교하고 설명한다. 모델별 성능 결과에 대해 검토하며 당뇨병 예측 성능이 가장 높이 측정된 머신러닝 기법을 제안한다. 마지막으로 본 연구의 결론과 시사점을 도출한다.

## Ⅱ. 연구 방법

### 2.1. 데이터 설명

연구에 사용된 데이터 셋은 독일의 Frankfurt 병원의 당뇨병 데이터이다. 데이터는 총 2000개의 행과 9개의 열로 구성되어 있으며 행은 환자의 수치 데이터, 열은 혈당, 혈압, BMI, 당뇨병 여부 등을 가리킨다. 당뇨병 여부를 제외한 나머지 데이터는 연속적인 데이터이며 당뇨병 여부는 1과 0으로 구성되어 있다.

본 연구에서는 분류 모델을 이용하여 당뇨병 여부를 예측하고자 한다. 그래서 해당 데이터는 연구에 적합하다. 당뇨병 여부를 제외한 8개의 열을 feature로 사용할 것이고 이를 통해 당뇨병 여부를 예측하고자 한다.

### 2.2. 상관관계 분석

피어슨 상관계수로 각각의 feature가 당뇨병 여부와 어떤 관계를 갖고 있는지 분석하였다. 피어슨 상관계수는 두 변수간의 상관 관계를 계량화한 값이며 코시-슈바르츠 부등식에 의해 +1과 －1 사이의 값을 갖는다. 피어슨 상관계수는 일반적으로 절댓값이 0.7 이상이면 강한 상관관계, 0.3 이상이면 뚜렷한 상관관계, 0.1 이상이면 약한 상관관계 그리고 0.1 미만이면 무시해도 좋을 상관관계라고 해석된다.[11]

피어슨 상관계수로 데이터의 feature를 분석해 본 결과는 표 1과 같다.

**Table.1** Pearson Correlation Coefficient

| feature | *Pearson* Correlation Coefficient |
|---|---|
| Glucose | 0.458 |
| BMI | 0.277 |
| Age | 0.237 |
| Pregnancies | 0.224 |
| DiabetesPedigreeFunction | 0.155 |
| Insulin | 0.121 |
| SkinThickness | 0.076 |
| BloodPressure | 0.076 |

따라서 피어슨 상관계수의 값이 0.2 이상인 ‘Glucose’, ‘BMI’, ‘Age’, ‘Pregnancies’를 최종 feature로 사용할 것이다.

### 2.3. 데이터 전처리

본 연구에서는 이상치를 탐색하기 위해 IQR 방법을 사용하였다. IQR 방법이란 전체 데이터를 오름차순으로 정렬한 후 25%, 50%, 75%, 100%로 4등분한다. 여기서 25%와 75% 사이의 값을 IQR (Interquartile Range)라고 한다. 이상치는 다른 데이터들에 비해 아주 큰 값이나 작은 값을 갖는 데이터를 말하며 통계적으로는 1.5 IQR을 벗어나면 이상치로 판단한다.[12] 이상치 데이터가 포함될 경우 왜곡된 분석 결과를 얻게 되므로 정확한 결과의 도출을 위해 데이터 분석하기 전에 이상치를 제거하는 과정이 필수적이다.[13] 즉, 이상치 데이터는 모델의 성능에 악영향을 미친다. 따라서 본 연구 데이터에서는 총 2000개의 데이터 셋 중 106개의 데이터에서 이상치가 탐색되어 이를 제거하였다.

### 2.4. 모델 생성

모델 생성에 앞서 데이터 셋을 학습 데이터 60%, 검증 데이터 20%, 테스트 데이터 20%로 나누어 학습을 진행하였다. 학습 데이터는 모델을 생성하여 학습할 때 필요한 데이터이다. 검증 데이터는 생성한 모델이 적합한지 검증할 때 사용하며 테스트 데이터는 모델의 성능을 평가할 때 사용한다. 본 연구에서는 앞서 언급한 비율대로 데이터를 분할하였기에 학습데이터 1136개, 검증데이터와 테스트 데이터 각각 379개로 연구를 진행하였다.

Voting은 일반적으로 동일한 훈련 세트를 가지고 여러 모델을 훈련하는 방법을 의미한다. 따라서 Voting은 서로 다른 알고리즘이 도출해 낸 결과물에 대해 투표를 하는 방식이다. 또한, voting은 두 가지 방식이 있는데 결과물에 대한 최종 값을 투표하여 결정하는 ‘hard vote’와 결과물이 나올 확률값을 다 더해서 각각의 확률을 구한 뒤 최종값을 도출하는 ‘soft vote’가 있다. 본 연구에서는 soft vote를 이용하여 로지스틱 회귀모델(Logistic Regression)과 KNN을 이용하여 Voting 모델을 구현하였다.

Stacking Ensemble 모델은 다양한 알고리즘을 조합하여 구성할 수 있으며, 개별 모델이 예측한 데이터가 training set으로서 최종 모델에서 예측하는 데 쓰여 각 알고리즘의 장점을 취하면서 약점을 보완할 수 있다. 본