# Identifying Key Health Indicators and Sociodemographic Factors Associated with Diabetes Risk

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

## 1 Introduction

By analyzing both numeric and categorical factors, we aim to identify the most influential risk factors and develop predictive models for early diabetes detection. Logistic regression outperforms other models in evaluation metrics, learning curves and ROC. A list of influential features is identified from the final model.

## 2 Data Overview and Preparation

### Dataset description and feature selection

We used the Diabetes Clinical Dataset compiled by ZIYA, which contains clinical and demographic information for 100,000 individuals. Our target variable, *Diabetes*, is a binary indicator of whether an individual has been diagnosed with diabetes (1) or not (0).

To identify relevant health and demographic variables associated with diabetes, we used a heatmap to visualize the correlation between each feature and diabetes status. Using a threshold of |correlation| > 0.1 to identify meaningful association, Age, Hypertension, Heart Disease, Bmi, HbA1c level and Blood glucose level are retained.

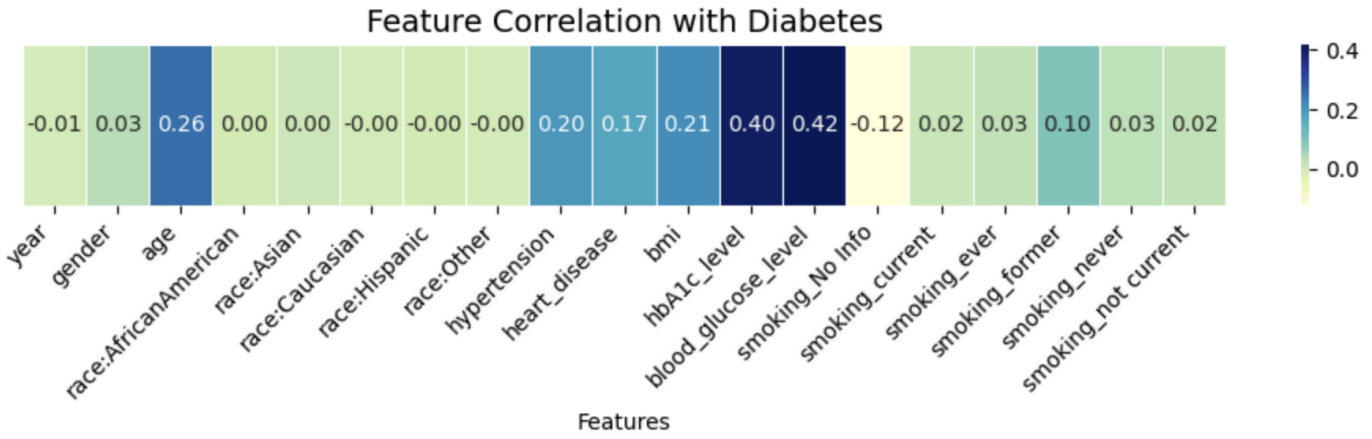


**Figure 1:** Heatmap showing the high correlation in blue

The distribution proves the potential of selected features to predict *Diabetes* in our model.

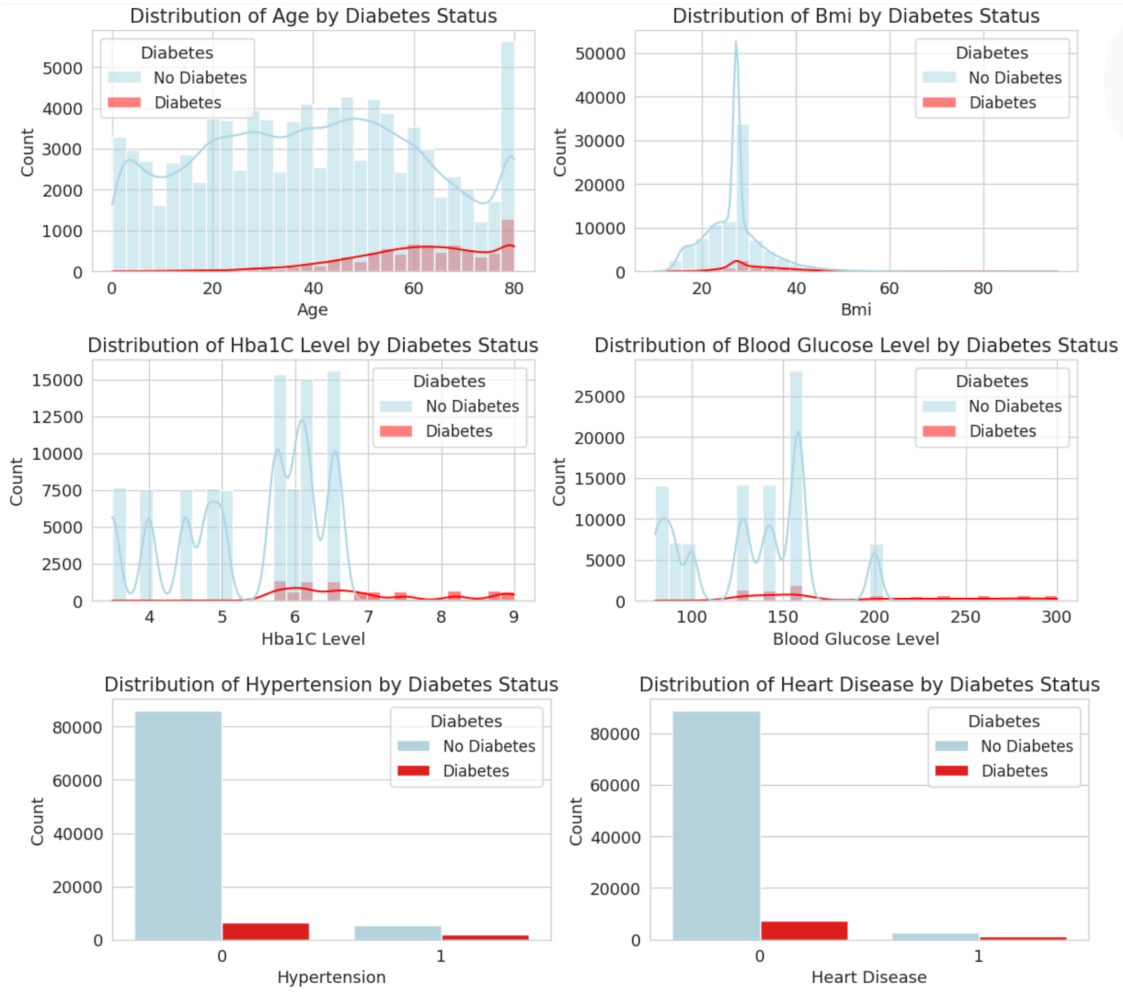


**Figure 2:** Distribution plots of continuous and categorical key features

## Data cleaning and feature engineering

We first confirmed that there are no missing values in the dataset, so no imputation was needed. To address scale differences across features, we applied standardization using StandardScaler().

## Baseline model accuracy and class imbalance

In this dataset, approximately 91.5% of individuals do not have diabetes, which results in a 91.5% accuracy in a baseline model that predicts all individuals as non-diabetic. To handle this imbalance and ensure consistent class proportions across data splits, we used *stratified sampling* when splitting the data into training (80%), validation (10%), and test (10%) sets.

## 3 Analysis

To train models on diabetes classification, we compared four classifiers: Logistic Regression, Decision Tree, Support Vector Machine and K-Nearest Neighbors .

We tuned four models using both accuracy and recall as scoring metrics during hyperparameter tuning with 3-fold cross-validation via GridSearchCV.

| **Logistics Regression** | **Decision Tree** | **SVM** | **KNN** |
|---|---|---|---|
| C = 0.01 | criterion = 'gini' | C = 10 | n_neighbors = 9 |
| penalty = 'l1' | max_depth = 3 | kernel = 'rbf' | weights = 'uniform' |
| solver = 'saga' | min_samples_split = 2 | gamma = 'auto' | metric = 'manhattan' |

**Figure 3**: Selected best hyperparameters

We assessed models using accuracy, recall and precision. Logistic Regression offers a strong balance between accuracy and recall, making it a reliable option for general diabetes prediction while maintaining interpretability through its coefficients. While other models perform less better.

| **Model** | **Accuracy*** | **Precision**** | **Recall***** |
|---|---|---|---|
| Logistics Regression | 0.9592 | 0.8708 | 0.6106 |
| Decision Tree | 0.9715 | 1.0000 | 0.6647 |
| SVM | 0.9685 | 0.9872 | 0.6376 |
| KNN | 0.9659 | 0.9411 | 0.6388 |

*reflects the overall proportion of correct predictions
**measures the model's ability to correctly identify individuals who actually have diabetes
***evaluates how many of the predicted positive cases have diabetes.

**Figure 4**: Model evaluation metrics on the validation set

We compared training and validation performance using learning curves. Logistic Regression demonstrated the most stable generalization, with training and validation accuracy closely aligned as the dataset size increased. In contrast, Decision Tree, SVM and KNN showed clear signs of overfitting.

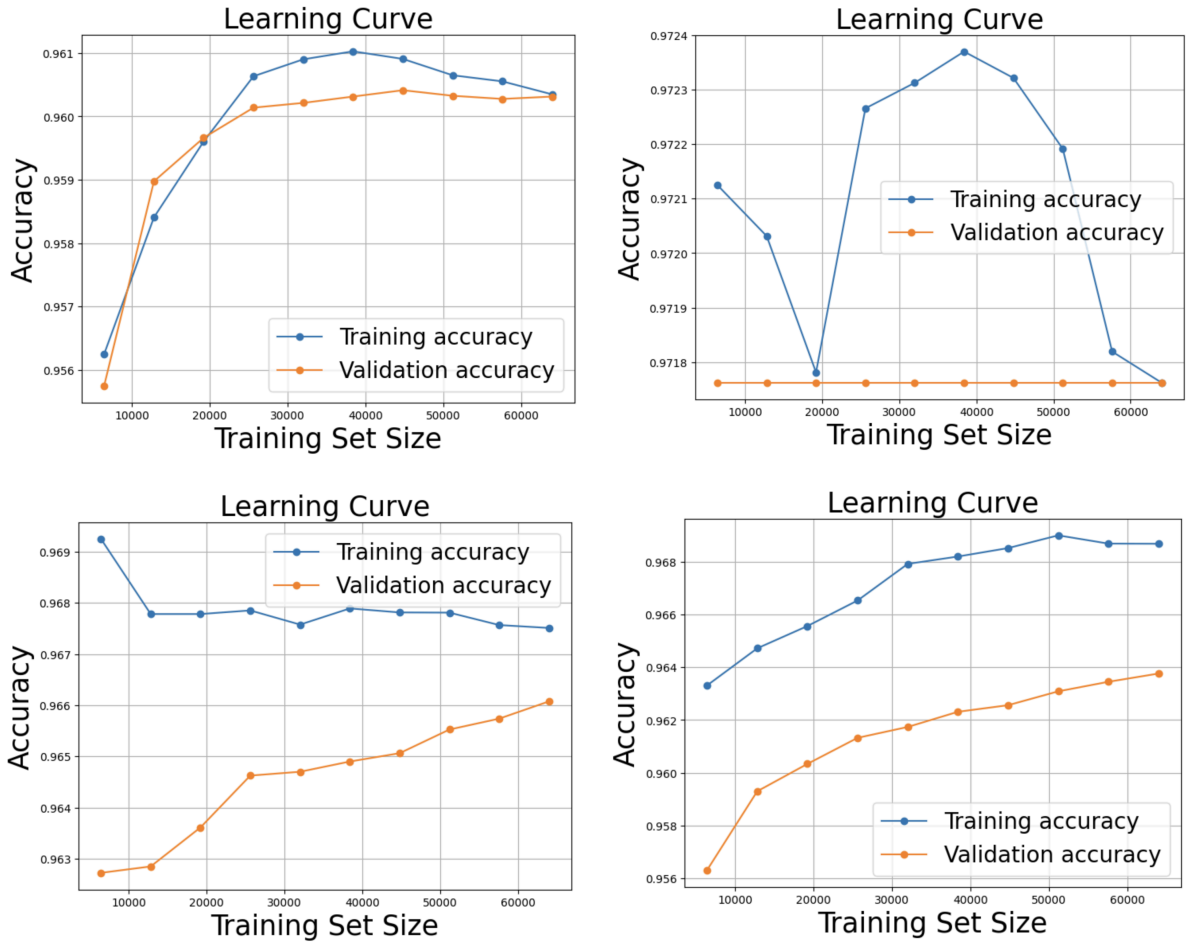


**Figure 5**: Learning curves for illustrating the overfitting

Given this comparison, we selected Logistic Regression as our final model for its balance of performance and robustness.

$$P(diabetes = 1|X) = \frac{1}{1+exp(5.005-\Sigma_i w_i x_i)}$$

$$\sum_i w_i x_i = 2.53 * HbA1c\ level + 1.35 * Blood\ Glucose\ Level + 1.04 * Age + 0.61 * Bmi + 0.20 * Hypertension + 0.16 * Heart\ disease$$

**Figure 6**: Final logistic regression equation with feature coefficients to predict probability

The ROC curves show consistent performance across training, validation, and test sets, with an AUC of 0.96 on all three. This indicates that the final logistic regression model perform robust with minimal overfitting and excellent generalization.

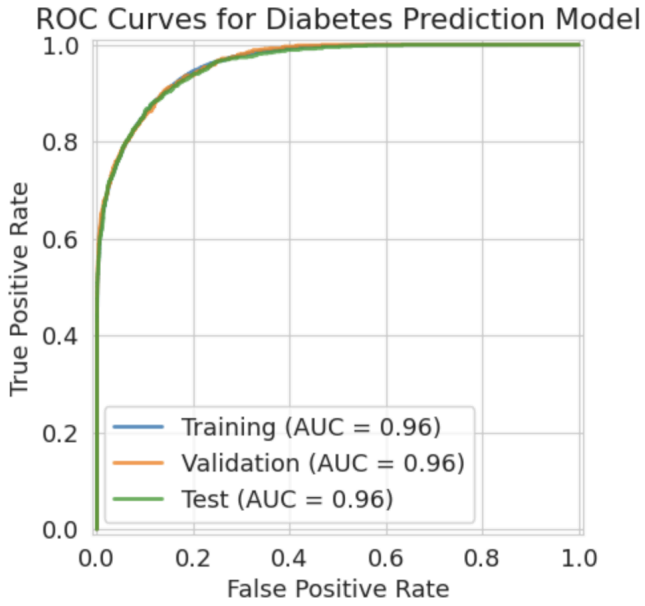


**Figure 7**: ROC Curves of the Final Logistic Regression Model

The final model identified HbA1c level and blood glucose level as the most influential features for predicting diabetes, which aligns with medical understanding of glycemic markers. Other meaningful contributors include age and BMI, suggesting that both biological markers and demographic factors impact diabetes risk.

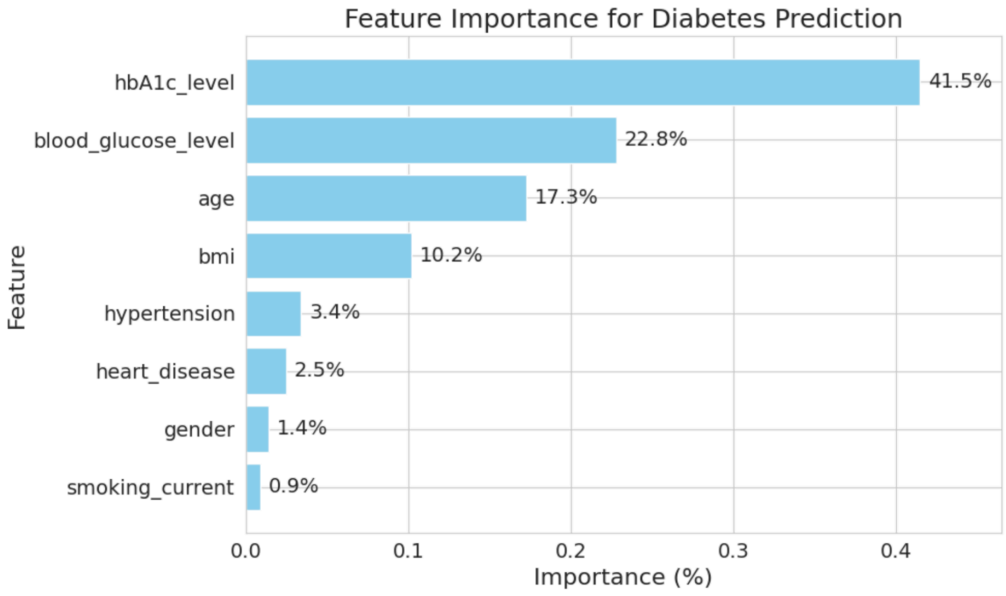


**Figure 8**: Each feature importance

## 4 Conclusion

Our research identifies the importance of each risk factor in diabetes and evaluates different model performance using stratified datasets. Future efforts include introducing interaction terms, adding clinical variables or differentiate different diabetes types/severity.

## Contribution

| Member | Proposal | Coding | Presentation | Report |
|---|---|---|---|---|
| **Alice Kang** | **1** | **1** | **1** | **1** |
| **Qidi An** | **1** | **1** | **1** | **1** |
| **Manwen Jia** | **1** | **1** | **1** | **1** |
| **Leon Li** | **1** | **1** | **1** | **1** |
| **Jiahao Chen** | **1** | **1** | **1** | **1** |