**CSE2035 C프로그래밍**

(설계 프로젝트 : 데이터 분석 프로그램)

담당교수: 정성원

1. **설계 문제 및 목표**

이번 프로젝트에서는 학생들이 “데이터 분석 프로그램”을 제작해야 한다. 프로그램은 아래와 같은 요구 사항을 갖는다.

|  |
| --- |
| 기본구현   * 파일 입력 받기 (function 0: ReadData()) * 성별 기준 생존자 통계 (function 1: StatsBySex()) * 임금 기준 생존자 통계 (function 2: StatsByFair()) * 승객 등급 기준 통계 (function 3: StatsByPClass()) * 가족 관계 기준 통계 (function 4: StatsByFamily()) * 나이대별 기준 통계 (function 5: StatsByAge()) * 결측 데이터 분석 (function 6: FindMissingValue()) * 상관계수 행렬 계산 (function 7: CalCorrelationMatrix())   추가구현   * 기본 구현을 제외한 다른 통계 혹은 분석 데이터를 계산한다. |

이 프로그램의 제작 과정은 C프로그래밍 수업의 기본 과정을 따라야 하고, 공지한 형식(C파일, 문서)에 맞추어 제출해야 한다.

1. **요구사항**

2.1 합성

1. 1~6번을 쉽게 나누기 위해서 if else 대신 case 문을 사용.

calloc, malloc, realloc 이후 free 해주기.

csv 파일에서 구조체로 데이터를 읽어오기 위해서 FILE\* ,fscanf 등 이해 필요

구조체를 배열의 포인터로 생각하기 위해서 구조체에 대한 이차원 포인터를 정의, 동적 할당을 통해서 잡아주기

상관 계수, 최근접 이웃 알고리즘에 대한 추가 이해

2.2 분석

0번 함수의 입력에서 N(총 사람 수)를 알기 위해서 output 을 총 사람수로 제작 필요

나머지 함수에서는 printf의 구현이 중요, output은 void 형태로 제작

2.3 제작

전체적으로 while 과 if else, switch 을 사용해 반복할 수 있게 제작. If를 사용한 이유는 switch 안에서 break; 는 다른 역할을 할 것이라 생각해서 if 로 7인 경우에는 다르게 제작함.

0번 함수- Output 을 N으로 하며, 각 데이터를 fscanf 로 받는다.

1번 함수- 성별로 데이터를 나누기 위해서 Enum data 를 if 문에서 어떻게 사용하는지 파악, 모든 데이터에 대해서 for if 문을 사용해서 num, survive 로 구문하며 dead 는 num – survive 형태로 구현

2번 함수- 총 9가지의 나누기를 모두 int형으로 결정(이후에 int 형 배열을 통해서 접근했으면 더 편할 것 같다는 생각을 했다.) 이후에 switch 문을 통해 sibsp + parch 의 경우에 따라서 데이터를 나눈 후 이후에는 1번 함수와 비슷하게 진행

3번 함수- 나이에 따라서 정하는 함수, 2번 함수에서 생각했던 배열을 통해 접근을 구현함. 나누는 수에 따른 크기의 배열을 malloc을 통해서 할당해준 이후, sizediv라는 나누는 수에 따라서 나이를 나눈 후, 1번 함수와 비슷하게 구현

4번 함수- 상관관계를 구하는 함수, 총 7 개를 비교하기 위해서, cor matrix 를 7\*7 배열의 형태로 제작, 이후에 평균값(x bar, y bar)을 matrix\_mean 에 저장, pow sqrt 함수를 활용하여 상관계수를 구하고, corr\_matrix에 저장, print

5번 함수- KNN에 대해서 구하는 함수, 각 항목의 max 와 min 값을 찾으며, 배열에 저장, P Q 에 각 학습 데이터, 테스트 데이터를 넣는다.(구조체의 원소로 넣은 이유는 \*Q=\*dataset 이라 하면 dataset 자체의 값이 변경될까봐 이렇게 구현) 이후 double 형태로 변환(변환하지 않았을 때 이상한 값이 도출), 계산. 이후 거리에 따라서 값을 bubble sort로 정렬, 정렬한 값에서 가장 짧은 거리가 모두 0.00000 형태가 나와서 정렬값의 두 번째 값인 data[1]부터 접근했더니 더 납득할 만한 결과가 나옴. 두 번째 값부터 계산함(이에 대한 원인은 찾아내지 못함)

6번 함수- 테스트와 학습을 나누는 기준을 테스트를(r+1)\*N/nFolds-(r)\*N/nFolds 의 크기로 잡은 이후 진행 이후의 과정은 5번과 동일함.

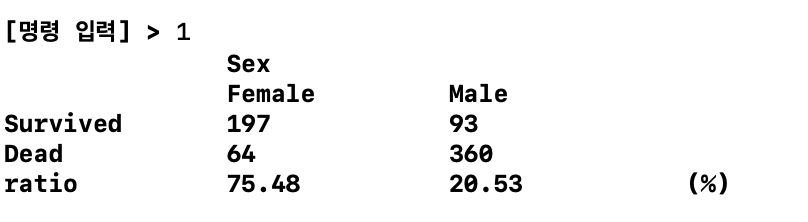
2.4 시험

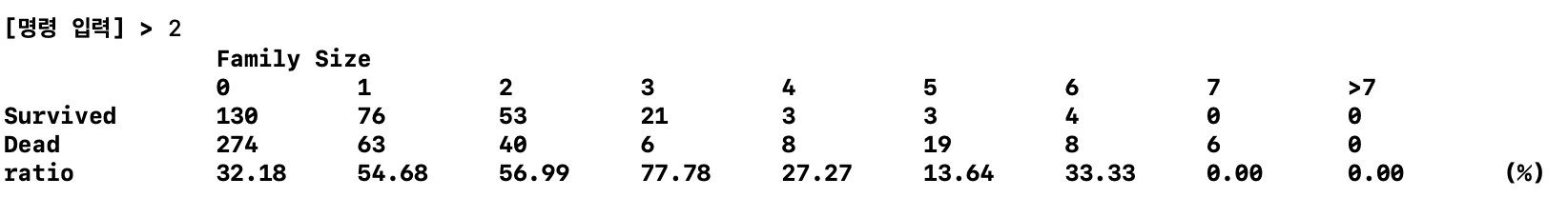
1, 2, 3, 4 번의 경우에 대해서는 잘 돌아감. 3번의 경우에 대해서 sizediv로 먼저 나눈 이후 i를 곱하는 행위에 대해서는 약간 애매하다는 것을 느끼고 이를 먼저 곱하고 나누는 형식으로 변경함

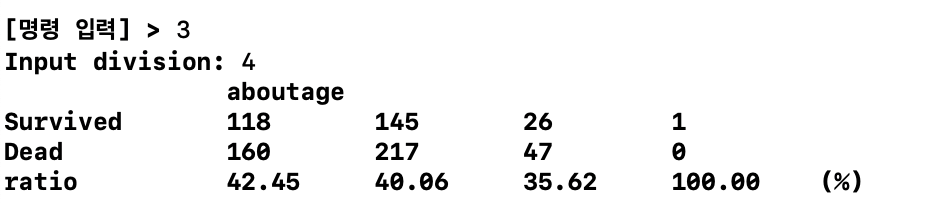
5번의 경우에서는 위에서 설명한 문제처럼 data[1]부터 접근하는 방식을 택했는데, 이에 대한 원인을 발견하지는 못했음.

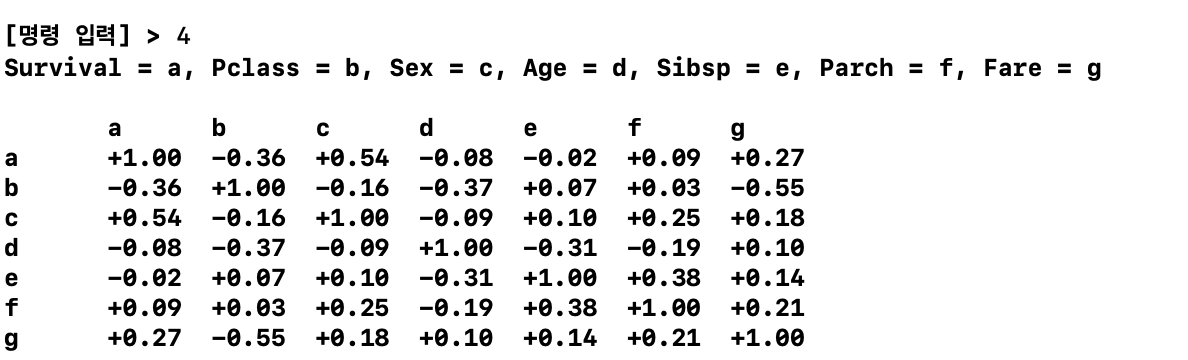
2.5 평가

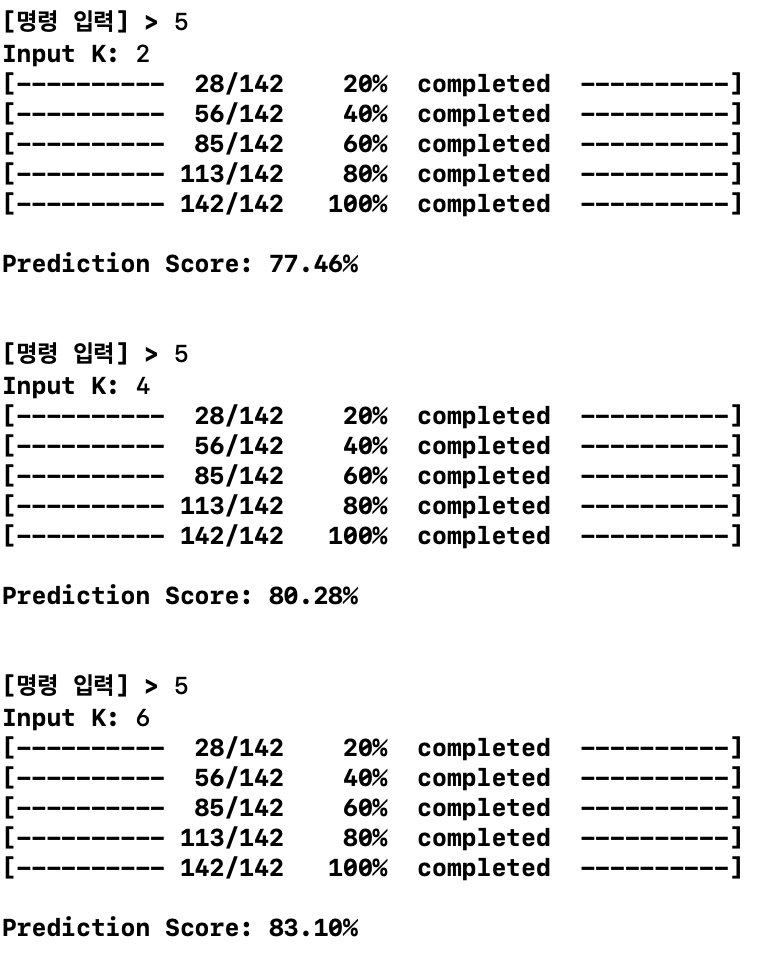
5번을 제작하면서 별로 효율적인 함수라고 생각하지 못함. 6번에서 다시 적용하기 위해서 함수를 부를 수도 있는데, 5번이 효율적이지 못해서 그 방법을 택하지는 못함. 아래는 1번부터 6번까지의 결과임.

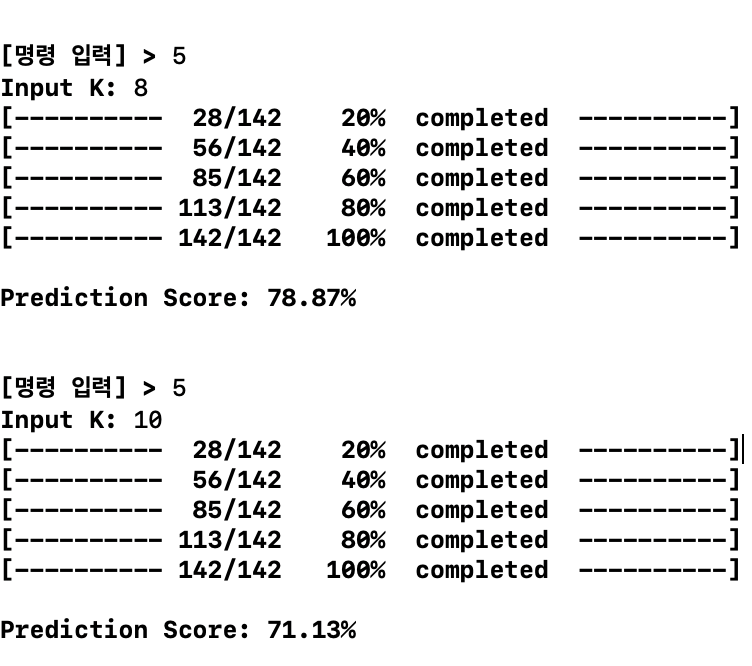


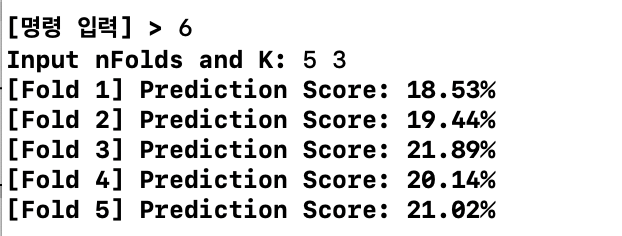












2.6 보건 및 안정

제약 사항에 대해서 구조체 사용, 동적할당 사용, 전역변수 사용 금지 등의 제한 조건은 지킴. 하지만 cspro 서버에서 코딩을 하는데에 한계를 느껴 xcode 를 통해서 코딩을 진행, 1, 2, 3, 4번에 대해서는 개인적으로 괜찮다고 평가. 하지만 0번 함수에서 파라미터로 \*\*형태를 받지 못한 점과 5, 6 번 함수에서 재사용성이 현저히 낮다는 점에서 개인적으로 아쉬움.

2.7 생산성과 내구성

최대한 다른 데이터에 대해서도 적용할 수 있도록 구성함. 하지만 구조체의 형태는 동일해야 한다는 문제가 있음 (상관계수를 구할 때, KNN에서 데이터 7개, 6개를 비교한다는 점에서 형태가 다르다면 이를 적용하지 못하는 프로그램) 또한 무한히 반복해도 한번 받은 dataset의 값은 변경하지 않도록 코드를 작성함. 이에 대해서 문제가 발생하지는 않을 것임을 예측 가능.

1. **기 타**

3.1 환경 구성

2.6에서 말한 것처럼 linux 서버가 아닌 mac 의 xcode를 통해서 구현했다는 차이점이 존재. 하지만 linux 서버에서도 동일하게 작동함을 확인 한 이후 시행함.

3.2 참고 사항

5번에서 거리 data 에 접근할 때 배열의 1번(두 번째 원소)부터 접근했다는 점에서 약간의 오류가 있음. 또한 이에 따라 6번의 값에도 문제가 있을 수 있음을 인지.

6번에서 데이터 값이 이상하게 나옴. 이것도 원인을 파악하지는 못함. 또한 정확하게 나눠지는 결과에서도 프로그램이 바로 끝나는 결과가 나타남.

3.3 팀 구성

20201635 전찬

3.4 수행기간

2020/12/06 (금요일) 23:59:59 까지