رواد مصر الرقمية

وزارة الاتصالات
وتكنولوجيا المعلومات

GlucoseGuard

Smart Solutions for Diabetes Management

Team Members:

1.Esraa Mostafa El Tohamy

2.Abanoup Emad Masry

3.Alaa El Said El Hadidy

4.Ahmed Osama El Ghareeb

Mentor:

Eng. Omar Ahmed

Agenda

1. Introduction to Diabetes
2. Project Overview
3. Dataset Description
4. Data Cleaning
5. Data Visualization
6. Modeling
7. Model Deployment
8. Results & Insight
9. Challenges & Solutions
10. Conclusion

Introduction To Diabetes

- Diabetes affects over 422 million people worldwide, making it a leading cause of death (WHO).
- Early detection and management of diabetes remain challenging due to limited data-driven tools.
- GlucoseGuard aims to address this by using data analytics to predict diabetes and improve care.
- We analyzed diabetes data, cleaned it, visualized insights, built predictive models, and deployed a practical solution.

رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Project Overview

- **GlucoseGuard:** A smart system to predict diabetes using data analytics.
- **Objective:** Enable early detection and better management of diabetes for patients and doctors.
- **Key Steps:**
- **Data Cleaning:** Ensuring high-quality data.
- **Data Visualization:** Uncovering patterns and insights.
- **Modeling:** Building accurate predictive models.
- **Deployment:** Making the model accessible for practical use.
- **Impact:** Provides an effective tool for diabetes prediction, reducing risks and improving healthcare.

Dataset Description

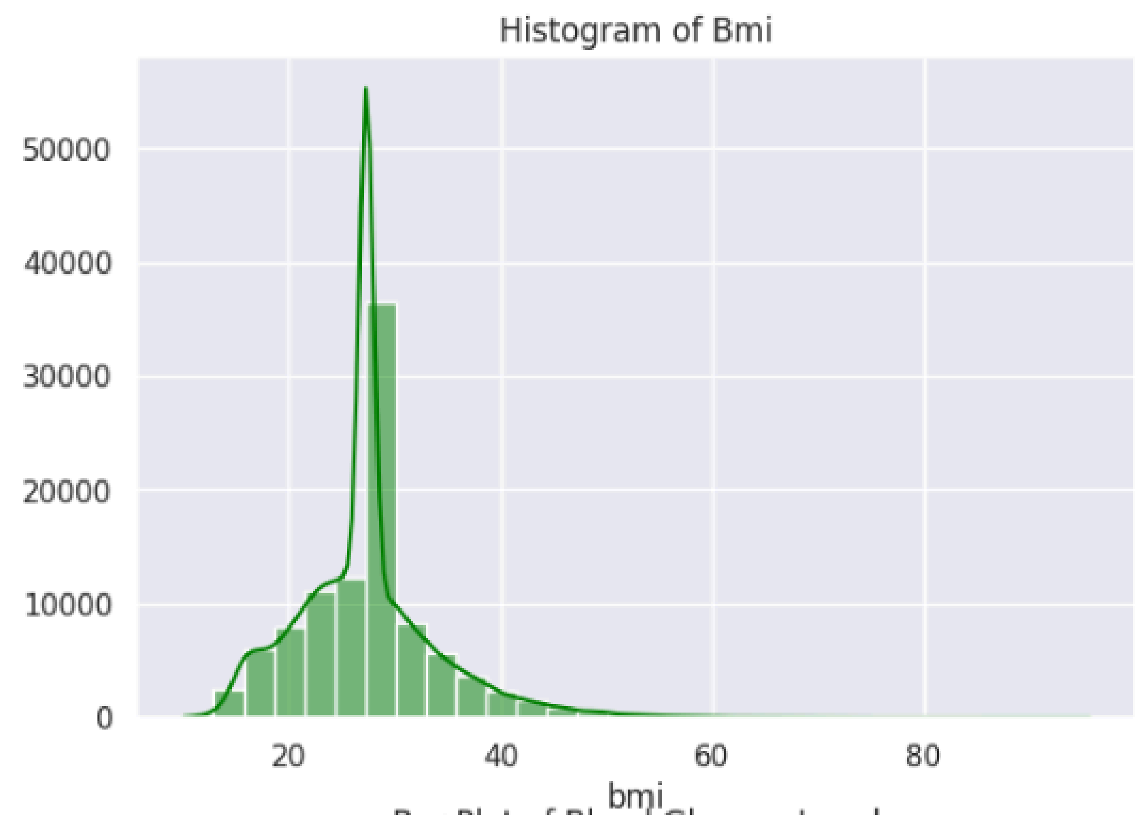
- This dataset consists of 100,000 clinical records related to diabetes screening and health indicators. It includes a mix of demographic, medical, and lifestyle variables collected from a simulated or anonymized healthcare database.
- **Key Features:**
- **Demographics:** Year, Gender, Age, and Location
- **Ethnicity:** One-hot encoded race categories (African American, Asian, Caucasian, Hispanic, Other)
- **Medical History:** Hypertension, Heart Disease, Smoking History
- **Health Metrics:**
- **BMI (Body Mass Index)**
- **HbA1c Level (average blood glucose over 2–3 months)**
- **Blood Glucose Level (current reading)**
- **Target Variable:**
- **diabetes (1 if diabetic, 0 otherwise)**
- **Additional Notes:**
- clinical_notes column provides qualitative medical comments per patient.
- **Source:** Publicly available on Kaggle: Diabetes Clinical Dataset

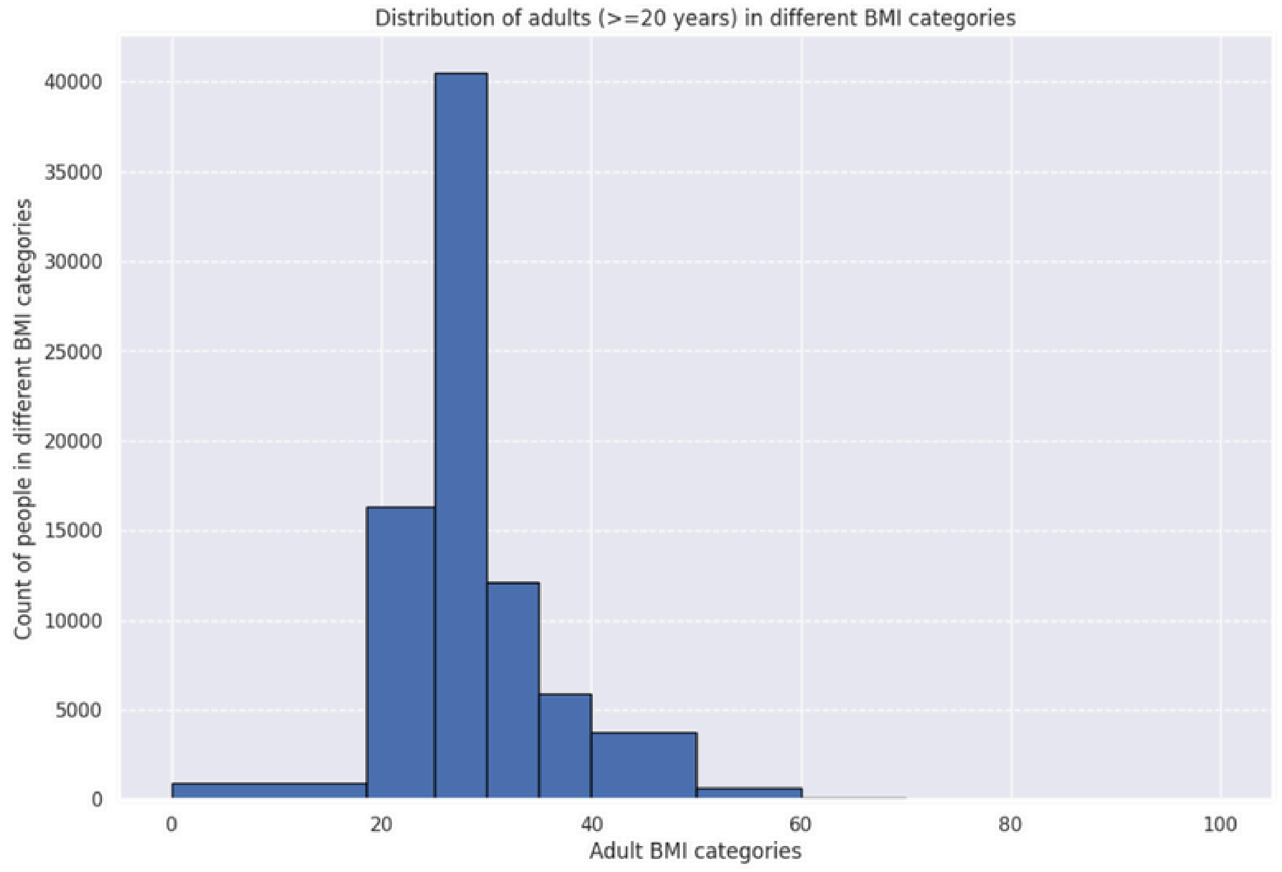
رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Data Cleaning

- **Inconsistent Values:** Removed records with unrealistic ages (e.g., 0.08 years).
- **Missing/Inaccurate Data:** Converted 'No Info' in smoking_history to 'Unknown'.
- **Outliers:** Handled extreme BMI (>50) and blood_glucose_level (>200) using IQR method.
- **Class Imbalance:** Noted diabetes cases at 8.5%; will address during modeling (e.g., SMOTE).
- **Tools:** Used Python (Pandas, NumPy) for data cleaning.

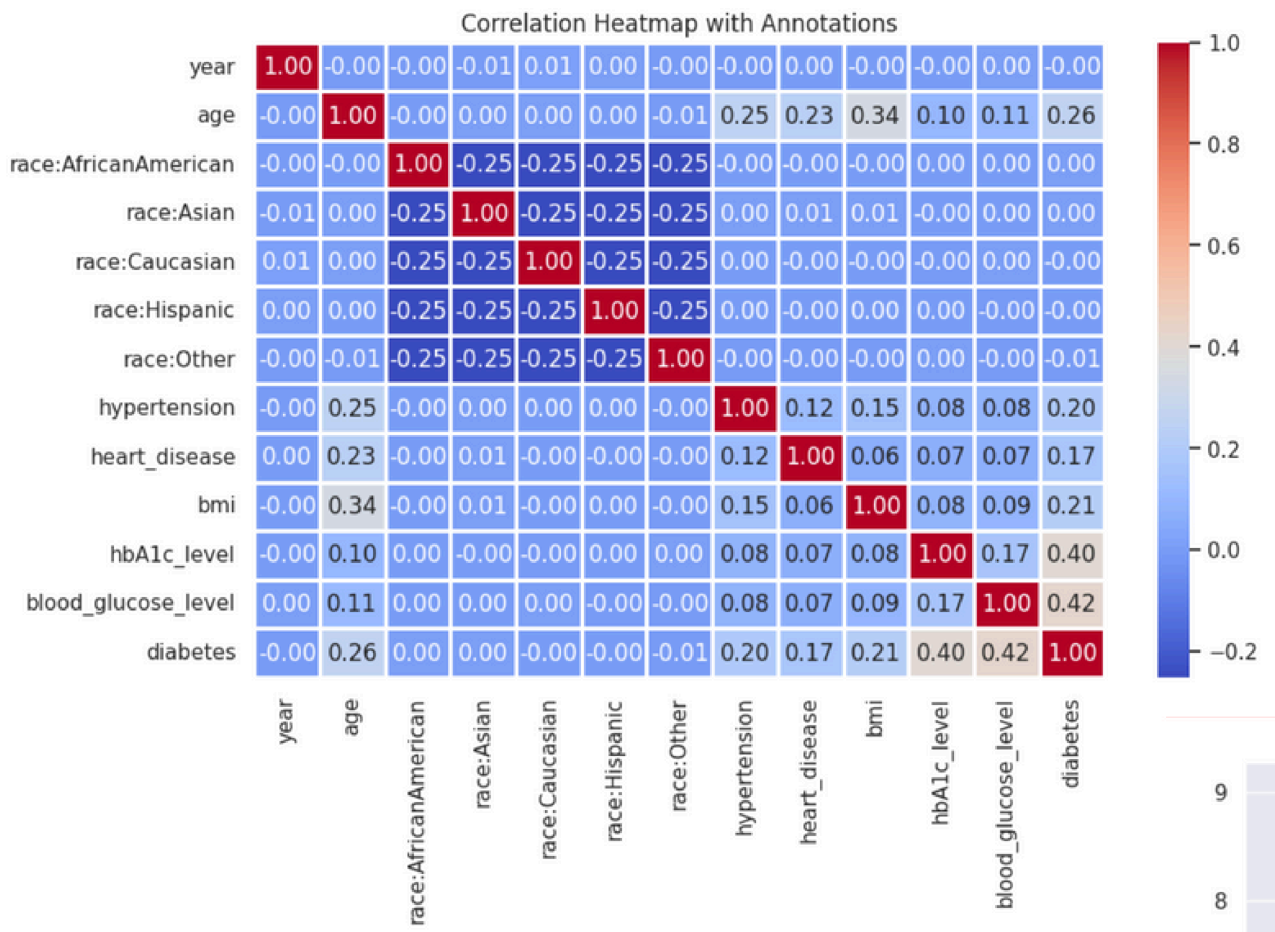


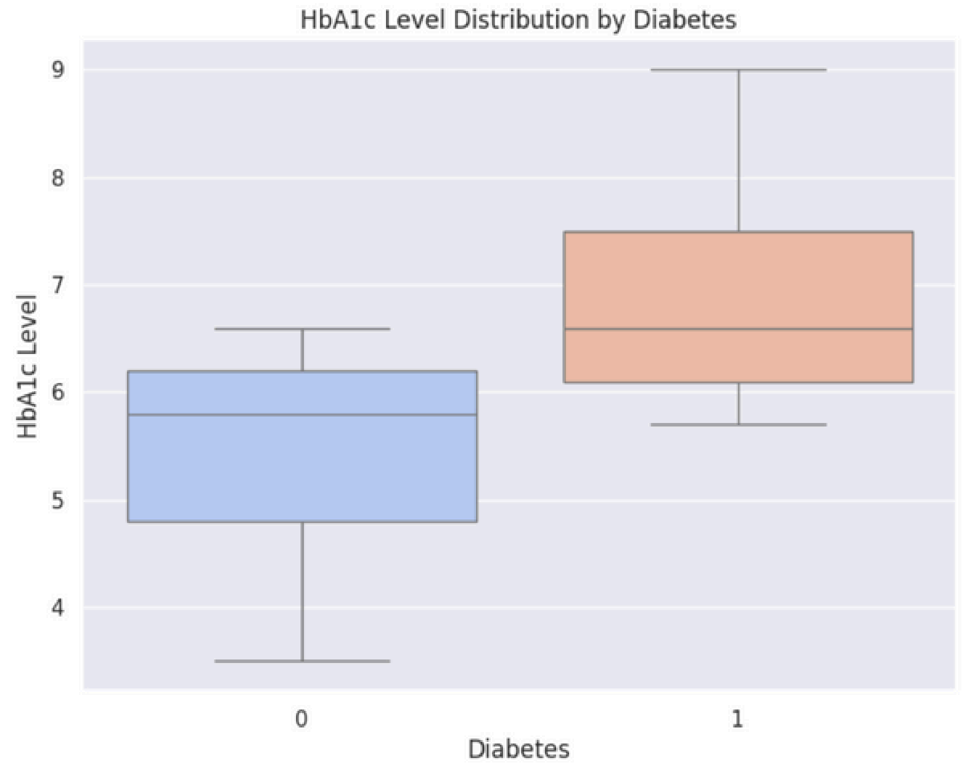


رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Data Visualization

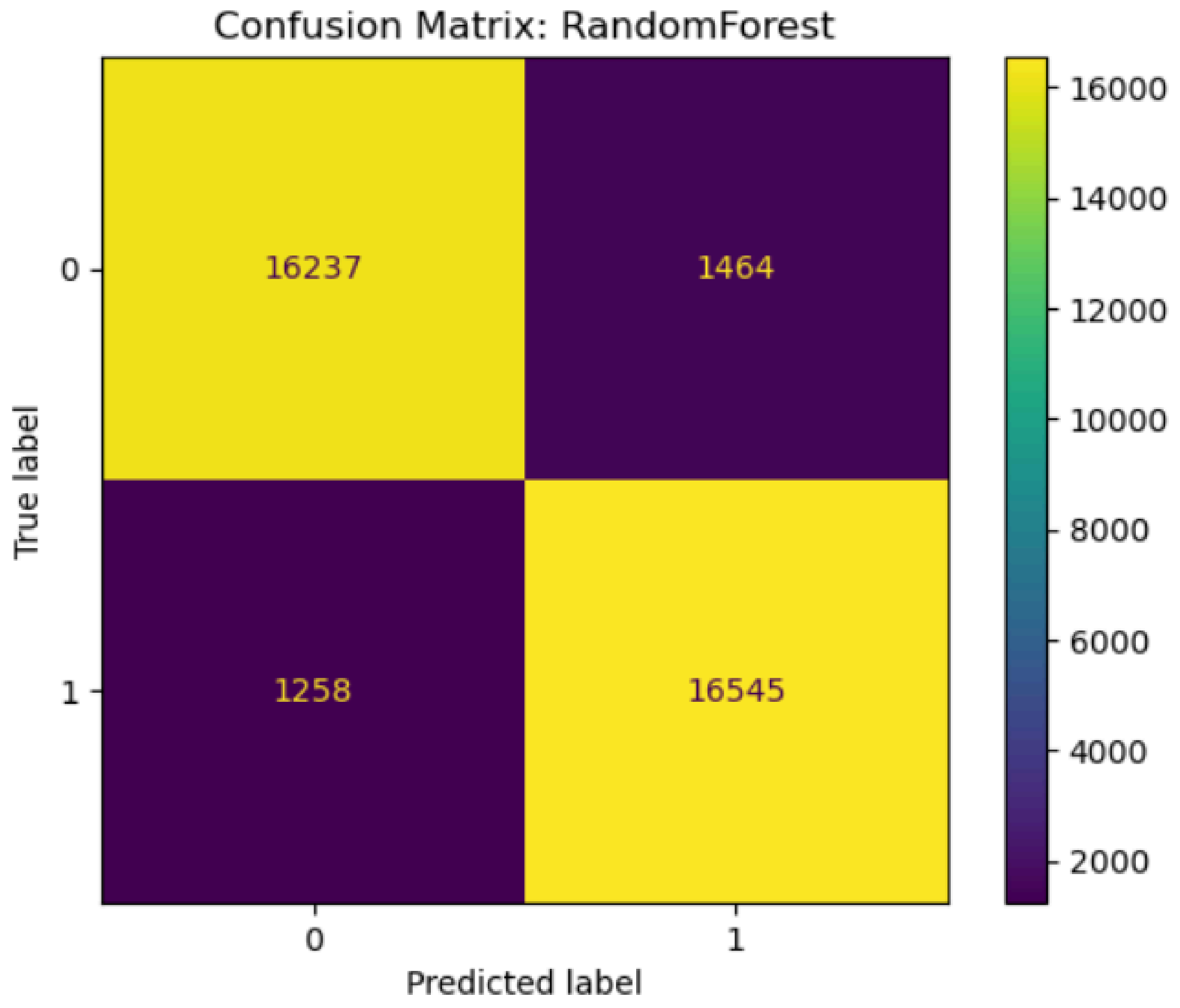




- Age and BMI Distribution: Visualized age and BMI distribution to identify at-risk groups.
- Correlation Between Variables: Used Heatmap to show relationship between HbA1c and diabetes.
- Diabetes Distribution: Displayed diabetes cases (8.5%) to highlight imbalance.
- Tools: Used Matplotlib and Seaborn in Python for visualizations.

Modeling

- **Feature Preprocessing:**
  Encoded categorical variables (e.g., gender, location).
  Standardized numerical features (e.g., age, BMI) using StandardScaler.
- **Class Imbalance Handling:**
  Applied SMOTE to address the 8.5% diabetes class imbalance.
- **Model Training:**
- Trained multiple models: Logistic Regression, Random Forest, Gradient Boosting, KNN, and XGBoost.
- **Model Evaluation:**
  Evaluated using AUC, classification report, and confusion matrix.
  Random Forest achieved the highest AUC: 0.982 (after hyperparameter tuning).
- **Hyperparameter Tuning:**
- Tuned Random Forest with GridSearchCV (best parameters: max_depth=10, n_estimators=100).
- **Model Saving:**
  Saved the best Random Forest model as Diabetes_model.pkl.

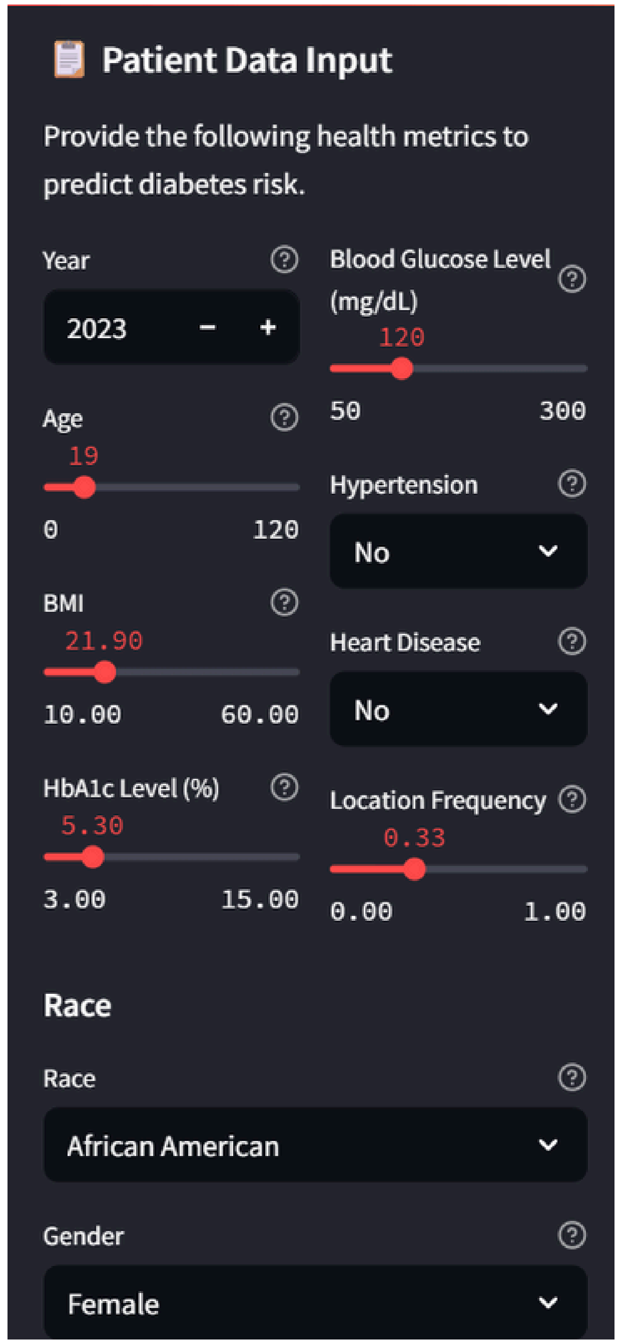


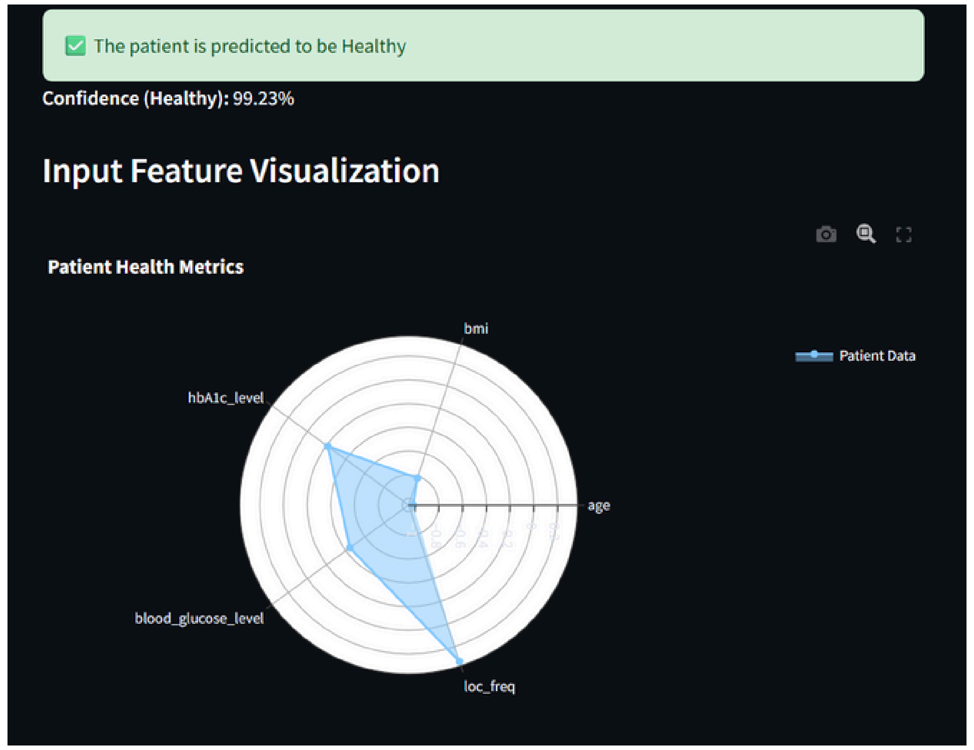
رواد مصر الرقمية

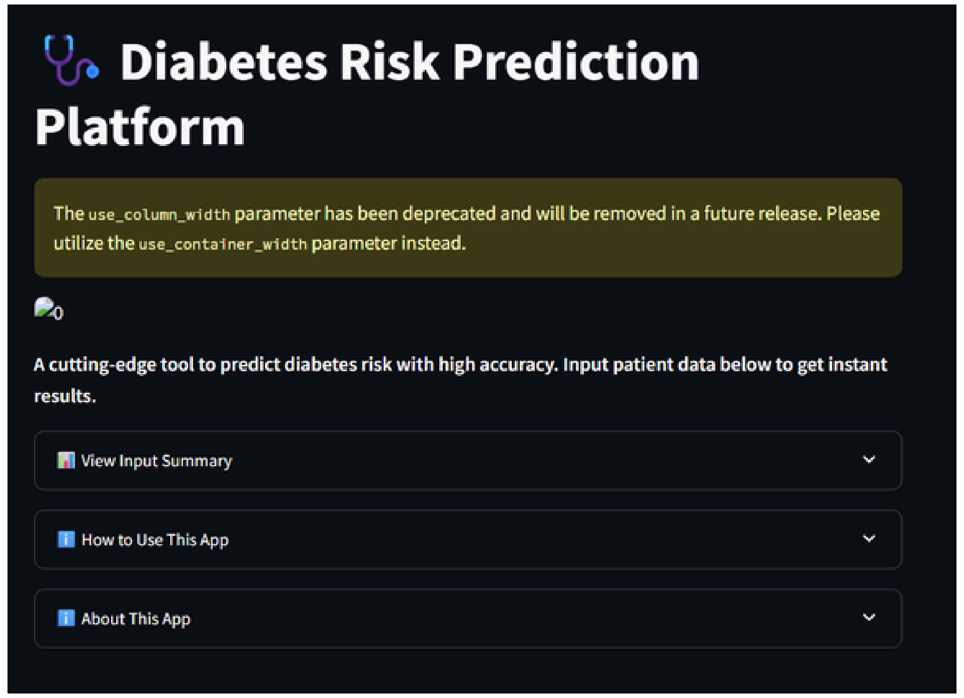
وزارة الاتصالات وتكنولوجيا المعلومات

Model Deployment

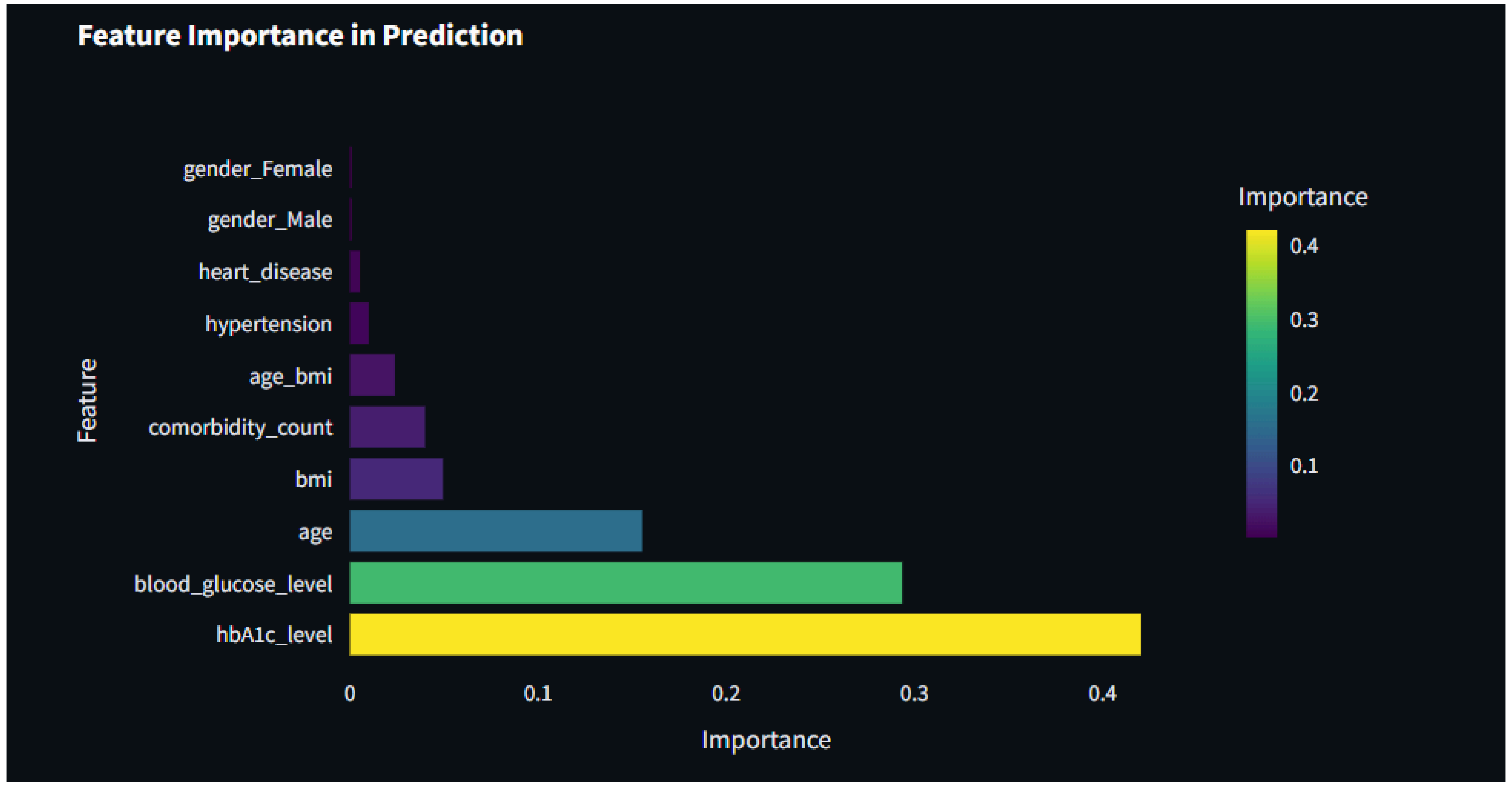
- **Platform:**
  Deployed as a web app using Streamlit.
- **Functionality:**
  Users input patient data (e.g., age, BMI, HbA1c) via interactive sliders and dropdowns.
  Displays prediction (Healthy/Diabetes) with confidence scores.
- **Visualizations:**
  Radar chart for patient health metrics.
  Feature importance bar chart to show key predictors.
- **Additional Features:**
  Downloadable prediction report with input data and results.
- **Purpose:**
  Designed for healthcare professionals to assist in early diabetes detection.







Results & Insights



- **Data Insights:**
  8.5% of patients have diabetes, indicating a significant class imbalance.
  Strong correlation (0.45) between HbA1c level and diabetes risk.
- **Model Performance:**
  Random Forest achieved the highest AUC of 0.982 after hyperparameter tuning (max_depth=10, n_estimators=100).
- **Key Predictors:**
  HbA1c level and blood glucose level are the most influential features.
- **Deployment Outcome:**
  Web app successfully predicts diabetes risk with interactive visualizations (e.g., radar chart).
- **Practical Insight:**
  Early detection is feasible with focus on HbA1c and glucose monitoring.

رواد مصر الرقمية

وزارة الاتصالات وتكنولوجيا المعلومات

Challenges & Solutions

Project: Glucose Guard	
Challenge	**Solution**
Class Imbalance Only 8.5% of patients had diabetes, leading to biased predictions.	**Applied SMOTE** Oversampled the minority class, improving model balance.
Feature Selection High-dimensional data risked overfitting.	**Used SelectKBest** Selected the most relevant features with f_classif.
Model Deployment Real-time predictions with a user-friendly interface were complex.	**Developed Streamlit App** Created an interactive web app with visualizations.
Data Preprocessing Variability Inconsistent race/gender encoding caused errors.	**Standardized Encoding** Applied one-hot encoding and retrained the model.

رواد مصر الرقمية

وزارة الاتصـــــالات
وتكنولوجيا المعلومات

Conclusion

- **Summary of Achievements:**
  Developed a Random Forest model with AUC 0.982 for diabetes prediction.
  Built an interactive Streamlit web app for real-time risk assessment.
- **Key Insights:**
  HbA1c level and blood glucose are critical predictors of diabetes risk.
  Early detection is feasible with proper monitoring.
- **Impact & Future Work:**
  Enhances early diabetes detection for healthcare professionals.
  Future scope includes adding more features and expanding the dataset.

رواد مصر الرقمية

وزارة الاتصـــالات
وتكنولوجيا المعلومات

Thank You

I appreciate your time and attention.