20개 선택해서 rf, xgboost 같은거 돌려보고 10개 선택해서 훈련, 테스트?

조합 파일

- 기본적으로 subject_id는 포함

	파일	사용 컬럼
1	omr.csv	BMI (kg/m2), Systolic BP
2	patients.csv	isMale, anchor_age
3	labevents.csv	Creatinine, Hemoglobin, Platelet Count, Sodium, Potassium, Chloride, Urea Nitrogen, Chloride, Hematocrit, Bicarbonate, Anion Gap, labitem_id,charttime
4	admissions.csv	eGFR
5	vitals_chartevents.csv	Mean blood pressure, pulse rate, body temperature - 이 안에서도 전처리, 하나만 선택?
6	hosp/transfers.csv.gz	careunit
7	hosp/drgcodes.csv.gz	drg_code
8	hosp/diagnoses_icd.csv.gz	seq_num(1:주, 2:부)
9	aki_results_stage_v2.csv	aki
10	comorbidity_counts.csv	comorbidity_count
11	medication_presence_with_total.csv	total_medications

	itemid	label
112	220473	Taurin
161	220799	ZSpecific Gravity (urine)
497	224015	Urine Source
498	224016	Urine Color
915	224876	Urine Appearance
1310	225454	Urine Culture
1883	226566	Urine and GU Irrigant Out
1934	226627	OR Urine

1938 226631 PACU Urine
2101 227059 UrineScore_ApacheIV
2232 227471 Specific Gravity (urine)
2243 227489 GU Irrigant/Urine Volume Out
2249 227519 Urine output_ApacheIV

1. 병합 -
2. 만명 추출

```
# 주 진단과 부 진단 구분
primary_diagnosis = diagnoses_icd[(diagnoses_icd['seq_num'] == 1) &
(diagnoses_icd['subject_id'].isin(aki_patient_ids))]
secondary_diagnosis = diagnoses_icd[(diagnoses_icd['seq_num'] > 1) &
(diagnoses_icd['subject_id'].isin(aki_patient_ids))]

# 주 진단 코드 통계
primary_counts =
primary_diagnosis['icd_code'].value_counts().reset_index()
primary_counts.columns = ['icd_code', 'count']

# 부 진단 코드 통계
secondary_counts =
secondary_diagnosis['icd_code'].value_counts().reset_index()
secondary_counts.columns = ['icd_code', 'count']

# DRG 코드와 연결하여 분석
drg_aki = drgcodes[drgcodes['subject_id'].isin(aki_patient_ids)]
drg_primary = drg_aki[drg_aki['drg_type'] == 'APR']

# DRG 코드별 주 진단 코드 통계
drg_primary_counts = drg_primary.groupby(['drg_code',
'description'])['subject_id'].nunique().reset_index()
drg_primary_counts.columns = ['drg_code', 'description', 'patient_count']

# 결과 확인
print("주 진단 코드별 환자 수:")
print(primary_counts.head())
```

```
print("\n부 진단 코드별 환자 수:")
print(secondary_counts.head())

print("\nAKI 환자의 DRG 코드별 환자 수:")
print(drg_primary_counts.head())
```

Pulse Rate 관련 항목:

	itemid	label	abbreviation	linksto	category	unitname	param_type	lownormalvalue	highnormalvalue
2	220045	**Heart Rate**	HR	chartevents	Routine Vital Signs	bpm	Numeric	NaN	NaN
3	220046	Heart rate Alarm - High	HR Alarm - High	chartevents	Alarms	bpm	Numeric	NaN	NaN
4	220047	Heart Rate Alarm - Low	HR Alarm - Low	chartevents	Alarms	bpm	Numeric	NaN	NaN
3857	229770	Resting Pulse Rate (COWS)	Resting Pulse Rate	chartevents	Toxicology	NaN	Text	NaN	NaN

Body Temperature 관련 항목:

	itemid	label	abbreviation	linksto	category	unitname	param_type	lownormalvalue	highnormalvalue
337	223761	**Temperature Fahrenheit**	Temperature F	chartevents	Routine Vital Signs	°F	Numeric	NaN	NaN
338	223762	Temperature Celsius	Temperature C	chartevents	Routine Vital Signs	°C	Numeric	NaN	NaN
505	224027	Skin Temperature	Skin Temp	chartevents	Skin - Assessment	NaN	Text	NaN	NaN
767	224642	Temperature Site	Temp Site	chartevents	Routine Vital Signs	NaN	Text	NaN	NaN
790	224674	Changes in Temperature	Changes in Temperature	chartevents	Toxicology	NaN	Text	NaN	NaN
1814	226329	Blood Temperature CCO (C)	Blood Temp CCO (C)	chartevents	Routine Vital Signs	°C	Numeric	NaN	NaN
2097	227054	TemperatureF_ApacheIV	TemperatureF_ApacheIV	chartevents	Scores - APACHE IV (2)	°F	Numeric	NaN	NaN
2776	228242	Pt. Temperature (BG) (SOFT)	Pt. Temperature (BG) (SOFT)	chartevents	Labs	NaN	Numeric	NaN	NaN
3466	229236	Cerebral Temperature (C)	Cerebral T (C)	chartevents	Hemodynamics	°C	Numeric	NaN	NaN

Mean Blood Pressure 관련 항목:

	itemid	label	abbreviation	\
26	220181	**Non Invasive Blood Pressure mean**	NBPm	
1975	226765	MapApacheIIScore	MapApacheIIScore	
1976	226766	MapApacheIIValue	MapApacheIIValue	
1981	226771	PotassiumApacheIIScore	PotassiumApacheIIScore	
1982	226772	PotassiumApacheIIValue	PotassiumApacheIIValue	
1985	226775	SodiumApacheIIScore	SodiumApacheIIScore	
1986	226776	SodiumApacheIIValue	SodiumApacheIIValue	
2066	227023	MAP_ApacheIV	MAP_ApacheIV	
2067	227024	MapScore_ApacheIV	MapScore_ApacheIV	
3265	228872	HM II- Mean BP	HM II- Mean BP	
3888	229827	Mean BP (VAD)	Mean BP (VAD)	

	linksto	category	unitname	param_type	lownormalvalue	\
26	chartevents	Routine Vital Signs	mmHg	Numeric	NaN	
1975	chartevents	Scores - APACHE II	NaN	Numeric	NaN	
1976	chartevents	Scores - APACHE II	mmHg	Numeric	NaN	
1981	chartevents	Scores - APACHE II	NaN	Numeric	NaN	
1982	chartevents	Scores - APACHE II	NaN	Numeric	NaN	
1985	chartevents	Scores - APACHE II	NaN	Numeric	NaN	
1986	chartevents	Scores - APACHE II	NaN	Numeric	NaN	
2066	chartevents	Scores - APACHE IV (2)	mmHg	Numeric	NaN	
2067	chartevents	Scores - APACHE IV (2)	NaN	Numeric	NaN	
3265	chartevents	Hemodynamics	mmHg	Numeric	NaN	
3888	chartevents	Durable VAD	mmHg	Numeric	NaN	

	highnormalvalue
26	NaN
1975	NaN
1976	NaN
1981	NaN
1982	NaN
1985	NaN
1986	NaN
2066	NaN
2067	NaN
3265	NaN
3888	NaN

Mean Blood Pressure 항목이 존재합니다.

환자의 기본 정보

- **Age**: 나이 (`admissions.csv` 또는 `patients.csv`)
- **Gender**: 성별 (`patients.csv`)
- **BMI**: 체질량 지수 (신장, 체중이 있어야 계산 가능)

2. 생체 신호 **(Vital Signs from `chartevents.csv`)**

- **Systolic Blood Pressure (SBP)**: 수축기 혈압
- **Diastolic Blood Pressure (DBP)**: 이완기 혈압
- **Mean Arterial Pressure (MAP)**: 평균 동맥압
- **Heart Rate (HR)**: 심박수
- **Respiratory Rate (RR)**: 호흡수
- **Oxygen Saturation (SpO2)**: 산소포화도
- **Temperature**: 체온

3. 실험실 검사 결과 **(Laboratory tests from `labevents.csv`)**

- **Serum Creatinine**: 혈청 크레아티닌 (AKI 진단의 핵심 요소)
- **Blood Urea Nitrogen (BUN)**: 혈액 요소 질소
- **Glomerular Filtration Rate (eGFR)**: 추정 사구체 여과율 (BUN 및 Creatinine로 계산 가능)
- **Hemoglobin**: 헤모글로빈
- **White Blood Cell (WBC) count**: 백혈구 수치
- **Platelet Count**: 혈소판 수치
- **Sodium**: 혈중 나트륨 수치
- **Potassium**: 혈중 칼륨 수치
- **Bicarbonate (HCO3)**: 중탄산염 수치

4. 입원 및 수술 정보 **(Admissions & Procedures from `admissions.csv` and `procedures_icd.csv`)**

- **Length of ICU Stay**: 중환자실 체류 기간
- **Surgery History**: 수술 이력 (과거 수술 유무)

5. 약물 사용 정보 **(Medications from `prescriptions.csv`)**

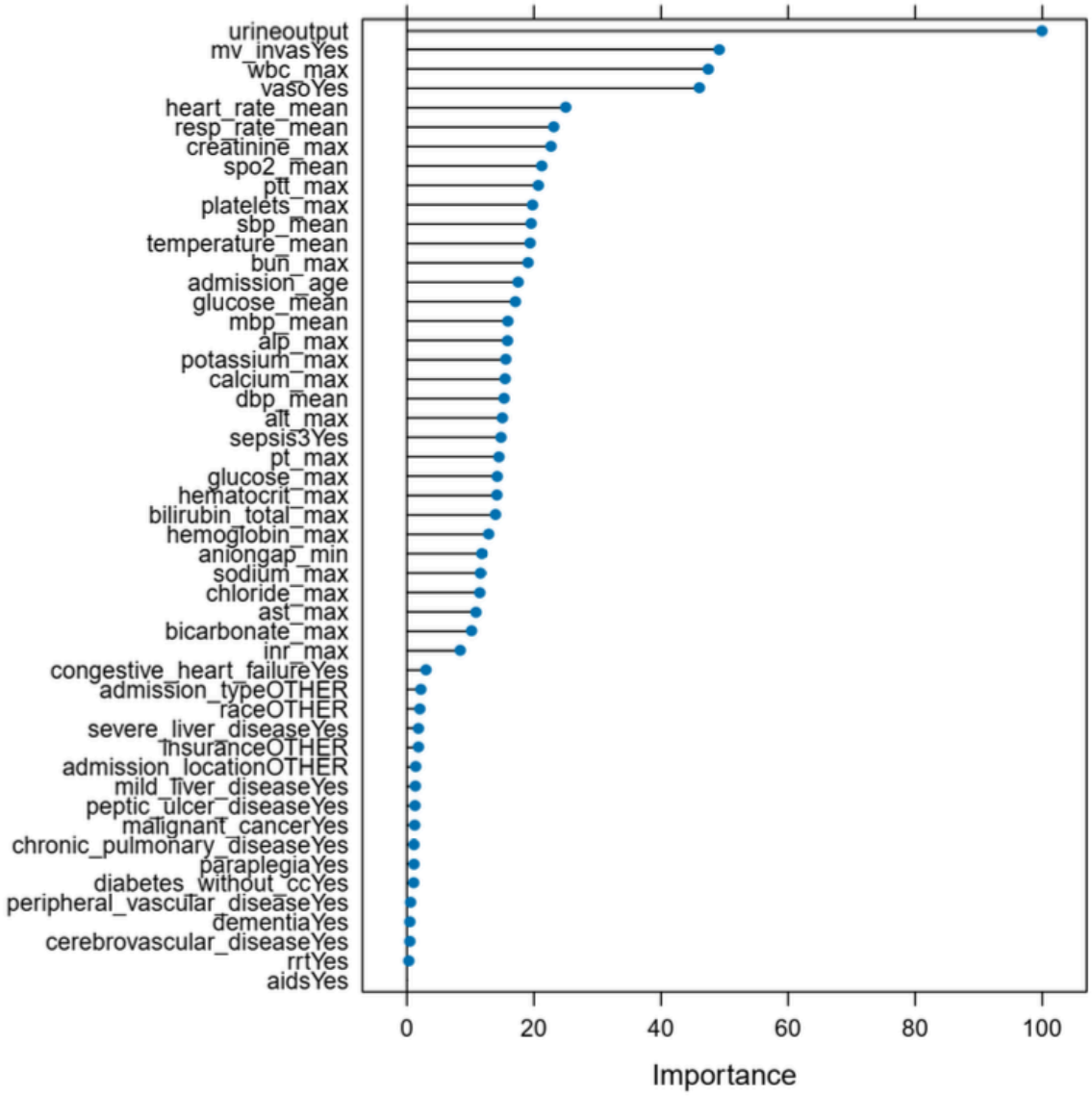
- **Diuretics Use**: 이뇨제 사용 여부
- **Vasopressors Use**: 혈관수축제 사용 여부

이거는 careunit unique정보들 중….

주어진 데이터는 MIMIC-3 데이터셋의 환자 care unit (치료 단위) 정보와 각 단위의 환자 수를 나타내는 것으로 보입니다. AKI (급성 신손상) 진단 모델링을 위해 이 정보를 사용할 때, 다음과 같은 care unit들이 특히 중요할 수 있습니다:

1. 중환자실 관련 unit:
    - Medical Intensive Care Unit (MICU)
    - Surgical Intensive Care Unit (SICU)
    - Medical/Surgical Intensive Care Unit (MICU/SICU)
    - Cardiac Vascular Intensive Care Unit (CVICU)
    - Trauma SICU (TSICU)
    - Neuro Surgical Intensive Care Unit (Neuro SICU)
2. 이유: 중환자실 환자들은 AKI 발생 위험이 높습니다.
3. 응급실 관련 unit:
    - Emergency Department
    - Emergency Department Observation
4. 이유: 급성 질환으로 내원한 환자들 중 AKI 위험이 있는 경우가 많을 수 있습니다.
5. 내과계 unit:
    - Medicine
    - Medicine/Cardiology
    - Nephrology (데이터에 없지만 관련성 높음)
6. 이유: 내과적 질환, 특히 심장 질환이나 신장 질환과 관련된 unit에서 AKI 발생 가능성이 높습니다.
7. 외과계 unit:
    - Surgery/Trauma
    - Surgery
    - Cardiac Surgery
8. 이유: 수술 후 AKI 발생 위험이 있습니다.
9. 특수 unit:
    - Transplant
    - Hematology/Oncology
10. 이유: 이식 환자나 항암치료 환자의 경우 AKI 위험이 높을 수 있습니다.

이러한 unit들의 환자 데이터를 중점적으로 분석하면 AKI 진단 모델의 정확도를 높일 수 있을 것입니다. 또한, 각 unit별 환자 수도 중요한 정보가 될 수 있으므로, 이를 가중치로 활용하거나 데이터 분포를 이해하는 데 사용할 수 있습니다.



상단 중요 변수들**:**

1. **urineoutput**: 소변량. 환자의 소변 배출량을 측정한 값으로, 신장 기능을 평가하는 중요한 지표입니다.
2. **mv_invasYes**: 침습적 기계 환기 사용 여부. 환자가 침습적인 방법으로 기계 환기를 받았는지 여부를 나타냅니다.
3. **wbc_max**: 백혈구 수치의 최고값. 백혈구 수는 면역 반응을 나타내며, 감염이나 염증 상태를 반영합니다.
4. **vasoYes**: 혈관 수축제 사용 여부. 혈관 수축제를 사용한 여부로, 이는 혈압을 유지하거나 관리하는 데 중요한 요소입니다.

5. **heart_rate_mean**: 평균 심박수. 환자의 심박수의 평균값을 나타내며, 심장 건강과 관련이 깊습니다.
6. **resp_rate_mean**: 평균 호흡수. 환자의 분당 호흡 횟수로, 호흡 기능을 평가하는 데 사용됩니다.
7. **creatinine_max**: 크레아티닌 수치의 최고값. 신장 기능을 평가하는 데 중요한 지표로 사용됩니다.
8. **spo2_mean**: 평균 산소포화도. 혈액의 산소포화도를 측정한 값으로, 호흡 기능과 관련이 있습니다.
9. **ptt_max**: 부분 트롬보플라스틴 시간(PTT)의 최고값. 혈액 응고 시간을 나타내며, 출혈성 질환 평가에 사용됩니다.
10. **platelets_max**: 혈소판 수치의 최고값. 혈액 응고에 중요한 역할을 하는 혈소판의 수를 나타냅니다.

중간 중요 변수들**:**

11. **sbp_mean**: 평균 수축기 혈압. 심장이 수축할 때의 혈압 수치를 의미하며, 고혈압 등의 상태를 평가하는 데 사용됩니다.
12. **temperature_mean**: 평균 체온. 환자의 체온을 측정한 값으로, 발열 여부 등을 평가할 수 있습니다.
13. **bun_max**: 혈중 요소질소(BUN) 수치의 최고값. 신장 기능을 평가하는 지표로 사용됩니다.
14. **admission_age**: 입원 당시 환자의 나이. 나이는 환자의 전반적인 건강 상태와 예후를 예측하는 데 중요한 변수입니다.
15. **glucose_mean**: 평균 혈당 수치. 혈중 포도당 농도를 나타내며, 당뇨병 및 대사 질환과 관련이 있습니다.
16. **mbp_mean**: 평균 동맥혈압. 평균 혈압 수치를 나타내며, 순환기계 상태를 평가하는 데 사용됩니다.
17. **alp_max**: 알칼리성 인산화효소(ALP) 수치의 최고값. 간과 뼈의 기능을 평가하는 지표입니다.
18. **potassium_max**: 칼륨 수치의 최고값. 혈액 내 칼륨 수치는 심장 기능 및 근육 기능과 관련이 있습니다.
19. **calcium_max**: 칼슘 수치의 최고값. 칼슘 수치는 뼈 건강 및 신경 기능을 평가하는 데 중요합니다.
20. **dbp_mean**: 평균 이완기 혈압. 심장이 이완될 때의 혈압 수치를 나타냅니다.
21. **alt_max**: 알라닌 아미노전이효소(ALT) 수치의 최고값. 간 기능을 평가하는 지표입니다.
22. **sepsis3Yes**: 패혈증 여부. 환자가 패혈증을 진단받았는지 여부를 나타냅니다.

하단 중요 변수들**:**

23. **pt_max**: 프로트롬빈 시간(PT)의 최고값. 혈액 응고 시간을 측정한 값입니다.
24. **hematocrit_max**: 최고 혈액 농축률. 적혈구의 비율로, 산소 전달 능력과 관련이 있습니다.
25. **bilirubin_total_max**: 총 빌리루빈 수치의 최고값. 간 기능 및 담도계 이상을 평가하는 데 사용됩니다.

26. **hemoglobin_max**: 헤모글로빈 수치의 최고값. 혈액 내 산소 운반 능력을 평가하는 지표입니다.
27. **aniongap_min**: 최소 음이온 간격. 산-염기 균형을 평가하는 데 사용됩니다.
28. **sodium_max**: 나트륨 수치의 최고값. 혈액 내 나트륨 수치를 측정하며, 전해질 불균형을 평가할 수 있습니다.
29. **chloride_max**: 염화물 수치의 최고값. 체내 전해질 균형을 평가하는 데 사용됩니다.
30. **ast_max**: 아스파르테이트 아미노전이효소(AST) 수치의 최고값. 간 손상을 평가하는 지표입니다.
31. **bicarbonate_max**: 중탄산염 수치의 최고값. 혈액의 산-염기 균형을 유지하는 데 중요한 역할을 합니다.
32. **inr_max**: 국제정상화비(INR)의 최고값. 혈액 응고 기능을 평가하는 지표입니다.

기타 질병 관련 변수들**:**

33. **congestive_heart_failureYes**: 울혈성 심부전 여부. 환자가 울혈성 심부전을 앓고 있는지 여부입니다.
34. **raceOTHER**: 환자의 인종이 'OTHER'(기타)로 분류된 경우.
35. **severe_liver_diseaseYes**: 중증 간질환 여부. 환자가 심각한 간 질환을 앓고 있는지 여부입니다.
36. **insuranceOTHER**: 보험 유형이 'OTHER'(기타)로 분류된 경우.
37. **mild_liver_diseaseYes**: 경증 간질환 여부. 환자가 경증 간 질환을 앓고 있는지 여부입니다.
38. **peptic_ulcer_diseaseYes**: 소화성 궤양병 여부.
39. **malignant_cancerYes**: 악성 암 여부. 환자가 악성 암을 앓고 있는지 여부입니다.
40. **chronic_pulmonary_diseaseYes**: 만성 폐 질환 여부.
41. **paraplegiaYes**: 하지 마비 여부.
42. **diabetes_without_ccYes**: 합병증 없는 당뇨병 여부.
43. **peripheral_vascular_diseaseYes**: 말초 혈관 질환 여부.
44. **dementiaYes**: 치매 여부.
45. **cerebrovascular_diseaseYes**: 뇌혈관 질환 여부.
46. **rrtYes**: 신장 대체 요법 여부. 환자가 신장 대체 요법을 받았는지 여부입니다.
47. **aidsYes**: 에이즈(AIDS) 여부.

요약**:**

그래프에 표시된 변수들은 주로 환자의 생리적 측정값(혈압, 심박수, 체온 등), 검사 결과(크레아티닌, 빌리루빈 등), 그리고 환자의 병력(심부전, 암, 폐 질환 등)과 관련된 데이터를 포함하고 있습니다. 이 값들은 환자의 상태를 예측하는 데 중요한 역할을 하며, 중요도가 높은 변수들이 모델 예측에 가장 큰 영향을 미칩니다.

이 그래프는 아마도 **병원 입원 중 발생할 수 있는 주요 결과(예: 중환자실 입원, 사망 위험, 급성 신손상 등)**를 예측하는 데 사용된 것 같습니다.