살려야한다

응급실 → 중환자실

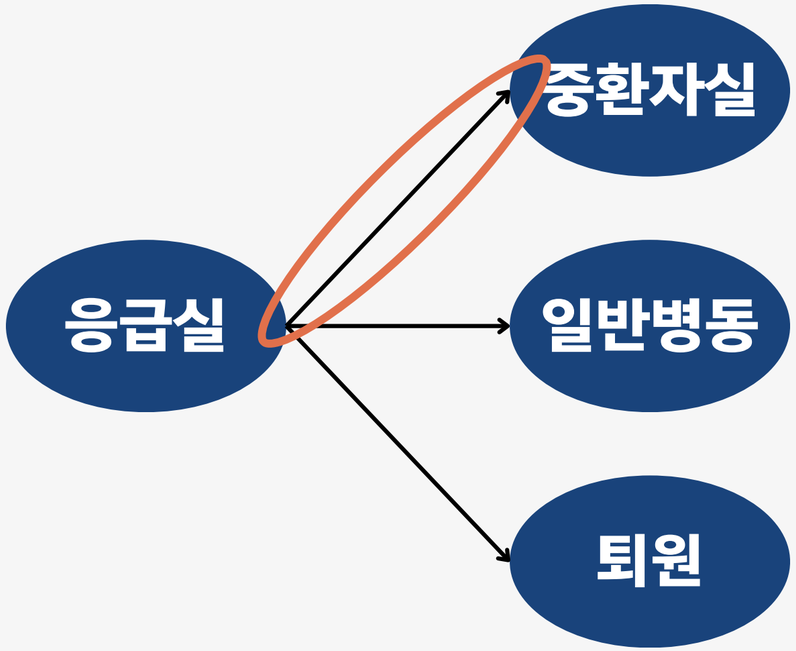
환자의 심정지 예측 & 요인 분석

6조 | 박찬혁 부형진 정수진 정필규

목차

주제 선정

질병 프로세스를 이용한 최적의 의료 처치 제안



응급실 입원 환자 중 가장 빠른 처치가 필요한 환자는 중환자실 입원 예정자

최적의 기준이 모호

자주 나오는 질병 카테고리 기준으로 의료 처치 확인

케이스 중복 / 개수 문제

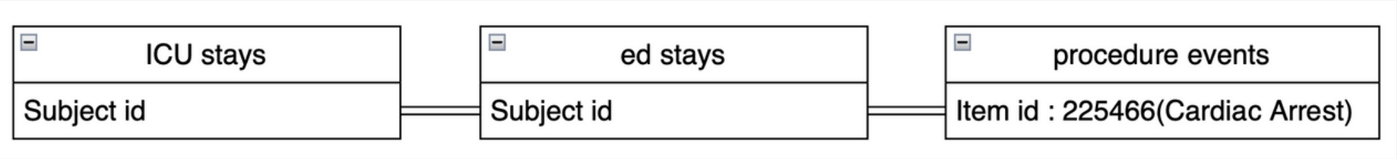
응급실에서 중환자실로 이동한 환자

심정지(=가장 위급한 상황) **예측 & 요인 분석**

전처리

1) 응급실&중환자실 환자 중 심정지를 일으킨 환자 수 파악

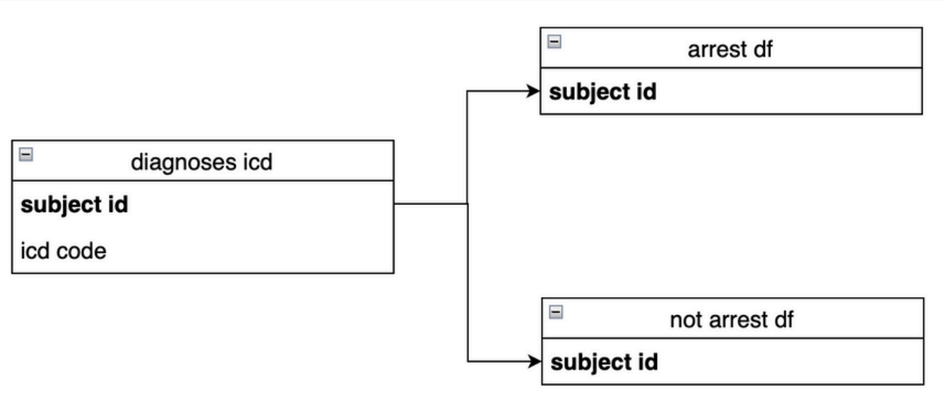
ICU stays & ed stays : Subject id 기준으로 병합한 후 중복을 제거
이들 중 **심정지(itemid=225466)**를 일으킨 환자를 파악



2) 심정지 환자와 비슷한 질병 코드를 가진 환자 파악

(심정지 일어나지 않은 환자 중 심정지 환자와 비슷한 조건 확보)

icd code를 기준으로 비슷한 질병 코드를 가진 사람 파악
= **자카르 유사도**를 기준으로 0.4 이상인 경우



109명 심정지가 온 집단

694명 심정지 환자와 비슷한 질병 코드를 가진 집단

심정지 환자 중 사망한 사람의 수: 86

전처리

3) 환자 데이터에 vital sign 32개를 연결

d items 파일에서 심정지와 관련된 핵심 vital sign 32개의 특성을 추출
핵심 vital sign을 기존 환자 데이터에 연결

4) 심정지 발생 시점 추가

charttime & starttime 기준

stay_id & charttime이 일치하는 경우 cardiac_arrest = 1
없을 경우, 새로운 행을 생성하여 starttime 시점에 심정지가 발생한 것으로 추가

핵심 vital signs
→ 일부 vital sign 사용(결측치 확인 & mean 제거)

1. Arterial Blood Pressure diastolic
2. Arterial Blood Pressure systolic
3. Heart Rate
4. Non Invasive Blood Pressure diastolic
5. Non Invasive Blood Pressure systolic
6. O2 saturation pulseoxymetry
7. Respiratory Rate
8. Temperature Celsius

전처리

5) 시계열 데이터셋 변환 & 스케일링

데이터셋의 문제점

1. 시간별로 측정한 값들이 서로 다르고 규칙성이 존재하지 않음
2. 해당 환자가 아예 측정하지 않은 feature 값도 존재함
3. 심정지 환자의 수가 매우 적어 데이터 불균형 문제 존재

1. stay_id 기준으로 각 feature가 90% 이상 결측인 경우 전체를 결측치 처리,
   나머지 결측치에 대해서는 선형 보간법, 평균 대치법을 사용해 결측치 처리.
   -> 결측치 전체를 -1로 mapping
2. 딥러닝 모델 활용을 위한 **MinMaxScaler** 스케일링
3. 오버샘플링으로 **SMOTE** -> 데이터 불균형 문제 해소

각 feature의 결측치 비율

Arterial Blood Pressure diastolic	0.647391	0.773428
Arterial Blood Pressure systolic	0.647298	0.773404
Heart Rate	0.316938	0.290137
Non Invasive Blood Pressure diastolic	0.697356	0.545089
Non Invasive Blood Pressure systolic	0.697231	0.544968
O2 saturation pulseoxymetry	0.344468	0.320273
Respiratory Rate	0.319299	0.298891
Temperature Celsius	0.934841	0.964294

stay_id 기준 전체 측정치가 결측값인 경우

Arterial Blood Pressure diastolic	472
Arterial Blood Pressure systolic	472
Non Invasive Blood Pressure diastolic	16
Non Invasive Blood Pressure systolic	16
O2 saturation pulseoxymetry	6
Temperature Celsius	602

사용 모델

1) Attention 기반 LSTM 모델

입력 데이터

- 시계열 데이터 (time_steps, num_features) 형태
- time_steps: 시간 차원의 고정 길이
- num_features: Feature 수

-1로 작성한 미측정값 결측치는 마스킹 → 해당 값을 모델에서 사용하지 않도록 처리

LSTM 레이어 : 시간 축의 상관관계를 학습하여 시계열 데이터의 패턴을 추출, return_sequences=True로 모든 시간 스텝의 출력을 반환

Softmax를 사용해 각 시간 스텝의 Attention 가중치를 계산
→ 각 feature의 측정 시간 별 중요도 확인

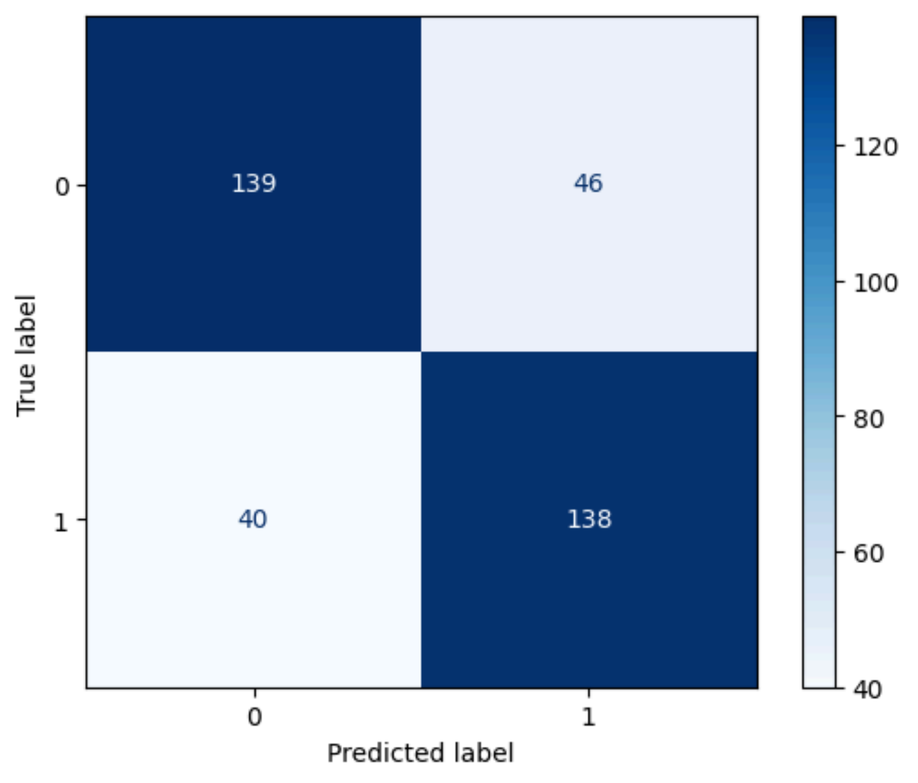
loss function : binary_crossentropy
Optimizer : adam

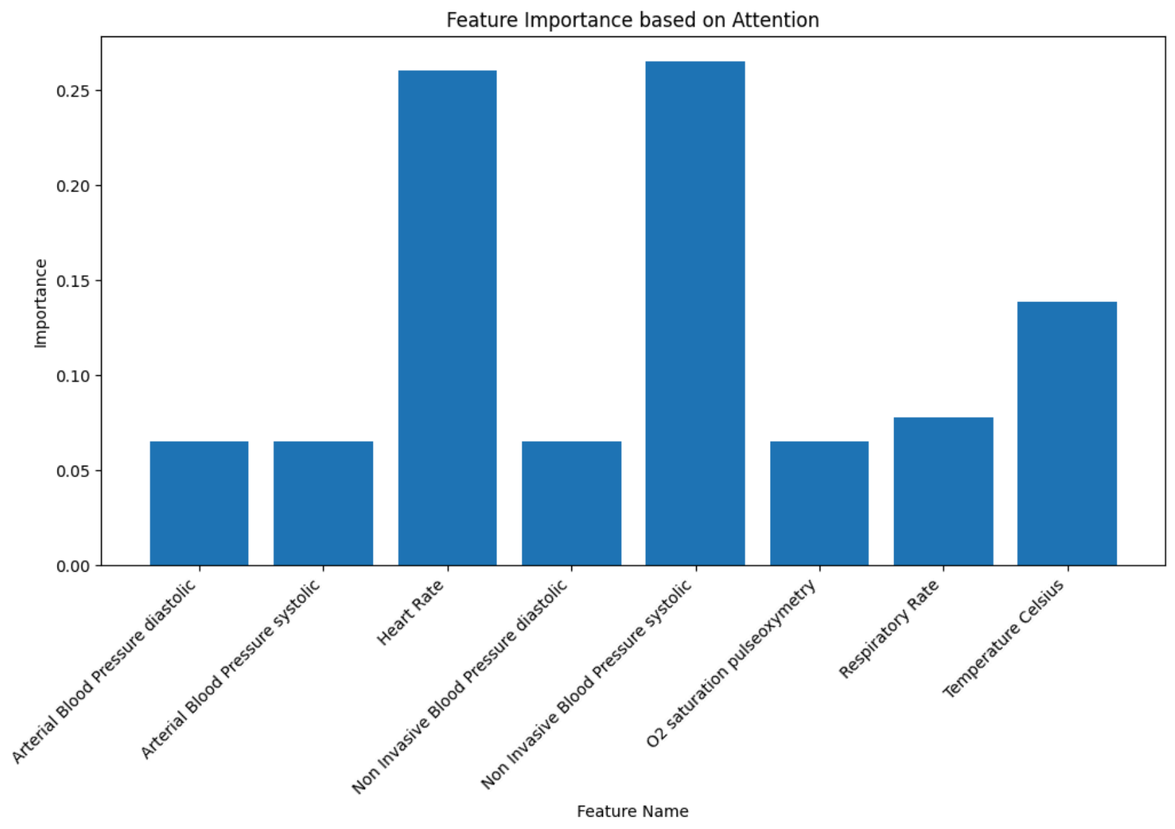
모델 적용

1) 심정지 발생 1시간 전 데이터 필터링

필터링된 데이터에서 stay_id별 행 개수의 평균값(**5**개)으로 개수 통일

Accuracy (정확도): 0.76
Sensitivity (민감도, Recall): 0.78
F1 Score: 0.76



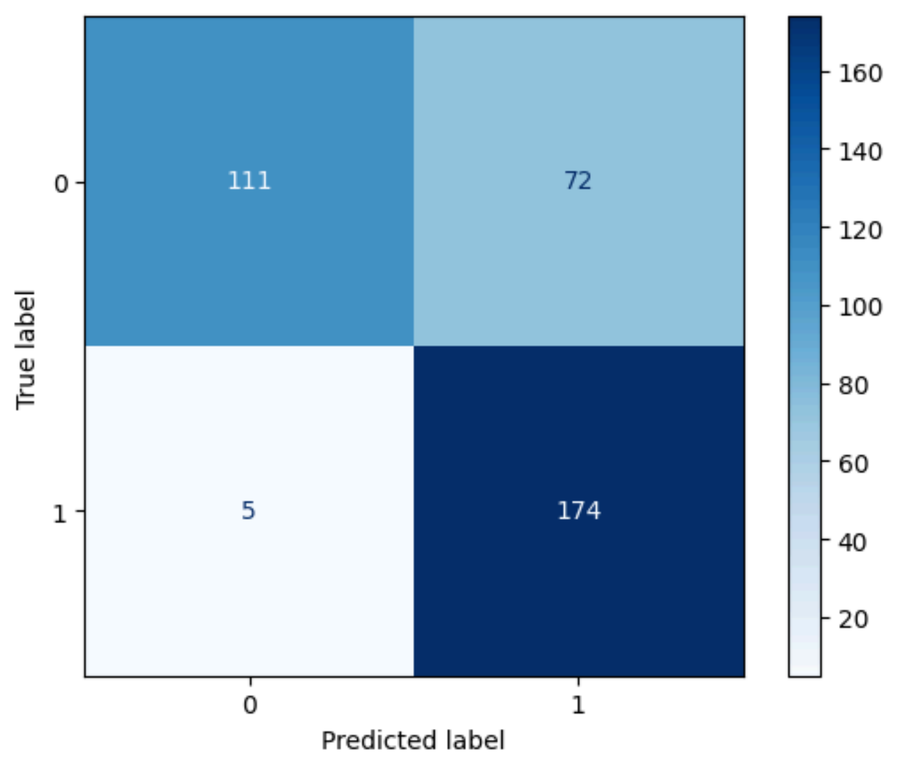


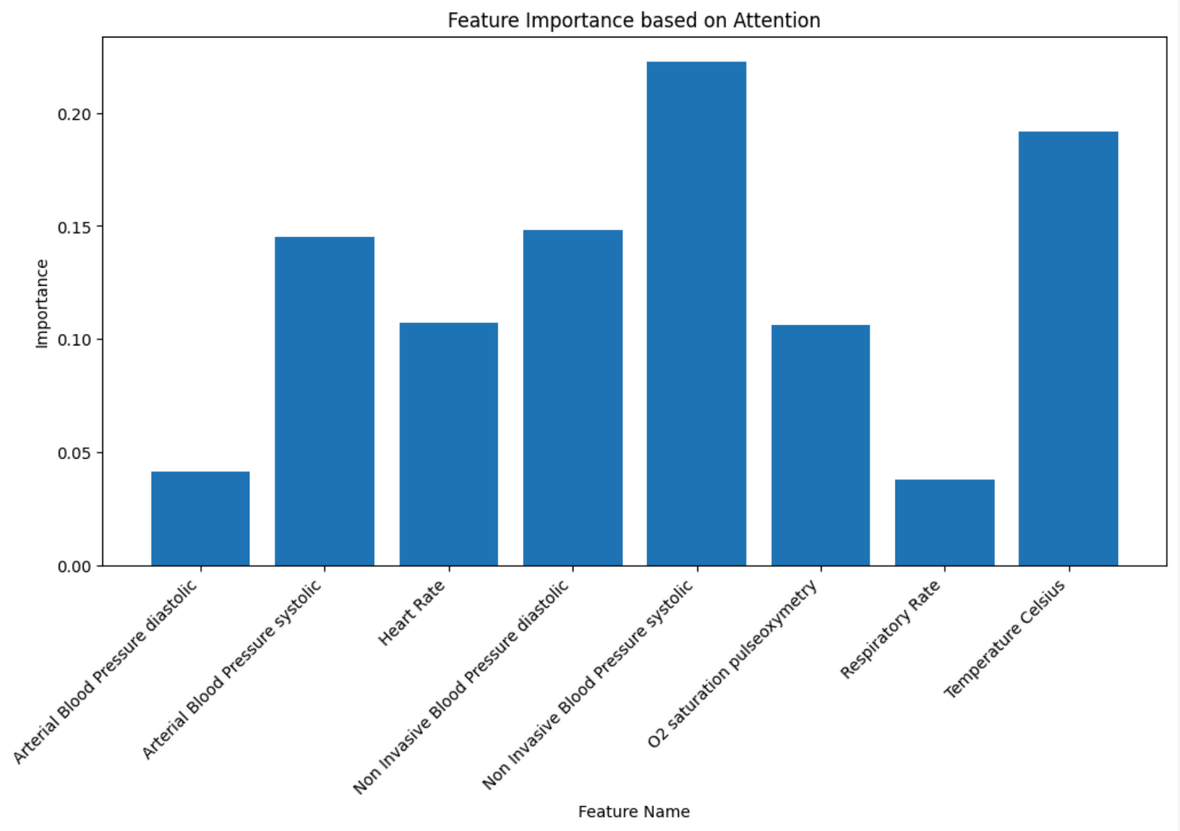
모델 적용

2) 심정지 발생 2시간 전 데이터 필터링

필터링된 데이터에서 stay_id별 행 개수의 평균값(**9**개)으로 개수 통일

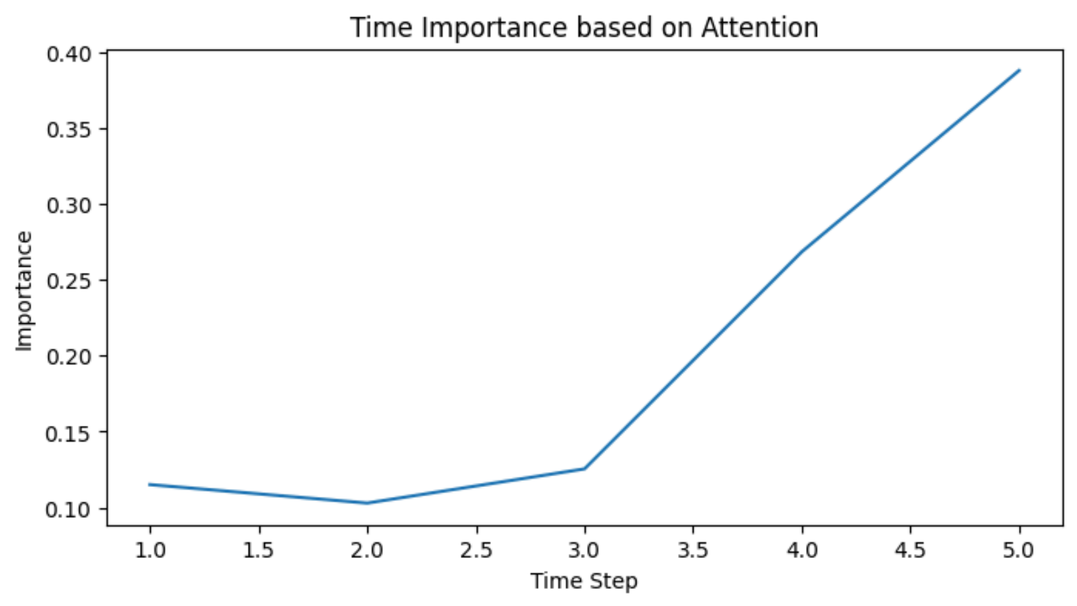
Accuracy (정확도): 0.79
Sensitivity (민감도, Recall): 0.97
F1 Score: 0.82



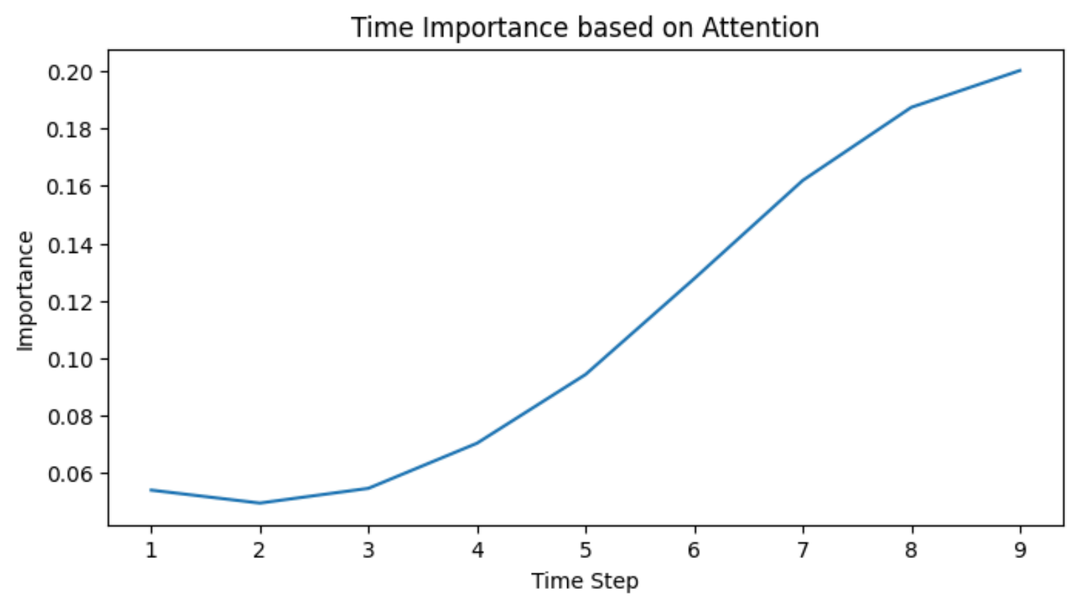


모델 적용

1) 심정지 발생 1시간 전 데이터 필터링

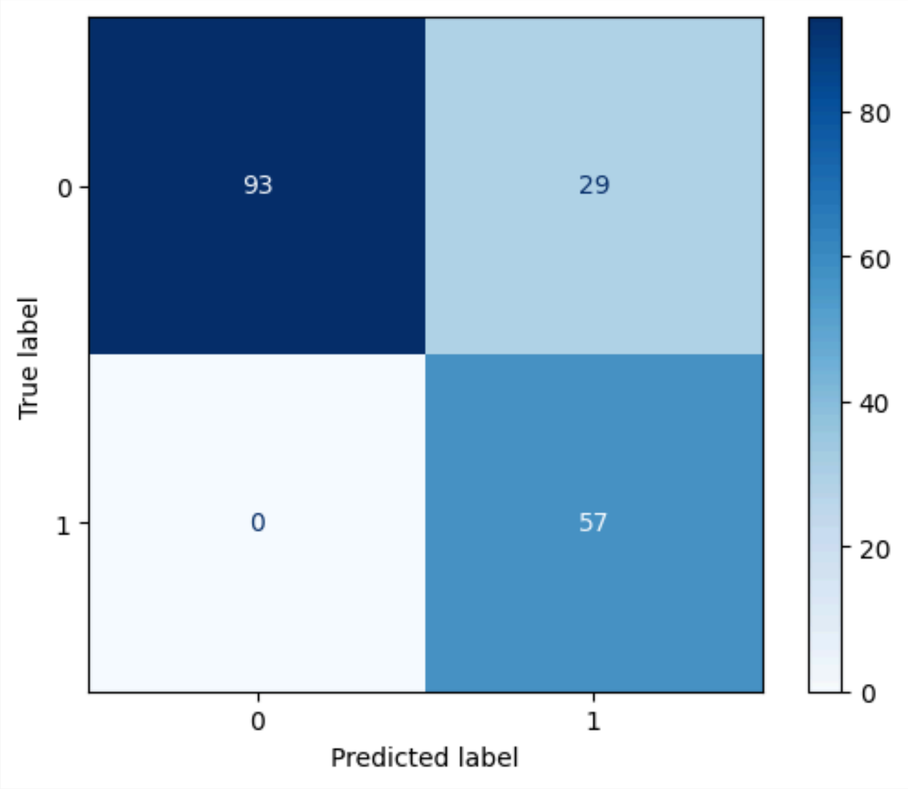


2) 심정지 발생 2시간 전 데이터 필터링

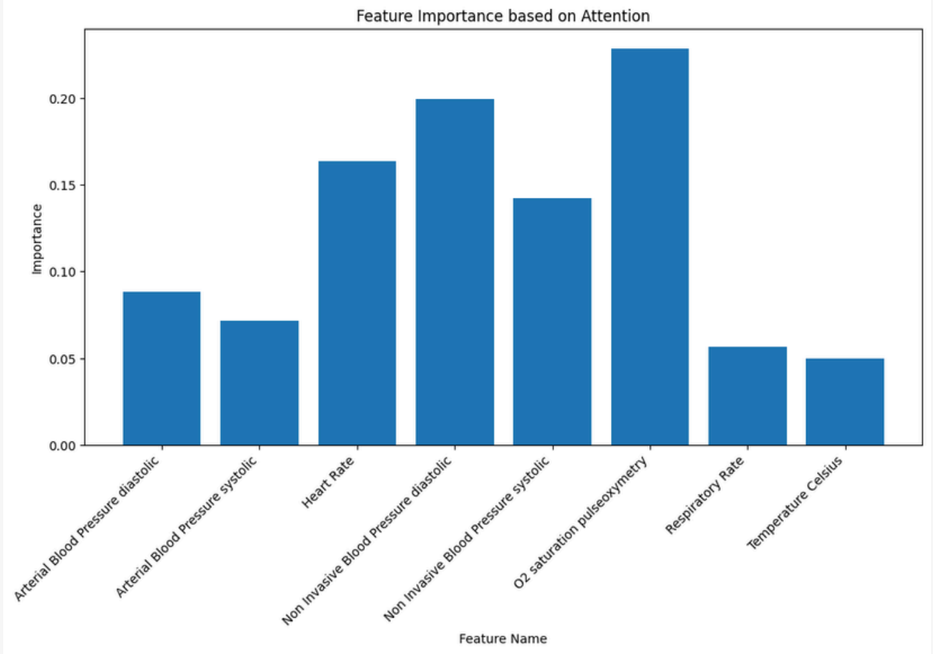


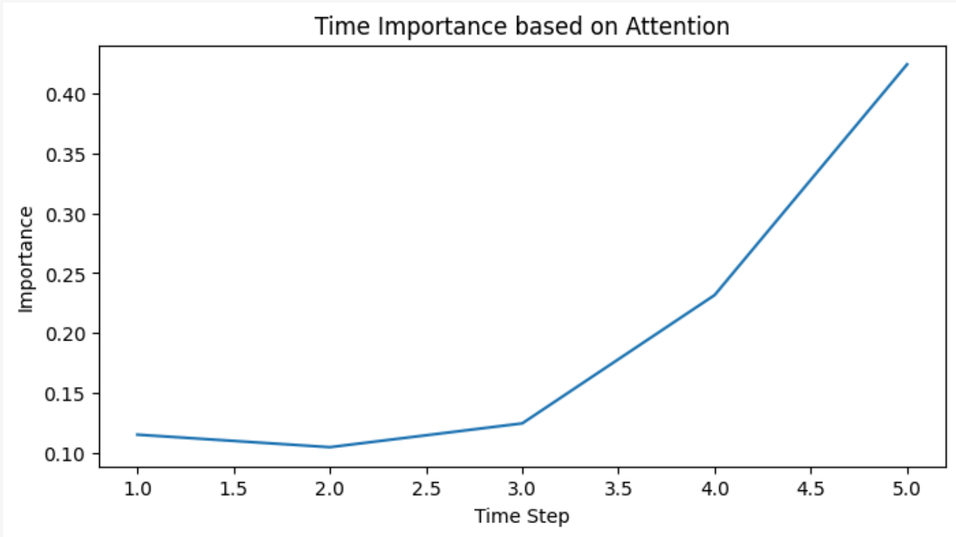
심정지 예측 모델을 활용한
응급실 내 심정지 예측

모델 적용



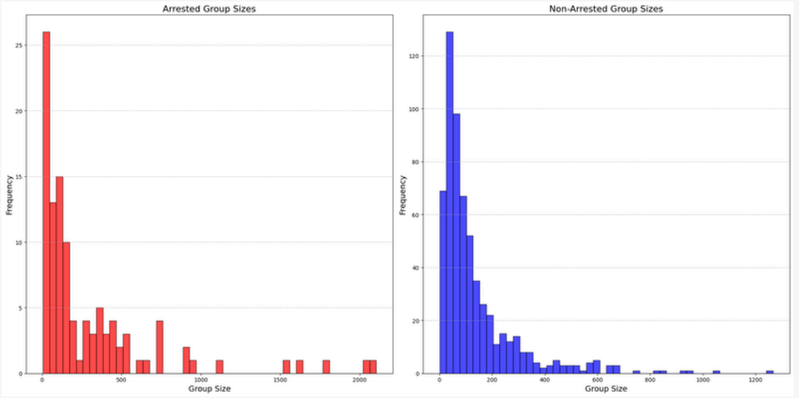
Accuracy (정확도): 0.84
Sensitivity (민감도, Recall): 1.00
Specificity (특이도): 0.76
F1 Score: 0.80





한계점

심정지 환자, 비심정지 환자들의 stay_id별 바이탈 사인 개수 편차



1) 적은 데이터 셋

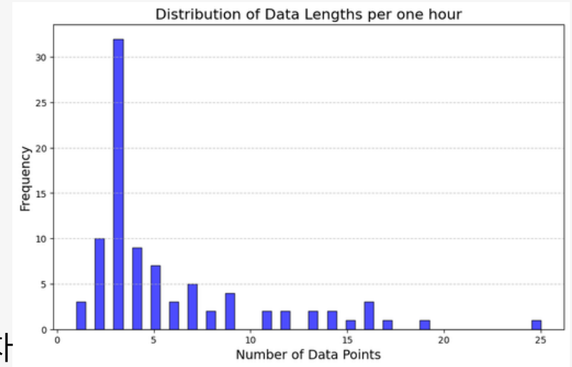
응급실에서 중환자실로 이동한 환자 중에서 심정지가 일어난 경우로 나누다 보니 109명의 환자로 분석을 하게 되어, 상대적으로 낮은 데이터 셋을 사용하게 됨.

2) vital sign의 결측치 & 개수 편차

vital sign 자체 결측치도 매우 많고 stay_id별 데이터의 개수 편차 또한 많아 일관적인 모델 적용이 어려움

3) 시간 데이터의 비연속성

바이탈이 랜덤한 시간마다 측정하고, 동일 시간 내에 검사한 횟수도 사람마다 달라 고정 샘플링을 진행하게 되었고, 이는 데이터를 왜곡했을 가능성이 존재



한 시간 내 시행한 vital 체크 수의 편차

THANK YOU.

QA

결측치를 선형보간법으로 채웠는데, vital 데이터에 결측치를 보간하는 것이 적절하다고 생각하시는지?

- 해당되는 데이터가 충분하다면 결측치를 제거하고 사용하는 것이 가장 이상적이라고 생각했지만, 응급실에서 중환자실로 이동한 심정지 환자의 수가 적다보니 해당하는 바이탈 데이터의 개수도 적었습니다. 주요 vital sign의 결측치가 최소 30%정도 있다 보니, 제거했을 경우 가용할 수 있는 데이터의 양이 너무 적었습니다. 따라서 최대한의 데이터를 살리기 위해 하나의 심정지 케이스 당 결측치가 90%가 넘는 경우에는 그 수치는 예측에서 제외하도록 하고 나머지 데이터들은 결측치를 보간하여 사용하는 방식을 채택하였습니다.

심정지를 예측하는데 더 긴 시간의 데이터는 필요 없을지 궁금합니다.

- 처음에는 좀 더 긴 시간의 데이터를 통해 예측을 진행하고자 하였으나, 중환자실로 이송 후 심정지가 온 시간이 환자마다 너무나도 상이하였습니다. 동일한 time step을 가져야하는 LSTM에서 모델을 돌리기 위해 각 환자의 데이터를 임의로 늘려야 하기 때문에 데이터가 왜곡될 우려도 있었고, 실제 긴 시간으로 모델을 돌려본 결과 변수와 시간의 중요도를 파악할 수 없어 짧은 시간의 데이터로 측정하였습니다. 또한 심정지의 경우 짧은 시간 내 vital 값의 변화에 반응하기 때문에 1시간에서 2시간 내의 데이터가 더욱 적절할 수 있을것이다라고 판단하였습니다.

QA

여러 vital sign 중, 핵심 vital sign을 대해서 어떤 기준으로 선택하였는지 궁금하고, 오버샘플링 메소드인 SMOTE, 결측치 보완방식인 linear interpolation 모두 선형적인 euclidean distance기반 방식인데 다차원인 해당 데이터셋이 다차원임에도 euclidean distance 기반 method를 사용하는 것이 적절한 approach인가요?

- 우선 혈류와 관련된 모든 feature를 포함하였습니다. 그리고 domain knowledge를 활용해 심정지와 관련 있는 특성을 더 보완하여 최종적으로 32개의 feature를 추출했습니다. 데이터 분석에서는 32개 중 일부 특성만을 사용했습니다.
- 결측치 보완의 경우에는 각 vital sign을 기준으로 독립적으로 채웠습니다. 예를 들어 Heart Rate의 경우 Heart Rate만을 가지고 결측치를 채웠습니다. 바이탈의 경우 연속적인 지표성을 띄기 때문에 선형보간법으로 채워넣는것이 가장 적절할 것이라고 생각하였습니다. 그리고 SMOTE의 경우 다른 업샘플링기법보다 데이터의 분포를 잘 유지하기 때문에, 심정지 환자의 주요 vital 요인을 잘 보존할 수 있을 것이라고 생각하여, SMOTE를 사용하였습니다.

-자카르 유사도를 통해 비슷한 질병을 분류를 어떻게 했는지 자세히 알고싶어요,
-"자카르 유사도"를 활용한 정확한 이유와 활용 방법이 궁금합니다.
-비슷한 질병 코드를 찾는 방법으로 자카드 유사도를 선택하셨는데, 코사인 유사도 등 다른 방법도 있었을텐데 적절하게 질병 코드가 분류되었는지도 평가하셨나요?

- 환자마다 여러개의 질병코드를 가지고 있었기 때문에 중환자실 데이터에서 각 stay_id 기준으로 질병코드를 묶은 후 심정지 환자와의 질병코드 집합들을 기준으로 하여, 비심정지 환자 중 질병 집합의 자카드 유사도가 0.4인 환자를 기준으로 추출하였습니다. 이를 활용한 이유는 심정지가 오지 않은 환자 중 질병이 유사한 대조군을 뽑기 위해서였고, 질병 코드가 하나의 Set으로 표현될 수 있어 자카드 유사도를 사용한 것입니다. 또한 질병 코드로는 어떤 질병이 더 위험한지에 대한 가중치를 판단하기가 애매하고, 공간상의 거리로 표현하기도 매우 애매하다고 판단하여, 집합의 특성을 이용한 자카드 유사도가 더욱 유용할 것이라고 예측되었습니다.

QA

심정지 여부를 예측하는 모델의 활용성이 어떻게 되는가?(=프로젝트에서 진행한 모델링이 구체적으로 어떤 목적을 달성해주는지, 어떻게 쓰일건지에 대한 계획이 궁금합니다)

- 저희는 심정지가 입원 여부와는 상관없이 바이탈 사인이 달라지는 특정 시점에 일어난다고 판단하였기 때문에 비슷한 증상을 가진 환자들의 중환자실 데이터를 통한 심정지 시점 예측 분석은 관련 질환으로 급하게 응급실에 입원한 환자들의 바이탈 사인을 통해 미리 심정지 시점을 예측하는 것에 활용할 가능성이 있다고 생각합니다. 이는 의료진이 바이탈 측정기의 비상사인을 통해 조치를 취하는 것보다 더 빠른 조치를 취함으로써 환자들의 빠른 회복 또는 예방에 사용 가능하다고 보고 있습니다.

검사 횟수가 환자마다 다른 상황이라면 예측 결과가 동등한 조건에서 나올수 없을 것 같습니다. 이 문제도 고려하셨나요?

- 실제로 시간 내 검사 횟수가 환자마다 완전이 상이하고, 전체 검사 횟수 또한 환자마다 달랐습니다. 이 문제를 고려해서 저희는 시계열 모델 활용 시 환자별 검사 주기를 1시간,2시간 안으로 제한한 후 이 안에 받은 검사 횟수 또한 동일하게 통일하는 과정을 거쳤습니다. 이를 통해 환자 간 검사 횟수 차이에 의한 편향을 줄이고, 환자별로 다른 검사 빈도가 예측 모델의 결과에 영향을 미치는 것을 최소화 하여 더욱 정확한 예측 결과 도출이 가능했던 것 같습니다.

Q A

심정지 예측 요인 분석에서 가장 중요한 바이탈 사인은 무엇이었으며, 이 결과는 임상적으로 어떤 의미를 가지나요?

- 예측 요인 분석에서 가장 중요한 바이탈 사인은 Heart Rate(심박수),Respiratory Rate(호흡수) 였습니다. 심박수가 비정상적으로 빠를 경우(빈맥), 이는 저혈량 쇼크나 심근경색과 같은 심정지로 이어질 수 있는 심각한 상태를 반영할 수 있습니다 또한 호흡수는 심정지로 이어지는 악화 경로에서 초기 경고 신호로 자주 나타나기 때문에 의료진이 이를 적극적으로 모니터링하고 개입하면 심정지를 예방할 수 있는 기회를 제공합니다.

데이터 불균형 기법인 smote 기법의 단점에 대해서 고려하신 부분이 있을까요??

- 학습 데이터와 유사한 합성 데이터를 많이 생성하면 모델이 과적합될 가능성이 높아질 수 있는 가능성에 대해 고려했습니다. 저희 데이터의 경우 데이터의 차원이 많이 높지 않았고 그로 인한 차원의 저주의 영향을 적게 받아 모델의 성능이 하락하는 문제는 크게 발생하지 않은 것 같습니다.