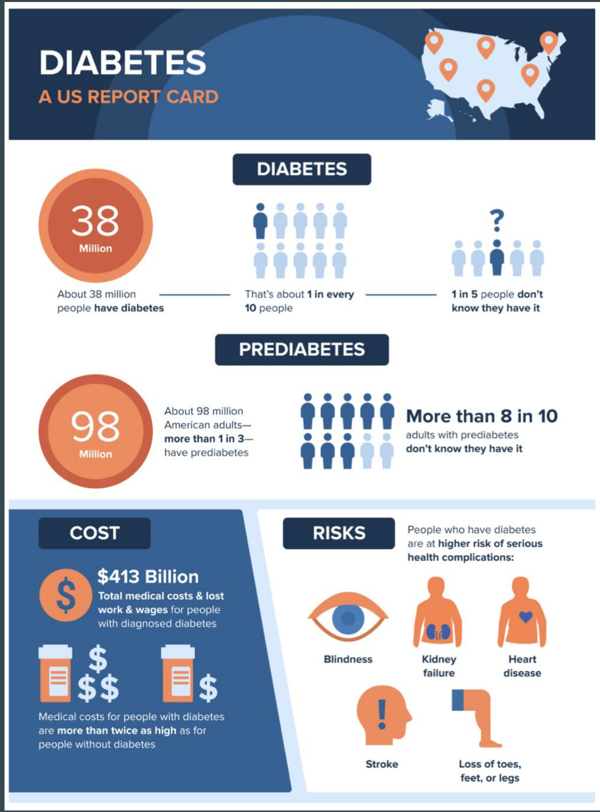
# Diabetic Retinopathy Screening

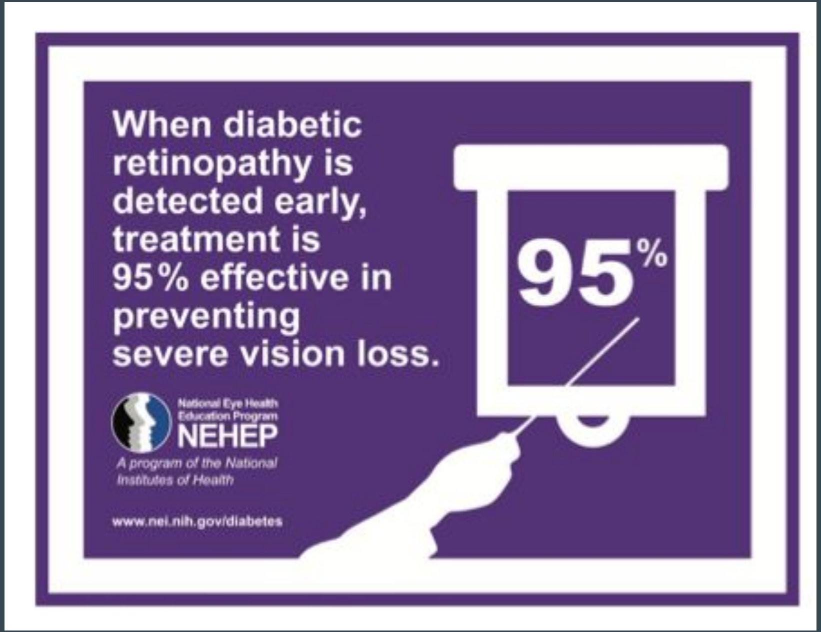
## BIS568 FINAL PROJECT

Jiaye Chen, Luning Yang, Mengmeng Du, Yawen Wei

Dec 14, 2023

# Problem Specification - Background



# Problem Specification - Goal & Population

- **Goal of the Predictive Model**
  - Predict diabetic retinopathy(DR) in patients with diabetes

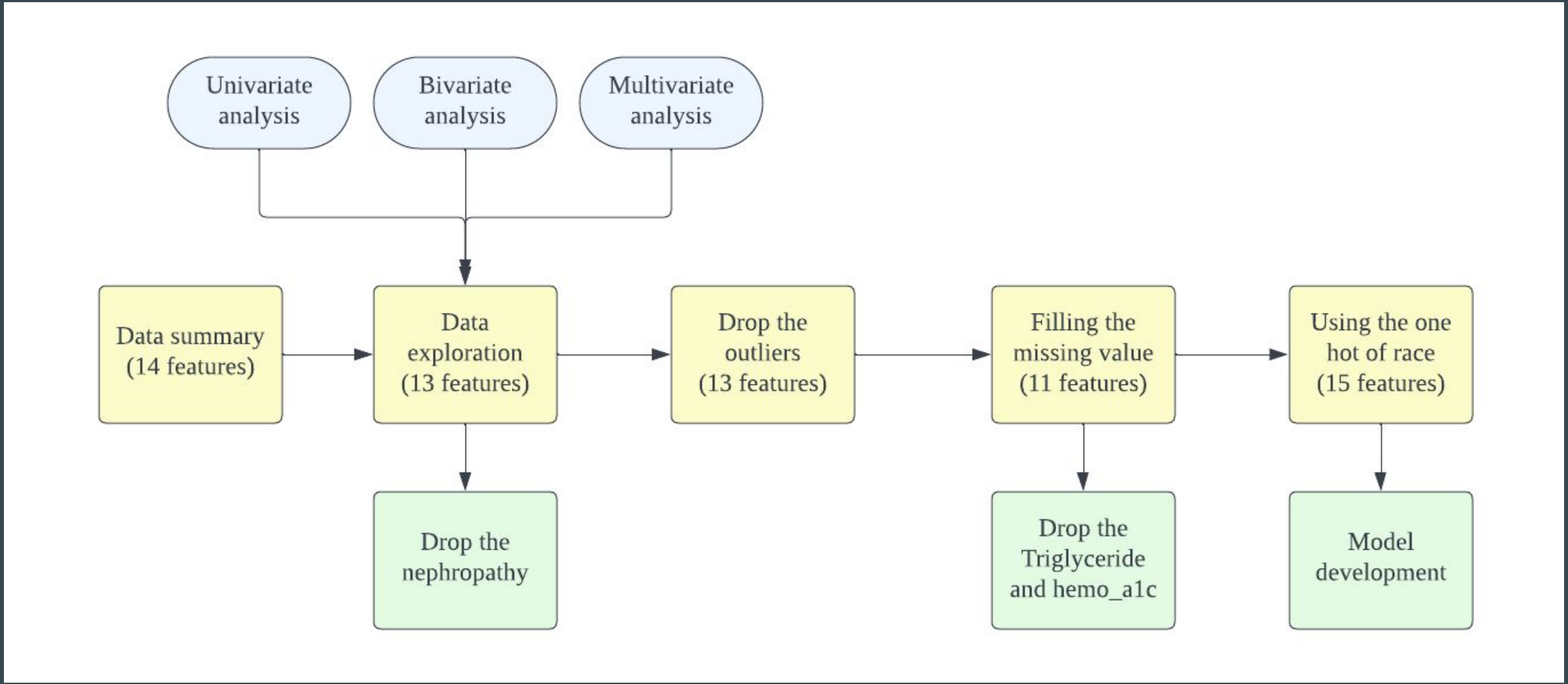
- **Patient Population**
  - Diabetic patients
    - with more than one record in MIMIC database

- **Target Cohort**
  - Diabetic patients with diabetic retinopathy(DR): 133
- **Control Cohort**
  - Diabetic patients without any complications: 1160

# Data Preparation

| | | |
|---|---|---|
| Demography | gender | Patient gender |
| | race | Patient race |
| | age | Patient age |
| Conditions | duration days | Diabetes/Diabetic retinopathy duration |
| | neuropathy | Neuropathy |
| | stroke | Stroke |
| | Insulin dependence | Insulin dependent diabetes mellitus |
| | nephropathy | Nephropathy |
| Clinical measurement | triglyceride | Triglycerides |
| | hemo | Hemoglobin |
| | hemo_a1c | Hemoglobin A1C |
| | bun | Blood urea nitrogen |
| | sbp | Systolic blood pressure |
| | dpb | Diastolic blood pressure |
| Predictor | predictor | Diabetes/Diabetic retinopathy |

# Data Preparation and Preprocessing



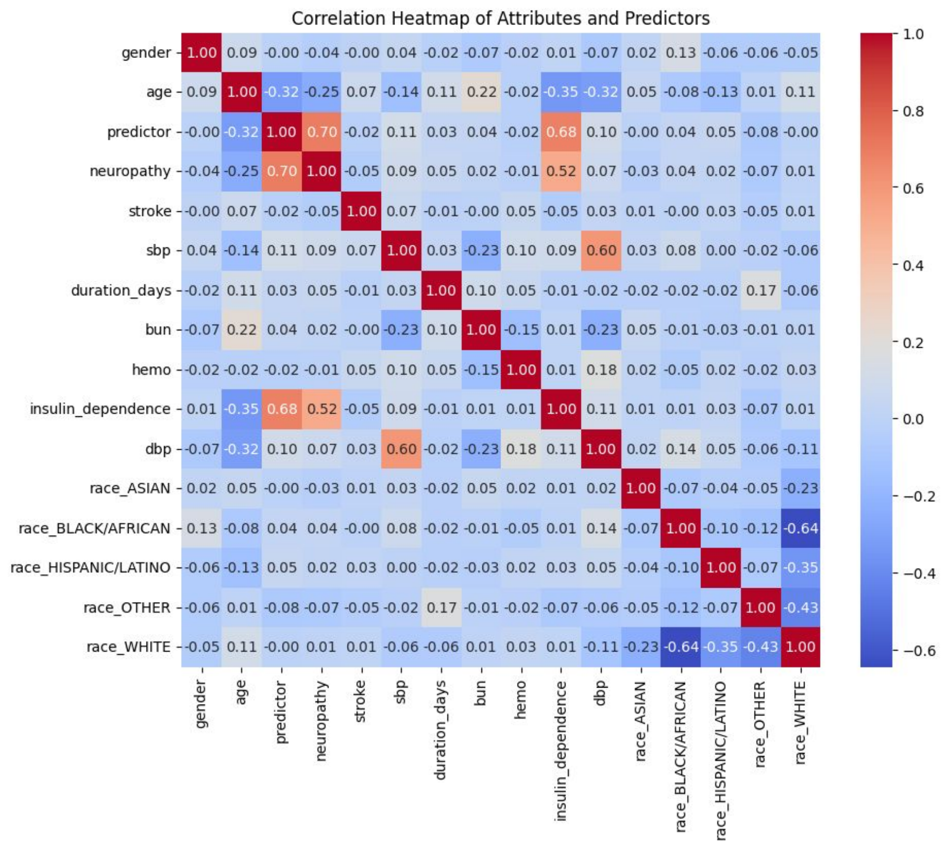
# Bias Assessment

- The data only reflects part of the patient in this area.
  - Our Prevalence based on MIMC dataset = 1.47%
  - 2021 Suffolk County Adjusted Prevalence of Any Diabetic Retinopathy (DR) or Vision Threatening DR from CDC = 4.04%
- Oversampling
  - Original cohort: 133 DR vs 1160 Non-DR
  - New cohort: 852 DR vs 1160 Non-DR

```
from imblearn.over_sampling import SMOTE
# Apply SMOTE to oversample the minority class (positive samples)
smote = SMOTE(sampling_strategy='auto')  # 'auto' adjusts to balance classes
X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)
```

- After implementing our model, we will continuously monitor of our ML model's performance.

# Correlation Matrix




- Predictor is strongly correlated with "neuropathy" (0.70) and "insulin_dependence" (0.68), suggesting that it is closely related to whether or not a person is having a insulin-dependent diabetes.
- Neuropathy is NOT a results from DR.

# Model Development

**Cleaned Data**

- 15 attributes
- 1 predictor
- 1293 patients

**Train-Test Splitting**

- 70 % Training
- 30% Testing
- Stratified Sampling

Resampling By SMOTE
Balance in neg- and pos+

**Inputted Data**

- From **1160** neg- and **133** pos+
- To **1160** neg- and **852** pos+

**Model Training**

- Logistic Regression
- Random Forest
- XGBoost

Parameter Tuning
by Grid Search

**Model Evaluation**

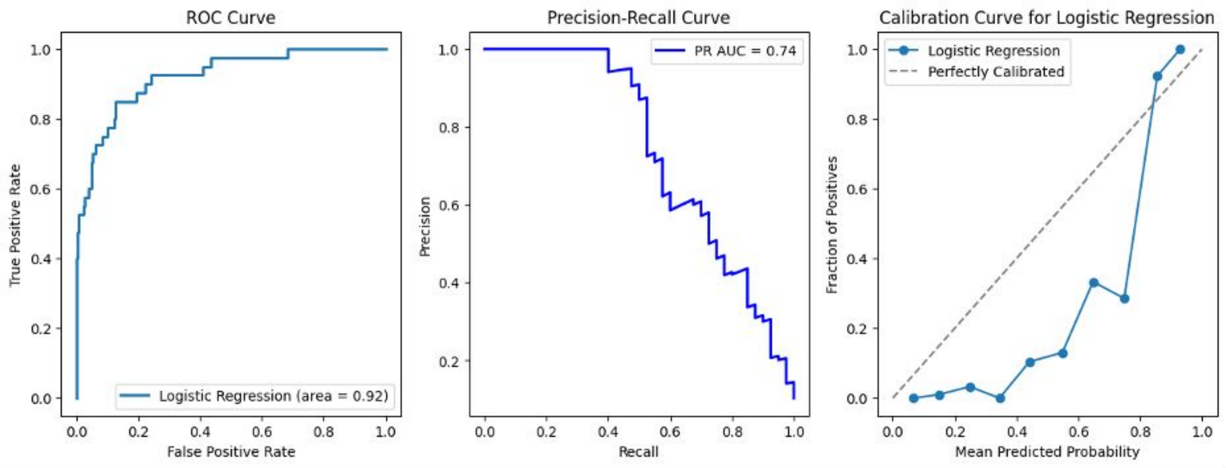
- ROC-AUC
- PR-AUC
- Confusion Matrix

**Interpretability Analysis**

- SHAP plots

# Model Performance and Evaluation

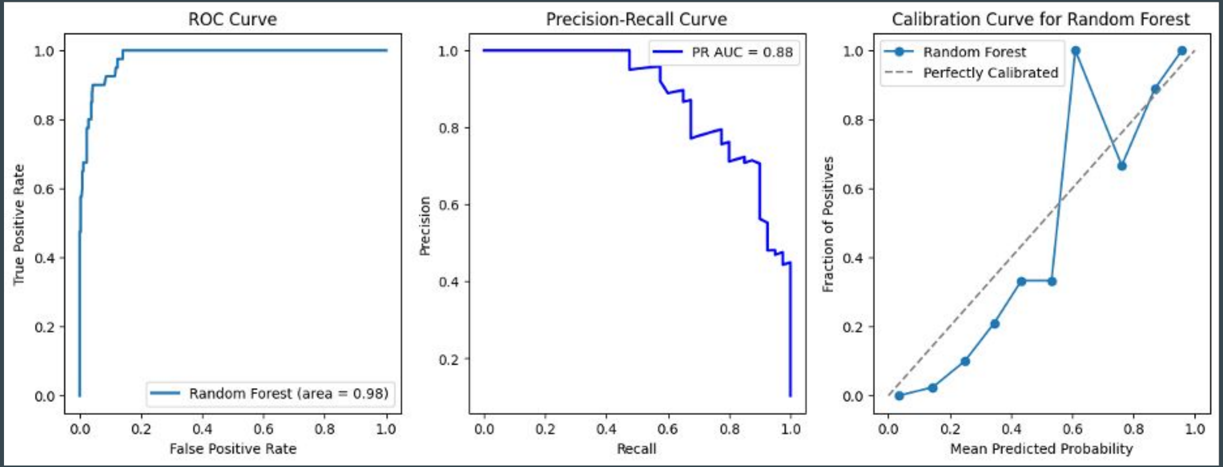
- **Logistic Regression**
  - ROC-AUC: 0.92
  - PR-AUC: 0.74
- Random Forest
  - ROC-AUC: 0.98
  - PR-AUC: 0.88
- XGBoost
  - ROC-AUC: 0.98
  - PR-AUC: 0.87




- Training and Cross-Validation
  - Best Model: **Random Forest**
  - Best Cross-Validation Accuracy: 0.961
  - Test Accuracy: 0.954

# Model Performance and Evaluation

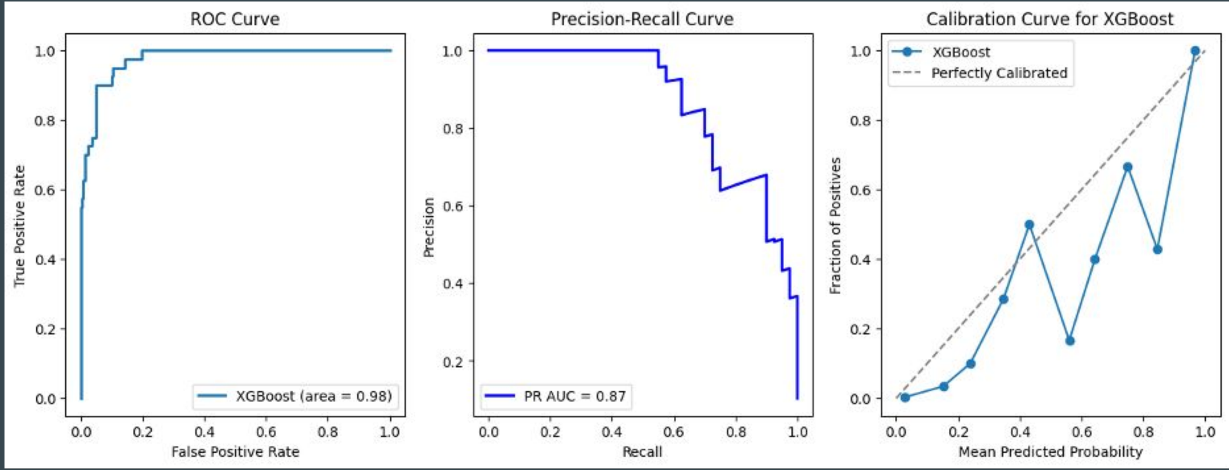
- Logistic Regression
  - ROC-AUC: 0.92
  - PR-AUC: 0.74
- Random Forest
  - ROC-AUC: 0.98
  - PR-AUC: 0.88
- XGBoost
  - ROC-AUC: 0.98
  - PR-AUC: 0.87




- Training and Cross-Validation
  - Best Model: **Random Forest**
  - Best Cross-Validation Accuracy: 0.961
  - Test Accuracy: 0.954

# Model Performance and Evaluation

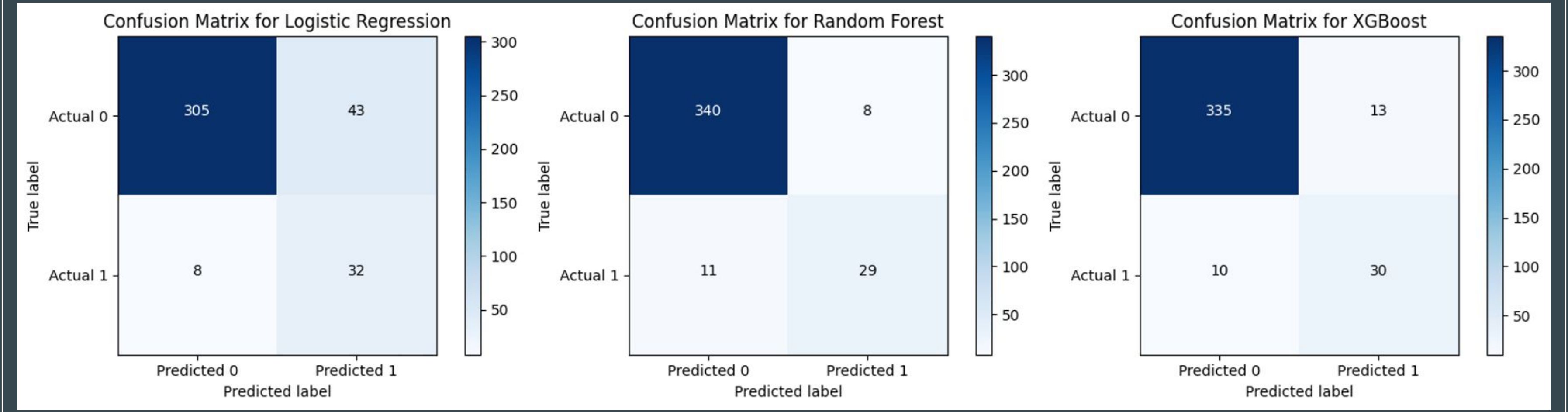
- Logistic Regression
  - ROC-AUC: 0.92
  - PR-AUC: 0.74
- Random Forest
  - ROC-AUC: 0.98
  - PR-AUC: 0.88
- XGBoost
  - ROC-AUC: 0.98
  - PR-AUC: 0.87




- Training and Cross-Validation
  - Best Model: **Random Forest**
  - Best Cross-Validation Accuracy: 0.961
  - Test Accuracy: 0.954

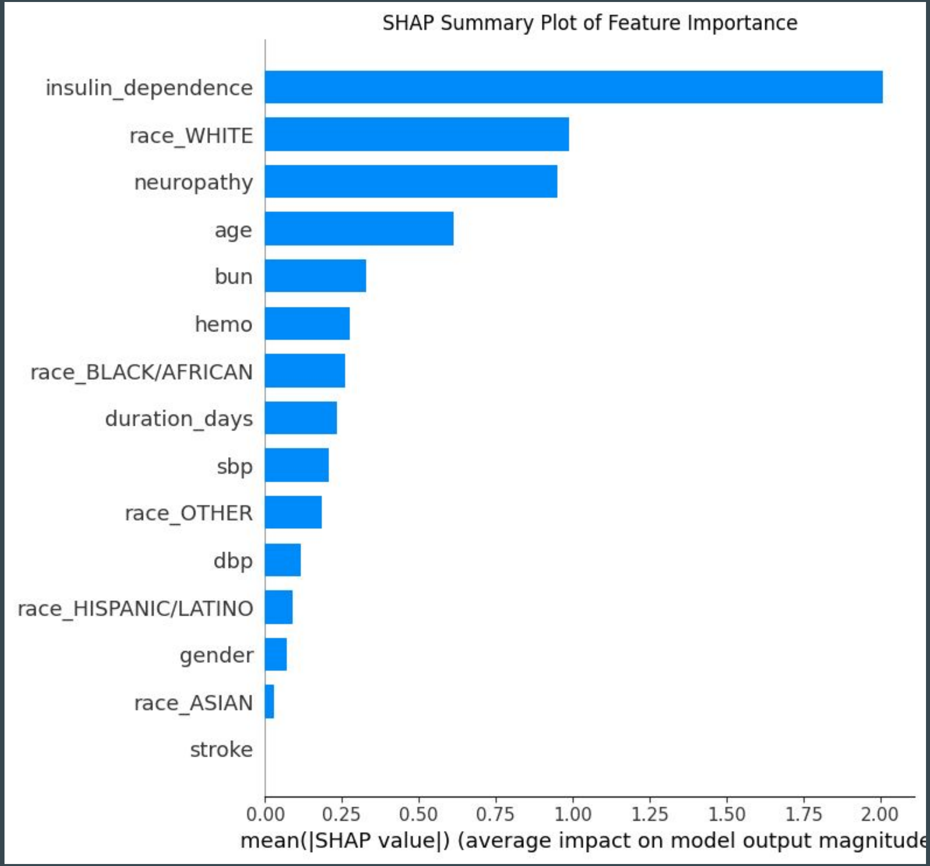
# Model Performance and Evaluation

- **Logistic Regression** outperforms in tolerating False Positive (FP)
- **XGBoost** outperforms Random Forest in minimizing False Negative (FN)



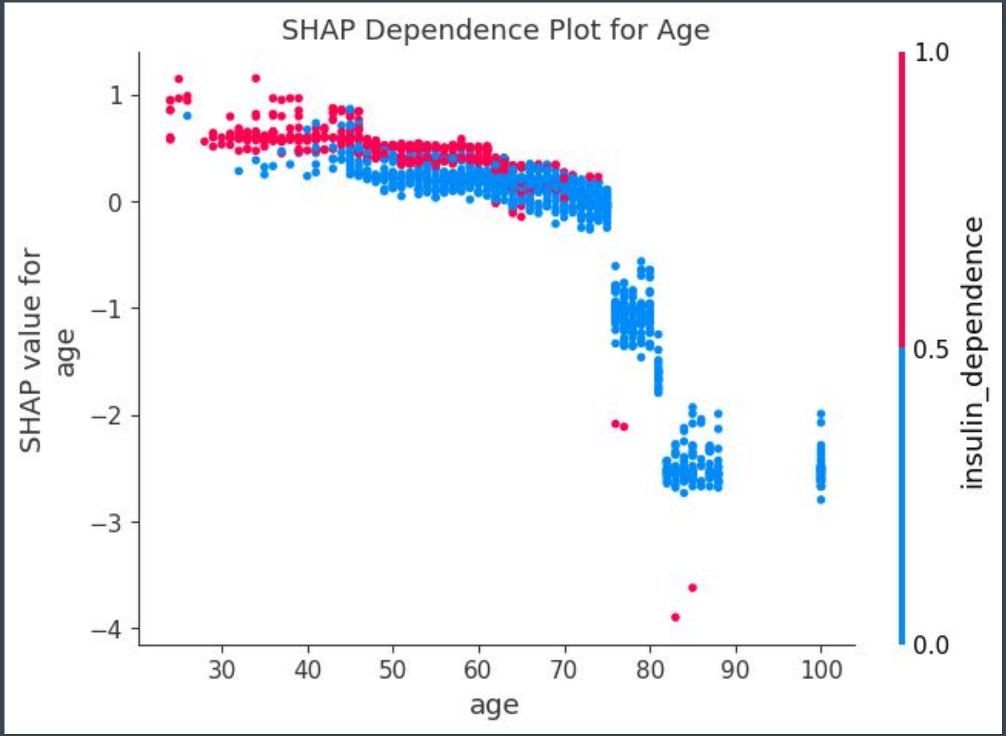
# Interpretability

- Insulin_dependenc show the most predictive power, followed by race_WHITE, and so on



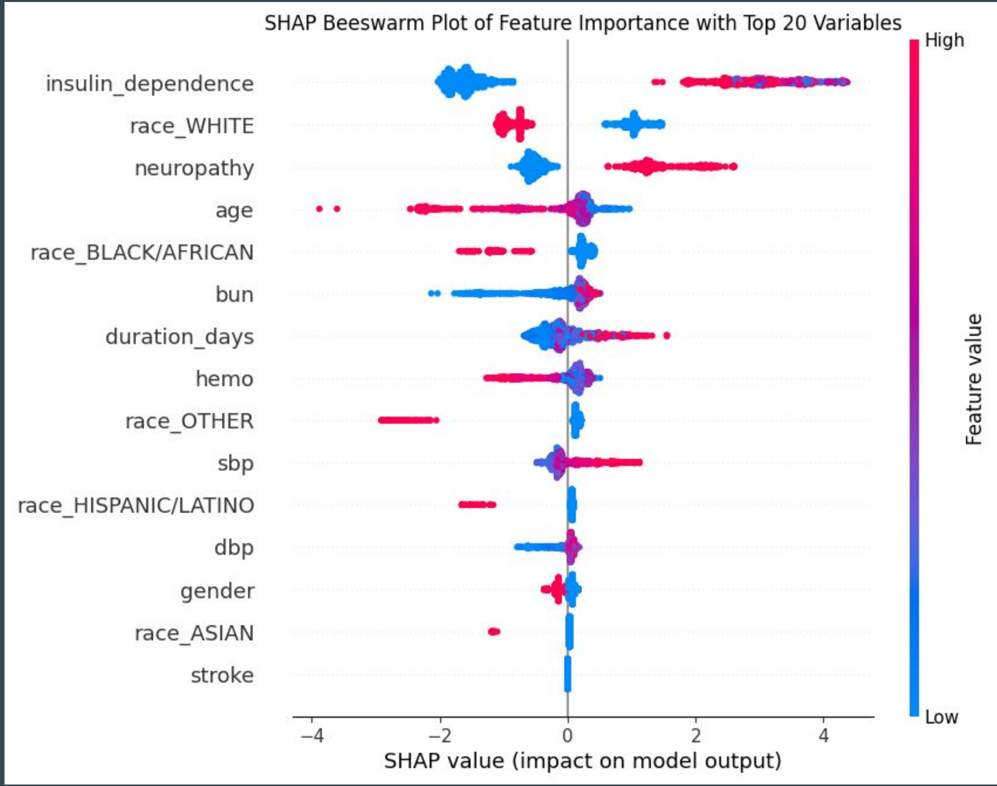
# Interpretability

- It is observed a downward trend in the SHAP dependence plot for the feature 'age,' indicating that smaller values of age contribute to higher SHAP values



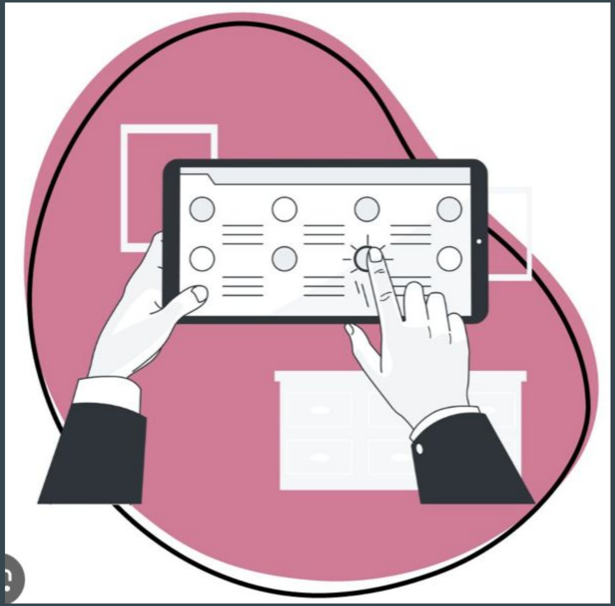
# Interpretability

- insulin_dependence not only has a significant impact on predictions (as indicated by the spread of dots along the x-axis) but also that higher values of insulin_dependence (red dots) are generally associated with an increase in the model's prediction value.



# Looking Forward

- Implementation Plan & Dissemination Strategy
    - User Interface
        - User-friendly Interface
    - Integration with Healthcare Systems
        - Electronic Health Records
        - Clinical Decision Support
        - Yale Health System



# References

Ogunyemi, Omolola I, et al. “Detecting diabetic retinopathy through machine learning on electronic health record data from an urban, safety net healthcare system.” JAMIA Open, vol. 4, no. 3, 2021, https://doi.org/10.1093/jamiaopen/ooab066.

“Diabetes Quick Facts.” Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 4 Apr. 2023, www.cdc.gov/diabetes/basics/quick-facts.html#:~:text=About%2038%20million%20people%20in,(and%20may%20be%20underreported).

“Diabetic Retinopathy Rule of Thirds Guides - Centers for Disease ...” Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, www.cdc.gov/visionhealth/pdf/factsheet.pdf. Accessed 14 Dec. 2023.

# Acknowledgement


- Jiaye Chen: Data preparation, Problem specification
- Luning Yang: Data preparation, Model development and performance evaluation
- Mengmeng Du: Data preparation, Data preprocessing
- Yawen Wei: Data preparation, Bias Assessment and Interpretability