JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea. Inst. Inf. Commun. Eng.) Vol. 23, No. 1: 399~406, Mar. 2019

Stacking Ensemble 모델을 이용한 당뇨병 예측

정주호[1] · 이나은[2] · 김수민[3] · 서가은[4] · 오하영[5*]

Diabetes Prediction Using Stacking Model

Juho Jung[1] · Naeun Lee[1] · Sumin Kim[1] · Gaeun Seo[1] · Hayoung Oh [2*]

[1]Undergraduate Student, Applied Artifical Intelligence, Sungkyunkwan University, Seoul, 03063 Korea
[2]*Associate professor Professor, Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

요 약

최근 전 세계적으로 당뇨병 유발률이 증가함에 따라 다양한 머신러닝과 딥러닝 기술을 통해 당뇨병을 예측하려고 하는 연구가 이어지고 있다. 본 연구에서는 독일의 Frankfurt Hospital 데이터로 머신러닝 기법을 활용하여 당뇨병을 예측하는 모델을 제시한다. IQR(Interquartile Range) 기법을 이용한 이상치 처리와 피어슨 상관관계 분석을 적용하고 Decision Tree, Random Forest, Knn, SVM, 앙상블 기법인 XGBoost, Voting, Stacking로 모델별 당뇨병 예측 성능을 비교한다. 연구를 진행한 결과 Stacking ensemble 기법의 정확도가 98.75%로 가장 뛰어난 성능을 보였다. 따라서 해당 모델을 이용하여 현대 사회에 만연한 당뇨병을 정확히 예측하고 예방할 수 있다는 점에서 본 연구는 의의가 있다.

ABSTRACT

With the recent increase in diabetes incidence worldwide, research has been conducted to predict diabetes through various machine learning and deep learning technologies. In this work, we present a model for predicting diabetes using machine learning techniques with German Frankfurt Hospital data. We apply outlier handling using Interquartile Range (IQR) techniques and Pearson correlation and compare model-specific diabetes prediction performance with Decision Tree, Random Forest, Knn, SVM, Bayesian Network, ensemble techniques XGBoost, Voting, and Stacking. As a result of the study, the XGBoost technique showed the best performance with 97% accuracy. Therefore, this study is meaningful in that the model can be used to accurately predict and prevent diabetes prevalent in modern society.

키워드 : Stacking, Ensemble, 당뇨병 예측, 기계학습, IQR

Keywords : Stacking, Ensemble, Diabetes Prediction, Machine learning=, IQR(Interquartile Range)

Received 29 January 2019, **Revised** 29 March 2019, **Accepted** 21 April 2019
(출판사에서작성)

*** Corresponding Author:** Hayoung Oh (E-mail: hyoh79@gmail.com Tel:+82-2-583-8585)
Global Convergence, Sungkyunkwan University, Seoul, 03063 Korea

Open Access http://dx.doi.org/10.6109/jkiice.2019.23.1.399 pISSN:2234-4772

Ⅰ. 서 론

1.1 선행연구

머신러닝과 딥러닝은 최근 발전함에 따라 많은 산업의 관심을 받고 있다. 이러한 기술은 의료계에서도 많이 활용되고 있는데, 웨어러블 기기를 이용한 헬스케어, 과거의 데이터를 이용하여 초기 질병을 예측하는 알고리즘, 의료상담을 대신해 주는 챗봇 로봇 등이 있다. 본 논문에서는 이 중에서도 내분비질환 중의 하나인 당뇨병을 예측할 수 있는 머신러닝 모델을 제시한다.

머신러닝과 딥러닝을 이용하여 당뇨병을 예측하는 연구들은 많이 선행되었었다. 연구들은 주로 3가지 종류가 있었는데, 당뇨병 발병 예측 분야, 당뇨병 환자들의 혈당 예측 분야, 당뇨병 예측의 정확도를 높이기 위한 데이터 필터링 분야의 연구들이 진행되었다. 당뇨병 발병을 예측하는 분야 중에는 [1]연구가 있다. 이 연구에서는 pipeline을 적용한 머신러닝 알고리즘과 적용하지 않은 머신러닝 알고리즘의 정확도를 비교하였다. pipeline을 적용하지 않은 알고리즘 중에서는 logistic regression이 96%로 가장 높은 정확도를 보였고, pipeline을 적용한 알고리즘 중에서는 AdaBoost Classifier 알고리즘이 98.8%로 가장 높은 정확도를 보였다. 이를 통해 다양한 머신러닝 모델들의 예측 정확도를 확인할 수 있었다. [2]연구에서는 PIMA Indian Dataset를 이용해 Decision Tree, Artificial neural network, Deep Learning, Naive Fayes를 당뇨병 예측에 활용하였고, 실험 결과 4가지 모델들은 90~98% 사이의 정확도가 나왔다. 이를 통하여 4가지 모델들이 당뇨병 예측에 유의미한 효과를 보인다는 사실을 알 수 있었고, 이러한 모델을 애플리케이션이나 웹 사이트의 형태로 구현하면 당뇨병의 조기 발견에 큰 도움을 줄 수 있다고 예상한다. [3]연구에서는 PIMA Indian datset과 연구에서 사용되는 dataset을 같은 머신러닝 모델들로 실험을 한 후 정확도를 비교하였다. 실험 결과, Random Forest 기법의 정확도가 해당 연구의 datset으로 실험하였을 때 94.10%로 가장 높았고, Pima dataset으로 실험하였을 때에도 75.00%로 가장 높았다. 국내의 datset을 이용하여 진행했던 연구도 있었다. [4]연구에서는 기존 연구에서 정상인과 환자를 분류하는 데에만 초점을 두었던 점을 보완하기 위하여 시계열 데이터를 바탕으로 예측 모델을 만들었다. 데이터는 질병관리본부에서 조사한 '한국인유전체역학주사(KoGES)' 중 안성(농촌), 안산(도시) 지역사회 코호트 자료를 사용하였고, 그중에서도 혈당, 당화혈색소, 혈액 요소질소, 등의 연속형 변수만을 이용하였다. 기존의 연구들과 비교하기 위해 K-NN, SVM, LR 모델들의 정확도와 정확도를 비교하였고, 실험 결과 RNN 방법으로 구축한 모델이 0.92로 정확도가 가장 높았다. 그러나 해당 연구는 지도학습의 형태이기 때문에 같은 데이터 형태를 대상으로 했을 때만 유의미하다는 한계점이 있다. 시계열 데이터를 이용한 연구 중에는 당뇨병 발병 예측이 아닌 당뇨병 입원 환자의 혈당을 예측하는 [5]연구에서는 LSTM 신경망과 연속 혈당 측정 데이터를 이용하여 환자의 30분 또는 60분 후의 혈당값을 예측하는 모델을 제시했다. 실험 결과 예측 성능이 가장 우수한 경우는 당화 혈색소를 변수로 입력한 모델이 가장 우수하였고, [6]기존의 Bi-LSTM 기반의 신경망 모델을 참조해 구현한 모델보다 정확도가 향상되었음을 보여주었다. 그러나 dataset의 크기가 작아 데이터에 노이즈가 너무 많다는 한계점이 있다. 당뇨병 환자들에게 중요한 저혈당 상태를 예측하는 모델을 제안하는 [7]연구에서는 CNN과 RNN을 접목한 CRNN을 사용하였다. 연구에서는 연속혈당 측정기를 이용하여 얻은 데이터를 전처리하고 알고리즘을 작동시켜 예측값을 구한 후, 환자와 데이터베이스에 전송하는 구조를 사용하였다. 이렇게 앞서 제시했던 당뇨병 발병 예측 모델과 혈당 예측 모델의 정확도를 높이기 위한 [8]연구에서는 인공신경망 모델이 당뇨병을 진단율이 높음을 실험을 통해 확인한 후 인공신경망의 구조를 어떻게 설정해야 정확도를 높이는 데 기여하는지에 관한 연구를 했다. 결과적으로, 데이터의 78%를 Training set으로 사용하고 22%를 Test set으로 사용한 것이 가장 높은 진단율을 도출하였으며, 히든 레이어를 2개 사용하고 각 히든 레이어에 노드를 8개, 4개 사용한 것이 가장 높은 86.98%의 진단율을 끌어냈다. 이 연구에서는 데이터의 수가 너무 적고 데이터에 대한 attribute 수도 너무 적어서 데이터의 빈값들을 채워도 정확도를 더 향상할 수는 없었지만, 나머지 약 13%의 부분을 의사와 협력한다면 충분히 보완할 수 있을 것으로 예상된다.

1.2 연구 동향

전 세계적으로 당뇨병 발생률이 증가함에 따라 다양

Table.1 Existing studies

	Name of Thesis	Published in	Research Topics	Characteristics	Dataset
1	Development of T2DM Prediction Model Using RNN	2019, Journal of Digital Convergence Vol. 17. No. 8, pp. 249-255	Propose a model using RNN to improve diabetes predictive diagnostic accuracy.	- Create a prediction model for diabetes using RNN by learning all the data over time.	Korean Genetic Survey Community Cohort (Ansan-Ansung) Data
2	Diabetes Prediction using Machine Learning Algorithms	2019, Aishwarya Mujumdar et al. / Procedia Computer Science 165, 292–299	Propose pipeline model to improve diabetes prediction accuracy.	- To predict diabetes with the highest accuracy logistic regression technique.	-
3	The Implemetation of Artificial Neural Network Model for Improving the Diagnosis Accuracy of Type 2 Diabetes	2018, Proceedings of Symposium of the Korean Institute of communications and Information Sciences , 849-850(2 pages)	Proposal of ANN-based Diabetes Diagnosis Model to Improve Diabetes Diagnosis Rate	- ANN, Bayesian network, logistic, and SVM models are used to find high accuracy diabetes prediction models. - Propose the ANN model with the highest accuracy as the diabetes prediction model.	Using Pima Indians data, data on 768 adult women living in Arizona, Phoenix, USA
4	Short Term Glucose and Hypoglycemia Prediction Using CGM and Convolutional Recurrent Neural Network	2020, The Korean Institute of Information Scientists and Engineers, 1556-1557 (2 pages)	Based on CGM blood glucose measurement data, propose a model that can proactively prevent hypoglycemia using CRNN.	- Propose a hypotensive prediction model using CRNN grafting CNN and RNN among neural network models.	-
5	A Prediction Algorithm of Hypoglycemia using Electrocardiogram based on Support Vector Machine	2020, The Institute of Electronics and Information Engineers , 1613-1615 (3 pages)	Propose SVM-based model to predict hypoglycemia based on data obtained using ECG	- Blood glucose data is obtained using electrocardiogram on behalf of blood collection or CGMS devices and hypoglycemia is predicted using SVM-based models.	D1NAMO Dataset (20 normal people and 9 Type-1 diabetics for 4 days)
6	Prediction of Blood Glucose in Diabetic Inpatients Using LSTM Neural Network	2020, Journal of KIISE 47(12), 1120-1125 (6 pages)	Propose LSTM neural network model to predict future blood glucose based on CGM blood glucose measurement data	- Propose prediction model is designed based on LSTM neural networks suitable for time series information prediction, such as changes in blood sugar.	16 patients at Soonchunhyang University Chanan Hospital
7	Convolutional Neural Network Based Glucose Level Prediction Using Electrocardiogram	2020, The Institute of Electronics and Information Engineers, 1616-1618 (3 pages)	Propose findings to predict diabetes using convolutional neural networks	- Based on real-time measured blood glucose data obtained by wearable equipment, hyperglycemia, hypoglycemia, and normality of diabetic patients are predicted using convolutional neural networks.	D1NAMO dataset (Fabien Dubosson et al, Informatics in Medicine Unlocked, Volume 13, pp.92-100, 2018.)
8	Improving the Accuracy of Diabetes Prediction using Filtering Techniques	2017, The Institute of Electronics and Information Engineers, 983-986(4 pages)	Propose filtering techniques to improve prediction accuracy of classification techniques used for diabetes prediction	- Using techniques such as Logic Regression, propose a filtering technique model to improve diabetes prediction accuracy.	Using the Pima Indians dataset (data from the National Academy on Diabetes Digestion- Kidney Disease collected by NIDDK in the United States)
9	2020_IS_Dual stack network for detecting diabetes mellitus and chronic kidney disease-1	2020, Information Sciences 547 (2021) 945-962	Propose model with DsNet predictable for diabetes and chronic kidney disease	- Analysis of human face images. - Predicted diabetes and chronic kidney disease using DsNet model with analyzed facial images.	Hong Kong Foundation for Research and Development in Diabetes, Prince of Wales Hospital , Guangdong Provincial Hospital of Traditional Chinese Medicine.
10	An improved noninvasive method to detect Diabetes Mellitus using the Probabilistic Collaborative Representation based Classifier	2018, Information Sciences 467 (2018) 477-488	Propose model using facial key block color feature based on ProCRC for diabetes prediction	- Using human face image. - Based on ProCRC, diabetes is predicted using facial key block color feature.	Prince of Wales Hospital, Hong Kong SAR, Guangdong Provincial TCM Hospital, Guangdong

한 머신러닝 기법을 이용하여 당뇨병을 예측하기 위한 연구가 이어지고 있다. 선행 연구를 보면 당뇨병 예측을 위해 머신러닝 기법인 Decision Tree, Random Forest, Neural Networks Model(CNN, RNN, KNN, ANN), Logistic Regression, SVM과 같은 모델들이 주로 사용된다. 구체적으로 당뇨병 예측 진단 정확도 향상을 위한 머신러닝 기법을 제안할 때 선행 연구에서는 이와 같은 머신러닝 기법들이 주로 사용된다. 본 연구에서는 앙상블 기법인 XGBoost, Voting, Stacking과 같은 모델들을 당뇨병을 예측하기 위해 추가로 사용한다. 선행 연구는 표 1을 통해 확인할 수 있다.

선행 연구에서는 머신러닝 기법을 사용하기 위해 Pima Indians 데이터를 주로 사용한다. Pima Indians 데이터는 미국의 NIDDK에서 수집한 당뇨병 소화-신장 질환에 대한 국가 학회의 데이터이다[9]. 그러나 Pima Indians 데이터는 당뇨병 속성변수들을 구분하는 기준이 명확하지 않아 선행 연구에서는 절반의 데이터를 제거했다[9]. 머신러닝 기법을 사용하기에 실질적으로 적합한 데이터 개수가 적은 것이 Pima Indians 데이터의 단점이다[10]. 본 연구에서는 독일의 Frankfurt Hospital의 2,000개 데이터를 추가로 활용하여 기존의 Pima Indians 데이터가 가지고 있는 단점을 보완하여 머신러닝 기법을 사용한다.

이러한 연구 동향에 따라, 본 연구의 구성은 다음과 같다. 독일의 Franfurt Hospital 데이터와 미국의 Pima Indians 데이터를 가지고 IQR 기법을 이용한 이상치 처리와 피어슨 상관관계 분석을 적용하고 설명한다. 분석한 데이터로 Decision Tree, Random Forest, KNN, SVM 기법과 선행 연구에서 주로 사용하지 않았던 앙상블 기법인 XGBoost, Voting, Stacking 머신러닝 기법을 사용하여 모델별 당뇨병 예측 성능을 비교하고 설명한다. 모델별 성능 결과에 대해 검토하며 당뇨병 예측 성능이 가장 높이 측정된 머신러닝 기법을 제안한다. 마지막으로 본 연구의 결론과 시사점을 도출한다.

Ⅲ. 연구 방법

3.1 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.

3.2 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 2와 같다.

Table.2 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.

3.3 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.

3.4 모델 생성

모델 생성에 앞서 데이터 셋을 학습 데이터 60%, 검증데이터 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.

3.4.1 Decision Tree (의사결정트리)

의사결정트리는 지도학습기법의 한 유형으로 각 단계마다 기준에 따라 데이터를 분류한다. 본 연구 데이터로 의사결정트리 모델을 생성한 결과는 그림 1과 같다.

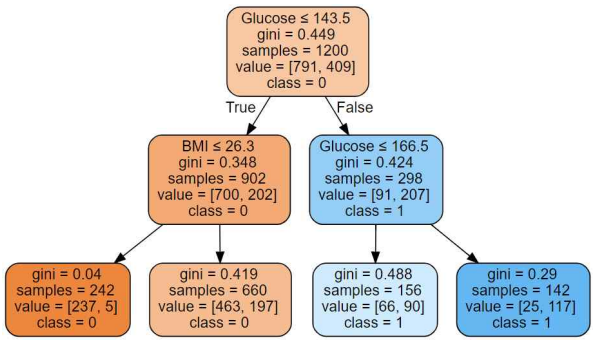

Fig.1 Decision Tree

그림 1에서 가장 위에 있는 노드를 root 노드라고 한

다. 그리고 root 노드 아래의 노드들은 leaf 노드라고 한다. 본 연구에서는 당뇨병 여부를 예측하는 것이기 때문에 binary split 기반의 지니 계수를 사용하였다. 지니 계수는 CART 알고리즘으로 산출되며 불순도를 계산하는 방법 중 하나이다. 그래서 이 값이 적을수록 분류가 잘 된 것으로 판단할 수 있다. 지니 계수의 수식은 아래와 같다. [14]

$$Gini(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (1)$$

그림 1에서 각 노드에 class가 부여된 것을 볼 수 있다. class가 0일 때는 당뇨병에 걸리지 않은 데이터 class가 1일 때는 당뇨병에 걸린 데이터이다. 따라서 의사결정트리로 데이터를 당뇨병 여부에 따라 예측하여 분류하였음을 할 수 있다.

3.4.2 Random Forest (랜덤 포레스트)

랜덤 포레스트는 의사결정트리 모형을 기저로 하며 무작위성을 최대로 부여함으로써 예측오차를 줄인다. 그런데 의사결정트리에서는 얼마나 나무를 성장시킬지, 그리고 성장한 나무를 어떤 식으로 가지치기할지 등을 연구자가 판단해야 한다는 어려움이 있다. 또한, 의사결정 트리는 모형 설명력이 높지만, 예측력이 떨어지는 편이다. 그래서 데이터 셋을 앞서 언급했듯이 훈련 데이터, 검증 데이터, 테스트 데이터로 분할하여 학습을 진행하는 것이 의사결정트리 모형에서의 일반적인 절차이다.[15] 랜덤 포레스트는 여러 개의 의사결정트리로 학습하여 의사결정트리의 예측력이 떨어지는 단점을 보완하기 위한 모델이다.

본 연구에서는 파이썬 scikit-learn 머신러닝 라이브러리를 이용하여 랜덤 포레스트 모델을 구축하였다.

3.4.3 KNN (K-Nearest Neighbor)

KNN 알고리즘은 k개의 이웃을 두고 각 데이터가 k개의 이웃 간의 거리를 측정하여 가장 가까운 이웃에 속하도록 하여 분류하는 알고리즘이다. 여기서 흔히 사용되는 거리 척도는 유클리드 거리이다.

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2}$$

$$P = (p_1, p_2, \ldots, p_n) \quad (2)$$
$$Q = (q_1, q_2, \ldots, q_n)$$

KNN 알고리즘의 특징은 방법이 간단하고 학습 단계에서 최소한의 처리를 한다는 것이다. 또한, KNN의 정확도는 우수하지만, 분류단계에서 실행 속도가 매우 느린 단점을 갖고 있다. [16]

3.4.4 SVM(Support Vector Machine)

SVM은 AT&T Bell 연구소의 Vapnik에 의해서 제안된 새로운 분류 및 회귀 기술이다. SVM은 클래스 별 데이터를 가장 잘 구분하는 최적의 하이퍼 플레인(Hyperplane)을 찾는 알고리즘이다. 하이퍼 플레인이란 데이터가 분포되어 있는 n차원의 feature 공간을 두 공간으로 나누는 n-1차원의 평면이다. 하이퍼 플레인과 가장 가까운 두 공간 내의 데이터와 하이퍼 플레인과의 거리인 margin을 최대화하는 구분 평면(Separating Hyperplane)을 학습하는 것이며 이로 인해 좋은 일반화 성능을 보인다. 여기서 하이퍼 플레인과 가장 가까우 데이터 포인터를 서포트 벡터 (Support Vector)라고 한다. 그러나 SVM은 많은 시간과 저장 공간이 소요되며 실제 구현에서 근사화된 알고리즘을 적용하여 이를 해결할 수 있지만, 분류 성능이 떨어지게 된다는 단점이 있다.[17]

3.4.5 Xgboost(extreme gradient boosting)

앙상블 학습(ensemble learning)은 머신러닝 기법의 예측력을 높이기 위해 복수의 모델을 활용하는 방법이다. 3.4.2의 랜덤 포레스트도 의사결정 트리를 활용한 대표적인 앙상블 학습 모델이다. 3.4.5의 부스팅(boosting)을 활용한 Xgboost, 3.4.6의 Voting 그리고 3.4.7의 Stacking도 앙상블 학습 모델이다. 부스팅은 복수의 분류기 가운데 예측력이 상대적으로 낮은 분류기들을 결합하여 예측력이 상대적으로 높은 분류기로 바꿈으로써, 전체 모델 내 분산을 줄이고, 예측력을 높이는 방법이라고 할 수 있다.[18]

본 연구에서는 grid search 방식을 이용하여 xgboost 모델의 하이퍼 파라미터를 찾았다. 하이퍼 파라미터란 모델에 사람이 초기에 넣어주는 변수를 의미한다. 하이퍼 파라미터의 종류로는 Learning Rate, Cost function, Traning Loop 등이 있다. grid search 방식은 모델에 적

용하고자 하는 하이퍼 파라미터 값들을 미리 지정하여 교차 검증을 통해 최적의 파라미터 값의 조합을 찾아내는 방식이다. [19] grid search로 도출된 최적의 파라미터의 값은 Learning Rate 0.01, Max_depth 10, Reg_alph 0, Reg_lambda 0.5이다.

3.4.6 Voting

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.[20] 본 연구에서는 soft vote를 이용하여 로지스틱 회귀모델(Logistic Regression)과 KNN을 이용하여 Voting 모델을 구현하였다.

3.4.7 Stacking

Stacking Ensemble 모델은 다양한 알고리즘을 조합하여 구성할 수 있으며, 이러한 조합을 통해서 각 알고리즘의 장점을 취하면서 약점을 보완할 수 있다.[21] 그래서 일반적으로 Stacking 모델은 다른 단일 모델보다 성능이 좋다. Stacking은 개별 모델이 예측한 데이터가 training set으로서 최종 모델에서 예측하는 데 쓰인다. 본 연구에서는 개별 모델로 SVM, 랜덤 포레스트, 로지스틱 회귀 그리고 최종 모델로 LightGBM을 사용했다.

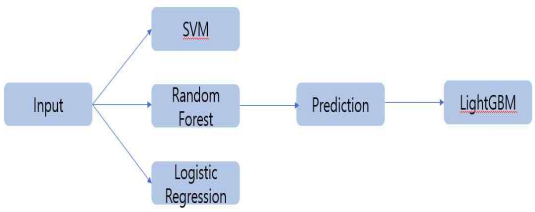

Fig.2 Stacking Structure

Algorithm 1 Training an Stacking

Require : Train data consisting of X and y.

```
X <- preprocessed feature
y <- label(2 classes)

#Train each model and predict y
kernel = "linear“
classifier <- SVC(formula, data, kernel)
svm_pred = classifier.predict(X)

library(RandomForest)
classifier   <- RandomForest(formula, data)
rf_pred = classifier.predict(X)

library(Logistic Regression)
classifier   <- Logistic Regression(formula, data)
lr_pred = classifier.predict(X)

#Create new array data from predicted y by each models
#Each of the anticipated array is added based on the line
new_array = array(svm_pred, rf_pred, lr_pred)

#Transpose the new_array
new_array = Transpose(new_array)

#Train LGBM model
library(RandomForest)
classifier   <- RandomForest(formula, n_estimator, random_state, data)
```

Fig.3 Stacking Pseudocode

Ⅳ. 실험 및 결과

4.1 모델 학습 데이터 및 평가 지표

모델 학습에 사용한 데이터셋은 총 3개로, Pima indians 데이터셋과 이상치를 처리한 데이터 셋과, Hospital Frankfurt, Germany 데이터셋이다. 각각 과거 환자의 기록 데이터를 이용해 새로운 환자의 당뇨병 발생 여부를 예측하는 것이 목표이고, 나아가 당뇨병의 걸릴 확률을 구하는 것을 최종 목표로 한다. 해당 데이터가 실제 환자의 상태를 나타내는 수치이기 때문에 각 환자마다 다른 특이점을 고려하여 평균에서 극단적으로 높은 이상치들은 제외하고 실험을 진행하였다.

모델 학습의 결과를 평가하는 지표로는 Accuracy와 F1-score를 사용했다.

Table.3 Evaluation Indicators

	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

Accuracy는 전체 데이터 중에서 제대로 분류된 데이터의 비율을 나타낸 것으로 다음과 같은 수식으로 나타낼 수 있다.

$$\frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives} \quad (3)$$

F1-Score은 precision과 recall지표의 조화평균을 의미하여 다음과 같은 수식으로 나타 낼 수 있다.

$$2 \times \frac{\mathrm{Pr}ecision \times Recall}{\mathrm{Pr}ecisiion + Recall} \quad (4)$$

이때 precision은 모델이 True로 예측한 데이터 중 실제로 True인 데이터의 수를 의미하고, 이는 다음과 같은 수식으로 나타낸다.

$$\frac{TruePositives}{TrueNegatives + FalsePositives} \quad (5)$$

또한 Recall은 실제로 True인 데이터를 모델이 True라고 인식한 데이터의 수를 의미하고, 이는 다음과 같은 수식으로 나타낼 수 있다.

$$\frac{TruePositives}{TruePositives + FalseNegatives} \quad (6)$$

4.2 모델 학습

앞서 준비한 3개의 데이터셋을 이용해 각 데이터셋마다 피어슨 상관관계 분석을 적용하고 Decision Tree, Random Forest, Knn, SVM, 앙상블 기법인 XGBoost, Voting, Stacking로 총 8개의 모델로 당뇨병을 예측했다. 각 모델별 하이퍼파라미터를 구하기 위해 하이퍼파라미터 최적화 방식으로 Grid Search, Bayesian Optimization, Random Search 기법을 혼용하여 각 모델별 하이퍼파라미터를 최적화했다.

4.3 모델 학습 결과

모델 학습 결과를 바탕으로, 최적의 파라미터를 산출하였다. 그리고 검증 데이터셋으로 실험한 결과는 표 5와 같다.

Table.4 F1-Scores Results

	XGBoost	LightGBM	RandomForest
Pima	0.77	0.8376	0.7662
Pima_IQR	0.7573	0.7756	0.7671
Frankfurt	0.9725	0.9875	0.9575
Frankfurt_IQR	0.9288	0.9367	0.9420

XGBoost의 경우 Randomized Grid Search 기법과 Bayesian Optimization 기법을 혼용해 Frankfurt 데이터셋을 기반으로 하이퍼파라미터를 최적화했고 이는 표 6의 결과로 확인 할 수 있다. 표 5에서 볼 수 있듯, 해당 데이터셋들은 결측지가 존재하지 않고 이상치만 존재함으로, 이상치를 처리하지 않은 Frankfurt 데이터셋이 가장 좋았으며, Pima indians 데이터셋을 IQR 기법으로 이상치 처리한 결과가 가장 좋지 않았다. 이는 Pima indians 데이터셋이 768개로 적은 데이터 수를 가지고 있음과 동시에 이상치처리로 인해 데이터 수가 더 적어졌기 때문으로 보인다.

LightGBM의 경우, boosting type을 'gbdt(Gradient Boosting Tree)' 로 고정하였다. 위에서 진행한 XGBoost과 동일한 방법으로 LGBM의 최적화를 진행하여 도출한 최적 하이퍼파라미터는 표 7과 같다. 표 5의 두번째 열 값은 해당 데이터셋에 대한 LightGBM의 정확도 값으로, 아무 처리도 하지 않은 Frankfurt 데이터셋, Frankfurt_IQR, Pima indians, Pima_IQR순으로 정확도가 낮아짐을 알 수 있다.

마지막으로 Random Forest로 예측한 결과는 표5의 세 번째 열에서 확인할 수 있다. 이 모델도 위의 두 모델과 같은 방법의 하이퍼파라미터 최적화 과정을 통해

표 8과 같이 최적의 하이퍼파라미터를 도출했다. 위의 두 모델과 같이 이상치처리를 하지 않은 Frankfurt 데이터셋에서 가장 높은 정확도를 보이며, 이번에는 이상치 처리를 하지 않은 Pima indians 데이터셋에서 가장 낮은 정확도를 보임을 알 수 있다.

Table.5 XGBoost Hyperparameters

max_depth	10
n_estimators	100
reg_alpha	0
reg_lambda	0.5
learning_rate	0.01

Table.6 LightGBM Hyperparameters

max_depth	5
min_data_in_leaf	10
num_iteration	1228
num_leaves	200
min_gain_to_split	2
bagging_fraction	0.698
feature_fraction	0.555

Table.7 RandomForest Hyperparameters

n_estimators	200
max_depth	12
min_sample_leaf	1
min_sample_split	2

표 5에서 처럼 전체 12개의 모델의 F1-Score에서 볼 수 있듯이 가장 좋은 성능을 보인 모델은 Frankfurt-LightGBM 데이터 셋, Frankfurt-XGBoost 데이터셋이다. 다시 말해, 전체 환자의 데이터 값을 보존한 상태로 LightGBM, XGBoost으로 학습한 모델을 최종 모델로 선정했다. 그림 11은 전체 모델별로 최적의 파라미터값을 넣은 결과를 비교한 그래프이다.

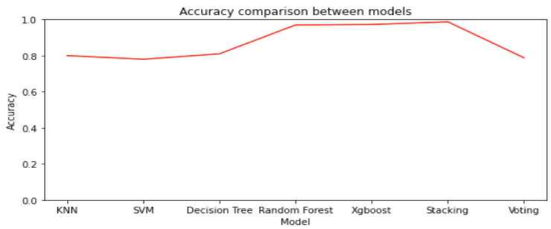

Fig.4 Accuracy Graph for the Models: Knn, SVM, Decision Tree, Random Forest, XGBoost, LightGBM, Voting

Ⅴ. 결 론

본 논문은 IQR이상치 처리 기법과 피어슨 상관계수를 적용해 환자별 당뇨병 발생 여부를 예측했다. 기존에 많은 연구에서 사용된 Pima Indians Diabetes 데이터와 새롭게 구한 Hosipital Frankfurt, Germany 당뇨병 환자 데이터에 IQR(Interquartile Range) 기법을 적용해 총 4가지 데이터셋을 생성했다. 이 4가지 데이터셋과 Decision Tree, Random Forest, Knn, SVM, 앙상블 기법인 XGBoost, Voting, Stacking 모델을 이용해 예측을 진행했다. 이중 정확도가 가장 높은 세 모델인 XGBoost, Voting, Stacking 모델을 선정하여 하이퍼파라미터를 최적화하여 비교 분석하였다. 결과로는 이상치 처리를 진행하지 않은 Stacking 모델로 예측한 결과의 accuracy와 f1-score 값이 가장 높아 이 모델을 본 연구의 최종 모델로 선정했다.

XGBoost, Voting, Stacking 모델 모두 IQR(Interquartile Range) 기법으로 이상치처리를 진행하지 않은 데이터셋의 정확도가 높다는 것을 알 수 있었다. 이는 데이터가 환자의 신체 상태라는 점에서 이상치를 처리하게 되면 연령대별로 다양성이 훼손되어 편향적인 데이터값을 갖게 되기 때문이라 판단된다. 즉, 당뇨병의 경우 특정 연령대(30 – 40대) 에서 많이 걸리는데 IQR기법을 사용하게 되면 해당 연령대에 가중치를 부여하기 때문에 60대 이후의 환자 데이터는 이상치로 처리하게 된다. 그러나 Frankfurt 데이터셋을 살펴보면 60,70,80대의 환자 수가 전체 환자 수의 20%로 많은 데이터들이 이상치로 분류되어 소거하기 때문에 IQR 기법을 사용하지 않고 보존된 데이터셋을 사용할 때 정확도가 높음을 확인할 수 있다. 본 연구가 오늘날 문제가 되는 당뇨병 예측의 정확도를 높였다는 점에서 의의를 갖지만, 본 연구에서 사용한 데이터셋의 수가 2000개로 통상적으로 머신러닝 연구에서 사용되는 데이터 수 보다 적다는 한계점을 갖는다.

REFERENCES

[1] A. Mujumdar and V. Vaidehi, “Diabetes Prediction using Machine Learning Algorithms”, *Procedia Computer Science* , vol. 165, pp. 292 – 299, 2019.

[2] H. Naz and S. Ahuja, “Deep learning approach for diabetes prediction using PIMA Indian dataset”, *Journal of Diabetes & Metabolic Disorders,* vol. 19, pp. 391 – 403, 2020.

[3] N. P. Tigga and S. Grag, “Prediction of Type 2 Diabetes using Machine Learning Classification Methods”, *Procedia Computer Science,* vol. 167, pp. 706 – 716.2020.

[4] J. S. Jang, M. J. Lee and T. R. Lee, “Development of T2DM Prediction Model Using RNN”, *Journal of Digital Convergence,* Vol. 17. No. 8, pp. 249-255, 2019

[5] S. H. Kim, H. B. Lee, S. W. Jeon, D. Y. Kim and S. J. Lee, “Prediction of Blood Glucose in Diabetic Inpatients Using LSTM Neural Network”, *Journal of KIISE* vol. 47, no. 12, pp. 1120-1125,. 2020.

[6] Q. Sun, M. V, Jankovie, L. Bally and S.G. Mougiakakou, “Predicting Blood Glucose with an LSTM and Bi-LSTM Based Deep Neural Network”, *2018 14th Symposium on Neural Networks and Applications IEEE*, pp. 1-5, 2018.

[7] C. H. Lim, H. S. Kang, Y. S. Lee, H. J. Lee and T. H. Eom, “Short Term Glucose and Hypoglycemia Prediction Using CGM and Convolutional Recurrent Neural Network”, *The Korean Institute of Information Scientists and Engineers,* pp. 1556-1557, .2020.

[8] K. B. Won and M. K. Kim, “The Implemetation of Artificial Neural Network Model for Improving the Diagnosis Accuracy of Type 2 Diabetes”, *Proceedings of Symposium of the Korean Institute of communications and Information Sciences,* pp. 849-850, 2018.

[9] S. H. Lee, T. H. Ahn, S. W. Song and Y. G. Jung, “Improving the Accuracy of Diabetes Prediction using Filtering Techniques”, *The Institute of Electronics and Information Engineers,* pp. 983-986, 2017.

[10] Y. R, Lee, E. S. Kim, J. U. Park, Y. W. Kim, H. S. Choi, and K. J. Lee, “A Prediction Algorithm of Hypoglycemia using Electrocardiogram based on Support Vector Machine”, *The Institute of Electronics and Information Engineers,* pp. 1613-1615, 2020.

[11] Documents for Peason Coefficient.[Internet]. Available:https://support.minitab.com/ko-kr/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/

[12] Documents for IQR.[Internet]. Available:https://bookdown.org/yuaye_kt/RTIPS/data-prep-2.html

[13] Y. J. Hong, E. H. Na, Y. H. Jung and Y. U. Kim, “Distributed Processing Environment for Outlier Removal to Analyze Big Data,” *Journal of Korean Computer Information Society Korean Computer Information Society*, vol. 24, no. 2, pp. 73-74, Jul. 2016.

[14] K. B. Park, “Possibility of Learning AI Decision Tree Algorithm in Social Studies Education,” *Korean journal of elementary education*, vol. 31, no. 4, pp. 133-143, 2020.

[15] J. E. Yoo, “Random Forest, ” *Education Evaluation Study*, vol. 28, no. 2, pp. 427-448, Jun. 2015.
ㄴ

[16] J. M Lee, “Artificial Intelligence : An Efficient kNN Algorithm,” *The KIPS Transactions : Part B*, vol. 11, no. 7, pp. 849-854, 2016.

[17] H. M. Je and S. Y. Bang, “Improving SVM Classification by Constructing Ensemble,” *Journal of the Information Society: Software and Application*, vol. 30, no. 3·4, pp. 251-258, Apr. 2003

[18] J. H. Han, D. G. Go and H. J. Choi, “Predicting and Analyzing Factors Affecting Financial Stress of Household using Machine Learning: Application of XGBoost,” *Korea Consumer Association*, vol. 30, no. 2, pp. 21-43, 2019.

[19] Documents for Grid Search.[Internet]. Available:https://databuzz-team.github.io/2018/12/05/hyperparameter-setting/

[20] Documents for Voting. [Internet]. Available:https://velog.io/@guns/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-%EC%8A%A4%ED%84%B0%EB%94%94-%EC%95%99%EC%83%81%EB%B8%94-Ensem

ble-Voting

[21] H. N. Eom, J. S. Kim and S. O. Choi, "Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model", *Journal of intelligence and information systems*, vol. 26, no. 2, pp. 105-129, 2020.