# A Study on the Prediction of Subway in Seoul by Weather

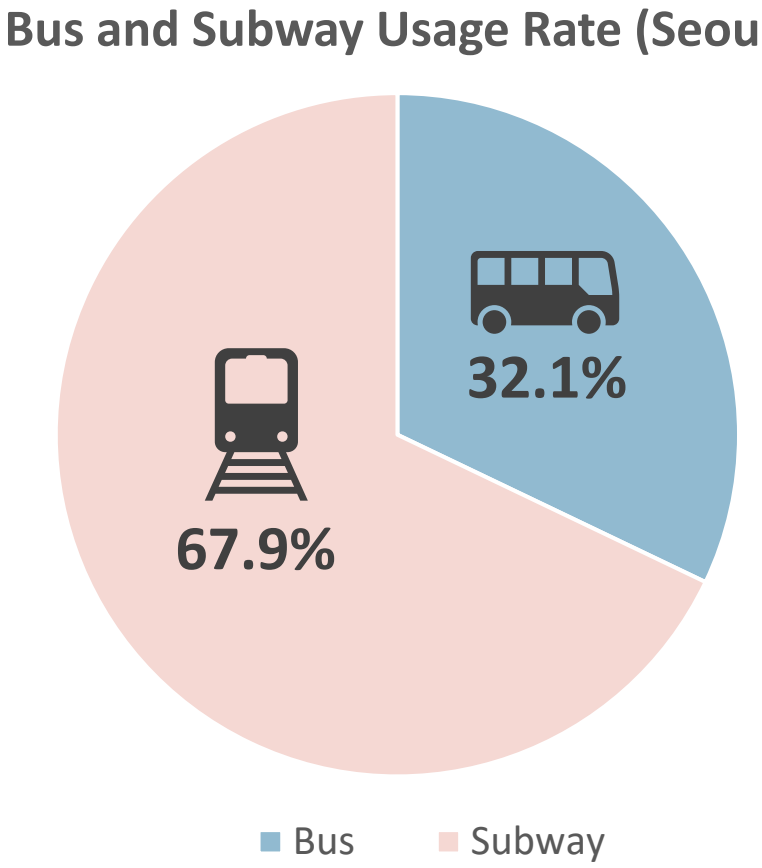
2023 summer QI AI program

Team C

SanGwon, YejinLee, HyeonjeongKim, JoeunKim

# Contents



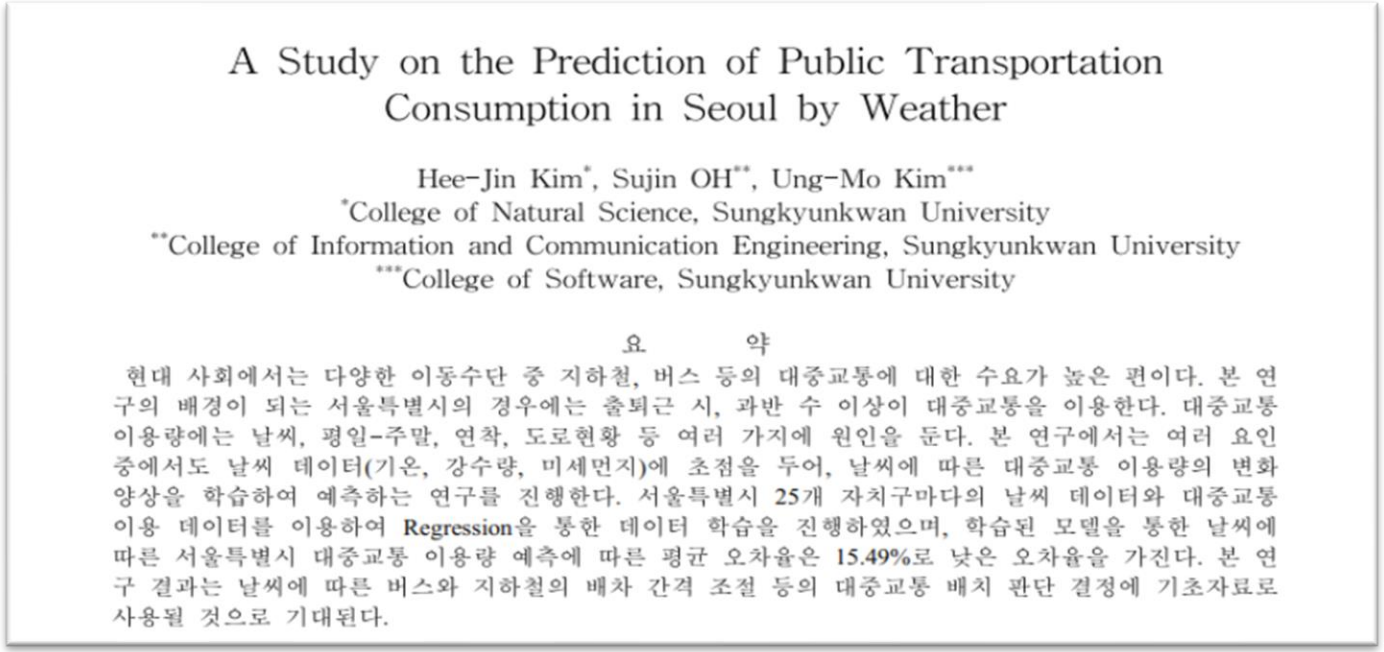



[Fig 1] Ministry of Land, Infrastructure and Transport

**Subway = Essential Transportation**

**⟶ Usage Prediction using Weather**

# Referenced Study



A Study on the Prediction of Public Transportation Consumption in Seoul by Weather

Hee-Jin Kim*, Sujin OH**, Ung-Mo Kim***
*College of Natural Science, Sungkyunkwan University
**College of Information and Communication Engineering, Sungkyunkwan University
***College of Software, Sungkyunkwan University

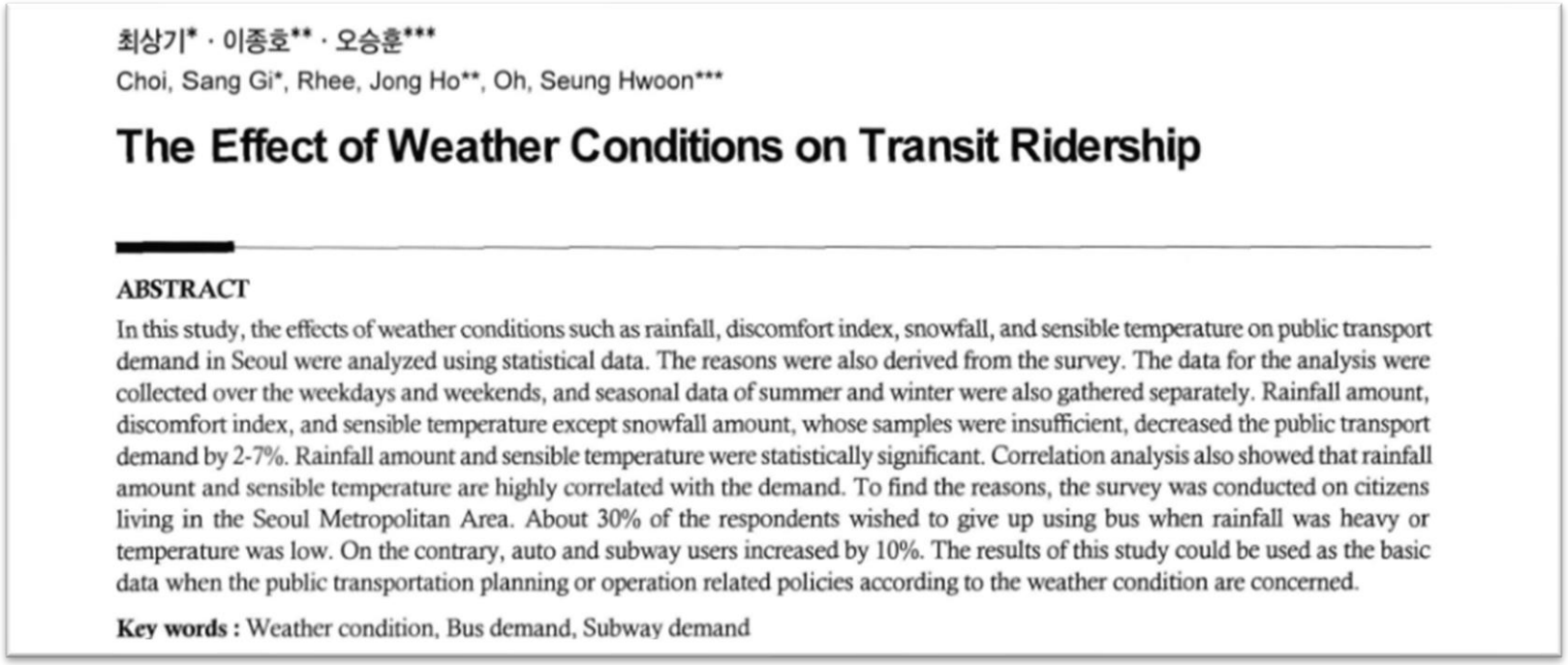
요 약

현대 사회에서는 다양한 이동수단 중 지하철, 버스 등의 대중교통에 대한 수요가 높은 편이다. 본 연구의 배경이 되는 서울특별시의 경우에는 출퇴근 시, 과반 수 이상이 대중교통을 이용한다. 대중교통 이용량에는 날씨, 평일-주말, 연착, 도로현황 등 여러 가지에 원인을 둔다. 본 연구에서는 여러 요인 중에서도 날씨 데이터(기온, 강수량, 미세먼지)에 초점을 두어, 날씨에 따른 대중교통 이용량의 변화 양상을 학습하여 예측하는 연구를 진행한다. 서울특별시 25개 자치구마다의 날씨 데이터와 대중교통 이용 데이터를 이용하여 Regression을 통한 데이터 학습을 진행하였으며, 학습된 모델을 통한 날씨에 따른 서울특별시 대중교통 이용량 예측에 따른 평균 오차율은 15.49%로 낮은 오차율을 가진다. 본 연구 결과는 날씨에 따른 버스와 지하철의 배차 간격 조절 등의 대중교통 배치 판단 결정에 기초자료로 사용될 것으로 기대된다.

[Fig 2] A Study on the Prediction of Public Transportation Consumption in Seoul by Weather
Hea-Jin Kim, Sujin Oh, Ung-Mo Kim

- Monthly
- Gaussian, Bagging, Random Forest
- Temperature + Rainfall + Fine dust

# Referenced Study



최상기* · 이종호** · 오승훈***

Choi, Sang Gi*, Rhee, Jong Ho**, Oh, Seung Hwoon***

## The Effect of Weather Conditions on Transit Ridership

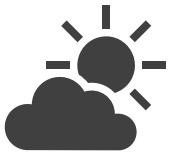
**ABSTRACT**

In this study, the effects of weather conditions such as rainfall, discomfort index, snowfall, and sensible temperature on public transport demand in Seoul were analyzed using statistical data. The reasons were also derived from the survey. The data for the analysis were collected over the weekdays and weekends, and seasonal data of summer and winter were also gathered separately. Rainfall amount, discomfort index, and sensible temperature except snowfall amount, whose samples were insufficient, decreased the public transport demand by 2-7%. Rainfall amount and sensible temperature were statistically significant. Correlation analysis also showed that rainfall amount and sensible temperature are highly correlated with the demand. To find the reasons, the survey was conducted on citizens living in the Seoul Metropolitan Area. About 30% of the respondents wished to give up using bus when rainfall was heavy or temperature was low. On the contrary, auto and subway users increased by 10%. The results of this study could be used as the basic data when the public transportation planning or operation related policies according to the weather condition are concerned.

**Key words :** Weather condition, Bus demand, Subway demand

[Fig 3] The Effect of Weather Conditions on Transit Ridership
Choi SangGi, Rhee JongHo, Oh SeungHwoon

- Daily
- Taking a Survey
- Derivation of Regression Equation



# Weather Data

*Before Preprocessing*

| Year | Number of Rows |
|---|---|
| **2016** | 8784 |
| **2017** | 8760 |
| **2018** | 8760 |
| **2019** | 8760 |

**Number of Columns : 27**

- Area code
- Area name
- Date & Time
- Temperature
- Rain fall
- Wind speed
- Humidity
- Vapor pressure

.
.
.

- Provided by the Korea Meteorological Administration
- From 0 a.m. to 12 p.m.

# Subway Data

*Before Preprocessing*

| Year | Number of Rows |
|---|---|
| **2016** | 201,300 |
| **2017** | 200,750 |
| **2018** | 200,750 |
| **2019** | 200,804 |

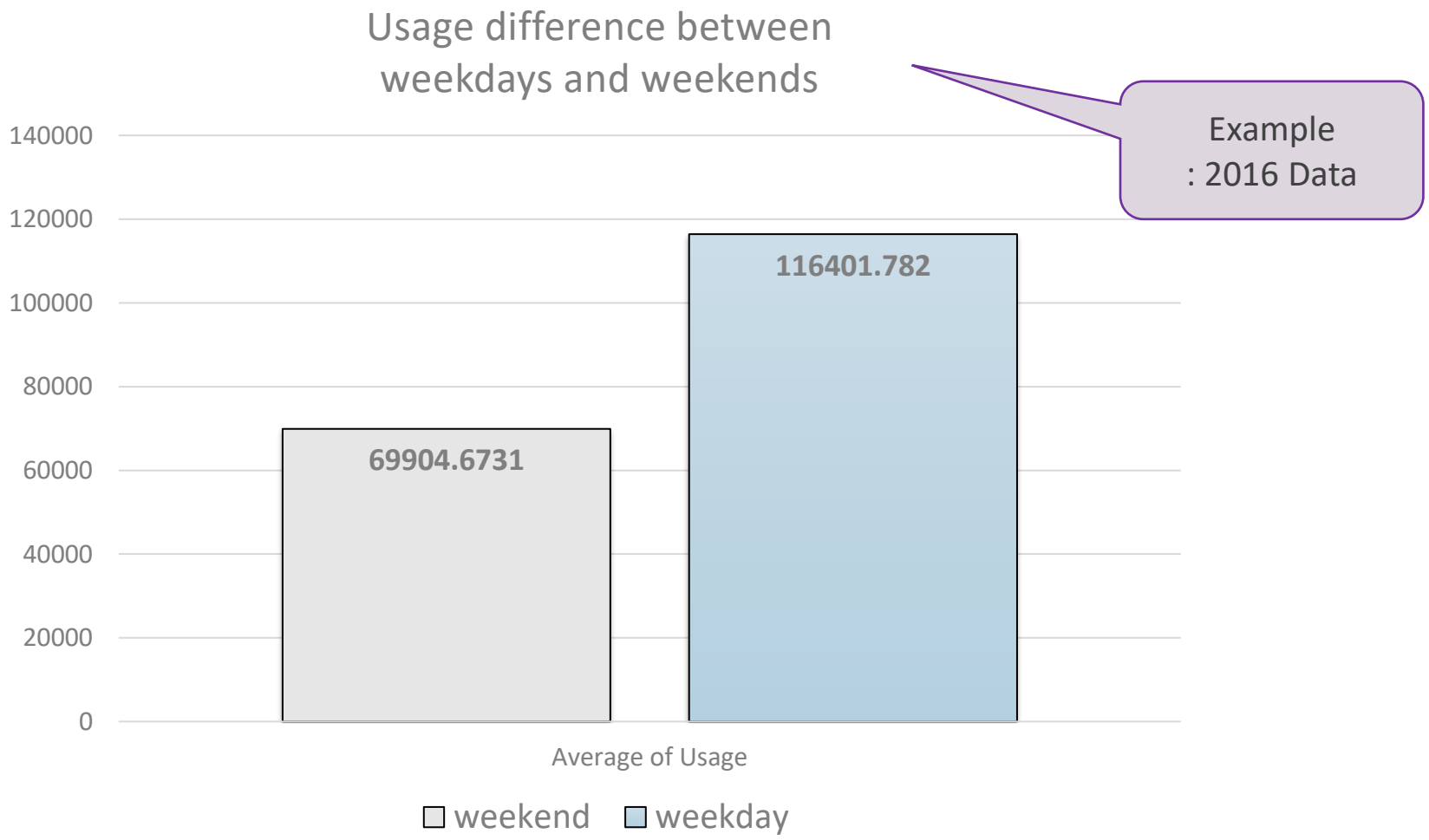
**Number of Columns : 24**

- Date
- Line
- Station Name
- Category
- 5 a.m. ~ 6 a.m.
- 6 a.m. ~ 7 a.m.

.
.
.
.

- all subway stations managed by the Seoul Transportation Corporation
- the number of people getting on and off from 5 a.m. to 12 p.m. by station from 2017 to 2019

# *Preprocessing : Why did we exclude holiday and weekend data?*



*Preprocessing: Why did we choose sensible temperature data?*

**Sensible Temperature** (Summer : May ~ Sep) =

$$-0.2442 + 0.55399Tw + 0.45535Ta - 0.0022Tw^2 + 0.00278TwTa + 3.0$$

Ta = Temperature (°c)
Tw = Wet bulb temperature (°c)
RH = Relative humidity (%)

**Daily highest sensible temperature**

**Sensible Temperature** (Winter: Oct ~ Apr) =

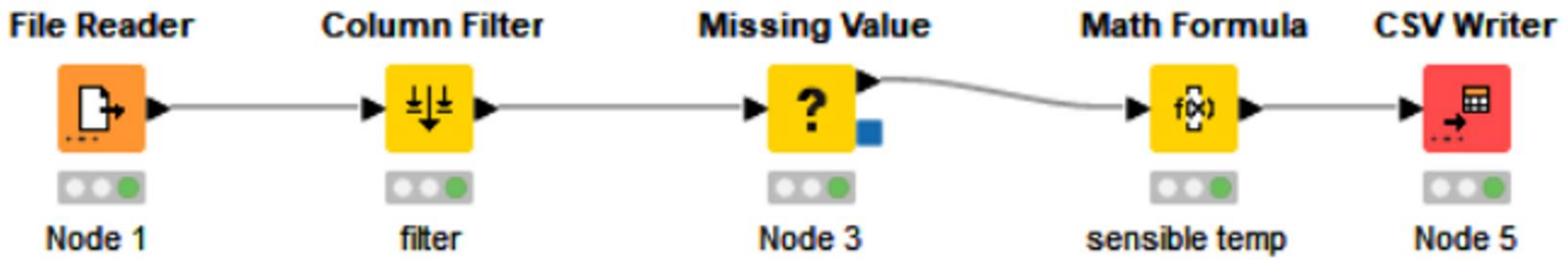
$$13.12 + 0.6215Ta - 11.37V^{0.16} + 0.3965V^{0.16}Ta$$

Ta = Temperature (°c)
V = 10-minute average wind speed (km/h)

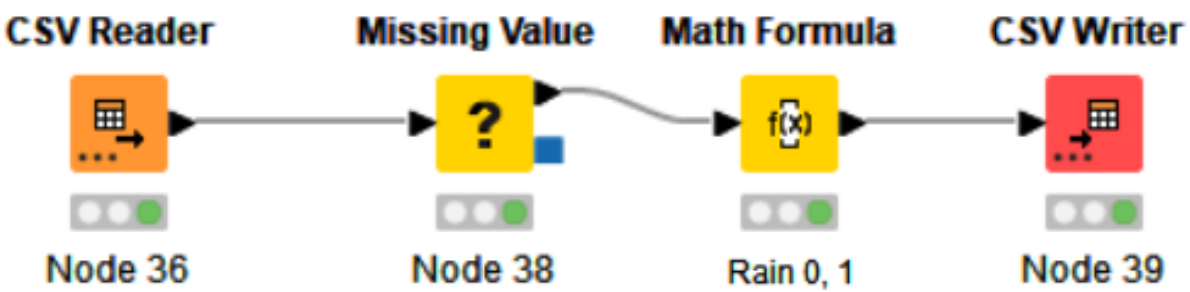
**Daily lowest sensible temperature**
**Temperature $\leq$ 10°c, Wind speed $\geq$ 1.3 m/s**

[Fig 4] Korea Meteorological Administration

# *Preprocessing with KNIME*




▲ Sensible-temperature Node



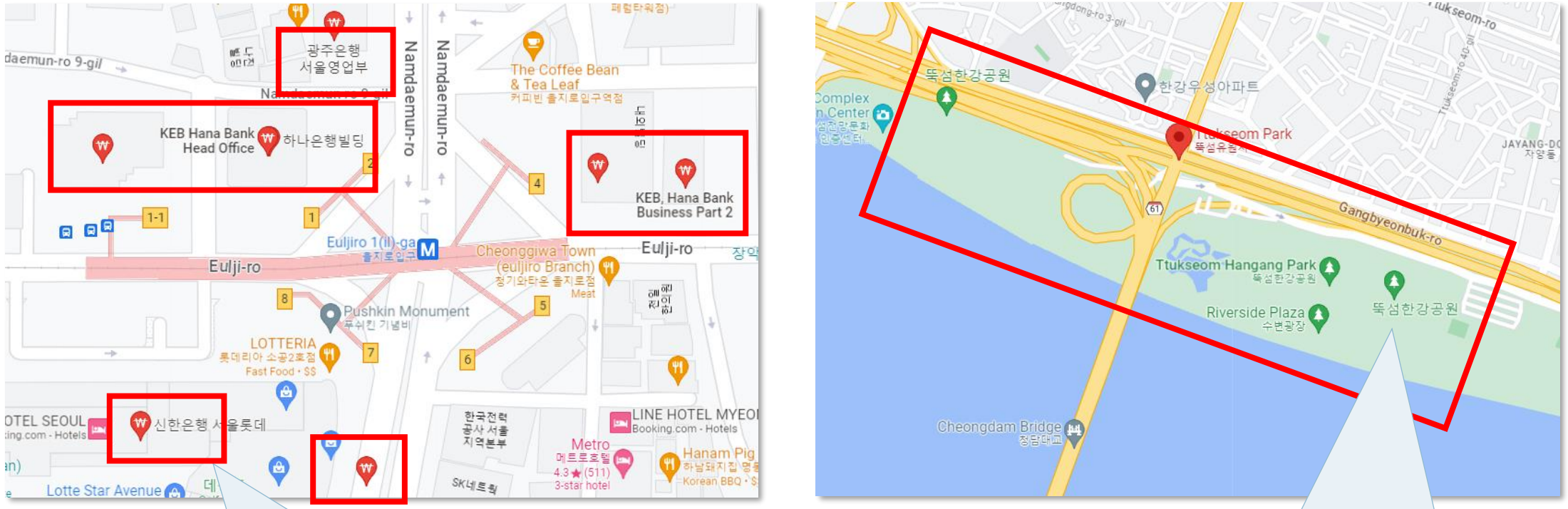

▲ Rain data Changing Node

| Rain(mm) |
|---|
| 0 |
| 0.3 |
| 0.1 |

| Rain(mm) |
|---|
| 0 |
| 1 |
| 1 |

**Rainy day → 1**
**Not rainy day → 0**

# *Preprocessing : Why did we select 'Euljiro 1(il)-ga'* & 'Ttukseom Park'?

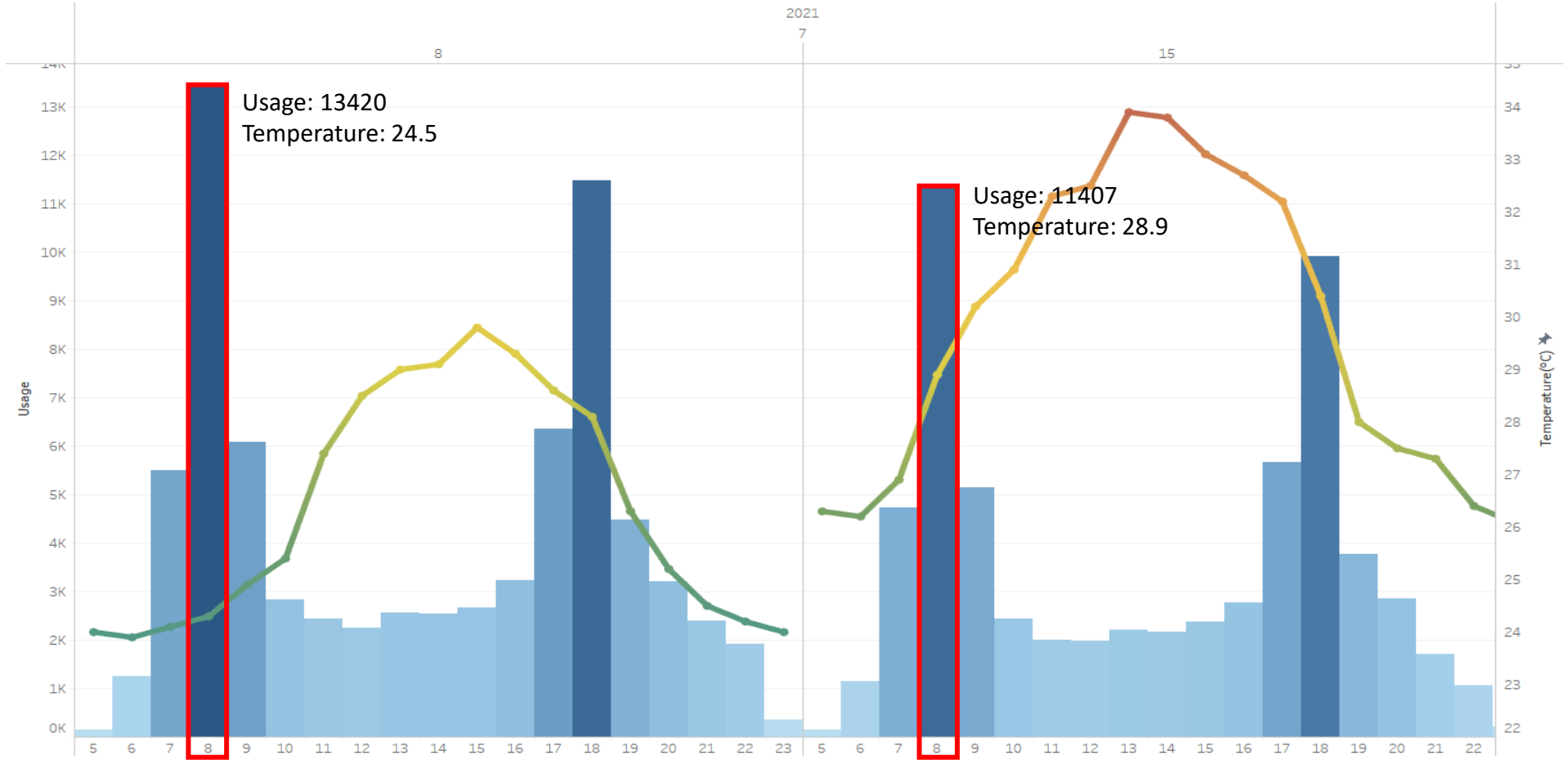



[Fig 5] Google Map

- Lots of financial companies
- Main users = office workers

- Nearby Han-River
- Main users = people who visit the park

# *Preprocessing : Why did we use past data?*




2020 ~ 2022 : Pandemic - Usage inconsistency → 2016 ~ 2019 : Pre-pandemic data

*After Preprocessing*

| Date | Rain | Wind | Humidity | Sensible-temp | Usage | Hour |
|---|---|---|---|---|---|---|
| | Rain (hour) | Wind (hour) | Humidity (hour) | Sensible-temp (hour) | Usage (hour) | |
| | Rain (week) | Wind (week) | Humidity (week) | Sensible-temp (week) | Usage (week) | |

Target Value

**Number of Our Data Sets : 4**

1. Euljiro 1(il)-ga boarding
2. Euljiro 1(il)-ga exiting
3. Ttukseom Park boarding
4. Ttukseom Park exiting

**AVG Number of Rows: 18455**
**Number of Columns: 17**

# Dataset #1. All features

**Features(15): Rain, Wind, Humidity, Sensible Temp, Hour, Rain(hour), Wind(hour), Humidity(hour), Sensible Temp(hour), Rain(week), Wind(week), Humidity(week), Sensible Temp(week), Usage(hour), Usage(week)**

# Dataset #2. Important feature Top 8

**Features(8): Hour, Usage(hour), Usage(week), Wind, Wind(hour), Wind(week), Humidity(hour), Humidity(week)**

# Linear Correlation

: A statistical measure that quantifies the strength of the linear relationship between two continuous variables.

| | Euljiro 1(il)-ga boarding | Euljiro 1(il)-ga exiting | Ttukseom Park boarding | Ttukseom Park exiting |
|---|---|---|---|---|
| Hour | 0.612 | -0.580 | -0.459 | 0.487 |
| Humidity | -0.141 | 0.122 | 0.156 | -0.114 |
| Humidity(hour) | -0.243 | 0.198 | 0.192 | -0.199 |
| Humidity(week) | -0.154 | 0.142 | 0.224 | -0.077 |
| Rain | -0.028 | 0.011 | -0.040 | -0.073 |
| Rain(hour) | -0.033 | 0.014 | -0.041 | -0.074 |
| Rain(week) | -0.030 | 0.016 | 0.012 | -0.043 |
| Sensible_temp | 0.048 | -0.079 | 0.056 | 0.179 |
| Sensible_temp(hour) | 0.086 | -0.099 | 0.052 | 0.214 |
| Sensible_temp(week) | 0.053 | -0.083 | 0.030 | 0.154 |
| Usage(hour) | 0.644 | 0.510 | 0.641 | 0.332 |
| Usage(week) | 0.975 | 0.993 | 0.889 | 0.849 |
| Wind | 0.183 | -0.119 | -0.141 | 0.153 |
| Wind(hour) | 0.269 | -0.169 | -0.160 | 0.228 |
| Wind(week) | 0.193 | -0.130 | -0.138 | 0.181 |

**Dataset #1 ( All features)**

# Linear Correlation

: The closer the absolute value is to 1, the higher the correlation

| | Euljiro 1(il)-ga boarding | Euljiro 1(il)-ga exiting | Ttukseom Park boarding | Ttukseom Park exiting |
|---|---|---|---|---|
| Hour | 0.612 | -0.580 | -0.459 | 0.487 |
| Humidity | -0.141 | 0.122 | 0.156 | -0.114 |
| Humidity(hour) | -0.243 | 0.198 | 0.192 | -0.199 |
| Humidity(week) | -0.154 | 0.142 | 0.224 | -0.077 |
| Rain | -0.028 | 0.011 | -0.040 | -0.073 |
| Rain(hour) | -0.033 | 0.014 | -0.041 | -0.074 |
| Rain(week) | -0.030 | 0.016 | 0.012 | -0.043 |
| Sensible_temp | 0.048 | -0.079 | 0.056 | 0.179 |
| Sensible_temp(hour) | 0.086 | -0.099 | 0.052 | 0.214 |
| Sensible_temp(week) | 0.053 | -0.083 | 0.030 | 0.154 |
| Usage(hour) | 0.644 | 0.510 | 0.641 | 0.332 |
| Usage(week) | 0.975 | 0.993 | 0.889 | 0.849 |
| Wind | 0.183 | -0.119 | -0.141 | 0.153 |
| Wind(hour) | 0.269 | -0.169 | -0.160 | 0.228 |
| Wind(week) | 0.193 | -0.130 | -0.138 | 0.181 |

**Dataset #1 ( All features)**

## Linear Correlation

: The closer the absolute value is to 1, the higher the correlation

| | Euljiro 1(il)-ga Boarding | Euljiro 1(il)-ga Exiting | Ttukseom Park Boarding | Ttukseom Park Exiting |
|---|---|---|---|---|
| Humidity | -0.1414 | 0.1224 | 0.1558 | -0.1140 |
| Humidity(hour) | -0.2426 | 0.1981 | 0.1919 | -0.1990 |
| Humidity(week) | -0.1543 | 0.1419 | 0.2241 | -0.0770 |
| Rain | -0.0281 | 0.0108 | -0.0399 | -0.0730 |
| Rain(hour) | -0.0326 | 0.0145 | -0.0409 | -0.0740 |
| Rain(week) | -0.0303 | 0.0157 | 0.0115 | -0.0430 |
| Sensible_temp | 0.0485 | -0.0788 | 0.0558 | 0.1790 |
| Sensible_temp(hour) | 0.0859 | -0.0991 | 0.0521 | 0.2140 |
| Sensible_temp(week) | 0.0534 | -0.0829 | 0.0295 | 0.1540 |
| Wind | 0.1830 | -0.1194 | -0.1410 | 0.1530 |
| Wind(hour) | 0.2688 | -0.1689 | -0.1604 | 0.2280 |
| Wind(week) | 0.1934 | -0.1296 | -0.1377 | 0.1810 |

**Dataset #1 ( All features)**

# Random Forest Feature Importance

: The attributes by which 1500 models are initially split at the root node, which is the top node.

| | Euljiro 1(il)-ga boarding | Euljiro 1(il)-ga exiting | Ttukseom Park boarding | Ttukseom Park exiting |
|---|---|---|---|---|
| Hour | 252 | 215 | 256 | 283 |
| Humidity | 62 | 51 | 114 | 22 |
| Humidity(hour) | 128 | 175 | 140 | 64 |
| Humidity(week) | 71 | 105 | 175 | 11 |
| Rain | 0 | 1 | 4 | 4 |
| Rain(hour) | 5 | 1 | 1 | 3 |
| Rain(week) | 3 | 2 | 0 | 0 |
| Sensible_temp | 11 | 8 | 34 | 63 |
| Sensible_temp(hour) | 35 | 36 | 21 | 150 |
| Sensible_temp(week) | 17 | 11 | 7 | 24 |
| Usage(hour) | 213 | 278 | 248 | 211 |
| Usage(week) | 313 | 296 | 286 | 291 |
| Wind | 89 | 68 | 53 | 65 |
| Wind(hour) | 177 | 150 | 93 | 190 |
| Wind(week) | 124 | 103 | 68 | 119 |

**Dataset #1 ( All features) & Tree depth=10 & Number of Models=1500**

# Random Forest Feature Importance

: A higher count of splits for a particular attribute indicates that it is considered more important

| | Euljiro 1(il)-ga boarding | Euljiro 1(il)-ga exiting | Ttukseom Park boarding | Ttukseom Park exiting |
|---|---|---|---|---|
| Hour | 252 | 215 | 256 | 283 |
| Humidity | 62 | 51 | 114 | 22 |
| Humidity(hour) | 128 | 175 | 140 | 64 |
| Humidity(week) | 71 | 105 | 175 | 11 |
| Rain | 0 | 1 | 4 | 4 |
| Rain(hour) | 5 | 1 | 1 | 3 |
| Rain(week) | 3 | 2 | 0 | 0 |
| Sensible_temp | 11 | 8 | 34 | 63 |
| Sensible_temp(hour) | 35 | 36 | 21 | 150 |
| Sensible_temp(week) | 17 | 11 | 7 | 24 |
| Usage(hour) | 213 | 278 | 248 | 211 |
| Usage(week) | 313 | 296 | 286 | 291 |
| Wind | 89 | 68 | 53 | 65 |
| Wind(hour) | 177 | 150 | 93 | 190 |
| Wind(week) | 124 | 103 | 68 | 119 |

**Dataset #1 ( All features) & Tree depth=10 & Number of Models=1500**

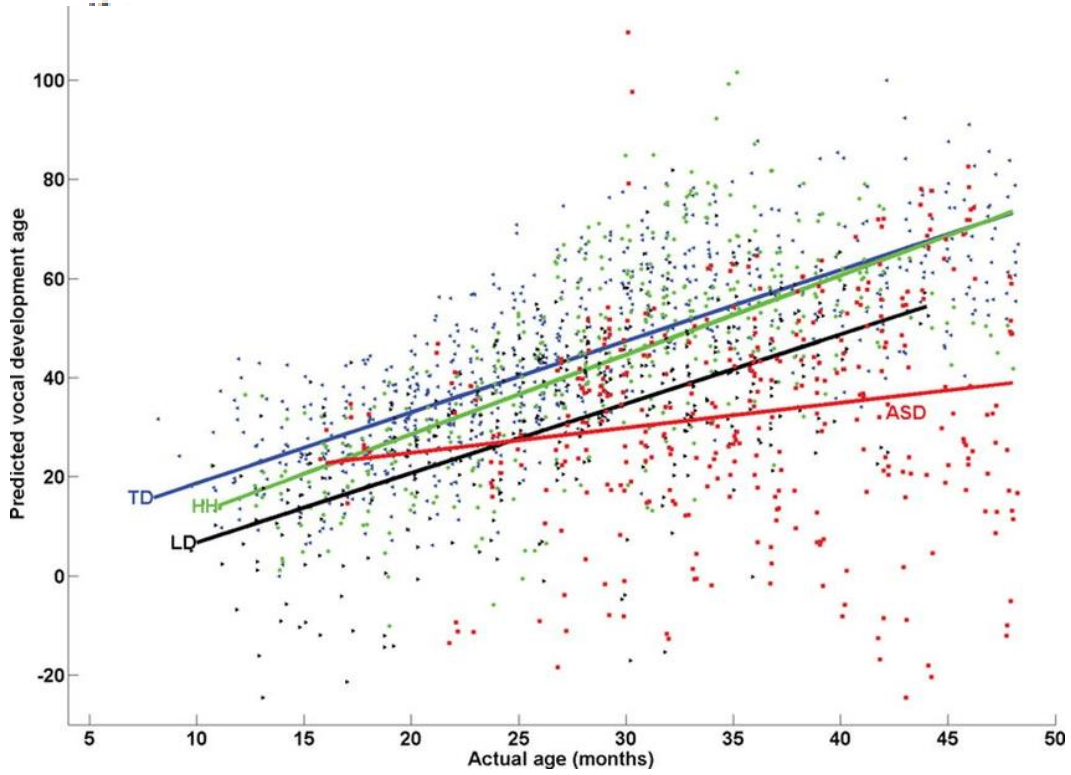
# Random Forest Feature Importance

: A higher count of splits for a particular attribute indicates that it is considered more important

| | Euljiro 1(il)-ga Boarding | Euljiro 1(il)-ga Exiting | Ttukseom Park Boarding | Ttukseom Park Exiting |
|---|---|---|---|---|
| Humidity | 62 | 51 | 114 | 22 |
| Humidity(hour) | 128 | 175 | 140 | 64 |
| Humidity(week) | 71 | 105 | 175 | 11 |
| Rain | 0 | 1 | 4 | 4 |
| Rain(hour) | 5 | 1 | 1 | 3 |
| Rain(week) | 3 | 2 | 0 | 0 |
| Sensible_temp | 11 | 8 | 34 | 63 |
| Sensible_temp(hour) | 35 | 36 | 21 | 150 |
| Sensible_temp(week) | 17 | 11 | 7 | 24 |
| Wind | 89 | 68 | 53 | 65 |
| Wind(hour) | 177 | 150 | 93 | 190 |
| Wind(week) | 124 | 103 | 68 | 119 |

**Dataset #1 ( All features) & Tree depth=10 & Number of Models=1500**

# *Multiple Linear Regression*




[Fig 6] Linear Regression, Wikipedia
[Fig 7] Multiple Linear Regressoin, CFI

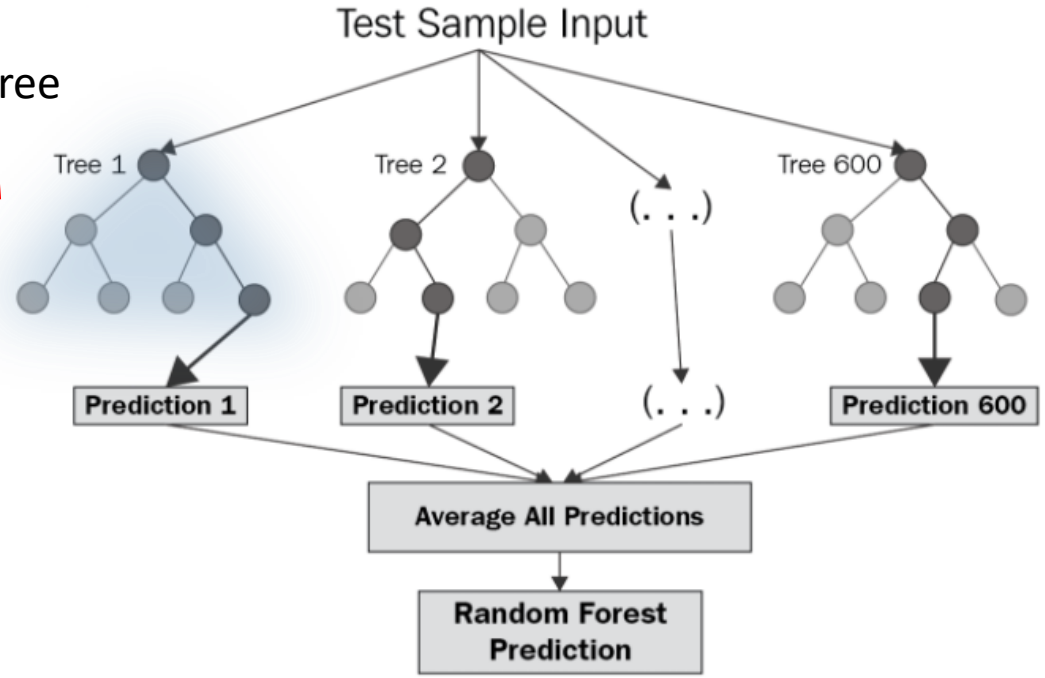
: To model the linear relationship between multiple independent variables and one dependent variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_n x_n + \epsilon$$

$\epsilon$ and $\beta_0$ are the error and y-intercepts, respectively, $\beta_n$ is the value representing the increment of the dependent variable for the independent variable $x_n$ and corresponds to the regression coefficient.

# *Random Forest Regression*

Decision tree

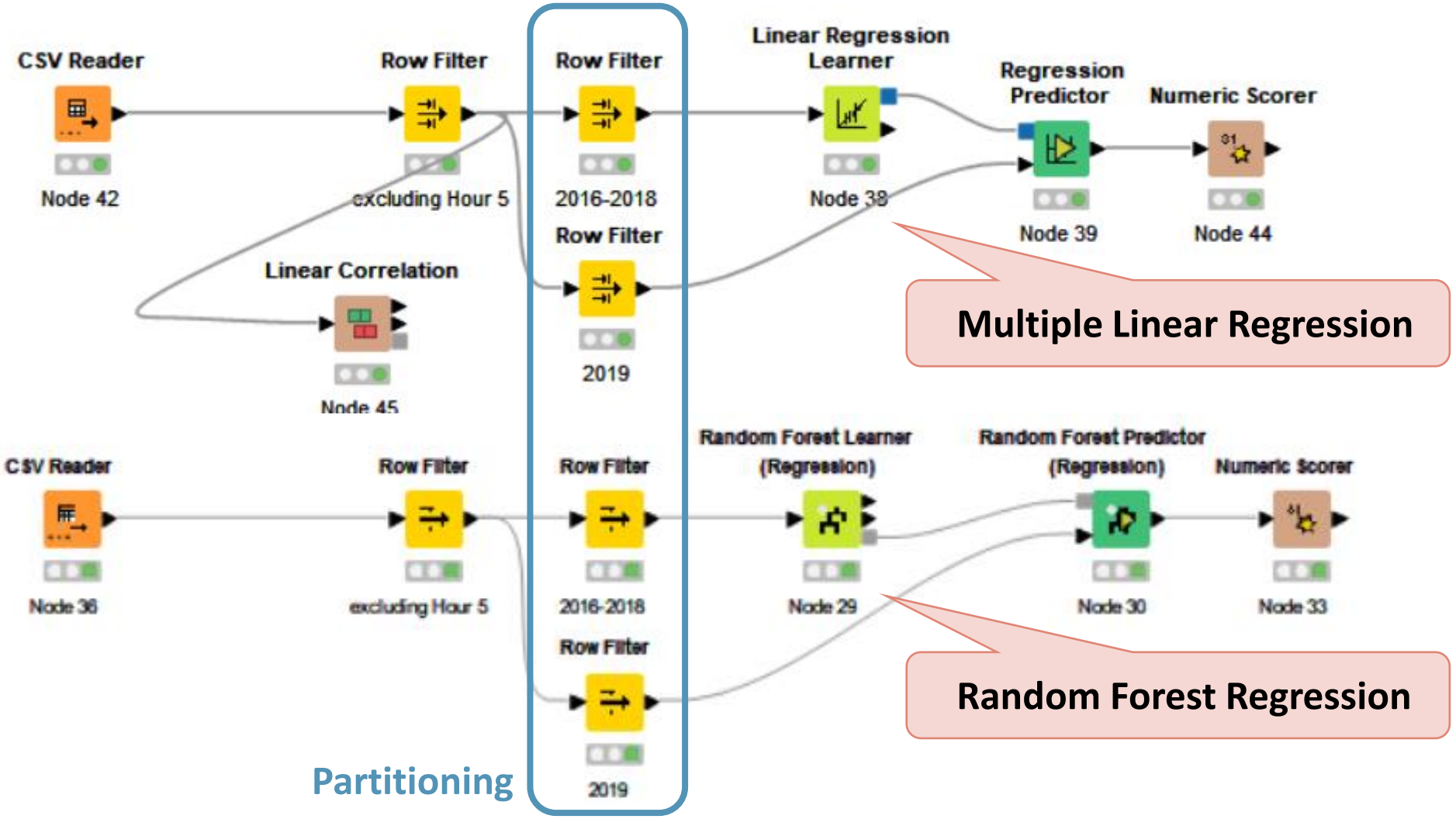



[Fig 8] Random Forest Regression, medium

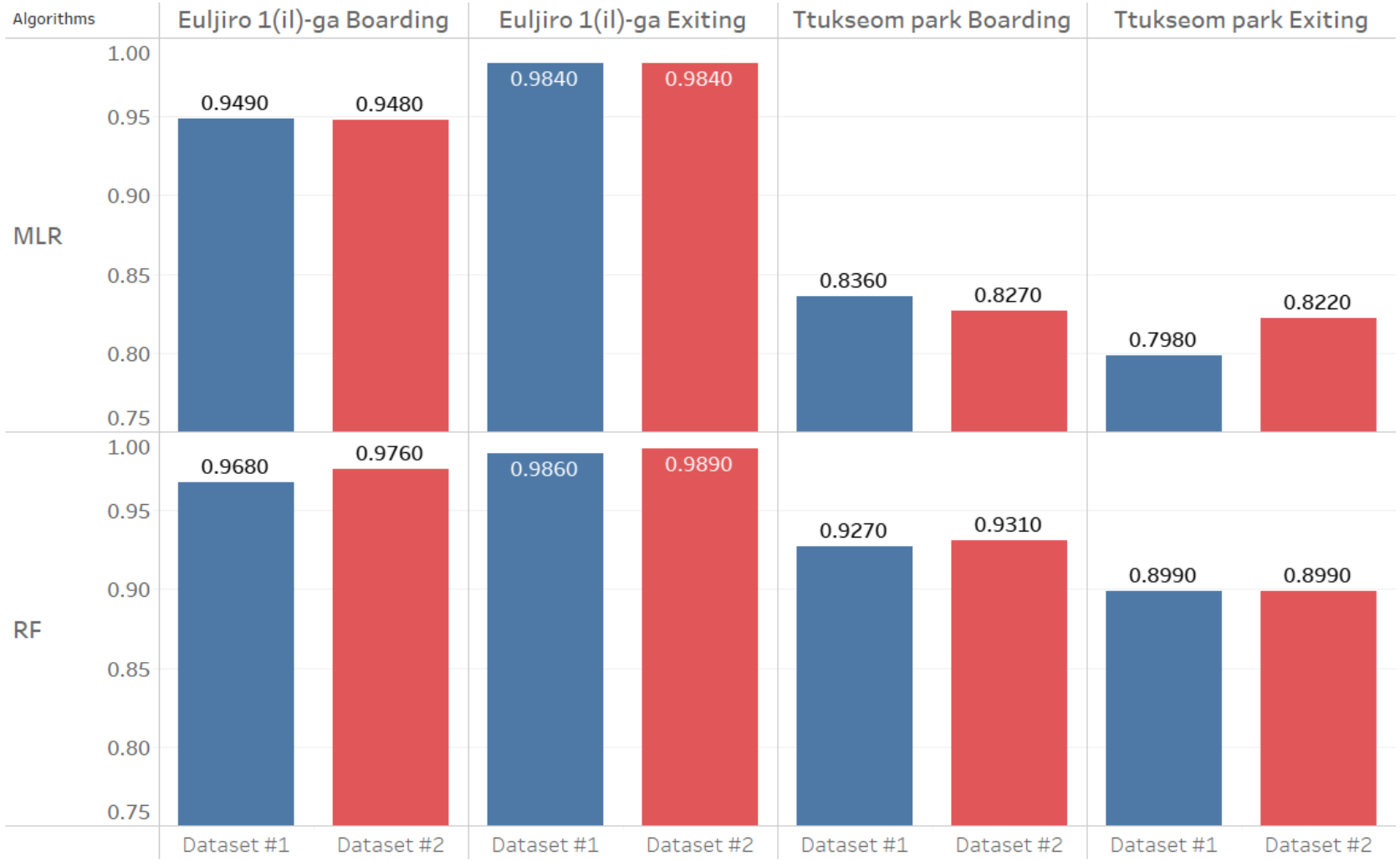
: An ensemble machine learning algorithm that combines multiple decision trees to make accurate predictions for regression tasks.

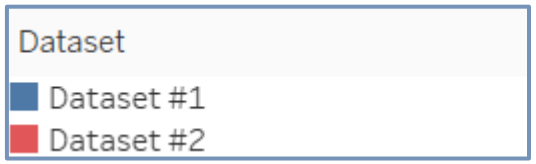
Providing stable predictive performance across diverse datasets and being beneficially employed in complex datasets with non-linear relationships.

# *Algorithms using KNIME*



# Dataset #1 (All features) VS Dataset #2 (Top 8 features)



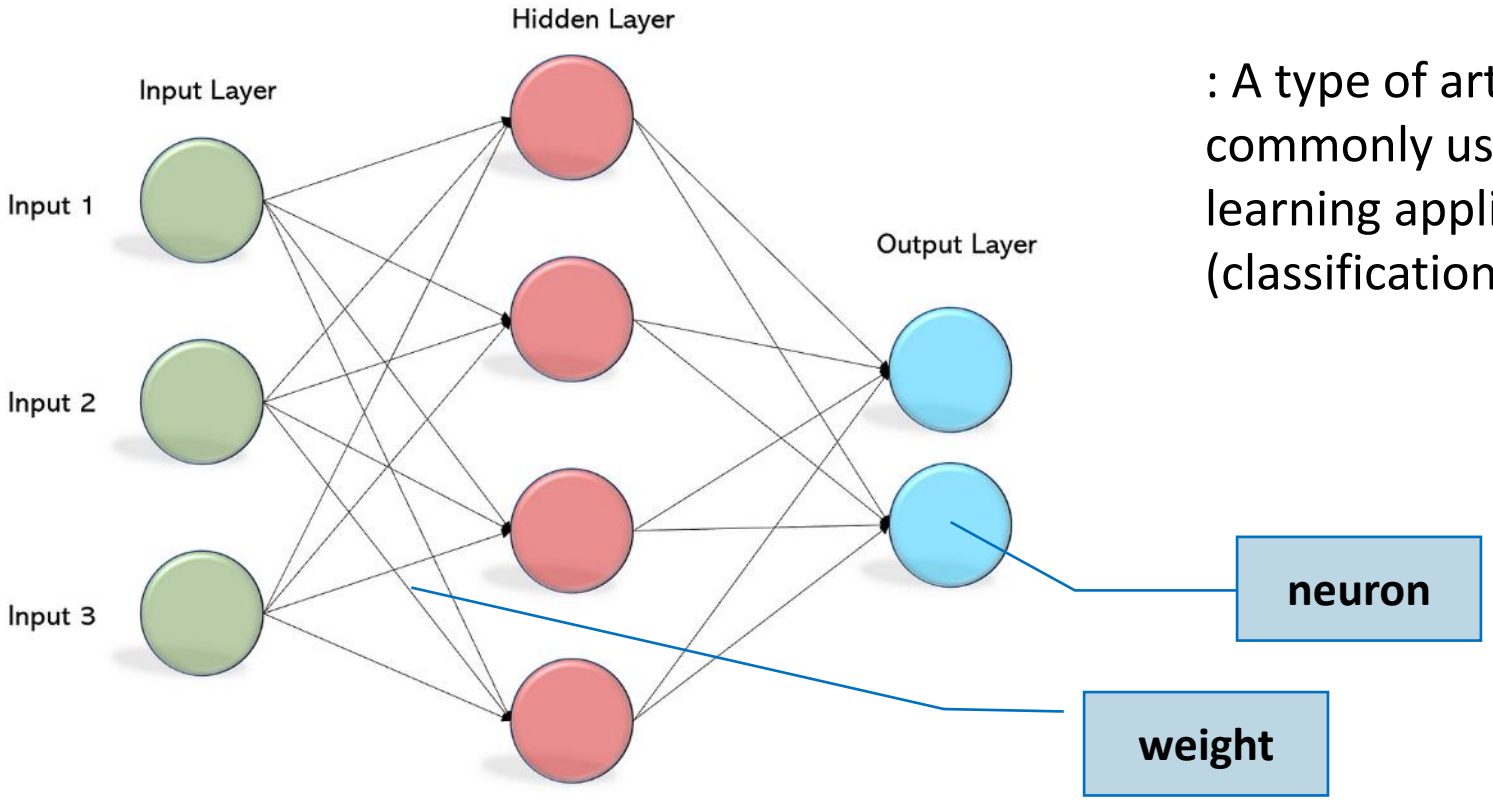

**$R^2$**

Almost

**Dataset #1** < **Dataset #2**

※ Because the number of each top 8 features included in dataset 2 is different, the value may be slightly reduced !

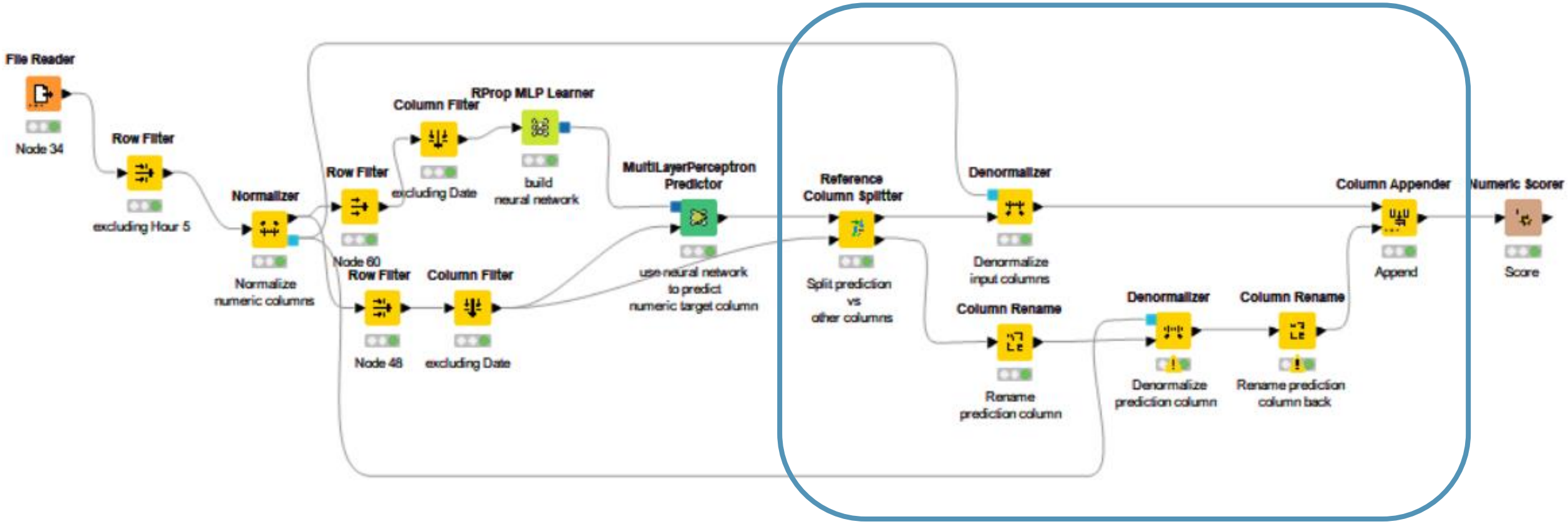
# *Multi Layer Perceptron*




[Fig 9] Muti Layer Perceptron, Wikipedia

: A type of artificial neural network commonly used in various machine learning applications.
(classification & regression)

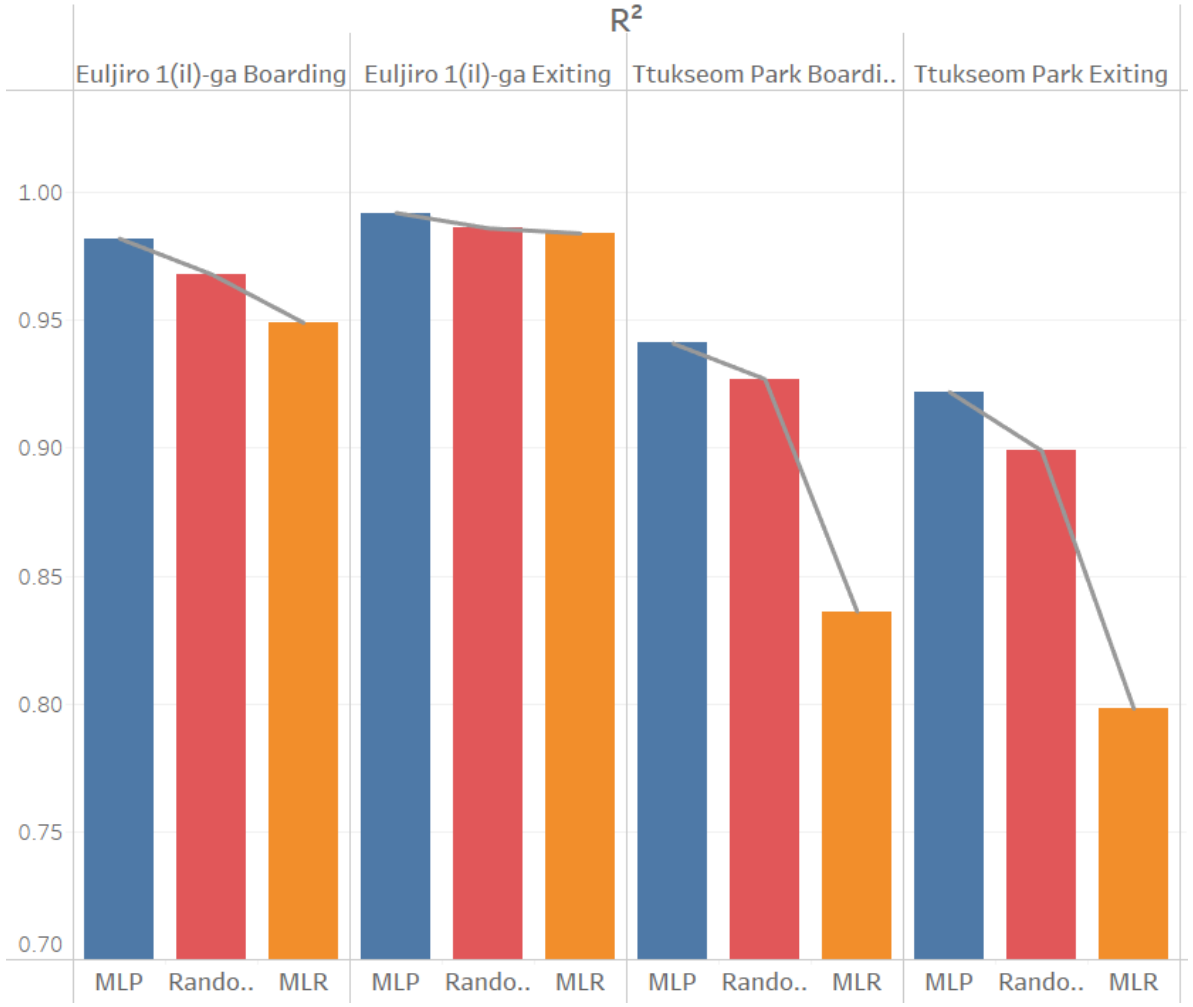
# *Algorithms using KNIME*




**Denormalize Result for Comparison**

# Analysis Method & Result





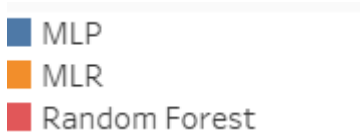

| Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|
| 500/3/32 | 500/6/16 | 500/5/24 | 1000/3/32 |

## $R^2$ score

| | Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|---|
| **MLP** | 0.982 | 0.992 | 0.941 | 0.922 |
| **Random Forest** | 0.968 | 0.986 | 0.927 | 0.899 |
| **MLR** | 0.949 | 0.984 | 0.836 | 0.798 |

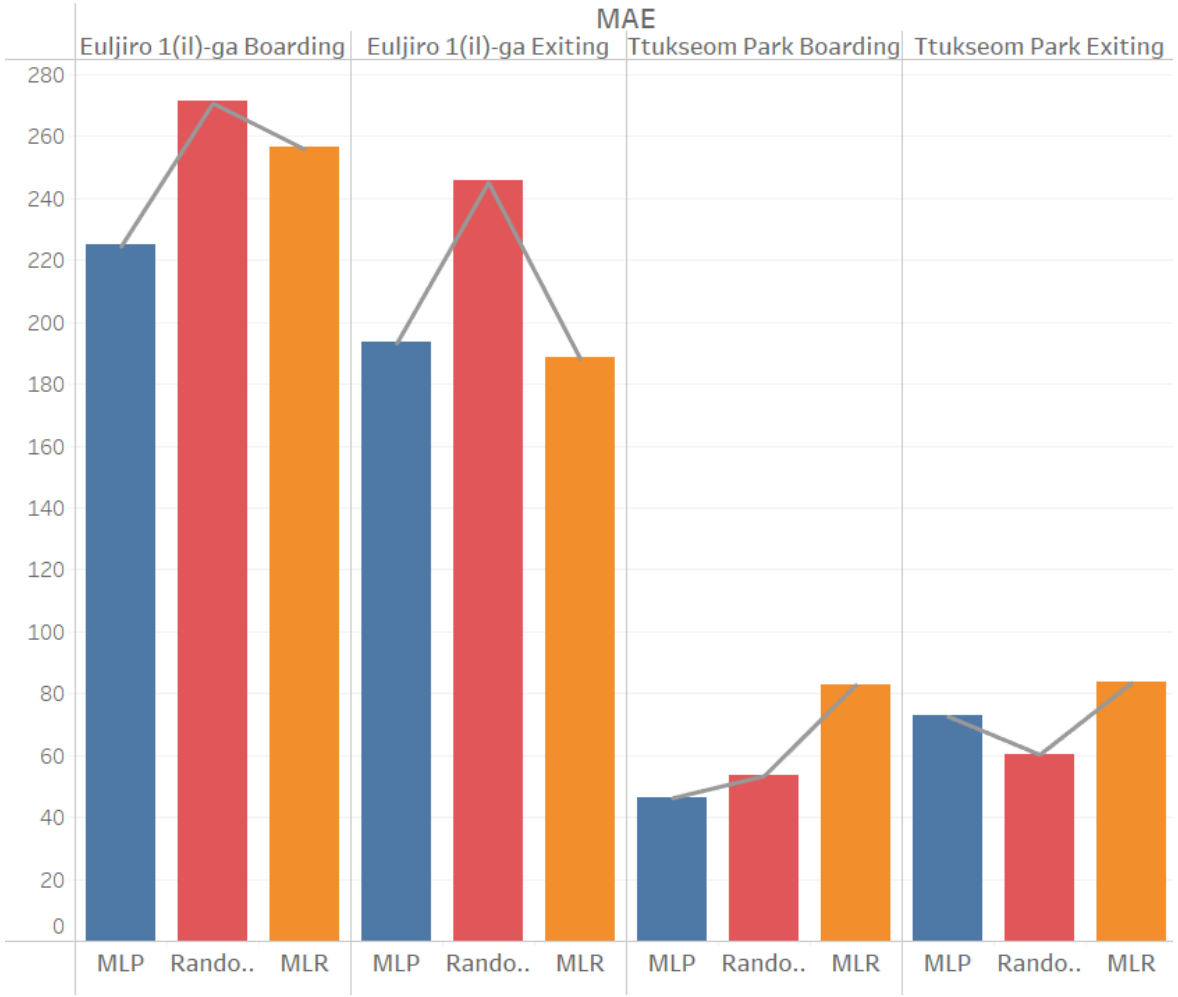
## Using Dataset #1 ( All features)

※ Random Forest‘s Tree depth=10 & Number of Models=1500



# Analysis Method & Result





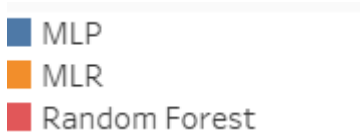

| Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|
| 500/3/32 | 500/6/16 | 500/5/24 | 1000/3/32 |

## MAE score

| | Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|---|
| **MLP** | 224.9 | 193.5 | 46.3 | 72.8 |
| **Random Forest** | 271.4 | 245.8 | 53.5 | 60.4 |
| **MLR** | 256.5 | 188.5 | 83.0 | 83.7 |

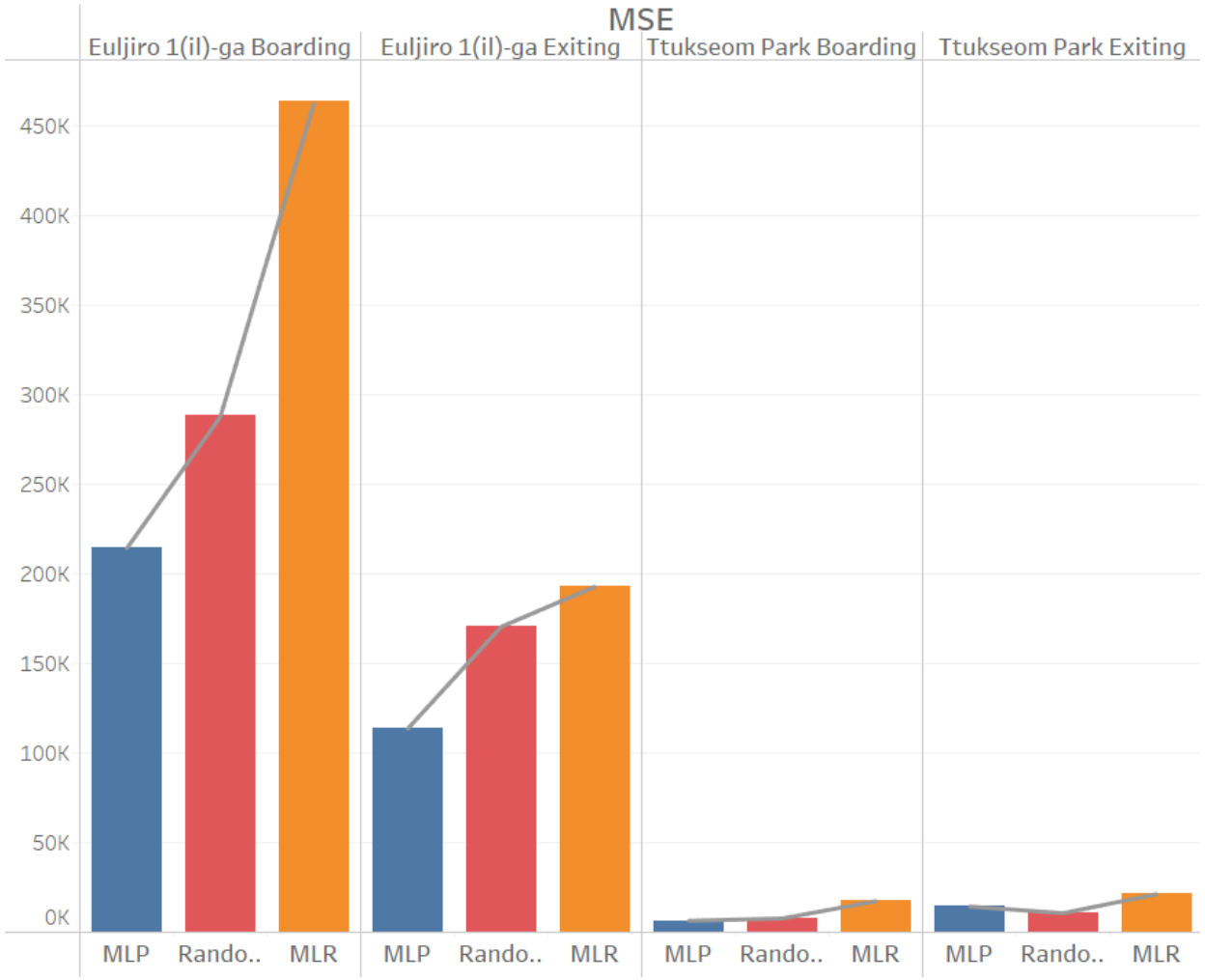
## Using Dataset #1 ( All features)

※ Random Forest‘s Tree depth=10 & Number of Models=1500



# Analysis Method & Result





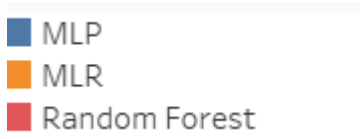

| Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|
| 500/3/32 | 500/6/16 | 500/5/24 | 1000/3/32 |

## MSE score

| | Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|---|
| **MLP** | 214,519 | 113,705 | 6,404 | 14,412 |
| **Random Forest** | 288,198 | 170,837 | 7,704 | 10,678 |
| **MLR** | 463,428 | 193,203 | 17,361 | 21,295 |

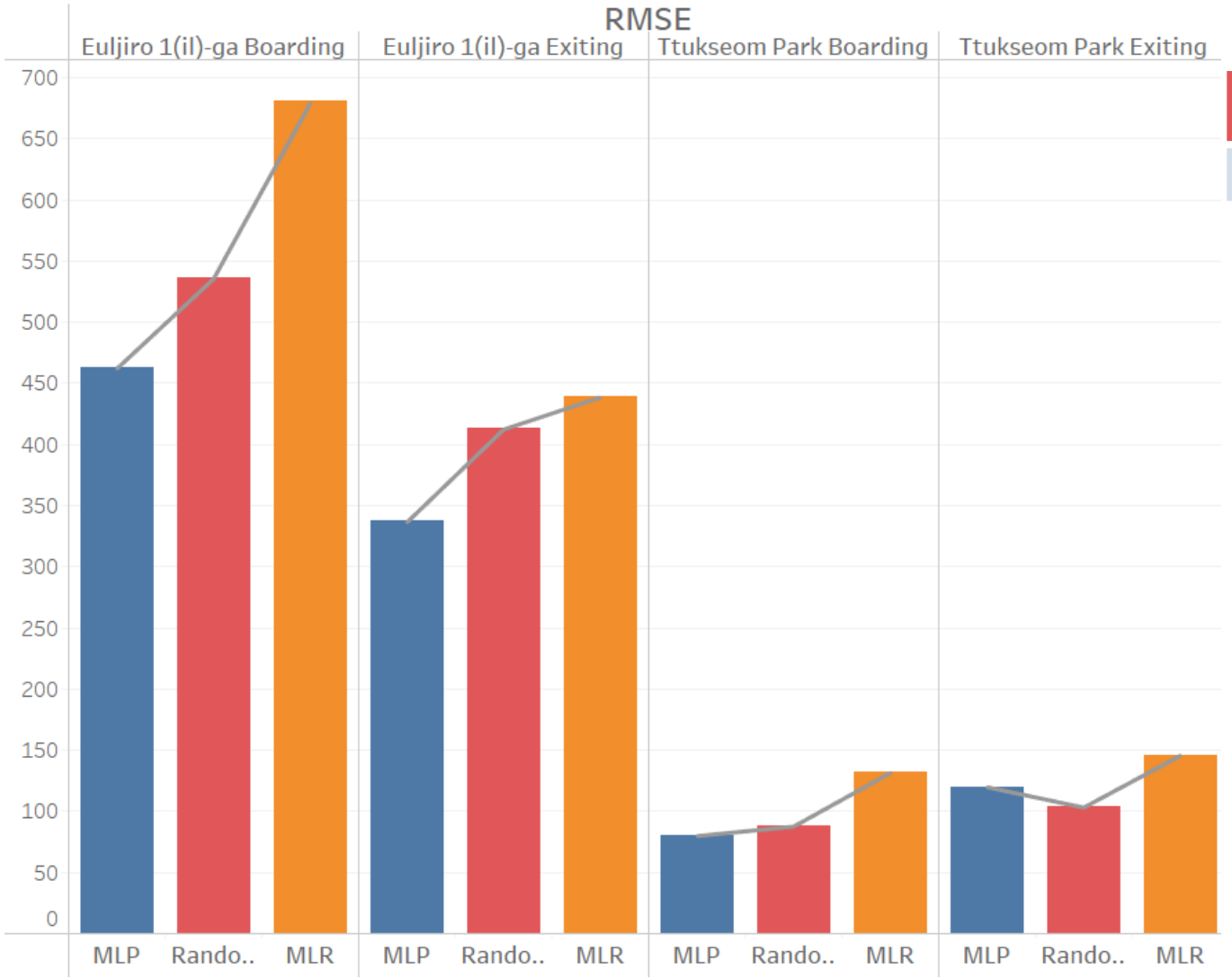
## Using Dataset #1 ( All features)

※ Random Forest‘s Tree depth=10 & Number of Models=1500



# Analysis Method & Result





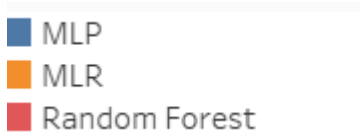

| Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|
| 500/3/32 | 500/6/16 | 500/5/24 | 1000/3/32 |

## RMSE score

| | Euljiro il(1) ga boarding | Euljiro il(1) ga exiting | Ttukseom park boarding | Ttukseom park exiting |
|---|---|---|---|---|
| **MLP** | 463.2 | 337.2 | 80.0 | 120.1 |
| **Random Forest** | 536.8 | 413.3 | 87.8 | 103.3 |
| **MLR** | 680.8 | 439.5 | 131.8 | 145.9 |

## Using Dataset #1 ( All features)

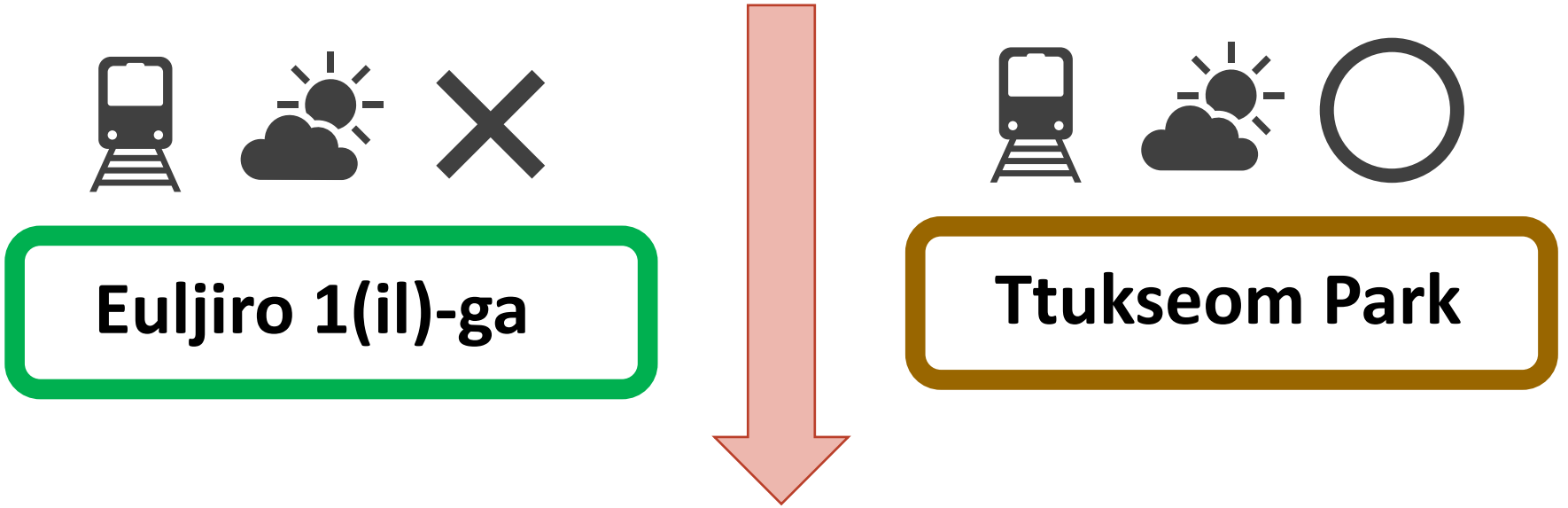
※ Random Forest‘s Tree depth=10 & Number of Models=1500



temperature was low. On the contrary, auto and subway users increased by 10%. The results of this study could be used as the basic data when the public transportation planning or operation related policies according to the weather condition are concerned.

[Fig 10] The Effect of Weather Conditions on Transit Ridership
Choi SangGi, Rhee JongHo, Oh SeungHwoon

▲ Referenced study said that weather affects subway usage




**Our research has demonstrated the need to consider different characteristics by region through analysis**


Q&A

# THANK YOU