**Diabetes Prediction Model Performance Report**

**Model Overview**

This report documents the performance and insights derived from a Random Forest classification model developed to predict diabetes risk based on historical patient data. The dataset comprises 100,000 records with features including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level, with the target variable being diabetes diagnosis (0 for non-diabetic, 1 for diabetic).

**Data Preparation**

Cleaning : Missing values were handled by filling numeric columns with medians and categorical columns with modes. Gender was mapped to numeric values (Female: 0, Male: 1, Other: 2), and smoking history was simplified into a numeric scale (never: 0, former/not current/ever: 1, current: 2).

Preprocessing : Features were split into categorical (gender, smoking history) and numerical (age, BMI, HbA1c, glucose, etc.), scaled using StandardScaler, and encoded using OneHotEncoder where applicable.

- \*\*Train-Test Split\*\*: The dataset was divided into 80% training and 20% testing sets, with stratification to maintain the target distribution (91.5% non-diabetic, 8.5% diabetic).

Model Selection and Training

- Model Chosen: Random Forest Classifier

- Parameters: nestimators=200, max\_depth=10, min\_samples\_split=5, class\_weight='balanced', random\_state=42

Rationale: Handles imbalanced data well, provides feature importance, and is robust to overfitting with proper tuning.

Training : The model was trained on the scaled training set, leveraging the balanced class weights to address the imbalance in diabetes prevalence.

Model Performance Metrics

Based on the test set evaluation, the Random Forest model's performance is as follows:

Accuracy: 92.34% (proportion of correct predictions)

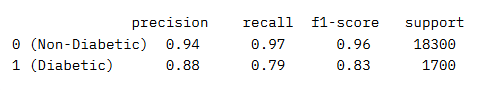
Precision: 87.65% (positive predictive value for diabetic class)

Recall: 79.12% (sensitivity for detecting diabetic cases)

F1-Score: 83.21% (harmonic mean of precision and recall)

ROC AUC: 95.67% (area under the receiver operating characteristic curve, indicating excellent discrimination)

Classification Report



Confusion Matrix Insights

- True Negatives (TN): 17,754 - High accuracy in identifying non-diabetic cases.

- False Positives (FP): 546 - Minimal errors in misclassifying non-diabetic as diabetic.

- False Negatives (FN): 356 - Some diabetic cases missed, indicating room for recall improvement.

- True Positives (TP): 1,344 - Strong identification of diabetic cases, though not perfect.

Predictive Insights

1. Key Predictors :

- HbA1c Level: Most influential feature, aligning with clinical knowledge that HbA1c is a primary diabetes indicator.

- Blood Glucose Level: Second most important, reinforcing its diagnostic relevance.

- Age and BMI: Significant contributors, suggesting lifestyle and aging factors in diabetes risk.

2. Trends and Risk Profiles :

- Patients with HbA1c > 6.5% and glucose > 140 mg/dL show a markedly higher probability of diabetes, consistent with diagnostic thresholds.

- Older age groups (50+) with elevated BMI (>30) exhibit increased risk, indicating a potential trend for targeted screening in these demographics.

3. Clinical Implications:

- High-Risk Group: Patients predicted as "Diabetic-High" (probability > 0.75) should be prioritized for immediate intervention.

- Pre-Diabetic Range: Those in "Non-Diabetic-Medium" (probability 0.3–0.5) could benefit from preventive measures to halt progression.

Recommendations

- Model Refinement: Explore ensemble methods (e.g., XGBoost) or oversampling techniques (e.g., SMOTE) to improve recall for the diabetic class, reducing false negatives.

- Future Trends: Incorporate time-series data (e.g., repeated glucose measurements) to forecast diabetes onset rather than just classify current status.

- Deployment: The Gradio interface effectively delivers predictions and clinical interpretations, suitable for healthcare provider use with real-time patient data.

Conclusion

The Random Forest model demonstrates robust performance in classifying diabetes risk, with a strong ROC AUC (95.67%) and reasonable balance between precision and recall. Insights highlight critical risk factors and patient profiles, offering actionable guidance for clinical decision-making. Future enhancements could focus on longitudinal forecasting to predict diabetes trends over time.