Final Project Report: Diabetes Predictive Modelling

This report summarizes the project dedicated to developing and preparing a machine learning model for predicting diabetes risk, outlining the process from data handling to deployment readiness.

---

1. Project Summary: Data, Model, and Deployment

The project aimed to build a robust predictive model for diabetes, utilizing a health dataset containing various clinical and demographic features.

Data Collection and Preprocessing

The raw dataset was initialized with 100,000 patient records. The preprocessing stage was critical for ensuring data quality and model readiness:

- **Invalid Data Removal:** 911 infant records (age < 1 year) were removed, as they were deemed illogical for an adult health analysis.
- **Missing Data Treatment:** A major challenge was the significant portion (approximately 36%) of 'No Info' values in the smoking_history feature. To retain these records and treat them as a distinct category, the 'No Info' values were kept as they were for subsequent one-hot encoding. This allowed the model to learn the predictive power of this missing/unknown status.
- **Feature Engineering and Normalization:** Categorical groupings for age and BMI were created to improve analysis clarity. Numerical features, including age, bmi, hbA1c_level, and blood_glucose_level, were normalized using StandardScaler for optimal model training.

Model Development and Deployment

The project involved selecting and training a classification model to predict the binary diabetes outcome.

- **Feature Selection:** Recursive Feature Elimination (RFE) with Logistic Regression was employed to identify and select the top 10 most influential predictors.
- **Model Tracking:** MLflow was used to manage the machine learning lifecycle, tracking key parameters and performance metrics such as accuracy.
- **Deployment Readiness:** For productionizing the solution, the final predictive model, the StandardScaler object, and the list of trained column names were persisted using pickle files. This setup ensures the model and its required preprocessing components can be easily loaded and deployed (likely via a platform like Databricks, as suggested by the file name).

2. Challenges and Key Insights

Challenges Faced

The primary challenges were centred on data cleaning and feature preparation:

- **Handling Illogical Ages:** Identifying and filtering out inaccurate age records was necessary to ensure the model learned from a clinically relevant adult population.
- **Managing High Missingness:** The substantial volume of 'No Info' values in the smoking_history posed a challenge. The decision was made to treat 'No Info' as a valid, distinct category rather than imputing, to avoid introducing an artificial bias and to allow the model to capture any predictive signal associated with the unknown status.
- **Outlier Treatment:** Outliers were observed in continuous variables like bmi (7.22%), hbA1c_level (1.32%), and blood_glucose_level (2.04%), which necessitated statistical treatment to prevent skewed model performance.

Key Insights from the Predictive Model

The Recursive Feature Elimination process provided valuable insights into the strongest predictors of diabetes risk:

- **Primary Metabolic Markers:** As expected, the model selected hbA1c_level and blood_glucose_level as top features, reinforcing their clinical importance in diagnosing and predicting diabetes.
- **Critical Risk Factors:** The model confirmed that established clinical risk factors—specifically age, the presence of hypertension, and heart_disease—are crucial independent predictors of diabetes.
- **Lifestyle Impact:** Certain categories of smoking_history (current, ever, and the No Info category) also emerged as top features, highlighting the statistical significance of lifestyle choices and the unknown status in predicting risk.

---

3. Recommendations for Healthcare Integration

The predictive model should be integrated into clinical workflows to support early detection and proactive patient management.

1. Real-time Risk Screening and Triage

- **System Integration:** Integrate the model's prediction pipeline directly into the Electronic Health Record (EHR) system. The model should automatically process new patient data upon entry and generate a diabetes risk score at the point of care.
- **Flagging High-Risk Patients:** Utilize the model as a triage tool during routine screenings. Any patient with a risk score above a predefined threshold should be

immediately flagged for priority follow-up, independent of the physician's initial assessment.

2. Personalized Preventative Care

- **Targeted Interventions:** Leverage the model's feature importance—especially the strong influence of age, bmi, and smoking_history—to recommend personalized preventative measures. For instance, a high-risk patient with a high BMI should receive immediate counselling on weight management and diet.
- **Education and Monitoring:** Use the model's output to initiate educational material focused on the patient's specific risk factors. Regular monitoring schedules should be automatically generated for high-risk individuals.

3. Optimized Diagnostic Pathways

- **Just-in-Time Diagnostics:** The model's prediction can justify the immediate scheduling of comprehensive follow-up tests, such as a Glucose Tolerance Test or a secondary blood panel, even if initial symptoms are mild. This optimizes the diagnostic pathway by focusing limited clinical resources on the patients most likely to benefit from early intervention.
- **Data-Driven Review:** Encourage healthcare professionals to use the model's output as a clinical decision support tool, providing a data-driven "second opinion" to confirm or challenge an initial diagnosis based on the patient's holistic risk profile.