Stat 451 Group 8

Identifying Key Health Indicators and Sociodemographic Factors Associated with Diabetes Risk

Alice Kang, Qidi An, Manwen Jia,
Leon Li, Jiahao Chen

Background

What is Diabetes?

Why Study Diabetes?

- 1 in 10 adults in the U.S. has diabetes
- Rising global prevalence → 422M+ cases worldwide
- Early prediction = better prevention & management

Background

Dataset Snapshot

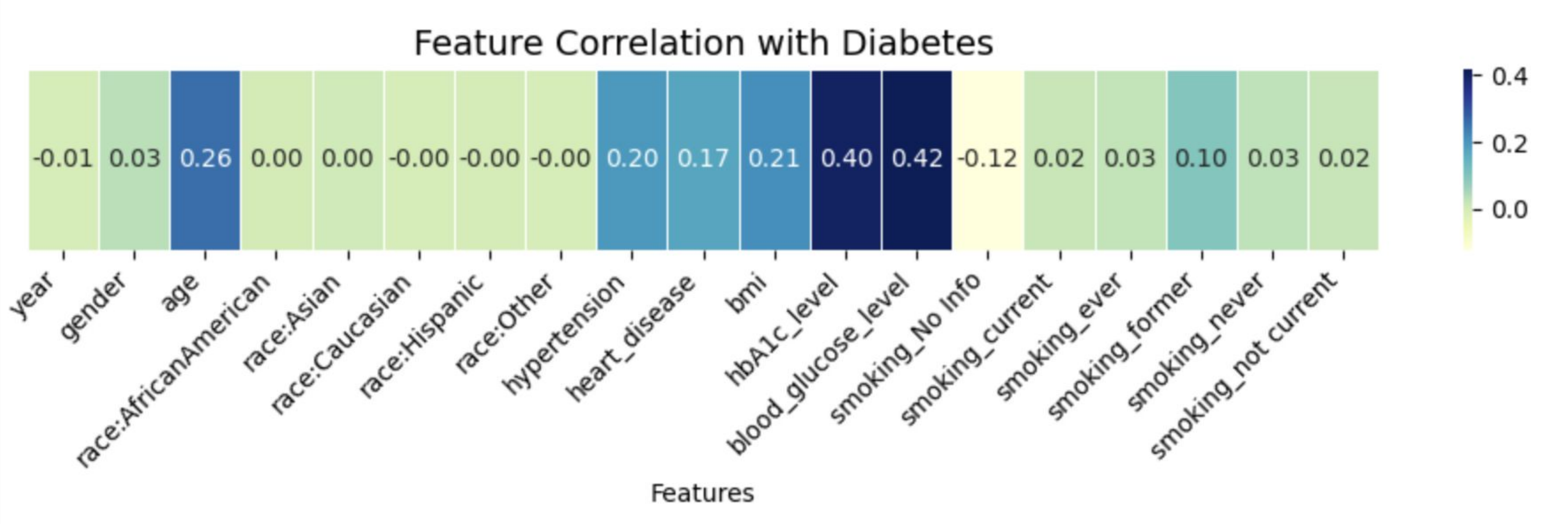
- 100,000 individuals
- Clinical + demographic features
- Source: *Diabetes Health & Demographics Dataset by ZIYA*

Background

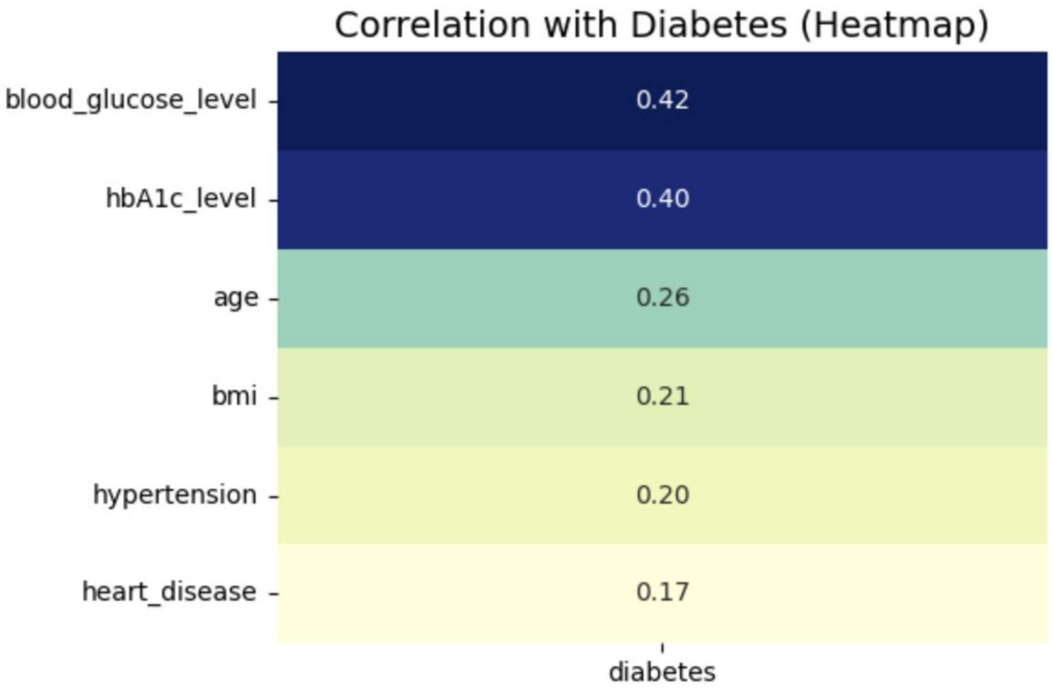
Our Project Goal:

- Predict diabetes status using health indicators
- Identify key risk factors from medical and demographic data

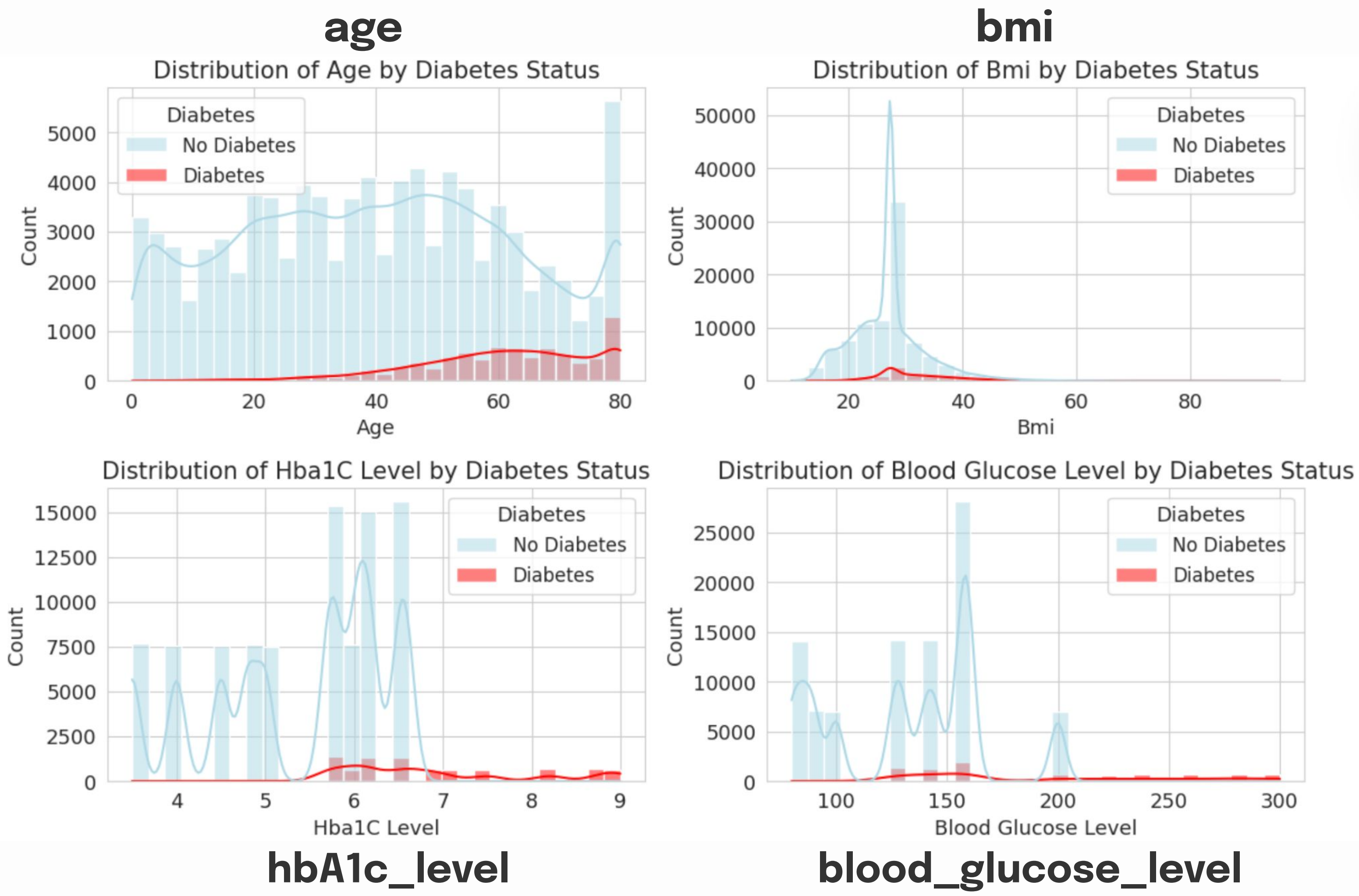
Heatmap of correlation coefficients of each characteristic with diabetes

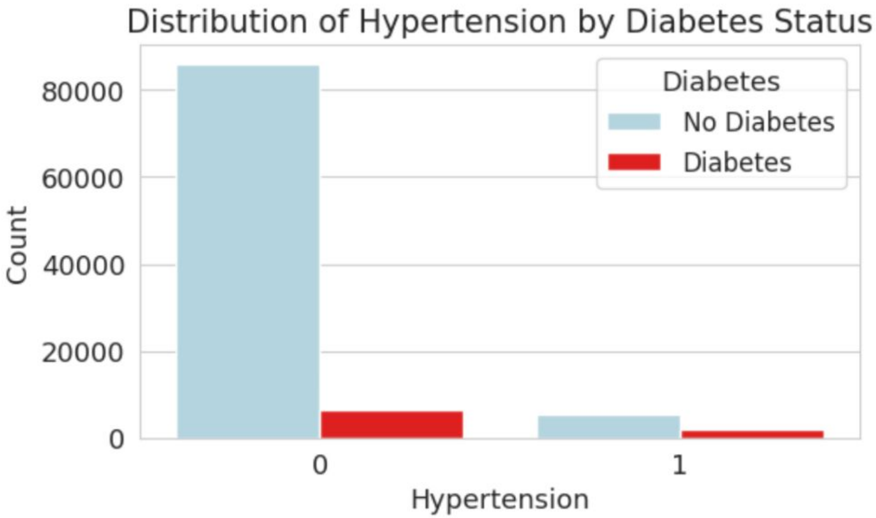


Heatmap of Selected Feature characteristic

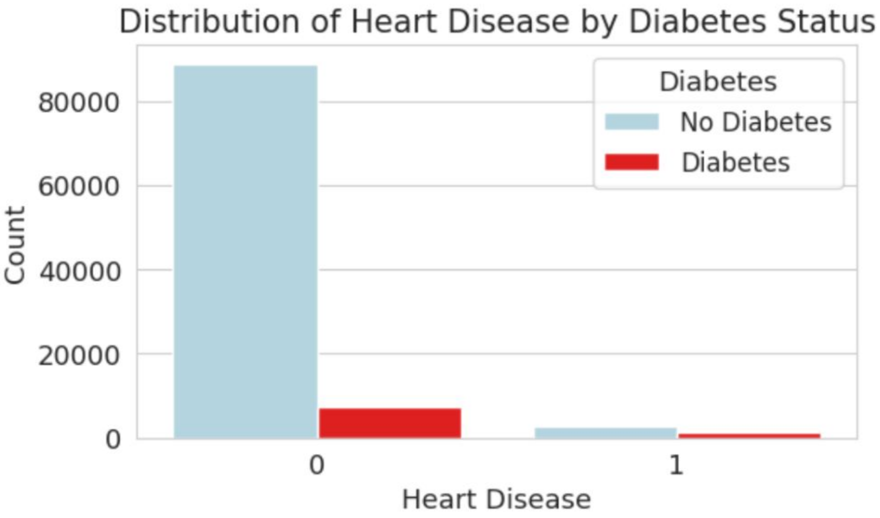


- ➔ Blood Glucose Level
- ➔ HbA1c Level
- ➔ Age
- ➔ Bmi
- ➔ Hypertension
- ➔ Heart Disease


age
bmi
Distribution of Age by Diabetes Status
Diabetes
No Diabetes
Diabetes
Count
Age
Distribution of Bmi by Diabetes Status
Bmi
Distribution of Hba1C Level by Diabetes Status
Hba1C Level
Distribution of Blood Glucose Level by Diabetes Status
Blood Glucose Level
hbA1c_level
blood_glucose_level



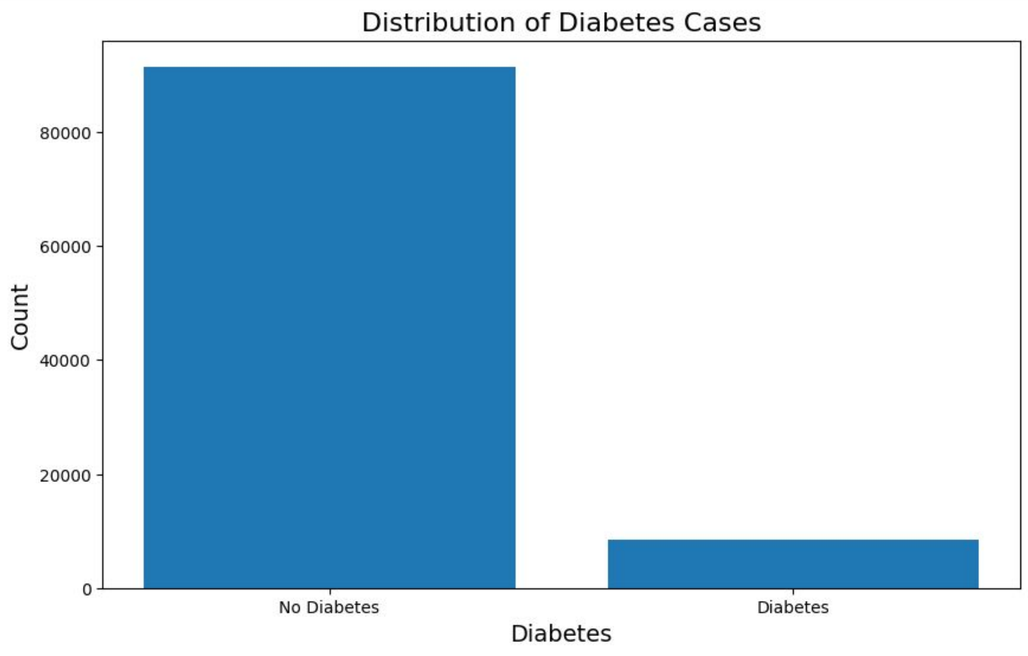
hypertension



heart_disease

Methods

- ➢ Rescaling : use standardscaler()
- ➢ Baseline Model Consideration:
  - ○ 91.5% of individuals do not have diabetes, and only 8.5% do.
  - ○ Predicting Probabilities of "no diabetes" gets 91.5% accuracy.



Methods

Stratified Sampling for Data Splitting

- ➢ Training Set : 80%
- ➢ Validation Set : 10%
- ➢ Test Set: 10%
- ➢ Stratify by diabetes proportion

```
Train set diabetes proportion: 0.0850
Validation set diabetes proportion: 0.0850
Test set diabetes proportion: 0.0850
```

Methods

Model Selection & Tuning

- Tested models: **Logistic Regression**, **Decision Tree**, **SVM**, **KNN**

- Each model was tuned twice:

  - Once with **accuracy** as the scoring metric
    i. measure the overall correctness of predictions

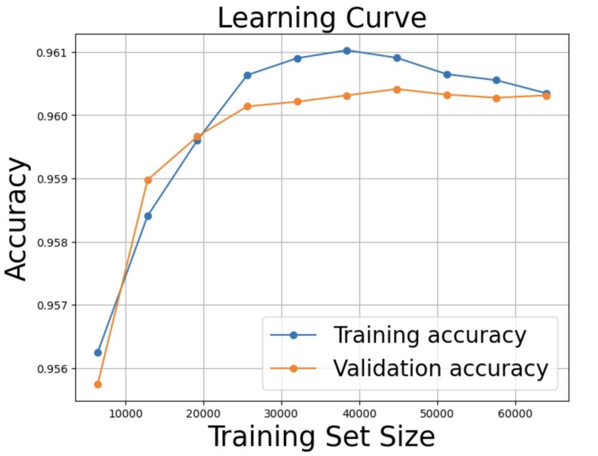
  - Once with **recall**, which is crucial for disease detection
    i. focuses specifically on finding all positive instances

Methods

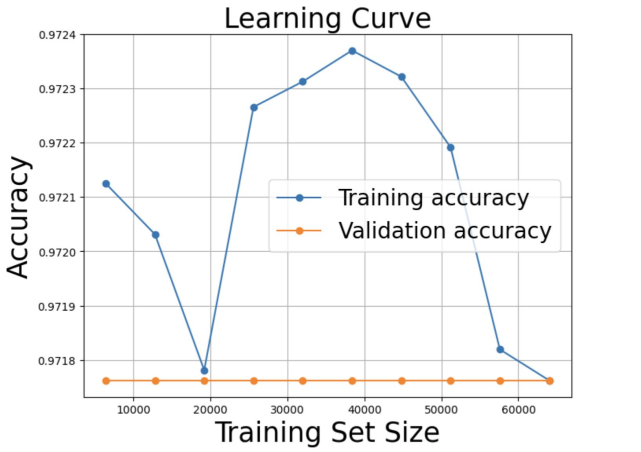
Model	**Accuracy score (optimizing for accuracy)**	**Accuracy score (optimizing for recall)**
Logistic	0.9619	0.9623**(Finalized Model)**
Decision Tree	0.9731	0.9570
KNN	0.9677	0.9698
SVM	0.9685	0.9685

Overfitting Curve

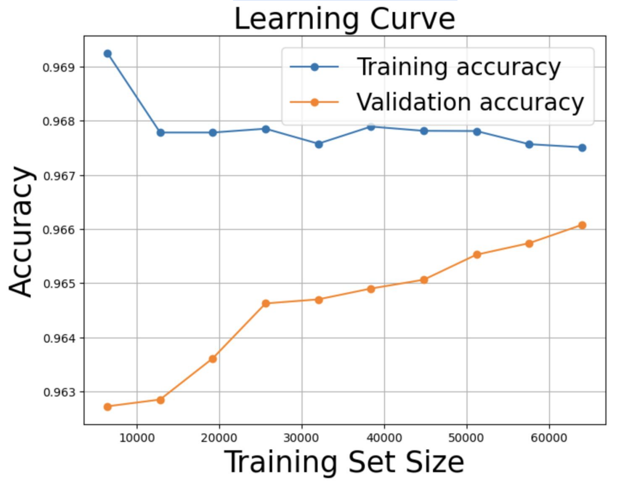
Logistic Model



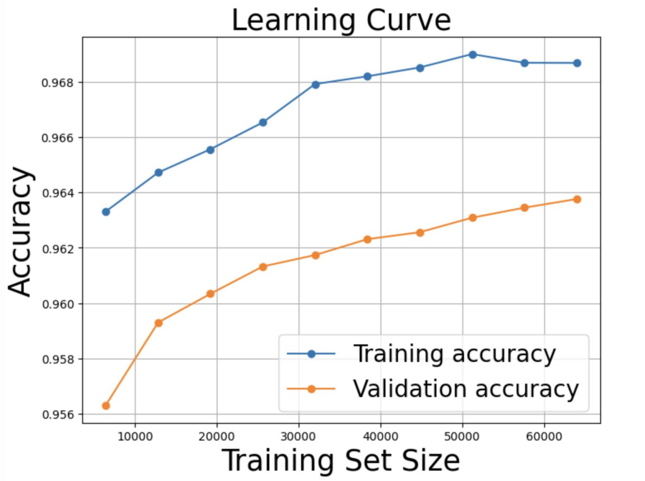
Decision Tree Model



SVM Model



KNN Model



Results

- Our final Model is **Logistic Regression Model.**

$$P(\text{diabetes} = 1 \mid X) = \frac{1}{1 + \exp\left(5.005 - \sum_i w_i x_i\right)}.$$

$$\sum_i w_i x_i = 2.53 \cdot \text{hbA1c\_level} + 1.35 \cdot \text{blood\_glucose\_level} + 1.04 \cdot \text{age} + 0.61 \cdot \text{bmi} + 0.20 \cdot \text{hypertension} + 0.16 \cdot \text{heart\_disease}$$

- Hyperparameters We Choose
  - c= **10** **(low regularization)**
  - penalty= **‘l1’** **(Lasso)**
  - solver= **‘liblinear’** **('liblinear' optimization algorithm)**

- Test Data Accuracy ≈ **95.8%** (> 91.5%)

Hyperparameter tuning:

- GridSearchCV
- 3-fold cross-validation
- Scoring metric: “Recall”

- Tuning parameters: C=[0.01,0.1,10] penalty=[‘l1’, ‘l2’] solver=[‘liblinear’, ‘sage’]

```
Fitting 3 folds for each of 16 candidates, totalling 48 fits
Logistic Best Params: {'C': 10, 'penalty': 'l1', 'solver': 'liblinear'}
Logistic Best CV Accuracy: 0.62
Logistic Validation Accuracy: 0.96
```

Test Data Confusion Matrix:

	Predicted 0(Negative)	Predicted 1(Positive)
Actual 0(No Diabetes)	9057	93
Actual 1(With Diabetes)	327	523

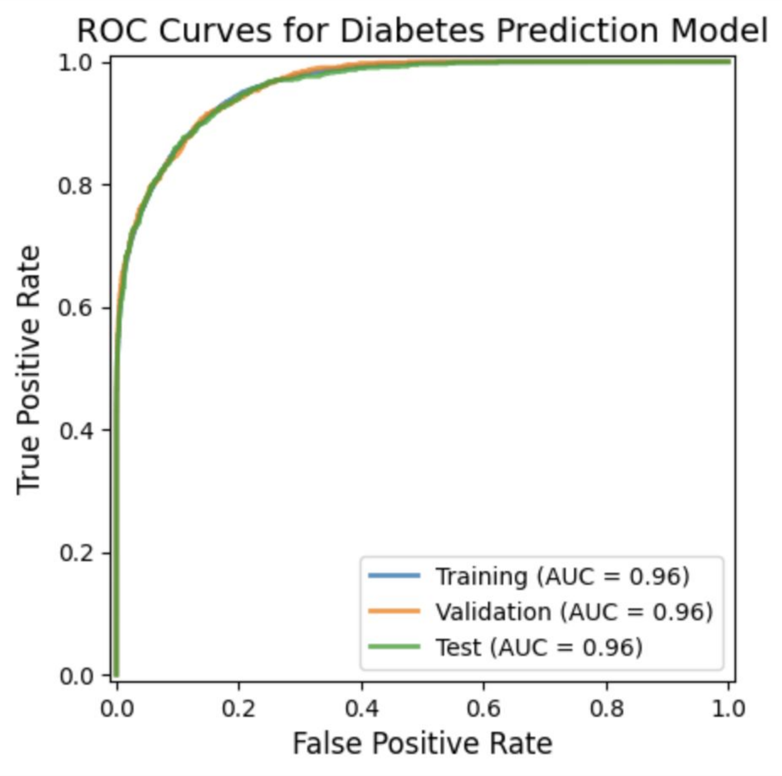
- **Precision** = 523/(523+93) ≈ **0.849**

High Confidence in Predicting Diabetes

- **Recall** = 519/(519+327) ≈ **0.613**

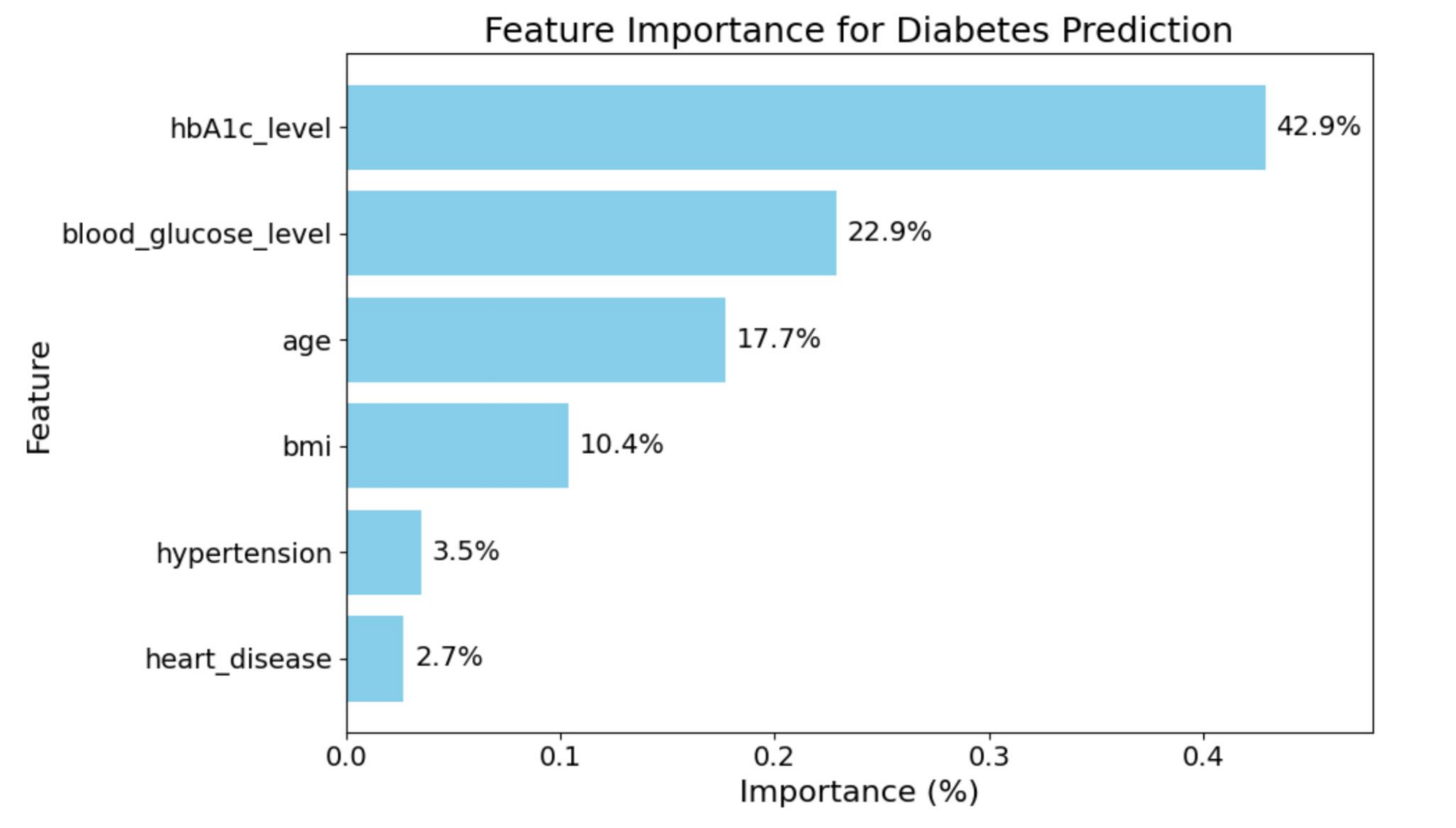
Identified Patients: 61.3% Missed Patients: 38.7%

ROC Curve Demonstrates Robust Generalization

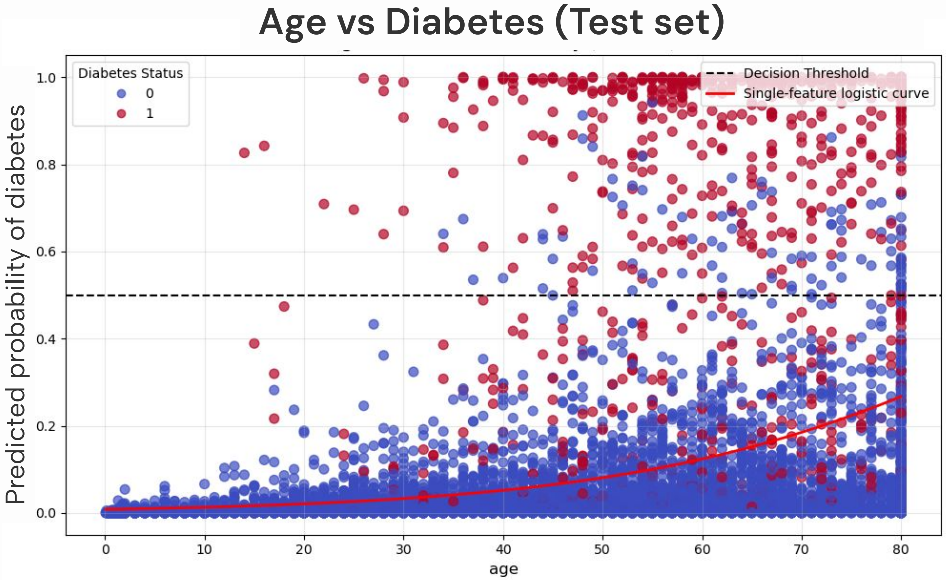


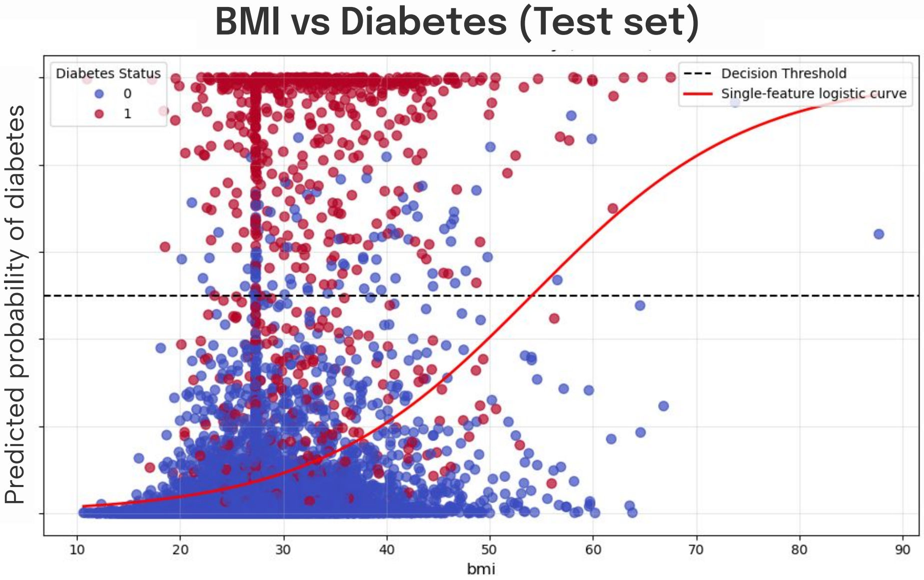
- **Consistent AUC** shows the model generalizes equally well to unseen data.

- **High TPR/FPR trade-off** allows tuning threshold for either fewer false alarms or fewer misses.


Feature Importance for Diabetes Prediction
hbA1c_level
blood_glucose_level
age
bmi
hypertension
heart_disease
42.9%
22.9%
17.7%
10.4%
3.5%
2.7%
Feature
0.0
0.1
0.2
0.3
0.4
Importance (%)

Visualize our logistic curves





- As age/BMI increases, the possibility of having diabetes will increase.
- Test data points are well-separated

Future Directions

- Interaction Terms
    - such as HbA1c * age, BMI * hypertension
- Include Other Potential Variables to Increase the Accuracy:
    - other medical conditions
    - family medical history
- Make new models to predict Type I, Type II, disease severity for other complex dataset

Thanks!

Do you have any questions? :)

Reference List

- World Health Organization (WHO). (2016). Global report on diabetes. Geneva: World Health Organization. https://www.who.int/publications/i/item/9789241565257

- Centers for Disease Control and Prevention (CDC). (2023). National Diabetes Statistics Report. https://www.cdc.gov/diabetes/data/statistics-report/index.html

- ZIYA. (2023). Diabetes Clinical Dataset (100K Rows). Available on Kaggle: https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows