# Predicting Diabetes from Health Factors

●●●

# Problem Overview

- 37 million people in the U.S. have diabetes
- 1 in 5 do not know they have diabetes
- Early detection of diabetes can help patients better manage their health and prevent the health complications from diabetes

Our goal is to create a model the can be used by doctors to predict diabetes.
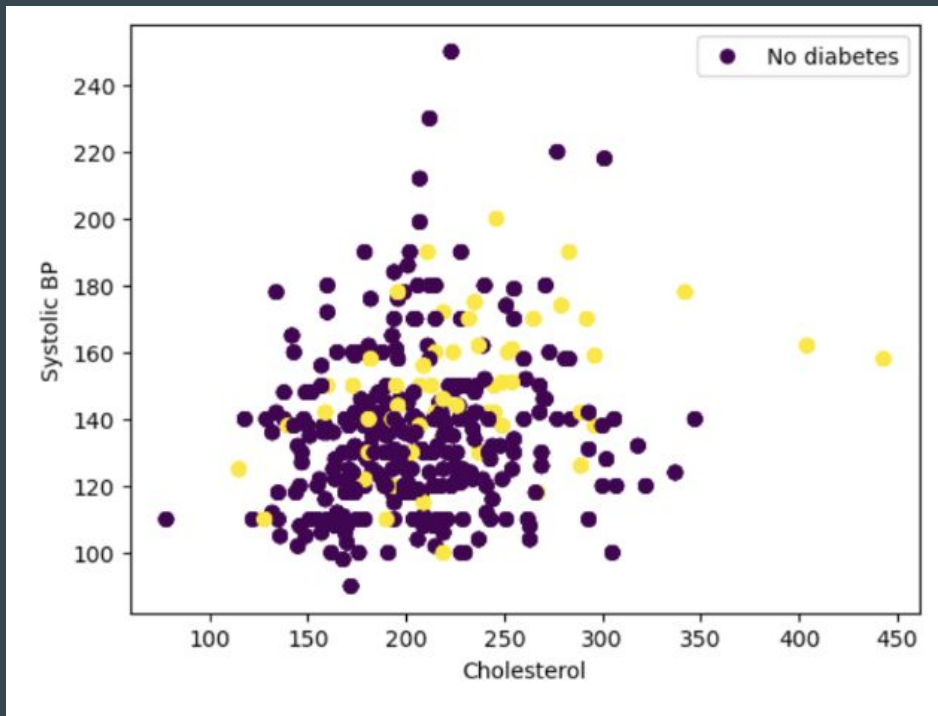
# The Data

- Collected from rural African American patients
- Health factors:
  - Cholesterol
  - Glucose
  - HDL cholesterol
  - Chol/HDL ratio
  - Age
  - Gender
  - Height
  - Weight
  - BMI
  - Systolic BP
  - Diastolic BP
  - Waist
  - Hip
  - Waist/hip ratio
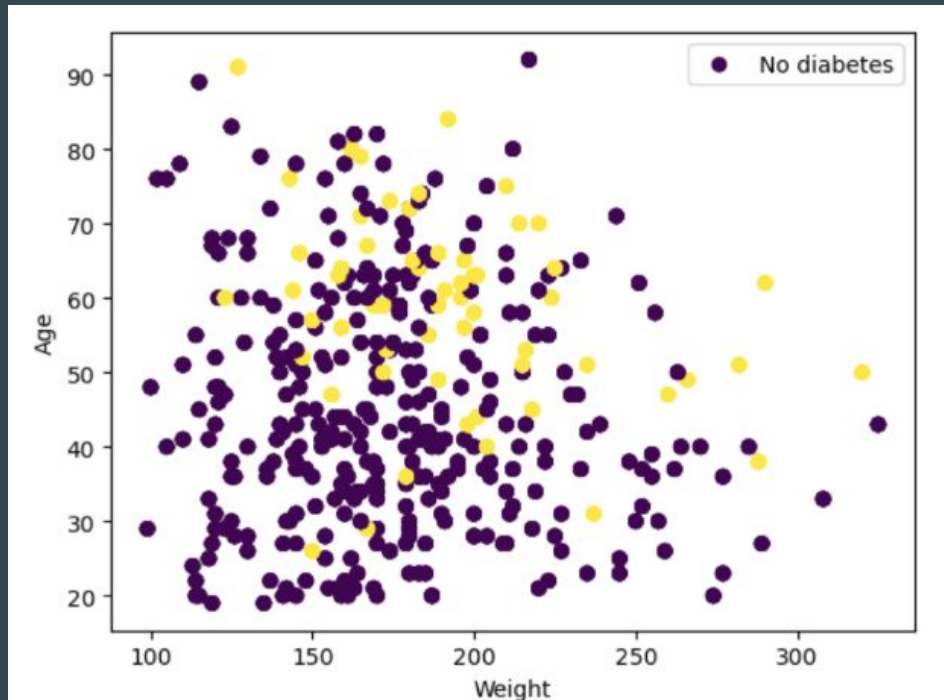  - Diabetes

# Correlation Heatmap
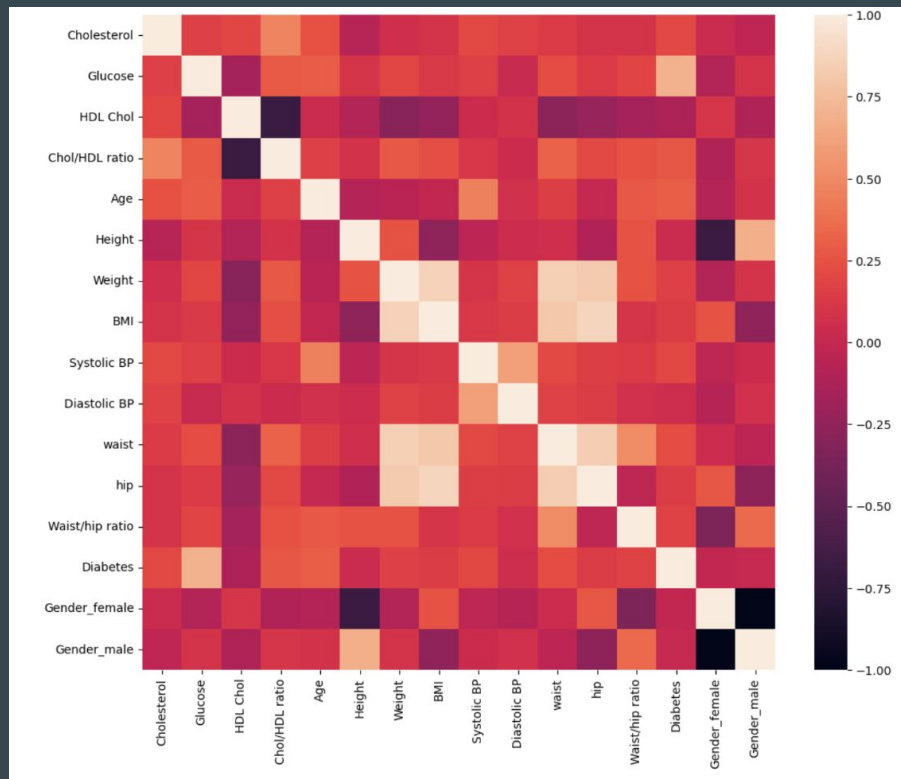
# Cholesterol vs. Systolic BP



- This table compares cholesterol levels and systolic blood pressure in patients with diabetes presence indicated by the color yellow.

# Weight vs. Age



- This table compares weight and age in patients with diabetes presence indicated by the color yellow.

# Model Testing

| Model 1: | ● Decision Tree |
| Model 2: | ● Logistic Regression |
| Model 3: | ● Gradient Boosting |

# Model Recommendation

## Logistic Regression Model

Parameters: solver = newton-cg, C = 10

### Model Performance:

| Accuracy | 0.99 |
| --- | --- |
| Precision | 1.0 |
| Recall | 0.94 |
| F1 Score | 0.97 |

# Model Performance

- To the right you can see a confusion matrix for the model.
- The model correctly predicted 84 cases of diabetes in the test dataset.
- The model correctly predicted 496 cases of non-diabetics in the test dataset.
- The model incorrectly predicted 5 false negatives and zero false positives.
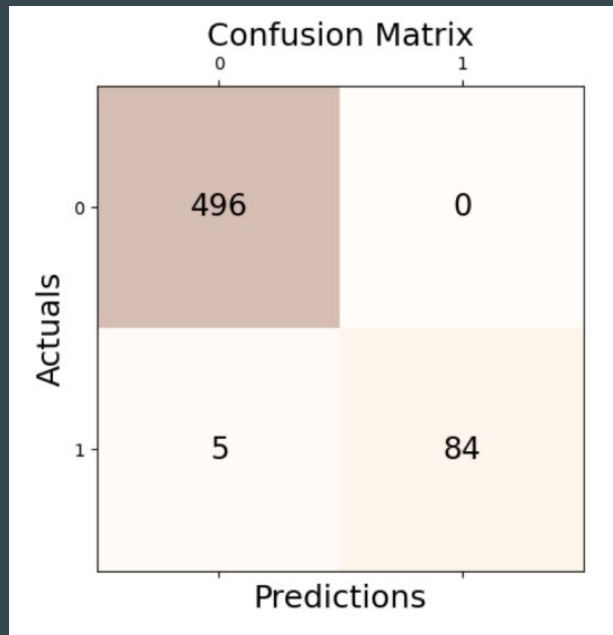
# Conclusion

- The logistic regression model was the best model.
- Our hope is that this model can be used by doctors to predict patients that should seek testing for diabetes.
- This model can help diagnose patients that are unaware they have diabetes.
- This model can help identify risk factors of diabetes.

# Future Considerations:

- Gather data from more patients
- Diversify the population sampled
- Include more health factors
- Expand utilization to diagnosing at risk populations for the prevention of diabetes