개별연구 최종 보고서

통계학과 이윤정

1. 서론

대한당뇨병학회에 따르면 한국 30세 이상 성인 약 7명 중 1명(13.8%)이 당뇨병을 가지고 있다. 당뇨병 유병률은 연령 증가에 따라 계속 증가하는 추세를 보이고 있다. 그중 65세 이상 성인에서는 약 10명 중 3명(27.6%)이다. 당뇨병은 만성질환일뿐더러, 심뇌혈관질환, 신장병증 등 건강에 직접적으로 영향을 끼치는 만성 합병증을 동반하기 때문에 당뇨병 초기에 발견하는 것이 중요하다. 하지만, 당뇨병은 오랜 기간 증상을 느끼지 못한 채 진행되기 때문에 초기에 발견이 매우 어렵다. 현재 당뇨병 진단은 혈당 검사를 통해 측정된 공복혈당을 기준으로 판정된다. 따라서, 공복혈당 이외의 추가적인 데이터를 통해 당뇨병 예측에 성공한다면 당뇨병 잠재환자들을 파악할 수 있을 것으로 예상된다. 이에 본 연구는 2016년부터 2018년까지 진행된 제 7기 국민건강영양조사 데이터를 기반으로 공복혈당을 제외한 데이터를 통해 당뇨병 예측 모형을 적합하고자 한다.

2. 본론

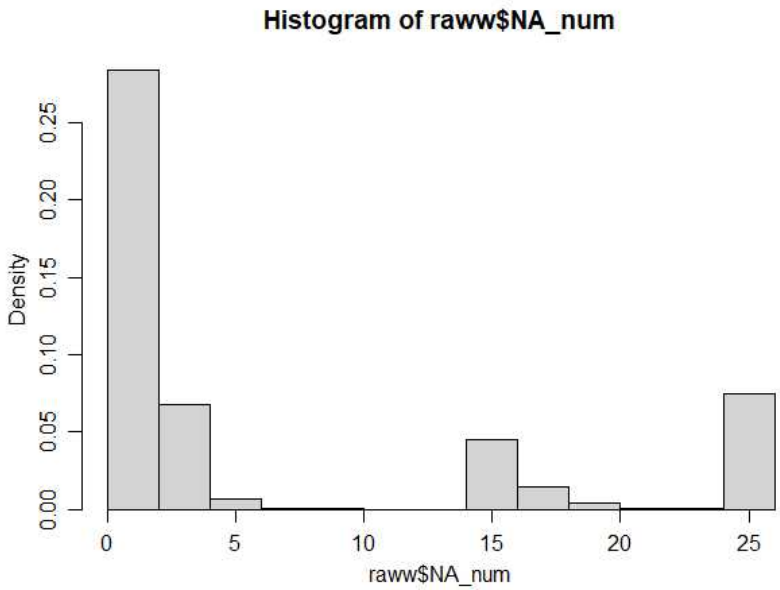
(1) 1차 변수 선택 및 전처리

국민건강영양조사 홈페이지에서 다운받을 수 있는 원시자료의 데이터는 가구원확인조사, 건강설문조사, 검진조사, 영양조사를 통해 수집된 데이터로 모형 적합을 위해선 1차 변수 선택 과정이 필요하다. 해당 과정은 선행 연구를 기반으로 진행되었으며, 선택된 변수는 총 26개로 다음과 같다. 변수 선택 전 원시자료 데이터는 총 7992 x 736 (행 x 열) 데이터로, 1차 변수 선택을 통해 만들어진 데이터 셋은 총 5675 x 26 (행 x 열)이다. 이때, 선형 연구에서 가장 중요한 변수는 HE_glu(공복혈당)이지만 모형에서 종속변수 Y로 사용될 HE_DM(당뇨병 유병여부)가 수치형인 HE_glu 변수를 cutoff하여 만든 범주형 변수이므로 종속변수와 직접적인 상관관계가 존재하므로 제외하였다.

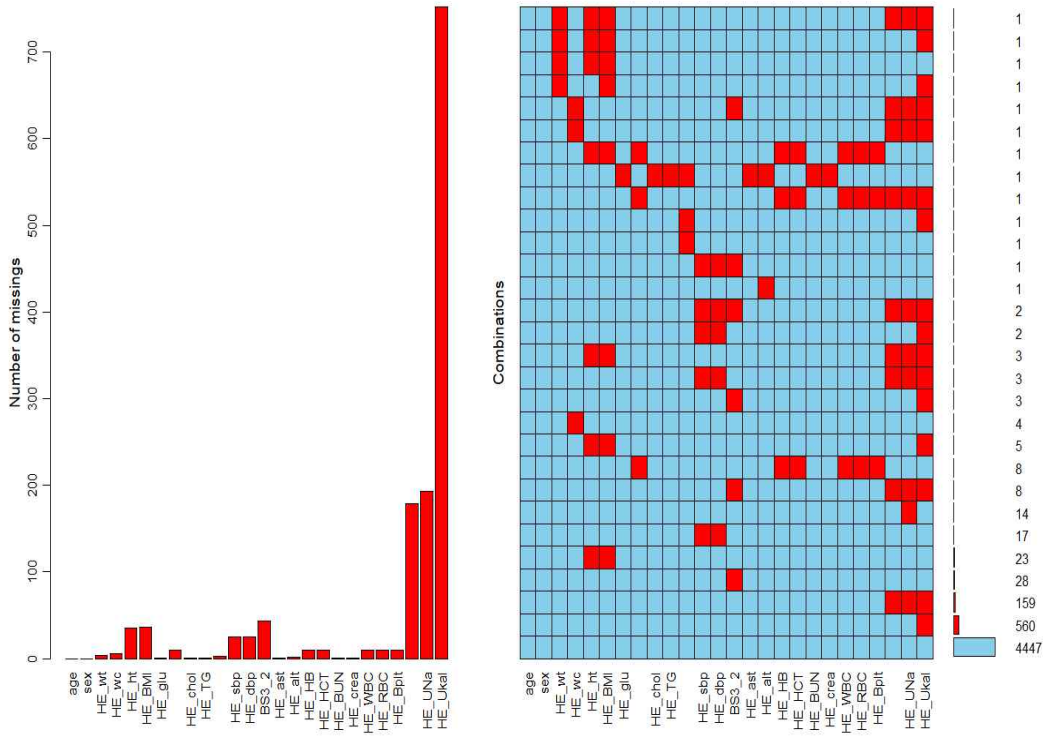
변수명	변수 설명	변수 종류		
age	만나이	수치형	1 ~ 79세	1 ~ 79
			80세 이상	80
sex	성별	범주형	남자	1
			여자	2
HE_wt	체중	수치형		
HE_wc	허리둘레	수치형		
HE_ht	신장	수치형		
HE_BMI	체질량 지수	수치형		
HE_HbA1c	당화혈색소	수치형		
HE_chol	총콜레스테롤	수치형		
HE_TG	중성지방	수치형		
HE_HDL_st2	HDL-콜레스테롤	수치형		
HE_sbp	최종 수축기 혈압	수치형		
HE_dbp	최종 이완기 혈압	수치형		
BS3_2	하루평균 흡연량	수치형		

HE_ast	AST(SGOT)	수치형		
HE_alt	ALT(SGPT)	수치형		
HE_HB	헤모글로빈	수치형		
HE_HCT	헤마토크리트	수치형		
HE_BUN	혈중요소질소	수치형		
HE_crea	혈중크레아티닌	수치형		
HE_WBC	백혈구	수치형		
HE_RBC	적혈구	수치형		
HE_Bplt	혈소판	수치형		
HE_UCREA	요크레아티닌	수치형		
HE_UNa	요나트륨	수치형		
HE_Ukal	요칼륨	수치형		
HE_DM	당뇨병 유병여부	범주형	정상	1
			공복혈당장애	2
			당뇨병	3

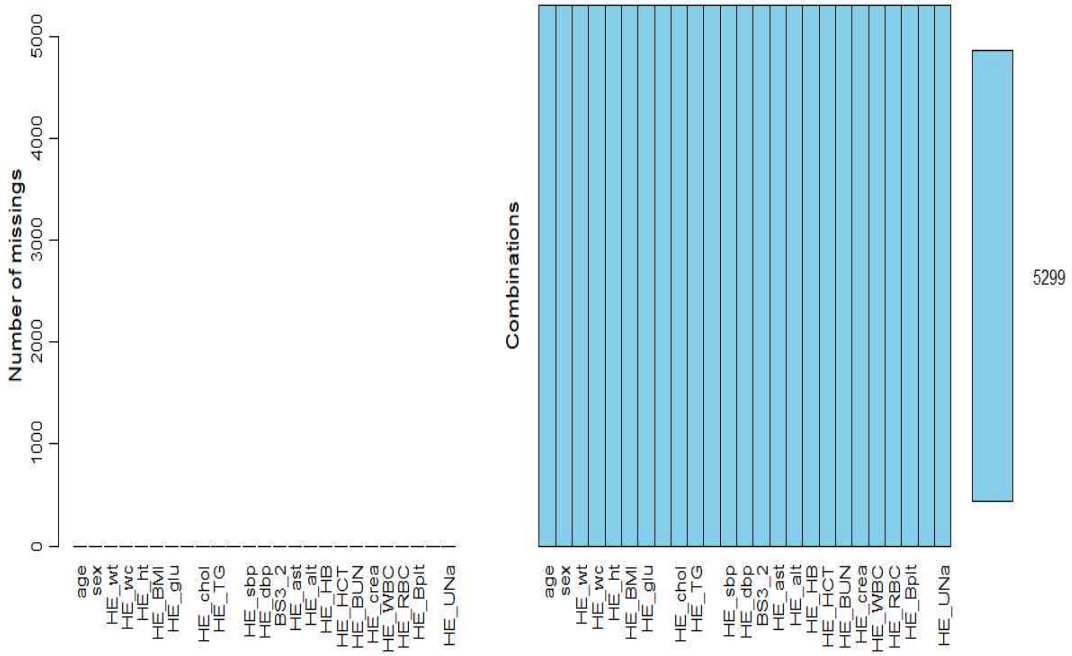
1차 변수 선택 과정을 거친 데이터를 분석한 결과, 검진 및 설문조사를 통해 수집된 데이터이므로 응답자별(행) 미응답변수가 존재하였다. 응답자별 미응답변수 개수를 column(NA_num)으로 만들어 히스토그램을 그려본 결과, 10개를 기준으로 분포가 양단됨을 알 수 있었다. 따라서, 본 연구에서는 26개의 변수 중 미응답변수가 10개보다 많은 경우 해당 응답자는 모형 적합에 부정적인 영향을 미친다고 판단되어 해당 행을 제거하였고, 총 377개의 행이 제거되었다.



이후, 변수 별 결측치 개수를 분석한 결과 HE_Ukal 변수의 결측치 개수가 700개 이상이므로 신뢰할 수 없다고 판단되어 제거하였다.

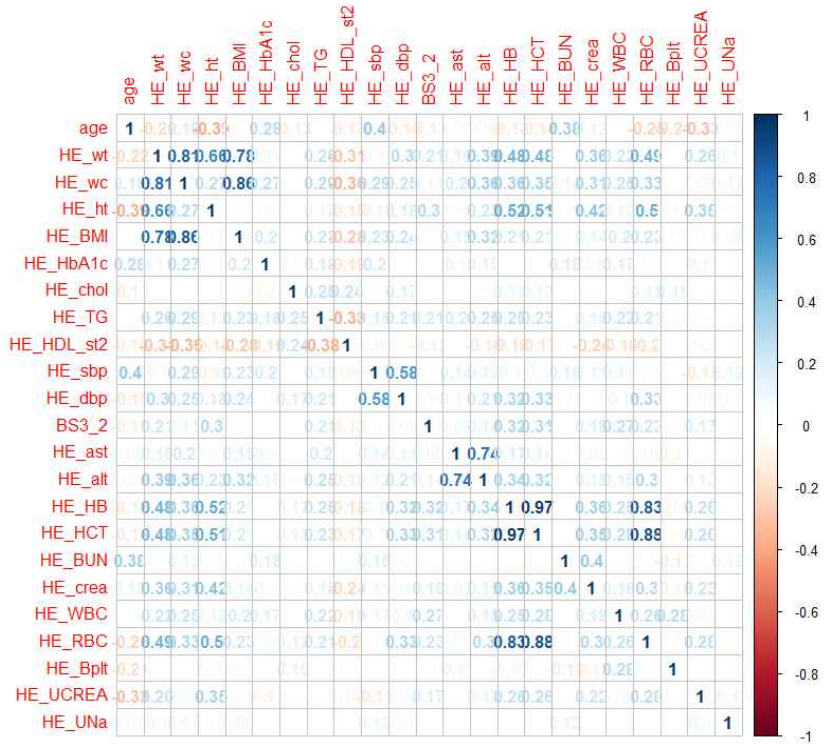


하지만, HE_Ukal 변수가 제거된 데이터 셋 역시 결측치가 존재하므로 모형 적합을 위해서는 결측치 처리 과정이 필요하다. 그러므로, mice 라이브러리를 통해 결측치 대체 과정을 진행하였다. 결측치 대체 방법에는 여러 방법이 존재하지만, mice는 연속형과 factor형 모두 가능하며 정확도가 높으므로, 해당 라이브러리를 사용하였다. 해당 그래프를 통해 mice를 적용 후 데이터는 결측치가 존재하지 않음을 알 수 있다.



(2) 2차 변수 선택 - Multi Logistic Regression 기법

모형 적합에 들어가기에 앞서 다중 공선성을 방지하기 위해 변수 간의 상관계수를 구하였다. 이때, 절댓값 0.7 이상인 경우 해당 변수 간 다중공선성이 존재한다고 보아 변수 간 다중공선성 문제를 해결하기 위해 HE_wt, HE_wc, HE_ast, HE_HB, HE_HCT 변수를 삭제하였다. 그러나, 예측을 위해 사용된 모델들은 tree 기반의 모델로 다중공선성 영향이 적어 변수를 삭제하여도 효과가 나타나지 않았다. 그러므로, 다중공선성을 고려하지 않고 모든 변수를 사용하기로 결정하였다.

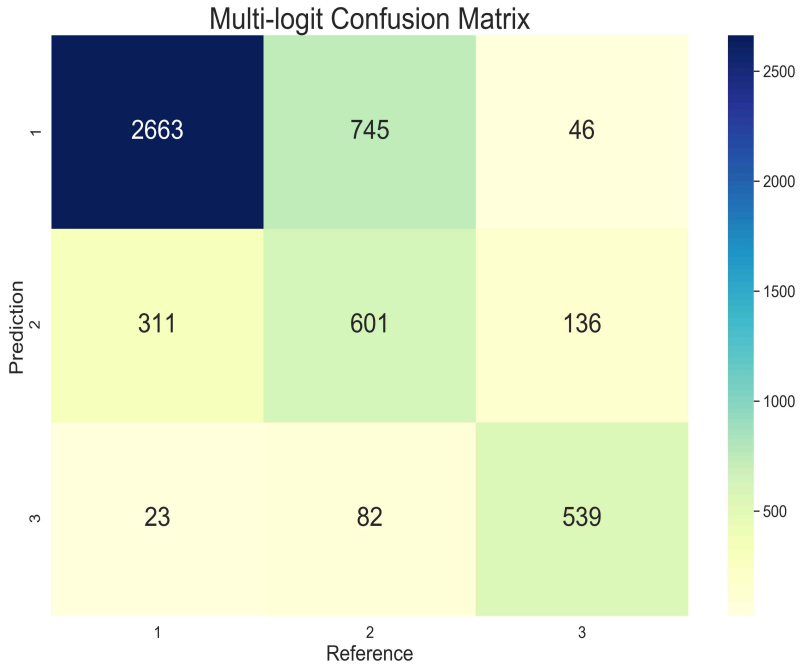


이후, 해당 데이터셋에서 회귀 모형을 통해 모형 적합에 사용될 독립변수를 선택하였다. 이때, 종속변수 Y인 HE_DM의 levels = 3이므로 다중 로지스틱 회귀모형을 사용하였고, HE_DM은 HE_glu 변수를 cutoff하여 만든 범주형 변수이므로 일종의 순서형 변수라고 볼 수 있다. 그러므로 sratio 옵션을 사용하여 연속비 로짓 모형을 사용하였다. 적합한 회귀식을 유도하기 위해 전진선택법, 후진제거법, 단계선택법을 적용해본 결과 후진제거법과 단계선택법으로 적합된 회귀모형은 동일하였고, 전진선택법으로 적합한 회귀모형만 차이가 존재하였다. 그러므로, 두 모형 간 우도비 검정을 사용하였다. 검정 결과, p-value=0.6689로 0.05보다 크므로 단계선택법으로 적합된 회귀모형을 채택하였다. 채택된 변수들은 다음과 같다.

항	Estimate	
	:1	:2
intercept	2.608e+01	1.585e+01
age	-1.020e-02	3.374e-03
HE_wc	-6.643e-03	-7.017e-02

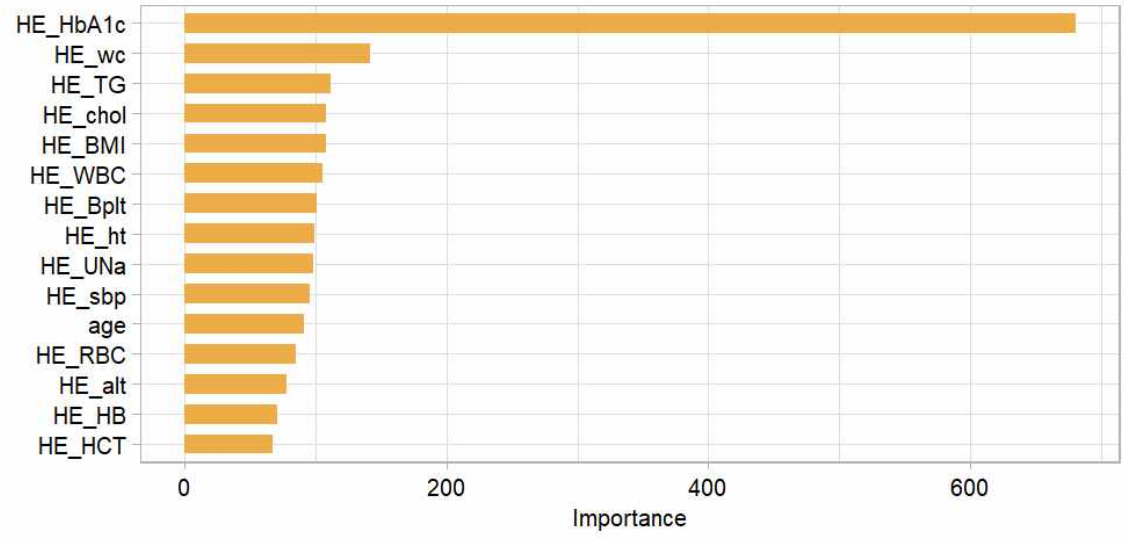
HE_ht	-2.474e-02	7.817e-03
HE_BMI	-7.996e-02	1.842e-01
HE_HbA1c	-3.004e+00	-3.488e+00
HE_chol	2.272e-03	1.541e-02
HE_TG	-1.772e-03	-2.353e-03
HE_sbp	-1.149e-02	-8.719e-05
HE_alt	-1.455e-04	1.037e-02
HE_HB	-6.490e-01	-6.091e-01
HE_HCT	1.432e-01	2.281e-01
HE_WBC	6.842e-02	-3.853e-02
HE_RBC	5.627e-01	3.674e-01
HE_Bplt	-7.601e-04	3.599e-03
HE_UNa	-3.974e-04	3.450e-03

적합된 다중 로지스틱 회귀모형의 Confusion Matrix는 다음과 같으며, 해당 모형의 정분류율은 73.965%이다.

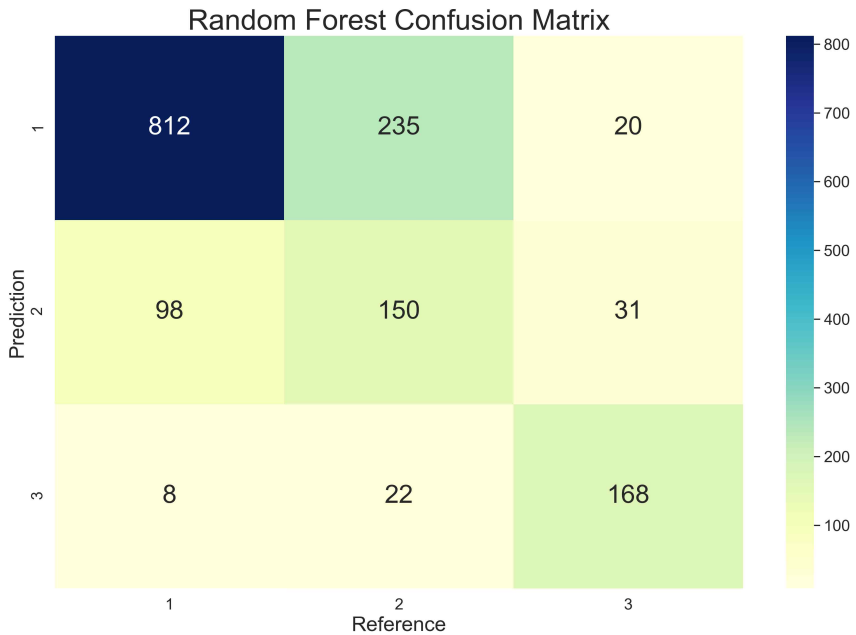


(3) Random Forest 기법

다항 로지스틱 회귀모형을 통해 선택된 변수를 기반으로 랜덤 포레스트 모형을 적합하였다. 랜덤 포레스트 기법은 다수의 의사결정나무모델에 의해 예측을 종합하는 앙상블 방식으로, 예측 변동성이 줄어들어 과적합을 방지 가능하다는 장점이 있다. 적합된 랜덤 포레스트 모형의 변수 중요도는 다음과 같다. 당뇨병과 연관성이 높은 HE_HbA1c(당화혈색소) 변수의 중요도가 확연히 높은 것을 확인할 수 있다.

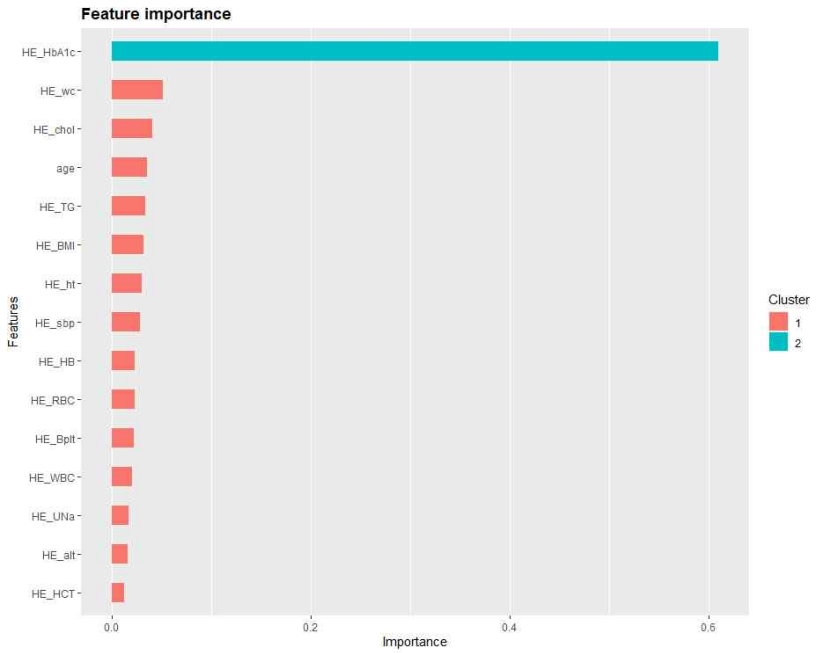


적합된 랜덤 포레스트 모형의 Confusion Matrix는 다음과 같으며, 해당 모형의 정확도는 72.67%이다.

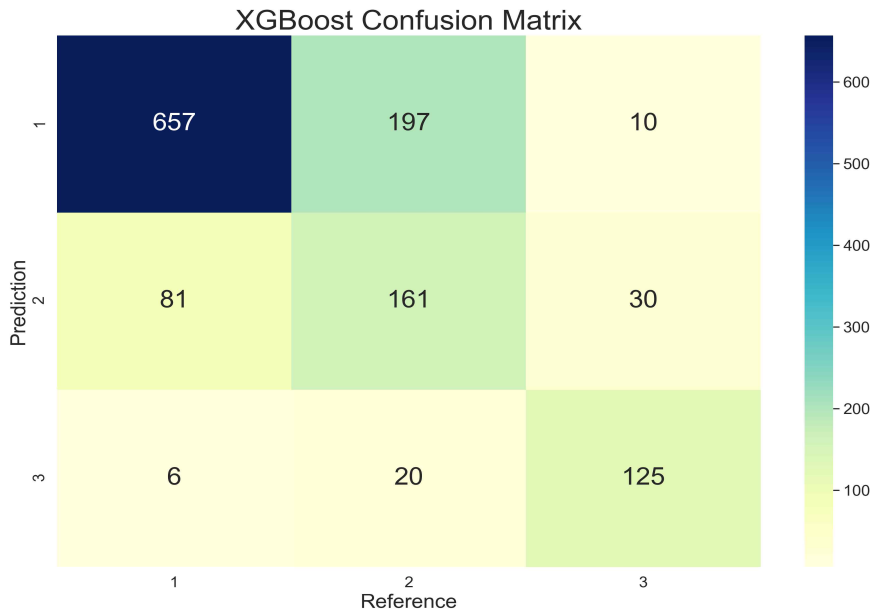


(4) XG Boost 기법

다항 로지스틱 회귀모형을 통해 선택된 변수를 기반으로 XGBoost 모형을 적합하였다. XGBoost 기법은 의사결정나무 기반의 앙상블 기법으로, 회귀 혹은 분류와 같은 예측 분석에 사용되며, 과적합을 방지 가능하다는 장점이 존재한다. 적합된 XGBoost 모형의 변수 중요도는 다음과 같다. 랜덤 포레스트 모형과 같이 당뇨병과 연관성이 높은 HE_HbA1c(당화혈색소) 변수의 중요도가 확연히 높은 것을 확인할 수 있다.

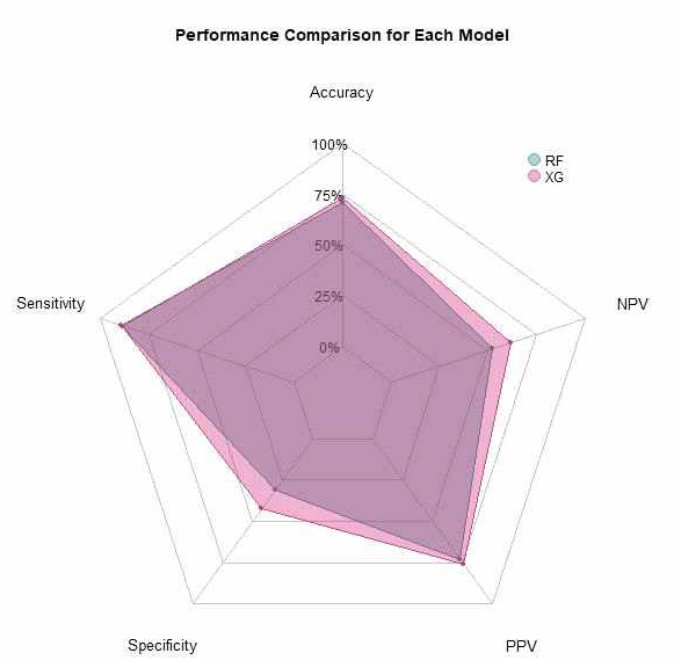


적합된 XGBoost 모형의 Confusion Matrix는 다음과 같으며, 해당 모형의 정확도는 73.27%이다.



3. 결론

본 연구에서는 제 7기 국민건강영양조사 데이터셋을 활용하여 당뇨병 진단에 직접적으로 연관된 공복혈당을 제외한 데이터를 통해 당뇨병 예측 모형을 적합하여, 당뇨병 잠재환자 예측에 기여하고자 하였다. 예측 모형을 적합하기 위해 과적합을 방지하고자, Random Forest 기법과 XGBoost 기법을 사용하였으며, 해당 모형을 비교한 결과 XGBoost 기법을 사용한 모형이 좀 더 정확도가 높았다.



하지만, 두 모형 모두 정확도가 70%대로 기대한 수치에 도달하지 못하였다. 연구 과정을 고찰한 결과 다음과 같은 한계점으로 인해 해당 결과가 도출되었다고 보았다.

1. 데이터셋의 변환 과정 누락

sex 변수를 제외한 데이터는 모두 수치형으로 해당 변수들의 범위가 제각각이므로 데이터 간 왜도와 첨도를 확인하여 일정 기준 이상 시 변환 과정을 거쳐야 했다. 하지만, 해당 과정이 생략되어있기 때문에 모델이 잘 학습하지 못하였을 가능성이 존재한다.

2. 종속변수 선택

본 연구에서 종속변수로 사용된 HE_DM은 공복혈당을 의미하는 HE_glu 변수를 cutoff하여 범주화한 변수로 로지스틱 모형을 적합하는 과정에서 종속변수가 순서형 변수임을 고려하면서 모형 적합과 해석에 어려움이 존재하였다. 이때, 종속변수를 HE_DM이 아닌 HE_glu로 모형을 적합시켰다면 모형 적합이 좀 더 수월했을 것으로 예측된다.

3. XGBoost 기법 옵션 설정

XGBoost 기법은 모형 적합 시 옵션에 따라 정확도가 최대 20%까지도 달라질 수 있다. 그렇기 때문에, 파라매터 튜닝을 통해 해당 모형의 적합한 옵션을 찾아야 한다. 하지만, 해당 과정

이 부족하였기 때문에 모델이 잘 학습하지 못하였을 가능성이 존재한다.

향후 연구에서는 앞서 서술한 한계점을 보완하고, 다양한 모형 적합 기법을 사용하여 좀 더 정확한 모형을 예측하고자 한다. 그뿐만 아니라, 정확도가 높은 모형 적합에 성공한다면 해당 모형과 다른 변수를 통해 당뇨병 환자의 합병증도 예측할 수 있는 모형을 적합하고자 한다.

4. 참고문헌

김윤정, 2018, 건강 지표 연관성을 활용한 웹 인터페이스 구축 및 질병 예측 연구
이수경, 2012, 노인 만성질환자의 건강관련 삶의 질 영향요인 분석 및 예측모델 개발

5. 별첨

```
> summary(fit1)

Call:
vglm(formula = HE_DM ~ age + HE_wc + HE_ht + HE_BMI + HE_HbA1c +
    HE_chol + HE_TG + HE_sbp + HE_alt + HE_HB + HE_HCT + HE_WBC +
    HE_RBC + HE_Bplt + HE_UNa, family = sratio(parallel = F),
    data = NA_data)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  2.608e+01  1.231e+00  21.182  < 2e-16 ***
(Intercept):2  1.585e+01  2.271e+00   6.981 2.94e-12 ***
age:1         -1.020e-02  3.933e-03  -2.594 0.009486 **
age:2          3.374e-03  8.580e-03   0.393 0.694100
HE_wc:1       -6.643e-03  9.778e-03  -0.679 0.496912
HE_wc:2       -7.017e-02  1.934e-02  -3.629 0.000285 ***
HE_ht:1       -2.474e-02  5.947e-03  -4.161 3.17e-05 ***
HE_ht:2        7.817e-03  1.149e-02   0.681 0.496182
HE_BMI:1      -7.996e-02  2.529e-02  -3.162 0.001568 **
HE_BMI:2       1.842e-01  4.990e-02   3.690 0.000224 ***
HE_HbA1c:1    -3.004e+00  1.194e-01 -25.164  < 2e-16 ***
HE_HbA1c:2    -3.488e+00  1.758e-01 -19.837  < 2e-16 ***
HE_chol:1      2.272e-03  1.009e-03   2.251 0.024355 *
HE_chol:2      1.541e-02  2.066e-03   7.460 8.65e-14 ***
HE_TG:1       -1.772e-03  4.181e-04  -4.238 2.25e-05 ***
HE_TG:2       -2.353e-03  6.308e-04  -3.730 0.000191 ***
HE_sbp:1      -1.149e-02  2.390e-03  -4.807 1.54e-06 ***
HE_sbp:2      -8.719e-05  4.562e-03  -0.019 0.984752
HE_alt:1      -1.455e-04  2.597e-03  -0.056 0.955312
HE_alt:2       1.037e-02  4.829e-03   2.147 0.031829 *
HE_HB:1       -6.490e-01  9.738e-02  -6.665 2.65e-11 ***
HE_HB:2       -6.091e-01  1.951e-01  -3.121 0.001801 **
HE_HCT:1       1.432e-01  4.093e-02   3.498 0.000470 ***
HE_HCT:2       2.281e-01  7.943e-02      NA       NA
HE_WBC:1       6.842e-02  2.413e-02   2.835 0.004584 **
HE_WBC:2      -3.853e-02  4.715e-02  -0.817 0.413821
HE_RBC:1       5.627e-01  1.797e-01   3.132 0.001733 **
HE_RBC:2       3.674e-01  3.738e-01   0.983 0.325609
HE_Bplt:1     -7.601e-04  6.277e-04  -1.211 0.225931
HE_Bplt:2      3.559e-03  1.270e-03   2.803 0.005058 **
HE_UNa:1      -3.974e-04  7.713e-04  -0.515 0.606355
HE_UNa:2       3.450e-03  1.613e-03   2.138 0.032497 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y=1|Y>=1]), logitlink(P[Y=2|Y>=2])

Residual deviance: 5999.565 on 10262 degrees of freedom

Log-likelihood: -2999.782 on 10262 degrees of freedom

Number of Fisher scoring iterations: 8

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1', '(Intercept):2', 'HE_HbA1c:1', 'HE_HCT:2'
```

Multi logistic Regression - Stepwise 모형