Team 2
DATS 6103: Summary Report
Professor Dr. Ning Rui
Dec 11, 2024

Diabetes Risk Prediction

Introduction

Diabetes represents one of the most pressing global health challenges of the 21st century, characterized by its chronic nature and profound impact on individual and population health. The World Health Organization estimates that diabetes affects approximately 537 million adults worldwide, with projections suggesting this number could rise to 783 million by 2045. Beyond its immediate health implications, diabetes contributes significantly to cardiovascular diseases, kidney dysfunction, neurological complications, and reduced quality of life.

This project aims to build a predictive model using machine learning techniques to classify individuals as diabetic or non-diabetic. By analyzing key features such as age, BMI, blood glucose levels, and HbA1c levels, the project seeks to uncover patterns and insights that can assist in early detection of diabetes.

Smart Questions:

- What are the primary risk factors (e.g., blood glucose level, BMI, age, hypertension) for diabetes in this population, and how accurately can they predict the likelihood of a diabetes diagnosis?
- Are there any notable differences in diabetes prevalence based on gender, age, or smoking history? For example, does smoking history combined with a high blood glucose level increase the risk?
- How does the distribution of BMI and HbA1c levels differ between those with and without diabetes?
- How sensitive are the models to changes in certain variables (e.g., slight increases in blood glucose or BMI)? Can we identify an actionable threshold for intervention?

By addressing these questions, the project aims to support early detection and prevention strategies, contributing to improved healthcare outcomes and quality of life for at-risk individuals.

Literature Review

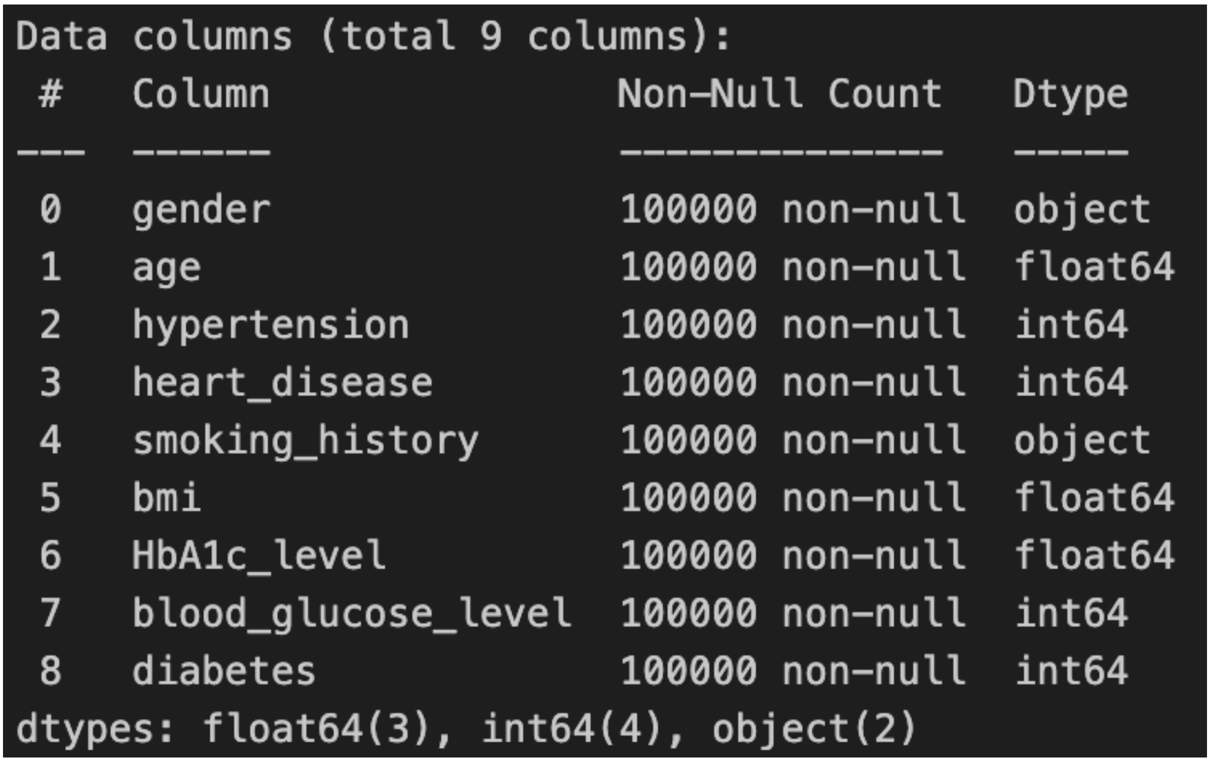
Diabetes represents one of the most pressing health challenges of the 21st century. According to the World Health Organization (WHO, 2021), diabetes affects over 537 million adults globally, with numbers projected to rise to 783 million by 2045. Beyond its direct health impacts, diabetes contributes to cardiovascular disease, kidney failure, and other comorbidities. Machine learning (ML) has emerged as a promising tool for improving diabetes prediction by leveraging medical and demographic data for early detection.

ML models have been increasingly applied to predict diabetes, leveraging algorithms like logistic regression, Random Forests, and XGBoost. A study by Kumar et al. (2019) compared these algorithms and concluded that ensemble methods (e.g., Random Forest and Gradient Boosting) outperformed linear models in accuracy and robustness. Furthermore, ML models can identify feature importance, offering insights into the factors most correlated with diabetes.

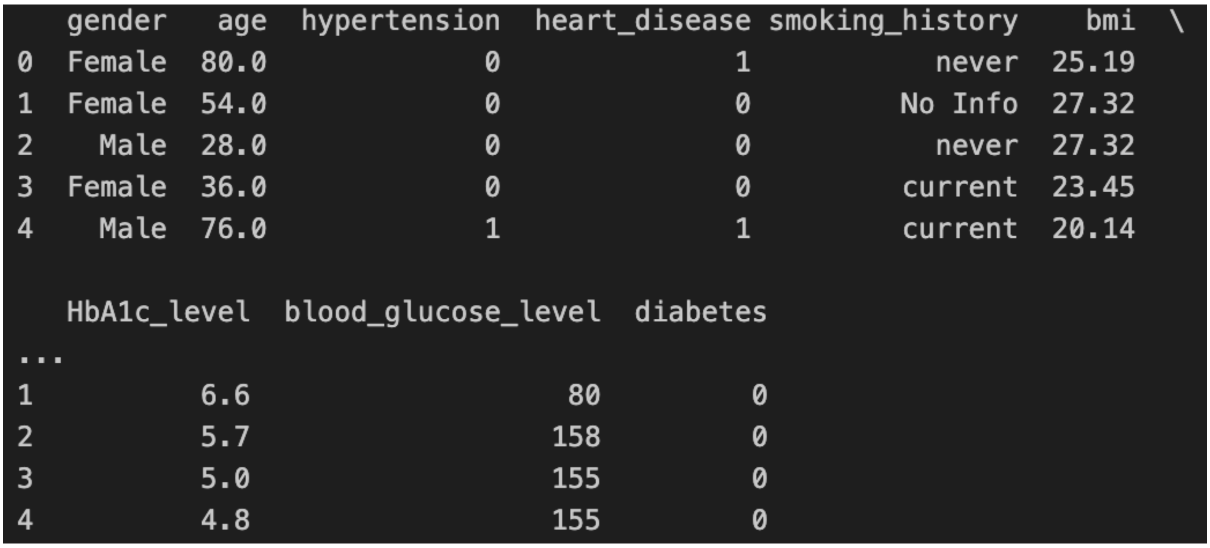
Description Of Data

The diabetes prediction dataset used in this project contains comprehensive demographic, clinical, and behavioral information from individuals, along with their diabetes status (positive or negative). The dataset includes the following features:

1. Age: Age of the individual in years.
2. BMI (Body Mass Index): A measure of body fat based on weight and height.
3. Hypertension: Whether the individual has high blood pressure (1 = Yes, 0 = No).
4. Heart Disease: Whether the individual has a history of heart disease (1 = Yes, 0 = No).
5. HbA1c Level: Average blood glucose levels over the past 2–3 months, represented as a percentage.
6. Blood Glucose Level: Blood sugar level in mg/dL.
7. Gender: Gender of the individual (e.g., Male, Female).
8. Smoking History:
   - Smoking History - Former: Indicates if the individual is a former smoker (1 = Yes, 0 = No).
   - Smoking History - Never: Indicates if the individual has never smoked (1 = Yes, 0 = No).
9. Diabetes (Target Variable): Indicates whether the individual is diabetic (1 = Yes, 0 = No).



```
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   gender               100000 non-null  object 
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64  
 3   heart_disease        100000 non-null  int64  
 4   smoking_history      100000 non-null  object 
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64  
 8   diabetes             100000 non-null  int64  
dtypes: float64(3), int64(4), object(2)
```

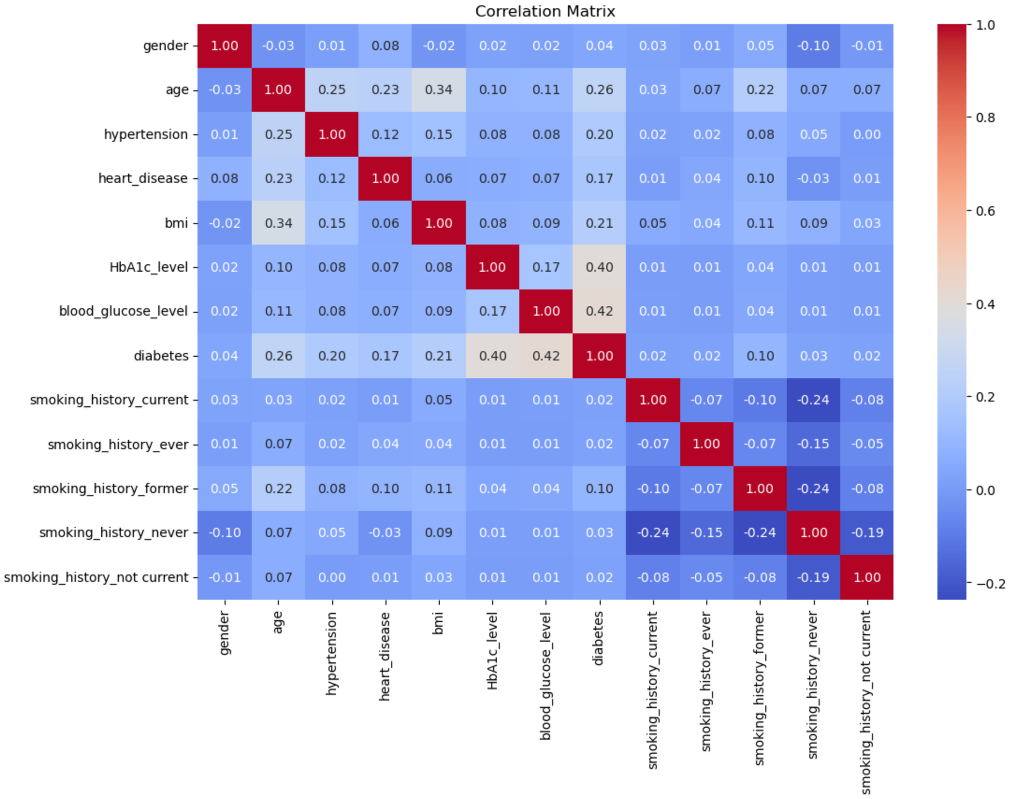


```
   gender   age  hypertension  heart_disease smoking_history    bmi  \
0  Female  80.0             0              1           never  25.19
1  Female  54.0             0              0         No Info  27.32
2    Male  28.0             0              0           never  27.32
3  Female  36.0             0              0         current  23.45
4    Male  76.0             1              1         current  20.14

   HbA1c_level  blood_glucose_level  diabetes
...
1          6.6                   80         0
2          5.7                  158         0
3          5.0                  155         0
4          4.8                  155         0
```

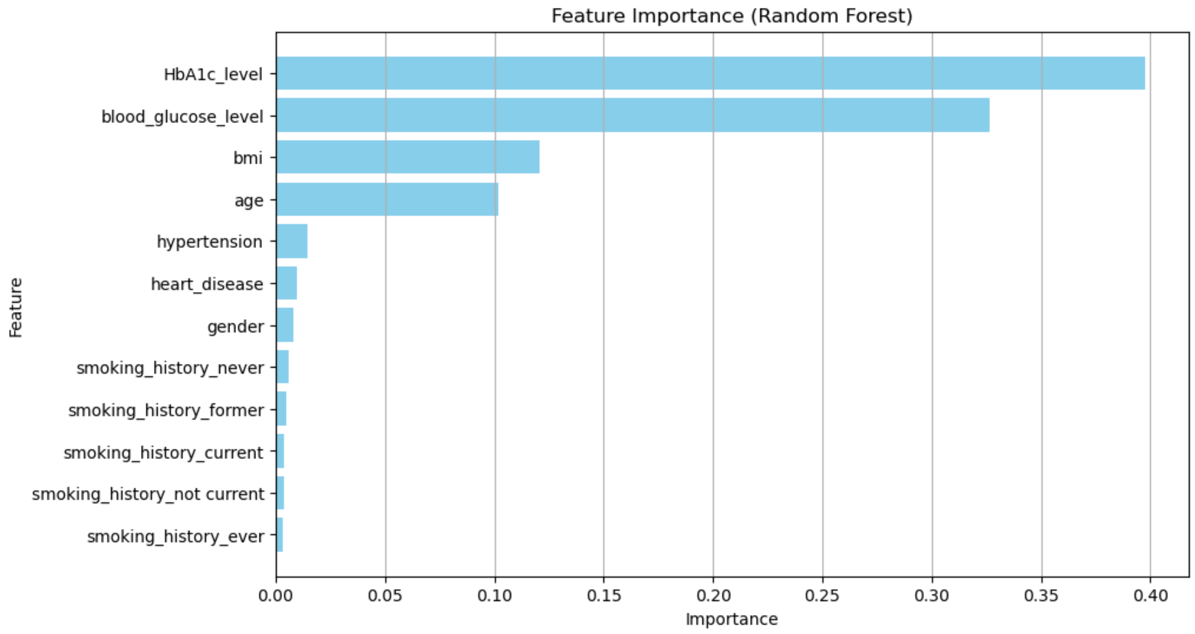
Smart Questions

Smart Question 1 : What are the primary risk factors (e.g., blood glucose level, BMI, age, hypertension) for diabetes in this population, and how accurately can they predict the likelihood of a diabetes diagnosis?

The analysis of the dataset, using both correlation metrics and feature importance from the Random Forest model, has identified several key factors that contribute significantly to the likelihood of a diabetes diagnosis.

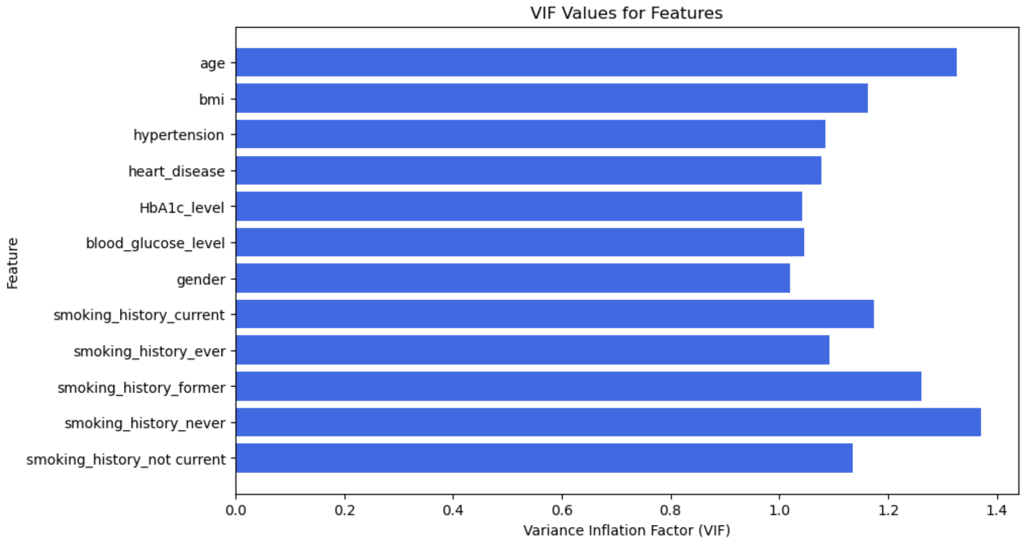


	gender	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes	smoking_history_current	smoking_history_ever	smoking_history_former	smoking_history_never	smoking_history_not current
gender	1.00	-0.03	0.01	0.08	-0.02	0.02	0.02	0.04	0.03	0.01	0.05	-0.10	-0.01
age	-0.03	1.00	0.25	0.23	0.34	0.10	0.11	0.26	0.03	0.07	0.22	0.07	0.07
hypertension	0.01	0.25	1.00	0.12	0.15	0.08	0.08	0.20	0.02	0.02	0.08	0.05	0.00
heart_disease	0.08	0.23	0.12	1.00	0.06	0.07	0.07	0.17	0.01	0.04	0.10	-0.03	0.01
bmi	-0.02	0.34	0.15	0.06	1.00	0.08	0.09	0.21	0.05	0.04	0.11	0.09	0.03
HbA1c_level	0.02	0.10	0.08	0.07	0.08	1.00	0.17	0.40	0.01	0.01	0.04	0.01	0.01
blood_glucose_level	0.02	0.11	0.08	0.07	0.09	0.17	1.00	0.42	0.01	0.01	0.04	0.01	0.01
diabetes	0.04	0.26	0.20	0.17	0.21	0.40	0.42	1.00	0.02	0.02	0.10	0.03	0.02
smoking_history_current	0.03	0.03	0.02	0.01	0.05	0.01	0.01	0.02	1.00	-0.07	-0.10	-0.24	-0.08
smoking_history_ever	0.01	0.07	0.02	0.04	0.04	0.01	0.01	0.02	-0.07	1.00	-0.07	-0.15	-0.05
smoking_history_former	0.05	0.22	0.08	0.10	0.11	0.04	0.04	0.10	-0.10	-0.07	1.00	-0.24	-0.08
smoking_history_never	-0.10	0.07	0.05	-0.03	0.09	0.01	0.01	0.03	-0.24	-0.15	-0.24	1.00	-0.19
smoking_history_not current	-0.01	0.07	0.00	0.01	0.03	0.01	0.01	0.02	-0.08	-0.05	-0.08	-0.19	1.00



Key Risk Factors Identified

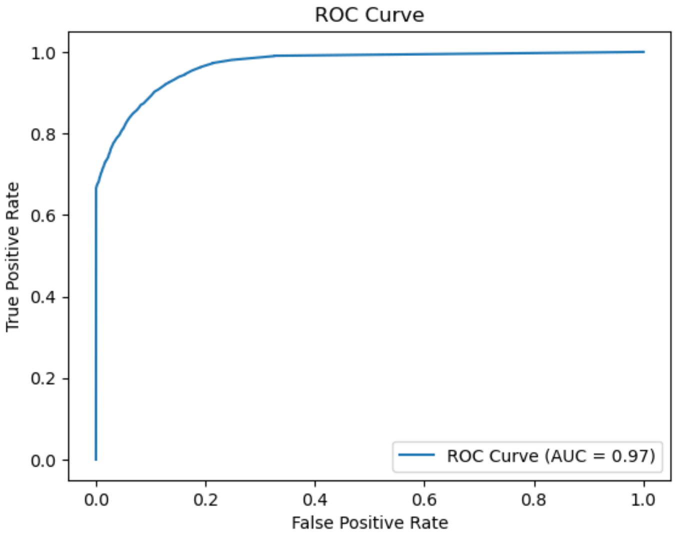
1. **Feature Importance Analysis**:
   - **HbA1c Level**: Identified as the most significant predictor with an importance score of **39.76%**. HbA1c provides an average measure of blood glucose levels over the past 2-3 months, making it critical for diabetes diagnosis.
   - **Blood Glucose Level**: The second most important feature (**32.62%**), directly correlating with the presence of diabetes.
   - **BMI**: With an importance of **12.09%**, BMI serves as a major risk factor, indicating the role of obesity in diabetes prevalence.
   - **Age**: Contributing **10.17%** to the model, age is a well-known risk factor, as diabetes prevalence increases with age.
   - **Hypertension**: While its importance is lower (**1.46%**), hypertension remains relevant as a co-morbidity linked to diabetes.
2. **Correlation Matrix**:
   - Strong correlations were observed between:
     - **HbA1c Level** and **Diabetes (r = 0.40)**.
     - **Blood Glucose Level** and **Diabetes (r = 0.42)**.
   - BMI and Age also showed moderate correlations with diabetes.
3. **Variance Inflation Factor (VIF)**:
   - All features have VIF values below 5, confirming no significant multicollinearity, ensuring the reliability of the model's predictions.



Model Performance

The Random Forest model, trained using SMOTE to address class imbalance, achieved strong performance metrics:

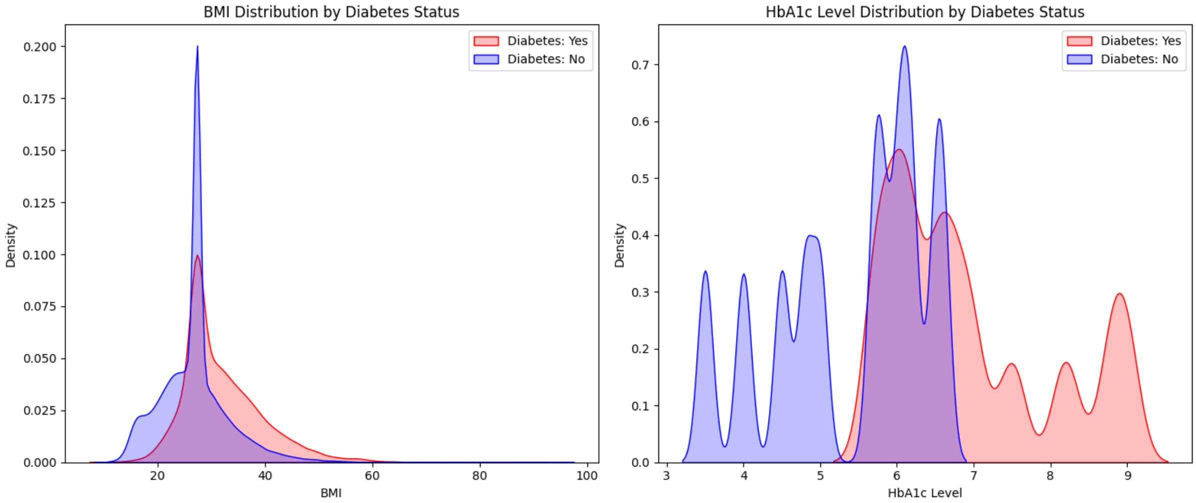
1. **Classification Metrics**:
   - **Accuracy**: 96.32% overall accuracy, reflecting the model's reliability in predicting diabetes.
   - **Precision**: 83% for Class 1 (Diabetic), indicating that 83% of predicted diabetic cases are correct.
   - **Recall**: 72% for Class 1, showing the model successfully identifies 72% of true diabetic cases.
   - **F1-Score**: 77% for Class 1, balancing precision and recall.
2. **ROC-AUC**:
   - The model achieved a high ROC-AUC score of **96.81%**, indicating excellent discriminatory power between diabetic and non-diabetic cases.



This analysis demonstrates that **HbA1c Level**, **Blood Glucose Level**, **BMI**, and **Age** are the primary risk factors for diabetes in this population. The Random Forest model, enhanced by

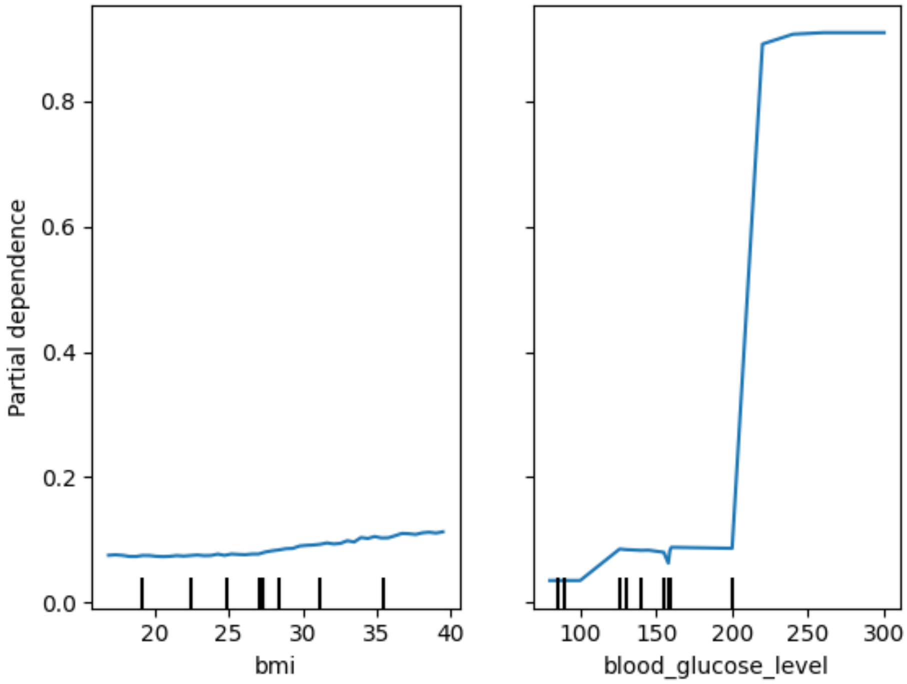
SMOTE, effectively predicts diabetes with high accuracy and precision, achieving an ROC-AUC of **96.81%**. However, efforts to improve recall (72%) may further enhance the model's utility in identifying at-risk individuals. These findings reinforce the importance of targeted interventions focused on key clinical and demographic factors for early diabetes detection.

Smart Question 2 : How does the distribution of BMI and HbA1c levels differ between those with and without diabetes?



Builds a diabetes prediction model using a Random Forest Classifier. It preprocesses the dataset by encoding categorical variables and handling missing values. The data is split into training and testing sets, and the model's performance is evaluated using metrics like confusion matrix, classification report, and AUC-ROC score. Partial Dependence Plots reveal the impact of BMI and blood glucose levels on predictions. Statistical summaries highlight differences in BMI and HbA1c levels between diabetic and non-diabetic individuals. Visualizations illustrate distributions of BMI and HbA1c levels, while threshold analysis suggests potential screening values for glucose and BMI based on the data.

Smart Question 3: How sensitive are the models to changes in certain variables (e.g., slight increases in blood glucose or BMI)? Can we identify an actionable threshold for intervention?



1. **Model Evaluation (ROC-AUC Score):**
   - **ROC-AUC Score:** 0.9635
     - This indicates that the Random Forest model has excellent performance in distinguishing between diabetic and non-diabetic individuals. A value closer to 1 suggests high discriminatory power.
2. **Partial Dependence Plots (PDP):**
   - The PDPs for **BMI** and **blood_glucose_level** illustrate how changes in these variables impact the model's predictions.
   - **Key Observations:**
     - **Blood Glucose Level:** A steep increase in predicted diabetes probability is observed as blood glucose level rises, especially beyond certain thresholds, indicating high sensitivity to this variable.
     - **BMI:** The relationship with diabetes probability may show a more gradual increase, suggesting moderate sensitivity compared to blood glucose.

       1.
3. **Threshold Analysis (Mean-Based):**
   - **Average Blood Glucose Level for Diabetic Individuals (Diabetes=1):** 194.09 mg/dL
   - **Average BMI for Diabetic Individuals (Diabetes=1):** 31.99
   - **90% Thresholds for High Risk:**
     - **Blood Glucose Level Threshold:** 174.69 mg/dL
     - **BMI Threshold:** 28.79
   - These thresholds are actionable points where individuals exceeding these values could be flagged for closer monitoring or intervention.
4. **Sensitivity Analysis:**
   - The PDPs show that **blood glucose level** is highly sensitive, as small increases beyond certain points lead to significant rises in diabetes probability.

- **BMI** has a more gradual effect, suggesting it is less sensitive but still an important risk factor.

5. **Actionable Thresholds for Intervention:**
   The calculated thresholds (174.69 mg/dL for blood glucose and 28.79 for BMI) represent high-risk points for diabetic individuals.

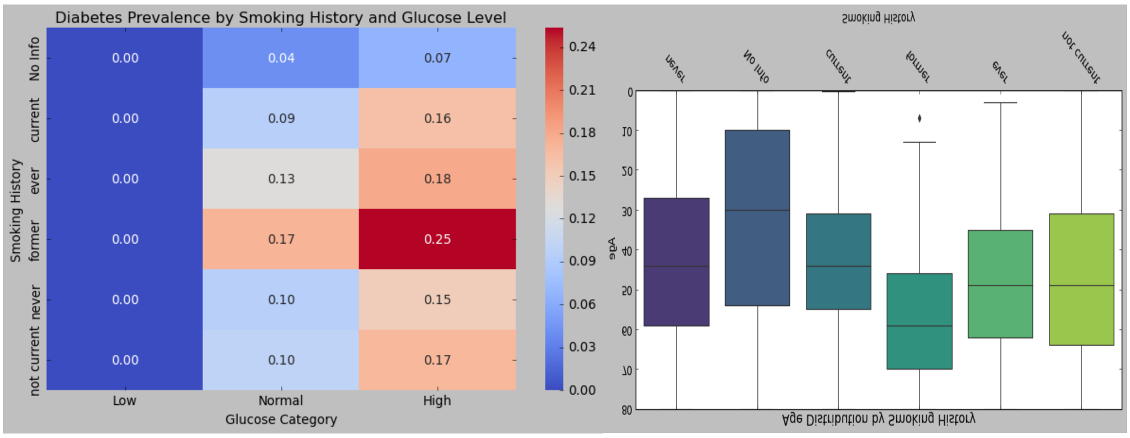
   These thresholds can guide preventive measures:

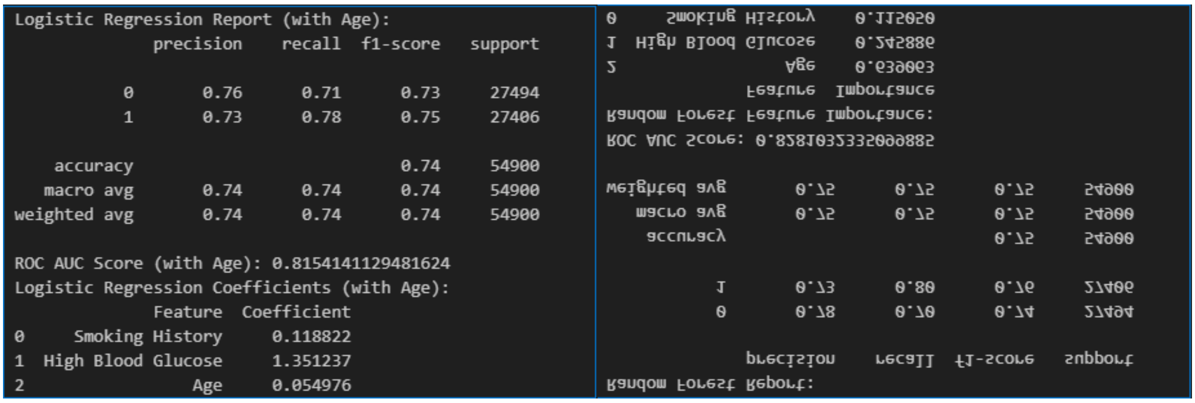
   - Individuals with blood glucose levels approaching or exceeding 174.69 mg/dL might benefit from medical evaluation.
   - Similarly, individuals with BMI exceeding 28.79 may need lifestyle interventions.

This analysis highlights how the model's predictions are influenced by key variables and identifies actionable insights for managing diabetes risk.

Smart Question 4: Are there any notable differences in diabetes prevalence based on gender, age, or smoking history? Does smoking history and age combined with high blood glucose level increase the risk?

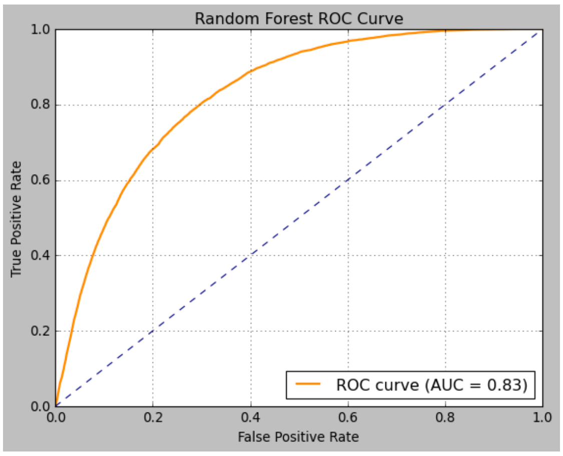
Diabetes prevalence is roughly similar across genders, with males having a slightly higher proportion of diabetes cases compared to females and others. Diabetes prevalence increases with age, with the elderly group having the highest prevalence. Former smokers have the highest diabetes prevalence, followed by those who have ever smoked.
The heatmap shows diabetes prevalence between different types of smokers over low, normal, and high blood glucose levels (80-100, 100-140, >140)mg/dL. The heatmap confirms that former smokers with both normal and high glucose levels are more at risk of diabetes compared to people with other categories, especially people who never smoked. The age plots vs smoking category confirms age's role in diabetes risk, as former smokers have the highest average age. Logistic regression and random forest were used to predict diabetes risk and identify key risk factors. Logistic regression provided a simple and interpretable model that quantified the impact of features like blood glucose levels on diabetes risk, offering clear insights into individual predictors. Meanwhile, random forest captures non-linear relationships and interactions between variables such as age. This leads to age being repeatedly used as a splitting variable which increases its importance in random forest.





```
Logistic Regression Report (with Age):
              precision    recall  f1-score   support

           0       0.76      0.71      0.73     27494
           1       0.73      0.78      0.75     27406

    accuracy                           0.74     54900
   macro avg       0.74      0.74      0.74     54900
weighted avg       0.74      0.74      0.74     54900

ROC AUC Score (with Age): 0.8154141129481624
Logistic Regression Coefficients (with Age):
              Feature  Coefficient
0     Smoking History     0.118822
1  High Blood Glucose     1.351237
2                 Age     0.054976
```

```
Random Forest Report:
              precision    recall  f1-score   support

           0       0.78      0.70      0.74     27494
           1       0.73      0.80      0.76     27406

    accuracy                           0.75     54900
   macro avg       0.75      0.75      0.75     54900
weighted avg       0.75      0.75      0.75     54900

ROC AUC Score: 0.8281032335000885
Random Forest Feature Importance:
              Feature  Importance
2                 Age    0.639063
1  High Blood Glucose    0.245886
0     Smoking History    0.115050
```



Limitations

Class Imbalance: Despite using SMOTE (Synthetic Minority Oversampling Technique) to address the imbalance between diabetic and non-diabetic cases, synthetic samples may not fully represent the complexity of real-world data. This could affect the model's ability to generalize.

Feature Limitations: The dataset lacks certain critical features that could enhance prediction accuracy, such as physical activity levels, dietary patterns, and family medical history. Additionally, some encoded features, like smoking history, may oversimplify their impact on diabetes prediction.

Overfitting Risk: The use of SMOTE and hyperparameter tuning, while enhancing performance, could lead to overfitting, particularly if cross-validation is not implemented thoroughly.

Threshold Trade-offs: Adjusting the classification threshold to improve recall for diabetic cases may reduce precision, leading to an increased number of false positives. This trade-off might have practical implications in real-world diagnostic settings

Static Dataset: The dataset is static and does not reflect temporal changes in health trends or advancements in diagnostic practices. Incorporating dynamic, real-time data could improve the robustness of the model.

Labels: The dataset lacks explanations of the categories of the smoking history variable names, such as current, not current, etc.… There is also no indication if the categories are mutually exclusive.

Conclusion

This project provided significant insights into the key risk factors associated with diabetes and showcased the potential of machine learning models in predicting diabetes risk. By combining exploratory data analysis (EDA), feature importance evaluation, sensitivity analysis, and predictive modeling, the project successfully identified critical predictors such as HbA1c levels, blood glucose levels, BMI, and age. These findings underscore the importance of addressing both clinical and demographic factors in diabetes risk assessment.

The use of techniques such as SMOTE for handling class imbalance and Random Forests for feature importance and prediction accuracy further highlights the practical utility of machine learning in healthcare applications. The model achieved strong performance metrics, including an accuracy of 96.32% and a ROC-AUC score of 96.81%, demonstrating its robustness and reliability.

However, the analysis also identified areas for improvement, particularly in enhancing recall for diabetic cases to reduce false negatives. Future research could focus on incorporating additional variables, such as dietary habits, physical activity, and family history, to improve predictive power and provide a more comprehensive understanding of diabetes risk. Furthermore, the integration of real-time data from wearable devices and longitudinal datasets could enhance the model's applicability in dynamic, real-world healthcare environments.

By leveraging these advancements, this project lays a foundation for developing actionable tools to aid early detection, personalized intervention, and effective diabetes prevention strategies.

References

1. Sonar, P., & JayaMalini, K. (2019). *Diabetes prediction using different machine learning approaches*. In *Proceedings of the 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 1289-1294). IEEE. https://ieeexplore.ieee.org/abstract/document/8819841?casa_token=hBQe3TnvUoIAAAAA:5ks4tZqepYwa-TjTShO6p_iNUbLkxPRx-4KBS1XEAvu4OdzHa-cK_MnpEvLSg8KQVJXN2j6g
2. Swain, A., Mohanty, S. N., & Das, A. C. (2016). *Comparative risk analysis on prediction of diabetes mellitus using machine learning approach*. In *Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT)*.IEEE. https://ieeexplore.ieee.org/abstract/document/7755319?casa_token=MtmTxTX20GMAAAAA:NddAriqQS60PXyQ3LrfgTRyYnDQJdEIpQmyD2MV6g43BQR-PwxEa-4ocf8a5Rj3yifH59nlO