# Predictive Modeling of Diabetes and Pre-Diabetes Risk Using Healthcare and Lifestyle Data

How well can a predictive modeling approach using features from healthcare and lifestyle data accurately identify individuals at risk of developing type 2 diabetes and pre-diabetes?

Kristina Ivanov & Kaitlyn Terrell

# What are the Risk Factors?

- Overweight or **obese**
- Age **35 or older**, risk increases with age
- **family history** of diabetes.
- Part of **at risk ethnic groups**: African American, Native American, Asian American, Hispanic/Latino, or Pacific Islander
- Not physically active, because of physical limitations or a **sedentary lifestyle**
- Have **prediabetes**
- A history of **gestational diabetes**, or gave birth to a baby weighing 9 pounds or more.[4]

# What is Prediabetes?

When blood glucose levels are **higher than normal** but not high enough to be diagnosed as diabetes.[2]

Occurs in individuals who have **insulin resistance** or beta cells in the pancreas that don’t make enough insulin to keep blood glucose at normal levels.[2]

The following test results show Prediabetes[2]

A1C—5.7 to 6.4 percent

FPG—100 to 125 mg/dL (milligrams per deciliter)

OGTT—140 to 199 mg/dL

# Background

**Why is diagnosing prediabetes important?**

Many people with prediabetes could develop diabetes within 5 years, which puts them at risk of serious health problems.[3]

**What are the long-term effects of type 2 diabetes?**

Heart attack, stroke, blindness, kidney failure, and loss of toes, feet, or legs.[3]

# Motivation

**What is the problem we want to solve?**

The rising global prevalence of type 2 diabetes demands for innovative approaches.

**Why are we wanting to make these predictions?**

To ensure that patients are aware of potential health risks and provide information which allows them to take preventative action against diabetes onset.

# Hypothesis

**Hypothesis:** Predictive modeling using medical and lifestyle data for early detection and risk stratification of type 2 diabetes.

**Objective:** Uncover patterns and key predictors through machine learning algorithms to enhance precision and reliability in diabetes development and pre-diabetes risk assessment.

**Anticipated Insights:** Identification of crucial relationships among features.Emergence of predictive patterns from the dataset.

# Data Sources

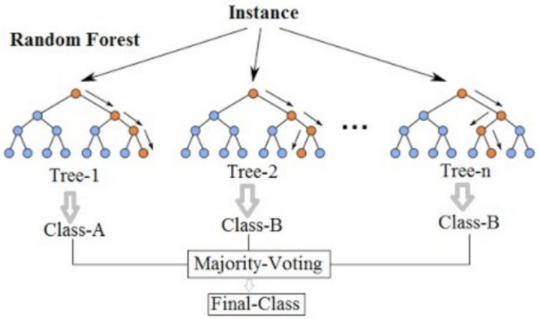
CDC Centers for Disease Control and Prevention
CDC 24/7: Saving Lives, Protecting People™

BRFSS™

- Diabetes
- High Blood Pressure
- High Cholesterol
- Cholesterol Check
- BMI
- Smoker
- Stroke
- Heart Disease or Attack
- Physical Activity
- Fruits
- Veggies
- Heavy Alcohol
- Any Healthcare
- General Health
- Mental Health
- Physical Health
- Difference in Walk
- Sex
- Age
- Education
- Income[6]

# Methods

## Random Forest Classifier:



https://williamkoehrsen.medium.com/random-forest-simple-explanation-377895a60d2d

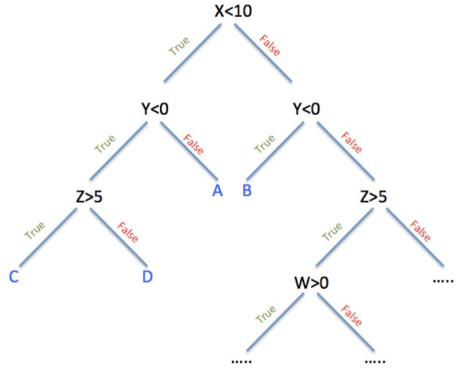
Pros:

- Random Forest can handle missing values well
- can be used for both classification and regression tasks

Cons:

- Challenging to interpret the reasoning behind specific predictions
- May not perform well on highly imbalanced datasets

## Decision Tree Classifier:



https://medium.com/machine-learning-bites/machine-learning-decision-tree-classifier-9eb67cad263e

Pros:

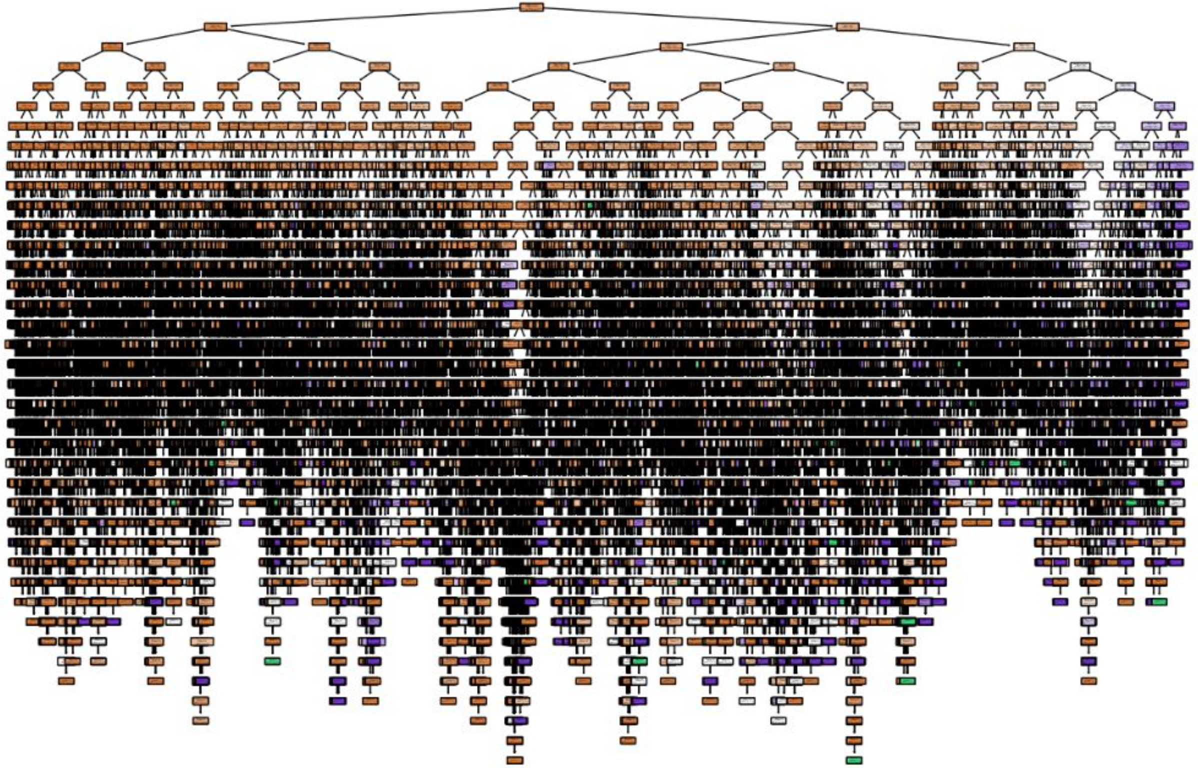
- Can handle both numerical and categorical features
- Do not require extensive data preparation
- Robust to outliers

Cons:

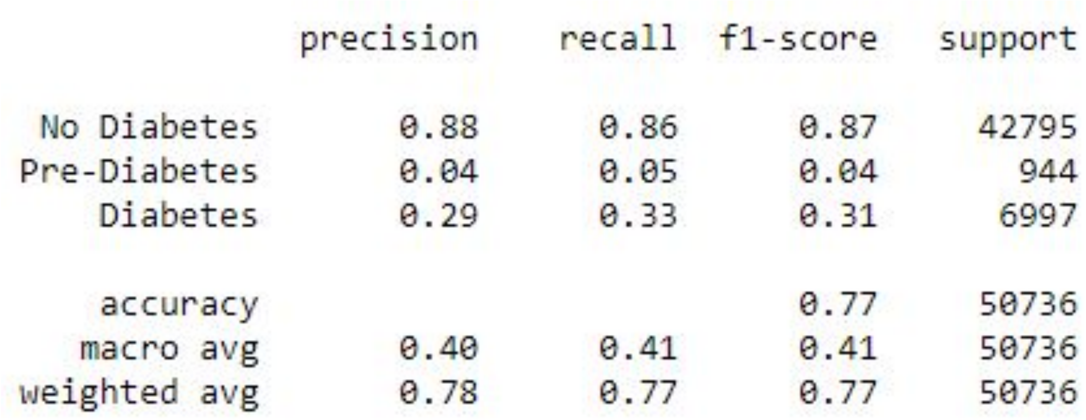
- Decision trees are prone to overfitting, especially when they are deep and complex
- Can have high variance, leading to different splits and structures

# Decision Tree Classifier

# Results: Decision Tree Classification



# Results: Decision Tree Classification

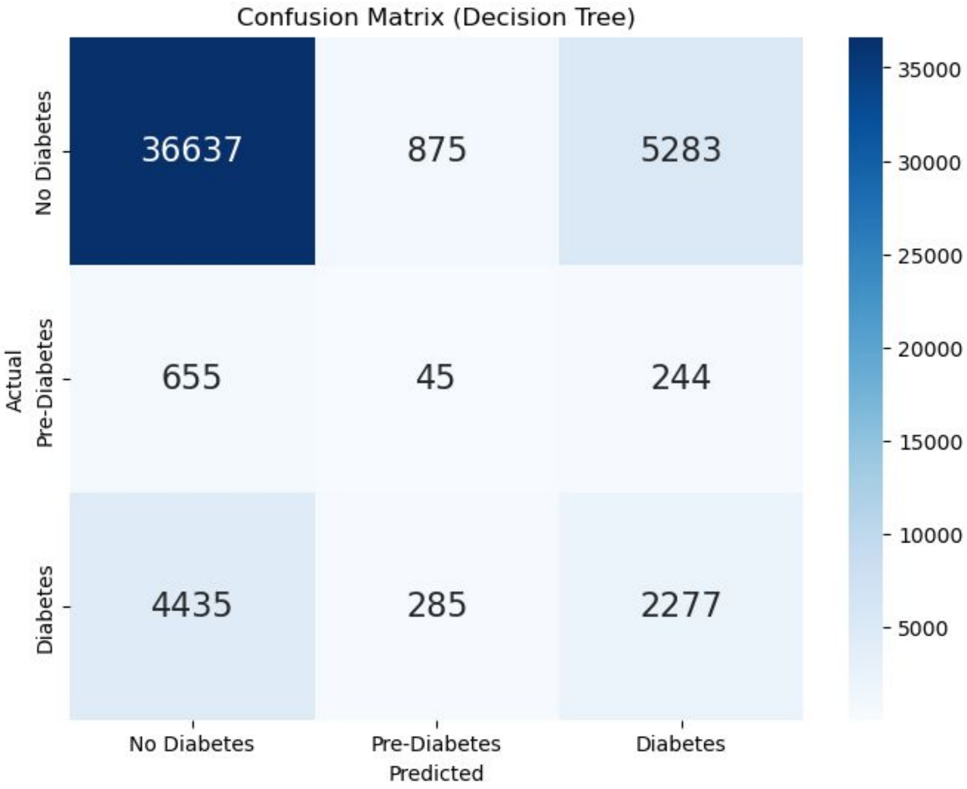


```
              precision    recall  f1-score   support

 No Diabetes       0.88      0.86      0.87     42795
Pre-Diabetes       0.04      0.05      0.04       944
    Diabetes       0.29      0.33      0.31      6997

    accuracy                           0.77     50736
   macro avg       0.40      0.41      0.41     50736
weighted avg       0.78      0.77      0.77     50736
```

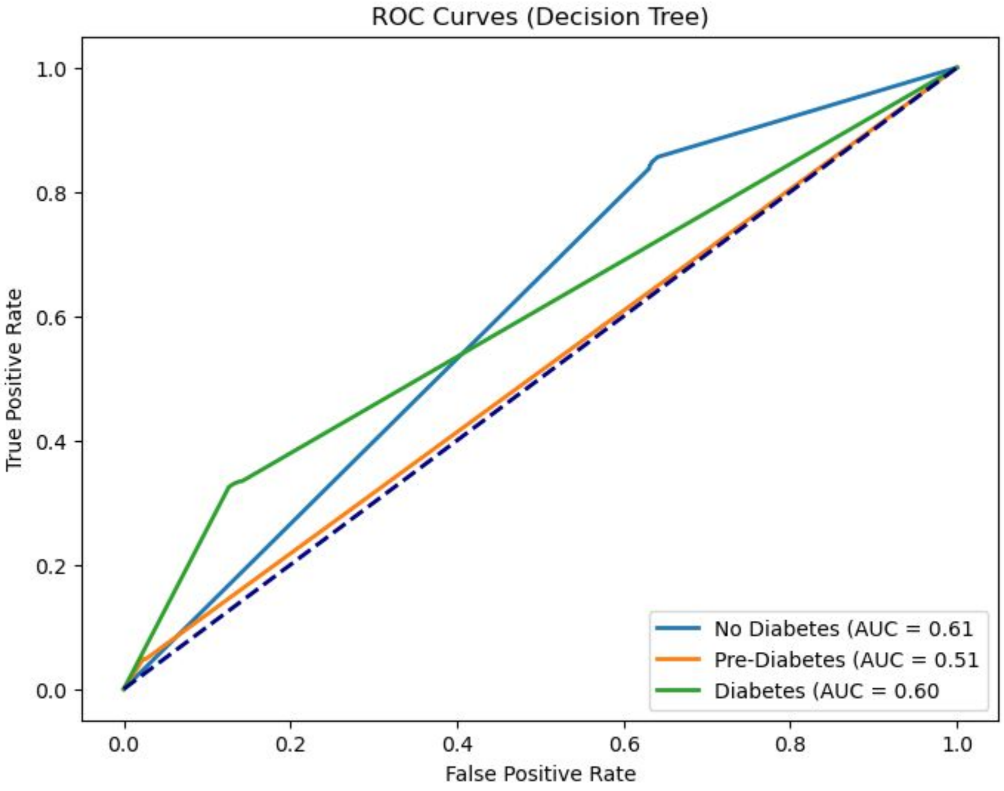
# Results: Decision Tree Classification



Confusion Matrix (Decision Tree)

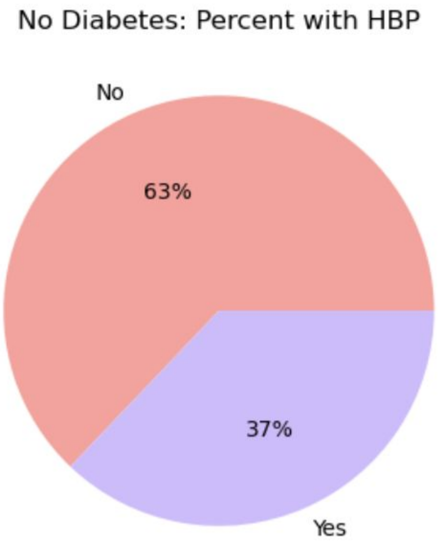
| Actual \ Predicted | No Diabetes | Pre-Diabetes | Diabetes |
|---|---|---|---|
| No Diabetes | 36637 | 875 | 5283 |
| Pre-Diabetes | 655 | 45 | 244 |
| Diabetes | 4435 | 285 | 2277 |

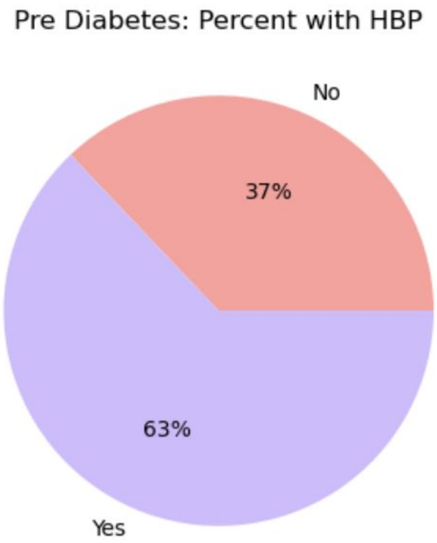
# Results: Decision Tree Classification

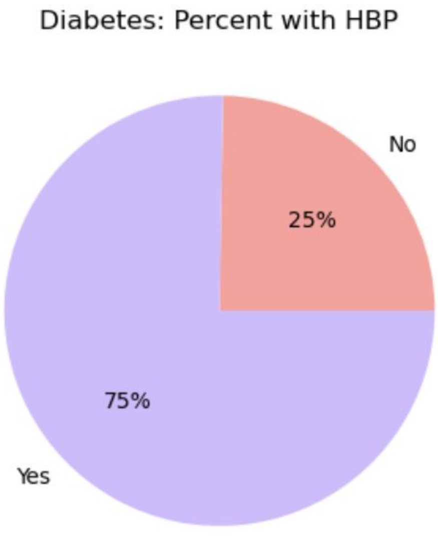


# Random Forest Classifier

# Results: Pre-processing data

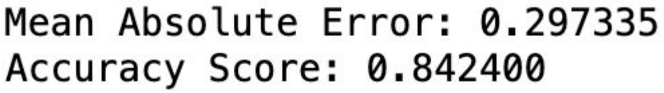

No Diabetes: Percent with HBP


Pre Diabetes: Percent with HBP

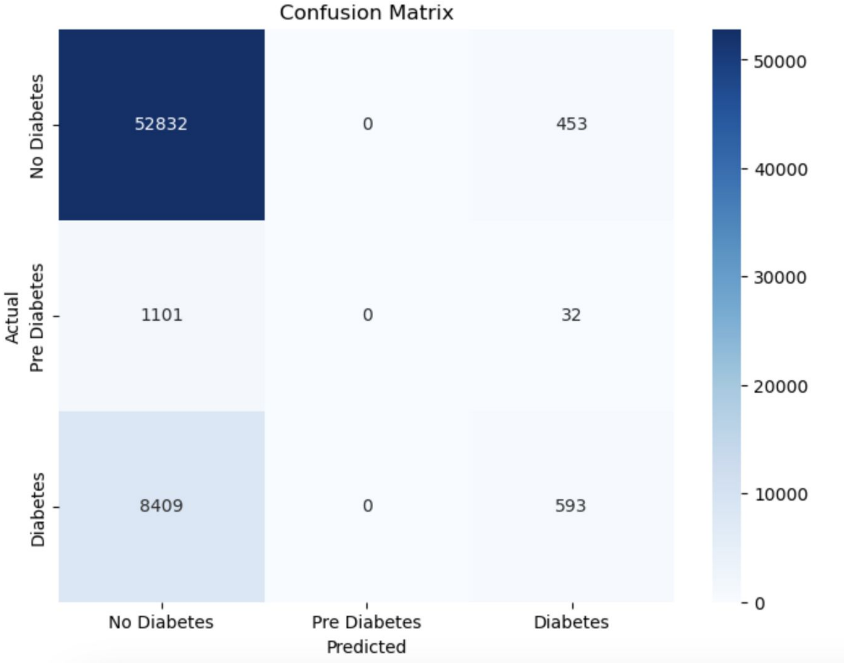

Diabetes: Percent with HBP

HBP: High Blood Pressure

# Results Forest Tree Classified (Discussion)



```
Mean Absolute Error: 0.297335
Accuracy Score: 0.842400
```



# Conclusions / Future Work

## Conclusions:

- Able to pre-processing the data beforehand and understand which characteristics stand out in groups
- Ran both Decision Tree and Random Forest classifiers
- Optimized both, but needs further optimization to improve accuracy

## Future Work:

- Assess alternative classifiers:
  - Alternative model may provide better accuracy
- Iterative improvement:
  - Continuous refinement of the model based on data that is fed in / user response
- Refining the classifier further:
  - Improve accuracy and efficiency

# References

1. Centers for Disease Control and Prevention. (2023). Behavioral Risk Factor Surveillance System Annual Survey Data. National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. Retrieved from https://www.cdc.gov/brfss/annual_data/annual_data.htm
2. Centers for Disease Control and Prevention. (2023). CDC Diabetes Health Indicators. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators. DOI: 10.24432/C53919
3. Centers for Disease Control and Prevention (2023). About Prediabetes and Type 2 Diabetes. National Diabetes Prevention Program. Retrieved from https://www.cdc.gov/diabetes/prevention/about-prediabetes.html#:~:text=Diabetes%20Is%20Serious%20and%20Common&text=Without%20intervention%2C%20many%20people%20with,Blindness
4. National Institute of Diabetes and Digestive and Kidney Diseases (2023). Risk Factors for Type 2 Diabetes. Retreived from https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes
5. National Institute of Diabetes and Digestive and Kidney Diseases (2023). Insulin Resistance & Prediabetes. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/prediabetes-insulin-resistance
6. Kaggle (2020). CDC Diabetes Health Indicators. Retrieved from https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/