Multivariate Data Analysis
Assignment #5

Dimensionality Reduction for MLR

과목명	다변량분석
담당교수	강필성 교수님
제출일	2019-05-30
이름	박지원
학과명	산업경영공학부
학번	2014170856

목차

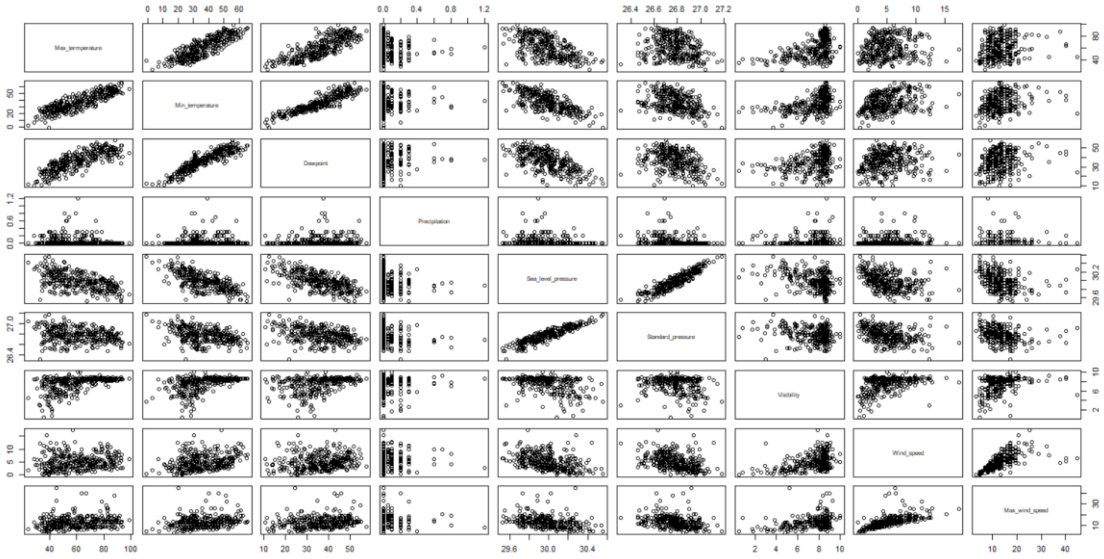
1. 모든 변수를 사용한 MLR 모델

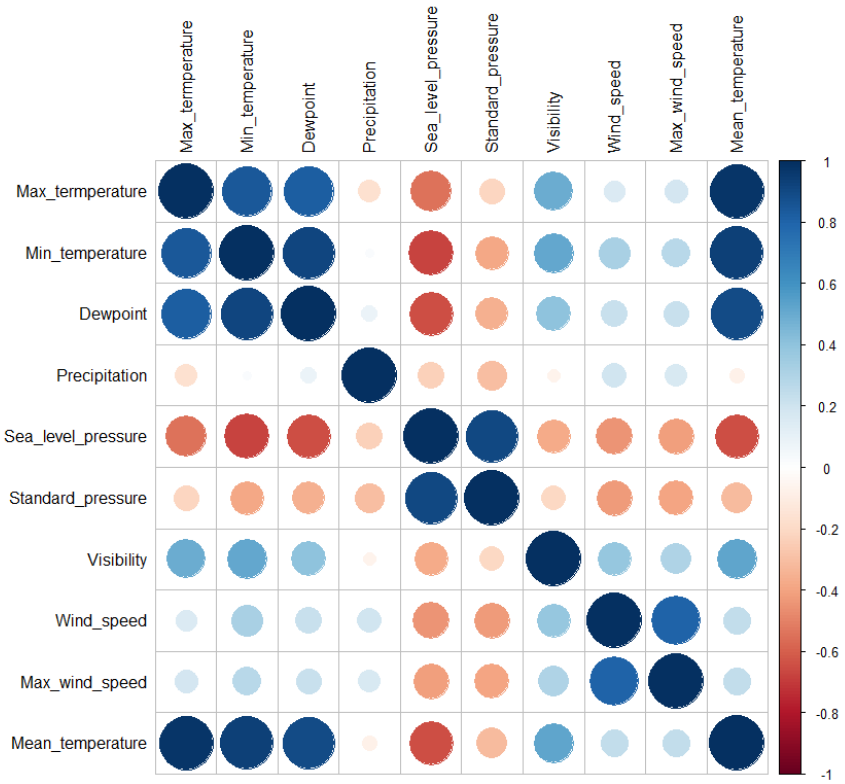
Dataset: Weather Ankara

이 데이터셋은 Ankara 지역의 날씨에 대한 정보를 포함하는 데이터셋이다. Mean_temperature 항목이 종속변수이며 최고기온, 최저기온을 포함한 나머지 9개의 변수들이 설명변수이다. 이 9개의 변수를 이용해 Mean Temperature를 예측하는 MLR 모델을 구축하고자 한다.

1.1. Exploratory Data Analysis

변수들 간의 상관관계를 파악하기 위해 scatterplot과 correlation plot을 그려보았다.





위의 두 plot으로부터 종속변수인 Mean temperature와 가장 상관관계가 높은 변수는 Max Temperature임을 알 수 있다. 다음으로 Min Temperature, 그리고 Dewpoint가 가장 상관관계가 컸다. Sea_level_pressure와는 비교적 강한 음의 상관관계가 있었다. 이 점이 변수 선택에 있어 영향을 미치는지 아래에서 함께 살펴보도록 하겠다. 독립변수간의 상관관계는 Min Temperature와 Dewpoint의 상관관계가 가장 컸다.

1.1. 모든 변수를 사용한 MLR 모델 구축

모델 구축에 앞서 변수들을 정규화하고 임의로 250 개의 Training dataset과 71 개의 Validation dataset으로 구분하였다. 모든 변수를 사용하여 MLR 모델을 구축했고 결과는 아래와 같다.

```
Call:
lm(formula = weather_target ~ ., data = weather_trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0726 -1.0676 -0.1012  0.9663  5.6825

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       48.902181   0.105780 462.302  < 2e-16 ***
Max_termperature   9.467611   0.296915  31.887  < 2e-16 ***
Min_temperature    4.347713   0.325961  13.338  < 2e-16 ***
Dewpoint           0.419139   0.282395   1.484  0.13906
Precipitation      0.002805   0.118801   0.024  0.98118
Sea_level_pressure -2.260351   0.777554  -2.907  0.00399 **
Standard_pressure   1.292050   0.613311   2.107  0.03618 *
Visibility         0.174968   0.142394   1.229  0.22037
Wind_speed         0.155136   0.161138   0.963  0.33664
Max_wind_speed     0.022691   0.136546   0.166  0.86816
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.669 on 240 degrees of freedom
Multiple R-squared:  0.9882,      Adjusted R-squared:  0.9877
F-statistic:  2225 on 9 and 240 DF,  p-value: < 2.2e-16
```

Adjusted R2는 0.9877이었다. 이는 데이터가 꽤 선형성을 띄고 있고, 9개의 변수이 전체 변동의 98.77%를 설명할 수 있음을 나타낸다. 유의수준 1%에서 통계적으로 유의미한 변수들을 파악하기 위해 P-value를 살펴보았다. P-value가 0.01보다 낮다는 것은 coefficient의 값이 0이라는 귀무가설을 기각할 수 있으며 변수가 유의미함을 나타내기 때문이다. P-value가 0.01보다 낮은 변수들은 Max_temperature, Min_temperature, 그리고 Sea_level_pressure였다. 종속변수와 상관관계가 컸던 변수들이 주로 유의미했다.

1.2. Validation Dataset에 대한 Performance Measures

위에서 구축한 모델을 이용해 Validation dataset에 대한 RMSE, MAE, MAPE를 산출했다. 결과는 아래와 같다.

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658

RMSE는 1.339973로, 차이의 제곱의 평균에 루트를 씌운 것이다. 부호의 영향을 제거하기 위해 제곱을 했기 때문에 MAE보다 큰 값이 계산되었다. MAE는 절대평균오차로 실제값과 예측값의 차이의 절댓값의 평균이다. 이 모델의 MAE는 1.042669로 평균적으로 이만큼의 차이가 있었음을 알 수 있다. MAE는 차이의 크기는 제공하지만 y의 스케일에 상관없이 계산된다. 이를 보완한 것이 MAPE로 이 모델의 값은 2.409658이었다. 이는 y값에 비해 얼마나 차이가 있었는지를 나타내는 것이다.

2. Exhaustive Search

2.1. Exhaustive Search를 수행하는 함수 구현 및 결과

Exhaustive Search를 수행하는 함수를 아래와 같이 직접 구현했다.

```
# Variable selection method 1: Exhausitve Search
x_idx <- c(1:9)
bestR2 <- 0

start_time <- proc.time()
for (i in 1:9){
  comb <- combinations(9,i,x_idx)
  for (j in 1:dim(comb)[1]){
    x_idx_selected <- comb[j,]
    tmp_x <- paste(colnames(weather_trn[x_idx_selected]), collapse=" + ")
    tmp_xy <- paste("weather_target ~ ", tmp_x, collapse = "")
    as.formula(tmp_xy)

    es_model <- lm(tmp_xy, data = weather_trn)
    s <- summary(es_model)
    R2_es <- s$adj.r.squared

    if(R2_es >= bestR2){
      bestR2 <- R2_es
      best_formula <- tmp_xy
       }
  }
}
end_time <- proc.time()
end_time - start_time
```

Training dataset에 대한 Adjusted R2 기준으로 가장 높은 값이 산출된 변수 집합은 {Max_termperature + Min_temperature + Dewpoint + Sea_level_pressure + Standard_pressure + Visibility + Wind_speed} 이었다. 전체 9개 변수에서 Precipitation과 Max Windspeed가 제외된 7개 변수였다. 이 변수 조합으로 구축한 모델 결과는 다음과 같다.

```
Call:
lm(formula = best_formula, data = weather_trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0737 -1.0642 -0.1116  0.9654  5.6535

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9029     0.1052 464.749  < 2e-16 ***
Max_termperature    9.4713     0.2810  33.703  < 2e-16 ***
Min_temperature     4.3391     0.3205  13.536  < 2e-16 ***
Dewpoint            0.4258     0.2758   1.544  0.12391
Sea_level_pressure -2.2618     0.7692  -2.941  0.00359 **
Standard_pressure   1.2913     0.6086   2.122  0.03488 *
Visibility          0.1744     0.1415   1.232  0.21919
Wind_speed          0.1702     0.1320   1.289  0.19854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.662 on 242 degrees of freedom
Multiple R-squared:  0.9882,      Adjusted R-squared:  0.9878
F-statistic:  2884 on 7 and 242 DF,  p-value: < 2.2e-16
```

2.1.1 Training Dataset에 대한 Adjusted R2 및 소요시간

	Adjusted R2	소요시간
Full Model	0.9877	
Exhaustive Search	0.9878	1.44

Adjusted R2값은 0.9878이었다. 모든 변수를 사용했을 때의 Adjusted R2는 0.9877이는데, 선택된 변수들만을 사용했을 때에는 Adjusted R2값이 증가했다. 이는 모델 구축에 있어 크게 유의미하지 않은 변수가 있었기 때문으로 예상된다. Exhausitve Search에 소요된 시간은 1.44초였다.

2.1.2 Validation Dataset에 대한 Performance Measures

학습한 모형을 이용하여 Validation dataset에 대한 RMSE, MAE, MAPE를 산출하였다. 결과는 아래와 같다.

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658
Exhaustive Search	1.337575	1.039726	2.403551

모든 변수를 사용한 MLR 모형의 결과와 비교했을 때, RMSE, MAE, MAPE가 모두 감소한 것을 알 수 있다. Exhaustive Search는 가능한 모든 변수 집합을 탐색해 Adjusted R2 기준 가장 최적의 조합을 선택하기 때문에 에러가 모두 낮아진 것으로 보인다.

3. Variable Selection Methods

3.1. Forward Selection을 사용한 변수 선택

우선 Forward Selection 방식을 사용하여 MLR 변수 선택 과정을 수행해보았다. Max_termperature, Min_temperature, Sea_level_pressure, Standard_pressure 그리고 Wind_speed가 선택되었다. 이 변수들로 모델을 구축한 결과는 아래와 같다.

```
Call:
lm(formula = weather_target ~ Max_termperature + Min_temperature +
    Sea_level_pressure + Standard_pressure + Wind_speed, data =
weather_trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1663 -1.0726 -0.0918  1.0451  5.9289

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9013     0.1055 463.628  < 2e-16 ***
Max_termperature    9.5078     0.2762  34.421  < 2e-16 ***
Min_temperature     4.5762     0.2700  16.948  < 2e-16 ***
Sea_level_pressure -2.7132     0.7261  -3.737 0.000232 ***
Standard_pressure   1.6083     0.5824   2.761 0.006192 **
Wind_speed          0.1796     0.1196   1.502 0.134485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.666 on 244 degrees of freedom
Multiple R-squared:  0.988,       Adjusted R-squared:  0.9878
F-statistic:  4017 on 5 and 244 DF,  p-value: < 2.2e-16
```

3.1.1 Training Dataset에 대한 Adjusted R2 및 소요시간 비교

	Adjusted R2	소요시간
Full Model	0.9877	
Exhaustive Search	0.9878	1.44
Forward Selection	0.9878	0.12

Training dataset에 대한 Adjusted R2는 0.9878이었다. 이 역시 Full Model보다 살짝 개선되었다. 소요 시간은 0.12초로, Exhaustive Search의 1.44초보다 짧았다. Exhaustive Search는 모든 변수 조합을 탐색하는 데 반해 Forward Selection은 부분만을 탐색하기 때문에 짧게 계산되었다.

3.1.2 Validation Dataset에 대한 Performance Measures

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658
Exhaustive Search	1.337575	1.039726	2.403551
Forward Selection	1.388217	1.064263	2.477107

Full Model, Exhaustive Search와 비교했을 때 RMSE, MAE, MAPE가 증가했다. Exhaustive Search와 다르게 모든 조합을 탐색하지 않고 빠른 시간내에 효율적으로 찾다 보니 소요시간은 개선되었지만 performance는 살짝 감소하였다.

3.2. Backward Elimination을 사용한 변수 선택

다음으로 Backward Elimination 방식을 사용해 변수 선택을 해보았다. Forward Selection과 마찬가지로 Max_termperature, Min_temperature, Sea_level_pressure, Standard_pressure 그리고 Wind_speed가 선택되었다. 이 변수 조합으로 모델을 구축한 결과는 아래와 같다.

```
Call:
lm(formula = weather_target ~ Max_termperature + Min_temperature +
    Sea_level_pressure + Standard_pressure + Wind_speed, data =
weather_trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1663 -1.0726 -0.0918  1.0451  5.9289

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9013     0.1055 463.628  < 2e-16 ***
Max_termperature    9.5078     0.2762  34.421  < 2e-16 ***
Min_temperature     4.5762     0.2700  16.948  < 2e-16 ***
Sea_level_pressure -2.7132     0.7261  -3.737 0.000232 ***
Standard_pressure   1.6083     0.5824   2.761 0.006192 **
Wind_speed          0.1796     0.1196   1.502 0.134485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.666 on 244 degrees of freedom
Multiple R-squared:  0.988,       Adjusted R-squared:  0.9878
F-statistic:  4017 on 5 and 244 DF,  p-value: < 2.2e-16
```

3.2.1 Training Dataset에 대한 Adjusted R2 및 소요시간 비교

	Adjusted R2	소요시간
Full Model	0.9877	
Exhaustive Search	0.9878	1.44
Forward Selection	0.9878	0.12
Backward Elimination	0.9878	0.11

Forward Selection과 같은 변수 조합이므로 Adjusted R2값 역시 동일했다. 단, 소요시간이 0.11초로 Exhaustive Search나 Forward Selection보다 빨랐다.

3.2.2 Validation Dataset에 대한 Performance Measures

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658
Exhaustive Search	1.337575	1.039726	2.403551
Forward Selection	1.388217	1.064263	2.477107
Backward Elimination	1.388217	1.064263	2.477107

Forward Selection과 변수 조합이 동일하여 같은 RMSE, MAE, MAPE값이 산출되었다.

3.3. Stepwise Selection을 사용한 변수 선택

다음으로 Stepwise Selection 방식을 사용해 변수 선택을 해보았다. Forward Selection, Backward Elimination과 마찬가지로 Max_termperature, Min_temperature, Sea_level_pressure, Standard_pressure 그리고 Wind_speed가 선택되었다. 이 변수 조합으로 모델을 구축한 결과는 아래와 같다.

```
Call:
lm(formula = weather_target ~ Max_termperature + Min_temperature +
    Sea_level_pressure + Standard_pressure + Wind_speed, data =
weather_trn)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1663 -1.0726 -0.0918  1.0451  5.9289

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9013     0.1055 463.628  < 2e-16 ***
Max_termperature    9.5078     0.2762  34.421  < 2e-16 ***
Min_temperature     4.5762     0.2700  16.948  < 2e-16 ***
Sea_level_pressure -2.7132     0.7261  -3.737 0.000232 ***
Standard_pressure   1.6083     0.5824   2.761 0.006192 **
Wind_speed          0.1796     0.1196   1.502 0.134485
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.666 on 244 degrees of freedom
Multiple R-squared:  0.988,       Adjusted R-squared:  0.9878
F-statistic:  4017 on 5 and 244 DF,  p-value: < 2.2e-16
```

3.3.1 Training Dataset에 대한 Adjusted R2 및 소요시간 비교

	Adjusted R2	소요시간
Full Model	0.9877	
Exhaustive Search	0.9878	1.44
Forward Selection	0.9878	0.12
Backward Elimination	0.9878	0.11
Stepwise Selection	0.9878	0.12

다른 방식들로 선택된 변수 조합과 같은 변수 조합이므로 Adjusted R2값 역시 동일했다. 이번에는 소요시간이 0.12초로 Forward Selection, Backward Elimination과 비슷하게 걸렸다.

3.3.2 Validation Dataset에 대한 Performance Measures

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658
Exhaustive Search	1.337575	1.039726	2.403551
Forward Selection	1.388217	1.064263	2.477107
Backward Elimination	1.388217	1.064263	2.477107
Stepwise Selection	1.388217	1.064263	2.477107

Forward Selection, Backward Elimination 방식과 변수 조합이 동일하여 같은 RMSE, MAE, MAPE값이 산출되었다.

4. Genetic Algorithm을 이용한 변수선택

4.1 GA를 이용한 변수 선택

Adjusted R2를 Fitness function으로 하는 Genetic Algorithm 기반의 변수 선택 함수를 아래와 같이 작성해보았다.

```
# Fitness function: F1 for the training dataset
fit_F1 <- function(string){
  sel_var_idx <- which(string == 1)
  # Use variables whose gene value is 1
  sel_x <- x[, sel_var_idx]
  xy <- data.frame(sel_x, y)
  # Training the model
  GA_lr <- lm(y ~ ., data = xy)
  s <- summary(GA_lr)
  R2 <- s$adj.r.squared
  #GA_lr_pred <- predict(GA_lr, type = "response", newdata = xy)

  return(R2)
}
```

작성한 함수를 이용하여 GA를 이용한 변수 선택을 수행했다. Population size는 100, crossover rate는 0.5, mutation rate는 0.01, maximum iteration은 100으로 설정했다. 선택된 변수들은 Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed였다. Exhaustive Search와 같은 결과였으며 Forward Selection, Backward Elimination, Stepwise Selection으로 선택된 변수들보다 2개 더 많은 변수가 선택되었다. 선택된 7개의 변수로 모델을 구축했다. 결과는 다음과 같다.

```
Call:
lm(formula = weather_target ~ ., data = GA_trn_data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0737 -1.0642 -0.1116  0.9654  5.6535

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        48.9029     0.1052 464.749  < 2e-16 ***
Max_termperature    9.4713     0.2810  33.703  < 2e-16 ***
Min_temperature     4.3391     0.3205  13.536  < 2e-16 ***
Dewpoint            0.4258     0.2758   1.544  0.12391
Sea_level_pressure -2.2618     0.7692  -2.941  0.00359 **
Standard_pressure   1.2913     0.6086   2.122  0.03488 *
Visibility          0.1744     0.1415   1.232  0.21919
Wind_speed          0.1702     0.1320   1.289  0.19854
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.662 on 242 degrees of freedom
Multiple R-squared:  0.9882,      Adjusted R-squared:  0.9878
F-statistic:  2884 on 7 and 242 DF,  p-value: < 2.2e-16
```

4.1.1 Training Dataset에 대한 Adjusted R2 및 소요시간 비교

	Adjusted R2	소요시간
Full Model	0.9877	
Exhaustive Search	0.9878	1.44
Forward Selection	0.9878	0.12
Backward Elimination	0.9878	0.11
Stepwise Selection	0.9878	0.12
Genetic Algorithm	0.9878	11.12

다른 방식들로 선택된 변수 조합과 같은 변수 조합이므로 Adjusted R2값 역시 동일했다. 이번에는 소요시간이 11.12초로 다른 방식들과 비교해서 가장 길었다.

4.1.2 Validation Dataset에 대한 Performance Measures

	RMSE	MAE	MAPE
Full Model	1.339973	1.042669	2.409658
Exhaustive Search	1.337575	1.039726	2.403551
Forward Selection	1.388217	1.064263	2.477107
Backward Elimination	1.388217	1.064263	2.477107
Stepwise Selection	1.388217	1.064263	2.477107
Genetic Algorithm	1.337575	1.039726	2.403551

Exhaustive Search와 변수 조합이 동일하여 같은 RMSE, MAE, MAPE값이 산출되었다. 결과를 종합해서 비교했을 때, Exhaustive Search와 Genetic Algorithm은 RMSE, MAE, MAPE 측면에서 가장 우수한 성능을 보였다. 하지만 소요시간이 각각 1.44초와 11.12초로 다른 방식에 비해 매우 길다는 단점이 있었다. Forward Selection, Backward Elimination, Stepwise Selection은 최적의 답을 찾지는 못하였지만 소요시간이 매우 짧았다.

4.2 GA의 하이퍼파라미터 변경

위에서는 Population size 100, crossover rate 0.5, mutation rate 0.01, maximum iteration은 100으로 설정했다. 여기서 하이퍼파라미터들을 하나씩 변경해보고 변수 선택 결과를 비교해 최종 결과에 가장 큰 영향을 미치는 하이퍼파라미터를 알아보고자 한다. 하이퍼파라미터들의 효과를 보기 위해 elitism 옵션은 제거하고 실행했다.

4.2.1 Population size 변경에 따른 결과 변화

Population size를 25, 50, 75, 100, 200으로 변경시켜 보았다. 모든 경우에 대해 선택된 변수 조합은 {Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed}였다.

4.2.2 Cross-over rate 변경에 따른 결과 변화

Cross-over rate를 0.1, 0.3, 0.5, 0.7, 0.9로 변경시켜 가며 결과를 살펴보았다. 모든 경우에 대해 {Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed} 조합이 선택되었다.

4.2.3 Mutation rate 변경에 따른 결과 변화

Mutation rate를 0.001, 0.005, 0.01, 0.1로 변경시켜 가며 결과를 살펴보았다. 이 때에도 모든 경우에 대해 {Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed} 조합이 선택되었다.

4.2.4 Maximum iteration 변경에 따른 결과 변화

Maximum iteration을 0.001, 0.005, 0.01, 0.1로 변경시켜 가며 결과를 살펴보았다.

Maximum iteration	선택된 변수 조합
5	Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Wind_speed
10	Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Wind_speed
20	Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed
100	Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed
200	Max_termperature, Min_temperature, Dewpoint, Sea_level_pressure, Standard_pressure, Visibility, Wind_speed

Maximum iteration이 5와 10이었을 때에는 Visibility가 선택되지 않았지만 20 이상부터는 Visibility가 선택되었다. 이는 Exhaustive Search가 찾은 최적해와 동일한 결과였다.

4.2.5. 결과 비교

다른 하이퍼파라미터들은 변경해도 결과에 변화가 없었던 반면, maximum iteration은 변화가 있었다. 이 데이터셋에 대해서는 maximum iteration이 결과에 가장 큰 영향을 미치는 하이퍼파라미터였다. 하지만 이러한 결과는 데이터셋이 거대한 편에 속하지 않았기 때문으로 판단된다. 변수의 개수가 훨씬 많은 경우라면 다른 하이퍼파라미터들이 더 큰 영향을 미칠 수 있을 것이다.