추진 배경

학회의 졍규 세션에서 매주 다른 알고리즘에 대해 학습하였는데, 이에 대한 이해도를 높이고자 직접 데이터를 찾아 실습하고자 하였다. 실습용 데이터로는 당뇨병 예측 데이터를 선택하였는데, 이는 세션 중 팀 과제였던 ‘논문 분석’의 영향을 받았다. 논문 분석은 데이터 분석과 관련된 논문을 선정하여 이를 읽고 이해한 뒤 해당 내용을 다른 학회원 모두에게 소개하는 방식이었는데, 여기서 본 팀이 분석한 논문이 심근경색증/협심증 예측 데이터였다. 당뇨병 데이터는 이와 같은 의학 데이터인 데다, 분류 알고리즘을 통해 특정 질환과 여러 x변수의 관계를 분석하는 형태에서도 동일하였다. 이에 당뇨병 데이터를 최종 데이터로 선정하여 실습해 보기로 하였다.

목표 및 내용

캡스톤 프로젝트에 있어 본 팀의 주된 목표는 알고리즘에 대한 이해도 향상이었다. 학습한 알고리즘 중 특히 분류 알고리즘에 초점을 맞추었다. 알고리즘 실습을 위해 구한 해당 데이터는 적절히 전처리를 거쳐야 하였으므로, 다양한 전처리 방법을 수행해 본 뒤 각각의 전처리 방법에 따라 결과가 어떻게 변하는지 확인해 보고 최적의 전처리 결과를 선정하였다. 이후

수행결과 및 기대효과

가장 성능이 높게 나타난 알고리즘은 최적의 파라미터를 대입한 Random Forest였다. 이를 바탕으로 당뇨병 발병에 가장 높은 영향을 미치는 요인을 분석해 본 결과, HbA1c\_level이 가장 큰 값을 가졌고, blood\_glucose\_level이 0.2621로 두 번째를 차지하였다. 이후 age, bmi, hypertension, heart\_disease, Male, Female이 차례로 뒤를 이었다. 이 결과를 적극적으로 이용한다면, 당뇨병의 예방 및 치료에 큰 도움을 줄 수 있을 것으로 기대된다.

후기

학회를 통해 데이터 분석이라는 분야를 처음 접해 보았기 때문에, 정말 처음부터 끝까지 순탄했던 기억이 없다. 단순히 알고리즘을 적용해 보고 그에 대한 결과를 확인해 볼 목적으로 시작한 실습이었음에도 불구하고 알고리즘을 적용하기 전 데이터를 전처리하는 과정에 ½ 이상의 시간을 투자한 것이나, 결과를 추출하기 위한 알고리즘 정확도의 지표를 잘못 선정하여 고친 것이 가장 기억에 남는 난관이었다. 그러나 부족한 부분을 보완하고자 팀원과의 교류를 활성화시킨 결과, 시간이 지날수록 보고서가 점차 완성된 형태를 갖추는 것이 뿌듯했다. 시간상의 부족이 있어 아쉬움이 남지만, 추후 활동에 대한 밑바탕이 되었음은 확신한다.