



바 이 오 마 커
기 반

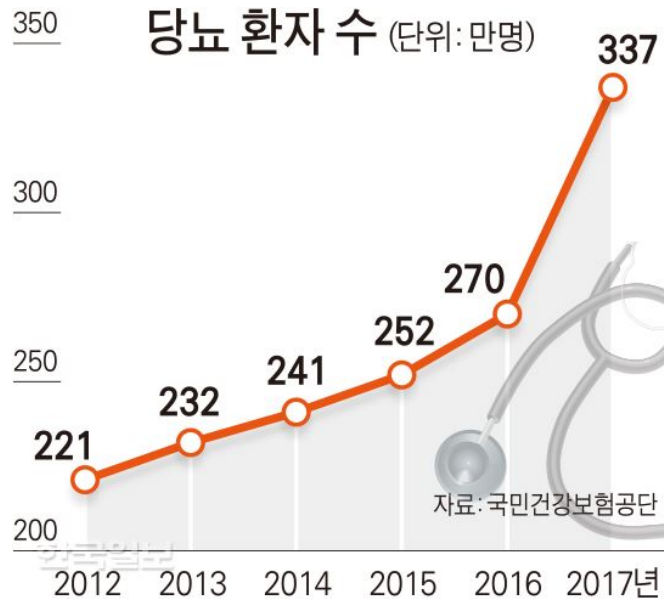
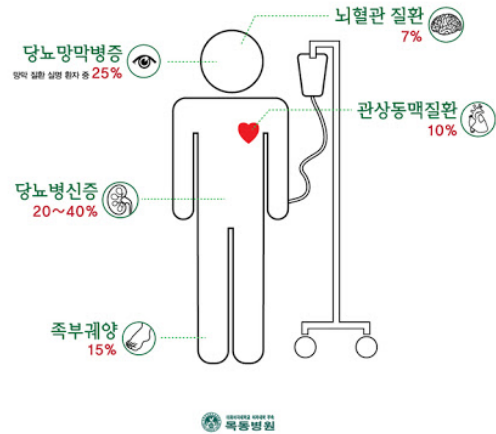
당 뇨 병
예 측 모 델

3 조

22000714 조 선 영

21600033 권 하 은

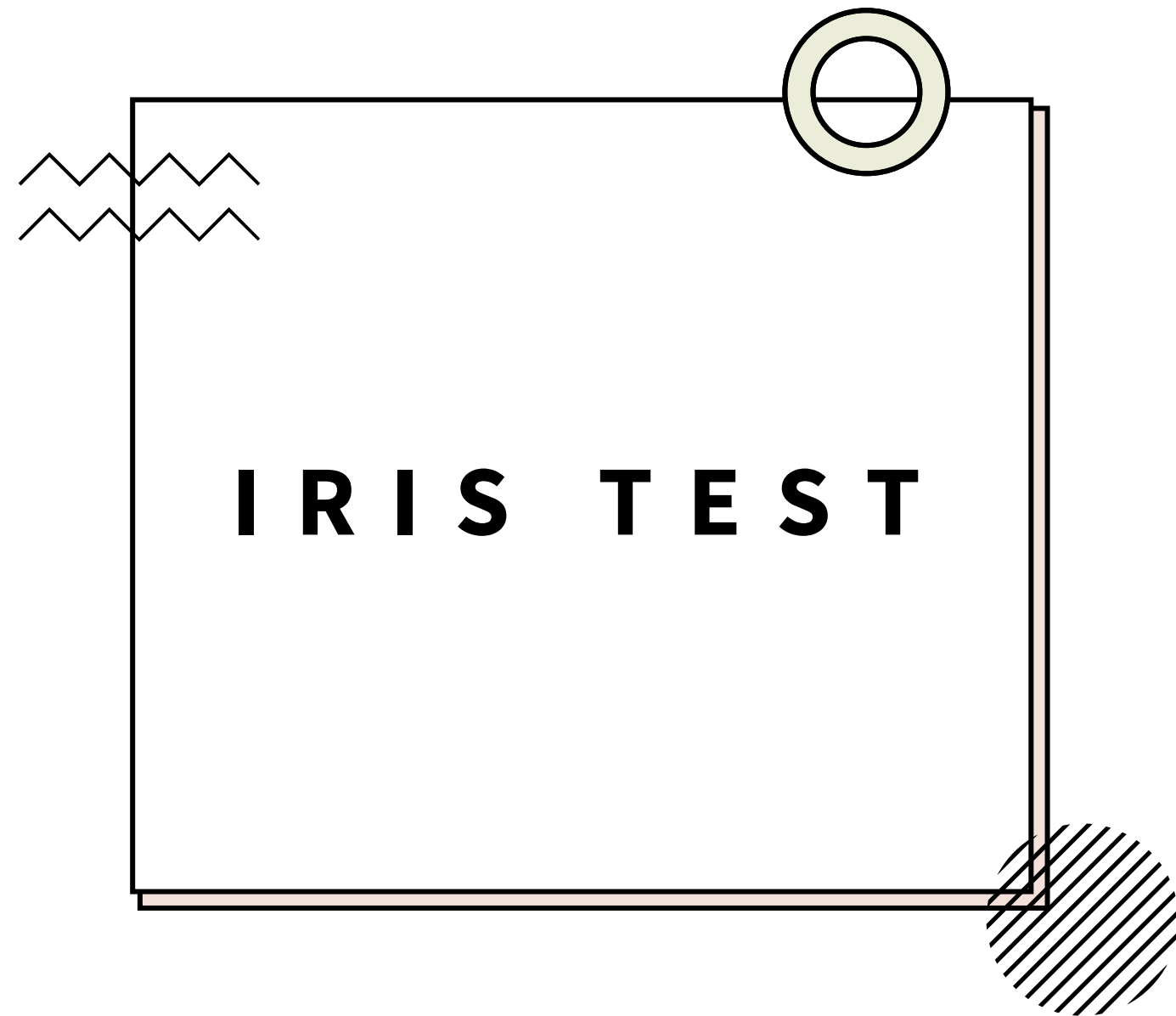
당뇨병 합병증 알림표



주제 소개 : 당뇨병

- 높은 혈당 수치가 오랜 기간 지속되는 대사 질환
- 세계 10대 사망 질환 → 당뇨병으로 인해 연간 400만명이 사망하고, 800조원의 치료비가 발생
- 진단 혹은 예측이 빠를수록 치료 가능성이 증가

➔ 기존 당뇨병 검사보다
간단하고 정확한 검사 필요



- 인슐린 저항성(Insulin resistance) 검사:
 - 제 2유형 당뇨병의 진단에 중요한 지표로 사용.
- 기존 인슐린 저항성 검사, 종류와 한계:
 - 인슐린 저항성 평가지표: 부정확하다
 - HOMA-IR: 공복인슐린의 변동폭이 크다
 - 인슐린 내성 검사: 시행하기 어렵다
- 본 프로젝트는 당뇨병 고위험군 피실험자의 **장내 미생물 데이터**를 활용하여 진행.
 - 혈액 샘플 + 변 샘플에서 얻은 데이터
 - 진단 뿐만 아니라 장 미생물총(Microbiota)을 통한 **치료에도 유의미한 결과** 검출 가능.
- 데이터 출처: iHMP Stanford School of Medicine (<http://med.stanford.edu/ipop.html>).



AI를 사용할 때

- 바이오 샘플의 10,780 개의 feature 를 **16개** 혹은 **17개의 유의미한 feature** 로 줄인다.

- Microbiota를 통한 실제 치료에 용이해짐

- **AUC**, 즉 분류 모델의 성능이 기존 10,780개 feature 이용 시 정확도 (0.98) 와 비슷한 수준을 유지하여 **정확한 결과를** 보장한다.

- 즉, **최소한의 정보, 최소한의 노력**으로 정확도가 높은 예측을 할 수 있다.

→ 당뇨병 **예방** 차원에서 큰 기대 효과 +
경제적 기대 효과 창출 가능



Introduction to
Adaboost




XGBOOST



Gradient
Boosting
Algorithm

Works



학 습 모 델

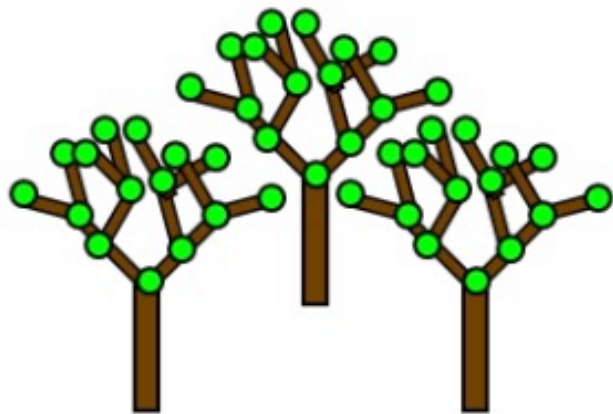
총 5가지의 학습 모델의 성능을
비교 분석



R A N D O M F O R E S T

Random Forest Classifier

Classification Technique

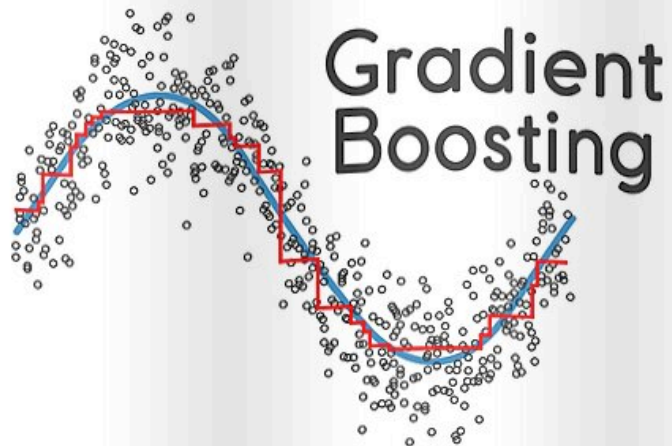


- 다수의 결정트리로부터 분류 또는 평균 예측치를 출력함으로써 동작.
- 예측의 변동성이 줄어들며, 과적합을 방지.



GRADIENT BOOSTING

- single leaf에서 학습을 시작.
- 타겟값에 대한 초기 추정 leaf이 tree를 타고가면서 error를 반영한 새로운 tree를 생성.
- 회귀 분석 혹은 분류 분석을 수행하는데 유용.
- 머신 러닝 알고리즘 중에서 가장 예측 성능이 높다고 알려짐.





A D A B O O S T

- 약한 학습기의 결과물들을 가중치를 두어 더하는 방법
- 잘못 분류된 것을 약한 학습기를 이용해 수정할 수 있음
- 다양한 상황에 적용(adaptive)할 수 있다.
- 이상점이 많은 데이터에 취약, 과적합 문제에는 덜 취약하다.





X G B O O S T

XGBoost

- Gradient Boosting 알고리즘을 분산 환경에서 실행할 수 있도록 구현해놓은 라이브러리.
- Regression, Classification 문제를 모두 지원.
- 성능과 자원효율이 좋아서 인기있는 알고리즘.

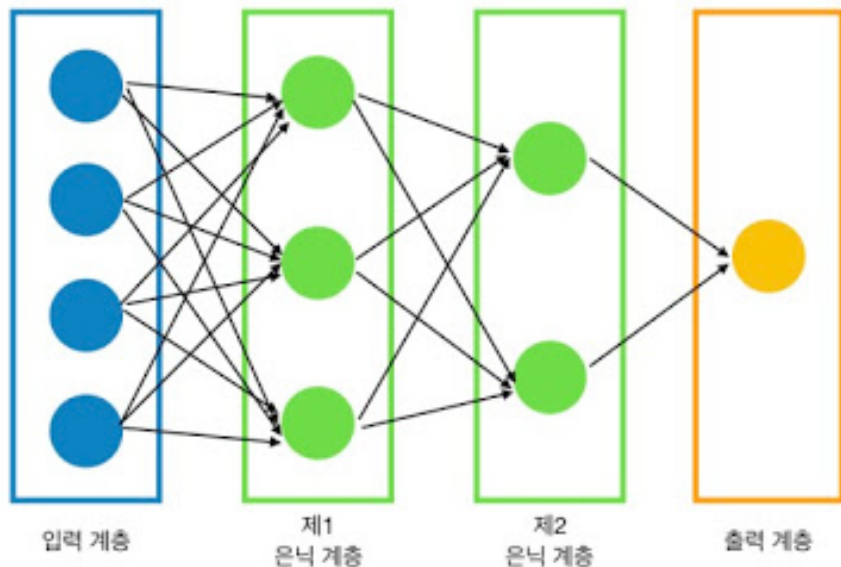




D N N

DNN

Deep Neural Network



- 연속형, 범주형 변수에 상관없이 모두 분석 가능한 알고리즘
- 입력 변수들 간의 비선형 조합이 가능하다.
- 예측력이 다른 머신러닝 기법들에 비해 상대적으로 우수한 경우가 많다.



코드 설명

```

header0_initialPadding = 10;

scrollTop() > header1_initialDistance) {
  header1.css('padding-top', 10) == header1_initialPadding +
  header1.css('padding-top', '' + $(window).scrollTop() - header1_initialDistance);

  header1.css('padding-top', '' + header1_initialPadding + 10);

  scrollTop() > header2_initialDistance) {
    header2.css('padding-top', 10) == header2_initialPadding +
    header2.css('padding-top', '' + $(window).scrollTop() - header2_initialDistance);

    header2.css('padding-top', '' + header2_initialPadding + 10);

    scrollTop() > header3_initialDistance) {
      header3.css('padding-top', 10) == header3_initialPadding +
      header3.css('padding-top', '' + $(window).scrollTop() - header3_initialDistance);

      header3.css('padding-top', '' + header3_initialPadding + 10);
    }
  }
}

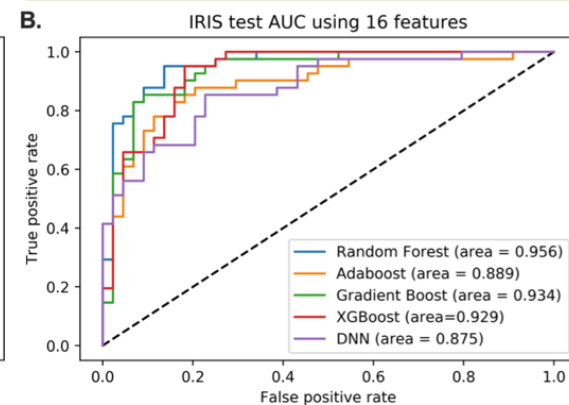
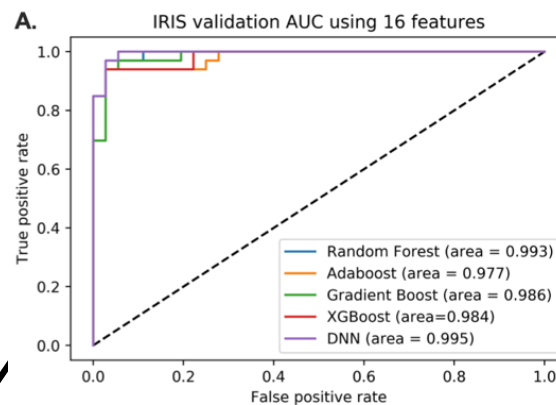
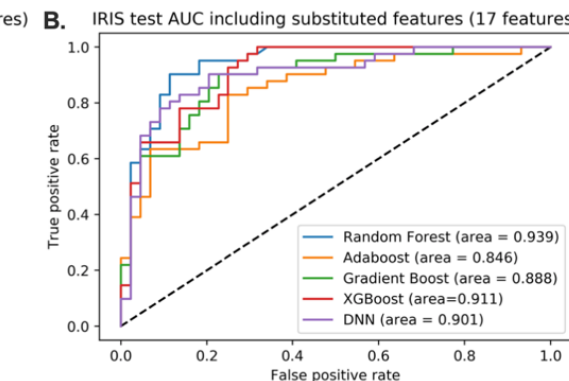
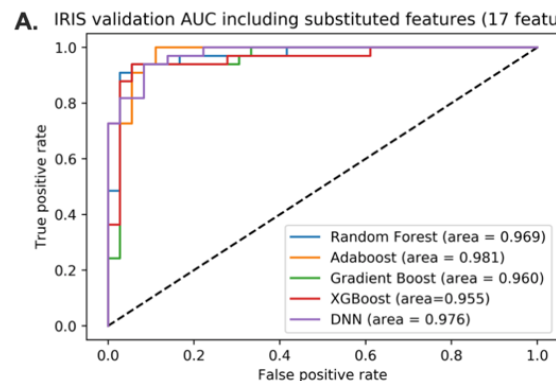
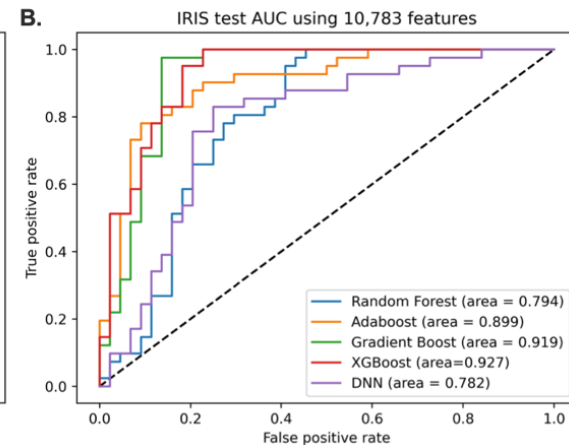
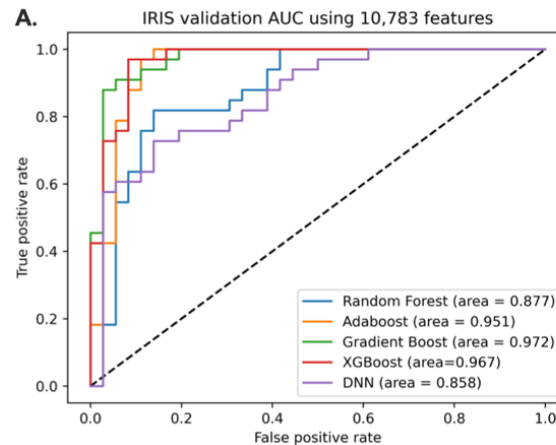
```

실행 결과

우 상단부터 10,783 개, 17개, 16개의 feature를 사용하여 Insulin Resistance와 연관성을 검사하였을 때의 AUC 그래프.

A는 validation set을 사용하여 제작한 AUC 그래프이며 B는 holdout dataset을 사용하여 제작한 AUC 그래프 이다.

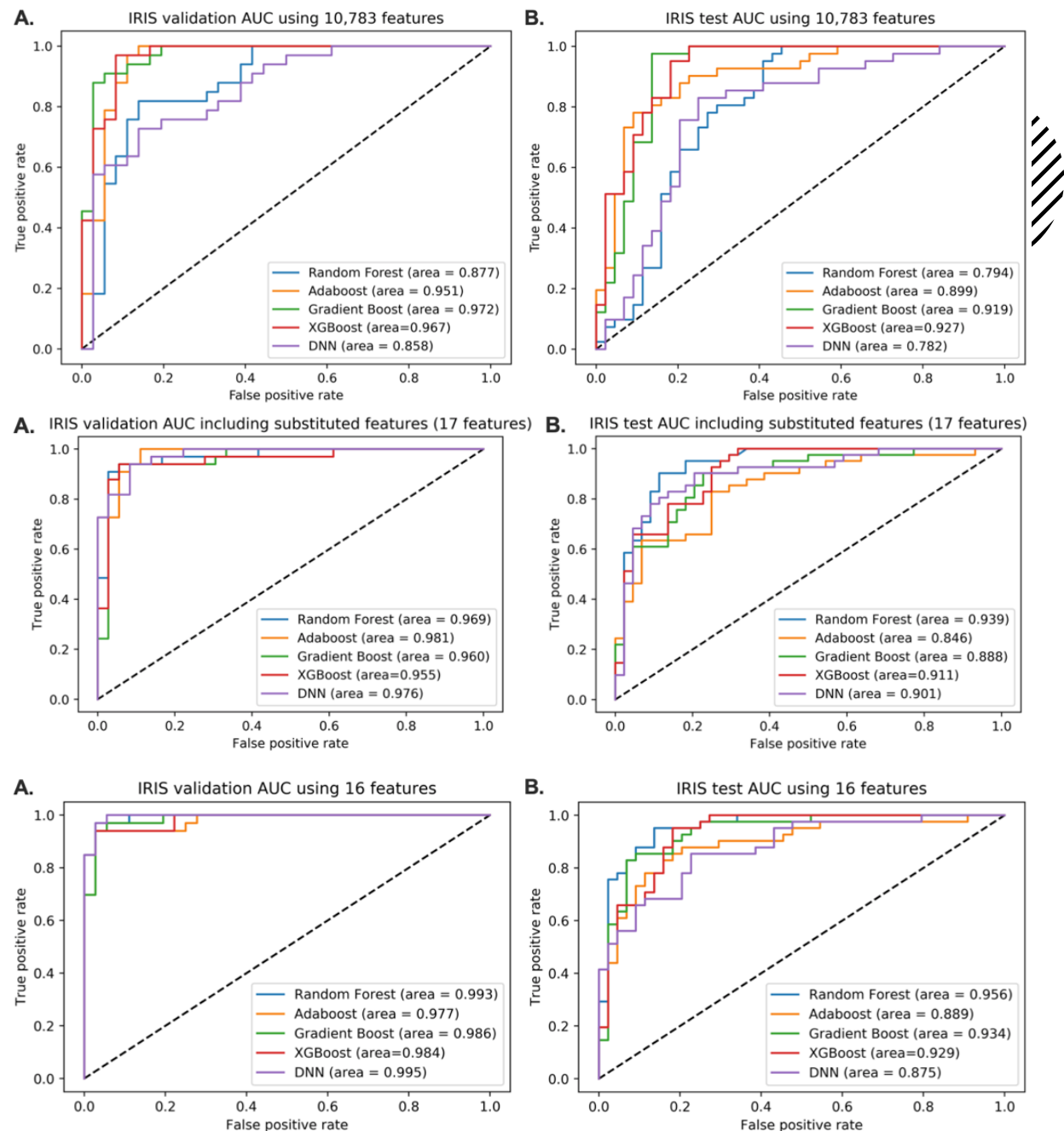
Feature reduction 을 수행하였을 때 True Positive rate, 즉 참 긍정성이 더욱 높은 수치를 기록하는 것을 확인할 수 있다.



- 10,780개 Feature 테스트에서 좋은 성능을 보여준 Gradient Boosting 이 Reduction 이후에도 가장 높은 수준의 AUC를 기록할 것으로 예상.

- 하지만 결과는 Random Forest 가 평균적으로 높은 AUC 수치를 기록함.

- 주 원인: feature의 개수가 많을 때 처리 성능이 떨어지던 Random Forest가 feature 개수의 감소로 예측 오차를 줄일 수 있었다.





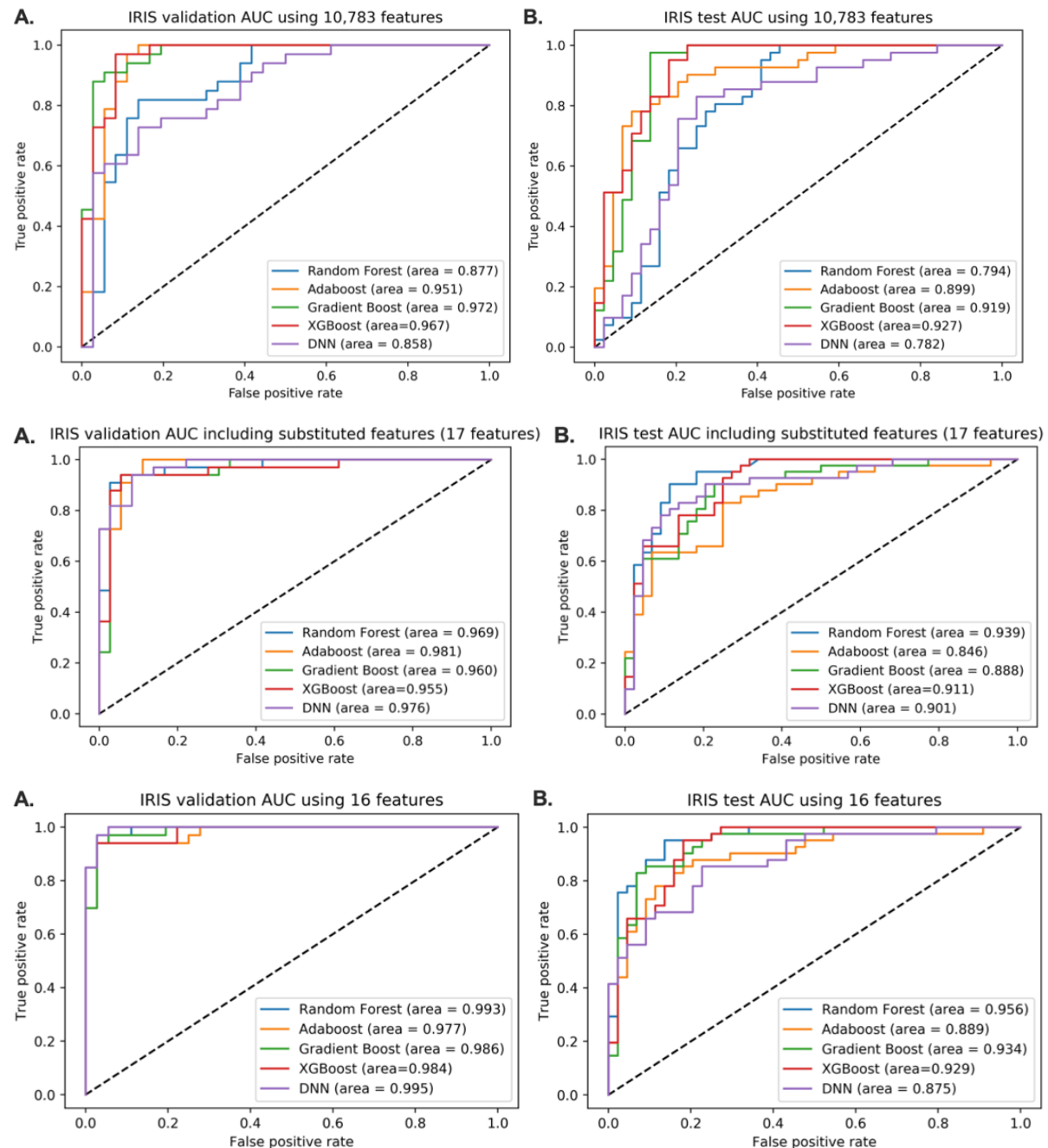
결론

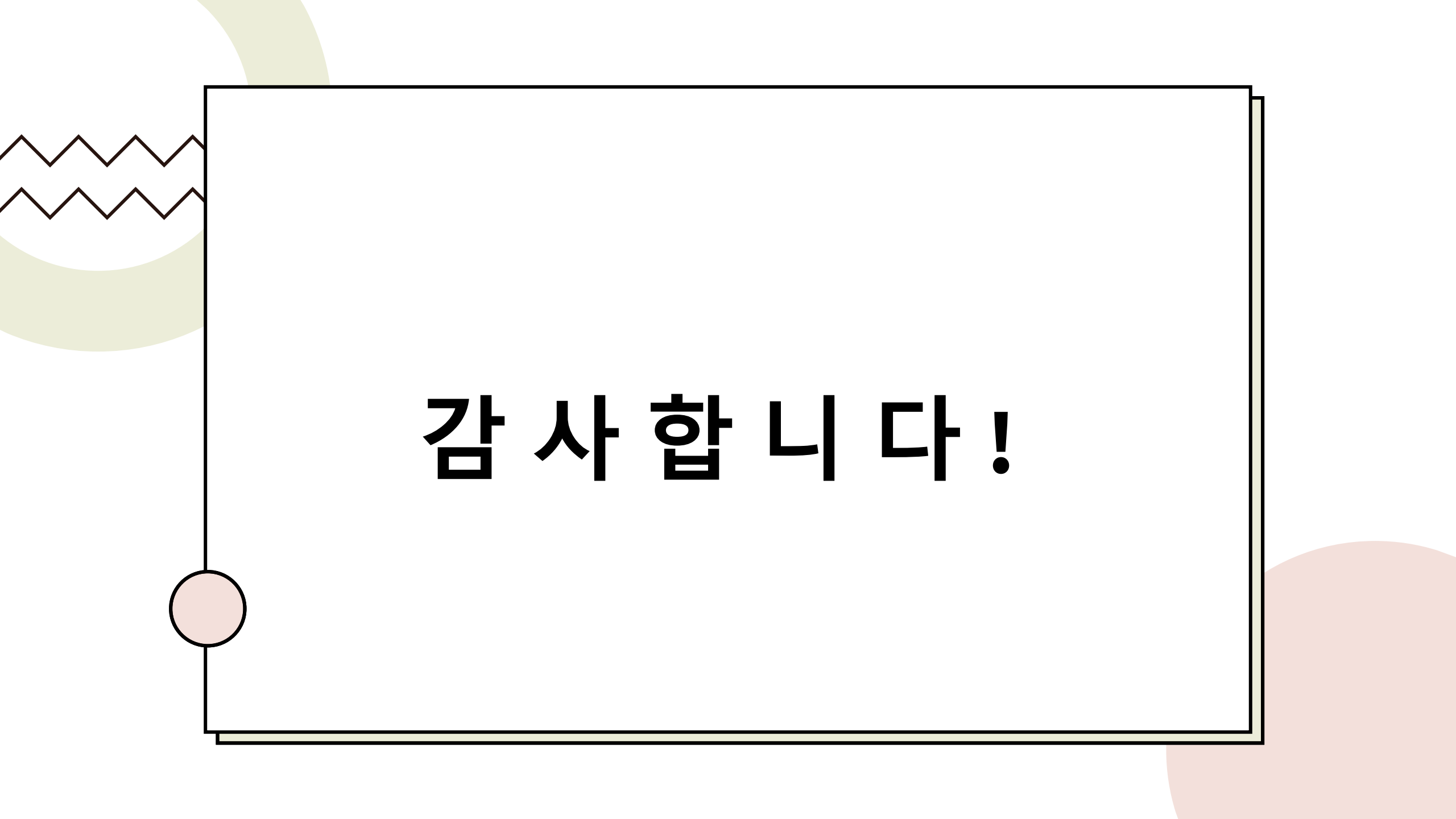
- 10,780 Features **VS** Feature Reduction(17개, 16개)

AUC 결과 신뢰도가 비슷하거나 After Feature Reduction 쪽이 나은 결과를 보여준다.

➔ 바이오 샘플에서 주요 지표 16개 혹은 17개만을 검출하더라도 높은 정확도로 인슐린 저항성을 측정할 수 있다.

➔ 즉, 제 2유형 당뇨병 초기 진단에 용이하게 사용될 수 있다.





감 사 합 니 다 !

코드 출처 논문:

Eunchong Huang, Sarah Kim and Taejin Ahn(2020), Deep Learning for Integrated Analysis of Insulin Resistance with Multi-omics Data, *Department of Life Science, Handong Global University*

데이터 출처:

iHMP Stanford School of Medicine (<http://med.stanford.edu/ipop.html>)

참고 문헌:

Park, G. M., & Bae, Y. C. (2019). Performance comparison of machine learning in the various kind of prediction. *The Journal of the Korea institute of electronic communication sciences*, 14(1), 169-178.
(학습 모델 성능 관련 정보)

김신곤, & 최동섭. (2008). 우리나라 당뇨병의 현황. *J Korean Med Assoc*, 51(9), 791-798.