|  |
| --- |
| REPORT |
| **과제 4**  사례기반추론과 의사결정나무 |

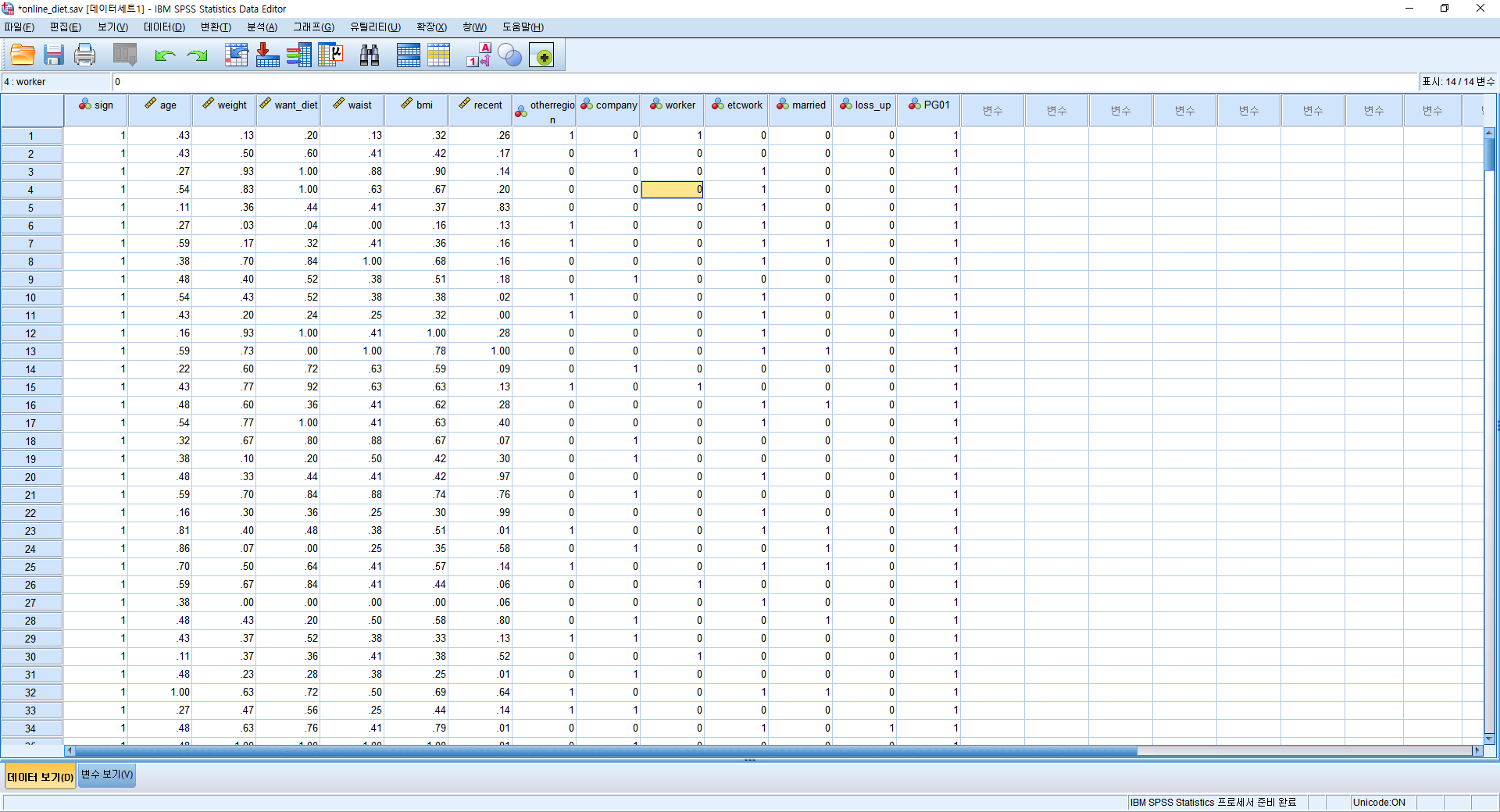




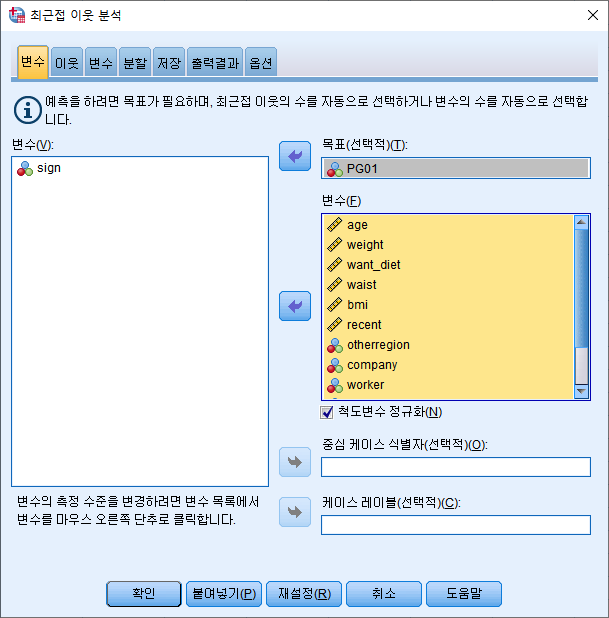
|  |  |  |
| --- | --- | --- |
| 제출일 | 2020. 04. 25 |  |
| 과목명 | 비즈니스애널리틱스개론 |  |
| 담당교수 | 안현철 |  |
| 전공 | 경영정보전공 |  |
| 학번 | 20175288 |  |
| 이름 | 최광삼 |  |

**목차**

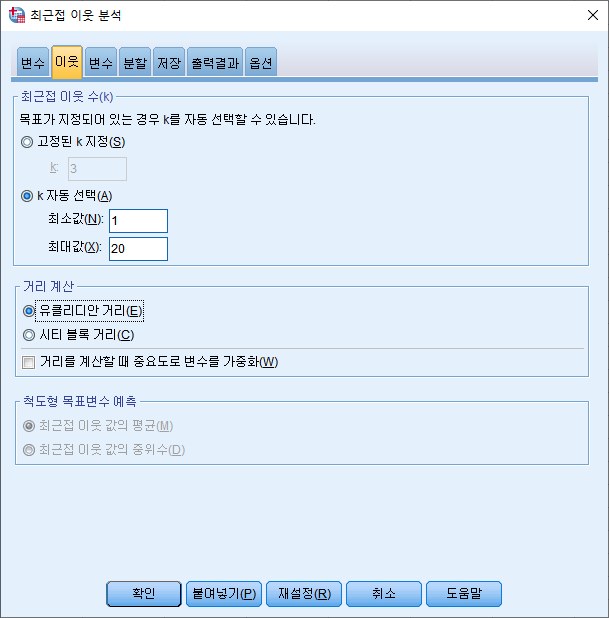
1. **사례기반추론**
2. SPSS 설정값
3. 결과
4. **의사결정나무**
   1. SPSS 설정값
   2. 결과
5. **사례기반추론**

사례기반추론은 과거경험에 기반해, 현재 주어진 문제에 대한 해결책을 제시하는 기법으로서 SPSS에서도 ‘분류분석 – 최근접 이웃’이라는 명칭으로 최근접 이웃 분석을 제공하고 있다. 이 보고서에서는 전략제품 구매여부 데이터셋을 이용하여 SPSS를 통한 사례기반추론을 진행하였다.

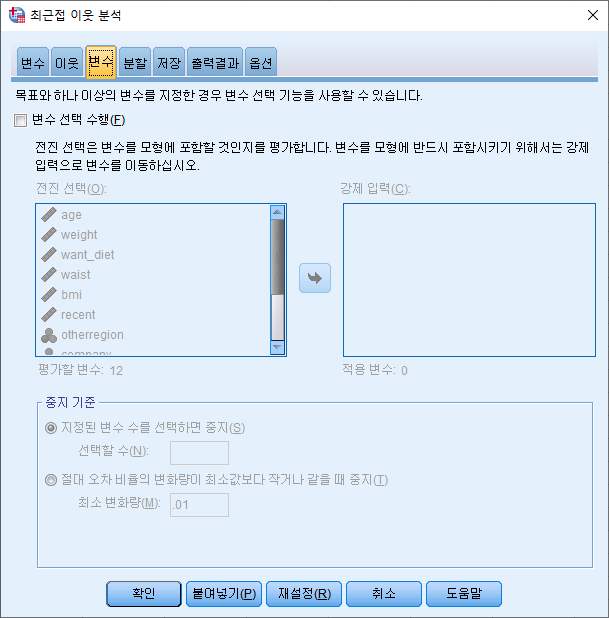
1. SPSS 설정값



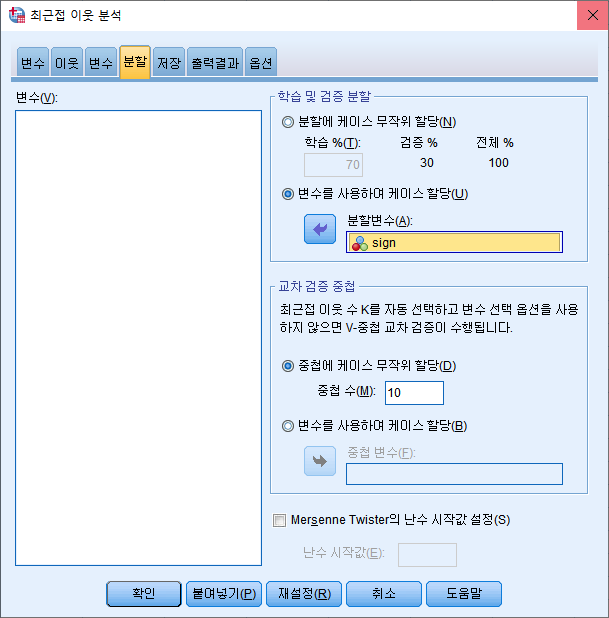
최근접 이웃 분석의 ‘변수’ 탭에서 예측을 위해 사용될 변수는 로지스틱스 회귀분석을 위하여 실시한 독립표본t검정과 카이제곱검정에서 분석에 사용되었을 때 유의미하다고 판단한 12개의 변수를 사용하였다. 또한 목표 변수에는 전략제품 구매여부(구매하였으면 1, 아니면 0)의 변수인 PG01을 사용하였다.



‘이웃’ 탭에서는 최근접 이웃 수(k)를 지정할 수 있다. k 자동선택을 선택한 후 최소값에 1, 최대값에 20을 넣어 최대 20번까지 사례를 비교하고 가장 적합한 사례를 찾을 수 있도록 설정하였다. 거리 계산법에는 기본값인 유클리디안 거리를 선택하였다.



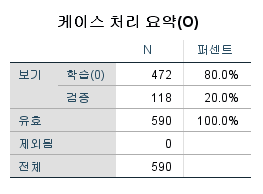
‘변수(특징)’ 탭에서는 사례기반추론을 하는데 있어 어떤 특징 변수를 사용할 것인지 선택할 수 있는데, 이미 12개의 유의한 변수를 사용하도록 지정하였으므로 특별히 설정하지 않았다.

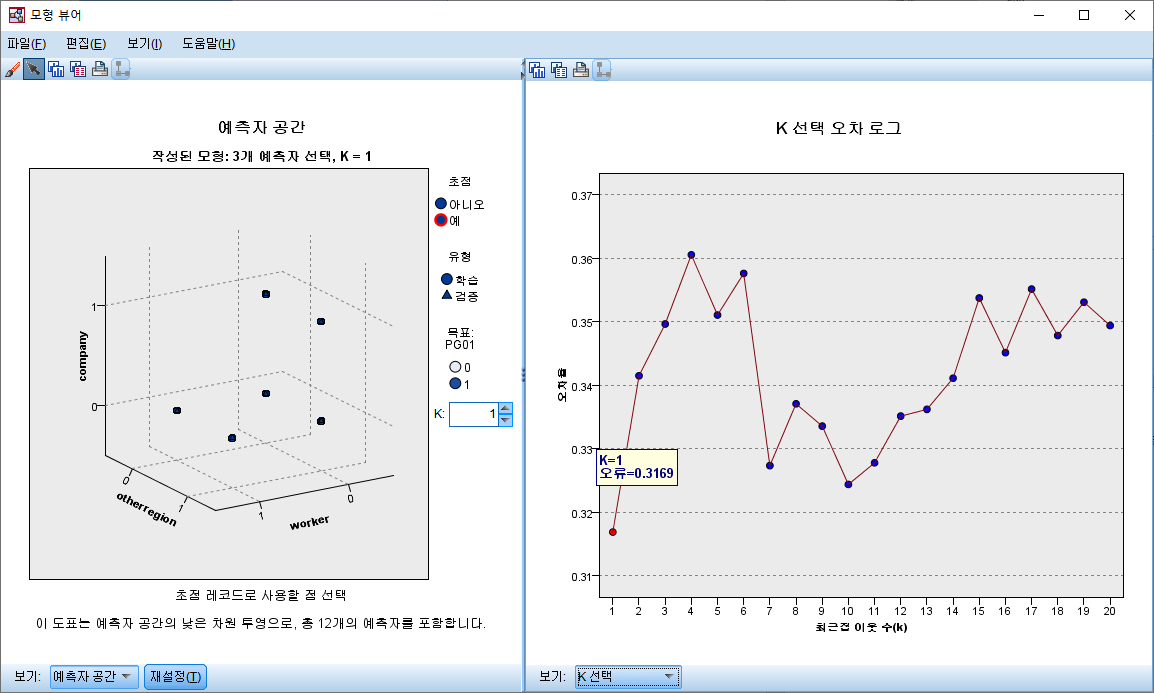


‘분할’ 탭에서 학습용 샘플과 검증용 샘플을 구분 짓게 할 수 있다. 이 데이터셋에서는 sign 변수를 통해서 1(학습용)과 0(검증용)으로 학습용 데이터와 검증용 데이터를 구분해 놓았기 때문에 변수를 사용하여 케이스 할당 기능을 사용하여 sign을 분할변수로 사용하였다.

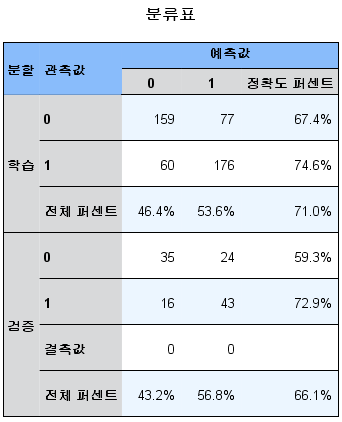
이상의 방법으로 설정이 다 끝난 뒤에는 확인을 눌러 최근접 이웃 분석을 실시하였다.

1. 결과





최근접 이웃 분석 결과 k가 1일 때 가장 정확하게 예측해내는 것으로 나타났다. K 선택 오차 로그를 살펴보면, 20번의 사례 중에서 첫 번째 사례에서 오류율이 0.3169로 예측돼 가장 적은 오차율을 보이는 것을 확인할 수 있다.

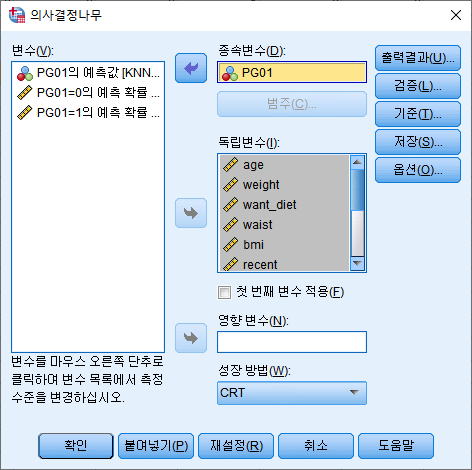


분류표를 살펴보았을 때는 k가 1일 때 학습용 데이터에서 71%, 검증용 데이터에서 66.1%의 정확도를 보이는 것으로 나타났다.

**II. 의사결정나무**

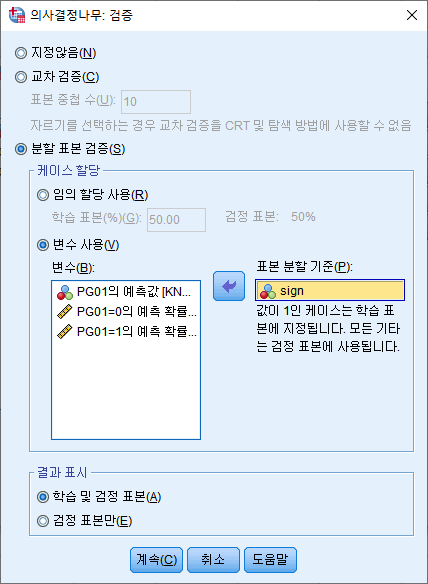
의사결정나무는 일련의 독립변수들을 활용하여, 분류를 하거나 예측을 수행하는 기법으로서 SPSS에서는 ‘분류분석 – 의사결정나무’이라는 명칭으로 의사결정나무 분석을 제공하고 있다. 의사결정나무 분석에는 앞서 실시한 사례기반추론에서 사용한 것과 동일한 전략제품 구매여부 데이터셋을 이용하여 분석을 진행하였다.

1. SPSS 설정값



의사결정나무 분석을 함에 있어 먼저 종속변수와 독립변수를 설정해주어야 한다. 종속변수에는 전략제품 구매여부를 1과 0으로 분류한 PG01 변수를 사용하였고, 독립변수에는 로지스틱스 회귀분석에서 독립검정t검정 및 카이제곱검정을 실시한 결과 분석에 사용했을 때 유의미하다고 판단한 12개의 변수를 사용하였다.

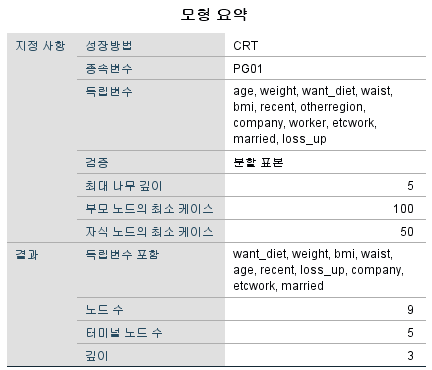
나무 성장 방법에는 CART 기법인 CRT를 적용하였다.



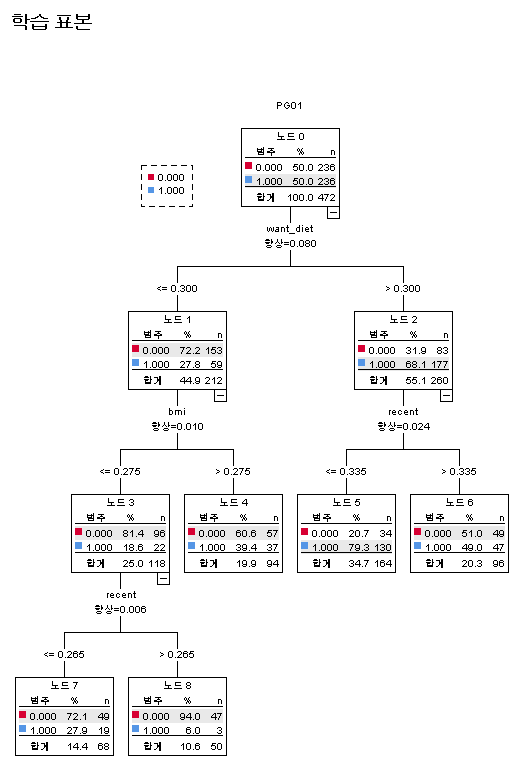
검증 설정에 대해서는 샘플을 미리 나누어 놓았으므로 분할 표본 검증을 선택하여sign 변수에 저장된 값(1: 학습용, 0: 검증용)을 사용하여 학습용 데이터와 검증용 데이터를 구분 짓게 하였다.

기준 등 나머지 설정에 대해서는 기본값으로 지정하여 의사결정나무 분석을 실행하였다.

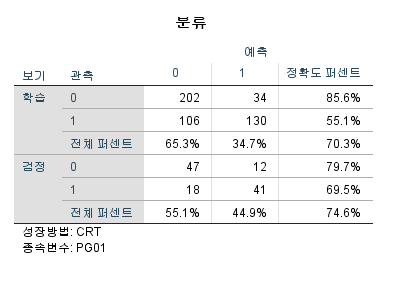
1. 결과



결과를 보면, 의사결정나무 분석에서 사용된 12개 변수 중 실제 분석에서는 9개의 변수를 사용하여 결과를 예측하였음을 알 수 있다.



분석된 의사결정나무에서는 want\_diet, bmi, recent 변수를 사용하여 want\_diet의 값이 0.3 이하인 경우와 0.3 초과인 경우를 1차로 분류하였다. want\_diet의 값이 0.3 이하인 경우에서는 다시 bmi가 0.275 이하인 경우와 0.275 보다 큰 경우를 둘로 나누었고, bmi가 0.275 이하인 경우에 있어서는 recent 변수의 값이 0.265 이하인 경우와 0.265 보다 큰 경우로 나누었다. Want\_diet의 값이 0.3 초과인 경우에 대해서는 recent 변수의 값이 0.335 이하인 경우와 0.335 초과인 경우를 둘로 나누었다.



의사결정나무를 통해서 도출된 예측 정확도는 학습용 데이터의 경우 70.3%의 정확도를 보였으며, 검정용 데이터의 경우는 74.6%의 정확도를 보였다.