안녕하십니까, 심장병의 원인 분석 이란 주제로 발표를 하게 된 201조의 000, 000 입니다.

시작하기에 앞서, 저희가 사용한 데이터의 출처에 대해 말씀드리겠습니다.

저희는 Keggle에서 데이터를 가져왔습니다. 데이터의 Raw데이터를 가져왔습니다.

데이터는 심장병의 여러 요인들에 대한 값과, 심장병의 위험군을 포함한 csv파일입니다.

해당 데이터에 대한 설명을 하겠습니다.

우선 age는 대상의 연령을 나타냅니다. Sex 성별, cp는 흉곽 통증의 유형이며, 0 = 전형적인 협심증, 1 = 비정형 협심증, 2 = 비각통, 3 = 무증상을 의미합니다. Trestbps는 휴식 상태 혈압, chol: 혈청 콜레스테롤 수치, fbs: 공복 혈당이 120 mg/dl 보다 큰지. restecg: 휴식 심전도 결과를 나타내며, 0 = 정상, 1 = ST-T파 이상, 2 = 좌심실 비대를 나타냅니다. 이때 ST-T파는 그림에서 알 수 있듯이, 보라색~T에 해당합니다. Thalach는 최대 심박수, exang는 운동으로 인한 협심증 유무, oldpeak는 운동으로 유도된 ST의 감소입니다. Slope는 최고 운동 ST 세그먼트의 기울기이며, 0 = 상향 기울기, 1 = 평평함, 2 = 하향 기울기를 의미합니다. ca값은 형광 투시를 통해 색이 칠해진 주요 혈관 수의 값이며 0, 1, 2, 3이 있습니다. thal: 지중해 빈혈 유형(유전병), 즉 헤모글로빈 결로 인한 빈혈의 유무이며, 1 = 정상, 2 = 고정 결함, 3 = 가역적 결함을 의미합니다.

마지막으로 target은 반응 변수이며, 1 = 심장병 고위험군, 0 = 심장병 저위험군을 의미합니다.

여기서 저희 먼저 설정한 설명 변수는 age(나이), cp(흉통 유형), thalach(최대 심박수), oldpeak(운동으로 유도된 ST 감소), exang(운동으로 인한 협심증), slope(최고 운동 ST 세그먼트의 기울기), ca(형광 투시를 통해 색이 칠해진 주요 혈관 수)입니다.

각 변수들은 심장병과의 연관이 있을 것으로 보이며 몇몇 변수의 경우 검사를 통해 얻을 수 있는 값이기에 모델의 변별력을 높여 줄 수 있을 것으로 보입니다.

그럼 다음으로 영가설과 대립가설을 설정하겠습니다.

영가설은 저희가 방금 말한 변수들이 심장 마비의 발생과는 관련이 없다. 이며, 대립가설은 해당 변수들이 심장 마비의 발생에 유의미한 영향을 끼친다는 것 입니다.

또한, 해당 내용은 심장병, 즉 사람의 생명과 관련된 내용이므로 유의수준을 5%가 아닌 3%로 설정하였습니다.

( => 오버피팅 문제 어케 해결할거냐 라고 나오면 k-fold든 random forest든… 말하기)

다음은 데이터 가공 계획입니다. 우선 결측값은 제거하겠습니다. 이상값의 경우 이상값의 개수와 수치를 판단 한 뒤에 제거 or 살리는 식으로 결정하겠습니다. 그 뒤 연속형 데이터에 대한 정규화를 진행합니다. 그리고 상관관계를 분석해 설명 변수의 조정을 할 것입니다.

학습시에 Logistic, RandomForest, SVM 사용할 예정이며, Logistic과 RandomForest의 경우 저희가 설정한 설명 변수 모델과 전체 변수를 사용하고 backward와 Purning을 한 모델과의 비교를 통해 어느정도의 성능을 내는지를 Confusion matrix와 함께 비교해 볼 것입니다.

해당 모델로 예측하고자 하는 것은 심장병 고위험군 분류 입니다. 건강검진 등에서 설명 변수를 값을 얻기 위한 검사를 진행 후 의사가 모델의 결과값을 보고 고위험군으로 분류된 환자에게는 설명과 함께 적절한 처방을 내려주도록 합니다.

모델의 설명 변수에서 특정값이 매우 높게 나타난 저위험군 환자에게도 해당 변수에 대한 설명과 해당 변수가 일으킬 수 있는 각종 병에 대한 사전 고지와 적절한 처방을 내릴 수 있을 것입니다.