LG U+ 프로젝트 데이터 모델링 보고서

Step1. 개요

1. 목적 : 사용자가 입력한 건강 정보 데이터를 바탕으로 건강 상태를 파악 후 맞춤 건강 정보 영상을 제공하는 서비스를 제작한다.

2. 한계 : 사용자의 건강 검진 및 건강 정보 데이터를 바탕으로 건강 상태를 예측하는 모델을 만들기에는 데이터 수집과 데이터 라벨링 하기가 까다롭다.

3. 대안 :
    - kaggle에서 병 위험군 분류 예측하기 쉬운 여러 데이터 셋을 가지고 머신 러닝 모델을 만든다.
    - 병 위험군 분류 모델의 성능이 좋으면 데이터 셋을 선택하여 모델을 만들고 성능이 안 좋으면 데이터 셋을 선택하지 않는다.
    - 만든 모델의 Feature Importance와 사용자에게 받기 쉬운 Feature인지를 고려하여 Feature Selection을 한 뒤 하이퍼 파라미터 튜닝 및 성능 검증을 통해 최적의 모델을 만든다.
    - 최종 모델의 Feature에 해당하는 건강 정보를 웹 UI를 통해 사용자에게 전달 받은 다음 전달 받은 건강 정보들을 만든 최종 모델에 적용하여 병 위험군인지 분류하여 사용자의 건강 상태를 파악한다.

4. 분류 성능 지표 : F1_Score사용
    - 사용 이유
        - 모델이 병 위험군을 잘 판정 하는 것에 대한 직관적으로 해석 가능한 성능 지표이다.
        - 불균형 데이터에서도 모델의 전반적인 성능을 공정하게 평가할 수 있다.

- F1_Score : 정밀도와 재현율의 조화평균이고 식은 다음과 같다.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- 정밀도 : 모델이 Positive로 예측한 데이터 중에서 실제로 Positive인 데이터의 비율

재현율 : 실제 Positive인 데이터 중에서 모델이 Positive로 올바르게 예측한 비율

5. 모델 선택 과정
   - 병 위험군 분류하기 위해 데이터 전처리 이후 여러 머신러닝 모델을 사용 후 F1_Score가 가장 높은 모델을 선택

Step2-1. 수면 장애 위험군 분류 데이터 (Sleep Disorder)

▼ 데이터 구성 및 전처리

1. Data : Sleep Health and Lifestyle Dataset (kaggle)

2. Data Info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           155 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```

3. Data Imbalance 확인

Sleep Disorder label	count
0	219
1	155

4. 데이터 전처리 과정

- NaN값 처리 : Sleep Disorder값이 NaN이면 0, NaN이 아니면 1로 처리 한 뒤 Sleep Disorder Label 칼럼 추가하고 Sleep Disorder칼럼 삭제

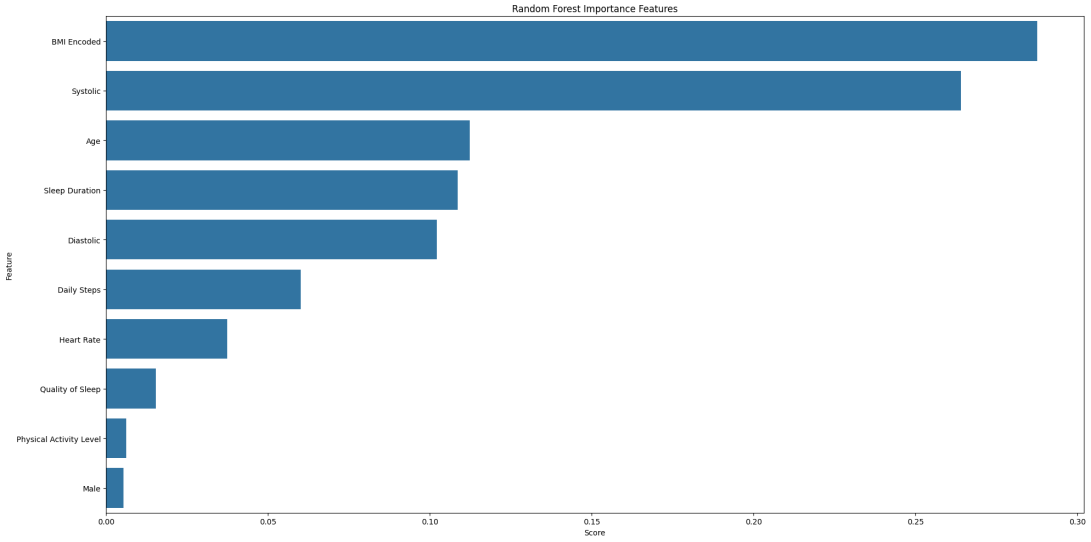
- Blood Pressure의 데이터 값과 형식은 125/88 (Str)이다. 이를 'Systolic'(수축기 혈압), 'Diastolic'(이완기 혈압)으로 분리하고 각각 Int값으로 들어가도록 처리

- 범주형 변수인 'Gender'와 'BMI Category' 칼럼에 각각 원-핫 인코딩, 라벨 인코딩 적용

Step2-2. 수면 장애 위험군 분류 모델 (Sleep Disorder)

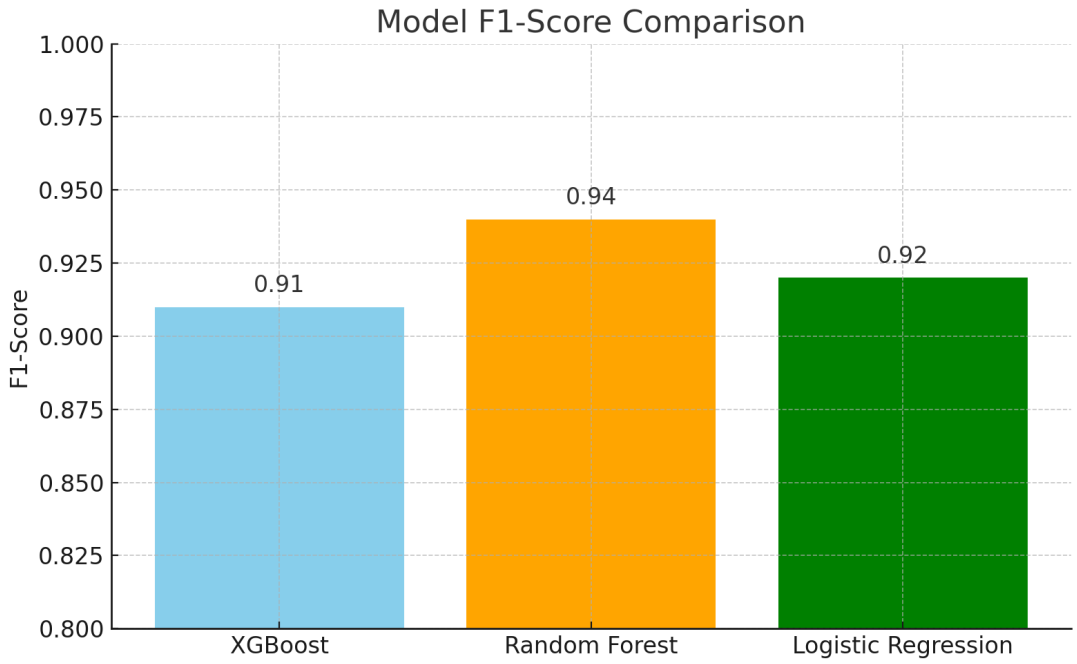
▼ 데이터 모델링

1. 선택 모델 : Random Forest
2. 모델링 과정
    - 하이퍼 파라미터 튜닝 : n_estimators, max_depth에 for loop방식을 이용하여 최적의 파라미터 선택
    - Feature Selection : Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



    - 최종 모델에 선택된 Feature : 'BMI Encoded', 'Age', 'Sleep Duration', 'Systolic', 'Diastolic', 'Daily Steps'
    - 모델 평가 지표 : F1_Score = 0.94

3. 다른 분류 모델과 성능 비교 (F1_Score)
    - 데이터가 많지 않아 Logistic Regression을 사용했고 분류 모델의 Baggig및 Boosting방식의 대표 알고리즘인 Random Forest와 XGBoost을 사용했다. 이 중 F1_Score가 가장 높은Randon Forest를 최종 선택했다.



Step3-1 심혈과 질환 위험군 분류 데이터 (Cardiovascular Disease)

▼ 데이터 구성 및 전처리

1. Data : Cardiovascular Disease dataset

2. Data info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 70000 entries, 0 to 69999
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           70000 non-null  int64
 1   age          70000 non-null  int64
 2   gender       70000 non-null  int64
 3   height       70000 non-null  int64
 4   weight       70000 non-null  float64
 5   ap_hi        70000 non-null  int64
 6   ap_lo        70000 non-null  int64
 7   cholesterol  70000 non-null  int64
 8   gluc         70000 non-null  int64
 9   smoke        70000 non-null  int64
 10  alco         70000 non-null  int64
 11  active       70000 non-null  int64
 12  cardio       70000 non-null  int64
dtypes: float64(1), int64(12)
memory usage: 6.9 MB
```

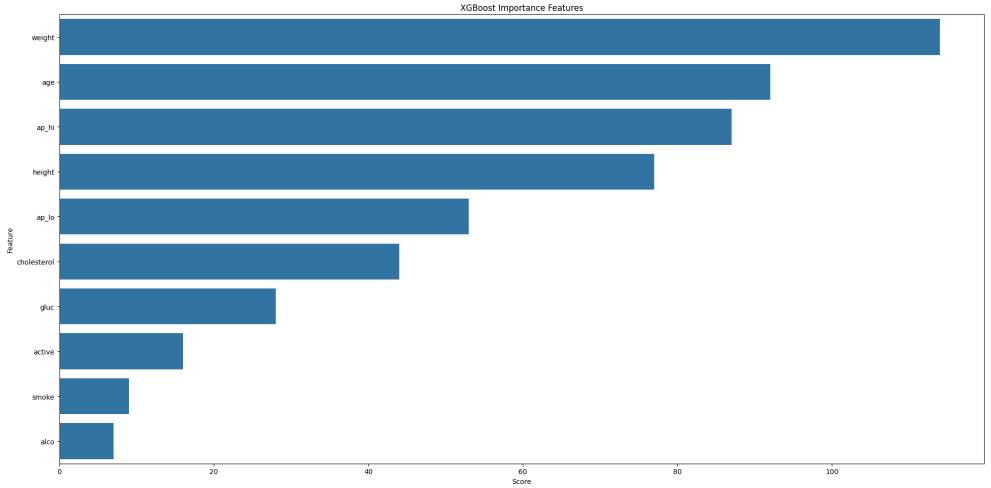
3. Data Imbalance확인

cardio	count
0	35021
1	34979

Step3-2. 심혈관 질환 위험군 분류 모델 (Cardiovascular Disease)

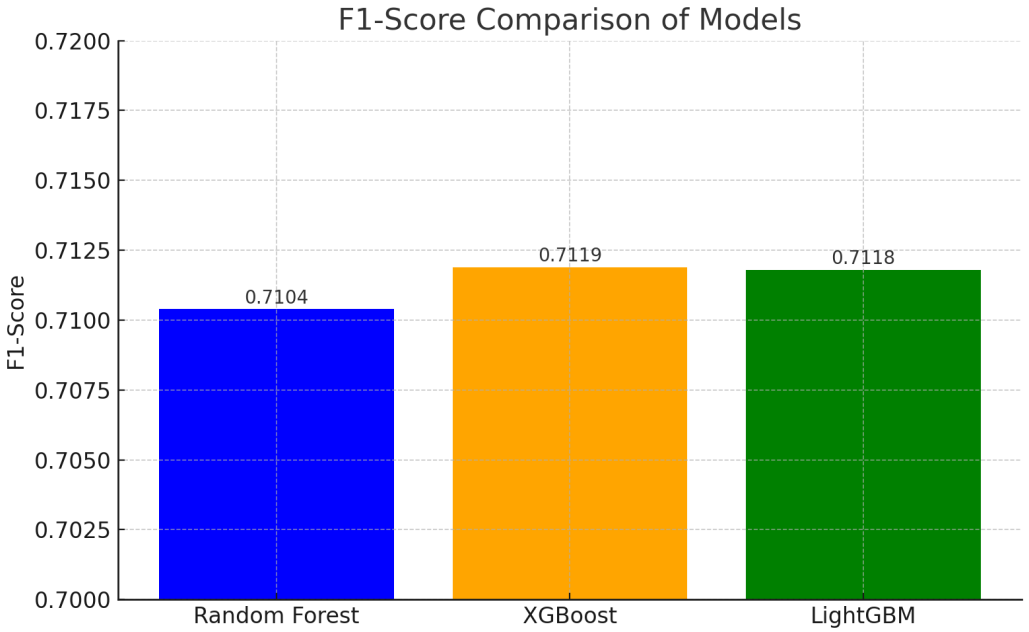
▼ 데이터 모델링

1. 사용모델 : XGBoost
2. 모델링과정
    - 하이퍼 파라미터 튜닝 : n_estimators, learning_rate, max_depth, reg_alpha에 for loop방식을 이용하여 최적의 파라미터 선택
    - Feature Selection : Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



- 최종 모델에 선택된 Feature : 'ap_hi', 'ap_lo', 'age', 'weight', 'height', 'cardio'
- 모델 평가 지표 : F1_Score = 0.7119

3. 다른 분류 모델과 성능 비교 (F1_Score)
    - 분류 모델의 Baggig및 Boosting방식의 대표 알고리즘인 Random Forest와 XGBoost을 사용했고 추가로 데이터가 많은 편이라 LightGBM 모델을 사용하여 F1_Score가 가장 높은 XGBoost를 최종 선택했다.



Step4-1. 당뇨 위험군 분류 데이터 (Diabetes)

▼ 데이터 셋 구성 및 전처리

1. Data: Diabetes Health Indicators Dataset (kaggle)

2.Data Info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 236378 entries, 0 to 236377
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   Diabetes_012          236378 non-null  float64
 1   HighBP                236378 non-null  int64
 2   HighChol              236378 non-null  float64
 3   CholCheck             236378 non-null  int64
 4   BMI                   236378 non-null  float64
 5   Smoker                236378 non-null  float64
 6   Stroke                236378 non-null  float64
 7   HeartDiseaseorAttack  236378 non-null  float64
 8   PhysActivity          236378 non-null  int64
 9   Fruits                236378 non-null  int64
 10  Veggies               236378 non-null  int64
 11  HvyAlcoholConsump     236378 non-null  int64
 12  AnyHealthcare         236378 non-null  int64
 13  NoDocbcCost           236378 non-null  float64
 14  GenHlth               236378 non-null  float64
 15  MentHlth              236378 non-null  float64
 16  PhysHlth              236378 non-null  float64
 17  DiffWalk              236378 non-null  float64
 18  Sex                   236378 non-null  int64
 19  Age                   236378 non-null  int64
 20  Education             236378 non-null  float64
 21  Income                236378 non-null  float64
dtypes: float64(13), int64(9)
```

3. data Imbalance 확인

Diabetes_012	
0.0	197191
2.0	33568
1.0	5619

4. 데이터 전처리 과정
    - 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: HvyAlcoholConsump을 Alco)
    - Diabetes_012 컬럼 값 처리: Diabetes_012값이 0,1,2로 구성 되어있는데, 기획한 서비스 내용에 맞게 1을 제거 후, 2를 1로 변경

Diabetes	
0	197191
1	33568

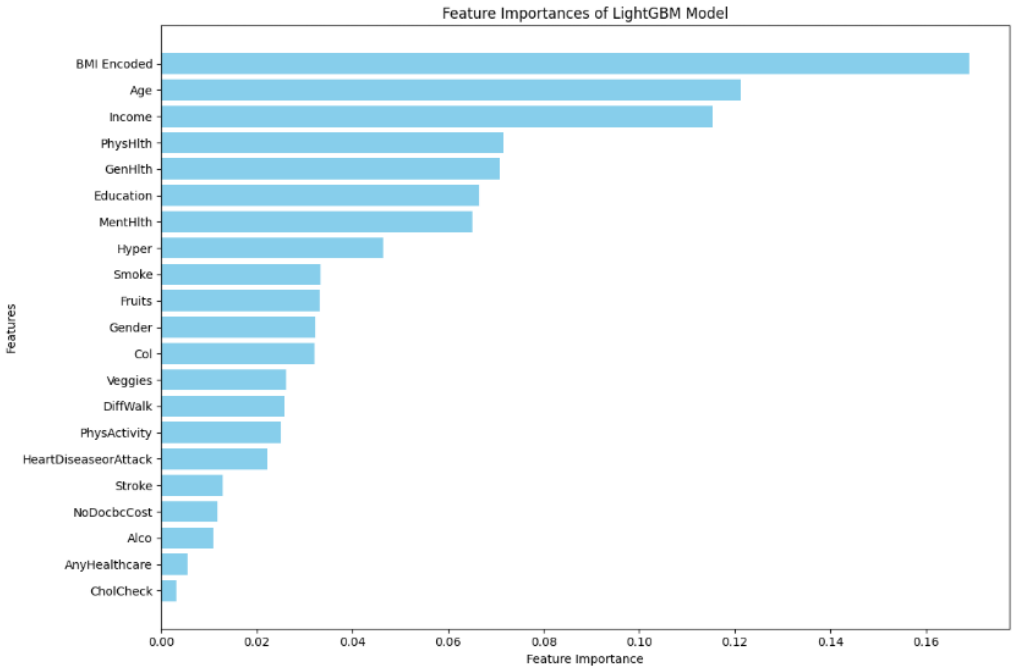
    - Smoker, HighChol, Diabetes_012의 데이터 값과 형식은 0/1 (float64)이다. 이를 int값으로 들어가도록 처리
    - 데이터 불균형 해결: SMOTE-Tomek을 사용해 데이터 불균형 해결

Step4-2. 당뇨 위험군 분류 모델 (Diabetes)

▼ 데이터 모델링

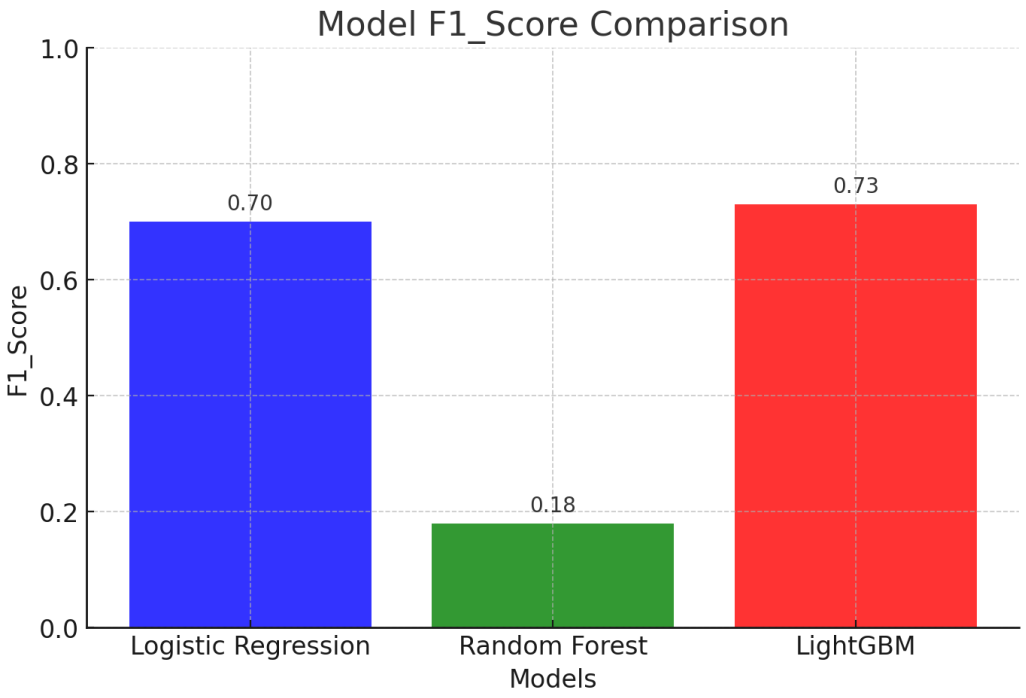
1. 사용 모델: LightGBM
2. 모델링 과정

- 하이퍼 파라미터 튜닝: RandomizedSearchCV을 사용하여 최적의 파라미터 선택
- Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



- 선택된 Feature: "Gender", "Age", "Hyper", "BMI Encoded", "Smoke", "Col", "Alco"
- 가중치: 선택 변수인 Diabetes의 예측 성능을 개선하기 위해 Diabetes 가중치 조정
- 모델 평가 지표: F1_Score = 0.73

3. 다른 분류 모델과 성능 비교 (F1_Score)
    - 데이터가 많아 대규모 데이터셋에 적합한 LightGBM을 사용했고 대중적인 분류 모델인 Logistic Regression과 Random Forest를 사용했다. 이 중 F1_Score가 가장 높은 LightGBM을 최종 선택했다.



Step5-1 폐암 위험군 분류 데이터 (lung cancer)

▼ 데이터 구성 및 전처리

1. Data: Survey Lung Cancer

2. Data Info확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Gender                 309 non-null    object
 1   Age                    309 non-null    int64
 2   Smoke                  309 non-null    int64
 3   YELLOW_FINGERS         309 non-null    int64
 4   ANXIETY                309 non-null    int64
 5   PEER_PRESSURE          309 non-null    int64
 6   CHRONIC DISEASE        309 non-null    int64
 7   Tired                  309 non-null    int64
 8   ALLERGY                309 non-null    int64
 9   WHEEZING               309 non-null    int64
 10  Alco                   309 non-null    int64
 11  COUGHING               309 non-null    int64
 12  SHORTNESS OF BREATH    309 non-null    int64
 13  SWALLOWING DIFFICULTY  309 non-null    int64
 14  CHEST PAIN             309 non-null    int64
 15  LUNG_CANCER            309 non-null    object
dtypes: int64(14), object(2)
```

3. data Imbalance 확인

LUNG_CANCER	count
YES	270
NO	39

4. 데이터 전처리 과정
    - 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: FATIGUE을 Tired)
    - 데이터 공백 제거
    - 범주형 변수인 'LUNG_CANCER'와 'Gender' 컬럼에 각각 문자를 이진 숫자 데이터로 처리, 라벨 인코딩 적용

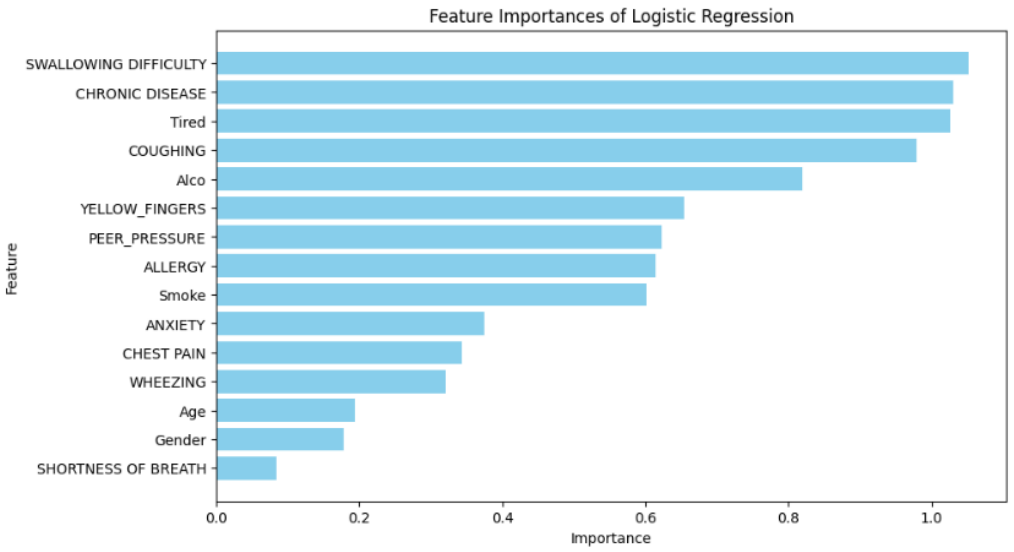
- 데이터 불균형 해결: SMOTE을 사용해 데이터 불균형 해결

Step5-2. 폐암 위험군 분류 모델 (lung cancer)

▼ 데이터 모델링

1. 사용 모델: Logistic Regression
2. 모델링 과정

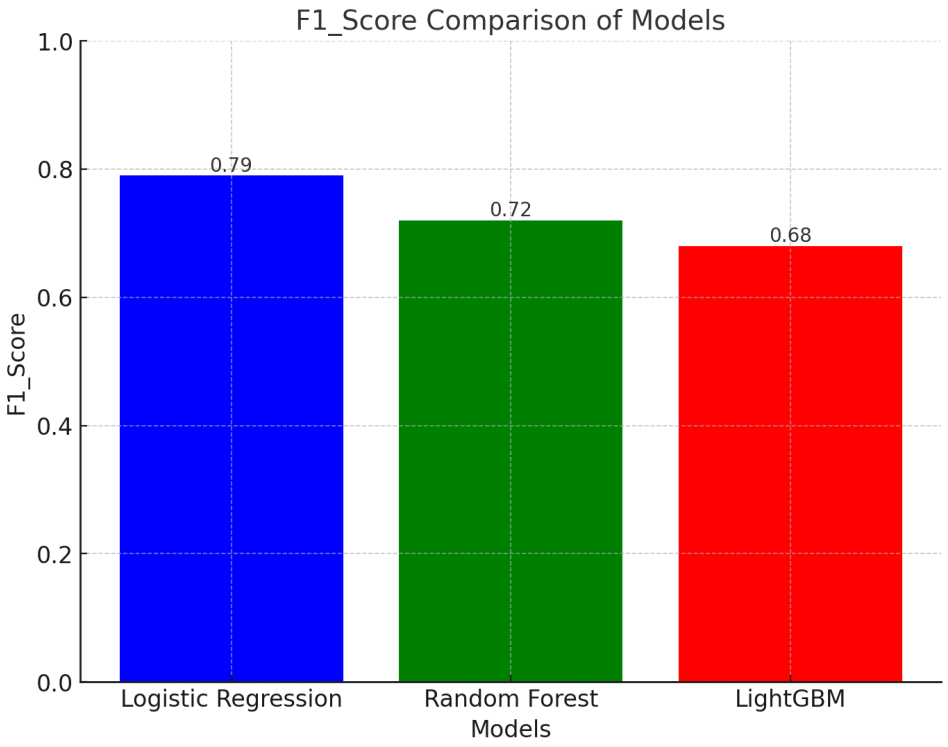
- 하이퍼 파라미터 튜닝: GridSearchCV을 사용하여 최적의 파라미터 선택
- Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



- 선택된 Feature: 'Gender', 'Age', 'Smoke', 'Tired', 'Alco'
- 모델 평가 지표: F1_Score = 0.79

3. 다른 분류 모델과 성능 비교 (F1_Score)

- 데이터가 많지 않아 소규모 데이터셋에 적합한 LogisticRegression을 선택했고 변수 중요도 평가 가능한 Random Forest와 카테고리형 변수 자동 처리 가능한 LightGBM을 사용했다. 이 중 F1_Score가 가장 높은 LogisticRegression을 최종 선택했다.



Step6-1. 간암 위험군 분류 데이터 (Liver_disease)

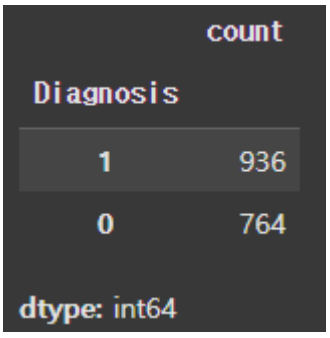
▼ 데이터 셋 구성 및 전처리

1. Data: Survey Lung Cancer

2. Data Info 확인

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1700 entries, 0 to 1699
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Age                 1700 non-null   int64  
 1   Gender              1700 non-null   int64  
 2   BMI                 1700 non-null   float64
 3   AlcoholConsumption  1700 non-null   float64
 4   Smoking             1700 non-null   int64  
 5   GeneticRisk         1700 non-null   int64  
 6   PhysicalActivity    1700 non-null   float64
 7   Diabetes            1700 non-null   int64  
 8   Hypertension        1700 non-null   int64  
 9   LiverFunctionTest   1700 non-null   float64
 10  Diagnosis           1700 non-null   int64  
dtypes: float64(4), int64(7)
```

3. Data Imbalance 확인



Diagnosis	count
1	936
0	764

dtype: int64

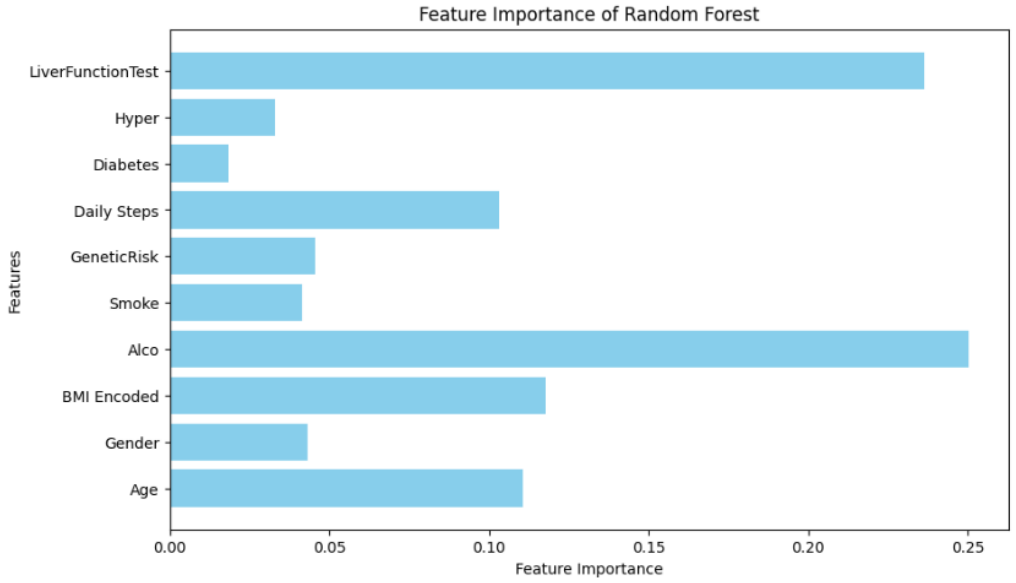
4. 데이터 전처리 과정
    - 데이터 컬럼 이름 정리: 원활한 작업을 위해 알아보기 쉬운 컬럼명 으로 변경 (예시: PhysicalActivity을 Daily Steps)
    - 데이터 공백 제거
    - Alco 컬럼의 값을 이진화

- Daily Steps 컬럼의 값을 신체 활동 기준으로 30분 걷기를 약 3000걸음으로 간주해서 계산, 1~5의 범주형 값으로 변환, int로 변환, Nan 값을 최소값(1) 으로 처리

Step6-2. 간암 위험군 분류 모델 (Liver_disease)

▼ 데이터 모델링

1. 사용 모델: Random Forest
2. 모델링 과정
    - 하이퍼 파라미터 튜닝: GridSearchCV을 사용하여 최적의 파라미터 선택
    - Feature Selection: Feature Importance를 계산 후 점수가 높은 것과 사용자에게 받기 쉬운 정보 인지를 고려하여 Feature를 선택



    - 선택된 Feature: 'Age', 'Gender', 'BMI Encoded', 'Alco', 'Smoke', 'Daily Steps', 'Hyper'
    - 모델 평가 지표: F1_Score = 0.72

3. 다른 분류 모델과 성능 비교 (F1_Score)
    - 데이터가 많지 않아 소규모 데이터셋에 적합한 LogisticRegression을 선택했고 변수 중요도 평가 가능한 Random Forest와 카테고리형 변수 자동 처리 가

능한 LightGBM을 사용했다. 이 중 F1_Score가 가장 높은 Random Forest을 최종 선택했다.