Basic Statistics

2023122054
오상호

Iris 데이터의 구조를 head() 확인한 결과 아래와 같이 도출되었다.

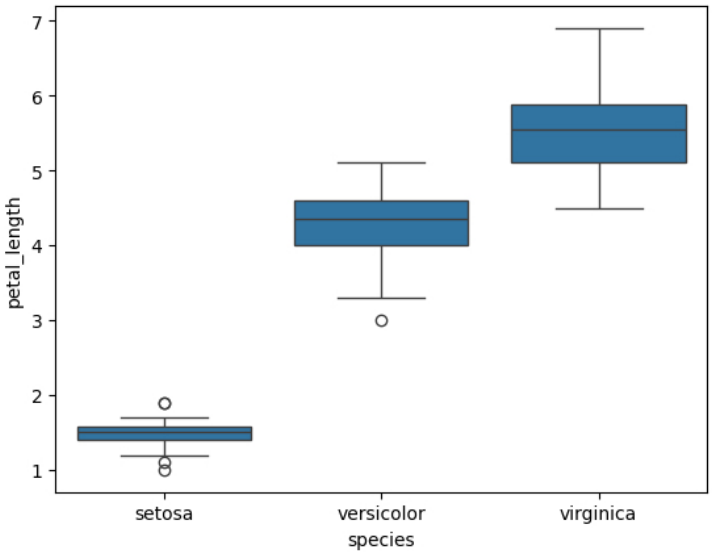
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

즉, Iris 데이터는 종 별 sepal_length, sepal_width, petal_length, petal_width로 구성되어 있다는 것을 알 수 있다. Iris 데이터의 종의 종류를 확인하기 위해 'species' 열에 대해 unique를 적용하여, Iris 데이터는 **'setosa', 'versicolor', 'virginica'** 와 같이 세 종류의 종으로 구성되어있다는 것을 확인할 수 있었다.

각 species의 개수와 petal length 통계량을 summary를 통해 도출한 결과

	count	mean	std	min	25%	50%	75%	max
setosa	50.0	1.462	0.173664	1.0	1.4	1.50	1.575	1.9
versicolor	50.0	4.260	0.469911	3.0	4.0	4.35	4.600	5.1
virginica	50.0	5.552	0.551895	4.5	5.1	5.55	5.875	6.9

위와 같이 도출되었다. 따라서 각 종별로 50개씩 동일한 개수로 이루어져 있으며, petal length의 max, min, IQR, mean 값을 확인해보았을 때, setosa, versicolor, virginica순으로 petal length의 분포가 점점 길어진다는 것을 예측할 수 있다. 또한, 해당 순서대로 std역시 증가한다. 이러한 petal length에 대해 Box plot으로 시각화를 해보면



summary를 통해서 예상한 것과 마찬가지로 setosa, versicolor, virginica순으로 petal length의 분포가 점점 길어진다는 것을 시각적으로 확인할 수 있었고, 해당 순서대로 std역시 증가한다는 것을 확인할 수 있었다. 이는 setosa의 petal length가 가장 짧으면서, 대부분의 개체들이 비슷한 길이를 가

진다는 것을 의미한다. 실제로 setosa->versicolor->virginica로 옮겨갈수록 box의 폭과 whisker의 폭이 넓어진다는 것을 확인할 수 있고, setosa->versicolor->virginica로 옮겨갈수록 점점 개체들의 petal length가 균일해지지 않는다는 것을 의미한다. 또한 setosa->versicolor->virginica로 옮겨갈수록 outlier가 없어지는 것을 확인할 수 있는데, 이를 통해서 q1-q3사이의 폭과 std가 커지면서 대부분의 개체가 whisker사이에 들어오게 된다는 것을 파악할 수 있다.

이러한 petal length의 길이에 대해 통계적으로 신뢰도 있는 예측을 하기 위해서는 ANOVA를 시행해야한다. 본 과제에서는 One-way ANOVA를 시행할 것이고, 이를 시행하기 위해서는 독립성, 정규성, 등분산성 조건을 만족하여야한다. 따라서 Iris 데이터에 대해 정규성 검정과 등분산성 검정을 시행하여야 One-way ANOVA를 수행할 수 있다.

종 별 Shapiro-Wilk(정규성 검정)을 시행하기 위해 다음과 같은 가설의 수립하였다.

H_0 : Data가 정규성을 띈다.

H_a : Data가 정규성을 띄지 않는다.

이러한 정규성 검정을 시행한 결과 아래와 같은 결과가 도출된다.

```
Species : setosa , Statistics : 0.9549767850318984 , p-value : 0.05481146719553462
Species : versicolor , Statistics : 0.9660044025433202 , p-value : 0.15847783815657984
Species : virginica , Statistics : 0.9621864428612805 , p-value : 0.10977536903223795
```

세 종의 p-value는 모두 0.05를 초과한다는 것을 알 수 있다. 그러므로 귀무가설을 기각하지 못하고, 세 종 모두 정규성 조건을 만족한다는 것을 알 수 있다.

이어서 세 종 간의 등분산성을 검정해야하기 때문에 등분산성 검정 (Levene)을 시행하였다.
해당 검정의 가설은 다음과 같다.

H_0 : 세 종 간의 Data는 같은 분산을 가진다.

H_a : 적어도 한 종의 Data는 다른 분산을 가진다.

아래는 등분산성 검정의 결과이다.

```
Statistics: 19.480338801923573, p-value: 3.1287566394085344e-08
```

p-value는 0.05미만으로 도출된다는 것을 알 수 있으므로, 귀무가설이 기각되고 대립가설이 성립한다는 것을 알 수 있다. 그러므로 Iris데이터는 등분산성 조건을 만족시키지 못한다.

원칙적으로는 등분산성 가정이 성립하지 않으므로 One-way ANOVA를 수행할 수 없지만 과제 명세상 이를 무시하고 One-way ANOVA를 수행하였다.

ANOVA의 가설은 다음과 같다.

H_0 : 세 Species 간 petal length는 동일하다.

H_a : 적어도 한 Species의 petal length는 다르다.

아래는 One-way ANOVA를 수행한 결과이다.

```
F-statistic: 1180.161182252981
p-value: 2.8567766109615584e-91
```

p-value는 0.05미만으로 도출된다는 것을 알 수 있으므로, 귀무가설이 기각되고 대립가설이 성립한다는 것을 알 수 있다. 그러므로 Iris데이터에서 세 Species 간 petal length는 동일하지 않다는 것을 결론으로 내릴 수 있다.

이러한 One-way ANOVA의 결과에 대해 통계적인 신뢰도를 높이기 위해 사후검정(Turkey HSD)를 수

행하였다. 결과는 아래와 같다.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
setosa	versicolor	2.798	0.0	2.5942	3.0018	True
setosa	virginica	4.09	0.0	3.8862	4.2938	True
versicolor	virginica	1.292	0.0	1.0882	1.4958	True

각 species 별로 pairwise하게 검정을 실시한 것이고 각각 귀무가설은 두 종간 petal length가 동일하다는 것이었다. 검정 결과 위처럼 귀무가설이 모두 reject되었고, 이에 따라 모든 species 간 petal length가 유의하게 다르다고 결론내릴 수 있었다.
해당 표의 경우, p-value를 나타낼 수 있는 정도에 한계가 있었기 때문에 정확한 p-value가 도출되지 않았는데, post.pvalues를 이용하여 print 한 결과 각각 2.13162821e-14, 2.13162821e-14 ,2.13162821e-14로 도출되었다. 즉, 유의수준 0.05 기준으로 모든 종 사이의 귀무가설이 기각되며, petal length가 유의미한 차이가 존재한다는 것으로 결론내릴 수 있다.

결론적으로 Boxplot의 결과를 통해서 setosa->versicolor->virginica로 옮겨갈수록 Petal Length가 길어진다고 시각적으로 예상할 수 있었으며, std와 whisker, outlier의 분포를 통해서 iris내의 종 내의 petal length의 균일성 역시 예상해 볼 수 있었다. 이를 예상이 통계적으로 신빙성이 있는지 확인하기 위해 One-way ANOVA를 설계하였고, 정규성 조건은 만족하고 등분산성 조건은 만족하지 않았으나, 과제 명세상 One-way ANOVA를 시행할 수 있다고 판단하여 수행하였다. One-way ANOVA를 수행한 결과, 세 Species 간 petal length는 동일하지 않다는 것을 결론내릴 수 있었다. 이러한 ANOVA결과를 바탕으로 사후검정을 실시하였는데, 사후검정은 각 species 별로 pairwise하게 petal length에 대해 검정하는 것이었다. 이 역시 모든 종별 조합에서 귀무가설을 기각할 수 있었고, ANOVA의 결론을 유의수준 0.05이내에서 모든 species 간 petal length가 유의하게 다르다고 결론내릴 수 있었다.
추가적으로 One-way ANOVA대신 Welch ANOVA를 사용한다면 등분산성 가정이 성립하지 않아도 더 정확한 ANOVA결과를 얻을 수 있으므로, 더 통계적으로 신뢰성 높은 결론을 얻을 수 있을 것으로 기대된다.

회귀분석 결과

Train: Test 비율을 7:3으로 설정하여 petal_length에 대해 문항에 주어진 조건대로 회귀 분석을 시행하였다. 시행 결과는 다음과 같다.

```
MSE: 0.12354507045731629
R^2: 0.9585127726373215
Intercept: 0.03688942292122954

Coefficients:
     feature      coef
 petal_width  1.472890
sepal_length  0.697919
 sepal_width -0.691295
```

Test set에서 $MSE \approx 0.1235$, $R^2 = 0.9585$로 도출되었다. 이는 모델이 petal_length 변동의 약 95.9%를 설명한다는 것을 의미하고 이를 통해 입력변수(sepal_length, sepal_width, petal_width)만으로도 petal_length가 잘 예측된다고 판단할 수 있었다.

추정된 coefficient들을 바탕으로 회귀분석 식을 세우면

$\widehat{Petal\ length} = 1.472890 \cdot petal\ width + 0.697919 \cdot sepal\ length - 0.691295 \cdot sepal\ width + 0.036889$

다음과 같다.

추정된 회귀계수는 다른 변수들을 고정했을 때의 영향이므로 각 변수별로 아래와 같이 분석할 수 있었다. (절편의 경우에는 다른 모든 변수들이 0일 때의 값이다.)

-petal_width(계수 1.4729): petal_width가 1 증가하면 petal_length는 평균적으로 약 1.47 증가한다. 세 변수 중 가장 영향이 크다.

-sepal_length(계수 0.6979): sepal_length가 1 증가하면 petal_length는 평균적으로 약 0.70 증가한다.

-sepal_width(계수 -0.6913): sepal_width가 1 증가하면 petal_length는 평균적으로 약 0.69 감소한다(다른 변수 통제 후 음의 관계).

다만, length와 width는 둘 다 꽃의 성장과 관련된 것이므로 입력변수(sepal_length, sepal_width, petal_width)간에 상관관계가 크게 존재할 수도 있다고 판단된다. 이 경우 다중공선성으로 인해 회귀계수 추정이 불안해지고 해석이 어려워질 가능성이 있다. 따라서 추후에 VIF를 통해 다중공선성의 정도를 파악하고, 다중공선성이 크다면 PCA나 Ridge, Lasso, Elastic Net등을 적용하여 모델의 안정성과 일반화 성능을 보완할 수 있을 것이다.