ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://dx.doi.org/10.24225/jkaia.2024.2.2.1

치매에 대한 예측 머신러닝데이터 관점에서

Prediction of Dementia from Machine Learning Data

Gaeun KIM[1], Myungae CHUNG[2], Kyunga KIM[3]

Received: November 11, 2024. Revised: December 02, 2024. Accepted: December 13, 2024.

Abstract

The main purpose of the study is to predict mental health status, that is, Alzheimer's diagnosis, by analyzing factors tailored to each data set using machine learning models. This study aims to find more relevant factors by analyzing unique factors existing in each data set. To this end, this study used decision tree models, random forest models, KNN models, SVM models, artificial neural network models, naive Bayesian models, logistic regression analysis, and XG boost models among the developed machine learning models. In the process of training the model using medical data, we went through trial and error, such as increasing variable values to increase the model performance index value that determines the degree of learning. In addition, we did not end this study by comparing the performance of the models but ended the study by finding out which variables are closely related to dementia prediction and their weights. These results can provide a foundation for what approach is needed when processing medical data. In addition, it will be helpful for research that predicts results through medical data and finds out which variables are closely related.

Keywords : Alsheimers, Machine Learning, Predicition, Model

Major Classification Code : Artificial Intelligence, Machine Learning, Predicition Analysis

1. Introduction

현재 대한민국을 비롯해, 전세계적으로 치매는 노인들에게 치명적인 병으로 인식되고 있다. 한 대학병원에서 조사한 데이터를 바탕으로 보면, 일반적으로 치매의 유병률은 65 세 이상 인구의 약 5~10%이고, 85 세 이상의 인구에서는 약 47%에 이르는 것으로 추정된다. 이처럼 치매는 특히 노인들에게 더 많이 나타나는 특징이 있다.

더구나 대한민국은 현재 고령화 사회이며, 고령화 속도는 OECD 평균에 비해 2 배 이상 빠르게 진행되고 있다. 통계청 자료에 따르면, 22 년 기준 대한민국의 65 세 이상 고령인구는 전체 인구의 17.5%를 차지하고, 향후, 25 년에는 20.6%를 기록해 초고령사회로 진입할 것이라고 한다. 또한, 23 년의 한 기사에 따르면, 대한민국의 치매 환자는 노인 인구 대비 10.2%를 차지하고, 22 년 3 월 기준 전국 추정 치매 환자 수는 약 88 만 명으로 추정 치매율은 10.3%에 이른다고 한다. 치매 환자 수는 꾸준하게 증가하여, 24 년에는 100 만 명, 30 년에는 135 만 명, 39 년에는 200 만 명, 40 년에는 317 만 명, 50 년에는 300 만 명이 넘을 것으로 예측된다.

1 First Author. undergraduate student, Department of Big Data Medical Convergence, Department of Bio-Convergence, Eulji University, Republic of Korea, Email: kae6121@naver.com
2 Second Author. professor, Department of Big Data Medical Convergence, Department of Bio-Convergence, Eulji University, Republic of Korea, Email: machung@eulji.ac.kr
3 Corresponding Author. MIIC, Korea, Email: kyungakim@naver.com

치매 유형에는 알츠하이머, 혈관성 치매, 기타 치매들이 존재하는데, 알츠하이머가 전체 치매의 76.04%를 차지하고, 혈관성 치매가 8.57%, 기타가 15.37%를 차지한다.

이러한 배경을 고려해, 본 연구는 60~90 세 사이의 알츠하이머 진단을 받지 않은 1389 명과 알츠하이머 진단을 받은 760 명의 연령, 교육 수준, 흡연, BMI, 알코올 소비량 등에 관한 데이터를 가지고 분석하였다. 이 연구는 머신 러닝 모델을 활용하여 알츠하이머에 영향을 미치는 다양한 요인을 조사하고, 이를 통해 알츠하이머 예방에 기여하는 것을 목표로 합니다. 나아가 알츠하이머의 원인과 관련된 변수들을 파악하여 치매 환자들을 위한 효과적인 지원 방안을 모색하고자 합니다.

2. The Main Text of a Thesis

2.1. The Purpose of the Study

연구의 핵심 목표는 주어진 데이터셋에 맞는 적절한 모델을 분석하고 데이터의 복잡한 관계와 특성을 명확하게 하는 것이다. 또한 예측 모델의 성능을 비교 분석해 제대로 학습이 된 모델이 어떤 모델인지 알아보는 것이다.

2.2. Research Data

해당 데이터는 Kaggle 에서 가져온 데이터로, 환자 2194 명의 광범위한 건강 정보가 담겨있다. 이 데이터셋은 크게 인구 통계적 세부 정보, 라이프 스타일 요인, 병력, 임상 측정, 인지 및 기능 평가, 증상 및 알츠하이머병 진단의 변수들이 있다.

2.3. Selection of Variables

해당 데이터는 크게 환자 ID, 인구 통계 세부 정보, 라이프 스타일 요인, 병력, 임상 측정, 인지 및 기능 평가, 증상, 진단 정보, 기밀정보의 정보들이 있다. 세부적으로, 인구 통계 세부 정보에는 연령, 성별, 민족성, 교육 수준의 내용이 있고, 라이프 스타일 요인에는 BMI, 흡연, 알코올 소비량, 신체활동, 다이어트 품질 점수, 수면의 질 점수가 포함된다. 병력에는 가족병력 알츠하이머병, 심혈관 질환, 당뇨병, 우울증, 두부 손상, 고혈압의 정보들이 포함되어 있고, 임상 측정에는 수축기 혈압, 이완기 혈압, 콜레스테롤 총량, 콜레스테롤 LDL (저밀도 지단백 콜레스테롤), 콜레스테롤 HDL(고밀도 지단백 콜레스테롤), 콜레스테롤, 중성지방이 포함된다. 인지 및 기능 평가에는 MMSE 간이 정신 상태 검사 점수로 인지 장애를 판단하는 내용, 기능 평가는 기능 장애를 판단하는 내용, 행동 문제, ADL (일상생활 활동 점수)가 포함되어 있다. 증상에는 혼란, 방향감각 상실, 성격 변화, 작업완료의 어려움, 건망증이 있다. 기밀정보에는 담당 의사에 관련된 기밀 정보가 있다. 이러한 변수들은 설명 변수이자 X 로 이용될 예정이다. 진단 정보는 알츠하이머 진단 여부에 대한 정보로 결과 변수이자 Y 로 이용될 예정이다.

이렇게 구성된 데이터셋을 train-set 과 test-set 로 나누어서 모델을 학습시켰다. 이때 두 데이터셋의 비율은 8:2 로 구성해서 train-set 이 8, test-set 이 2 가 되도록 데이터를 랜덤으로 구분하여 학습 및 테스트를 진행하였다.

2.4. Data Preprocessing

Kaggle 에서 가져온 위의 데이터의 범위가 0 부터 10, 4 부터 10, 60 부터 120 등 변수마다 각각 다른 데이터 범위를 가지기 때문에, 정확한 분석을 위해서 데이터의 범위를 0 부터 1 사이로 함축시키는 데이터 전처리 과정을 거쳤다. 이 과정에서는 Min Max-scaler 를 이용해서 데이터를 일정 범위로 조절해 서로 다른 규모를 가진 데이터를 하나의 규모에 맞게 표준화하였다. 즉, 여러 변수를 동일한 가중치로 고려하기 위한 과정이라고 보면 된다.

2.5. Directing Data

2.5.1. Analyzing the Density of Variables

데이터 전처리 작업을 하기 전에, histogram 과 밀도 곡선, boxplot 을 이용해 각 변수의 밀도 분포를 살펴보았다. 여러 변수들 중, 나이 변수를 먼저 살펴봤다. 나이 변수는 60~90 세 사이의 데이터만 포함하고 있고, 이 중에서도 85 세 이상의 데이터가 가장 많다. 이러한 전체적인 분포를 histogram 과 밀도 곡선을 이용해 알아보았고, boxplot 을 이용해 데이터의 중앙값, 최댓값,

최솟값 등과 비교해서 면밀하게 살펴보았다. 그 외의 데이터들도 분포가 고르게 되어 있고 데이터 내부적으로도 고르게 되어있어 분석할 필요가 없었다.

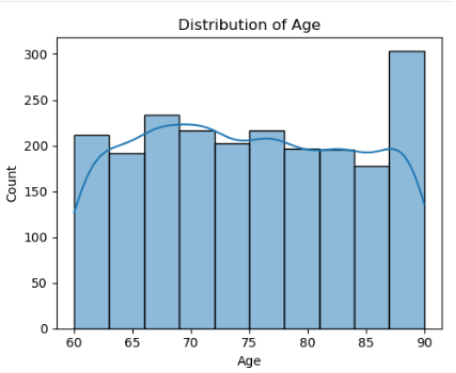

Figure 1: Age 변수의 히스토그램과 밀도 곡선

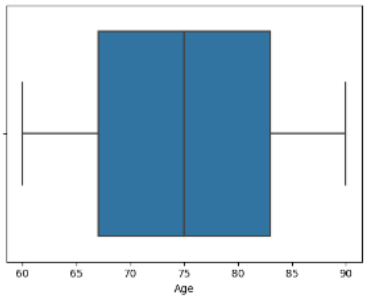

Figure 2: Age 변수의 boxplot

2.5.2. Improving Model Performance

모델 분석의 정확도를 높이기 위해 Grid search 를 진행했다. Grid search 는 모델 하이퍼 파라미터(hyperparameter)에 넣을 수 있는 값들을 순차적으로 입력한 뒤에 가장 높은 성능을 보이는 하이퍼 파라미터들을 찾는 탐색 방법이다. 여기서 말하는 하이퍼 파라미터는 모델 생성시, 사용자가 직접 설정하는 변수로, 예를 들어, 랜덤 포레스트 모델을 만든다면, 트리의 개수는 몇 개까지 할 것인 것, 트리의 깊이는 얼마나 할 것인지 등을 정하는 것이다. 본 연구는 각 모델에 맞는 최적의 파라미터를 찾아 입력시켜, 가장 적합하게 모델의 성능을 측정하였다.

3. Research

3.1. Model

3.1.1. Model Used for Learning

앞서 전처리를 한 데이터를 바탕으로 다양한 머신 러닝 및 딥 러닝 모델들 중에 연구 목표에 적합한 모델을 선정하여, 모델을 학습시키고 결과를 비교하였다.

본 연구에서 사용된 모델은 DT(Decision-Tree) 모델, RF(Random-Forest) 모델, NN(Neural-Network) 모델,

SVM(Support-Vector-Machine) 모델, KNN(K-Nearest – Neighbors) 모델, NB(Naive-Bayes) 모델, LR(Logistic Regression) 모델, XG-Boost 모델로 총 8 가지의 여러 모델의 성능 평가를 진행하였다.

3.2. How the Model Works

3.2.1. Decision Tree Model

머신 러닝의 학습 방법 중 지도학습 방식이며, 분류 방식에 해당하는 알고리즘 모델이다. 학습을 통해서 데이터 내의 규칙을 찾아내서 트리구조를 기반으로 하여 데이터를 분류하는 방식이다. 세부적으로 작동 방식을 살펴보면, 먼저 데이터를 구분할 수 있는 특성을 선택한다. 그리고 선택된 그 특성을 바탕으로 데이터를 둘 이상의 그룹으로 나눈다. 이 과정을 계속 반복해서 트리를 성장시켜 나가다가 분할이 더 이상 유용하지 않거나, 노드가 특정 조건을 만족하면, 분할을 멈추고 리프 노드를 생성해 종료시키는 작동 방식을 갖는 모델이다.

3.2.2. Random Forest Model

여러 개의 의사결정나무(Decision Tree)를 배깅(bagging)이라는 앙상블 기법을 기반으로 학습하고 조합해 최종 예측을 만드는 알고리즘 모델이다. 세부적으로 살펴보면, 먼저 원본 데이터에서 중복을 허용해서 랜덤하게 여러 개의 샘플 데이터를 생성하는 부트스트랩 샘플링(Bootstrap Sampling)을 진행해야 한다. 그 후, 부트스트랩 샘플 데이터에 대해 개별 의사 결정 트리를 학습시키고 그 예측을 결합하는 방식이다. Random forest 모델은 decision tree 모델과 달리, 데이터에

민감하지 않다는 일관성이 있고 과적합 가능성이 줄어든다는 차이점이 있다.

3.2.3. Neural Network Model (deep learning)

인공 신경망 모델은 인간의 뉴런을 모방한 딥러닝이자 머신러닝 모델이다. 입력 데이터에서 중요한 패턴을 학습하고 이를 바탕으로 예측, 분류 혹은 회귀작업을 수행한다. Neural Network 모델은 모델에 데이터를 입력하는 입력 층, 데이터를 처리하고 학습하는 주요 작업이 이루어지는 은닉 층, 예측 결과를 출력하는 출력층으로 구성된다. 이러한 방식은 비선형 처리가 가능하고 이미지, 음성 등의 여러 용도의 데이터를 효과적으로 학습할 수 있고, 중요한 특징을 자동으로 학습할 수 있어 적은 전처리로 좋은 성능을 낼 수 있다는 장점이 있다.

3.2.4. Support Vector Machine Model

데이터 포인트 간의 최적의 결정 경계를 찾아 분류나 회귀작업을 수행하는 지도학습 알고리즘이다. SVM 의 목표는 두 클래스 간의 최대 폭을 가진 결정 경계를 찾는 것이다. SVM 은 특히 이진 분류 문제에서 성능이 뛰어나며, 고차원 데이터에서도 효과적으로 작동하여 텍스트 분류나 이미지 분류 같은 다양한 문제에 활용된다.

3.2.5. K-Nearest Neighbors Model

지도학습 알고리즘 중 하나로, 분류와 회귀작업에 사용된다. 이 데이터는 예측하고자 하는 새로운 데이터 포인트가 주어지면, k 개의 가장 가까이 있는 데이터들의 정보를 비교 분석해서 예측을 수행한다. 그래서 KNN 은 데이터 간의 거리 측정에 의존하며, 데이터 분포에 민감한 모델이라는 특징이 있다.

3.2.6. Naive Bayes Model

이즈 정리를 기반으로 하는 통계적 분류 알고리즘이다. 베이즈 정리는 주어진 데이터 포인트가 특정 클래스에 속할 확률을 계산하는 과정에서 사용된다. 나이브 베이즈 모델은 해당 변수들이 서로 독립적이라고 가정하기 때문에, 계산이 단순하다는 특징이 있다.

3.2.7. Logistic Regression Model

머신 러닝의 비지도학습 방식 중 회귀 방식에 해당하는 알고리즘으로, 주로 종속 변수의 분류 문제에서 사용된다. 선형 회귀와 달리, 로지스틱 회귀는 결과 값이 연속된 수치가 아니라 0 과 1 사이의 확률 값으로 제한된다는 특징이 있다. 이 알고리즘은 두 데이터 요인 간의 관계를 찾는 데이터 분석 기법으로, 관계를 사용해 다른 요인을 기반으로 이러한 요인 중 하나의 값을 예측한다.

주로 입력과 가중치를 조합해서 값을 계산한 후, 로지스틱 함수 혹은 시그모이드 함수를 사용해서 선형 조합을 확률 값으로 변환한다. 그러고 난 후, 확률 값을 기준으로 이진 분류의 결정 경계를 설정하고 일반적으로 0.5 보다 크면 양성으로 0.5 보다 작으면 음성으로 분류해 훈련 데이터에 대해 최적화하는 알고리즘을 사용해 조정하여 학습을 진행한다.

3.2.8. XG Boost Model

여러 개의 의사결정 트리에 기반한 앙상블 학습 알고리즘이다. XG-boost 모델은 여러 개의 약한 결정 트리를 순차적으로 학습한 후, 각 단계에서 이전 단계의 예측 오류를 이는 방향으로 학습한다. 즉, 각 트리는 이진 트리가 잘못 예측한 데이터에 가중치를 부여해 성능을 개선한다. 랜덤 포레스트 모델은 한 번에 다양한 데이터셋을 만들어 그 결과를 평균화하여 다양한 데이터셋에서 안정적 성능을 얻을 수 있는 알고리즘을 만드는 방식이고, 이와 달리 XG 부스트 방식은 오차가 줄어드는 방향으로 학습하는 방식인 그라디언트 방식에서 학습 시간이 오래 걸린다는 단점을 보완한 방식이다.

3.3. Model Performance Analysis

TP(True Positive): 모델이 실제로 양성 클래스를 양성이라고 정확하게 예측한 경우

FN(False Negative): 모델이 실제로 음성 클래스를 음성이라고 정확하게 예측한 경우

FP(False Positive): 모델이 실제로 음성 클래스를 양성으로 잘못 예측한 경우

TN(True Negative): 모델이 실제로 양성 클래스를 음성으로 잘못 예측한 경우

3.3.1. Confusion Matrix

각 모델의 정확도를 예측하기 위해 혼동 행렬의 정확도, 정밀도, 진양성율, 위양성율, 조화평균을 계산할 수 있다.

$$Accuracy = \frac{Sum\ of\ diagonals(TP)}{Total\ number\ of\ instance} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$F1\ Score = \frac{2*(Precision*Recall)}{Precision+Recall} \quad (5)$$

Figure 3: 정확도, 정밀도, 진양성율, 위양성율, 조화평균을 계산하는 수식

3.3.2. Model Performance Indicators

여러 예측 모델의 학습 결과를 한 눈에 비교하기 위해 모델의 성능을 판단하는 기준(accuracy, precision, recall, specificity, f1-score)을 토대로 결과를 비교한 표를 제작해보았다.

아래의 표는 본 연구에서 진행한 여러 모델의 학습 결과를 비교하기 위해 각 모델의 정확도, 정밀도, 진양성율, 위양성율, 조화평균에 대한 값을 한 눈에 비교할 수 있도록 작성한 표이다.

그 아래의 모델 별로 결과값을 보여준 그림은 각 모델의 정확도, 정밀도, 진양성율, 위양성율, 조화평균, 혼동행렬 값까지 포함된 결과이다.

Table 1: 예측 모델 성능 비교표

Model	Accuracy	Precision	Recall	Specificity	F1 score
1. Decision Tree	0.960	0.965	0.919	0.982	0.942
2 Random Forest`	0.967	0.986	0.919	0.993	0.951
3. Neural Network	0.879	0.839	0.805	0.918	0.822
4. Support Vector Machine	0.912	0.878	0.866	0.936	0.872
5. K-Nearest Neighbors	1.000	1.000	1.000	1.000	1.000
6. Naïve Bayes	0.809	0.725	0.725	0.854	0.725
7.Logistic Regression	0.837	0.776	0.745	0.886	0.760
8. XG-boost	0.956	0.951	0.919	0.975	0.935

3.3.3. Model Accuracy

모델에 치매 데이터를 이용해서 학습시키고 예측 결과를 비교하기 위해서 각 모델의 학습 결과 중 F1-score 를 비교하려고 한다.

F1-score 는 모델은 Precision 과 Recall 성능을 동시에 고려한 조화평균 값이며, 0 과 1 사이의 값이며 1 에 가까울수록 모델 성능이 좋음을 나타낸다.

$$F1\ Score = \frac{2*(Precision * Recall)}{Precision + Recall}$$

Figure 4: F1 Score 계산하는 수식 그림

Table 2: 각 모델의 F1 Score 비교

Model	F1 Score
1. Decision Tree	0.942
2. Random Forest	0.951
3. Neural Network	0.822
4. Support Vector Machine	0.872
5. K-Nearest Neighbors	1.000
6. Naïve Bayes	0.725
7. Logistic Regression	0.760
8. XG-boost	0.935

4. Conclusion

4.1. Final Model Selection

F1-score 를 분석해 본 결과, K-Nearest Neighbors 모델, Random Forest 모델, Decision Tree 모델, XG-boost 모델, Support Vector Machine 모델, Neural Network 모델, Logistic Regression 모델, Naive Bayes 모델 순서대로 학습이 잘 된 모델이었다.

4.2. Weight Prediction

학습이 잘 된 예측 모델이 무엇인지 찾아내는 것으로 해당 연구를 마무리하기가 아쉬워서, 치매라는 변수에 가장 큰 가중치를 차지하는 변수가 무엇인지 알아내기 위해, 추가적인 분석을 진행했다.
이때, 이진변수가 아닌 변수들은 p-value 가 0.05 미만일 때 가중치를 분석한 값이 유의미하다고 해석할 수 있으므로 먼저, 카이 제곱 검증을 통해 p-value 를 비교했다. 우선, 이진변수가 아닌 구간형 변수들은 Age, BMI, Alcohol Consumption, Diet Quality, Sleep Quality, Physical Activity, Systolic BP, Diastolic BP, Cholesterol Total, Cholesterol LDL, Cholesterol HDL, Cholesterol Triglycerides, Functional Assessment, ADL, MMSE 이다. 이 변수들의 p-value 를 비교하면, 유의미한 변수들은 MMSE, Functional Assessment, ADL 이다.

이후에 가중치를 알아보기 위한 분석을 진행했는데, 결과 변수인 치매 진단과 양의 관계를 갖는 설명 변수는 Memory Complaints, Behavioral Problems 이다. 반대로, 음의 관계를 갖는 설명 변수는 Functional Assessment, ADL, MMSE, Education Level, Smoking, Cholesterol LDL 이다.

한편, 이진 변수들은 오즈비 분석을 통해서 설명 변수와 결과 변수 간의 상관성을 알아낼 수 있다. 오즈비 분석의 기준은 값> 1 이면, 설명 변수 증가할 때, 결과 변수 발생 확률의 증가를 의미하고, 값 = 1 이면, 설명력이 없다는 것을 말하고, 값 <1 이면, 설명 변수 증가할 때, 결과 변수 발생 확률이 감소한다는 것을 의미한다. 이를 토대로 이진변수를 분석하면, 값이 1 보다 큰 변수들은 Memory Complaints, Behavioral Problems 이고 값이 1 보다 작은 변수들은 Smoking, Education Level 이다.

=>이진변수에서 양의 관계로 유의미한 변수는 Memory Complaints, Behavioral Problems 이고, 음의 관계로 유의미한 변수: Smoking, Education Level 이다.

=> 구간 변수에서 양의 관계로 유의미한 변수는 없고, 음의 관계로 유의미한 변수는 Functional Assessment, ADL, MMSE, Cholesterol LDL 이다.

4.3. Finalization

연구 시작 전에는 Alzheimer′s disease diagnosis 와 Medical History 와 관련된 변수들만으로 모델을 학습시키고 예측하는 연구를 진행하려고 하였으나, K-fold 와 K-means 등의 정확도를 높이는 방법을 진행하였음에도 데이터의 불균형 때문인지(치매 진단을 받지 않은 사람의 데이터는 1389 개이지만, 치매 진단을 받은 사람의 데이터는 760 개로 양의 차이가 있다) 혼동 행렬, 정확도와 F1-score 를 비롯한 지표들의 값이 전반적으로 낮게 나왔다.

그래서 데이터의 불균형을 해결하고자 oversampling 과 under sampling 의 방법을 시도하였다. 그러나 under sampling 은 데이터의 양이 적은 쪽의 데이터의 양에 맞게 데이터 양이 많은 데이터를 줄이는 방식인데, 그렇게 되면 데이터가 너무 적어지기 때문에 진행할 수 없었다. 그래서 데이터의 양이 많은 쪽에 맞춰 데이터의 양이 적은 데이터를 늘리는 방식인 oversampling 을 시도했다.

또 다른 방안으로는, Kaggle 데이터에서 설명 변수를 Medical History 뿐만 아니라 다른 설명 변수들을 포함해 머신 러닝을 시켜보는 방안을 시도했다. 그 결과, 설명 변수가 많아지자 혼동 행렬 값이 좋아지고 정확성, F1-score 도 좋아져서, 본 연구를 이러한 방향으로 틀어서 진행하였다.

이렇게 틀어진 방향을 바탕으로 모델을 학습시킨 결과, K-Nearest Network 모델이 가장 훌륭한 학습 결과를 보여주었고, Decision Tree 모델과 Random Forest 모델의 F1-score 가 0.9 점대로 KNN 모델 다음으로 학습이 잘 되었다고 결론을 낼 수 있었다. 그러나 학습 결과가 0.9 점대로 매우 높아, 결과가 과적합 되었을 가능성이 있다고 판단된다.

추가적으로, 위의 결과만을 가지고 본 연구를 마무리하기가 아쉬워서, 부수적으로 결과 변수에 비중 있게 영향을 끼치는 설명 변수가 무엇인지를 분석했다.

그 결과, Memory Complaints 와 Behavioral Problems 가 증가하면, 치매 진단을 받을 확률이 증가한다는 것을 도출했다. 또한, Smoking, Education level 이 증가하면, 치매를 진단받는 확률이 낮아진다는 것을 도출했다. 그러나, Smoking 과 관련된 내용은 상식적인 면에서 이해되지 않은 결과였기 때문에 추후 연구를 통해서 확실하게 검증할 필요가 있다고 생각한다. 또한, Functional Assessment 기능 점수이고 ADL 은 일상생활 활동 점수, MMSE 는 간이 정신 상태 검사 지수, Cholesterol LDL 은 콜레스테롤 LDL 로 저밀도 지단백 콜레스테롤 지수를 말하는데, 이 변수들의 증가할 때, 치매 진단을 받을 확률이 낮아진다는 것을 알았다. 이때 Functional Assessment, ADL, MMSE 는 점수가 낮을수록 기능 장애, 일상생활 장애, 인지 장애의 가능성이 높음을 나타내는 지수라고 한다. 그래서 이러한 결과가 나오는 게 당연할 수 있다. 그러나 Cholesterol LDL 의 경우는 정상 기준이 130mg/dl 이고 그 이상을 넘어가면 질환을 의미한다는 상식선에서는 이해되지 않는 결과다. 그래서 앞서 말한 Smoking 변수와 더불어 Cholesterol LDL 변수도 추가적인 연구를 진행해야 한다.

그러나 본 논문에서는 치매와 관련된 데이터를 머신 러닝하는 과정에서 정확성과 F1-score 를 늘리는 방안을 고안한 것을 보여줌으로써, 향후 의학 데이터를 예측 모델에게 머신 러닝하는데 기여할 수 있을 것이라고 생각한다.

References

BBC Korea. (n.d.). *Korea's 'Ultra-fast aging' is underway… Why can't people experience problems?*

Jeong, B., Kim, J. H., & Heo, T. Y. (2020). A study on the application and comparison of statistical models and machine learning-based techniques for predicting the onset of dementia. *Journal of The Korean Data Analysis Society (JKDAS)*, *22*(5), 1819–1834.

Jo, M., Jung, S., & Ahn, J. (2023). Predicting dementia using ensemble machine learning models: Focusing on exercise and sleep information. *Journal of Tourism and Interdisciplinary Research, 43*(2), 91–109.

Jo, Y., Yoo, J., & Kim, J. (2024). A study on dementia prediction models and commercial utilization strategies using machine learning techniques: Based on sleep and activity data from wearable devices. *Journal of Business Administration Research*, *28*(2), 137–157

Kang Min-soo, Chung Myung-ae, Han Dong-hun, "NO-Core AI", pp15-85

Kang Sung-ki, (2023.08.23) "*The number of dementia patients surpassing 1 million next year... The annual management cost per person exceeds 20 million won"*, Dementia News

Kim, H., & Park, J. (2023). Machine learning-based dementia prediction: Data processing techniques. *Journal of Cognitive Science*, *45*(3), 101–120

Korea Institute of Science and Technology Information. (n.d.). Machine learning-based dementia prediction techniques using latent period-specific data processing

Lee, T., & Oh, H. (2021). Dementia prediction model based on gradient boosting. *Journal of the Korea Institute of Information and Communication Engineering*, *25*(12), 1729–1738.

Rabi el kharoua, "Alzheimer's Disease Dataset"

Ryu, S., Shin, D., & Jung, G. (2020). Predicting dementia risk using feature extraction and hyperparameter optimization with XGBoost. *IEEE Access, 8*, 177708–177720.