인공지능 기반 혈당 수치 예측

**1. 프로젝트 개발 배경**

IoT, 웨어러블 기기 등의 보급과 함께 헬스케어 및 의료 분야에서도 IT 기술을 접목하려는 시도가 늘어나고 있습니다. 특히 지속적인 관리가 필요한 만성질환의 경우, 단순히 처방이나 수술만으로 효과를 보기 어렵기 때문에 치료와 병행해 IT 기기를 통한 관리를 하는 것이 환자의 상태를 개선하는데 큰 도움이 될 수 있다는 점에서 각광을 받고 있습니다.

여러 질환 중 일상생활에서의 관리가 예후에 큰 영향을 주는 질병으로 당뇨병이 있습니다. 처방된 약을 꼬박꼬박 먹는 것뿐만 아니라, 일상생활에서의 식습관, 생활패턴까지 신경을 써야 하기 때문입니다.

중증 당뇨 환자를 위해 자동으로 주기마다 인슐린을 주입하는 기기나, 혈당 측정기도 시중에 나와있지만, 매우 불편하기 때문에 대다수의 환자에게는 이러한 기기를 적용하지 않고 있습니다. 그 외의 경우는 기기를 통한 투약, 측정 대신 경험이나 관행에 의한 처방 등이 이뤄지고 있습니다.

경험, 관행은 결국 전문가인 의료진의 의사결정과 피드백이 장기간에 걸쳐 형성된 것이므로, 완전히 효과나 근거가 없는 것은 아닙니다. 그러나 이러한 경험이 충분한 전문의가 아닌 경우나, 환자 본인의 경우는 경험의 부족으로 인해 잘못된 판단을 내릴 수도 있습니다

**2. 프로젝트 목적**

인공지능을 활용하여 혈당의 정확한 예측을 통해 오진율을 줄이고 보다 정확한 처방을 내릴 수 있도록 도움을 주는 것을 목적으로 하고 있습니다.

**3. 프로젝트 내용**

3-1. 데이터

의료 데이터의 경우 각종 규제, 절차를 비롯해 민감한 개인정보이므로 쉽게 확보하기 어려웠으나, 다행히 당뇨병에 인공지능 관련 기술을 접목하고자 하는 연구를 진행하시던 교수님의 도움으로 제한된 조건하에 데이터를 이용할 수 있었습니다.

해당 데이터는 입원 환자의 차트 기록 및 검진 데이터였는데, 차트의 경우 입원기간 동안 매일 기상, 아침, 점심, 저녁, 수면 시간대의 혈당 측정 기록과 투약내용, 검진 데이터는 나이, 키, 체중, 혈압, 공복혈당, 당화혈색소 등의 신체 정보가 포함됐습니다.

대부분의 데이터는 이미 csv 파일에 숫자로 저장되어 가공이 쉬웠으나, 투약내용의 경우 비정형 텍스트 데이터이기 때문에, 마이닝이 필요했습니다.

가령 인슐린 제재의 경우 처방에 따라 4U (단위), 8U 등 단위가 다르고, 또한 속효성의 경우 투약 즉시, 지효성의 경우 종류에 따라 6시간, 8시간 등 장기간에 걸쳐 효과가 나타나는 등 차이가 있습니다. 따라서 텍스트를 가공해 투약된 제재와 양을 파악하고, 이를 의약 데이터와 연계하여 실효 인슐린 투여량을 추정할 필요가 있었습니다.



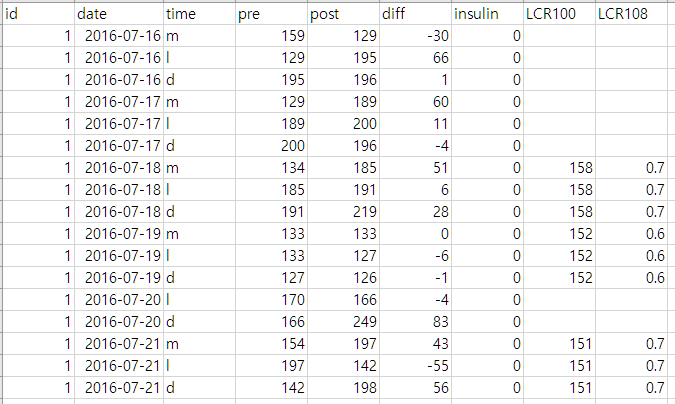
[사진1] 혈당 측정 데이터 및 투약 데이터 예시

위의 예시에서 DH-LTP 22U의 경우, DH-LTP라는 제재를 22단위 투약했다는 의미이며, DH-AOTP 6-8-8의 경우, DH-AOTP라는 제재를 아침에 6단위, 점심 8단위, 저녁 8단위 투약했다는 의미입니다.

또한 단위의 기재에서도, 위처럼 U를 붙이거나, 생략하는 경우도 있고 동시에 DH-HUR4단위처럼 단위를 붙이는 경우도 있어 패턴이 명확하지 않습니다.

3-2. 데이터 전처리

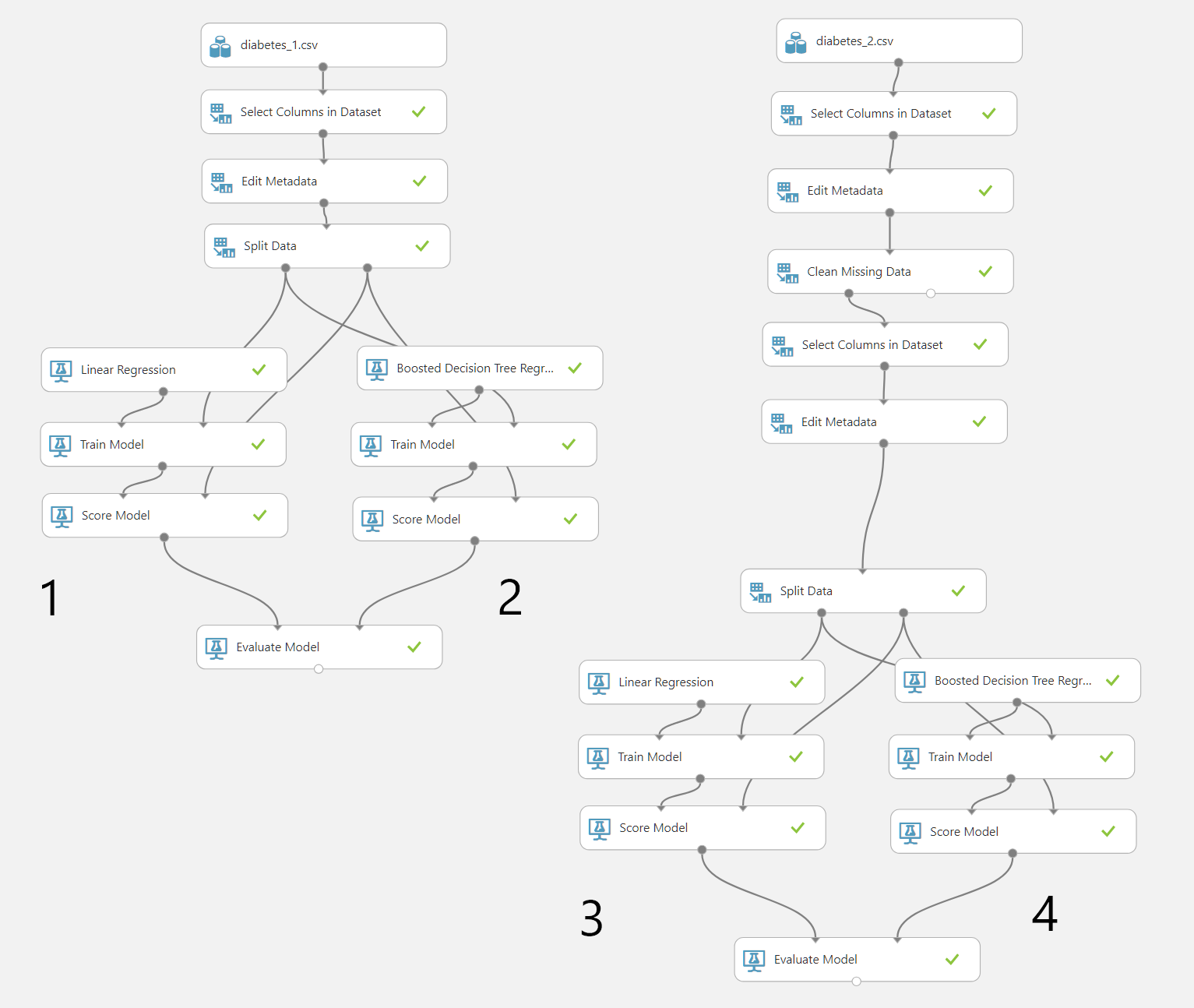
데이터의 양을 고려할 때 수작업으로 전처리를 하는 것은 적절하지 않았기 때문에, 정규표현식을 비롯한 텍스트 처리 기술을 활용하여 토큰화 등을 진행하고, 잘못 추출된 케이스를 보완할 수 있는 규칙을 추가하며 개선하는 방식으로 전처리 알고리즘을 설계했습니다.



[그림2] 전처리 이후 데이터 형식

3-3. 모델

모델은 post, 즉 다음 시점의 혈당 수치를 예측하며 입력으로 현재 혈당 수치, 유효 인슐린, BMI, 공복혈당을 활용하는 것이 목표이므로, 일반적인 경우 Regression 모델을 활용합니다. 그러나 데이터를 제공한 교수님이 제안한 의견으로 실제 의료진의 입장에서는 특정한 환자 개개인을 예측하는 것 이전에, 이 환자가 어떤 특성의 집단에 속하는지 분류하고, 이후 집단의 특성을 고려해 혈당을 예측하기 때문에 Classification이 포함되면 의료진이 더 이해하기 쉽다는 것이 있어, 이를 반영해 분류-회귀로 구성된 2중 모델을 설계했습니다.



[그림3] Azure Machine Learning Studio 상의 모델 구조도

실제 머신러닝은 클라우드 플랫폼인 Azure의 Machine Learning Studio를 활용해 설계, 학습 등을 진행했습니다. 시행착오 끝에 분류는 총 7개 그룹(카테고리)로 환자를 구분하고, 각 그룹마다 Regression 모델을 학습시켜 총 7개의 회귀 모형으로 환자를 예측하는 것으로 구성했습니다.

3-4. 주요성과

분류 이후 각 Regression 모델의 정확도 등을 살펴보면, R제곱값이 0.3이하로 실질적인 예측의 정확도는 낮게 나왔습니다. 머신러닝 모델의 구조나 최적화의 문제도 분명 있을 것이나, 근본적으로 데이터의 부족 및 오류도 원인이 됐습니다. 가령 혈당의 경우 인슐린 제재의 투여량 외에도 식사량이나 식단, 영양소 구성에 따라 크게 변동이 있고, 특히 차트에 간식 등이 기록된 경우도 있으나 간식, 식단 정보가 누락된 것이 데이터의 대다수여서 이를 반영할 수 없다는 한계가 있었습니다. 외국의 유사한 당뇨병 머신러닝 모델 연구 사례에서는, 17명의 환자에 대해 식단을 비롯한 운동, 활동량 등의 세세한 정보를 측정하고 환경을 통제하여 높은 정확도를 달성하기도 했는데, 이처럼 정확하고 통제된 데이터가 아닌, 누락과 오류가 많은 데이터를 기반으로 모형을 설계, 학습한다는 것이 어려운 점이었습니다.

실제 의료현장에서도 의료진이 환자의 매일매일의 혈당을 예측하는 일은 어렵고, 특히 이번 데이터처럼 입원 환자의 경우에는 외진 환자에 비해 저혈당, 고혈당 상황에서 바로 개입이 가능하기 때문에 오히려 관리가 다소 느슨해져서 혈당 변동폭이 커지는 경우도 있다는 의견을 들었습니다.

3-5. 결론 (발전방향)

의료 분야에 머신러닝, 딥러닝을 도입하는 시도가 늘어나고 있지만, 이와 같이 데이터 확보의 어려움, 확보한 데이터도 가공이 어렵거나 누락된 정보 등이 많고, 실제 현장에서의 상황 등에 따라 예측이 어려워지는 등 예상하지 못한 어려움에 봉착하는 경우도 많다고 합니다.

이번 프로젝트는 정확도의 측면에서는 다소 아쉬움이 있으나, 실제 의료 현장의 데이터를 가공하고 활용한 경험과 전문가인 의료진의 의견을 듣고 협업한 경험이 큰 도움이 될 것이라 봅니다.

**4. 주요 사용기술**

- 언어 : Python

- 모델 : 자연어처리(NLP), 머신러닝(Regression, Classification)  
- 프레임워크 : Azure Machine Learning Studio

**5. 주요 소스코드**

- 투약 데이터 텍스트 처리 관련 코드



- 데이터를 가공해 시계열 예측을 수행하는 모델에 적합한 형태로 처리하는 코드

