

대학 컴퓨터 실습 교양과목에서의 학업성취 요인에 대한 연구

김완섭

숭실대학교 베어드학부대학

요약

본 연구는 실습 중심의 컴퓨터 교양과목에서의 학생들의 학업성취에 영향을 미치는 요인을 발견하기 위한 것이다. 교과목 운영을 통한 교육성과 즉 학생들의 학업성취도를 향상시키기 위해서는 학업성취도에 영향을 미치는 요인들을 분석하고 그 결과를 교육에 반영하여 개선하는 순환 과정이 필요하다. 특히 컴퓨터 실습을 중심으로 하는 실용 교과목은 이론 중심의 과목들과 특성이 다르므로 그 요인에 대한 연구가 요구된다. 본 연구에서는 요인을 발견하기 위한 분석으로 로지스틱 회귀분석과 데이터마이닝 분야의 결정트리 분석을 수행하였다. 실험을 위한 데이터로는 서울소재 S대학의 교양필수과목에서 시행된 MOS 자격증 시험결과를 사용하였다. 로지스틱 회귀분석을 통해서 담당교수, 수강인원, 수업시간, 그룹(강의기간) 순으로 중요성을 파악할 수 있었다. 데이터마이닝의 의사결정트리 분석을 통해서 그 외에 학번, 재수강 여부, 강의실 환경의 추가 요인을 발견할 수 있었고, 특히 다양한 요인들이 학업성취에 복합적으로 영향을 미치는 것을 트리 모형을 통해 파악할 수 있었다. 분석 결과의 트리모형을 결과로 제시하였으며, 수식을 제안하여 여러 개의 트리 모형으로부터 요인의 중요도를 수치화하여 제시하였다.

키워드 : 컴퓨터 교양교육, 학업성취 요인, 로지스틱 회귀분석, 데이터마이닝, 의사결정트리

A Study on Factors of the Academic Achievement in Computer Training Courses as the Liberal Arts in University

Wanseop Kim

Baird University College, Soongsil University

ABSTRACT

The purpose of this study is to find out the factors of the students' achievement on the computer training courses which are based on computer practice. In order to improve the academic achievement of the students, it is necessary to analyze the factors affecting academic achievement and apply the results of the analysis to education. In particular, it is necessary to study for finding out factors of the academic achievement in practical computer training courses, because these courses are different from other courses focusing on the theory. In this study, in order to find out the factors, the logistic regression analysis and the decision tree analysis which is the field of data mining were performed. For the experimental data, the test results of the MOS certification of the S university in Seoul were used. Through logistic regression analysis it is found that the factors of the professors, class size, lecture time, group(lecture period) are important in order. Through decision tree analysis of data mining, it is found that there are some additional factors ; entrance year, whether the course is retaken, and the

논문투고 : 2013-10-07

논문심사 : 2013-10-08

심사완료 : 2013-10-29

classroom environment, and these various factors effect the academic achievement compositively as identified through the model tree. The tree model was presented as a result of the analysis, and the importance of the factors is expressed numerically from multiple tree models by using the proposed mathematical formula.

Key words : Computer Liberal Arts, Academic Achievement Factor, Logistic Regression, Data Mining, Decision Tree

1. 서론

교과목 운영을 통해 학생들의 학업성취도를 향상시키기 위해서는 교육에 영향을 미치는 요인을 분석하고 그 결과를 교육에 반영하는 순환 과정이 요구된다. 초·중등학교 및 대학 교육에서 학업성취도의 요인을 분석하는 다양한 연구들이 진행된 바 있다[1][3][4][10]. 그러나 대부분의 연구는 일반적인 교실 수업을 대상으로 연구가 진행되었다. 본 연구에서는 실습 중심의 컴퓨터 수업에 대하여 학업성취도의 요인을 분석하는 연구를 수행하고자 한다. 컴퓨터 실습을 중심으로 하는 실용 교과목은 이론 중심의 과목들과 특성이 다르기 때문에 그 요인에 대한 연구가 요구된다.

본 연구에서는 서울소재 S대학교의 교양필수 과목인 '컴퓨터활용1' 교과목에서 획득한 데이터를 통해 컴퓨터 교양교육에서의 요인을 분석하는 연구를 수행하였다. S대학교의 컴퓨터활용 교과목은 마이크로소프트(Microsoft)사의 MOS(Microsoft Office Specialist) 자격증 시험으로 학업성취도를 평가하고 있다. 일반적으로 대학의 강좌에서 분반 별로 객관적인 교육성과 데이터를 획득하는 것이 쉽지 않으며, 특히 컴퓨터 관련 교양 교과목에서 실무적인 컴퓨터 활용 능력에 대한 교육성과를 평가하는 것은 더욱 그렇다. 그러나 S대학교의 '컴퓨터활용1' 강좌에서는 CBT(Computer Based Test) 방식으로 실시되는 MOS 자격증 평가 도입을 통해 각 학생들의 실용적 컴퓨터 활용 능력을 객관적으로 평가할 수 있었고, 각 학생들의 획득 점수를 학업성취 측정 자료로 사용하여 컴퓨터 실습 교과목의 학업성취도 요인 탐색 연구를 진행하였다.

본 연구에서는 2013년 '컴퓨터활용1' 교과목을 수강한 3,179명의 학생들의 성적 데이터를 토대로 분석을 수행하였다. 학업성취에 영향을 미치는 요인을 발견하기 위한 분석 방법으로는 로지스틱 회귀분석과

의사결정트리 분석을 사용하였다. 로지스틱 회귀분석은 요인을 탐색하는 방법으로 많이 사용되고 있다[5][12]. 본 연구에서는 로지스틱 회귀분석을 적용하여 교육성과에 영향을 미치는 다양한 요인들의 영향력을 파악할 수 있었다. 또한 회귀분석만으로는 요인들 간의 계층적인 관계를 규명하는데 한계가 있기 때문에 본 연구에서는 데이터마이닝(Data Mining) 분야의 대표적인 분류 분석기법인 의사결정트리(Decision Tree) 분석을 수행하였다. 데이터마이닝 분야의 분석 방법들은 종속변수에 대한 요인을 탐색하는데 유용하게 사용되고 있다. 대표적으로 의사결정트리 기법은 요인들을 탐색하는 연구에 많이 활용된다[13][14][15][16]. 그 외에 신경망 알고리즘의 경우 값을 예측하는 모델을 생성할 수 있으며 오지영, 이수정은 다층 인식 모형(MLP) 신경망을 사용하여 컴퓨터 활용 능력의 정도를 예측하고 그 요인을 발견하는 연구를 수행하였다[11].

의사결정트리 분석은 종속변수에 영향을 미치는 요인들의 관계를 트리 형태로 제공해주며, 그 결과는 직관적으로 해석될 수 있고 의사결정에 유용한 규칙들을 발견하는데 큰 장점이 있다[6][20][21]. 본 연구에서는 의사결정트리 분석을 통해 학업성취도에 영향을 미치는 요인들 간의 관계를 트리 형태로 획득할 수 있었으며, 도출된 모델을 통해 교육성과에 영향을 미치는 중요한 요인들을 발견하고 강좌 운영에 적용할 수 있는 규칙들을 도출하였다. 특히 결정트리 분석에서는 CHAID, Exhausted-CHAID, CRT 그리고 QUEST 알고리즘을 적용하여 결정트리 분석이 입력 변수들 간의 상호관계에 취약한 문제를 보완하고자 하였다. 입력에 사용된 다양한 요인들 중에서 특히 담당교수, 실습실 환경, 수강인원, 강의시간, 강의기간(그룹) 등의 요인들이 복합적으로 학업성취도에 영향을 주는 것을 파악할 수 있었다.

본 논문의 2장에서는 관련 연구를 정리하였고, 3장에서는 분석에 사용한 데이터 및 분석방법을 소개하였다. 4장과 5장에서는 데이터를 사용하여 분석한 과정과 결과를 설명하였다. 4장에서는 로지스틱 회귀분석을 수행하여 교육성파에 영향을 미치는 요인들을 발견하고, 요인들의 중요성을 정리하였다. 5장에서는 데이터마이닝의 의사결정트리 알고리즘을 사용하여 분석한 결과를 정리하였다. CHAID, CRT 및 QUEST 의사결정트리 알고리즘을 수행하여 도출된 트리 모형을 분석하였다. 6장에서는 트리 모형으로부터 요인의 중요도를 수치화하는 수식을 제안하고 요인의 중요도를 수치화하여 제시하였다. 마지막으로 7장에서는 연구의 내용을 정리하고 향후 연구 방향들을 제시하였다.

2. 관련 연구

다양한 교육 분야에서 학업성취도에 영향을 주는 요인을 탐색하는 많은 연구들이 진행된 바 있다. 학업성취도의 요인은 교실 내적 요인과 교실 외적 요인이 모두 고려될 수 있다. 금지현, 이용환은 초등학생의 실과 교육 학업성취도가 부모의 사회적 수준, 양육의 태도 등 교실 외적 요인의 영향을 받음을 제시하였다[1]. 박동준, 최수영은 초등학교 수학 과목의 학업성취도의 요인으로 가정환경, 사교육, 컴퓨터활용도, 학습태도 등 교실 내적·외적 요인의 영향을 받음을 제시하였다[4]. 그 외에도 학생의 심리적 요인, 사전지식, 학교적응력 등 다양한 요인들에 대한 연구가 진행되었다[9][12][19].

대부분의 학업성취도의 요인을 탐색하는 연구의 경우 교실수업 과목 또는 전반적인 학업성취에 대하여 연구를 수행하였다. 그러나 대학에서 실습을 중심으로 하는 컴퓨터 수업의 경우 일반적인 교실수업과는 그 특성이 다르기 때문에 이에 대한 별도의 연구가 필요하다. 컴퓨터 실습실 여건(노후도, 빔프로젝터 설치 수 등), 실습실의 크기(대형/중형/소형), 학생의 전공(IT관련학과 여부), 수강인원 등 더욱 다양한 요인이 학업성취도에 영향을 줄 수 있기 때문이다. 특히 대학에서 컴퓨터 교양수업의 경우 다양한 학과, 학년, 나이의 학생들이 섞여서 수업을 듣기 때문에 이러한 다양한 요인들이 학생의 학업성취에 어떻게

영향을 미치는가를 연구하는 것은 효과적인 컴퓨터 실습 교양수업을 위해 필요하다.

소연희는 학습성취도에 영향을 주는 요인으로 학습자의 특성, 교육자의 특성, 교육의 내용 그리고 수업환경의 네 가지로 제시하였다[7][8]. 첫째 학습자의 특성은 일반적으로 학습자의 인지능력, 사전지식, 적성, 인구통계학적 특성 및 흥미를 포함한다. 컴퓨터 실습 교양교육에서는 학생의 전공, 학년, 성별 등의 특성이 고려될 수 있다. 즉, 컴퓨터학부 정보통신학과 등 IT 분야와 관련이 높은 학과일수록 컴퓨터 실습 교양과목에 대한 학습능력이 우수할 것이고 이러한 특성이 학습성취도에 영향을 미칠 수 있음을 고려할 수 있다. 두 번째, 교육자의 특성에는 담당교수의 나이, 성별 등의 인구통계 속성이나 강의경력 및 능력 등이 해당된다. 세 번째, 교육내용에는 교과내용, 과제내용 등이 포함될 수 있다. 네 번째 수업환경은 강의실의 시설 등의 유형적인 요소와 평가방법, 자율성의 정도 등의 무형적인 요소를 포함한다. 특히 컴퓨터 실습 수업의 경우 실습실 및 컴퓨터 장비의 노후 정도, 실습실의 크기 등이 교육성파에 영향을 미칠 수 있다. 또한 수업환경에서는 실습실, 강의시간 등을 고려할 수 있다. 일반적으로 대학 1학년 학생들의 경우 1교시 수업과 그 후의 수업을 비교할 때 학생들의 출석률 및 수업참여 태도에 차이가 발견되고 있기 때문에 수업시간 변수가 요인의 하나로 분석되어야 할 것으로 인식하였다.

대학 수업에서 학생들의 학업성취도의 요인을 발견하기 위한 다수의 연구들이 진행된 바 있다. 김현철은 설문을 통한 대학생들의 학업성취의 요인을 탐색한 연구를 실시하였다[2]. 이 연구에서 학교만족도, 학과만족도, 강의만족도 및 재수여부, 입학희망 우선순위, 출석률 등의 변수를 고려하여 학업성취의 요인을 탐색하였고, 분석 결과 학생의 학업성취도는 대입 전형요소보다 대학생들의 만족도에 높은 영향을 받는다고 제시하였다[2]. 그 외에 주영주 등은 전문대학 공학계열 컴퓨터 수업을 대상으로 학업 만족도와 학업성취도 간의 관계를 연구하였고, 학습내용의 유용성과 학생의 학습몰입 정도가 학업성취에 영향을 주는 것을 제시하였다[18]. 전병호는 대학교의 정보교육 특강 수강자들의 설문과 시험성적 데이터를 토대로

교육성과에 대한 요인을 탐색하는 연구를 수행하였고 컴퓨터 자기효능감, 혁신성, 동기, 자기조절 효능감 및 교수자 특성이 교육성과에 유의한 영향을 미친다고 제시하였다[17].

지금까지 수행된 학업성취도에 대한 요인 탐색 연구들은 대부분 일반적인 교실 수업을 대상으로 실시되었다. 또한 설문 조사 자료를 토대로 분석을 수행되었기 때문에 학생들의 객관적인 학업성취 자료를 사용하지 못하는 한계를 가진다. 본 연구는 실습을 중심으로 하는 컴퓨터 교양 수업을 대상으로 요인을 탐색하고자 하였으며, 수업의 평가 요소로 실시된 마이크로소프트사의 국제공인 자격증 점수를 학업성취 데이터로 사용하여 객관적인 분석을 실시하고자 하였다. 또한 컴퓨터 실습 교양 수업을 대상으로 그 요인을 탐색하는 연구로 진행하였다.

3. 연구 방법

3.1 자료의 준비와 기초 분석

본 연구에 사용된 데이터는 서울소재 S대학교의 교양필수 과목인 '컴퓨터활용1'의 2013년 진행된 모든 분반의 수업환경 및 수강생들의 정보로서 3,179명의 학생들의 정보로 구성되어 있다. 자료는 아래의 (그림 1)과 같은 형태로 구성되어 있다. 학업성취의 요인으로 사용될 수 있는 15개의 입력변수들과 학업성취를 의미하는 3개의 종속변수로 사용할 수 있는 변수들로 구성되어 있다.

| 그룹 | 분반코드 | 강의실코드 | 강의실 건물 | 강의실 크기 | 수강 인원 | 수업시간 | 재수강여부 | 담당교수 | 소속대학 | 강의과목 | 학년 | 학번 | 나이 |
|----|------------|-------|--------|--------|-------|------|-------|------|-------|-------|-----|------|----|
| B | 2150533064 | 02208 | 경상관 | 중형 | 43 | 오후3시 | 재수강안 | 교수1 | 인문대학 | 중어중문 | 2학년 | 2012 | 19 |
| B | 2150533064 | 02208 | 경상관 | 중형 | 43 | 오후3시 | 재수강안 | 교수1 | 인문대학 | 기독교학과 | 2학년 | 2012 | 20 |
| C | 2150533065 | 02208 | 경상관 | 중형 | 42 | 오후3시 | 재수강안 | 교수1 | IT대학 | 컴퓨터학부 | 2학년 | 2010 | 23 |
| C | 2150533065 | 02208 | 경상관 | 중형 | 42 | 오후3시 | 재수강안 | 교수1 | 인문대학 | 철학과 | 2학년 | 2012 | 20 |
| B | 2150533064 | 02208 | 경상관 | 중형 | 43 | 오후3시 | 재수강안 | 교수1 | 사회과학대 | 인문홍보 | 4학년 | 2010 | 22 |
| C | 2150533065 | 02208 | 경상관 | 중형 | 42 | 오후3시 | 재수강안 | 교수1 | 인문대학 | 국어국문 | 4학년 | 2007 | 25 |
| C | 2150533065 | 02208 | 경상관 | 중형 | 42 | 오후3시 | 재수강안 | 교수1 | 사회과학대 | 사회복지 | 4학년 | 2010 | 21 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 사학과 | 1학년 | 2013 | 20 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 불어불문 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 생활체육 | 1학년 | 2013 | 20 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 불어불문 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 사학과 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 독어독문 | 1학년 | 2013 | 18 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 생활체육 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 생활체육 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 국어국문 | 1학년 | 2013 | 20 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 사학과 | 1학년 | 2013 | 19 |
| A | 2150533002 | 10209 | 벤처관 | 중형 | 55 | 오후1시 | 일반안 | 교수1 | 인문대학 | 사학과 | 1학년 | 2013 | 19 |

(그림 1) 실험 데이터의 일부

학업성취를 나타내는 자료로는 학생들의 '시험점수'(수치형), '점수등급'(범주형), '합격여부'(이분형)의 3개의 변수를 준비하였다. 그리고 학업성취의 요인으로 사용될 수 있는 자료로 학생의 학년, 나이, 소속대학, 소속학과 그리고 수업시간, 강의실 등 15개의 변수를 준비하였다. 학업성취에 영향을 미치는 요인들은 일반적으로 네 가지로 구분된다. 첫 번째는 학습자의 특성, 두 번째는 교육자의 특성, 세 번째는 수업의 내용, 그리고 네 번째로는 수업의 내용이다. 네 번째 특성인 수업의 내용의 경우 모든 분반에서 기본적으로 동일한 교재를 사용하였으며 커리큘럼 상에도 큰 차이가 없었기 때문에 분석에서 사용할 수 있는 변수를 획득할 수 없었다. 물론 과제나 별도의 보충 자료, 동영상 자료 제공 등 담당교수 별로 수업의 내용의 약간의 차이가 존재하지만 객관적인 자료가 아니므로 실험에서는 제외하였다. 본 연구에서는 각 특성 별로 사용할 수 있는 변수들을 <표 1>과 같이 준비하였다.

<표 1> 입력 변수의 분류

| 구분 | 해당 변수들 |
|--------|--|
| 학습자 특성 | 학년, 학번(입학년도), 나이, 소속대학, 소속학과, 재수강여부 |
| 교육자 특성 | 담당교수코드, 성별, 나이 |
| 수업 환경 | 분반, 수강인원, 강의실코드, 강의실규모, 강의시간, 그룹(강의기간) |

주요 변수별로 학업성취에 미치는 영향을 파악하기 위하여 일반적으로 교육성과의 요인으로 인식되는 몇 개의 요인에 대하여 합격률을 <표 2>에서 <표 9>의 표로 정리하였다. 참고로 학생들의 최종 합격률은 약 75%이지만 본 실험에서는 1차 시험에 대한 합격률, 즉 제시문의 결과를 제외한 합격률을 사용하였다.

<표 2>는 소속대학별 합격률을 정리한 표이다. <표 2>를 통해 수강학생의 소속대학별로 합격률에 편차가 있음을 확인할 수 있다. 그러나 소속학과의 특성에 의하여 합격률의 차이가 발생했다고 단정할 수는 없다. 각 소속대학별로 강의가 진행된 기간이 다르고 담당교수도 동일하지 않기 때문에 다른 요인으로 인해 이러한 합격률 편차가 나타날 수 있기 때문이다. 일반적으로 컴퓨터 기술과 관련성이 높은 IT 대학이나 공과대학 등 자연계열 학생들이 합격률이

높을 것으로 생각하지만 <표 3>에서는 이러한 인식이 맞지 않음을 알 수 있다. IT대학, 공과대학, 자연과학대학 소속 학생들보다 경영대학, 경제통상대학, 사회과학대학 학생들의 합격률이 오히려 높은 점을 확인할 수 있다. 이를 통해 컴퓨터 실습 교양교육에 있어서는 학생들의 가지고 있는 IT에 대한 전문성이나 전공 교과목 수강을 통한 사전지식의 내용이 학업성취에 크게 영향을 미치지 않음을 알 수 있다.

<표 2> 학생의 소속대학별 합격률 비교

| 소속대학 | 응시자 | 합격자 | 불합격자 | 합격률 |
|--------|-----|-----|------|-------|
| IT대학 | 613 | 378 | 235 | 61.7% |
| 경영대학 | 304 | 193 | 111 | 63.5% |
| 경제통상대학 | 215 | 141 | 74 | 65.6% |
| 공과대학 | 758 | 452 | 306 | 59.6% |
| 금융학부 | 62 | 38 | 24 | 61.3% |
| 법과대학 | 125 | 65 | 60 | 52.0% |
| 사회과학대학 | 261 | 181 | 80 | 69.3% |
| 인문대학 | 561 | 271 | 290 | 48.3% |
| 자연과학대학 | 280 | 156 | 124 | 55.7% |

<표 3> 담당교수 별 합격률 비교

| 교수코드 | 응시자 | 합격자 | 불합격자 | 합격률 |
|------|-----|-----|------|-------|
| 교수01 | 782 | 520 | 262 | 66.5% |
| 교수02 | 65 | 27 | 38 | 41.5% |
| 교수03 | 170 | 95 | 75 | 55.9% |
| 교수04 | 129 | 78 | 51 | 60.5% |
| 교수05 | 278 | 161 | 117 | 57.9% |
| 교수06 | 53 | 22 | 31 | 41.5% |
| 교수07 | 52 | 32 | 20 | 61.5% |
| 교수08 | 295 | 110 | 185 | 37.3% |
| 교수09 | 280 | 186 | 94 | 66.4% |
| 교수10 | 279 | 147 | 132 | 52.7% |
| 교수11 | 311 | 189 | 122 | 60.8% |
| 교수12 | 110 | 66 | 44 | 60.0% |
| 교수13 | 375 | 242 | 133 | 64.5% |

<표 3>을 통해 담당교수 별로 합격률에 큰 편차를 나타냄을 볼 수 있다. 이를 통해 컴퓨터 교양 강좌에서 담당교수의 교육 역량이 학업성취에 매우 중요한 요소임을 판단할 수 있다. 그러나 담당교수가 교육한 학생들의 소속대학, 소속학과가 다르며 강의 기간과 시간이 달라서 동일한 교육 조건이 아니기 때문에 담당교수의 역량이 절대적이라고 단정할 수는 없다. 본 연구에서 진행한 로지스틱 회귀분석과 의사

결정트리 분석을 통해 다른 변수들을 함께 고려하여 요인들의 계층적 관계와 중요도를 파악할 수 있으며 이 결과는 4장과 5장에서 설명한다.

그 외에 강의기간(A/B/C그룹), 강의시간 그리고 강의실 환경에 대한 내용에 대하여 합격률을 분석하여 표로 정리하였다. 강의 기간의 경우 일반적인 강의와 달리 S대학교에서는 집중력 있는 강의와 실습실의 효율적인 활용을 위하여 16주 강의를 아닌 5주 단위 강의(주당 75분 2회 강의로 5주간 진행됨)로 진행하고 있다. 즉, A그룹은 개강 후 첫 5주 수업을 진행하여, B그룹은 그 후 5주(중간고사 기간과 겹침) 동안 진행되고, C그룹은 그 후 5주간(종강 기간에 겹침) 진행된다. 수업 기간별로 합격률에 큰 편차가 있으며 수업 기간이 학업성취도에 영향을 주는 것으로 판단된다. 그 외에 수업시간, 실습실 크기, 수강학생, 재수강 여부에 대하여 합격률을 비교하였다. 강의시간에 따라 합격률의 편차가 큰 것을 확인할 수 있는데 특히 오전 9시 수업과 저녁 7시 수업의 합격률이 저조함을 확인할 수 있다. <표 8>을 통해 학년이 높을수록 대체로 합격률이 높은 것을 확인할 수 있고, <표 9>를 통해 재수강반 학생들의 합격률이 상대적으로 높은 것을 확인할 수 있다.

<표 4> 수강학생의 소속대학별 합격률 비교

| 그룹 | 응시자 | 합격자 | 불합격자 | 합격률 |
|-----|-------|-----|------|-------|
| A그룹 | 1,102 | 665 | 437 | 60.3% |
| B그룹 | 1,051 | 593 | 458 | 56.4% |
| C그룹 | 1,026 | 617 | 409 | 60.1% |

<표 5> 강좌의 수업시간에 따른 합격률 비교

| 수업시간 | 응시자 | 합격자 | 불합격자 | 합격률 |
|-------|-------|-----|------|-------|
| 오전 9시 | 1,230 | 693 | 537 | 56.3% |
| 오후 1시 | 1,575 | 938 | 637 | 59.6% |
| 오후 3시 | 104 | 81 | 23 | 77.9% |
| 오후 4시 | 51 | 41 | 10 | 80.4% |
| 오후 7시 | 219 | 122 | 97 | 55.7% |

<표 6> 클래스의 실습실에 따른 합격률 비교

| 실습실 | 응시자 | 합격자 | 불합격자 | 합격률 |
|--------|-----|-----|------|-------|
| 02208 | 476 | 268 | 208 | 56.3% |
| 10209 | 643 | 411 | 232 | 63.9% |
| 10B106 | 714 | 445 | 269 | 62.3% |
| 11310 | 695 | 392 | 303 | 56.4% |
| 12427 | 651 | 359 | 292 | 55.1% |

<표 7> 실습실 크기에 따른 합격률 비교

| 구분 | 응시자 | 합격자 | 불합격자 | 합격률 |
|----|-------|------|------|-------|
| 대형 | 1,409 | 837 | 572 | 59.4% |
| 중형 | 1,770 | 1038 | 732 | 58.6% |

<표 8> 수강학생의 학년에 따른 합격률 비교

| 학년 | 응시자 | 합격자 | 불합격자 | 합격률 |
|-----|-------|------|------|-------|
| 1학년 | 2,842 | 1649 | 1193 | 58.0% |
| 2학년 | 86 | 58 | 28 | 67.4% |
| 3학년 | 112 | 68 | 44 | 60.7% |
| 4학년 | 137 | 98 | 39 | 71.5% |

<표 9> 수강학생의 재수강반 수강여부에 따른 합격률 비교

| 구분 | 응시자 | 합격자 | 불합격자 | 합격률 |
|------|-------|-------|-------|-------|
| 일반반 | 3,040 | 1,766 | 1,274 | 58.1% |
| 재수강반 | 139 | 109 | 30 | 78.4% |

4. 로지스틱 회귀분석을 통한 요인 탐색

4.1 분석의 방법

본 장에서는 로지스틱 회귀분석을 수행하여 컴퓨터 실습 교양과목의 학업성취도에 영향을 미치는 요인들을 탐색하였다. 실험에는 SPSS Statistics 18 프로그램을 사용하였다. 실험에서 ‘시험합격여부’ 변수를 종속변수로 사용하였고, 분석에서 합격은 1, 불합격은 0으로 코딩되었다. 또한 대부분의 설명변수가 범주형이므로 가변수를 사용하였다. 가변수의 코딩방법에 따라 회귀계수의 해석이 달라지는데, 본 연구에서는 가장 일반적인 방법인 기준 셀 코딩방법(reference cell coding)을 사용하였다.

4.2 분석 수행 및 해석

회귀분석을 이용하여 중요 요인을 발견하고자 하는 경우, 각 변수의 표준화 회귀계수를 이용하거나 전체 모형에서의 각 회귀계수의 유의확률의 정도로 중요도를 평가하는 경우가 일반적이다. 하지만 본 연구에서는 설명변수들이 대부분 여러 범주를 가진 범주형 변수이므로 위의 방법들을 이용하는 것에 어려움이 있어 SPSS 패키지에 있는 변수선택 툴을 이용하였다.

진진단계선택법의 경우, 변수선택의 원리는 다음과 같다. 절편 항을 포함한 모형에서 시작하여 각 단계마다 모형에 포함되지 않은 설명변수 중 기준 통계량에 대한 가장 작은 유의수준을 가진 변수를 선택하여, 기준 절단 값(default 0.05)보다 작으면 모형에 넣는다. 또한 각 단계마다 이미 전 단계에서 모형에 포함된 변수들에 대해서 변수 제거 기준에 충족하는지를 검사하여 기준 통계량에 대한 가장 큰 유의수준을 가진 변수가 기준 절단 값(default 0.10)을 초과하면 모형에서 제거한다. 검사를 계속하다가 전에 가정한 모형을 만나거나 더 이상 선택 또는 제거 기준을 만족하는 변수가 없을 때에 변수선택을 종료한다.

이러한 변수선택의 원리를 통해 첫 단계에서 모형에 포함된 변수일수록 중요한 의미를 가지는 요인이라고 판단할 수 있다. 왜냐하면 모형에 포함되지 않은 변수들 중에 가장 작은 유의수준을 가진 변수를 선택하기 때문이다. 또한 선택되어 모형에 포함된 변수들만이 종속변수에 유의한 영향을 주는 요인이라고 생각할 수 있다. 이는 각 회귀계수의 유의확률의 정도로 중요도를 평가하는 방법과 같은 원리이다. 본 연구에서는 SPSS의 Wald 통계량을 기준으로 하는 진진단계선택법을 사용하여 변수 선택을 수행하였다.

변수 선택 결과는 (그림 2)에 표시하였다. 로지스틱 회귀분석을 통해 변수들이 선택된 순서를 통해 학업성취에 영향을 미치는 요인들을 파악할 수 있다. 변수가 선택된 순서를 통해 그 중요도를 확인할 수 있다. 변수가 선택된 순서는 ‘담당교수’, ‘수강인원’, ‘수업시간’ 그리고 ‘그룹(강의기간)’ 순이다.

첫 번째로 ‘담당교수’가 선택되었는데, 이는 담당교수의 교육역량이 학업성취에 미치는 영향이 큰 것을 보여준다. 특히 학생들의 소속대학, 소속학과, 학년 등의 학습자의 특성에 해당하는 요소보다 담당교수가 중요한 변수로 선택된 것은 주목할 만하다. 일반적으로 컴퓨터 전공교육의 경우 수강학생의 컴퓨터에 대한 기존의 전문적 지식, 프로그래밍 능력 등이 학업성취에 큰 영향을 미치지만 컴퓨터 교양과목의 경우 전문적인 기반 지식을 요구하지 않기 때문에 담당교수의 교육역량과 학생들의 학습태도가 오히려 더 중요한 요소임을 알 수 있다.

반정식에 포함된 변수

1 - 100 행(138 중에서)

| | B | S.E. | Wals | 자유도 | 유의확률 | Exp(B) |
|-------------------|--------|-------|---------|-----|------|--------|
| 1 단계 ^a | | | 103.334 | 12 | .000 | |
| 담당교수 | | | | | | |
| 담당교수(1) | .087 | .132 | .434 | 1 | .510 | 1.091 |
| 담당교수(2) | -.940 | .274 | 11.789 | 1 | .001 | .390 |
| 담당교수(3) | -.362 | .188 | 3.694 | 1 | .055 | .696 |
| 담당교수(4) | -.174 | .210 | .685 | 1 | .408 | .841 |
| 담당교수(5) | -.279 | .163 | 2.955 | 1 | .086 | .756 |
| 담당교수(6) | -.942 | .299 | 9.920 | 1 | .002 | .390 |
| 담당교수(7) | -.129 | .305 | .178 | 1 | .673 | .879 |
| 담당교수(8) | -1.118 | .162 | 47.843 | 1 | .000 | .327 |
| 담당교수(9) | .084 | .166 | .254 | 1 | .614 | 1.087 |
| 담당교수(10) | -.491 | .161 | 9.260 | 1 | .002 | .612 |
| 담당교수(11) | -.161 | .159 | 1.029 | 1 | .310 | .851 |
| 담당교수(12) | -.193 | .223 | .753 | 1 | .386 | .824 |
| 상수항 | .599 | .108 | 30.753 | 1 | .000 | 1.820 |
| 2 단계 ^b | | | 72.407 | 22 | .000 | |
| 수강인원 | | | | | | |
| 수강인원(1) | .285 | .848 | .113 | 1 | .737 | 1.330 |
| 수강인원(2) | -.323 | .465 | .483 | 1 | .487 | .724 |
| 수강인원(3) | -1.838 | .456 | 16.283 | 1 | .000 | .159 |
| 3 단계 ^c | | | | | | |
| 담당교수(11) | -.727 | .272 | 7.129 | 1 | .008 | .483 |
| 담당교수(12) | -.267 | .276 | .934 | 1 | .334 | .766 |
| 상수항 | 2.194 | .445 | 24.283 | 1 | .000 | 8.967 |
| 4 단계 ^d | | | 8.570 | 2 | .014 | |
| 그룹 | | | | | | |
| 그룹(1) | .154 | .199 | .598 | 1 | .439 | 1.167 |
| 그룹(2) | -.429 | .182 | 5.528 | 1 | .019 | .651 |
| 수강인원 | | | 76.198 | 22 | .000 | |
| 수강인원(1) | -.803 | 1.066 | .567 | 1 | .451 | .448 |
| 수강인원(2) | -.311 | .519 | .358 | 1 | .550 | .733 |
| 수강인원(3) | -1.699 | .545 | 9.716 | 1 | .002 | .183 |

- a. 변수가 1: 단계에 진입했습니다 담당교수, 담당교수.
- b. 변수가 2: 단계에 진입했습니다 수강인원, 수강인원.
- c. 변수가 3: 단계에 진입했습니다 수업시간1, 수업시간1.
- d. 변수가 4: 단계에 진입했습니다 그룹, 그룹.

(그림 2) 로지스틱 회귀분석 분석 결과

두 번째로 ‘수강인원’이 선택된 것은 주목할 만하다. 이 결과는 수강인원이 많은 클래스의 경우 학업성취도가 낮아지고, 수강인원이 적을수록 학업성취도가 높아짐을 의미한다. 그러나 이 결과만으로는 수강인원이 몇 명 이하가 효과적인 것을 파악할 수 없으므로 다음 장에서 데이터마이닝의 의사결정트리 분석을 통해 효과적인 수업을 위한 수강인원 기준을 파악해보고자 한다.

세 번째로 수업시간이 선택되었다. 이는 <표 5>을 통해서도 알 수 있지만 강의시간이 수업에 영향을 주는 큰 요인인 것을 알 수 있다. 특히 9시 수업의 경우가 그 이후 강의에 비해 학생들의 지각률이 높고 수업 집중도가 저조하기 때문으로 이해될 수 있다. 특히 이 과목은 P/F 과목이기 때문에 수업 시간에 대한 영향이 큰 것으로 이해된다.

네 번째로 ‘그룹’이 선택된 것은 특기할 만하다. ‘그룹’은 해당 클래스의 수업이 진행된 기간을 의미한다.

다. 참고로 S대학교의 컴퓨터활용 수업은 집중적 강의 및 효과적 강의실 운용을 위해 5주 단위의 수업을 진행한다. 2013년의 경우 A그룹은 개강 후 첫 5주간인 2013.3.4.(월)~4.5.(금), B그룹은 그 후 5주간인 2013.4.8.(월)~5.10.(금) 그리고 C그룹은 마지막 5주간인 2013.5.13.(월)~6.14.(금) 기간에 진행되었다. 즉, ‘그룹’이 선택된 것은 강의 기간에 따라 합격률의 편차가 큰 것을 의미한다. 이는 본 1학년 대상의 컴퓨터 교양과목의 학업성취도가 학습 환경에 크게 영향을 받음을 보여준다. A그룹의 경우 교육성과(합격률)가 가장 높고, B그룹의 경우 교육성도가 가장 낮으며, C그룹의 경우 중간 정도에 해당함을 확인할 수 있다. A그룹의 경우 합격률이 상대적으로 높는데 이는 신입생들의 수업참여도, 집중력이 학기 초에 가장 높은 것으로 이해할 수 있다. B그룹의 경우 수업 기간이 중간고사에 겹쳐져서 다른 과목의 시험으로 인한 집중력 저하로 본 컴퓨터 교양수업에 대한 수업참여도가 떨어지는 것으로 이해할 수 있다.

지금까지 로지스틱 회귀분석을 통한 여러 가지 요인들의 중요도를 파악하고 분석하였다. 그러나 회귀분석의 경우 중요한 변수들 간의 계층적 관계를 분석해줄 수 없기 때문에 요인들의 영향력을 명확하게 이해하는데 한계가 있다. 다음 장에서는 데이터마이닝의 의사결정트리 분석을 통해 추가적인 분석을 수행하고 분석한다.

5. 의사결정트리 분석을 통한 요인 탐색

5.1 의사결정트리 분석 소개

본 장에서는 의사결정트리 분석을 사용하여 학업성취 요인을 탐색하는 실험을 수행하고 그 분석 결과를 해석하고자 한다. 의사결정트리 분석은 데이터마이닝의 대표적인 분류 기법 중 하나로서 하나의 종속 변수에 대하여 여러 개의 입력변수(요인)들이 계층적으로 어떠한 영향을 주는 가를 나무 형태의 시각화 모델로 표시해준다. 의사결정트리 기법은 분석된 결과가 이해하기 쉽고 의사결정에 유용한 규칙들을 도출하기 쉽기 때문에 많은 분야에서 특정 결과에 영향을 미치는 요인들을 발견하고 의사결정을 위한 분석 도구로 사용되고 있다. 트리 모형에서 가지(Branch)

를 분리하는데 사용된 변수를 종속변수에 영향을 주는 변수로 이해할 수 있으며 상위에 선택된 변수일수록 중요한 요인으로 이해할 수 있다. 또한 회귀분석의 경우 변수들 간의 계층적인 관계를 분석할 수 없지만 결정트리 분석의 경우 전체 데이터를 계층적으로 소규모의 데이터로 분리하면서 특정 그룹에 대한 특성을 분석할 수 있는 장점이 있다.

결정트리 분석은 전체 레코드를 반복적으로 분할하여 각 세분화된 각 영역의 순수도(Purity)를 향상시키며 트리 모형을 구축한다. 이러한 과정에서 불순도를 제거할 수 있는 변수를 선택하는데, 상위 노드에 위치하는 변수일수록 종속변수를 잘 분류하는 변수이며, 다시 말해 종속변수에 큰 영향을 주는 요인으로 해석할 수 있다. 결정트리 분석은 트리 모델을 생성하는 과정에서 가지(Branch)를 분리하는 변수를 선택하는 방법과 가지의 개수를 설정하는 방법 등에 따라 여러 가지 알고리즘들로 구분된다.

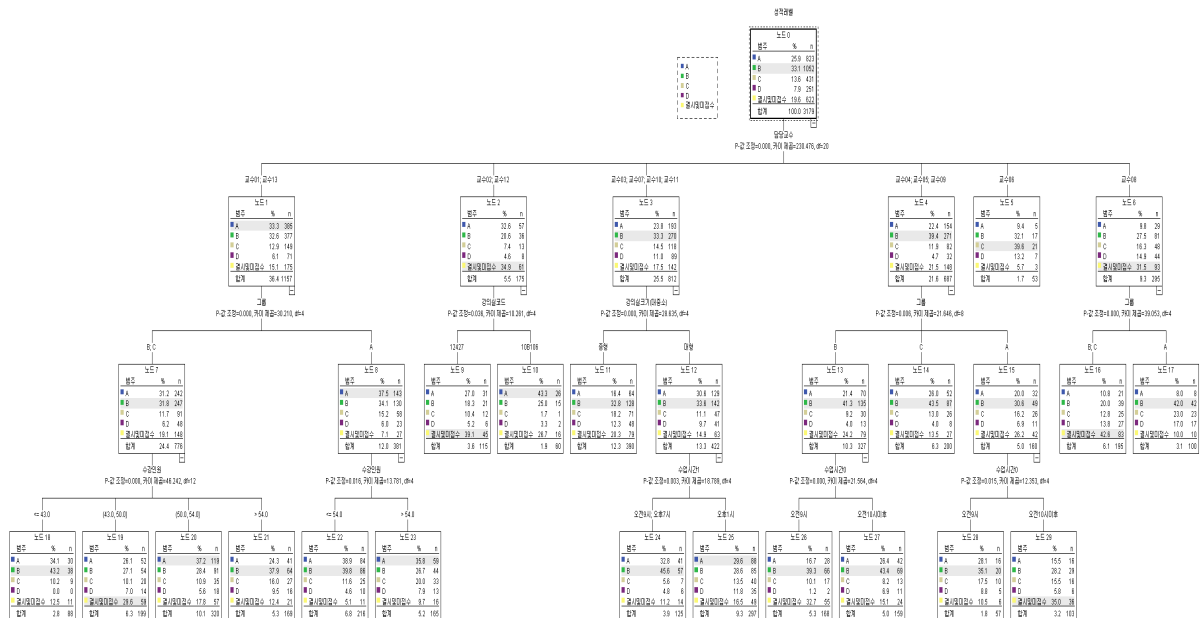
5.2 의사결정트리 분석 결과

본 연구에서는 SPSS Statistics 18에서 지원하는

CHAID, Exhausted-CHAID, CRT 그리고 QUEST를 적용하여 분석을 수행하였다. 실험에서 결과 모델의 단순화를 위해 트리의 깊이는 3으로 제한하였다. 본 실험에서는 학업성과를 의미하는 ‘점수등급’ 변수를 추가로 생성한 후 종속변수로 사용하였다. 시험 결과의 원 자료(Raw Data)에는 ‘시험점수’(연속형), ‘시험 합격여부’(이분형) 변수가 존재한다. 그러나 결정트리의 경우 종속변수로 범주형 자료만 적용 가능하므로 연속형 변수인 ‘시험점수’를 사용할 수 없고, ‘시험합격여부’ 변수의 경우 단지 ‘합격’, ‘불합격’ 결과를 포함하고 있어 학업성과의 정도를 표현하기에 정보의 손실이 발생하기 때문에 <표 12>를 기준으로 시험 점수를 등급화한 범주형의 ‘점수등급’ 변수를 생성하였다.

(그림 3)은 CHAID 결정트리 알고리즘을 수행한 결과이다. ‘담당교수’가 가장 중요한 요인으로 선택되었다. 시험 합격률에 따라 담당교수를 6개의 그룹으로 구분하고 있다. 실습을 중심으로 하는 컴퓨터 교양과목의 경우 학생의 전공보다 담당교수의 역량이 중요함을 확인할 수 있다. 두 번째 분리에서는 강의 기간을 의미하는 ‘그룹’이 많이 선택되었다.

A그룹의 경우 B그룹, C그룹의 경우보다 교육성과



(그림 3) CHAID 결정트리 분석 결과

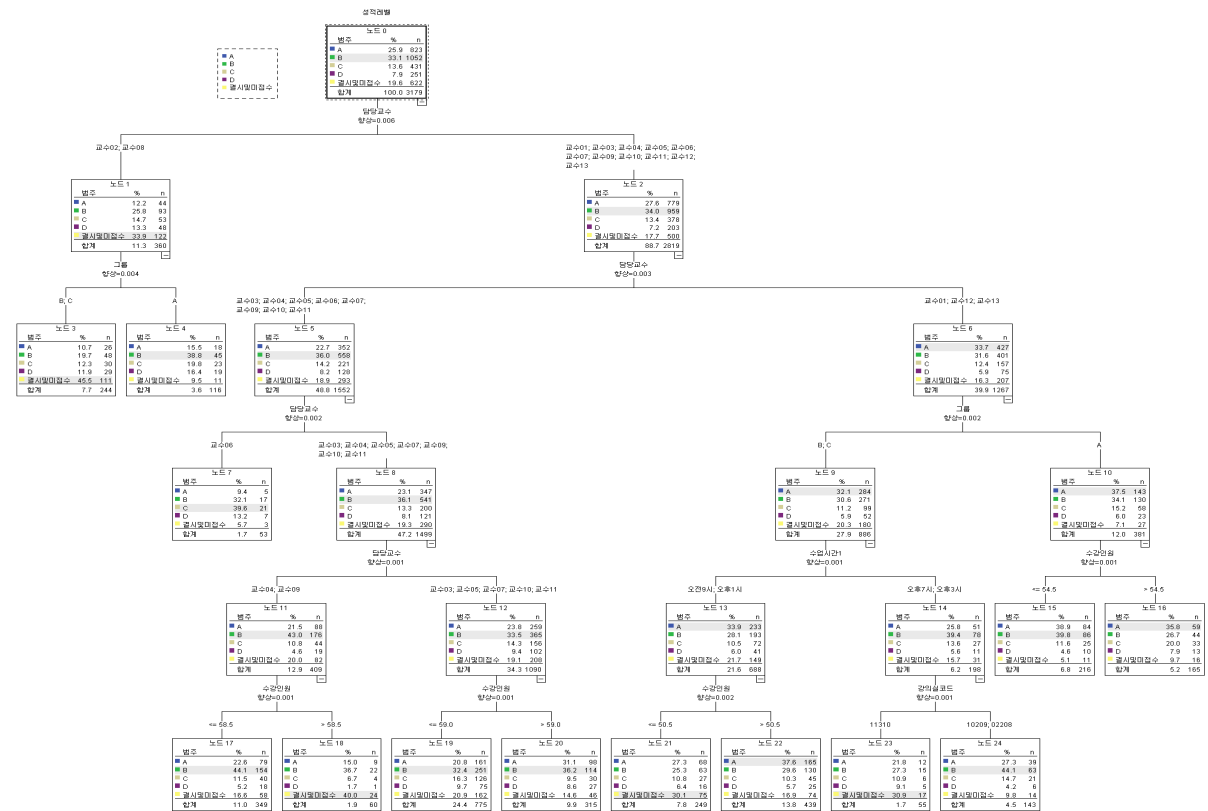
가 높는데, 이는 신입생들의 경우 학기 초(3월) 수업에 대한 참여도가 상대적으로 좋기 때문으로 이해할 수 있다. 반면, B그룹의 경우 교육성도가 가장 낮는데, 이는 수업 기간이 중간고사 기간과 겹쳐 있어 학생들의 본 컴퓨터 교양수업에 대한 참여도가 부족했기 때문으로 이해할 수 있다. 그 외에도 '강의실코드', '강의실크기'가 선택되었는데 이는 실습실의 노후도, 교수와 학생과의 거리, 프로젝트의 설치 개수 등의 수업환경 요인이 학업성취도에 영향을 주는 것으로 분석된다.

<표 12> 시험점수에 따른 '점수등급' 구분 기준

| 시험점수 | 점수등급 |
|--------------|--------|
| 800이상 1000이하 | A등급 |
| 700이상 800미만 | B등급 |
| 400이상 700미만 | C등급 |
| 400미만 | D등급 |
| 결시 또는 미점수 | 결시및미점수 |

세 번째 분리에서 '수강인원', '강의시간'이 선택되었다. '수강인원'의 경우 소규모 그룹일수록 학업성취도가 높음을 확인할 수 있다. 구체적으로는 43명 미만이 수강한 분반의 성적이 다른 그룹에 비해 합격률 및 높은 점수 등급(A,B)을 받은 비율이 높음을 확인할 수 있다. '강의시간'의 경우 오전 9시 수업을 하는 경우가 10시 이후에 하는 수업에 비해 상대적으로 학업성취도가 낮은 것을 확인할 수 있다. 물론 9시 수업 분반이 절대적으로 학업성취도가 낮은 것은 아니지만 <표 5>를 통해 전체 데이터의 분포에서도 볼 수 있듯이 9시 수업 분반, 저녁 7시 분반의 합격률이 다른 시간대에 비해 낮음을 통해 강의 시간에 학업성취에 영향을 주고 있음을 파악할 수 있다.

(그림 4)는 CRT 결정트리 알고리즘을 사용하여 분석한 결과이다. (그림 3) CHAID 알고리즘의 분석 결과가 동일하게 '담당교수'가 첫 번째 분기 변수로 선택되었다. 종속변수에 대한 몇 요인들의 영향력이 비

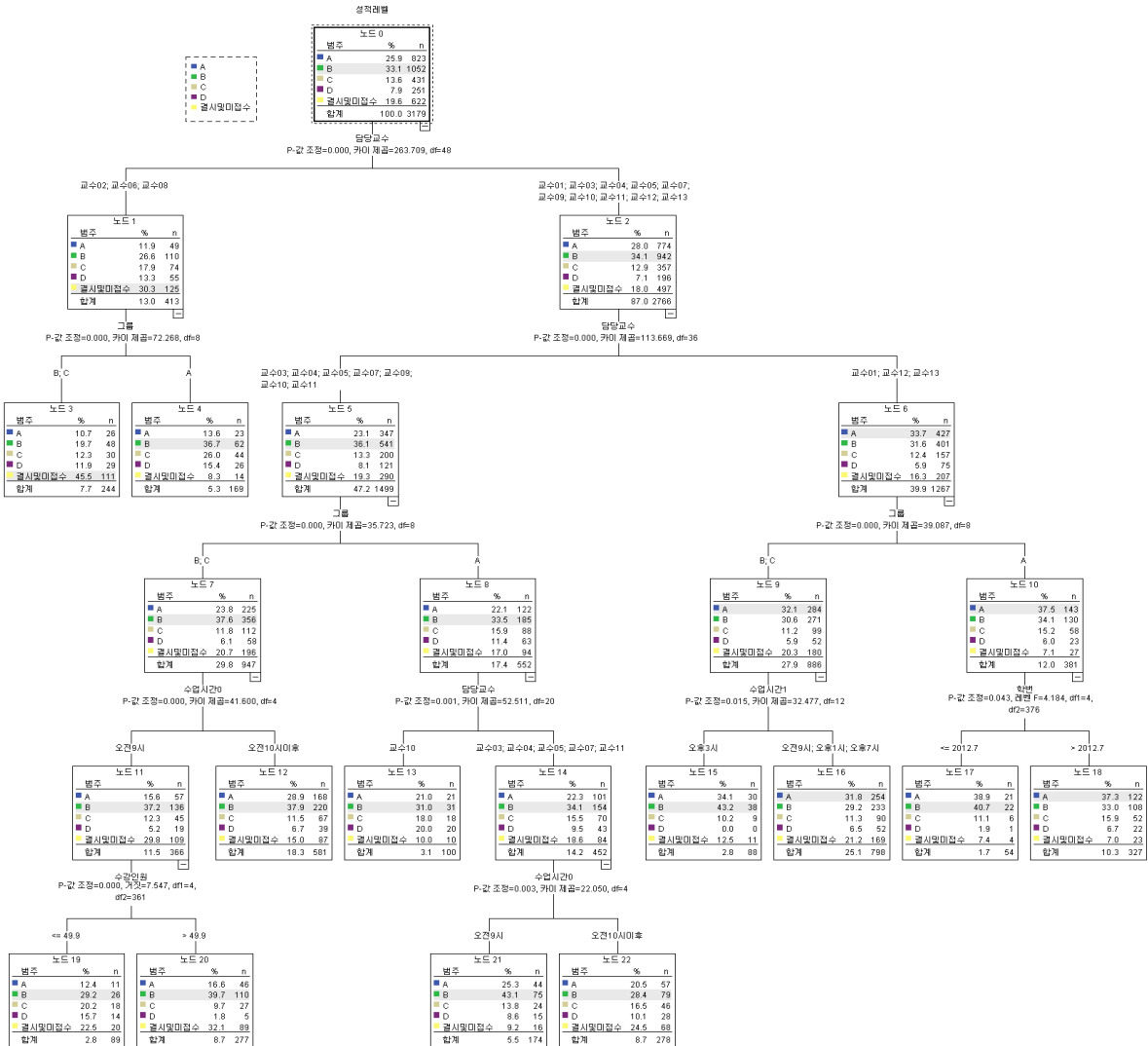


(그림 4) CRT 결정트리 분석 결과

숫할 경우 알고리즘에 따라 선택되는 변수에 차이가 나타나지만, 본 실험의 경우 CHAID, CART 그리고 QUEST 알고리즘에서 모두 '그룹' 변수를 첫 번째 분리 변수로 선택한 것을 볼 때 '그룹'이 가장 중요한 요인임을 확인할 수 있다. CART 알고리즘의 분석 결과의 경우 모든 분리에서 2개의 하위 그룹으로 분리됨을 볼 수 있는데 이는 알고리즘이 2개로 분기되도록 설계되었기 때문이다. 두 번째 및 세 번째 레벨에서 '그룹'이 '담당교수' 외에 추가로 선택된 것을 통해 '그룹'의 학업성취에 영향을 주는 중요한 요인임을

다시 확인할 수 있다. 마지막으로 네 번째 및 다섯 번째 레벨에서 '수강인원', '강의실코드'가 선택되었다. 수강인원이 적은 반이 수강인원이 많은 반보다 대체로 합격률 및 성적 레벨이 높음을 확인할 수 있다. 또한 강의실 환경에 따라 합격률 및 성적 레벨이 차이가 있음을 다시 확인할 수 있었다. 선택된 변수에 약간의 차이는 있지만 전반적으로 CHAID 알고리즘의 분석 결과와 유사한 결과를 보여주고 있다.

(그림 5)는 QUEST 알고리즘의 수행 결과이다. 이 결과에서도 '담당교수'가 가장 의미있는 변수로 선택



(그림 5) QUEST 결정트리 분석 결과

되어 첫 번째 분기에서 사용되었다.

두 번째 레벨 및 세 번째 레벨에서 ‘그룹’이 중요한 변수로 선택되어 사용되는 것을 확인할 수 있다. 네 번째 및 다섯 번째 레벨에서는 ‘강의시간’, ‘수업시간’, ‘수강인원’ 및 ‘학번’이 의미있는 변수로 선택되었다. 특히 다른 알고리즘에서 선택되지 않았던 ‘학번’이 의미있는 변수로 사용된 것은 주목할 만하다. 2012~13 학년의 학생들이 2011 이상의 학번보다 컴퓨터 교양과목에서 학업 성취도가 높게 나타난다. 작은 그룹에서 발생한 현상이기는 하지만 나이가 적을수록 컴퓨터를 다루는 능력이 뛰어나기 때문으로 해석된다.

본 장에서는 컴퓨터 교양과목에서의 학업 성취도의 요인 및 그들 간의 관계를 발견하기 위하여 SPSS Statistics 18.0에서 지원하는 결정트리 알고리즘들을 적용하고 그 결과를 설명하였다. CHAID, CART 및 QUEST의 적용 결과 그림 제시하고 설명하였다. Exhausted-CHAID 알고리즘도 수행하였으나 CHAID 알고리즘과 매우 유사하여 결과를 별도로 제시하지는 않았다.

5.3 의사결정트리 분석의 종합적 해석

이번 절에서는 위의 세 알고리즘을 종합해보고자 한다. 각 노드에 포함된 속성은 조금씩 다르지만, 본 연구에서는 학업성취에 영향을 주는 중요한 변수를 찾는 데에 목적이 있기 때문에 ‘분류기준변수’의 관점으로 보아야 한다. 이 관점으로 본다면, 세 알고리즘에서 선택된 변수가 상당히 유사하다는 것을 알 수 있다. 세 알고리즘에 의해 분석된 결과가 다르다면 해당 데이터는 특정 몇 개의 변수들에 의해서는 명확하지 않게 분류된다고 할 수 있다. 즉, 종속변수가 특정 변수에 의해 받는 영향력이 적다고 할 수 있다. 본 연구에서는 세 알고리즘 모두 비슷한 변수를 선택하였기 때문에 선택된 변수가 종속변수에 큰 영향을 미치는 요인이라고 판단할 수 있다.

세 알고리즘 모두 첫 번째 분류기준변수로 ‘담당교수’를 선택하였고, 두 번째 분류기준변수로 대부분 ‘그룹’을 선택하였다. ‘담당교수’가 첫 번째 분기 변수로 선택된 것은 수업을 담당하는 교수의 역량에 의해 학생들의 성과가 달라질 수 있음을 의미한다.

이는 컴퓨터 교양과목의 경우 학생들의 전공에 따른 기본지식의 차이보다 교수의 역량이 중요함을 의미한다. ‘그룹’ 변수가 선택된 이유는 3장에서 설명한 것과 같이 B그룹의 시험시기가 중간고사 기간과 맞물려 B그룹 학생들의 성적이 저조하였고, 상대적으로 신학기를 맞아 의지가 충만한 A그룹의 학생들의 성적이 좋았기 때문이라고 판단된다. 시험의 시기가 학생들의 성과에 큰 영향을 준다는 것을 알 수 있다.

그 외에 ‘수강인원’, ‘강의시간’, ‘실습실’ 등이 선택되었다. ‘수강인원’의 경우 40명 정도의 소규모 분반이 50명 이상의 분반에 비해 대체로 높은 학습 성취도를 나타내는 것을 확인할 수 있었다. 이는 컴퓨터 교양과목의 경우 효과적인 교육을 위해서는 수강인원을 40명 이하로 운영하는 것이 효과적인 것으로 이해된다. ‘강의시간’의 경우 아침 9시, 저녁 7시 등 이른 수업이나 늦은 야간 수업의 경우 학생들의 참여도가 낮아 학업성취도가 떨어지는 것으로 분석되었다. ‘실습실’의 경우 모든 결정트리에서 발견되어 중요한 요인으로 파악된다. 실습실의 노후 정도, 프로젝트의 설치 개수, 교수와 학생 간의 간격 등의 환경적 요인이 컴퓨터 교양과목의 학업성취도에 영향을 주는 것으로 이해된다.

6. 결정트리 종합을 통한 요인 중요도 분석

6.1 다중 의사결정트리 종합의 필요성

의사결정트리 분석은 데이터마이닝의 대표적인 분석 기법으로서 특정 사건의 요인과 규칙을 발견하기 위한 목적으로 다양한 분야에서 사용되고 있다. 그러나 결정트리 분석방법의 특성을 제대로 이해하고 못하고 분석에 사용할 경우 잘못된 해석을 내릴 수 있는 여지가 많다. Selwyn은 결정트리 분석이 입력데이터의 특성에 영향을 받는 문제가 있음을 제시하고 있다[21]. 또한 박현경, 송문섭은 여러 요인들이 비슷한 변별력을 갖는 경우 분기 선택 방법에 따라 상이한 결과가 도출될 수 있는 문제를 제시하였다[6].

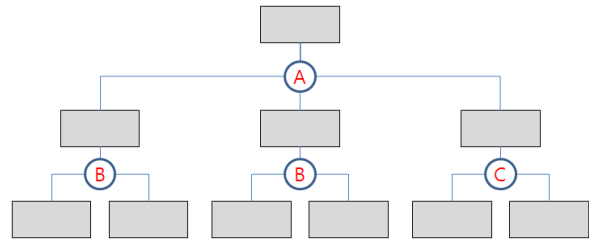
결정트리 분석이 갖는 첫 번째 한계는 중요한 요인이 결과 트리에 표시되지 않을 수 있다는 점이다. 실제로 영향력이 큰 중요한 요인이지만 산출된 트리 모형에는 표시되지 않을 수 있다. 이러한 문제는 변수들

간의 높은 상관성, 즉 다중공선성(Multicollinearity) 때문에 발생한다. 예를 들어, 두 변수 A, B가 종속변수 T에 높은 영향을 주는 경우에도 이 두 변수가 매우 높은 상관성을 갖는다면 이들 중 하나의 변수만 분기 변수로 선택되어 트리에 나타나게 되는 것이다.

결정트리가 갖는 두 번째 한계는 어떤 결정트리 알고리즘을 적용하는가에 따라서 상이한 결과가 도출될 수 있다는 점이다. 이는 주로 알고리즘이 분기 변수를 선택하는 평가 기준의 차이에서 발생한다. 가장 많이 사용되는 C5.0, CART 그리고 CHAID 알고리즘에 대해서 비교해보면, C5.0의 경우 Entropy Gain을 분기 기준으로 사용하며, CART의 경우 Gini Index와 분산의 차이를 사용하고, CHAID의 경우 t검정과 F검정을 사용하는 점에서 차이가 있다. 그 외에도 분기 방법에 따라 모형이 변경될 수 있다. CART는 하위노드로 분기할 때 단순화된 트리를 생성하기 위하여 항상 이진(Binary) 분기만을 사용하는 것이 특징이다. 또한 C5.0의 경우 상업화된 프로그램에 따라 범주형 변수의 분기 방식에 차이가 있다. 기본적으로 알고리즘에서는 모든 범주값에 따라 하위 노드를 분리하지만, 트리가 지나치게 커지고 복잡해지는 문제를 해결하기 위하여 유사성이 높은 범주값들을 묶어 소수의 노드로 분할하는 방법을 사용하기도 한다. 이러한 차이에 따라 기본적인 의사결정트리 알고리즘의 방식은 같다고 하더라도 결과가 상이할 수 있다. 본 연구에서는 이러한 문제를 해결하기 위한 하나의 방법으로 여러 알고리즘들을 수행한 후 그 결과들을 종합하여 해석함을 통해 변수의 중요성을 산출하는 방법을 다음 절에서 제시한다.

6.2 단일 의사결정트리에서의 요인 중요도 분석

의사결정트리 분석을 통해서 도출된 트리모형을 발견하는 데 목적을 둘 수도 있지만 트리모형으로부터 요인의 영향력의 정도를 측정하고 비교하기를 원하는 경우가 많다. 본 절에서는 하나의 결정트리 분석을 통해 얻어진 트리모형으로부터 변수의 중요도를 산출하는 수식을 제안한다.



(그림 6) 간단한 결정트리 예시

일반적으로 결정트리 분석으로 도출된 트리모형으로부터 요인의 중요성을 비교할 때 트리에서 변수가 사용된 레벨(깊이)이나 횟수를 고려한다. 예를 들어, (그림 6)에서 레벨 1의 분기점에서 사용된 A가 가장 중요한 요인이며, 그다음으로 B, C의 순서로 중요하다고 본다. 같은 레벨에 발견되었으나 표시 횟수를 고려하여 B가 C보다 2배 더 중요하다고 볼 수 있다. 그러나 트리가 더욱 복잡해질 경우 트리로부터 변수의 중요성을 직관적으로 판단하기는 어렵다.

본 연구에서는 결정트리로부터 변수의 중요도를 산출하는 수식으로 아래와 같이 식 (1)을 제안한다.

$$IMP_x = \frac{\sum_{All\ Branch\ b\ with\ x} \left(\frac{1}{BranchCount(b)} \right) \left(\frac{1}{2} \right)^{depth(b)-1}}{\sum_{n=1}^{depth} \left(\frac{1}{2} \right)^{n-1}} \quad (1)$$

식 (1)에서 IMP_x 는 변수 X의 중요도를 의미하며, 0에서 1사이의 값으로 계산된다. depth는 의사결정트리의 깊이이며, BranchCount(b)는 결정트리에서 분기점 b가 속한 레벨의 전체 분기점의 개수이다. 마지막으로 depth(b)는 분기 b가 해당된 레벨(깊이)를 의미한다. 식 (1)에 의해 산출된 모든 변수의 중요도의 합계는 1.0이 된다. 예를 들어, (그림 6)의 트리모형에서 변수 B의 중요도는 식 (2)에 의해 0.22로 계산된다.

$$IMP_B = \frac{\left(\frac{1}{3} \right) \left(\frac{1}{2} \right)^{2-1} + \left(\frac{1}{3} \right) \left(\frac{1}{2} \right)^{2-1}}{\left(\frac{1}{2} \right)^{1-1} + \left(\frac{1}{2} \right)^{2-1}} = 0.22 \quad (2)$$

6.3 다중 의사결정트리에서의 요인 중요도 분석

하나의 의사결정트리 알고리즘만을 적용하여 분석하는 경우 중요한 요인을 발견하지 못할 수 있다. 따라서 의사결정트리 분석을 수행하는 경우 여러 개의 의사결정트리 알고리즘을 적용하고 종합하면 하나의 알고리즘만 적용하는 경우보다 좀 더 신뢰할 수 있는 결과를 얻을 수 있다. 본 연구에서는 여러 개의 결정트리 모형에서 요인의 중요도를 계산하는 수식으로 식 (3)을 제안한다.

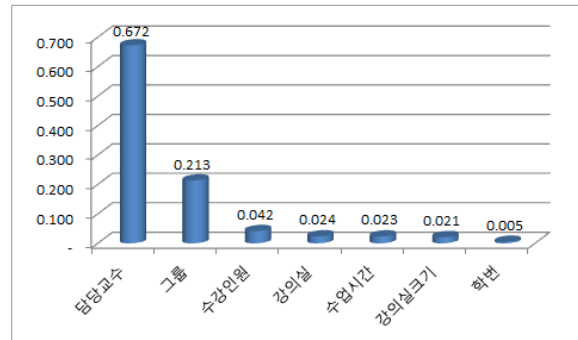
$$CIMP_x = \frac{1}{TreeCount} \sum_{all_used_tree} IMP_{t,x} \quad (3)$$

식 (3)에서 CIMP_x는 여러 알고리즘의 결과를 종합한 변수 X의 종합적 중요도를 의미한다. 그리고 TreeCount는 분석에 사용한 의사결정트리 알고리즘의 개수를 의미하고, IMP_{t,x}는 알고리즘 t를 사용한 분석 결과 트리를 식 (1)에 적용하여 계산된 변수 x의 중요도이다.

3장의 CHAID, CRT 및 QUEST 알고리즘을 적용한 실험 결과에 식 (3)을 적용하여 3개의 의사결정트리를 종합한 변수 중요도를 산출한 결과를 <표 13>에 정리하였다. ‘담당교수’가 0.672로 종합적으로 가장 중요한 변수이고, 다음으로 ‘그룹’이 0.213으로 중요한 변수로 계산되었다. 그 외에 그룹, ‘수강인원’, ‘강의실’, ‘수업시간’, ‘강의실크기’, ‘학번’ 순으로 중요한 변수로 계산되었다. 하나의 알고리즘만 적용할 경우 ‘강의실’, ‘학번’, ‘수업시간’, ‘강의실크기’ 요인들은 발견되지 않을 수도 있었지만 여러 개의 의사결정트리의 적용과 해석을 통해 충분한 요인들을 발견할 수 있었다. 제안한 수식을 통해 산출된 변수의 중요도를 <표 13> 및 (그림 7)에 제시하였다.

<표 13> 선택된 요인의 중요도 계산 수치

| 요인 | CHAID | CRT | QUEST | 종합 |
|-------|-------|-------|-------|-------|
| 담당교수 | 0.625 | 0.731 | 0.661 | 0.672 |
| 그룹 | 0.188 | 0.194 | 0.258 | 0.213 |
| 수강인원 | 0.063 | 0.046 | 0.016 | 0.042 |
| 강의실 | 0.063 | 0.008 | 0 | 0.024 |
| 수업시간 | 0 | 0.022 | 0.048 | 0.023 |
| 강의실크기 | 0.063 | 0 | 0 | 0.021 |
| 학번 | 0 | 0 | 0.016 | 0.005 |



(그림 7) 요인별 중요도 수치 차트

7. 결론 및 향후 연구

본 연구에서는 로지스틱 회귀분석과 데이터마이닝의 결정트리 분석을 통하여 실습을 중심으로 하는 컴퓨터 교양과목의 학업성취도에 중요한 영향을 미치는 요인을 탐색하였다. 로지스틱 회귀분석을 이용하여 시험의 합격/불합격 여부를 종속변수로 설정하고 이에 관련이 있는 여러 독립변수 중 전진단계변수선택 방법을 활용하여 교육성과에 유의한 영향을 주는 변수들을 발견하였다. 또한 CHAID, CRT 및 QUEST의 여러 의사결정트리 알고리즘을 수행하여 상위 단계에 있는 분류기준변수들을 통해 중요한 요인들을 발견하였다. 회귀분석에서는 ‘담당교수’, ‘수강인원’, ‘수업시간’, ‘그룹’(강의기간) 순으로 시험의 합격 여부에 영향을 준다는 결과를 얻을 수 있었고, 의사결정트리 분석에서는 ‘담당교수’, ‘그룹’, ‘수강인원’, ‘강의실’(실습실의 수업 환경), ‘수업시간’, ‘강의실크기’, ‘학번’ 등이 중요한 요인으로 선택되었다. ‘담당교수’, ‘그룹’, ‘수강인원’은 회귀분석과 의사결정트리 분석 모두에서 선택된 것으로 보아 학업성취도에 영향력이 큰 요인이라고 판단할 수 있다.

본 연구를 통하여 대학의 컴퓨터 실습 교양과목에서의 학업성취도에 영향을 미치는 요인들을 탐색하였다. 도출된 결과를 토대로 아래와 같이 교육환경 개선을 위한 방안들을 정리하였다.

- **그룹(강의기간)** - 현재 A/B/C그룹으로 5주 단위로 진행되는 방식의 경우 B그룹에 수업을 수강하는 학생들의 학업성취도가 저조한 문제가 있다. 또한

5주 수업이 심도있는 강의가 되기 위해서는 강의 기간이 부족하다는 평가가 있다. 따라서 수업을 8주 단위로 진행하여 A/B그룹의 형태로 운영하는 방안을 고려할 수 있다.

▪ **수강인원** - ‘수강인원’은 회귀분석에서 두 번째 중요한 요인으로 선택되었고, 의사결정트리 분석에서는 세 번째 중요한 요인으로 선택되었다. 컴퓨터 실습 교양수업에서 수강인원을 소규모로 운영하는 것이 효과적임을 보여준다. 현재 수강인원을 컴퓨터 설치대수에 따라 수강하도록 하여 한 분반에 50명 이상의 학생들이 수강하고 있는데, 효과적인 수업운영을 위하여 40명 이내로 제한하는 것을 고려할 수 있다.

▪ **강의시간** - 강의시간은 학업성취도에 중요한 요인으로 파악되었다. 특히 오전 9시 수업이나 오후 7시 수업의 경우 지각 및 결석이 많아 결과적으로 학업성취도가 낮아지는 것으로 판단된다. 오후 7시 강의는 개설하지 않는 것을 고려할 수 있다.

▪ **담당교수** - ‘담당교수’는 학업성취도의 가장 중요한 요인으로 분석되었다. 담당교수의 역량을 개발할 수 있도록 담당교수 워크숍을 개최하여 효과적인 강의 방법 및 강의 자료를 공유하는 것을 고려할 수 있다.

본 연구는 실습을 중심으로 하는 컴퓨터 교양교육의 학업성취도에 영향을 미치는 요인들을 탐색하고 몇 가지 학업성취도 향상을 위해 개선할 수 있는 점들을 도출하여 제안하였다. 향후 연구에서는 이를 실제 교육환경에 적용했을 때의 향상 효과에 대한 연구를 수행하고자 한다. 그 외에도 대학의 컴퓨터 실습 교양과목에 온라인 교육 또는 온라인과 오프라인을 혼합하는 블렌디드 교육 방식이 학업성취도에 미치는 영향에 대해서 연구를 진행하고자 한다.

참 고 문 헌

[1] 금지현, 이용환 (2008), 실과에서 학업성취도와 그 영향요인에 관한 구조방정식모형 분석, **한국실과교육학회지**, 21-2, 315-332.

[2] 김현철 (2010), 대학생의 학업성취(Ⅱ) : 학업성취도에 대한 새로운 예측변수의 탐색, **한국교육**, 32-2, 247-274.

[3] 김현철 (2012), 엔트로피 지수 분해를 통한 중학생 학업성취도 격차 요인의 연구, **한국교육논단**, 11-1, 1-23.

[4] 박동준, 최수영 (2009), 수학 학업성취도의 영향 요인 분석 연구, **한국수학교육학회**, 23-2, 383-398.

[5] 박소연, 이홍직 (2013), 청소년의 인터넷 중독에 영향을 미치는 요인에 관한 연구, **한국전자통신학회 논문지**, 8-2, 291-299.

[6] 박현경, 송문섭 (2003), **데이터 마이닝에서 변수중요도에 관한 연구**, 서울대학교 석사학위논문.

[7] 소연희 (2006), 효과적인 교실수업에 영향을 미치는 요인 탐색, **교육방법연구**, 18-1, 1-22.

[8] 소연희 (2009), 수업에 대한 자율성 지각과 다중지능 및 교과흥미가 학업성취에 미치는 영향, **한국아동교육학회**, 18-2, 5-18.

[9] 송찬욱, 유성모 (2011), 학업성취 관련 심리요인들과 가정 환경변인이 학업성취에 미치는 영향, **뇌교육연구**, 8-1, 27-53.

[10] 염명숙, 김광복 (2009), 사회과 사이버가정학습이 사회과 학업성취도와 학습태도에 미치는 영향, **정보교육학회논문지**, 13-4, 509-516.

[11] 오지영, 이수정(2008), 신경망을 이용한 초등학교 컴퓨터 활용 능력 예측, **정보교육학회논문지**, 12-3, 267-274.

[12] 이명숙, 이규민 (2009), 전문계 고등학생의 학업성취, 자아개념, 학교적응 및 학교부적응행동간의 관계, **교육과학연구**, 40-1, 167-193.

[13] 임은정, 김성진, 신현욱 (2013), 데이터마이닝 기법을 이용한 시각장애인의 취업결정요인 분석 연구, **장애와 고용**, 23-1, 273-302.

[14] 이주리 (2009), Data Mining을 이용한 초등학교 학생의 삶의 만족도에 대한 보호요인 및 위험요인 탐색, **아동학회지**, 30-1, 11-25.

[15] 이해주, 전의현 (2013), 데이터마이닝을 이용한 학업성취 결정요인 탐색, **아동교육**, 22-2, 5-18.

[16] 이해주, 전의현 (2012), 의사결정트리 기반의 분석을 통한 청소년의 컴퓨터 사용 유형별 관련

- 변수 추출, **컴퓨터교육학회논문지**, 15-2, 9-18.
- [17] 전병호 (2010), 대학교의 정보교육 만족과 성과의 영향변인에 관한 연구, **정보교육학회논문지**, 14-4, 477-488.
- [18] 주영주, 정애경, 이상철, 한애리 (2011), 전문대학 공학계열 컴퓨터 수업에서 학업만족도와 학업성취도 관련변연 연구, **전기학회논문지**, 60-2, 56-62.
- [19] 허인숙 (2002), 사회과 교육에서 사전지식을 고려한 학습과 개념도의 활용, **시민교육연구**, 34-2, 235-254.
- [20] Han, J. & Kamber, M. (2006), *Data Mining Concept and Techniques*, Morgan Kaufmann.
- [21] Selwyn Piramuthu (2008), Input Data for Decision Tree, *Expert System with Application*, 34-2, 1220-1226.

저 자 소 개

김 완 섭



2000년 2월 : 숭실대학교 컴퓨터 학부 (공학사)

2003년 8월 : 숭실대학교 컴퓨터 학과 (공학석사)

2006년 8월 : 숭실대학교 컴퓨터학과 인공지능 박사수료

2007년~현재 : 숭실대학교 베어드 학부대학 컴퓨터활용 교수

관심분야 : 컴퓨터 교양교육, 프로그래밍 교육, 인공지능, 데이터마이닝

e-mail : wskim92@ssu.ac.kr