decrease | 2.4% | 99.4% |
| Suwon IC ~Giheung IC | Increase | 30.7% | 96.3% |
| | Maintain | 94.2% | 94.3% |
| | Decrease | 7.4% | 96.9% |
| Giheung IC ~Osan IC | Increase | 31.8% | 95.9% |
| | Maintain | 93.7% | 82.4% |
| | Decrease | 4.5% | 90.5% |
| Osan IC ~Anseong JC | Increase | 39.7% | 89.7% |
| | Maintain | 97.5% | 95.5% |
| | Decrease | 16.7% | 79.2% |
| Anseong JC ~Anseong IC | Increase | 31.0% | 98.6% |
| | Maintain | 95.5% | 97.3% |
| | Decrease | 1.6% | 99.2% |
| Total | Increase | 39.5% | 97.0% |
| | Maintain | 95.4% | 94.7% |
| | Decrease | 8.0% | 94.8% |

- As the number of congestion points and sections increases, the travel speed has a negative correlation that decreases.

- Accuracy of DSRC based > Accuracy of VDS based

- The actual section speed change was judged to be "increased" if it continued to increase for 15 minutes, "decreased" if it continued to decrease, and "maint ained" in other cases. The evaluation index was judged to be accurate, and the analysis results are shown in Table 3.

# LIMITATION

- It is set for some sections, and it is necessary to develop it using data from many sections when it is actually applied.

- Detailed standards should be prepared to improve the accuracy of determining changes in congestion situations.

- It will be necessary to develop a model that can determine the c hange in identity in specific situations such as unexpected situati ons.

# CASE OF KOREA

Case 2 : A traffic congestion prediction technique bas
ed on KNN using highway traffic Informtion

# DATASET

1. VDS(Interval speed per kilometer) & DSRC(3 to 4 km section speed)

- Since it is difficult to collect TCS data in the single section by time zone in local mountainous areas with low traffic volume, VDS data were mixed and used based on DSRC data.

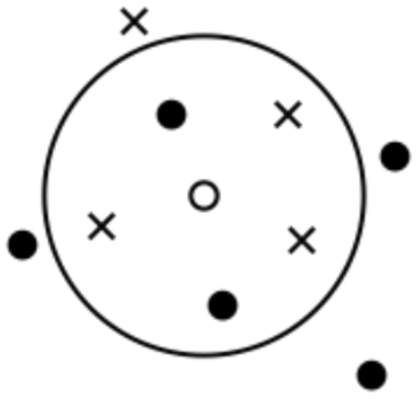
2. TCS (traffic volume)

- It is suitable for predicting the required time in large cities due to the characteristics of the section with a large change in the required time caused by various causes of congestion (vehicle increase, weather, accident, etc.).

[표 3-2] 예측에 사용되는 데이터 테이블

| 구분 | Table Name |
|---|---|
| TCS 1시간 이력 데이터 적재 | GET_ITCSTR_INOUTS_TRA |
| VDS 15분이력 데이터 적재 | SCREEN_VDS15MINTEXT |
| VDS 5분이력 데이터 적재 | SCREEN_VDS5MINTEXT |
| DSRC 5분 이력 데이터 적재 | GET_T_SMHS_DSRC_LINK5MIN_FTMS |
| TCS실시간 데이터 호출 | GET_ITCSTR_IOLANE_TRA_REA_T |
| VDS 15분실시간 데이터 호출 | EX_GETVDS15MIN_FTMS |
| VDS 5분실시간 데이터 호출 | EX_GETVDS5MIN_FTMS |
| DSRC 실시간 데이터 호출 | GET_T_RLRL_DSRC_5MIN |
| 기상 이력/실시간 데이터 호출 | GET_T_BSWT_WEATHER |

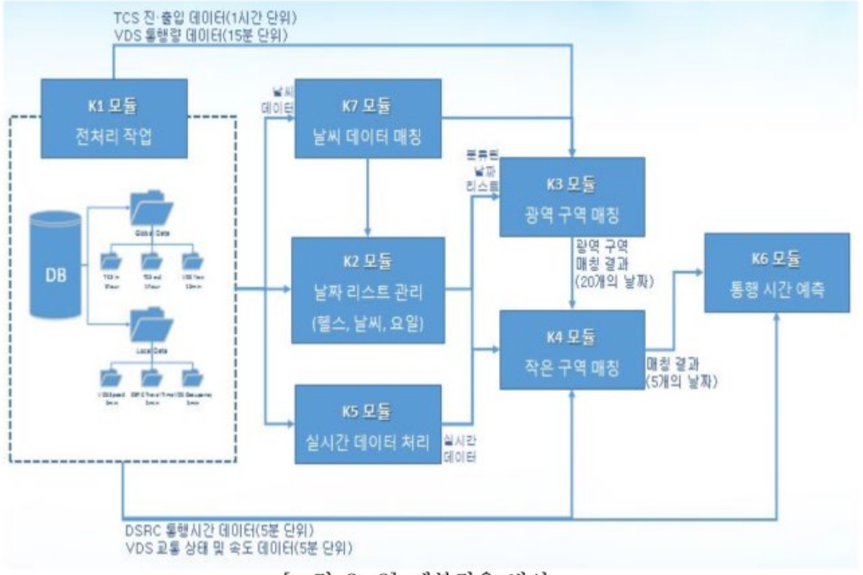
# METHODS



## K-Nearest Neighbors

After measuring the similarity based on the Euclidean distance score, the distance for the speed of the same section time zone is calculated and used.

# METHODS




[그림 3-2] 세부적용 방안

- The predicted date and the past dates with the same weat her conditions are selected.

- 20 to 30 dates are selected by judging the similarity of the traffic volume.

- By judging the similarity of VDS and DSRC data, the most s imilar dates are selected, and the travel time is predicted ba sed on the speed matrix of these dates.

# RESULT

[표 4-2] 신뢰도 비교 결과

| 구 분 | 교통예보 (현재시간 기준) | | | | | |
|---|---|---|---|---|---|---|
| | KNN | A사 | B사 | C사 | D사 | E사 |
| 평 균 | 91 | 87 | 86 | 86 | 86 | 82 |
| 3월 | 92 | 88 | 87 | 88 | 88 | - |
| 4월 | 91 | 87 | 85 | 86 | 84 | - |
| 5월 | 92 | 87 | 85 | 85 | 83 | - |
| 6월 | 90 | 89 | 87 | 87 | 89 | - |
| 7월 | 91 | 88 | 85 | 85 | 87 | - |
| 8월 | 91 | 86 | 85 | 86 | 84 | - |
| 9월 | 91 | - | - | - | 85 | 82 |
| 10월1주 | 90 | - | - | - | 89 | 85 |
| 10월3주 | 89 | - | - | - | 83 | 79 |

- Congestion by section was predicted by configuring a system using the R language for large-capacity data processing and DBMS, and based on this, the verification was based on the required time for each section, and the accuracy was 91%.

# LIMITATION

- It can be seen that the accuracy of future changes in the state of congestion is h igh because similar forms of historical data are found with the current state to be compared, which is the limitation of the KNN algorithm, but when congestion doe s not appear, the accuracy is relatively low.

- Accuracy can be increased together only when there is a guarantee that basic d ata such as DSRC, VDS, and TCS are collected smoothly and accurately.

# REFERENCES


효율적인 교통관리를 위한 혼잡상황변화 유형 분류기법 개발, 심상우, 이환필, 이규진, 최기주

고속도로 통행정보를 활용한 KNN 기반 교통정체 예측 기법, 최재억

G. Meena, D. Sharma and M. Mahrishi, "Traffic Prediction for Intelligent Transportation System using Machine Learning," 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), 2020, pp. 145-148, doi: 10.1109/ICETCE48199.2020.9091758.


THANK YOU!