DAT 1기 캡스톤 프로젝트 보고서

당뇨병 예측 데이터를 활용한 모델 비교 및 주요 위험요인 선별

: 분류 알고리즘을 중심으로

DAT 1기

데분데분

김미영 이주은 조강국 하수연

Data Analysis & Technology

DAT 1기 캡스톤 프로젝트 보고서

당뇨병 예측 데이터를 활용한 모델 비교 및 주요 위험요인 선별

: 분류 알고리즘을 중심으로

Comparison of Models and Screening of Major Risk Factors

Using Diabetes Prediction Data

이 보고서를 캡스톤 프로젝트 보고서로 제출합니다.

2023년 06월 08일

DAT 1기

데분데분

김미영 이주은 조강국 하수연

Data Analysis & Technology

DAT 1기 캡스톤 프로젝트 보고서

당뇨병 예측 데이터를 활용한 모델 비교 및 주요 위험요인 선별

: 분류 알고리즘을 중심으로

국문 요약

본 보고서에서는 당뇨병 예측 데이터를 활용하여 다양한 전처리 기법을 적용한 뒤 그 차이를 비교해 보았다. 이후 Logistic Regression, Random Forest, SVM 모델의 예측 성능을 비교해 보았는데, 그 결과 Random Forest 모델이 가장 우수하였다. 모델을 비교하는 지표로는 AUC를 이용하였다. 또한, 가장 성능이 높은 Random Forest 모델을 통해 추정해 본 결과, 당뇨병에 영향을 미치는 위험요인들의 중요도 순위는 당화혈색소(HbA1c_level), 혈당 수치(blood_glucose_level), 나이(age)순이었다.

Abstract

This study applied various pre-processing techniques on diabetes prediction data and compared the differences. Since then, the predictive performance of the Logical Regression, Random Forest, and SVM models were compared, with the result of Random Forest as best. Comparison indicator which this study chose was AUC. In addition, as a result of estimating through the most performance Random Forest model, the three most important risk factors affecting diabetes were ordered: HbA1c_level, blood_glucose_level, and age.

DAT 1기

데분데분

김미영 이주은 조강국 하수연

Data Analysis & Technology

목 차

표 목차

그림 목차

당뇨병 예측 데이터를 활용한 모델 비교 및 주요 위험요인 선별

1. 서론

최근 비만, 부종, 인구 고령화 등의 요인으로 인해 당뇨병 유병률이 세계적으로 급증하고 있다. 당뇨병은 만성적 질환이기에 관리하지 않을 경우 중증의 합병증을 유발할 수 있다는 점에서 사회적으로 주목을 받고 있기도 하다. 대한당뇨병학회의 당뇨병 팩트 시트 논문(Diabetes Fact Sheet in Korea 2021)에 따르면, 2020년의 당뇨병 환자 수는 600만 명으로, 2010년에 비해 2배 증가한 바 있다. 이에 반해 당뇨병의 발생을 감하기 위해서는 당뇨병에 영향을 미치는 위험 요인을 파악하고, 사전에 예방할 필요가 있다.

머신러닝(machine learning)이란 인공지능의 한 분야로, 최근 데이터의 중요성이 증대되며 그 양과 질이 개선됨에 따라 크게 주목받고 있다. 의학 분야에서도 이미지 인식, 질병의 발병 예측 등으로 적극 활용되고 있다(조상아 외, 2021). 본 연구에서는 당뇨병에 영향을 미치는 요인을 알아보기 위하여 Random Forest를 활용한다. Random Forest의 경우, 타 모델과의 비교를 통해 해당 모델의 선정 이유를 서술한다.

본 보고서의 구성은 다음과 같다. 2절에서는 데이터를 소개한 뒤 전처리 과정을 설명하였으며, 3절에서는 2절의 최종 데이터에 세 가지 모델을 적용하여 본 뒤, 최적의 모델을 선정한다. 4절에서는 선정한 모델로 데이터 내 당뇨병과 각 변수들의 관계를 파악해 본다. 5절에서 결론을 맺은 뒤, 6절에서 한계점 및 제언을 서술함으로써 본 보고서를 마무리한다.

2. 데이터 소개

2-1. 데이터베이스 소개

본 보고서에서 다룬 데이터는 전자건강기록(EHR, Electronic Health Records)으로부터 수집되었다. 이는 당뇨병 예측 데이터로, gender, age, hypertension, heart_disease, smoking_history, bmi, HbA1c_level, blood_glucose_level, diabetes의 9개 열과 100,000개의 행으로 이루어져 있다.

2-2. 전처리

결측치는 없는 것으로 확인되어 별도의 결측치 처리 과정은 거치지 않았다. 한편, 실수형으로 저장된 age 변수의 데이터를 모두 int로 바꾸어 보기 편하게 하였다. 본격적인 전처리는 데이터의 열 중 Gender, smoking_history, diabetes를 다루는 것으로 구분한다. 이후 적절히 전처리 된 데이터로

정규화를 한다.

변수 중 Gender의 경우 자료 100,000개 중 여성이 58,552명, 남성이 41,430었는데, 그 외(Other)가 18명으로 표기되어 있었다. 해당 자료의 크기가 크지 않다고 판단하여 18개의 행을 모두 삭제하여 99,962개의 데이터를 남겼다. 여성과 남성의 경우 더미 변수를 생성함으로써 범주형 변수를 연속형 변수로 변환한다.

열 smoking_history에 대응하는 값으로는 No Info, never, former, current, not current, ever로 여섯 가지이다. No Info를 값으로 가지는 행이 35,810개로 전체 99,982개의 데이터 중 35.8%를 차지한다. 결측치로 간주하여 제거하기에는 그 비중이 크다고 판단하여 아래 Table 1과 같은 방안을 제시함으로써 직접 비교해 보았다. 이는 이후 전처리 과정으로 Diabetes 열을 다룬 뒤에 이루어져 한 번에 평가된다.

Table 1. Four Ways of Data Pre-processing; smoking_history

가. No Info 를 가지는 행을 모두 제거한 뒤 남은 값에 대해서는 더미 변수를 생성함
나. No Info 를 제거하지 않고, 5 개의 값 모두를 대상으로 더미 변수를 생성함
다. Label Encoding 방법을 사용하여 0 부터 n-1 까지의 연속적인 수치 데이터로 표현함
라. smoking_history 를 나타내는 열을 삭제함

* 가, 나의 경우 다중공선성 해결을 위해 ever 값을 제거함

데이터의 종속변수로는 Diabetes 열을 사용한다. 그러나 값을 출력해 본 결과, 당뇨병 환자가 아닌 데이터가 91,500개인 반면, 당뇨병 환자의 데이터 수는 8,500개로 그 비율이 10:1을 초과함을 확인했다. 범주형 데이터 내 두 값 간의 비대칭이 큰 경우, 전처리의 여부나 모델의 종류와 무관하게 모델의 성능이 높게 측정되는 경향이 있기에 이를 해결하고자 하였다. 먼저 SMOTE 기법을 통해 데이터를 Oversampling 한다. Oversampling이란, 적은 비율의 데이터의 수를 늘림으로써 데이터 불균형을 해결하는 것이다. Oversampling의 방법 중 하나인 SMOTE 기법은 적은 비율의 데이터를 기준으로 근접해 있는 데이터들과 일정 거리 떨어진 위치에 새 데이터를 생성한다. 이후 이를 Undersampling한 데이터와 비교해 보았다. 이는 앞서 네 가지 경우에 따라 전처리한 smoking_history에 각각 덧붙인 뒤, Logistic Regression 모델을 적용하여 그 성능을 평가하는 방식으로 진행된다. 성능 평가의 지표로는 Test 데이터에 대한 모델의 AUC(Area Under the Curve)를 사용하였으며, 결과는 Table 2와 같았다.

Table 2. Oversampling vs Undersampling

	Oversampling	Undersampling
가	0.8442	0.8626
나	0.8389	0.8705
다	0.8878	0.8874
라	0.8776	0.8885

최적의 결과에 한해 추가적으로 Train 데이터에 대한 AUC 값을 구하여 과적합이 발생하지 않는다는 사실 또한 확인하였다. 위 결과에 따라 smoking_history 열을 일괄 제거하였으며, Undersampling으로 데이터의 균형을 맞추기로 결정하였다.

마지막으로 표준화를 통해 각각의 값을 평균이 0이고, 분산이 1인 값으로 변환한다. 추후 사용할 모델인 Logistic Regression과 SVM의 경우 데이터가 가우시안(Gaussian) 분포를 가지는 것을 기반으로 구현되었기 때문이다.

3. 변수 선택 및 예측을 위한 분류 모형

3-1. Logistic Regression

반응변수 Y가 0 또는 1의 값을 갖는 이변량 질적변수(binary qualitative variable)일 때, 설명변수와 반응변수 간의 함수 관계를 로지스틱 회귀모형으로 적용시킨다. 이분형 자료가 관심의 범주인 $y = 1$에 속할 확률이 $Pr(Y = 1) = \theta$, 그 반대의 경우를 $Pr(Y = 0) = 1 - \theta$이라고 하자. 이때, 설명 변수 $X = (x_0, x_1, \cdots, x_p)$에 대한 베르누이 분포(Bernoulli distribution)를 따르는 반응변수 Y의 조건부 기대값을 다음과 같이 표현할 수 있다.

$$Y \sim Bernoulli(\theta),$$

$$Pr(Y = y|X = x) = [\theta(x)]^y \ [(1 - \theta(x))]^{1-y} \ y \in \{1, 0\},$$

$$E(Y|X = x) = Pr(Y = 1|X = x) = \theta(x).$$

반응변수의 조건부 기댓값은 성공확률이므로 $0 \le \theta(x) \le 1$의 제약을 가지게 되는데, 이런 경우 일반적인 선형회귀 모형을 적용하게 되면 확률값이 가지는 범위를 벗어나게 된다. 이러한 문제를 해결하기 위해 확률의 로짓변환(logit transformation)을 이용하여 선형화한다.

$$log(\theta(x)/(1 - \theta(x))) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p.$$

위 식을 정리하면 다음과 같고, 이는 0과 1 사이의 값을 가지는 확률의 성질을 만족하게 된다.

$$\theta(x) = (exp(\beta_0 + x^T\beta))/(1 + exp(\beta_0 + x^T\beta)) = \{1 + exp(-\beta_0 - x^T\beta)\}^{-1}.$$

로지스틱 회귀모형에서 모수를 추정할 때, 일반적으로 최대우도추정법 (maximum likelihood estimation)을 이용한다. n쌍의 $(xi, yi), i = 1, \cdots, n$의 자료를 가정하자. 이때, 아래와 같이 모수에 대한 우도함수(likelihood function)를 정의하고, 우도함수가 최대화되는 모수를 추정치로 구한다.

$$L(\beta_0, \beta_1, \cdots, \beta_p) = \prod_{i=1}^{n} Pr(Y = y_i \mid X = x_i) = \prod_{i=1}^{n} \theta(X_i)^{y_i} (1 - \theta(x_i))^{1-y_i}.$$

계산의 편의를 위해 위의 우도함수에 자연로그를 취해서 계산한다.

$$log\, L(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^{n} log(Pr(Y = y_i \mid X = x_i))$$

$$= \sum_{i=1}^{n} [\, y_i (\beta_0 + x_i^T \beta) + log\{1 + exp(\beta_0 + x_i^T \beta)\}].$$

로그우도함수식이 최대화되는 모수 β_0, β 를 구하기 위해 0, β 각각에 대해 편미분 하여 0이 되는 값을 구한다. 다음을 구할 때는 Newton-Raphson 방법 등을 이용한 반복 알고리즘을 통해 수치적(numerical)으로 추정한다.

$$U(\beta_0) = \frac{\partial L(\beta_0, \beta_1, \ldots, \beta_p)}{\partial \beta_0} = 0,$$

$$U(\beta) = \frac{\partial L(\beta_0, \beta_1, \ldots, \beta_p)}{\partial \beta} = 0.$$

3-2. Random Forest

랜덤포레스트(Random Forest)는 다수의 개별 의사결정트리를 이용해 하나의 강력한 모델을 구성하여, 개별트리를 이용해 얻은 결과값을 통합한다. 의사결정트리에서 발생하는 일부 문제를 해결하고자 Breiman(2001)이 고안했으며, 앙상블(Ensemble) 기법의 종류인 배깅(Bagging) 기법을 활용하였다.

의사결정트리는 의사결정 규칙을 나무의 형태로 도식화한 것으로, 크게 Decision 노드, Root 노드, Leaf 노드로 이루어져 있다. 각 노드는 기준점과 같다. Root 노드를 시작으로 Root 노드 하위에 위치한 Decision 노드로 분류기준(질문)이 주어지며, 기준 만족 여부에 따라 해당 노드에 저장된 데이터가 각 하위 노드로 분리 전달된다. 데이터가 더 이상 분리되지 않고 최종 출력이 주어질 때 그것을 Leaf 노드라고 부르며, Leaf 노드를 최대한 섞이지 않은 상태로 분류하는 것이 의사결정트리의 핵심이다.

노드에서 데이터는 불순도(Impurity)에 따라 분류된다. 불순도란, 한 범주 안에 서로 다른 데이터가

얼마나 섞여 있는가를 나타내는 척도로, 노드에 주어지는 분류기준을 설정할 때는 기준 노드의 불순도와 비교해 하위 노드의 불순도가 감소하도록 설정해야 한다. 불순도는 다음과 같이 표현할 수 있다.

$$(1) \quad Entropy(X) = -\sum_{k=1}^{m} p_k \ log_2 \ (p_k)$$

$$(2) \quad Gini(X) = 1 - \sum_{k=1}^{m} p_k^2$$

이때, P_k는 대상 집합 내에 위치한 데이터의 비율이다. (1)은 $Entropy$(엔트로피) 계수라 불린다. 불순도를 측정하며, 예를 들어 $Entropy$ 값이 1에 가까울수록 한 범주 안의 서로 다른 특성의 데이터가 5:5 비율에 수렴한다는 의미이다. (2)는 $Gini$(지니) 계수라 불리며, $Entropy$와 유사하게 불순도 측정 지표로 사용된다. 참고로, $Gini$ 계수를 활용한 대표적인 알고리즘으로는 CART 모델 등이 있다.

일반적으로 의사결정트리에서의 주된 분류기준은 분리 이전의 불순도와 분리 이후의 불순도의 차이를 나타내는 '정보획득(Information Gain)'이며, 의사결정트리는 정보획득이 최대가 되는 분기조건을 찾아 분기를 반복한다. 다시 말해, 하위 노드로 분류될 때 모든 분류지점 중에서 불순도를 최소화하는 지점을 최적의 분할 기준으로 선택하는 방식으로 학습을 진행한다.

$$\Delta \, Information \, Gain = Imp(T) - [p_L * Imp(T_L) + p_R * Imp(T_R)]$$

위의 식은 정보획득을 표현한 식으로, *Imp(T)*는 특정 노드 T에서의 불순도를 나타내는 함수이다. p_L과 p_R은 설명변수 내의 특정 지점을 기준으로 분류한 후 상위노드 T 대비 두 하위노드의 비율을 의미한다. *Imp*(T_L)과 *Imp*(T_R)은 특정 지점을 기준으로 분류한 이후의 각 하위노드의 불순도 함수이다. 위의 과정을 재귀적으로 반복하여 최하위 마디에 속한 자료의 수가 미리 지정한 최솟값에 도달하면 분류를 중단하게 되며, 이때 최솟값을 자료의 수에 비해 지나치게 작게 지정할 경우 최종 모형이 과적합 될 가능성이 있다. 또한, 표본에 따라 모형의 변화폭과 예측력의 차이가 심하게 나타나는 단점이 있다. 이 문제를 해결하고자, Breiman(2001)은 앙상블 기법인 배깅(Bagging)을 활용하였다.

배깅(Bagging)은 데이터를 랜덤으로 복원 추출하여 여러 개의 샘플 데이터를 생성한 다음, 의사결정트리 형태의 모델에 적용하여 학습시킨 뒤 도출된 결과를 집계하는 방법이다. 각각의 샘플 데이터는 동일한 유형의 알고리즘을 기반으로 분류 예측에 사용되며, 도출된 예측치의 평균값을 활용한다. 이는 학습 데이터가 많지 않더라도 충분한 학습효과를 주어 과대 적합 등의 문제 해소에 도움을 준다.

랜덤포레스트는 배깅과 유사한 작동 구조를 보이나, 변수 선택에 있어 랜덤하게 선택하도록 하여, 개별 결정 트리들의 상관성을 줄여 예측력을 향상하도록 도모한다. Breiman(2001)은 그의 논문을 통해

개별 트리들의 상관성이 줄어들면 랜덤포레스트의 예측력이 좋아짐을 증명했으며, 상관성을 줄이고자 개별 결정 트리를 분리할 변수의 후보를 랜덤으로 선택하도록 하였다. 다만 이는 변수 후보를 한 번만 랜덤하게 추출하여 활용하는 것이 아니다. 각 분리에 대해 변수 후보를 랜덤 선택하는 것으로, 일부 분리에서 동일한 변수가 선정된다 하더라도 다른 분리에서 다시 랜덤으로 변수를 선택하기 때문에, 서로 다른 두 트리가 같아질 가능성은 적어지며, 훈련 데이터의 변동에 민감하게 반응하지 않아 안정된 예측 성능을 얻을 수 있게 된다. 통상적인 랜덤포레스트 모델이 작동하는 원리는 아래와 같다.

단계 1. 복원추출을 허용하여, 임의의 학습 데이터 $D = \{(x_i, y_i) \mid i = 1,2,\ldots,n\}$를 이용해 총 B개의 붓스트랩 샘플 데이터셋 $D_1, D_2, D_3, \ldots, D_B$를 생성한다.

단계 2. b = 1,2, … , B에 대하여 다음을 반복한다.

(a) $Out\ of\ Bag, O_b = D - D_b$를 보관한다.

(b) D_b를 이용하여 모델 $M(x, D_b)$를 학습시킨다.

(c) O_b를 이용하여 성능을 계산한다.

단계 3. 단계2를 통해 얻은 성능을 집계하여 최종모형 M(x)을 만든 뒤 M(x)의 성능을 계산한다.

이러한 랜덤포레스트도 단점이 없는 것은 아니다. 첫째, 다수의 개별 의사결정트리를 활용해 모델을 구축하기 때문에, 입력 변수에 따른 출력 변수의 변화를 해석하기 힘들다. 이는 앙상블 기법을 활용한 모델 대부분이 갖는 문제점이기도 하다. 둘째, 보통 앙상블 기법은 의사결정트리가 아닌 다른 형태의 모형에도 적용할 수 있지만, 랜덤포레스트는 의사결정트리를 이용하여 만들어진 모델이므로 결정트리 이외의 다른 예측 모형에는 적용할 수 없다.

결정트리가 지나치게 복잡하거나, 많은 규칙과 질문으로 수많은 분리가 발생한다면, 해석의 어려움 뿐만 아니라 과적합 문제도 발생할 수 있다. 이를 해결하는 방법은 '가지치기(Pruning)'로, 사전 방식과 사후 방식으로 나뉘며, 쉽게 말해 일부 분리를 의도적으로 제어함으로써 과적합 발생을 방지하는 것이다. 다만 본 분석에서는 사용하지 않았으며, 대신 '하이퍼 파라미터'를 설정해 과적합 문제를 해결하고자 하였다. 하이퍼 파라미터의 대표적인 종류로는 Table 3과 같다.

Table 3. Typical Parameters of Random Forest

n_estimators	결정트리의 개수 지정. Default = 10
max_features	분리/분할에 사용되는 특징의 수. Default = auto
max_depth	결정트리의 최대 깊이. Default = None

* 이외에도 다양한 파라미터가 존재하며, 데이터/모델의 형태(분류/회귀 등)에 맞게 활용됨

파이썬과 등의 프로그램에서는 적절한 파라미터를 탐색해 주는 기능을 제공하며, 파이썬의 경우 sklearn에 포함된 GridSearchCV 또는 RandomizedSearchCV가 있다. 본 분석에서는 GridSearchCV를 활용했으며, 해당 기능은 사용자가 조정하고 싶은 파라미터를 선정한 뒤 입력하고자 하는 값을 리스트 형태로 입력하면 해당 값들을 이용해 다양한 경우의 수를 조합하여 최적의 파라미터 값을 출력한다. 본 분석에서는 상기의 파라미터 중 n_estimators, max_depth만을 활용하였다.

랜덤포레스트는 수많은 의사결정트리를 조합하므로, 전체 형태를 한눈에 파악하기 쉽지 않다. 따라서, 각 특성변수의 중요도를 요약하는 기능을 별도로 제공한다. 해당 기능은 어떤 특성을 사용한 노드가 모델 내 모든 트리에 걸쳐 평균적으로 불순도를 얼마나 감소시키는지 파악한 뒤 중요도 점수를 부여한다. 중요도 점수가 높은 변수는 점수가 낮은 변수보다 예측 등에 있어서 상대적으로 높은 중요성을 갖고 있음을 의미한다.

3-3. SVM

SVM은 분류 및 예측 기법에 사용되는 기계 학습법의 일종으로서, 두 그룹의 데이터를 구분시키는 초월평면(hyperplane)을 계산하는 방법이며, 이때 마진(margin)을 최대화하는 방법이다. 마진은 초월평면과 직교 방향으로 가장 가까운 데이터 샘플과의 거리를 의미하며 마진이 큰 초월평면을 찾는 이유는 기계학습법에 의해 두 그룹을 구분 짓는 분류기, 즉 초월평면을 계산한 후에 새로운 데이터인 Test 데이터로 분류할 때 보다 더 정확한 예측을 할 수 있다는 가정 때문이다. 초월평면과 가장 가까운 데이터 샘플들을 support vertor라고 하는데, SVM은 이 support vector들만을 이용하여 분류기를 모델링한다. SVM에서 가장 널리 사용되는 커널함수로는 선형커널(linear kernel), 다항커널(polynomial kernel), RBF 커널(RBF kernel) 등이 있다.

주어진 샘플 S_i 의 특징 벡터를 $x_i = < f_1(S_i), \dots, f_n(S_n) >$ 로 정의하면 선형 커널 함수에 기반을 둔 이진 분류는 위의 식과 같이 표현된다. 일반적으로 SVM의 학습데이터 S 는 $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ 과 같이 표현되는데, 여기에서 $y_i \in Y = \{+1, -1\}$ 는 분류기의 출력공간(샘플의 유형: +1 은 당뇨병 환자 샘플, -1은 정상 환자 샘플)을 나타낸다. 이와 같은 SVM을

이용한 샘플의 이진 분류의 개념을 도식화하면 다음 그림과 같다.

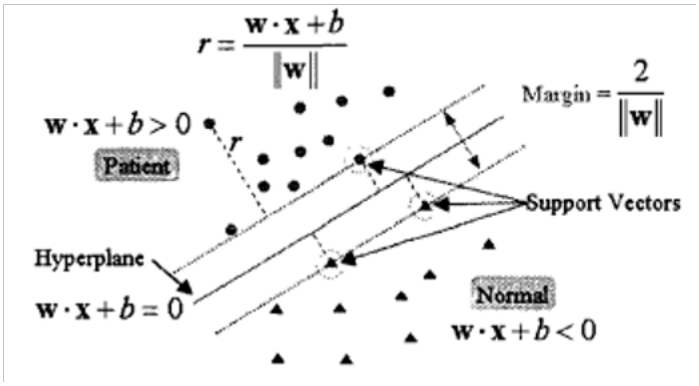

Figure 1. Binary Classification of Samples

즉, SVM을 이용한 분류모형을 설명하기 위해 학습데이터(learning data set) $\mathrm{D} = \{(\mathrm{x}_i, \mathrm{y}_i)|\mathrm{x}_\mathrm{i} \in \{-1, +1\})_{i=1}^{\mathrm{n}}$가 다음과 같은 제약식을 만족한다고 가정한다.

$$w^T x_i + b \geq +1, \quad for\ y_i = +1$$

$$w^T x_i + b \leq -1, \quad for\ y_i = -1$$

위 제약식을 다음과 같이 하나의 식으로 표현된다.

$$y_i(w^T x_i + b) \geq 1, \quad for\ i = 1, \ldots, n(1)$$

제약식(1)은 학습데이터가 선형분리 가능한 경우에 해당되고 선형분리 불가능한 데이터인 경우 슬랙 변수(slack variable) $\xi_i(\geq 0)$을 사용하여 다음과 같이 나타낼 수 있다.

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad for\ i = 1, \ldots, n$$

일반적으로 데이터가 선형분리 불가능하다고 판단하여, 학습 시의 에러를 어느 정도 허용하는 soft margin 방법을 사용하는데 다음과 같은 최적화 문제로 해를 구한다.

$$minimize \quad Q(w, b, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_\mathrm{i}$$

$$subject\ to \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, n$$

$$\xi_i \geq 0, \ i = 1, \ldots, n$$

여기서 C는 마진(margin)의 최대화와 분류 오류율의 최소화 사이 트레이드-오프(trade-off)를 결정

하는 모수이다. 커널 함수(kernel function) $K(x_i, x_j)$을 도입하여 위의 최적화 문제를 다음과 같이 나타낼 수 있다.

$$maximize \qquad Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$subject\ to \qquad \sum_{i=1}^{n} \alpha_i y_i = 0, \qquad i = 1, \dots, n$$

$$C \geq \alpha_i \geq 0\,,\ i = 1, \dots, n$$

여기서 α_i는 라그랑지 배수(Lagrange multiplier)로써 , 모든 학습데이터마다 하나씩의 양수 값을 갖게 되는데, support vector 외에 다른 데이터는 모두 0의 값을 갖게 된다. 또한 K는 커널이라 불리며, 연구 목적과 데이터 분포에 따라 선형 커널뿐만 아니라 radial basis function(RBF)과 같은 여러 가지 비선형 커널도 많이 사용된다. 그리하여 최종적으로 다음 식과 같이 분류기의 식을 구할 수 있게 되는데, 일반적으로 Test 데이터를 이 식에 적용하였을 때, 양수 값이 나오면 양성, 음수 값이 나오면 음성으로 판별한다. 따라서 서포트 벡터 x_i를 사용하여 최적의 분리 평면은 다음과 같이 입력벡터 x의 결정함수 식으로 나타낼 수 있다.

$$F(x) = sgn(\sum_{i=1}^{n} \alpha_i \alpha_j K(x, x_i) + b)$$

자주 사용되는 커널함수 $K(x_i, x_j)$ 로는 Table 4와 같이 3가지가 있다.

Table 4. Kernel

Linear Kernel	$K(x_i, x_j) = x_i^T x_j$
Polynomial Kernel	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$
RBF(Radial Basis Function) Kernel	$K(x_i, x_j) = \exp(-\gamma \left\|\|x_i - x_j\right\|\|^2)$

여기서 γ, r 그리고 d는 커널의 형태를 결정하는 모수들이다. SVM의 성능에 영향을 미치는 모수로는 패널티 모수 C , 커널모수 γ그리고 커널함수 $K(x_i, x_j)$등이 있다.

4. 당뇨병 예측 모델들의 성능 비교 및 주요 위험요인 선별

4-1. 최적의 모델 선정

먼저 Logistic Regression의 경우, 아무 규제 없이 수행하였을 때 AUC 점수가 0.8885로 도출되었다. 한편, L1과 L2 규제를 사용하여 모델의 가중치를 제한한 결과 AUC 값이 각각 0.8826, 0.8838로 나왔다. 그리고 Random Forest를 이용, 파라미터 조정 없이 모델을 수행하였을 때의 Test AUC 가 0.9034로 도출되었으나, Train AUC가 0.99845로 과적합 문제가 발생하였다. 이에 GridSearchCV 를 활용해 최적의 파라미터를 탐색하였고, n_estimators와 max_depth 두 가지 파라미터를 각각 200, 11로 선정하였다. 이 경우 Train의 AUC가 0.9219, Test AUC는 0.9093으로 도출되어 과적합 문제를 완화하였다. 마지막으로 SVM을 이용하여 AUC 점수를 평가하기 위해 선형 커널과 비선형 커널인 RBF(가우시안)과 Poly(다중) 커널을 사용하였으며, 선형커널에서는 0.8874, poly(다중) 커널에서는 0.8968, RBF(가우시안) 커널에서는 0.8990이 나왔다. Test AUC는 각각의 커널에서 0.8887, 0.9006, 0.8983으로 과적합이 발생하지 않는 것으로 보인다. 결과를 통해 비선형 커널이 가장 높은 Train 점수를 보여준 것으로 나타난다. 따라서, SVM의 경우 주어진 당뇨병 데이터에 대해 비선형 커널인 poly(다중) 커널을 사용한 모델이 가장 적합하며, 이를 통해 당뇨병 데이터가 비선형 분류가 가능한 특성을 가지고 있을 것이라고 추론할 수 있다. 결과적으로, 세 가지의 분류 알고리즘 중 최적의 파라미터를 대입한 랜덤포레스트의 점수가 0.9093으로 도출되어 최적의 모델로 선정되었다.

4-2. 가장 높은 위험 요인

4-1에서 선정한 최적의 모델인 Random Forest를 활용하여 당뇨병 발병에 가장 큰 영향을 미치는 요인을 찾고자 변수 중요도를 추출한다. 그 결과, HbA1c_level이 0.3542로 가장 큰 값을 가졌고, blood_glucose_level이 0.2621로 두 번째를 차지하였다. 이후 age, bmi, hypertension, heart_disease, Male, Female이 차례로 뒤를 이었다. 각 요인별 중요도와 그 순위는 Table 5와 같다.

Table 5. Feature Importance

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	Female	Male
Feature Importance	0.1972	0.0258	0.0149	0.1404	0.3542	0.2621	0.0026	0.0029
Ranking	3	5	6	4	1	2	7	8

5. 결론

본 보고서를 위한 연구에서는 당뇨병 예측 데이터를 이용하여 적절히 전처리한 뒤, Logistic

Regression, Random Forest, SVM의 세 가지 알고리즘을 실습해 보았다. 먼저 아무런 파라미터 조정 없이 Random Forest를 모델링한 결과, Train AUC 0.9984, Test AUC 0.9034으로 과적합이 발생하였다. 이후 GridSearchCV를 사용해 최적의 파라미터를 탐색하였고, n_estimators와 max_depth 두 가지 파라미터를 각각 200, 11로 선정하였다. 그 결과, Train AUC 0.9219, Test AUC 0.9093으로 과적합을 완화하였다. 본 연구는 세 가지 알고리즘을 실습해 AUC 0.9093이 나온 GridSearchCV를 사용해 최적의 파라미터를 조정한 Random Forest를 최적의 모델로 선정하였고, 이 모델을 활용해 당뇨병의 주요 위험 요인을 선별하였다. 전처리와 모델 평가에 있어서는 모두 예측 성능 AUC를 이용하였다. 결과적으로 당뇨병은 당화혈색소(HbA1c_level)와 혈당 수치(blood_glucose_level)가 각각 0.3542, 0.2621로 가장 유의미한 관계를 가짐을 알게 되었다

6. 한계 및 제언

6-1. 한계

본 보고서에 다룬 당뇨병 예측 데이터는 위험요인으로 사용할 만한 독립변수가 부족하였다는 점에서 한계가 있다. 전처리 과정에 있어 그 값들이 의미하는 바가 명확하지 않다고 판단하여 열 smoking_history를 일괄 제거한 점 또한 위 한계의 발생에 일조하였다.

당뇨병은 본래 1형과 2형으로 발병 원인 등의 성질이 명확히 다르다. 당뇨병은 혈당을 조절하는 인슐린과 관련이 있는데, 1형 당뇨의 경우 인슐린이라는 호르몬 자체의 부족으로 혈중 포도당이 쌓이게 되는 것으로, 자가면역기전이 그 원인이다. 반면 2형 당뇨는 정상적으로 분비된 인슐린이 제대로 된 역할을 하지 못하여 발생한다. 위 두 종류의 당뇨에 있어 가장 큰 차이는 발생 시기에 있는데, 1형 당뇨는 소아 및 청소년에게서 주로 발견되지만 2형 당뇨는 주로 30세 이상의 사람에게서 서서히 진행된다. 본 보고서에서 다룬 데이터는 당뇨병의 종류를 구분하지 않고 수집한 것이기에, 발병 요인 중 age에 해당하는 데이터가 무의미해질 수 있다는 점에서 한계가 있다.

6-2. 제언

본 연구는 의료 데이터를 적극적으로 이용하여 질병의 예방과 치료에 큰 도움을 줄 수 있음을 보여준다. 머신러닝을 활용하여 각종 질병에 영향을 미치는 주요 요인을 식별함으로써, 환자와 의료진은 질병을 예방할 수 있게 된다. 이는 더 나아가 보건 정책 수립에도 활용될 수 있으며, 전세계 국민의 건강 유지에 크게 기여할 것으로 전망된다. 이렇듯 의료 데이터의 종합적인 분석과 머신러닝의 적용은 당뇨병을 비롯한 다양한 질병의 관리에 대한 새로운 가능성을 보여준다. 한편, 앞서 한계에서 언급한

바와 관련하여 당뇨를 1형과 2형을 구분하는 후속 연구를 제시한다. 1형과 2형 당뇨의 구분이 가능하다면, 회귀분석을 사용하여 하나의 데이터셋 내 동일 변수가 각각의 당뇨에 미치는 영향이 어떻게 다른지 비교해 볼 수 있을 것이기 때문이다.

참고문헌

김한상, 조상아. "보건의료 빅데이터 기반 인공지능 활용 전략." HIRA Research 1, 2 (2021): 196-207.

박윤미. "로지스틱 회귀분석에서 희소주성분회귀법의 효율성 연구." 국내석사학위논문 韓國外國語大學校 大學院, 2015. 서울

엄재홍, 장병탁, “SVM 앙상블을 이용한 심혈과질환 질환단계 예측”, 서울대학교 컴퓨터공학부(바이오지능연구실), 2006

임진수, 오윤식, 임동훈, “유방암 분류 성능 향상을 위한 배깅 서포트 벡터 머신, 경상대학교 (생물학과 및 정보통계학과), 2014

조백환, 이종실, 지영준, 김광원, 김인영, 김선일, “특징점 선택방법과 SVM 학습법을 이용한 당뇨병 데이터에서의 당뇨병성 신장합병증의 예측”, 한양대학교 의용생체공학과, 성균관대학교 의과대학 내분비 대사 내과, 2007