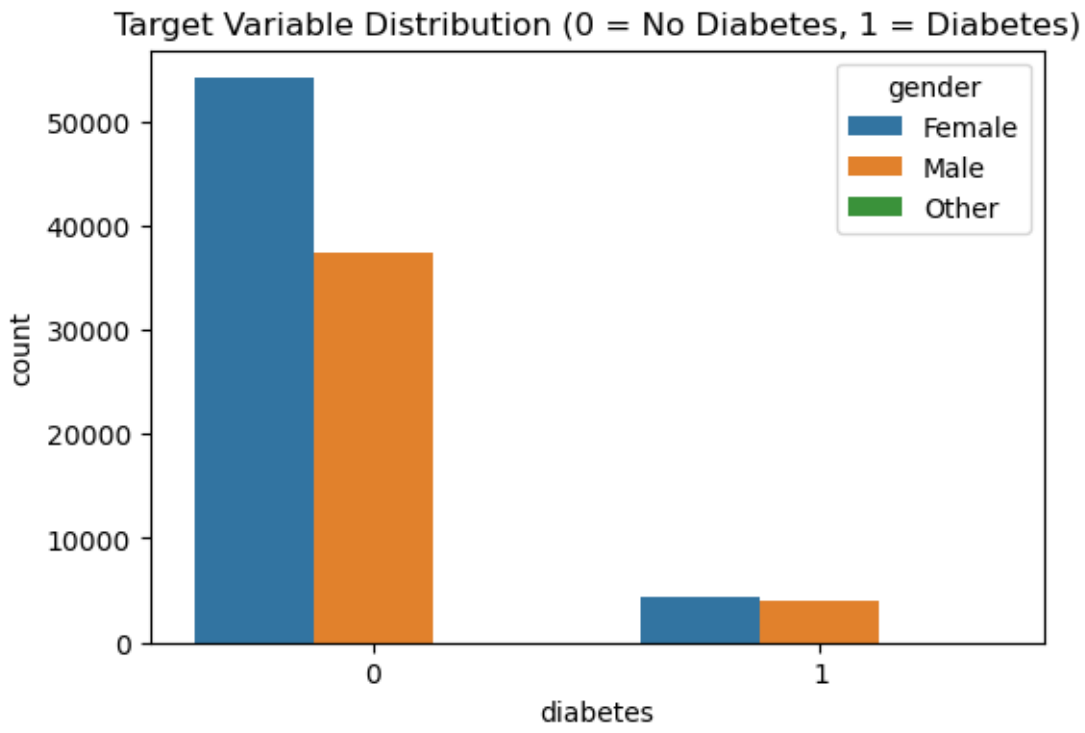
1). Target Variable Distribution

- The x-axis represents the diabetes status: 0 indicates no diabetes, and 1 indicates diabetes.
- The y-axis shows the count of individuals in each category.
- The bars are grouped by gender: Female (blue), Male (orange), and Other (green).
- From the plot, it is clear that most individuals do not have diabetes (category 0), with females outnumbering males in both categories.
- The count of individuals with diabetes (category 1) is significantly lower and roughly balanced between females and males.
- The "Other" gender category has very low counts in both diabetes categories.

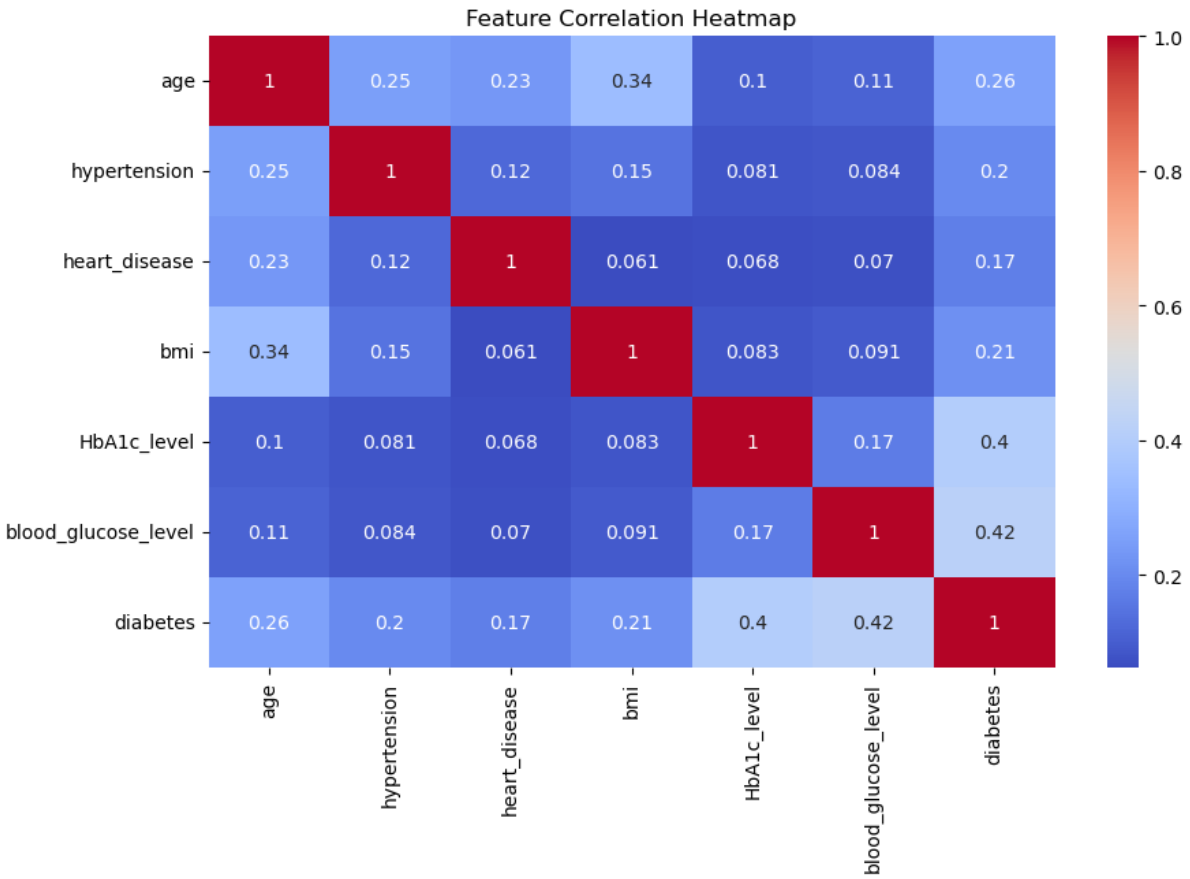


2). Feature Correlation

- The heatmap uses color intensity to represent correlation values, with **red indicating stronger positive correlation (1.0)** and **blue indicating weaker or no correlation**.
- Diagonal cells have a correlation of 1.0 since each feature perfectly correlates with itself.
- Notable correlations with diabetes (last column and row) include:
    - **blood_glucose_level (0.42)**: The strongest positive correlation, indicating higher glucose is strongly associated with diabetes.
    - **HbA1c_level (0.40)**: Also strongly correlated with diabetes, a known long-term indicator of blood sugar control.
    - **bmi (0.21) and age (0.26)**: Moderate positive correlations with diabetes.
    - Other features like hypertension (0.20) and heart disease (0.17) show weaker but positive correlations.

Interpretation:

This heatmap helps identify important predictors for diabetes, such as glucose and HbA1c levels, validating clinical understanding. It guides feature importance and selection for model building.



3). Model Evaluation and Confusion Matrix

- Model Accuracy: 0.95905 → The model correctly predicts ~95.9% of cases overall.
- ROC AUC Score: 0.9617 → Excellent discriminative ability between the two classes (close to 1).

Classification Report (per class performance):

- Class 0 (No Diabetes):
    - Precision: 0.97 → 97% of predicted non-diabetic cases were correct.
    - Recall: 0.99 → 99% of actual non-diabetic cases were correctly detected.
    - F1-score: 0.98 → High balance between precision and recall.
    - Support: 18,292 samples.
- Class 1 (Diabetes):
    - Precision: 0.86 → 86% of predicted diabetic cases were correct.
    - Recall: 0.62 → Only 62% of actual diabetics were correctly detected.
    - F1-score: 0.72 → Lower than Class 0 due to recall.
    - Support: 1,708 samples.

Averaged Scores:

- Macro Avg: Treats both classes equally (useful for imbalanced data).
    - F1-score: 0.85
- Weighted Avg: Weights classes by their frequency.
    - F1-score: 0.96 (closer to overall accuracy due to Class 0 dominance)

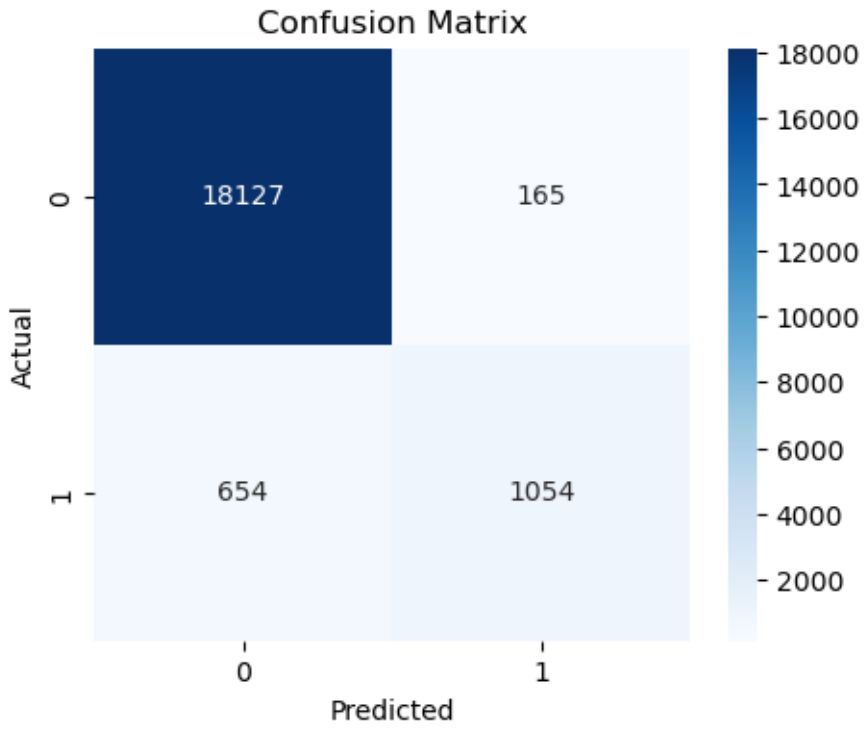
True Negatives (TN) = 18,127

- Correctly predicted negatives (healthy people).
- Very high number → model does well at identifying negatives.

2. **True Positives (TP) = 1,054**
   - Correctly predicted positives (people with diabetes).
   - Decent number, but smaller than TN, suggesting fewer positive cases in data.
3. **False Negatives (FN) = 654**
   - Positive cases misclassified as negative (diabetic patients missed).
   - Important in healthcare: missing a diabetic patient can have serious consequences.
4. **False Positives (FP) = 165**
   - Negative cases misclassified as positive (healthy people flagged as diabetic).
   - Less critical than FN, but could cause unnecessary anxiety or tests.

Model is **very good at predicting non-diabetic patients**, but **misses a significant number of diabetic patients**.

In medical applications, **recall (sensitivity)** is critical to reduce false negatives. This model may need tuning or resampling to improve recall.



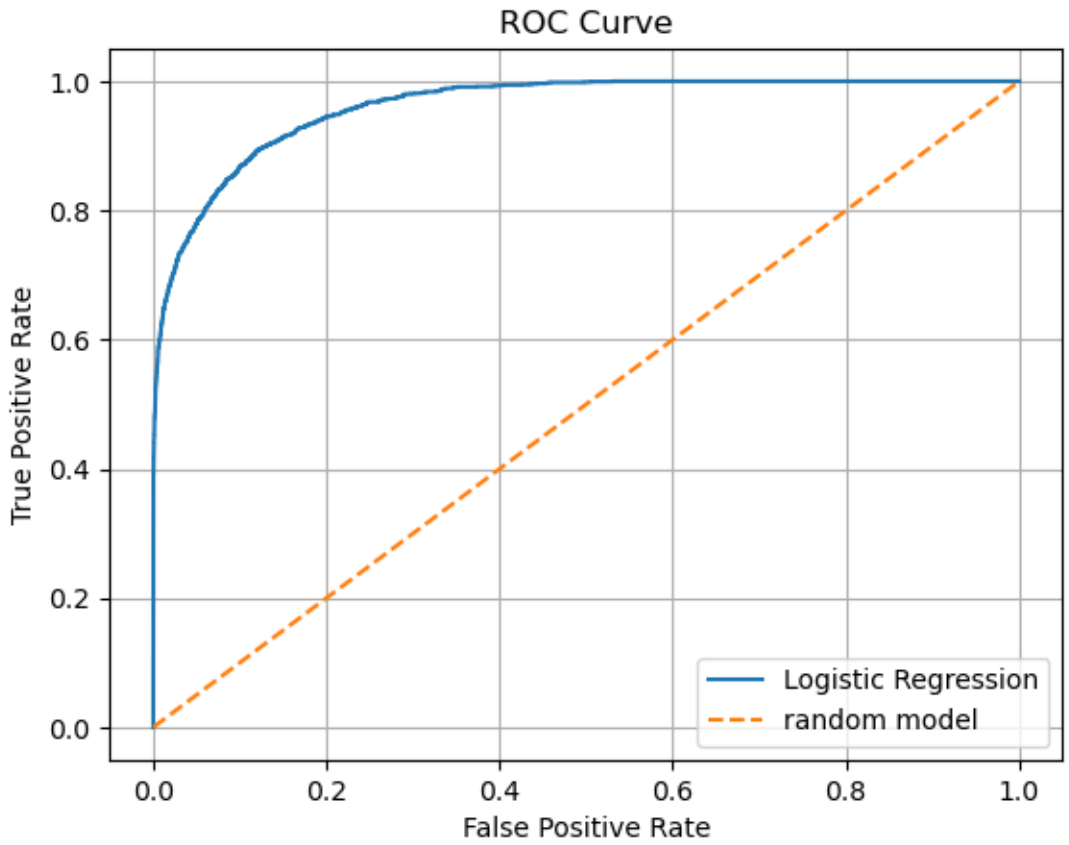
4) Model Interpretation and Visualization

What the Plot Shows:

- **X-axis**: False Positive Rate (FPR)
- **Y-axis**: True Positive Rate (TPR)
- **Blue Curve**: Performance of the Logistic Regression model
- **Orange Diagonal Line**: Baseline performance of a random classifier (no predictive power)

 The **blue ROC curve bows toward the top-left corner**, which is ideal. It indicates that the model achieves a **high TPR** while keeping the **FPR low**—a sign of strong discriminative ability.

- The **orange diagonal line** represents random guessing. The farther the blue curve is from this line, the better the model is at distinguishing between the two classes.
- The **area under the curve (AUC)**—though not explicitly labeled—would likely be **high (close to 1.0)**, suggesting excellent model performance.

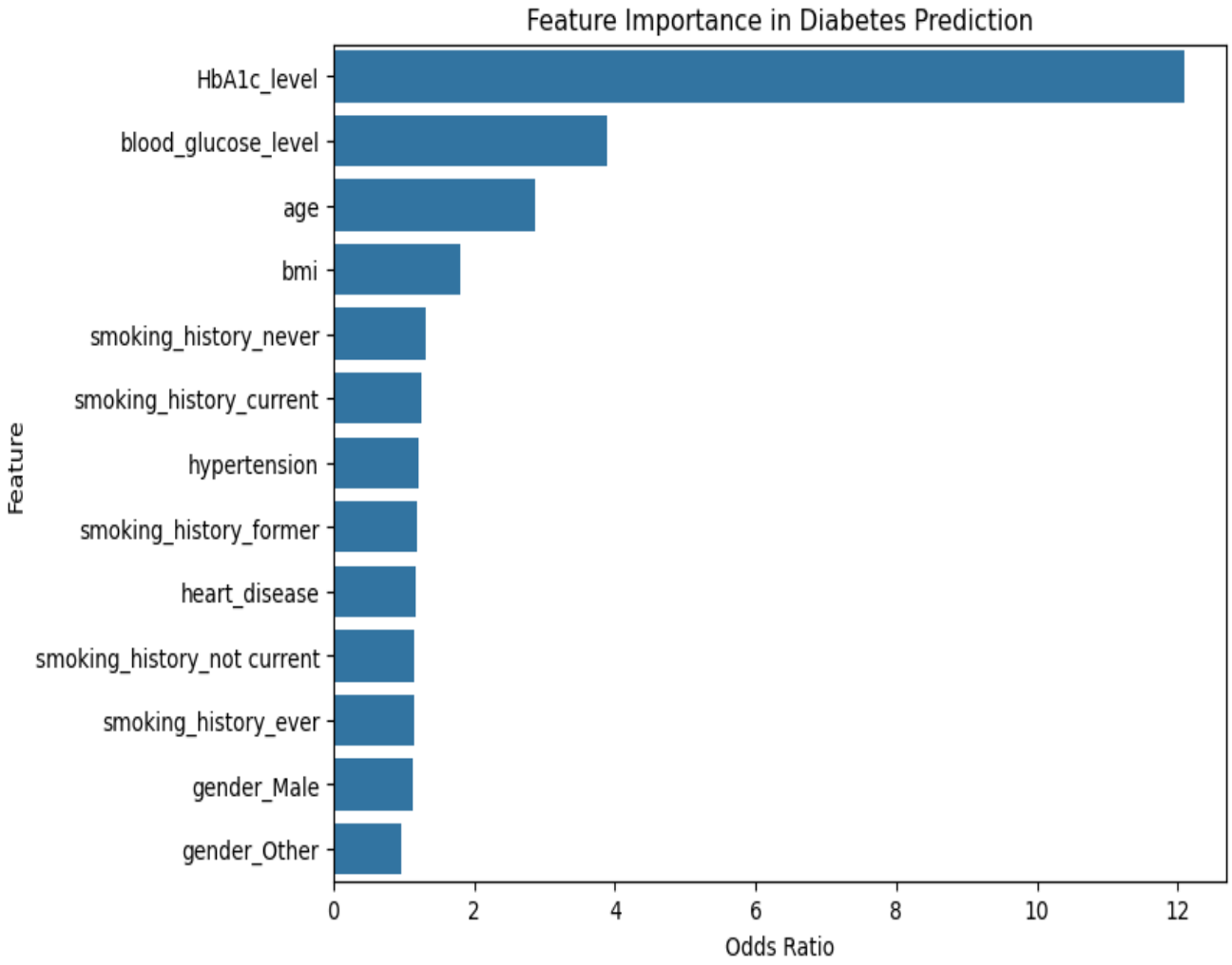


5). Insights from Model

- **X-axis**: **Odds Ratio** – a measure of how strongly each feature influences the odds of having diabetes.
- **Y-axis**: **Feature** – the variables (predictors) used in the model.
- **Bars**: Show the magnitude of the odds ratio for each feature. A higher bar indicates a greater impact on the prediction.
- **Top Predictors**:
  - 🔹 **HbA1c_level** is by far the most important feature, with an odds ratio >12. This aligns with clinical knowledge since HbA1c is a key indicator of long-term blood sugar levels.
  - 🔹 **blood_glucose_level** and **age** also have strong predictive power, suggesting current glucose levels and age are also strong indicators of diabetes risk.
- **Moderate Predictors**:
  - 🔹 **BMI (Body Mass Index)** – consistent with known risk factors.
  - 🔹 **Smoking history (never, current, former, etc.)** – present but less impactful than blood markers or age.
- **Less Influential Features**:
  - 🔹 **Gender (Male, Other)**, **heart_disease**, and some smoking-related categories have **lower odds ratios**, indicating **weaker predictive power** in this model.
- **Model Interpretability**:
  - Odds ratios are commonly used in **logistic regression**, which this chart likely comes from.
  - An odds ratio >1 means the feature increases the odds of diabetes, while <1 would reduce it (none appear <1 here, but we can't confirm that without directionality).

Conclusion:

- **HbA1c_level** is the dominant predictor, making it a critical input for any diabetes-related model.
- Features like **blood glucose**, **age**, and **BMI** are also important and should be retained.
- Some variables (e.g., certain smoking categories or gender) have lower influence and may be candidates for exclusion if simplifying the model is a goal.
- This kind of analysis helps in **feature selection** and **model explainability**, especially important in healthcare applications.


Feature Importance in Diabetes Prediction
HbA1c_level
blood_glucose_level
age
bmi
smoking_history_never
smoking_history_current
hypertension
smoking_history_former
heart_disease
smoking_history_not current
smoking_history_ever
gender_Male
gender_Other
Feature
0
2
4
6
8
10
12
Odds Ratio