DATA MINING TERM PROJECT

**HUFSURANCE**



201400944 김홍범

201601785 신지해

201601220 박상아

**목 차**

**1. 프로젝트 목표**

**2. 타깃 업무 및 변동사항**

**3. Data Preprocessing 전처리**

**4. 연관성 분석**

**5. 군집 분석(clustering)**

**6. 의사결정나무 분석**

**7. 프로젝트 결과 및 소감**

1. 프로젝트 목표

올해로 국민건강보험 전국민 가입이 30주년을 맞았다. 그렇다면 이처럼 국민의 삶에 필수불가결한 국민건강보험은 적절한 방향으로 운영되고 있을까? 실제 우리나라의 경우 국민이 받는 의료혜택은 보험료의 62%정도에 불과하고, 국가별 의료비 공공지출 비율이 OECD 평균인 73%를 밑도는 58.2%라고 한다. 따라서 점진적인 개혁이 요구된다고 할 수 있으며, 본 프로젝트는 건강보험의 수요자인 국민의 관점에서 보험 운영의 개선점을 제시하려 한다.

2. 타깃 업무 및 변동사항

본 프로젝트의 조원별 타깃 업무는 다음과 같다.

1. 김홍범 : 클러스터링, 프로젝트 발표

2. 신지해 : 데이터 전처리, 의사결정 나무

3. 박상아 : 데이터 전처리, 연관성 분석

또한, 최종 발표 이후 변동사항은 다음과 같다.

1. 최종 발표 당시 미비했던 클러스터링을 추가했다.

2. 의사결정나무의 목표변수인 진료의 ‘저가형’, ‘고가형’분류를 보다 균등하게 조정했다. (기존 기준 30만원 -> 10만원 하향)

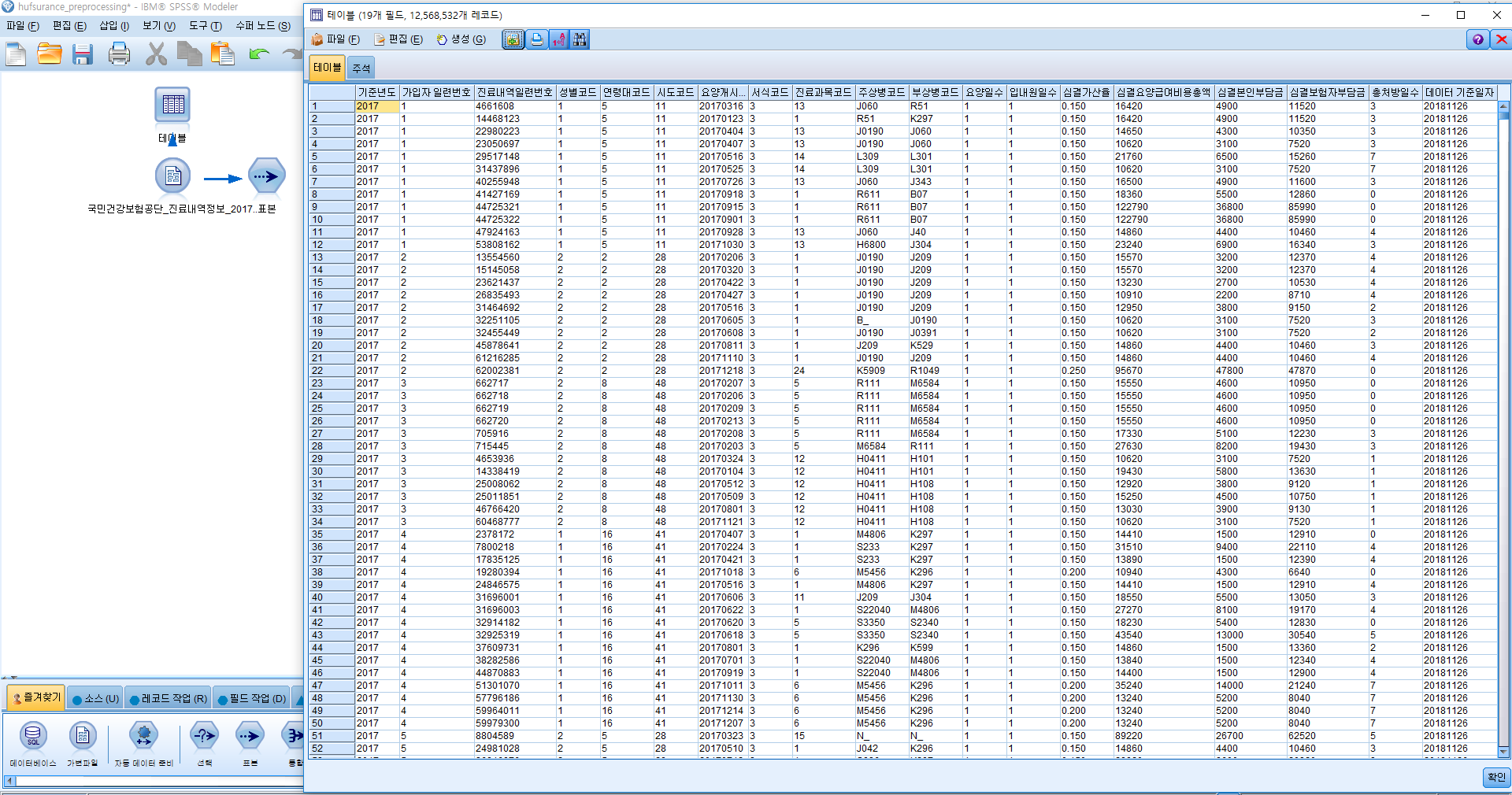
3. 의사결정나무 분석에 QUEST알고리즘을 추가 적용하였다.

4. 의사결정나무 분석의 목표변수를 ‘총액’ 기준과 ‘보험자 부담’기준으로 두 번 진행했다.

3-1. Data Preprocessing 전처리 (백만개 표본 추출)

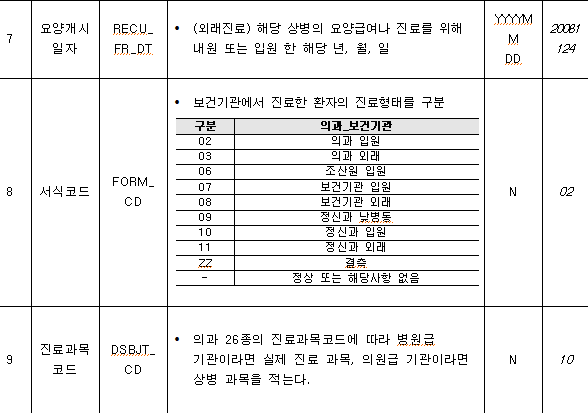
결과 : ‘aaa\_million.csv’

1) row data : 19개 필드, 12,568,532개 레코드

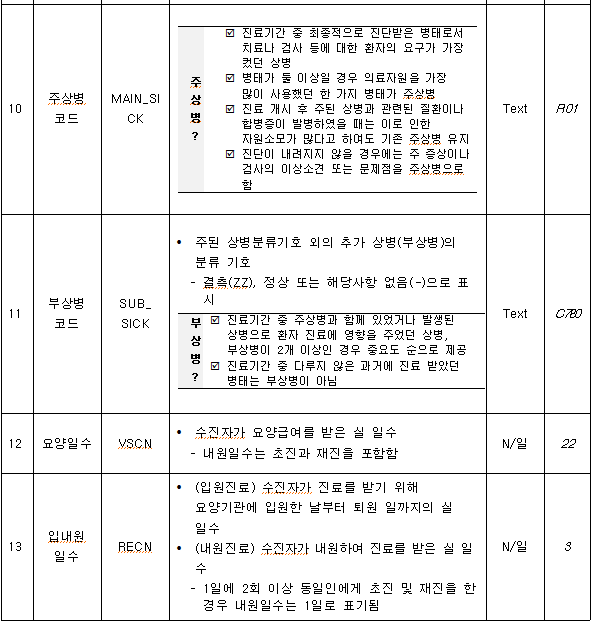


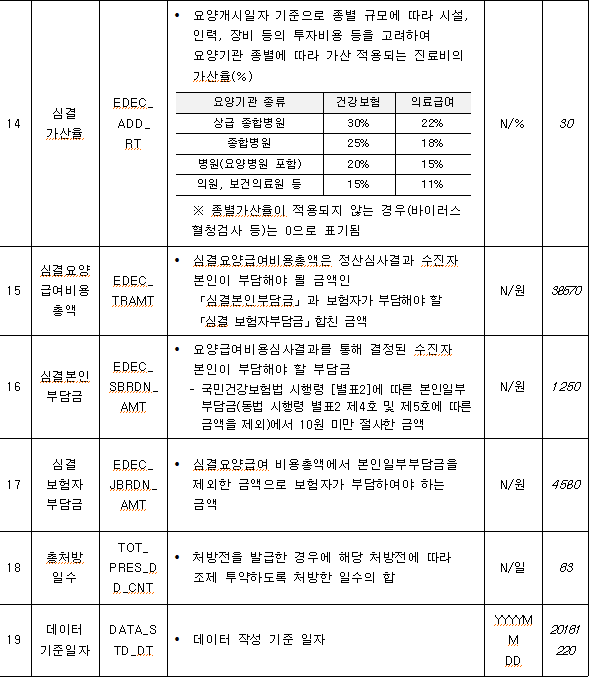
2)각 필드에 대한 주석



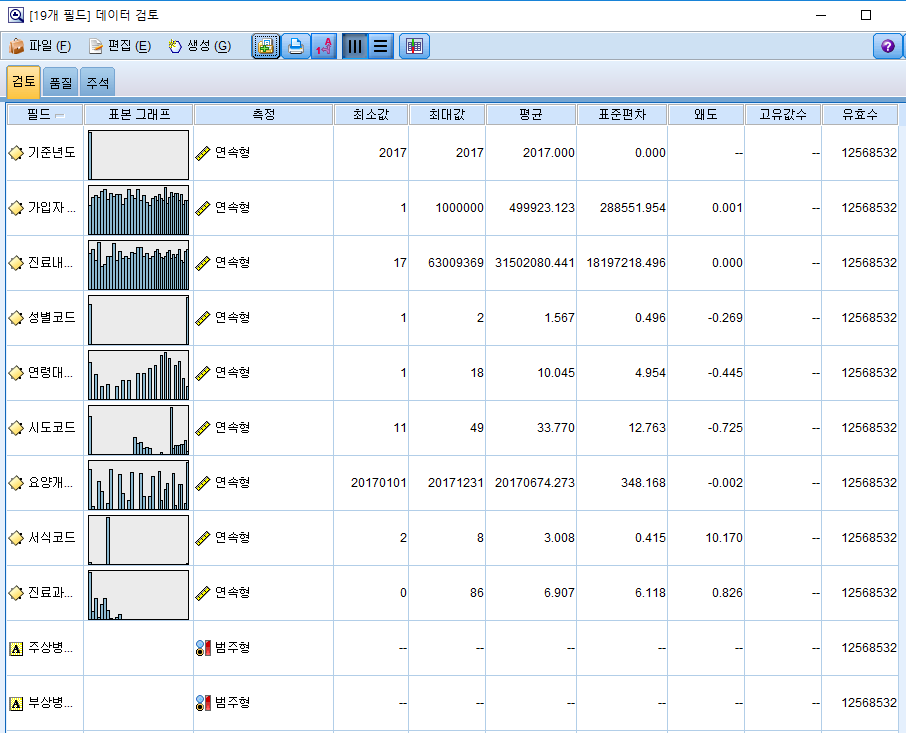








3) row data의 데이터 검토 : 모든 attribute의 유효수가 12,568,532이므로 missing value가 없다.





데이터 검토 결과를 통해 각 attribute를 점검하면 다음과 같다.

1. 첫번째와 마지막 attribute인 *기준년도*와 *데이터 기준일자*는 무의미하므로 삭제한다.

2. *가입자 일련번호* : 총 1,000,000명의 환자에 대한 데이터이다.

3. *연령대코드* : 특정 연령대에 대한 진료 빈도수에 유의미한 분포를 보인다. 대략적으로 0-9세의 영유아, 미취학 아동 및 초등 저학년의 진료빈도가 높고, 청장년층으로 갈수록 빈도가 감소하다가 중장년층 및 노년층을 향하며 빈도가 증가한다.

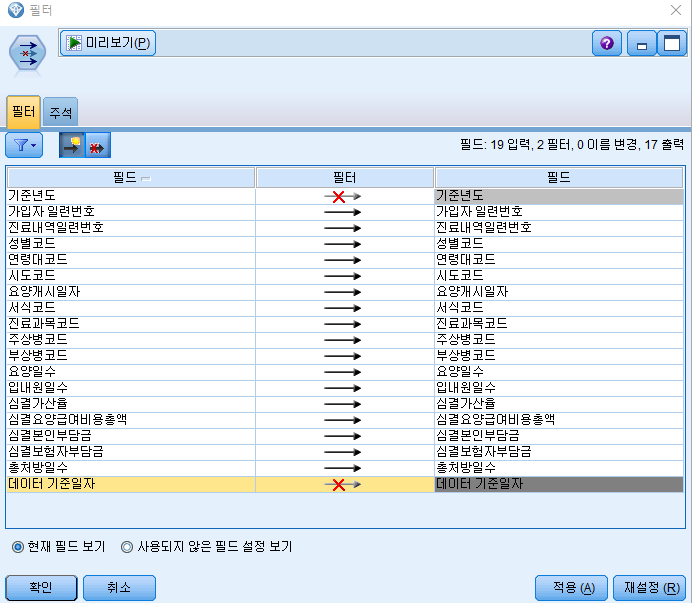
4. *요양개시일자* : 요양기관에 내원한 일자인 요양개시일자의 분포는 일정한 cycle을 형성한다.

5. *시도코드, 서식코드, 진료과목코드, 심결가산율* : 다음의 네 가지 attribute는 공통적으로 특정 항목의 빈도가 월등히 높다. 이후 범주형으로 분류된다.

6. *요양일수, 입내원일수, 심결요양급여비용총액, 심결본인부담금, 심결보험자부담금, 총처방일수*

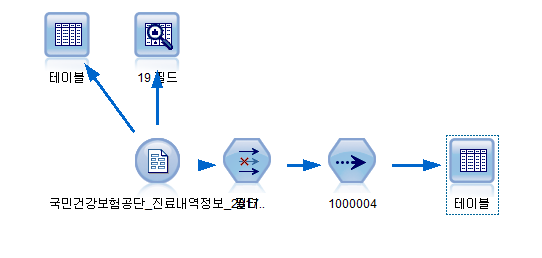
: 다음 항목들은 공통적으로 왜도가 높다. Outlier 판별과 데이터 분리를 전제로 한 분석이 요구된다.

3) 분석에 사용되는 attribute목록



4) 약 백 만개의 표본 추출(가입자 일련번호 79818번까지)

: row data의 용량이 커 원활한 프로그램 작동이 어렵기 때문에, 처음부터 백 만개의 표본을 추출하여 이후 분석을 진행한다. 형성된 테이블은 ‘aaa\_million.csv’로 저장한다.



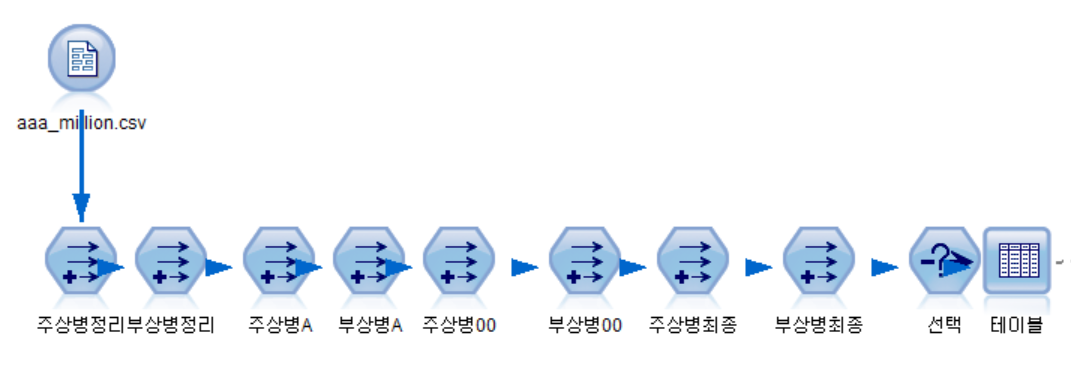
3-2. Data Preprocessing 전처리 (환자별 데이터의 통합)

결과 : ‘aaa\_million\_2.csv’, ‘aaa\_million\_byPatient.csv’ 사용 : 연관성 분석, 클러스터링

“주상병코드”와 “부상병코드”필드는 KOICD 질병분류 정보센터, ‘한국표준질병-사인분류 7차 개정안’의 코드를 반영하고 있다. 이는 상당히 세분화된 분류이기에 데이터 분석에 어려움을 야기했다. ([3. 연관성 분석]참조) 따라서 “주상병코드”와 “부상병코드”를 아래와 같은 21가지 항목으로 분류하여 각각 “주상병최종”, “주상병최종” 필드를 형성했다. 또한, 어떤 분류에도 포함되지 않는 1000여개의 케이스 (0.1%)는 선택노드를 통해 제거했다.

* A00-B99 I. 특정 감염성 및 기생충성 질환
* C00-D48 Ⅱ. 신생물
* D50-D89 Ⅲ. 혈액 및 조혈기관의 질환과 면역메커니즘을 침범한 특정 장애
* E00-E90 Ⅳ. 내분비, 영양 및 대사 질환
* F00-F99 Ⅴ. 정신 및 행동 장애
* G00-G99 VI. 신경계통의 질환
* H00-H59 Ⅶ. 눈 및 눈 부속기의 질환
* H60-H95 Ⅷ. 귀 및 유돌의 질환
* I00-I99 Ⅸ. 순환계통의 질환
* J00-J99 Ⅹ. 호흡계통의 질환
* K00-K93 ⅩⅠ. 소화계통의 질환
* L00-L99 ⅩⅡ. 피부 및 피하조직의 질환
* M00-M99 ⅩⅢ. 근골격계통 및 결합조직의 질환
* N00-N99 ⅩⅣ. 비뇨생식계통의 질환
* O00-O99 ⅩⅤ. 임신, 출산 및 산후기
* P00-P96 ⅩⅥ. 출생전후기에 기원한 특정 병태
* Q00-Q99 ⅩⅦ. 선천기형, 변형 및 염색체이상
* R00-R99 ⅩⅧ. 달리 분류되지 않은 증상, 징후와 임상 및 검사의 이상소견
* S00-T98 ⅩⅨ. 손상, 중독 및 외인에 의한 특정 기타 결과
* V01-Y98 ⅩⅩ. 질병이환 및 사망의 외인
* Z00-Z99 ⅩⅩⅠ. 건강상태 및 보건서비스 접촉에 영향을 주는 요인
* U00-U99 ⅩⅩⅡ. 특수목적 코드

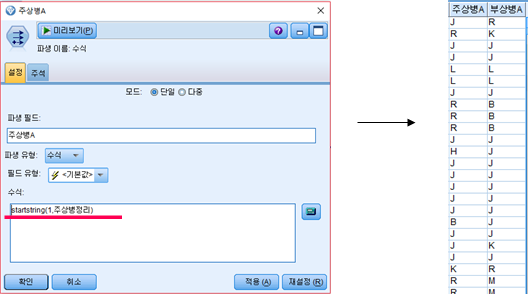
“주상병최종”, “주상병최종” 필드를 파생하는 과정은 다음과 같다.



1). 3자리 정리 : 파생 노드를 통하여 각각의 코드를 3자리로 정리하였다

.

2) 알파벳으로 정리 : 파생 노드를 통하여 첫 글자로 정리하였다.

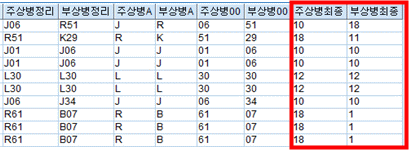


3) 숫자 코드 정리 :　파생 노드를 통하여 1단계에서 숫자로 정리하였다.

4) 최종 코드 정리 : 한국질병 분류 센터의 정보에 따라 코드 정리를 위해 *주(부)상병A* 와 *주(부)상병00 필드*를 이용한 수식을 작성하였다.

스크린샷이(가) 표시된 사진

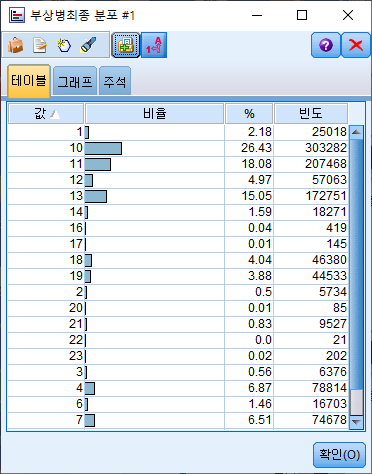
자동 생성된 설명

5) 얻게 된 결과는 다음과 같다. 

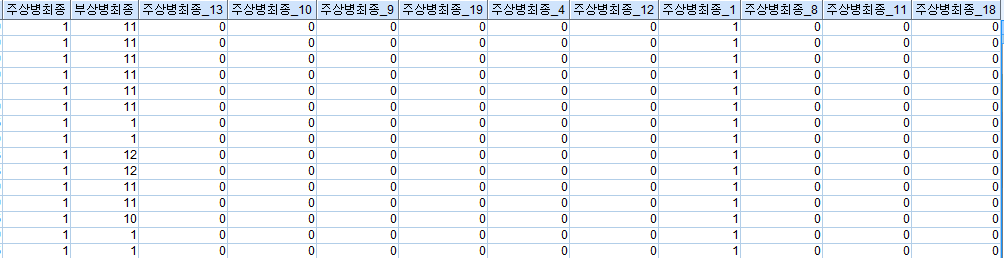
단, 코드 오류로 22가지 분류에 포함되지 않은 약 1600개의 row는(‘주상병최종’과 ‘부상병최종’이 0)는 선택노드를 통해 삭제한다. 결과는 ‘aaa\_million\_2.csv’에 저장한다.

이후 환자별 데이터가 필요한 연관성 분석 및 클러스터링을 위한 환자별 데이터 통합을 진행했다. 우선, 주상병, 부상병의 경우는 상위 10개의 케이스를 선별하여 이분변환했다. 구체적인 과정은 다음과 같다.

1) 이분형 변수 생성 : 이분형 변수를 생성한다. 총 21개의 변수를 생성하면 너무 많은 값이 생성되므로 *주·부상병최종 필드에서 상위 10개만* 선별한다.



2) 이분변환 결과는 다음과 같다.





그리고는 아래와 같이 통합 노드를 사용하여 “가입자 일련번호” 즉, 환자 한 명을 기준으로 “요양일수”, “입내원일수”, “심결요양급여비용총액”, “심결본인부담금”, “심결보험자부담금”, “총처방일수” 및 주요 주-부상병 최종의 이분형 값의 합을 구했다. 통합 결과로 앞선 필드명에 ‘\_sum’ 접미어를 결합한 필드들이 생성되었다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

그 다음 가입자 일련번호를 기준으로 ‘aaa\_million\_2.csv’파일과 전체 외부 결합한 뒤, 각 환자별 첫 번째 레코드만 남기고 삭제했다.

스크린샷이(가) 표시된 사진

자동 생성된 설명스크린샷이(가) 표시된 사진

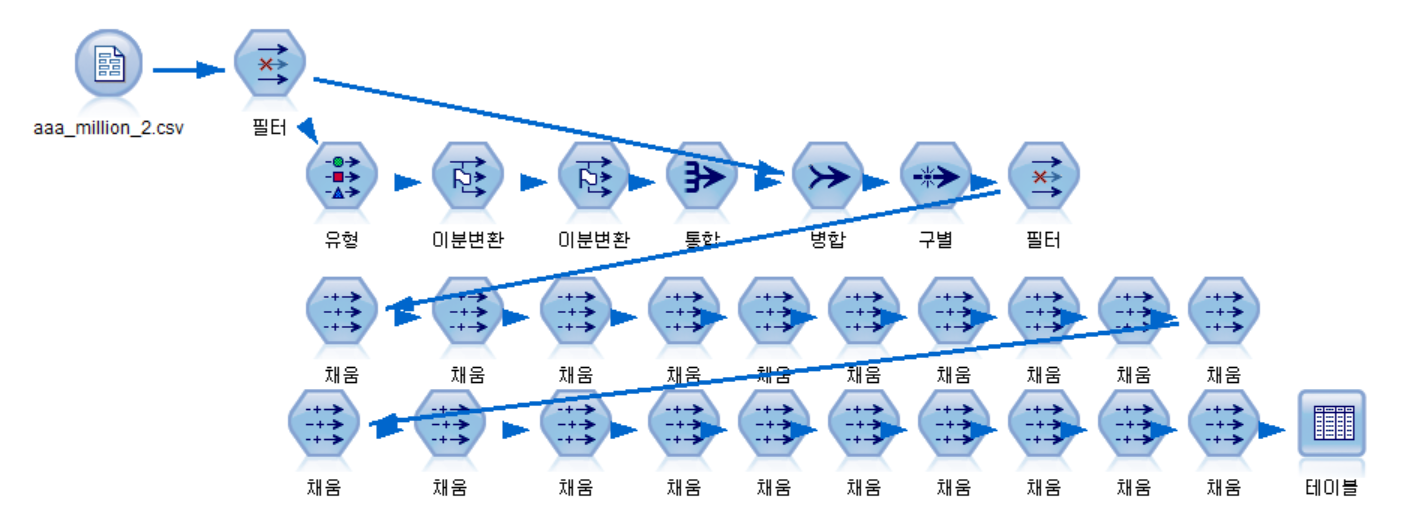
자동 생성된 설명

한편, 주-부상병 최종의 합은 이분형으로의 재조정이 필요하므로, 아래와 같이 채움노드를 이용한다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

결과 스트림은 다음과 같으며, 파일은 ‘aaa\_million\_byPatient.csv’로 저장한다.



3-3. Data Preprocessing 전처리 (연속된 진료내역 통합)

결과 : ‘aaa\_million\_illness.csv’, 사용 : 의사결정 나무, 신경망분석

기존의 데이터는 특정 상병에 대한 진료가 서로 인접한 날 반복되었을 경우 각각의 진료내역별로 row를 형성했다. 이는 특정 상병에 연관된 금액 및 내원 일수를 진료내역 건 별로 분산하여 정확한 파악에 오류를 줄 수 있다고 판단했다. 따라서 아래와 같이 “가입자 일련번호”, “주상병코드”, “부상병코드”, “서식코드”, “심결가산율”, “진료과목코드”를 키 필드로, “요양일수”, “입내원일수”, “심결요양급여비용총액”, “심결본인부담금”, “심결보험자부담금”, “총처방일수”를 통합 필드로하여 특정 환자의 서로 인접한 진료내역을 통합했다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

통합 결과로 “요양일수”, “입내원일수”, “심결요양급여비용총액”, “심결본인부담금”, “심결보험자부담금”, “총처방일수”에 \_sum 접미어를 결합한 필드들이 생성되었고, 이는 기존 ‘aaa\_million\_2.csv’파일과 아래와 같이 ‘전체 외부 결합’하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

그리고는 아래와 같이 필터를 적용하였다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

이후, 앞서 다룬 키 필드 기준으로 중복된 값들을 구별 노드를 통해 제거하고, ‘aaa\_million\_illness.csv’파일이 형성되었다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

스크린샷, 컴퓨터이(가) 표시된 사진

자동 생성된 설명

(18개 필드, 484,730개 레코드)

하늘이(가) 표시된 사진

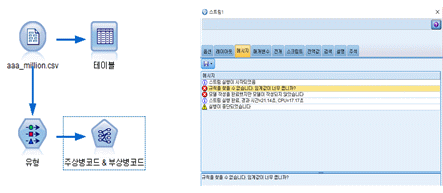
자동 생성된 설명

4. 연관성 분석

1) 목표 : 주상병 코드와 부상병 코드를 각각 변수로 설정하고, 주상병과 부상병 사이의 연관성을 파악하고자 한다.



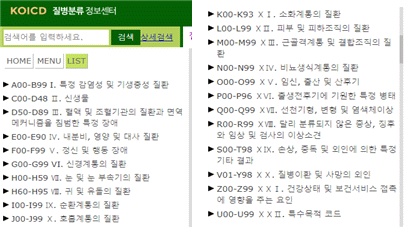
2) 연관성 분석에는 Apriori를 사용하였고, 주상병코드와 부상병코드를 변수로 지정하였다. 각각 지지도, 신뢰도, 전항값 수를 10, 75, 5를 기본으로 설정하였다. 이후 값을 조정하며 연관성을 분석하였다.

연관성 분석한 결과는 다음과 같다.

1. 결과를 얻기 위해 지지도, 신뢰도를 하향했다. 유의미한 결과를 얻을 수 없었다.

2. 각 코드는 알파벳 하나와 숫자 3~4개의 조합으로 이루어져 있다. 크게 분류하자면 *알파벳은 큰 범주의 질병 코드*이고, *구체적인 질병은 각각 숫자가 다르게* 부여된다.

3. A~Z, 000~999 까지의 조합으로 이루어진 코드는 *구체적이고 다양하게 부여*되었다. 유의미한 결과를 얻기에는 코드 정리가 필요했다.

4. 위의 질병 코드를 한국질병분류정보센터의 정보에 따라 정리했다.

3) 한국 질병 분류 정보센터에는 21개, 다음과 같이 질병 코드를 간소화했다.

4) 질병 코드를 간소화한 데이터 파일, ‘aaa\_million\_2.csv’를 바탕으로 만들어진 환자별 데이터, ‘aaa\_million\_byPatient.csv’파일을 바탕으로 연관성 분석을 진행했다.

지도이(가) 표시된 사진

자동 생성된 설명



연관성 분석의 결과는 다음과 같다.

1. 필드 : 전항을 주상병 상위 10개 필드로, 후항을 부상병 10개 필드로 지정한다,
2. 분석 실시 : 기본값으로 Apriori 분석을 실시한다.
3. 결과 : 비교적 높은 신뢰도로 결과가 생성되었다.

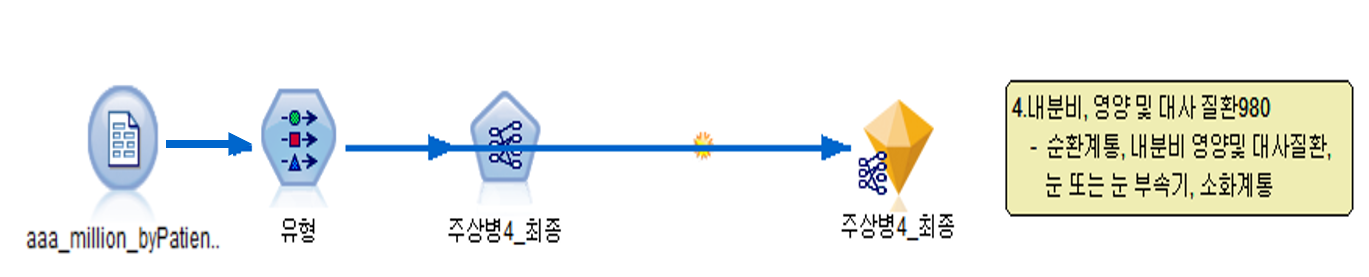
하지만 대부분의 후항 또는 전항값은 10코드와 관련이 있었다. 이는 호흡기와 관련 된 질환으로 속한 구체적 질병으로는 바이러스성 감기, 인플루엔자성 호흡기 질환이다. 1년간의 기록에서 이 데이터가 의미하는 바는, *특별한 질병과의 연관성 보다는 누구나 쉽게 걸릴 수 있는 질병인 호흡기 질환, 즉 질병 코드 10번의 빈도수가 많다는 것을 주로 파악할 수 있다.* 따라서 특별한 질병과의 연관성을 얻기 위해서는 다른 방법이 필요했다.

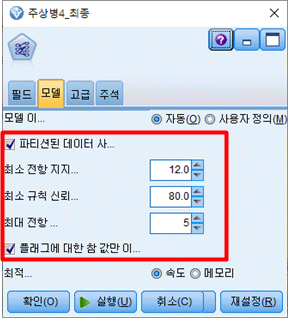
5) 빈도수가 아닌 특별한 질병과의 연관성을 파악하기 위해 전항값에 단독 코드를 입력한다. 후항값에는 부상병코드 필드 전부를 적용한다. 유의미한 값을 얻기 위해, 호흡기 질환(코드10)을 제외, 빈도수 상위 3가지 질환 코드를 연관성 분석한다.

6) 분석 과정과 결과는 다음과 같다.

1. ***코드4.내분비, 영양 및 대사 질환***

- 유의미한 결과, 즉 특정 소수의 결과를 얻기 위해 지지도와 신뢰도를 반복적으로 상향 조정하며 특정 값을 얻는다.







a. *주상병4*는 *내분비, 영양 및 대사 질환*이다. 여기에는 당뇨병, 갑상선 장애, 비만 및 기타 과영양 등이 포함되어 있다.

b. *부상병9*는 *순환계통*의 질환이며, *부상병11*은 *소화계통*의 질환이다.

c. *부상병4*는 같은 코드의 질병이므로 연관 질병이 당연하므로 제한다.

**d. 지지도 13% , 신뢰도 85% 의 결과, 401 인스턴스**

e. 결과 : 지지도 13%와 신뢰도 85%까지 상향조정한 결과, 위와 같은 결과를 얻었다. 총 401명, 전체의 13%에 해당한다. 또한 이 규칙의 신뢰도가 85%인 점으로 상당히 믿을만한 결과라는 것을 알 수 있다.

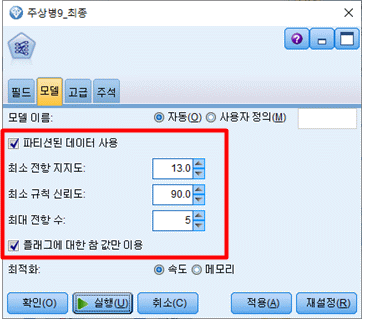
주상병4 에 포함되어 있는 당뇨병, 비만 및 기타 과영양과 대사 장애의 질병을 가지고 있는 사람들이 있다. 이들 중 85%의 비율로 부상병11(소화계통)과 부상병9(순환계통)을 함께 앓고 있는 것으로 나타났다. 부상병9(순환계통)에 포함된 질병으로는 고혈압 또는 저혈압성 질환, 뇌혈관 질환, 동맥 또는 세동맥 질환 등이 포함 되어 있다. 또한 부상병11(소화계통)의 질병에는 구강, 식도 및 십이지장, 복막 또는 간으로 주로 섭취와 관련된 기관의 문제로 보인다. 영양, 즉 섭취와 관련된 질병인 주상병4와 연관성을 가지고 있는 부상병9는 섭취 패턴에 따라 혈관 또는 혈압 문제를 야기할 수 있다는 것은 대부분 알고 있는 사실일지 모른다. 하지만 부상병11의 경우에서, 섭취 패턴에 따라 식도나 십이지장 또는 복막에 질병을 앓고 있다는 사실을 유추하기 힘들 것이다.

***2. 코드 9.순환계통의 질환***

- 유의미한 결과, 즉 특정 소수의 결과를 얻기 위해 지지도와 신뢰도를 반복적으로 상향 조정하며 특정 값을 얻는다.

개체이(가) 표시된 사진

자동 생성된 설명





주상병9를 분석한 결과는 다음과 같다.

a. *주상병9*는 순환계통의 질환이다. 여기에는 고혈압 또는 저혈압성 질환, 뇌혈관 질환, 동맥 또는 세동맥 질환 등이 포함 되어 있다.

b. *부상병9*는 같은 계열의 질병으로 제외한다.

c. *부상병4*는 내분비, 영양 및 대사 질환이다. *부상병 7*은 눈 및 눈 부속 기관의 질환이다. *부상병 11*은 소화계통의 질환이다.

d. 부상병4는 주상병4의 내용에서 언급한 바와 같다.

**e. 지지도 13%, 신뢰도 90%, 인스턴스 400이상의 결과**

f. 결과 : 지지도 14%, 신뢰도 90%까지 상향 조정한 결과, 위와 같은 결과를 얻었다. 총 431명과 445명으로 약 14%에 해당하며, 신뢰도가 무려 90% 이상인 믿을만한 결과이다.

주상병9(순환계통)에 포함되어 있는 질병으로는 고혈압, 저혈압성 질환, 뇌혈관성 질환, 동맥 또는 세동맥 질환등이 있다. 또한 식도 정맥류나 림프절의 장애를 포함한다. 이들 중 90%의 비율로 부상병4(내분비, 영양 및 대사 질환), 부상병10(호흡계 질환), 부상병11(소화 계통)의 질환을 앓고 있는 것으로 나타났다. 부상병 4의 내용은 위에서 언급한 바와 같다. 부상병11(소화계통)에 포함되는 질병으로는 식도의 질환을 포함하고. 구강, 복막 및 간 등의 질환, 장의 질환 등이 나타난다. 부상병11의 식도 질환에서는 정맥의 영역은 포함하지 않는다. 식도염이나 식도 역류병 등을 포함한다. 공통적인 기관‘식도’를 제외하고는 그다지 공통점을 찾아 볼 수 없지만 연관성을 가지고 함께 발병했다. 부상병7(눈 및 눈 부속 기관)에 포함되는 질환은 말 그대로 눈꺼풀, 망막, 녹내장, 백내장 등의 질병을 포함한다. 순환계통의 질환 보유자의 90% 이상이 질병코드7, 즉 눈과 관련한 질병으로 병원을 내원했다는 결과를 발견했다.

7) 결론

주상병4(내분비,영양 및 대사 질환)과 부상병9(순환계통), 부상병11(소화계통)의 질환. 그리고 주상병 9(순환계통)과 부상병4(내분비,영양 및 대사 질환), 부상병7(눈 및 눈 부속기 질환), 부상병11(소화계통)의 질환은 높은 연관성을 나타내었다. 항목들 사이에는 간혹 발병하는 기관이 같은 질병도 포함하고 있었다. 하지만 대부분의 항목들에서는 큰 공통점을 찾아볼 수 없었다. 우리도 서로의 개연성을 생각해 낼 수 있는 질병들도 있었지만, 의외의 필드들도 많았다. 예를 들어, 순환계통 환자들의 90% 이상이 눈의 문제로 다시 병원을 찾는다던가, 내분비, 대사에 질환을 가지고 있는 환자들의 80% 이상이 식도나 간, 위의 문제로 다시 병원을 찾게 될 줄은 예측하지 못했을 것이다. 우리는 이를 합병증이라 부르곤 한다. 연관성 분석을 통해 전혀 예측하지 못한 합병증을 수치로 파악한 것이다.

국민 건강 보험 공단에서는 현재 질병이나 부상으로 인해 발생한 고액의 진료비로 가계에 과도한 부담이 되는 것을 방지하기 위하여, 국민들이 평소에 보험료를 내고 보험자인 국민건강보험공단이 이를 관리·운영하다가 필요시 보험급여를 제공하는 방법으로 운영한다. 보다 효과적이고 효율적인 재단 운영을 위해서 공단은 위와 같은 방법으로 이전의 데이터를 사용할 수 있다.

공단에서는 연관성 분석으로 얻게 된 데이터를 국민에게 제공한다. 특정 질병으로 보험료를 지급 받게 된 보험자에게 아직은 발병하지 않은 연관 질병에 대해서 미리 안내하고, 검진을 유도한다. 이를 통해 보험자는 건강 악화와 가계 부담을 줄일 수 있고, 공단에서는 보험 지급금을 줄일 수 있다. 상호 이익을 얻을 수 있는 방법이다.

5. 군집 분석(clustering)

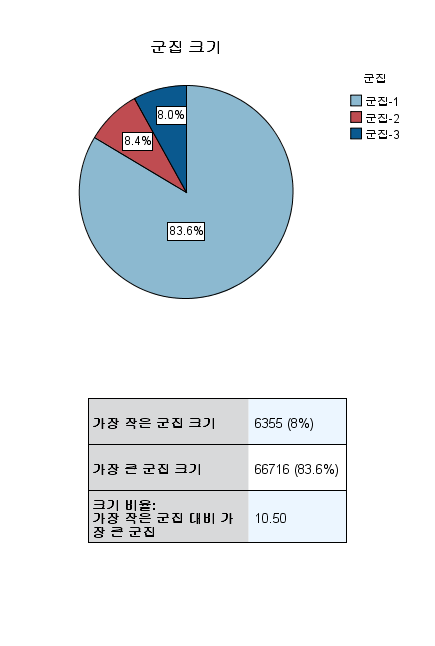
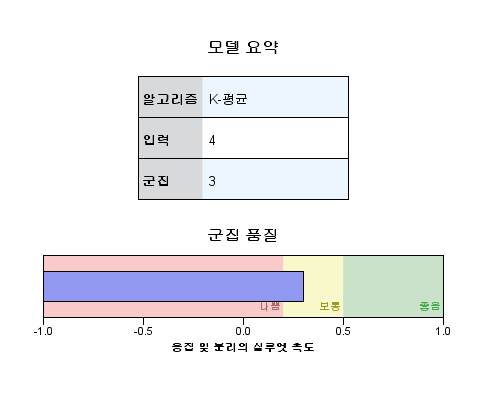
1) 목표 : 모든 변수들을 요인화하여 군집화시키는 것이 아닌 앞의 데이터 분석을 통한 목표 변수들(Targeted Variables)을 설정 하였다. 특히 어떤 목표 변수들이 목표변수인 심결본인부담금총액에 영향을 끼치는지 알기 위해 간단한 도메인 지식을 활용하였고 의외로 성별, 시도군 같은 목표 변수들은 결과치에 크게 영향을 끼치지 않을거라 생각하였다. 따라서 결정한 최종 목표변수들은 연속형 변수인 입내원일수, 총처방일수, 심결요양급여비용총액 마지막으로 명목형 목표변수인 연령대 코드를 통해 알아보고자 했다.

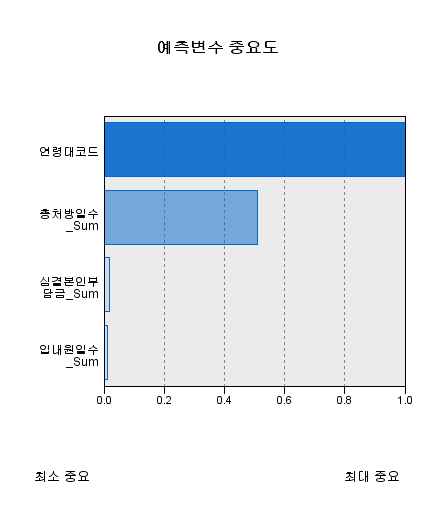
2) 방법 : K-means, Twostep, kohonen 3가지 클러스터링 방법의 비교를 통해 가장 적합한 방법을 선택하고 이를 적용할 것이다.

**3) K-means method**

K-means 방법은 직접 군집수를 적용할 수 있다. 따라서 3~5가지의 군집수를 적용해보았고 이를 통해 각각의 군집수에 따른 특징을 살펴 보았다.

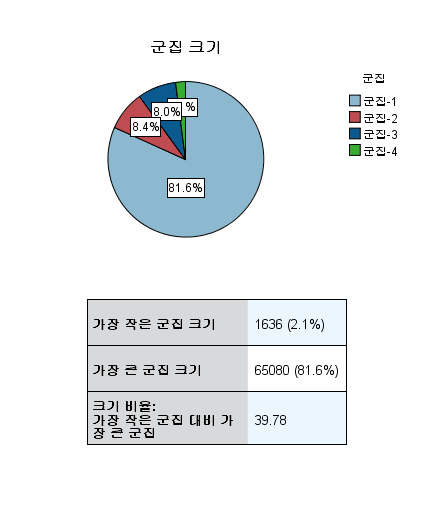
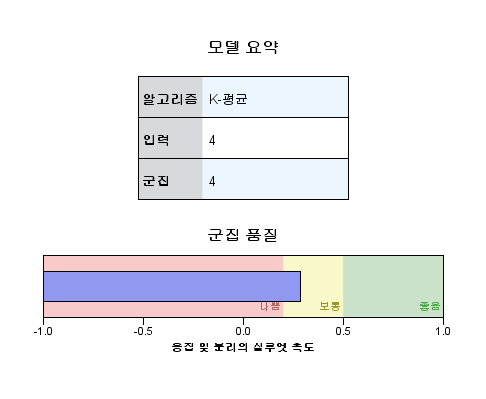
K-means -method (군집 수 : 3개)



군집 품질이 보통으로 나타났다. 또한 특이하게 연령대의 특정 변수들이 몰려서 분포되었는데 군집 2, 군집3이 각각 12, 11로 나타났다. 이에 대한 원인은 목표 변수 수치들이 4개 밖에 되지 않아 발견하지 못하였다. 

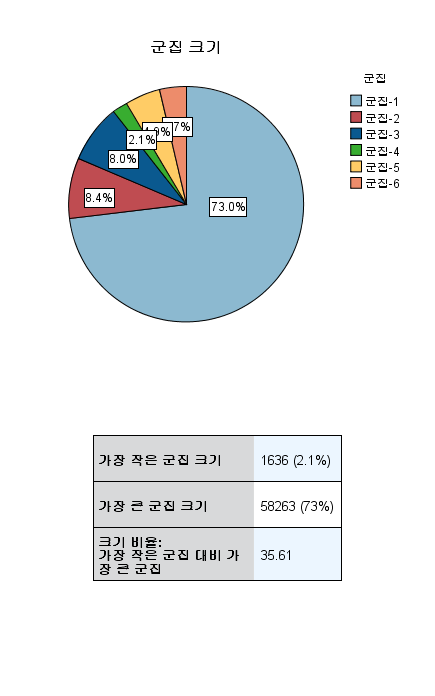
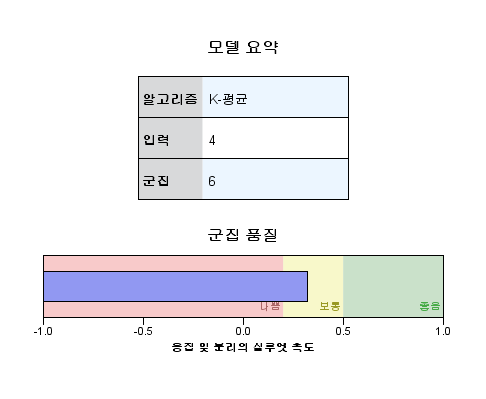
하지만 예측 변수 중요도를 보았을 때 연령대코드와 총 처방일수가 높은 수치로 나타난것으로 보아 특정 연령대 11,12 가 총처방일수가 높이 나타났음은 알 수 있었다.

k-means -method (군집 수 : 4개)



군집 품질수가 4개일 경우 역시 품질은 보통으로 나타났다. 이번 군집 역시 군집 2,3,4 가 각각 12, 11 ,17 연령대를 포함하게 군집화 되어있다. 이 역시 위의 결과와 마찬가지로 각각 총처방일수와 연관이 되어 있음을 유추할 수 있었다.

k-means -method (군집 수 : 6개)

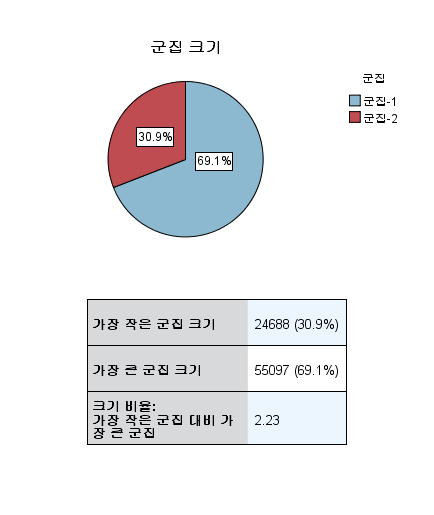
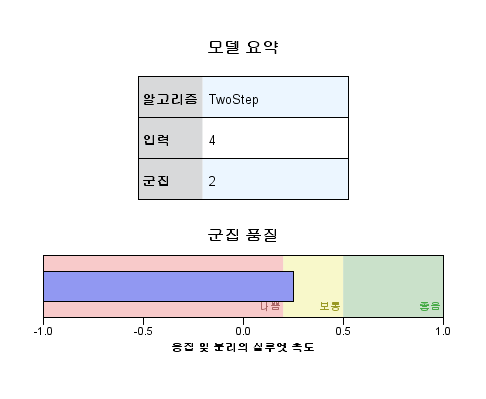


6개의 군집 품질 역시 보통으로 나왔다. 또한 결과 해석 역시 연령대를 기준으로 품질 이 크게 분리되어 있는데 각각 12,15,17,18로 군집화되었다. 이는 역시 총처방일수와의 연관성을 나타 내었다 볼 수 있다.

**4) Two-step method**

- 개체 수가 큰 자료의 군집화에 매우 효율적인 Two-step method를 사용해 보았다.

- 거리 측도 : 로그 우도, 군집 기준 : 슈바르츠 Bayesian(BIC)을 기준으로 하였다.



그 결과 적정한 군집수는 2개고 군집 품질은 보통으로 선정되었다. 거리 측도를 유클리디안으로 설정하고 분석한 결과 역시 큰 차이는 발견되지 않고 군집 품질 : 보통, 군집크기는 2이 설정되어 분석 되었다. 마찬가지로 군집 기준을 아케이케 정보 기준(AIC)으로 각각 다르게 설정하여도 항상 같은 품질, 같은 군집수의 결과값이 도출되었다.

군집은 각각 연령대, 총요양급여비용총액을 기준으로 군집화가 되었는데 나이가 들어감에 따라 비용이 증가하는 것은 엄연히 당연한 결과인데 이를 수치와 데이터군집으로 확인 할 수 있었다.

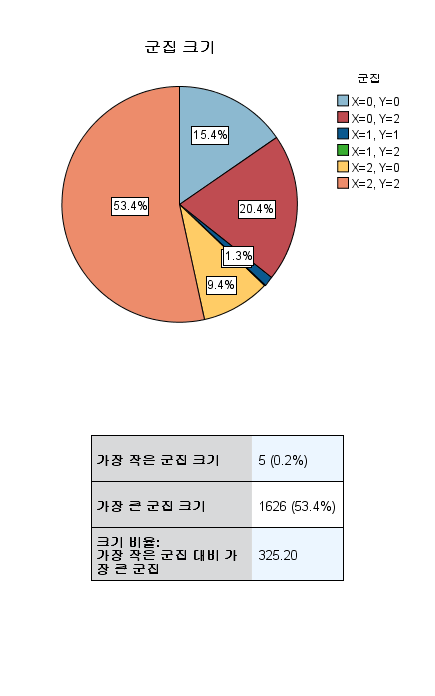
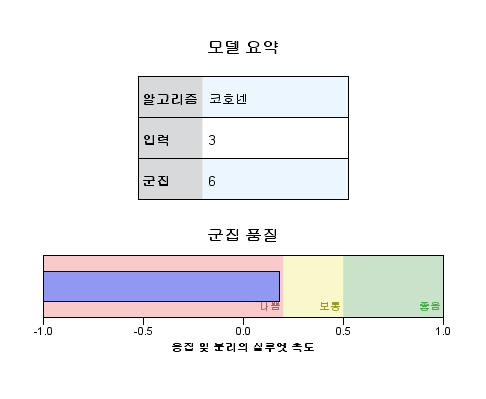
**5) Cohonen-som method**

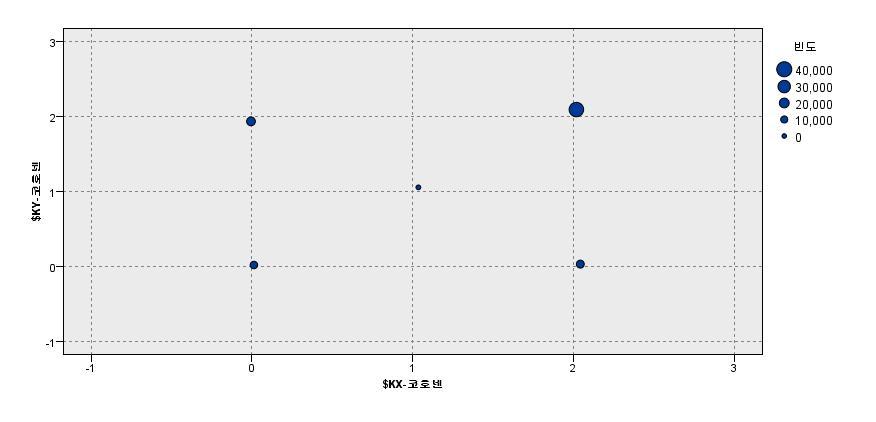
타 알고리즘에 비해 연산시간이 많이 소요되는 코호넨 알고리즘을 사용해 보았다. 군집 수 결정이 어렵지만 시각화(visualization)가 가능하고 군집간 관계를 파악하는게 쉽다.

크기 : x=3, y=3 학습결과

아래는 도표를 이용하여 결과치를 시각화한 그래프이다.

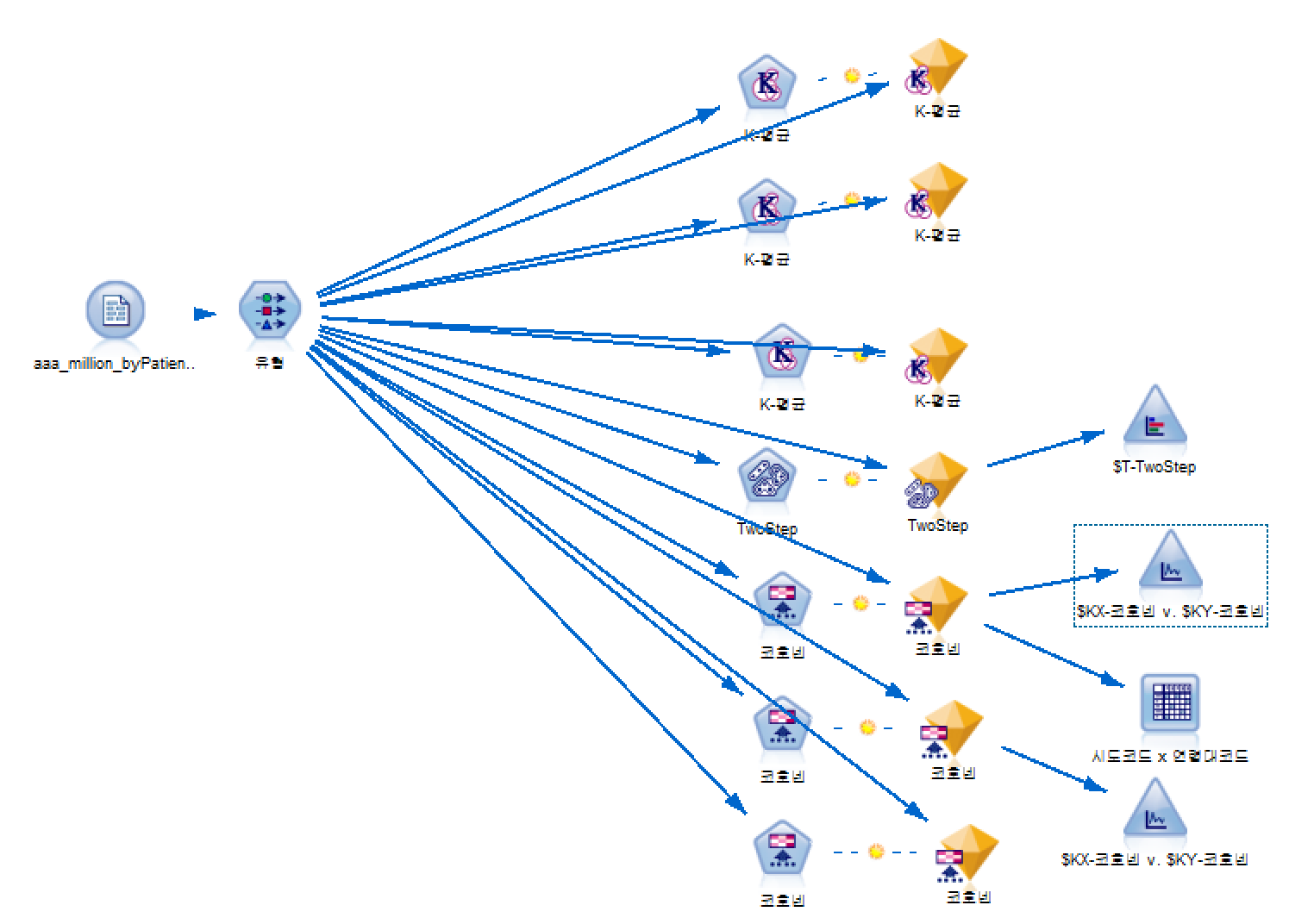
비슷한 방법으로 4 by 3 으로 학습한 그래프이다.





코호넨 군집분석 결과 2,2 의 군집의 크기가 제일 큼을 확인할 수 있었다.

클러스팅 스트림 구조



6) 개선할 점

클러스티링을 진행하면서 제일 중요한 것은 목표 변수라는 생각을 많이 하였다. 그러나 이 목표 변수는 분석가가 스스로 선택해야 하므로 분석에 관련된 도메인 지식이 제일 중요하다. 그러나 이번분석은 생소한 보험이라는 주제로 분석을 진행하다 보니 보험액 산출등에서 어떠한 요소가 결정적 요소인지를 간과하였다. 만일 다음 이에 관련된 연구를 진행한다면 직접 계리사님 등을 찾아 뵈어 미리 목표 변수를 설정하여야 할 것 같다.

6. 의사결정나무 분석

1) 목표

들어가기에 앞서 분석에 사용한 데이터는 ‘aaa\_milion\_illness.csv’로, 특정 상병에 연관된 금액 및 내원 일수가 진료내역별로 분산되는 오류를 막고자 특정 환자의 서로 인접한 진료내역을 통합한 데이터이다.

심결요양급여비용총액을 ‘저가형-고가형’으로 범주화하여 목표변수(Targeted Variable)로 설정하고, “서식코드”, “입내원일수\_Sum”, “진료과목코드”, “주상병최종”, “부상병최종”, “연령대코드”을 의사결정 규칙 형성에 대한 변수로 한다. 결과를 통해 저가형 및 고가형 의료행위에 영향을 주는 요인들을 분석한다.

이후 심결보험자부담금을 목표변수(Targeted Variable)로 설정하여 동일한 과정을 반복하고, 앞선 결과와 비교하여 국민건강보험공단의 보험료 산정의 타당성을 평가한다. 이는 진료내역의 raw 데이터를 기준으로 분석한 결과이기에 보험공단에서 제공하지 않는 구체적 산정기준을 파악할 수 있음에 의미가 있고, 나아가 보험의 소비자인 ‘국민’의 관점에서 주체적으로 보험료산출의 타당성을 검토할 수 있음에 의의가 있다.

2) 심결요양급여비용총액(심결본인부담금, 심결보험자부담금)의 범주화

스크린샷이(가) 표시된 사진

자동 생성된 설명

백 만개의 표본에 대한 데이터 검토의 결과이며, 데이터 분포가 과도하게 치우친 양상을 보인다. 따라서 이상 분석을 통해 노이즈를 제거하고 생성된 그룹을 검토한다. 비용과 관련된 변수인 심결요양급여비용총액, 심결본인부담금, 심결보험자부담금에 대한 이상 분석을 진행한 결과는 다음과 같다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

모델링 결과 2개의 피어 그룹이 생성되었다. 또한, 새로이 생성된 $O-Anomaly는 T(true) 혹은 F(false)로 True인 row는 노이즈이다. 이 결과를 바탕으로 $O-Anomaly가 true인 4,847개(1%)의 노이즈를 제거한다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

각각의 피어 그룹에 대한 데이터 검토를 살펴보자.

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그룹1의 데이터 검토 결과>

스크린샷이(가) 표시된 사진

자동 생성된 설명

<그룹2의 데이터 검토 결과>

다음 결과로 보아 대체적으로 그룹1은 ‘저가형’ 진료를, 그룹2는 ‘고가형’ 진료를 나타낸다. 그룹1은 데이터의 다수이며, 그룹1에서 모든 변수의 최댓값이 대략 10만원 이하이므로 ‘저가형’진료는 최소 10만원 이하이다. 따라서 본 분석에서 ‘저가형’과 ‘고가형’의 구분을 100,000(10만원)으로 한다.

3) 의사결정나무 분석 – 목표값 : “심결요양급여비용총액\_Sum”

1. 75%의 training set과 25%의 test set으로 진행한다.

스크린샷이(가) 표시된 사진

자동 생성된 설명스크린샷이(가) 표시된 사진

자동 생성된 설명

2. 범주형 목표필드인 "저가형고가형"을 생성한다. ‘심결요양급여비용총액\_Sum’을 기준으로 10만원 이하는 1, 이상은 2를 부여한다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

3. 목표필드는 “저가형고가형”, 입력필드는 "주상병최종", "부상병최종", "진료과목코드", "입내원일수\_Sum", "서식코드", “연령코드”으로 하여 분석을 실행한다.

4. 분석결과

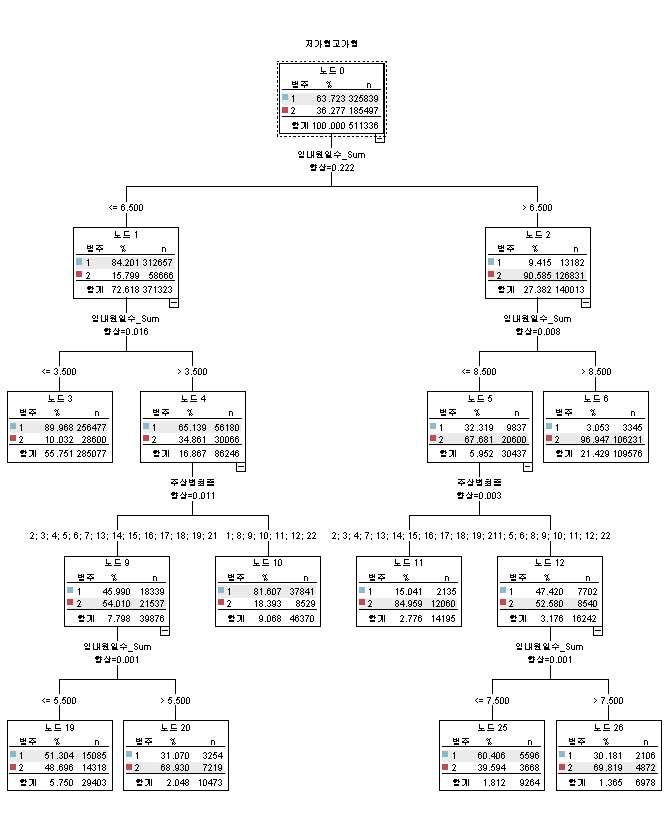
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1. C5.0 | 2. CART | 3. CHAID | 4. QUEST |
| Accuracy | 95.51% | 87.11% | 86.55% | 86.2% |
| 트리깊이 | 23 | 4 | 4 | 5 |
| 예측변수  중요도 | 입내원일수 | 입내원일수 | 입내원일수 | 입내원일수 |
| 서식코드 | 주상병최종 | 주상병최종 | 연령대코드 |
| 주상병최종 | 진료과목코드 | 부상병최종 | 부상병최종 |
| 부상병최종 | 연령대코드 | 연령대코드 | 서식코드 |
| 진료과목코드 | 서식코드 |  | 주상병최종 |
| 연령대코드 | 부상병코드 |  | 진료과목코드 |

형성된 모델의 Accuracy는 C5.0모델이 95.51%로 가장 높으나 트리깊이가 23일 정도로 상당히 복잡하기 때문에, 나머지 알고리즘을 중심으로 분석을 진행한다.

4-1. 분석결과(CART)

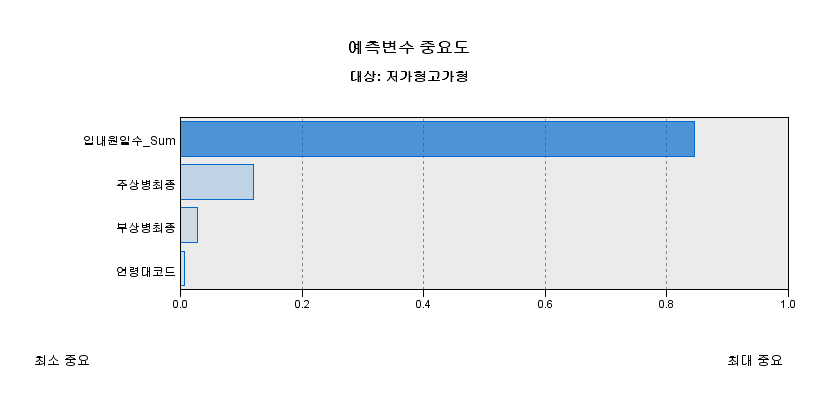
스크린샷이(가) 표시된 사진

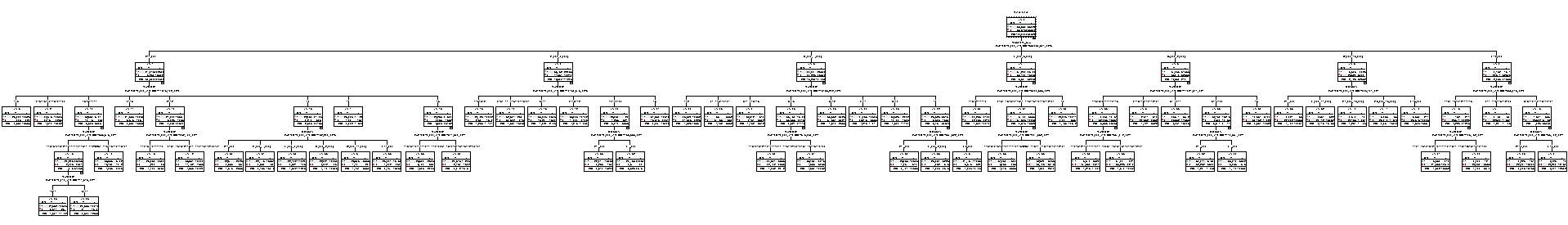
자동 생성된 설명



다음 결과에서 도출할 수 있는 유의미한 결과는 입내원 일수에 대한 정보이다. 저가형과 고가형 진료는 입내원 일수 6.5일을 기준으로 크게 나뉜다. 따라서 6.5일 이상의 입원에 대해 장기입원 지원을 추가적으로 고려할 필요가 있다.

4-2. 분석결과(CHAID)





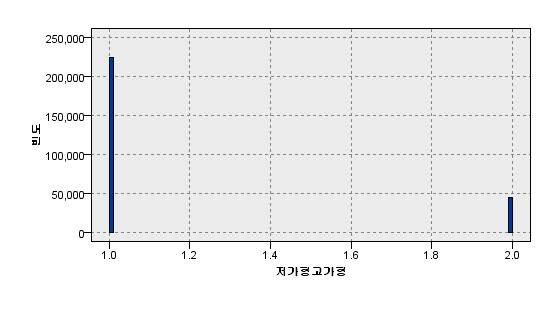
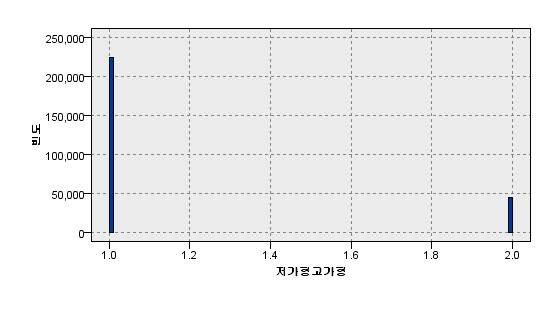
CHAID 알고리즘은 앞선 CART 알고리즘을 통한 입내원일수 분석을 보다 세밀하게 다룰 수 있다.

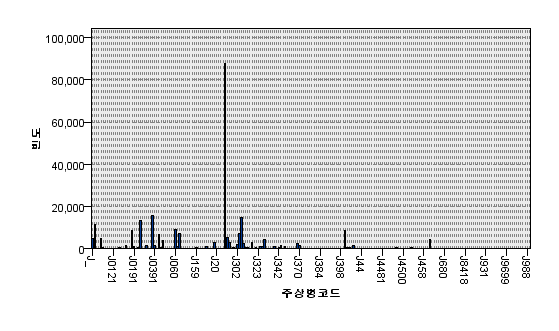
스크린샷이(가) 표시된 사진

자동 생성된 설명스크린샷이(가) 표시된 사진

자동 생성된 설명

우선 저가형 진료는 입내원일수가 1일 이내인 당일진료가 제일 많다. 당일-저가형 진료의 대표적인 질병은 주상병10인데 주상병10에 대한 추가적인 분석을 진행했다.

우선 주상병10은 저가형 진료 비율이 압도적이다. 



주상병10(호흡계통의 질환)의 구체적인 주상병 코드 분석결과는 다음과 같고, 상위 4개의 질병 목록은 아래와 같다. (상위 4개 항목 순서대로 나열)

J20.9 상세불명의 급성 기관지염

J03.90 재발성으로 명시되어 있지 않은 상세불명의 급성 편도염

J30.4 상세불명의 앨러지비염

J02.9 상세불명의 급성 인두염

분석결과 감기 및 알러지 질환이 다수를 차지했다. 특히 이 질병들은 ‘상세불명의 급성’ 질환 및 알러지성 질환인데, 이는 환절기 건강관리 및 생활위생관리와 밀접한 관련성이 있으므로, 적절히 관리 및 사전예방이 된다면 다수의 저가형 진료가 내역이 감소추세를 보일 것이다.

스크린샷이(가) 표시된 사진

자동 생성된 설명

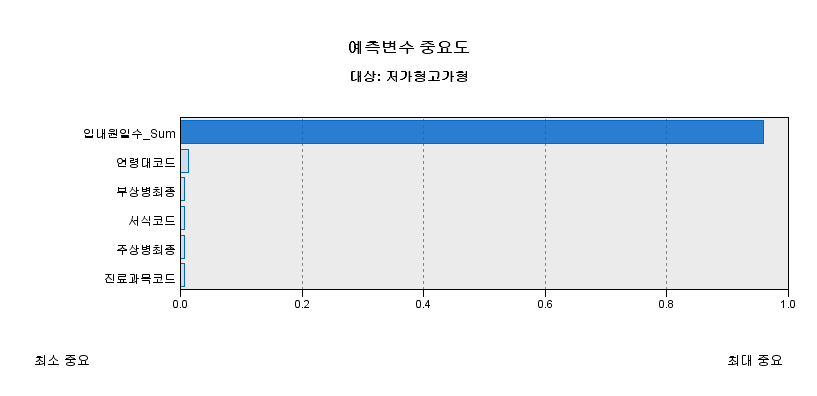
한편, 앞선 CART알고리즘에서 고가형 진료를 6.5일 기준으로 분리하였으나, 보다 세밀하게 분석하면 10일 이상, 특히 17일 이상에서 빈번하다. 따라서 고가형 진료에 대한 입원료 지원을 실행할 때 보다 세부적인 기준을 두어 진행하는 것이 바람직하겠다.

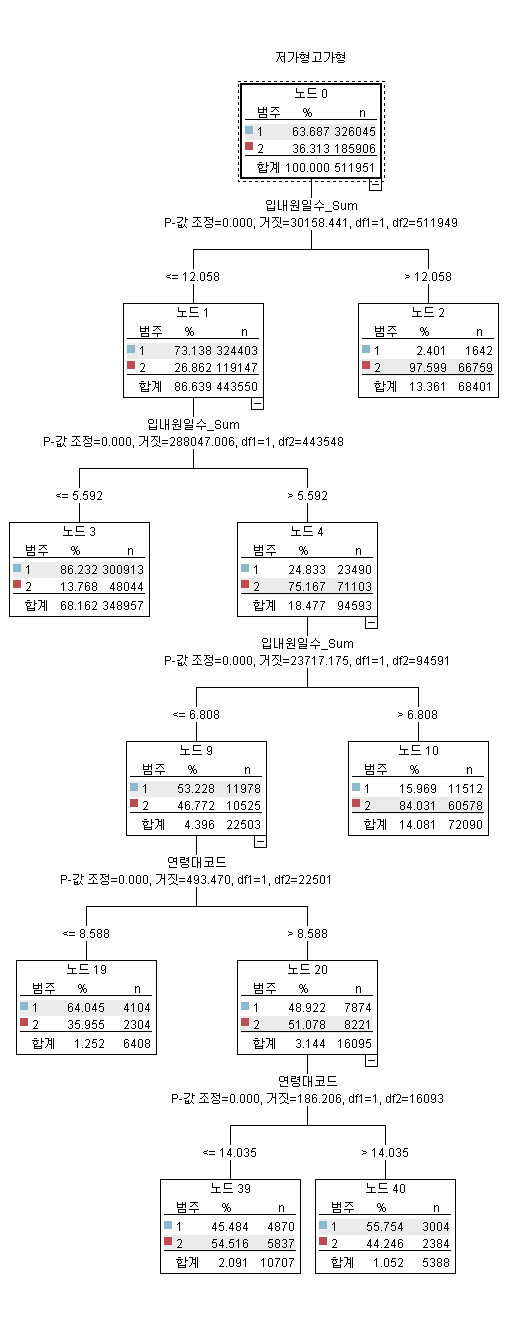
스크린샷이(가) 표시된 사진

자동 생성된 설명

추가적으로 고가형의 장기 입원이 가장 많이 나타나는 노드45의 하위 노드들을 살펴본다. 연령대 코드르 분류되며 연령대코드 13이상(60대 이상)에서 상대적으로 장기입원이 많다. 따라서 60세 이상 노년층에 대한 정기적인 건강검진을 확대 시행하여 고비용의 장기입원을 예방할 필요가 있다.

4-4. 분석결과(QUEST)





QUEST알고리즘 또한 입내원 일수에 대하 앞선 두 알고리즘의 결과를 반영한다. 하지만 예측변수 중요도가 입내원 일수에 치우쳐 입내원 일수 이외의 변수에 대한 유의미한 결과를 도출해내기 어렵다.

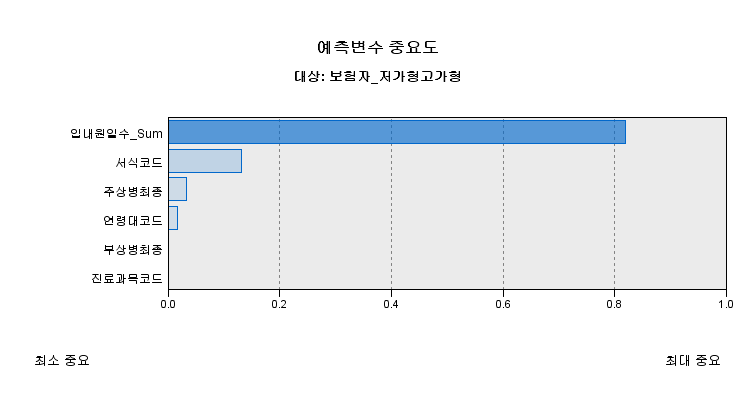
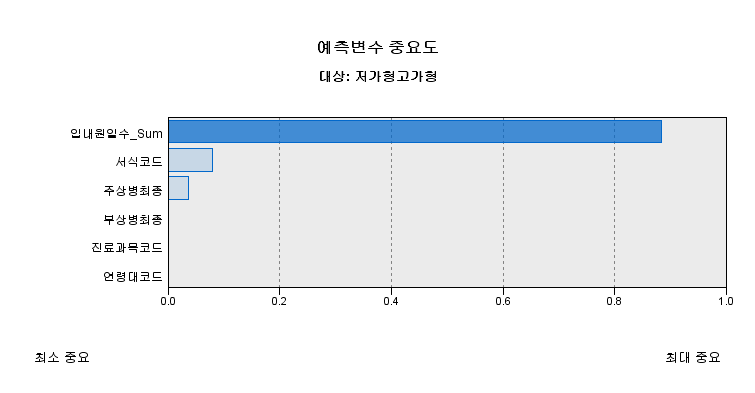
4) 의사결정나무 분석 – 목표값 : “심결보험자부담금\_Sum”

실재 건강보험공단의 보험료 지불 근거를 파악하기 위해 추가적인 분석을 진행했다. 1~3단계는 앞선 3)분석과 동일하며, 최종 결과는 다음과 같다.

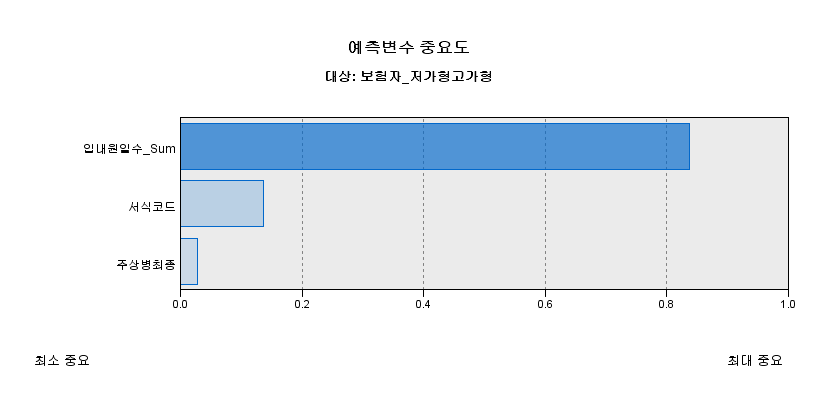
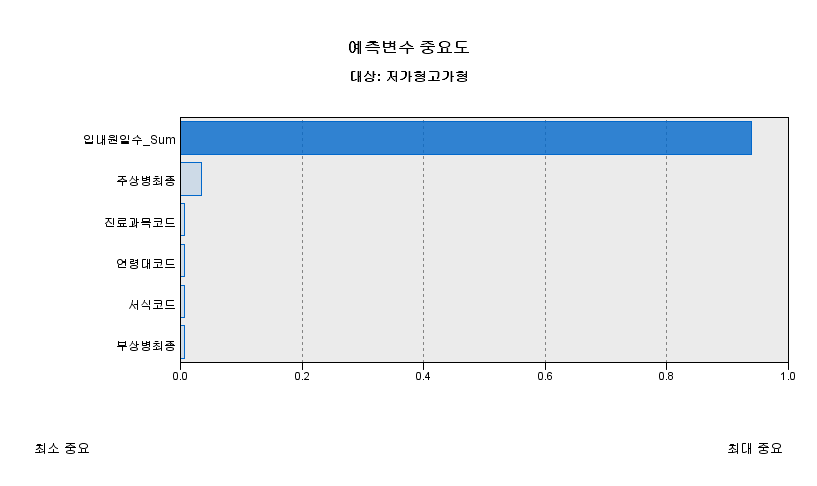
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 1. C5.0 | 2. CART | 3. CHAID | 4. QUEST |
| Accuracy | 98.64% | 95.93% | 94.88% | 95.9% |
| 트리깊이 | 22 | 2 | 4 | 4 |
| 예측변수  중요도 | 입내원일수 | 입내원일수 | 입내원일수 | 입내원일수 |
| 서식코드 | 주상병최종 | 주상병최종 | 서식코드 |
| 주상병최종 | 서식코드 | 연령대코드 | 부상병최종 |
| 연령대코드 |  | 부상병최종 | 주상병최종 |
| 부상병최종 |  | 진료과목코드 | 진료과목코드 |
| 진료과목코드 |  |  |  |

분석 결과 기존 총액에 대한 모델보다 모델 정확도가 상승했고, 모델별 예측변수 중요도에 변화가 나타났다. 따라서 각 모델별 예측변수 중요도의 변화를 비교하여 “실제 진료비” 산출 요소(목표변수-심결요양급여비용총액\_Sum)>와 “건강보험공단의 보험료” 산출 요소(목표변수-심결요양보험자부담금\_Sum)>을 파악한다.

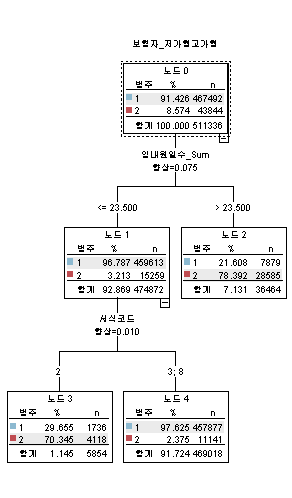
(\*왼쪽이 총액에 대한 분석 / 오른쪽이 보험공단(보험자) 지급액)



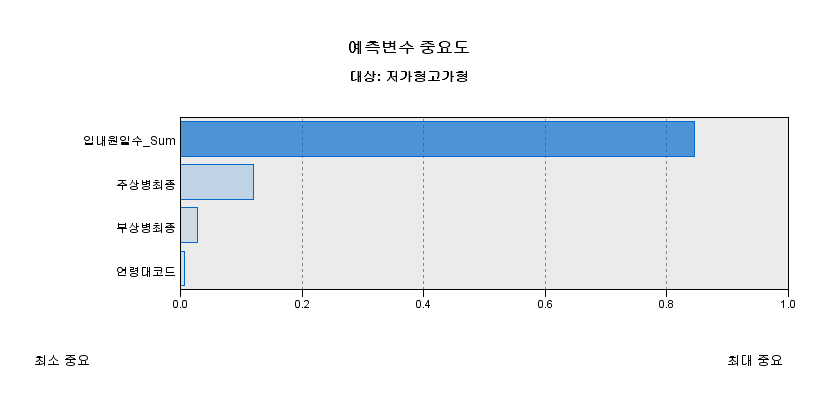
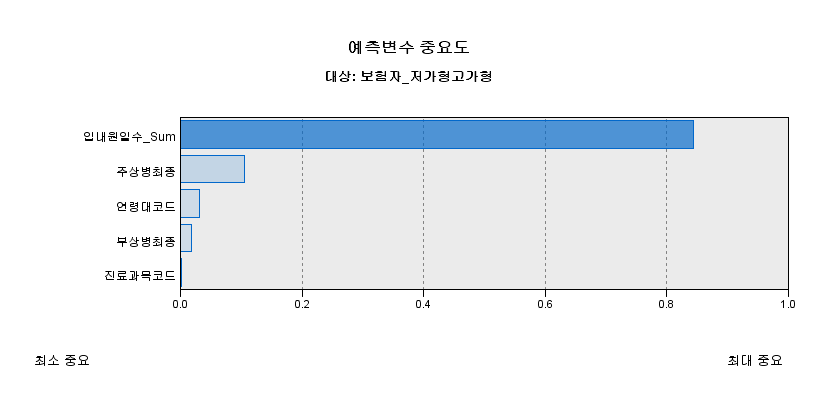
<C5.0 알고리즘> : 건강보험공단이 ‘연령대’에 대해 보험료 산정을 달리한다는 것을 알 수 있다.



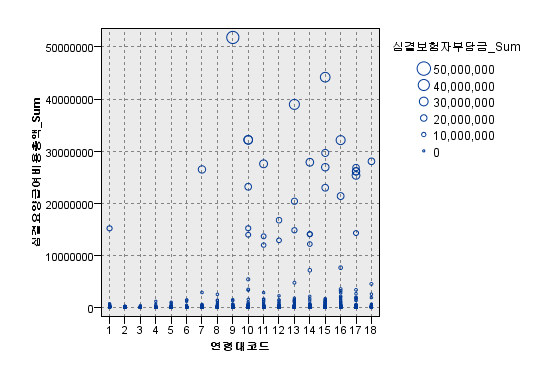
<CART 알고리즘> : ‘서식코드’에 변화가 관찰되는데, 이때 서식코드란 환자에 대한 진료형태의 구분이다.



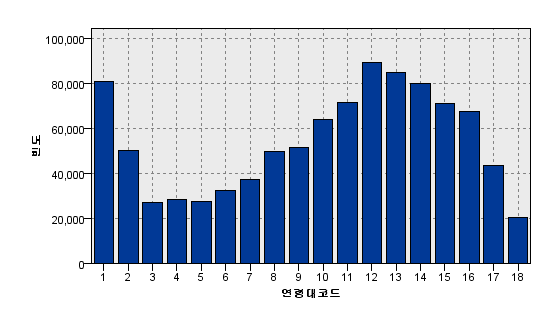
보다 구체적으로 살펴보면, ‘노드3’의 서식코드2는 의과입원을, ‘노드4’의 서식코드3과 4는 각각 의과외래 및 보건기관외래를 뜻한다. 즉, 건강보험공단의 보험료 산출 시 보건기관(시,도의 보건소) 진료 내역이 보다 유의미하게 반영된다. 따라서 공중보건 향상을 위한 보건소의 적극적 운영은 국민의 보험금 부담을 줄여줄 것이다.

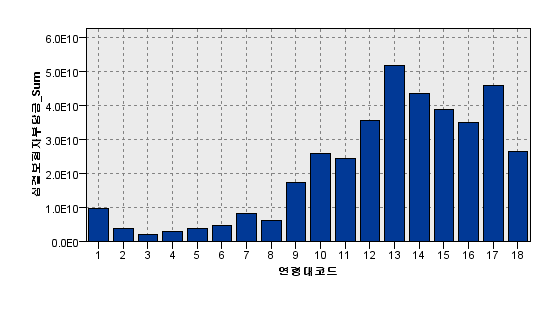
<CHAID 알고리즘> : 건강보험공단의 실재 보험료 지불 현황에 대한 분석에서 ‘연령대코드’에 대한 예측 중요도가 소폭 상승했다. 이는 앞서 언급한 고령환자(연령대코드13, 60세 이상)에 대한 보험료 추가적 지급이 어느정도 이루어지고 있다고 볼 수 있다. 그래프(a)를보면 실제로 고령환자에 대한 보험자 부담금이 높은 편이다. 다만, 아래의 그래프(c)와 같이 연령대코드14(65세)이후에 보험료 지급이 감소하므로 이들 고령자에 대한 적절한 지원이 요구된다.



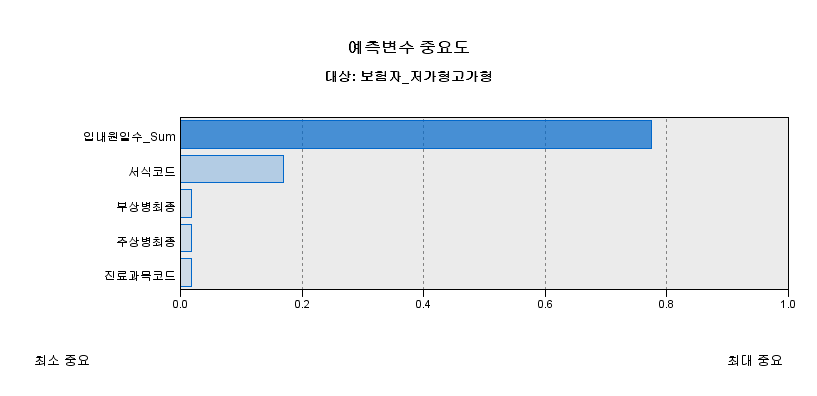
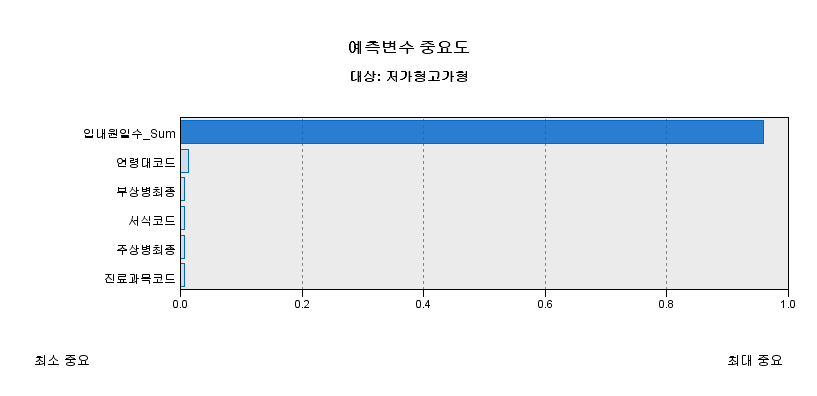
(a) X:연령대코드 / Y:심결요양급여비용총액 / Y2(bubble):심결보험자부담금



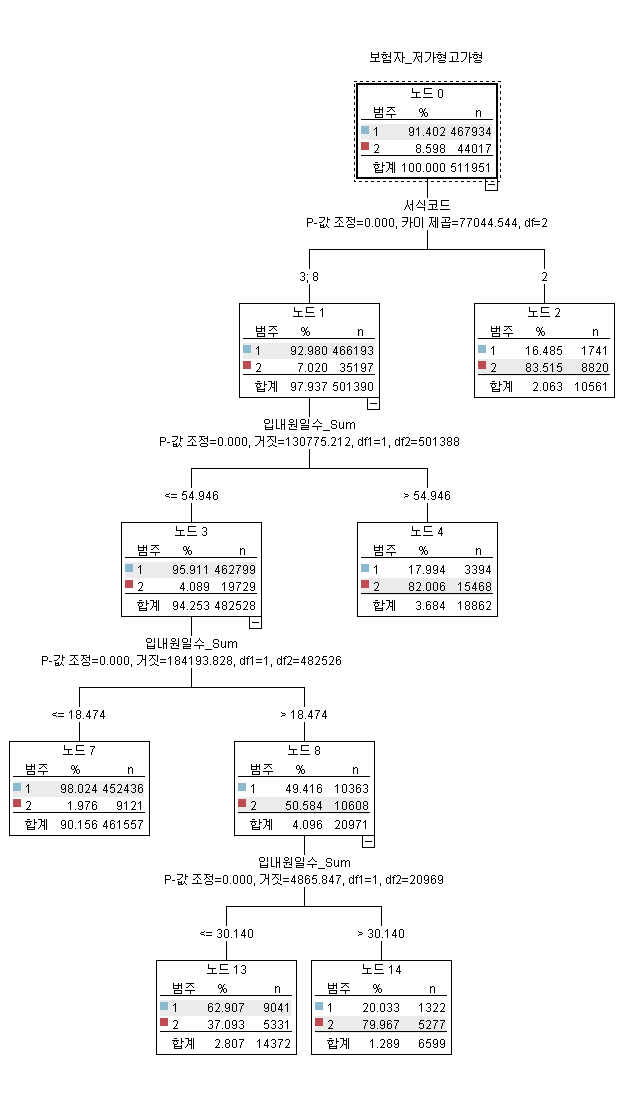
(b) X:연령대 코드 / Y:연령대 코드의 빈도



(c) X:연령대 코드 / Y:심결보험자부담금



<QUEST 알고리즘> : CART알고리즘과 마찬가지로 ‘서식코드’ 변수에 대한 변화가 나타났다.



위의 보험자부담금에 대한 의사결정나무를 보아 앞선 CART알고리즘과 동일한 결과를 얻을 수 있다.

5)전체 스트림 구조

지도, 텍스트이(가) 표시된 사진

자동 생성된 설명지도, 텍스트이(가) 표시된 사진

자동 생성된 설명

7. 프로젝트 결과 및 소감

데이터 마이닝이란 과목은 우리에게 매우 생소했다. 이전 과목들을 듣지 않고 처음 접해서 용어나 모든 것이 어려웠다. 직접 발표 주제를 정하면서, 양이 많기만 하면 좋은 데이터 인 줄 알았다. 하지만 발표 주제와 데이터를 직접 정하면서 분석할 만한 데이터와 아닌 데이터에 대해서 알게 되었다. 팀원과 함께 row data 전처리 하는 방식을 배워가며 연관성 분석이라는 역할을 맡았을 때, 언뜻 보기에는 비슷한 점이 많은 주상병, 부상병 코드이기에 단순한 작업이라고 생각했다. 하지만 생각보다 연관성 분석을 하는 과정이 매우 복잡했다. 또한 보기와는 다르게 수치적으로 나타난 연관성은 훨씬 낮았다. 유의미한 결과를 얻기 위해 여러 번 과정을 반복하며 어떤 방식으로 접근해야 할지 다양하게 생각하며 직접 데이터를 접했던 유일한 수업이였던 것 같다.

+) 초반 분석에 데이터에 대한 도메인 지식이 없이 분석 툴의 적용에만 초점을 맞춰 난항을 겪었다. 하지만 데이터를 이해하며 본질적으로 접근하니 문제점들이 해결되기 시작했고, 이를 통해 분석의 뚜렷한 목적과 이를 달성하기 위한 적절한 도구의 사용이 데이터마이닝의 핵심이라는 것을 알 수 있었다. 다만, 분석을 위한 도메인 지식을 학습했음에도 전문성이 부족한 영역이므로 분석의 정교함에 한계가 있었음에 아쉬움이 있다.