당뇨 위험요인 예측 모델 개발

과목명	데이터 마이닝	담당교수	이소현 교수님
요일	목 123	학과	경영정보학과
학번 / 이름	202012850 정유림 202012841 방가윤		

목차

1. 서론

1.1. 연구배경

식습관과 운동 부족으로 인한 비만 증가는 최근의 추세이며, 이로 인해 만성 생활습관 질환인 당뇨병의 발병도 증가하고 있다. 당뇨병은 한국인 사망 원인 상위권에 위치하며 다양한 합병증을 유발할 수 있는 병으로 알려져 있다. 이에 따라 당뇨병 연구와 당뇨병에 영향을 미치는 요소들에 대한 연구는 매우 중요하다.

당뇨병 예방과 각 요소에 대한 연구를 위해서는 다양한 분류 모델에 대한 이해가 필요하다. 특히 대규모 데이터에서 규칙이나 패턴을 찾아내는 데이터 마이닝 기법이 많이 활용되고 있다. 데이터 마이닝에서도 다양한 기법이 사용되며, RNN과 LSTM과 같은 시계열 데이터를 활용한 지도학습 방법을 적용한 모델들이 당뇨병 발병 예측이나 진단을 위한 시스템으로 제안되고 있다.

또한, Triglyceride, Cholesterol, FBS (Fasting Blood Sugar), BMI (Body Mass Index) 등이 당뇨병 발병 요인으로 알려져 있으며, 의사결정트리 알고리즘인 ID3, C4.5, 랜덤 포레스트 기법을 활용하여 발병 인자의 중요도와 발병 인자들 간의 관계를 분석하는 연구도 있다.

본 보고서에서는 k-최근접 이웃(K-Nearest Neighbors, KNN) 분석을 활용하여 당뇨병 예측을 수행하는 것을 목표로 한다. KNN은 지도학습의 일종으로, 주어진 데이터셋 내에서 새로운 데이터를 분류 또는 예측하는 데 사용된다. 이를 통해 우리는 당뇨병의 발병 여부를 예측하고, 개인들에게 적절한 조치 및 치료를 제공하는 데 도움을 줄 수 있다.

당뇨병 예측을 위한 KNN 분석을 통해 개인의 건강 상태를 파악하고 조치를 취함으로써, 당뇨병으로 인한 합병증 발생 가능성을 최소화하고 개인의 삶의 질을 향상시킬 수 있다. 따라서 이 연구의 결과는 당뇨병 예방 및 관리에 대한 정책 수립과 개인들의 건강 관리에 유용한 정보를 제공할 것으로 기대된다.

1.2. 연구 필요성

당뇨병은 우리나라 뿐만 아니라 전 세계적으로 건강 문제로 부각되고 있는 만성 질환이다. 이러한 당뇨병의 증가 추세는 현대 사회에서의 식습관 변화, 비만 문제, 운동 부족 등과 밀접한 관련이 있다. 이에 따라 당뇨병 예방과 조기 진단의 중요성이 더욱 부각되고 있으며, 이를 위한 연구의 필요성이 크게 대두되고 있다.

첫째, 당뇨병은 많은 사람들에게 심각한 건강 문제를 일으킨다. 당뇨병은 고혈당 상태가 지속되는 만큼, 혈관, 신장, 신경 등 다양한 장기에 영향을 미치며 합병증을 유발할 수 있다.

이로 인해 심혈관 질환, 신부전, 신경병증 등 다양한 합병증의 위험이 증가하게 된다. 따라서 당뇨병 예방과 조기 진단은 이러한 합병증 발생 가능성을 최소화하고 개인들의 건강을 보호하는데 중요한 역할을 한다.

둘째, 당뇨병은 사회적, 경제적 부담을 초래한다. 당뇨병은 개인의 건강 문제 뿐만 아니라 사회 전반에 큰 영향을 미친다. 당뇨병으로 인한 의료비, 휴가 일수 증가, 생산성 저하 등은 개인과 조직, 사회 전체에 큰 경제적 부담을 주는 결과를 초래할 수 있다. 이에 따라 정부와 보건 기관은 당뇨병 예방 및 조기 진단을 통해 사회적, 경제적인 비용을 줄이고 건강한 사회 구축을 위한 노력을 기울여야 한다.

셋째, KNN 분석을 활용한 당뇨병 예측은 개인 맞춤형 건강 관리에 큰 도움을 줄 수 있다. KNN 분석은 개인의 건강 상태를 파악하고 예방 조치를 취하는 데 유용한 도구로 활용될 수 있다. 개인의 건강 정보와 관련된 다양한 변수를 분석하여 당뇨병 발병 가능성을 예측함으로써, 개인들은 적절한 건강 관리와 예방 조치를 취할 수 있다. 이를 통해 개인의 건강 상태를 개선하고 당뇨병으로 인한 합병증 발생 가능성을 최소화할 수 있다.

넷째, KNN 분석은 당뇨병 예측에 대한 정확성과 신뢰성을 향상시킬 수 있는 분석 방법이다. KNN 분석은 주어진 데이터셋 내에서 가장 가까운 이웃들을 활용하여 예측을 수행하는 알고리즘으로, 데이터 간의 유사성을 고려하여 분류 또는 예측을 진행한다. 이를 통해 당뇨병 예측 모델의 정확성과 신뢰성을 높일 수 있으며, 개인들에게 더 정확하고 신뢰할 수 있는 예측 결과를 제공할 수 있다.

따라서, 당뇨병은 사회적으로 큰 문제로 부각되고 있으며, 이를 예방하고 조기 진단하기 위한 연구의 필요성이 크게 증가하고 있다. KNN 분석을 활용한 당뇨병 예측은 개인 건강 관리와 예방 조치에 유용한 도구로서 기능할 수 있다. 나아가 당뇨병 예측의 정확성과 신뢰성을 향상시키는 데 기여할 수 있다. 이를 통해 개개인의 건강을 지키고 사회적, 경제적 부담을 줄이는데 도움을 줄 수 있다. 따라서 본 연구는 당뇨병 예방과 조기 진단을 위한 중요한 연구로서 그 필요성을 가지고 있다.

1.3. 연구 목적

본 연구에서는 KNN 분석을 활용하여 당뇨병 예측을 수행하고, 이를 통해 개인들에게 적절한 건강 관리와 예방 조치를 제시하는데 기여하고자 한다. 연구 결과는 당뇨병 예방과 관리에 대한 정책 수립 및 개인 건강 관리에 유용한 정보를 제공할 것으로 기대된다.

2. 본론

2.1. 연구방법

2.1.1. 개념설명: K-최근접 이웃(KNN), 의사결정트리, 랜덤포레스트

k-최근접 이웃(k-nearest Neighborhood)

: KNN은 신호와 이미지 분류에 널리 사용되는 간단하고 직관적인 분류기다. K-최근접 이웃 알고리즘은 새로운 데이터가 주어졌을 때, 기존 데이터 중에서 가장 가까운 k개의 데이터를 기반으로 새로운 데이터를 예측하는 방식이다. 즉, 새로운 샘플을 기준 데이터와 비교하여 결정을 내린다. 이러한 유사한 특성을 가진 데이터는 유사한 범주에 속하는 경향이 있다는 가정하에 사용된다. 일반적으로 레이블이 지정되지 않은 특정 X에 대해 KNN 규칙은 훈련 데이터 세트에서 가장 가까운 k 개의 레이블이 지정된 데이터를 찾고, k 개의 데이터에서 가장 자주 나타나는 클래스를 X에 할당한다.

의사결정 트리(Decision Tree)

: 의사결정 트리는 의사 결정 분석에서 사용되며, 의사결정 과정을 시각적으로 이해하기 쉽게 보여줄 수 있다. 이름에서 알 수 있듯이 나무와 같은 모델을 사용하여 결정을 수행한다. 데이터 마이닝에서 널리 사용되는 도구 중 하나로, 특정 목표에 도달하기 위한 전략을 도출하는 데 사용된다. 결정 트리는 분류(Classification)와 회귀(Regression) 모두 가능한 지도 학습 모델로서, 각 레벨에서 (예/아니오) 질문을 연속적으로 말단 노드에 이를 때까지 진행하면서 학습한다. 특정 기준에 따라 데이터를 구분하는 모델로, 분기할 때마다 변수 영역을 두 개로 구분한다. 결정 트리에서 질문이나 정답을 담은 네모 박스를 노드라고 하며, 가장 처음 만나는 노드를 루트 노드, 가장 마지막 노드를 터미널 노드 또는 잎 노드라고 한다.

랜덤 포레스트(Random Forest)

: 의사결정 트리의 약점은 오버피팅(Overfitting)될 가능성이 크다는 것이다. 이를 해결하기 위해 랜덤 포레스트는 앙상블(Ensemble) 기계학습 모델로 여러 개의 결정 트리를 생성하여 새로운 데이터를 각 트리에 동시에 통과시킨다. 각 트리의 분류 결과에 대해 투표를 실시한 후, 가장 많은 득표를 받은 결과를 최종 분류 결과로 판단한다. 랜덤 포레스트는 많은 수의 트리를 생성함으로써 오버피팅을 예방할 수 있다. 따라서 일부 트리가 오버피팅될 수는 있지만, 전체 앙상블 모델의 예측력이 향상된다. 이러한 특징을 가진 랜덤 포레스트는 결정 트리의 한계를 극복하고 예측 성능을 향상시킬 수 있는 강력한 분류 모델이다.

2.1.2. 데이터 소개

이 보고서는 데이콘 당뇨병 위험 분류 예측 경진대회에서 제공되는 데이터를 사용하였다. 데이터 셋은 임신 횟수, 포도당 농도, 혈압, 피부 두께, 인슐린, 체질량 지수, 당뇨병 혈통 기능, 나이 등의 변수들로 구성되어 있다. 데이터 셋에는 총 8개의 입력 변수(Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age)와 1개의 출력 변수(Outcome)가 있다.

<변수 목록>

Pregnancies : 임신 횟수
Glucose : 포도당 농도
BloodPressure : 혈압
SkinThickness : 피부두께
Insulin : 인슐린
BMI : 체질량지수
DiabetesPedigreeFunction : 당뇨병 혈통 기능
Age : 나이
Outcome : 당뇨병 여부(0은 발병되지 않음, 1은 발병)

2.1.3. 분석과정

1. 데이터 전처리

- 결측치 처리: 데이터 셋에서 결측치를 확인하였으나 존재하지 않았다.
- 이상치 처리: 데이터 셋에서 이상치를 확인하였으나 존재하지 않았다.

2. 변수 정리 및 파생 변수 생성

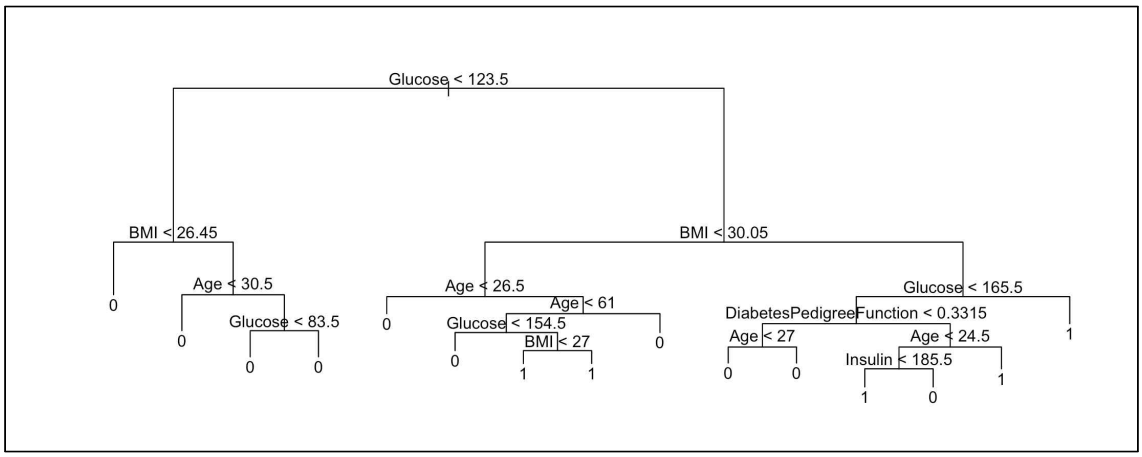
- 변수 변환 : 정제된 데이터라 변수를 그대로 유지했다.
- 데이터 정규화: 정제된 데이터라 데이터 정규화 과정은 생략했다.

3. 가설 수립

"당뇨에 가장 큰 영향을 주는 요소는 BMI요소일 것이다"라는 가설을 수립했다. 가설을 수립하게 된 배경은 "메디칼타임즈"에서 발표된 뉴스글을 활용하였는데 해당 뉴스글은 2023년 미국임상내분비학회(AACE)의 제2형 당뇨병(T2D) 관리 알고리즘 합의문을 다루고 있다. 뉴스글에 따르면 AACE는 당뇨병 관리를 위한 핵심 축으로 비만을 지목하면서 체중 관리의 중요성을 강조하고 있다. 다시말해, AACE의 합의문에서는 당뇨병 관리를 위해 체중 조절을 강조하고체중에 따른 약제 선택 기준을 제시하고 있다. 더불어 다양한 연구에서 생활습관 개입을 통한 체중 감량이 혈당, 이상지질혈증, 혈압, 심혈관 질환 등을 개선시키는 것으로 확인되었다. 뉴스글과 이전 연구 결과를 종합적으로 고려하면, 비만과 체중 관리는 당뇨병 예측 및 관리에 중요한 요소로 작용하는 것으로 판단된다. 이에 따라 가설 "당뇨 가장 큰 영향을

주는 요소는 BMI일 것이다."는 일부 검증되었다고 판단하였고, BMI와 당뇨병 간의 관련성을 더 근거화하기 위해 추가적인 분석을 하려고 한다.

2.1.4. 의사결정트리 분석과정 및 결과

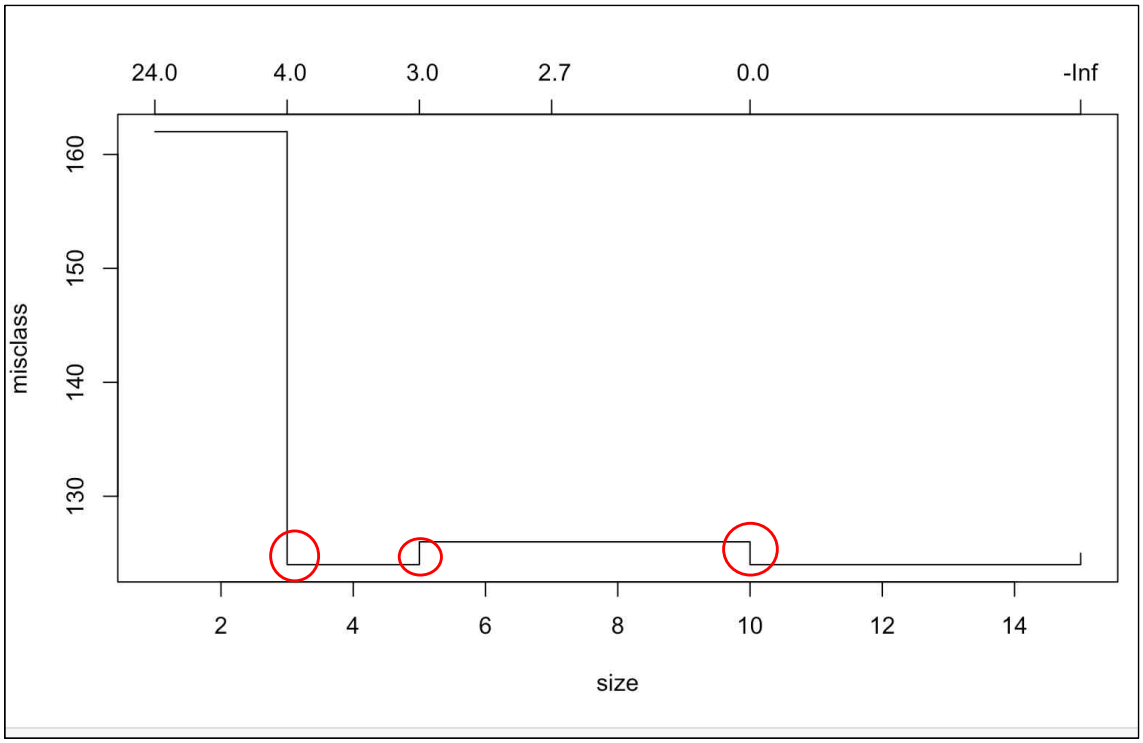


<최적값 찾기 전 의사결정 트리>

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 288  70
         1  13  85

               Accuracy : 0.818
                 95% CI : (0.7794, 0.8523)
    No Information Rate : 0.6601
    P-Value [Acc > NIR] : 4.987e-14

                  Kappa : 0.5547

 Mcnemar's Test P-Value : 7.906e-10

            Sensitivity : 0.9568
            Specificity : 0.5484
         Pos Pred Value : 0.8045
         Neg Pred Value : 0.8673
             Prevalence : 0.6601
         Detection Rate : 0.6316
   Detection Prevalence : 0.7851
      Balanced Accuracy : 0.7526

       'Positive' Class : 0
```

<최적값 찾기 전 accuracy>

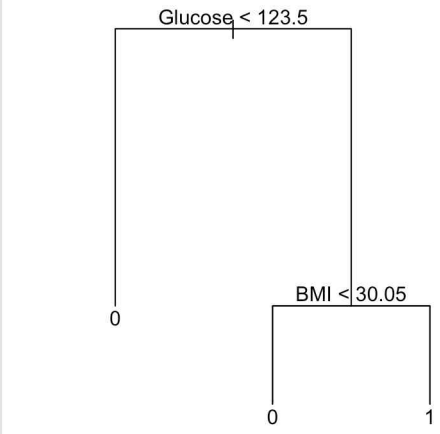
최적값 찾기 전 Accuracy 값은 0.818로 높은 편이지만, 트리의 깊이가 깊어 과적합의 가능성이 높아보인다. 이는 오히려 모델의 성능을 떨어트릴 수 있어 최적값을 찾는 과정이 필요하다.



<의사결정 트리에서 최적값 찾는 과정>

그래프에서 y값이 가장 낮아지는 지점이 3, 5, 10이다. 이러한 k값들을 모두 적용해보았다.

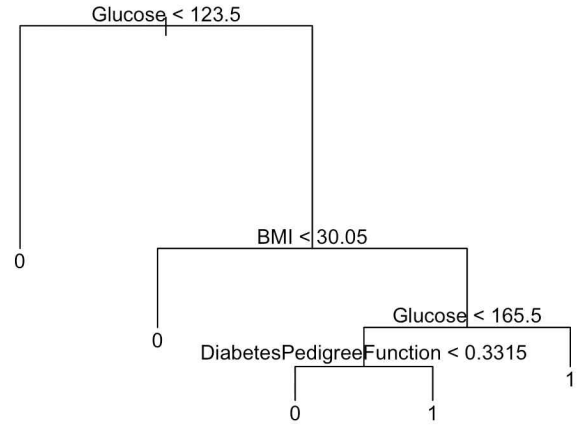
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 104  27
         1  19  46

               Accuracy : 0.7653
                 95% CI : (0.6996, 0.8228)
    No Information Rate : 0.6276
    P-Value [Acc > NIR] : 2.664e-05

                  Kappa : 0.4865

 Mcnemar's Test P-Value : 0.302

            Sensitivity : 0.8455
            Specificity : 0.6301
         Pos Pred Value : 0.7939
         Neg Pred Value : 0.7077
             Prevalence : 0.6276
         Detection Rate : 0.5306
   Detection Prevalence : 0.6684
      Balanced Accuracy : 0.7378

       'Positive' Class : 0
```



< k=3일 때 의사결정 트리>

k = 3 인 경우 accuracy는 0.7653으로 셋 중 가장 높지만 의사결정 트리가 너무 얇아 부적합하다.

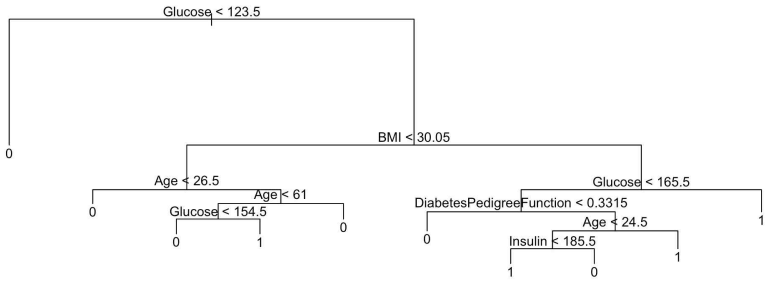
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 112  42
         1  11  31

               Accuracy : 0.7296
                 95% CI : (0.6617, 0.7904)
    No Information Rate : 0.6276
    P-Value [Acc > NIR] : 0.001643

                  Kappa : 0.3669

 Mcnemar's Test P-Value : 3.775e-05

            Sensitivity : 0.9106
            Specificity : 0.4247
         Pos Pred Value : 0.7273
         Neg Pred Value : 0.7381
             Prevalence : 0.6276
         Detection Rate : 0.5714
   Detection Prevalence : 0.7857
      Balanced Accuracy : 0.6676

       'Positive' Class : 0
```



< k=5일 때 의사결정 트리>

k=5인 경우 의사결정 트리의 형태의 깊이는 적당했으나 accuracy값이 0.7296으로 k=10인 경우보다는 정확도가 떨어진다.

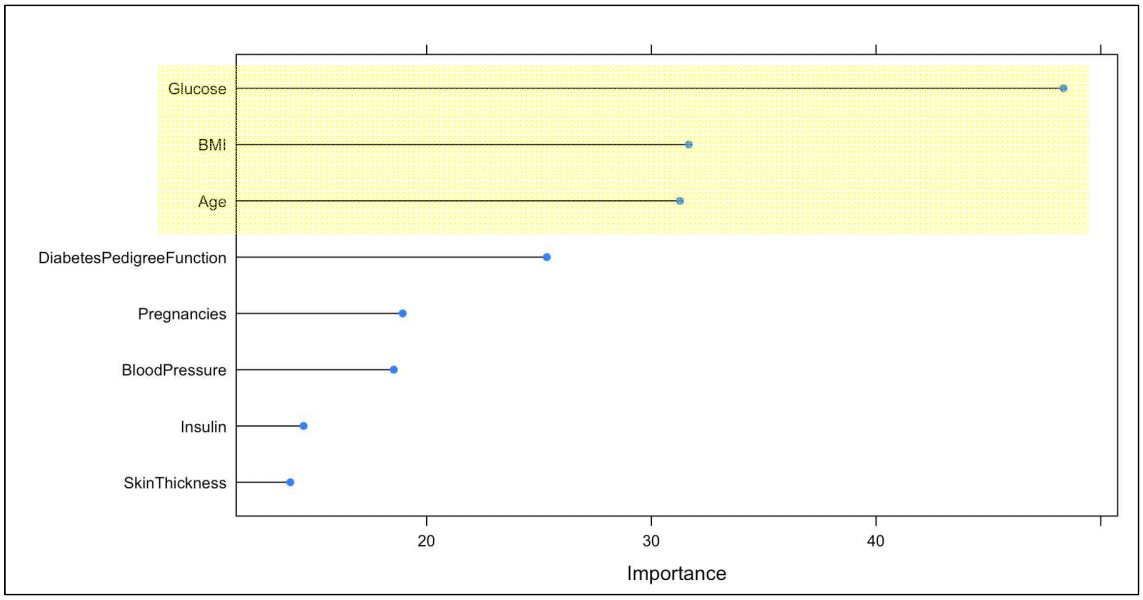
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 113  40
         1  10  33

               Accuracy : 0.7449
                 95% CI : (0.6779, 0.8043)
    No Information Rate : 0.6276
    P-Value [Acc > NIR] : 0.0003273

                  Kappa : 0.4045

 Mcnemar's Test P-Value : 4.11e-05

            Sensitivity : 0.9187
            Specificity : 0.4521
         Pos Pred Value : 0.7386
         Neg Pred Value : 0.7674
             Prevalence : 0.6276
         Detection Rate : 0.5765
   Detection Prevalence : 0.7806
      Balanced Accuracy : 0.6854

       'Positive' Class : 0
```



< k=10일 때 의사결정트리>

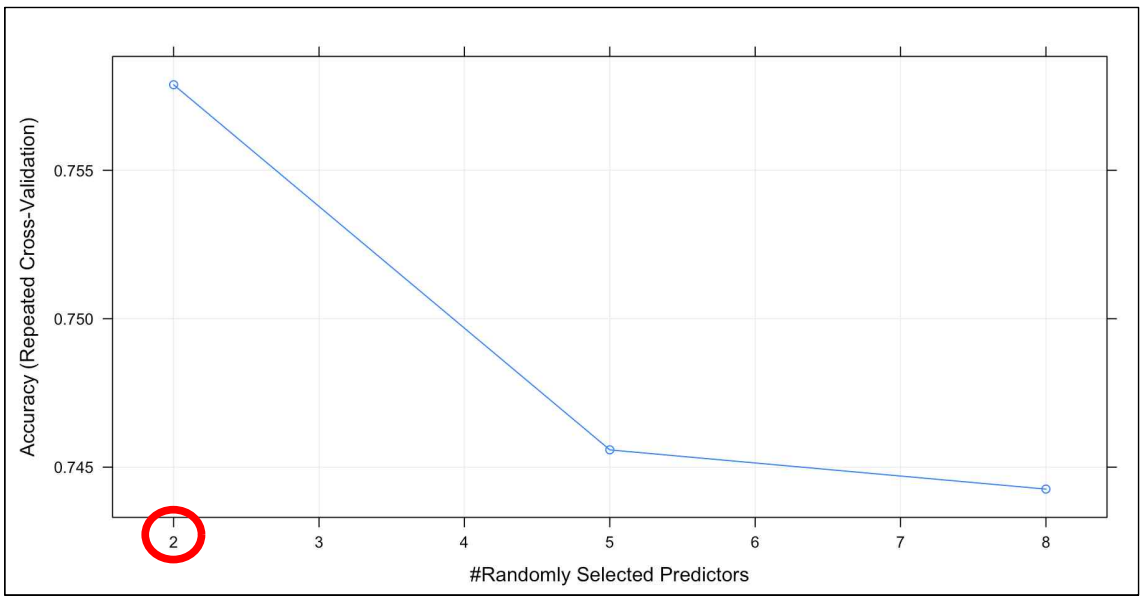
k=10인 경우 의사결정 트리의 형태는 깊이도 적당하고 변수도 다양하게 나타났다. 더 나아가 k=5일 때에 비해 accuracy가 0.7449로 더 높게 나타났다.

2.1.5. 랜덤포레스트 분석과정 및 결과



<랜덤 포레스트를 통해 당뇨병에 있어서 가장 영향력 있는 변수 조사>

위의 그래프를 보면 1위는 글루코스, 2위는 BMI수치, 3위는 나이로 글루코스 수치가 당뇨병에 미치는 영향이 가장 유의미하다는 것을 알 수 있다.



<랜덤포레스트 accuracy값 구하기>

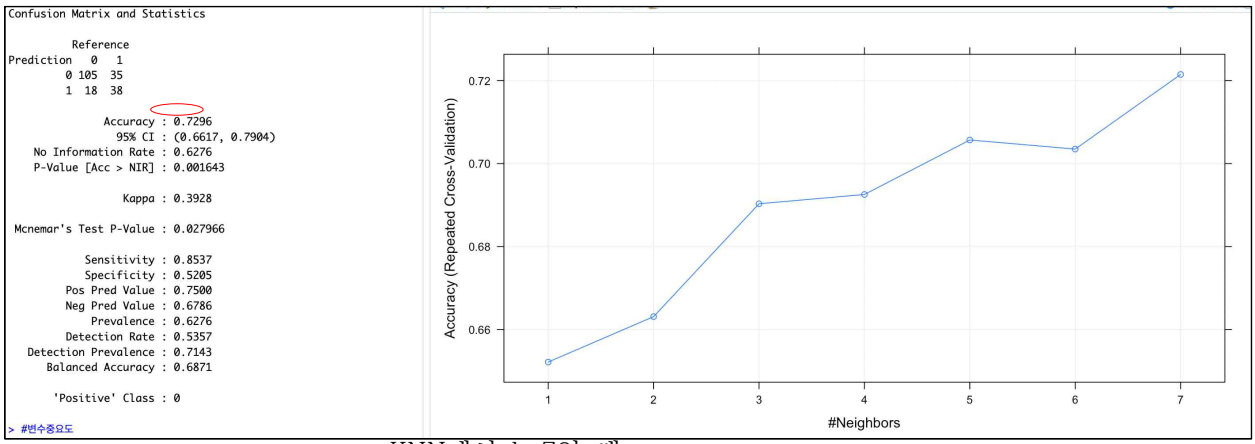
그래프를 보면 가장 랜덤포레스트 accuracy값이 가장 높을 때인 k가 2일 때 가장 크다는 것을 알 수 있다.

```
          Reference
Prediction   0   1
         0 104  29
         1  19  44

               Accuracy : 0.7551
                 95% CI : (0.6887, 0.8136)
    No Information Rate : 0.6276
    P-Value [Acc > NIR] : 9.841e-05

                  Kappa : 0.4611

 Mcnemar's Test P-Value : 0.1939

            Sensitivity : 0.8455
            Specificity : 0.6027
         Pos Pred Value : 0.7820
         Neg Pred Value : 0.6984
             Prevalence : 0.6276
         Detection Rate : 0.5306
   Detection Prevalence : 0.6786
      Balanced Accuracy : 0.7241

       'Positive' Class : 0
```

<랜덤포레스트 k=2일 때 accuracy>

결과값을 보면 k =2 일 때 정확도가 가장 높은 것을 알 수 있었다.

2.1.6. K-최근접 이웃(KNN) 분석과정 및 결과



<KNN에서 k=7일 때 accuracy>

결과값을 보면 k =7 일 때 accuracy가 0.7296으로 정확도가 가장 높은 것을 알 수 있었다.

3. 결론

3.1. 해석

"당뇨에 가장 큰 영향을 주는 요소는 BMI요소일 것이다"라는 예상과 달리, 당뇨에 가장 큰 영향을 끼치는 요소는 Glucose였다. 따라서 Glucose수치와 연관된 위험요인들을 관리하기 위한 애플리케이션을 개발해 당뇨병 환자들의 Glucose 수치를 조절하여 건강관리를 도울 수 있다.

또한 대학병원과 협업을 통해 당뇨와 관련된 합병증을 겪고 있거나, 중증도가 심한 환자들과 담당의가 소통할 수 있도록 한다. 의사들은 환자 개개인의 건강상태를 파악하고, 실시간으로 소통할 수 있다. 이를 통해 애플리케이션의 전문성을 더 높일 수 있고, 환자들의 건강관리 앱에 대한 신뢰를 높일 수 있다.

Glucose 수치에 영향을 미치는 위험요인에는 식단, 운동 부족, 비만, 유전, 스트레스, 약물 및 의약품 등이 있다. 이러한 위험요인들을 관리하기 위한 애플리케이션의 개요는 다음과 같다.

[애플리케이션 개요]

1. 혈당 모니터링: 사용자는 애플리케이션을 통해 혈당 측정 값을 기록할 수 있다. 이를 통해 사용자는 일상적인 식사, 운동, 스트레스 등과 혈당 수치 간의 상관 관계를 파악할 수 있다.

2. 식단 관리: 애플리케이션은 사용자에게 건강한 식단 관련 정보와 가이드라인을 제공한다. 사용자는 식사 기록을 할 수 있으며, 애플리케이션은 식단의 탄수화물 함량과 영양 성분을 분석하여 혈당 수치에 미치는 영향을 예측하고 조언을 제공한다.

3. 운동 계획 및 기록: 사용자는 애플리케이션을 통해 개인 맞춤형 운동 계획을 수립하고 운동 기록을 남길 수 있다. 운동은 혈당 조절에 도움이 되는데, 애플리케이션은 운동의 종류와 강도에 따른 혈당 변화를 분석하여 사용자에게 피드백을 제공한다.

4. 스트레스 관리: 애플리케이션은 스트레스 관리를 위한 기능을 제공한다. 사용자는 스트레스 상황을 기록하고 스트레스 관리 방법에 대한 조언과 심리적 지원을 받을 수 있다.

5. 데이터 분석 및 예측: 애플리케이션은 사용자의 혈당 수치, 식단, 운동, 스트레스 등의 데이터를 분석하여 개인 맞춤형 건강 조언을 제공한다. 또한, 개개인의 데이터를 기반으로 한 예측 모델을 활용하여 미래의 혈당 수치 예측을 제공한다.

6. 의사와의 소통 창구 마련: 환자는 애플리케이션을 통해 건강 상태를 게시한다. 건강 상태에 문제가 있다고 판별되는 경우 의사에게 알림이 가게 된다. 의사는 실시간으로 환자의 정보를 제공받아 적절한 진단을 내려준다.

3.2. 시사점

데이터 보안과 개인정보 보호: 사용자의 건강 정보와 개인정보를 보호하기 위해 강력한 보안 및 개인정보 보호 시스템을 구축해야 한다. GDPR와 같은 규정을 준수하고, 사용자의 동의를 얻는 절차를 강화해야 한다.

3.3. 한계점

1. 정확성과 신뢰성: 집에서 측정된 혈당 수치는 전문적인 의료 기기와 비교하여 정확성이 낮을 수 있다. 따라서 의료 전문가의 확인과 검사가 필요하다. 또한, 사용자의 부정확한 데이터 입력이나 잘못된 측정 방법으로 인해 예측이 부정확할 수 있다.

2. 개인 차이와 다양성: 각 개인의 신체 특성과 생활 습관은 혈당 수치에 영향을 미치는데, 애플리케이션은 일반적인 가이드라인을 기반으로 건강 조언을 제공한다. 그러나 각 개인의 차이와 다양성을 완전히 고려하기는 어렵다.

3. 기술적 제약과 접근성: 애플리케이션을 사용하기 위해서는 스마트폰 또는 태블릿과 인터넷 연결이 필요하다. 이는 일부 사용자에게 기술적 제약이나 접근성 문제를 일으킬 수 있다.

4. 의사와의 상호작용 부족: 애플리케이션은 사용자에게 일반적인 조언을 제공할 수 있지만, 의사와의 직접적인 상호작용이 부족하다. 중증도가 심한 환자나 합병증을 겪고 있는 환자들은 의료 전문가와의 실시간 소통이 중요하다.

4. 참고문헌

1. 김주호. "초기 당뇨병 예측을 위한 지능형 분류 알고리즘 비교 분석." 국내석사학위논문 전남대학교대학원, 2021. 광주

2. 당뇨병 관리 핵심은 '체중'…AACE, 주요 타깃 설정 2023년5월25일).메디컬타임즈.Retrievedfrom
https://www.medicaltimes.com/Main/News/NewsView.html?ID=1153485