생키다이어그램을 활용한 의사결정나무분석

배성윤[1], 문성민[2], 최경철[3], 방선주[4], 손상준[5], 홍창형[6], 신현정[7], 이경원[8]

[123] 아주대학교 라이프미디어협동과정, [2] 낭떼르대학교 언어과학, [47] 아주대학교 산업공학과,
[56] 아주대학교 정신건강의학과, [8] 아주대학교 미디어학과

AAGCC 글로벌문화콘텐츠학회

의료인문융합콘텐츠센터 The Center for Medical Humanities and Convergent Contents

DECISION TREE ABOUT

CURRENT PATH

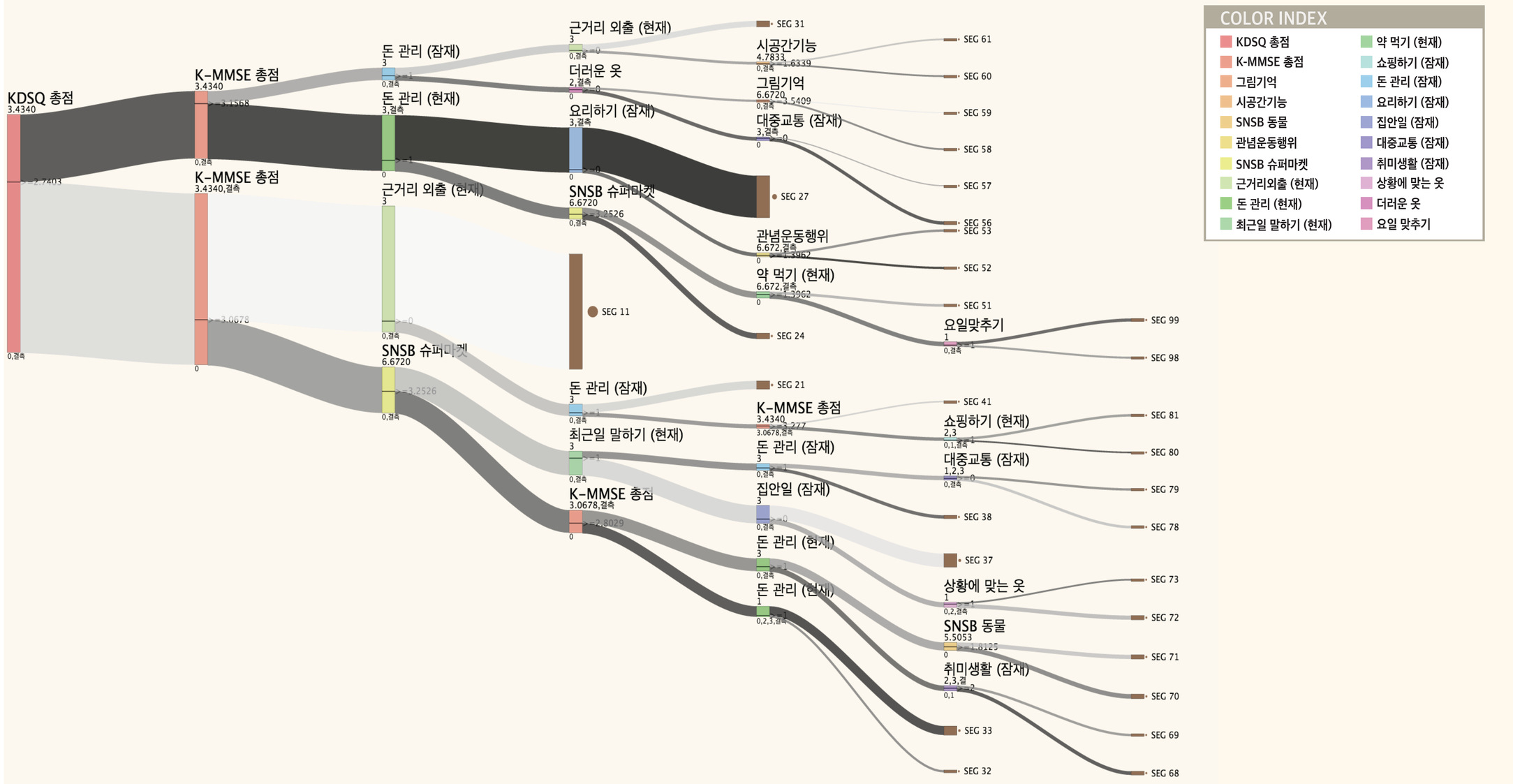

그림1. 생키다이어그램을 활용한 의사결정 나무 분석 시각화의 모습. 좌측에는 의사결정 나무의 시각화가 있고 오른쪽 위에는 결정 변수의 색인이 그려져 있다.

요약

의사결정 나무(Decision making tree)는 목표변수(Target variable)를 분류(Classification)하고 예측(Predict)하여 나무 구조로 나타낸다. 의사결정 나무 분석은 다른 계량적 분석 방법에 비해 쉽게 이해하고 활용할 수 있다. 그러나 방대한 데이터를 분석하면 나무구조가 복잡해 결과 해석이 어렵다. 본 연구에서는 의사결정 나무를 생키다이어그램(Sankey Diagram)으로 데이터가 많아져도 쉽게 분석할 시각화방법을 제시한다.

디자인 가이드 라인

시각화에 사용할 예제 데이터는 치매 환자(Dementia) 비율이 목표 변수이고 치매 환자의 심리검사들 점수가 입력 변수(Input variable)인 데이터를 활용하여 진행하였다. 목표 변수인 치매 환자 비율이 연속형이기 때문에 의사결정 나무 알고리즘 중 CHAID 를 사용한 회귀나무 분석(Regression tree)을 시각화로 나타내었다. 시각화 하는데 나타내어야 할 사항은 그림 2와 같다.

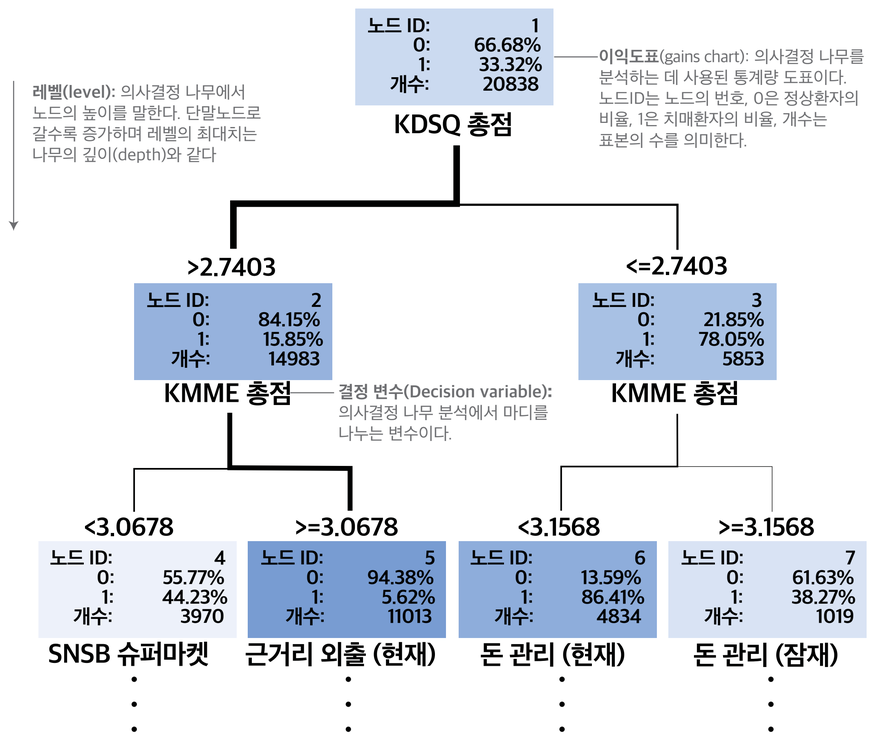

그림2. SAS에서 제공하는 Node-link diagram 형태의 의사결정 나무 분석 시각화의 일부이다. 라인이 굵을수록 포함된 환자의 수가 많다. 노드 위에 부등호는 결정 변수에서 분할 기준을 나타낸다.

시각화

그림1의 좌측을 보면 생키다이어그램을 활용한 의사결정 나무시각화를 볼 수 있다. 시각화에서 노드의 길이와 에지(edge)의 폭은 표본의 수를 나타낸다. 폭의 변화를 통해 표본수의 변화를 알 수 있다. 에지의 색상은 흰색에서 검은색으로 갈수록 치매 환자 비율이 높다. 그 외의 통계량은 그림3의 예시처럼 시각화의 상호작용을 통해 나타낸다. 시각화는 나무구조의 계층구조를 되어 있어 깊이를 알 수 있다. 각 노드의 레벨은 상호작용을 통해 노드의 가지(branch)를 보여준다. 마지막으로 결정변수는 노드의 위의 텍스트로 표기하고 색상으로 나타내었다. 같은 검사들은 같은 색상계열을 사용하여 그룹화하였다. 그리고 우측에 변수 색상 색인을 제공하여 총 결정 변수를 확인할 수 있다.

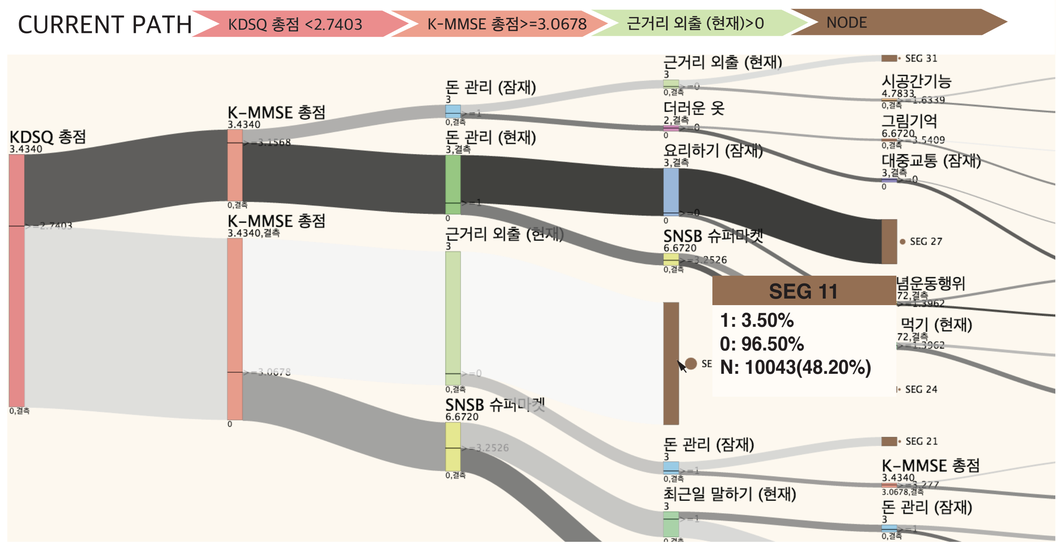

그림3. 노드 위에 마우스 커서를 올려놓으면 노드의 커서 바로 옆에 통계량의 수치가 나타난다. 화면 상단에 가지를 나타내서 선택한 노드까지 경로를 한눈에 파악하면서 어떤 기준값들을 노드가 가졌는지 알 수 있다. 가지의 나오는 변수의 개수로 선택한 노드의 레벨을 알 수 있다.

결론

본 연구에서 제작한 시각화는 기존의 그래프에서 직관적이지 못했던 정보의 표현을 시각적으로 개선하였다. 몇 가지 통계량과 결정 변수, 계층구조의 표현 개선을 통해 주요 단말 마디를 빠르게 찾을 수 있었다. 그러나 관측치의 수가 적어 노드의 길이가 너무 짧은 노드들은 시각화에서 그림만으로 비교하는 데 한계가 있어 비교하려면 마우스를 올려 통계량 정보를 봐야 하는 불편함이 있었다.