**AI-Powered Wellness Advisor for Proactive Diabetes Risk Prediction and Lifestyle Management**

**A Project Report**

**Abstract**

The increasing prevalence of Type 2 diabetes worldwide necessitates proactive and accessible health management tools. This project addresses this need by developing an AI-powered web application, the "Wellness Advisor," designed to provide individuals with an instantaneous diabetes risk assessment and personalized lifestyle guidance. The system leverages a Random Forest machine learning model trained on a comprehensive health dataset of 100,000 records. Key data preprocessing steps, including one-hot encoding for categorical features and standardization for numerical features, were implemented. Model optimization was performed using GridSearchCV to achieve the highest possible accuracy. The final model was integrated into a user-friendly interface built with Streamlit. Evaluation of the model on unseen test data demonstrated an excellent overall accuracy of 97%. However, a detailed analysis revealed a critical performance insight: while the model exhibits perfect precision in identifying diabetic patients, its recall for this class is 67%, indicating a risk of false negatives. This report details the project's architecture, implementation, results, and discusses the implications of its performance for real-world application.

**1. Introduction**

**1.1. Background**

Type 2 diabetes is a chronic metabolic disorder that poses a significant global health challenge. Its gradual onset means many individuals are unaware of their condition until complications arise. Early detection and proactive lifestyle modifications are the most effective strategies for preventing and managing the disease. However, access to regular medical screenings and personalized dietary advice is often limited by cost and availability, creating a gap between public health goals and individual action.

**1.2. Problem Statement**

A significant portion of the at-risk population lacks accessible tools for immediate, preliminary risk assessment. Furthermore, generic health advice is often ineffective as it is not tailored to an individual's specific health profile. This project aims to solve this by creating a data-driven tool that bridges the gap between clinical data and personal, actionable health guidance.

**1.3. Project Objectives**

The primary objectives of this project were to:

1. **Develop a High-Accuracy Predictive Model:** To train, evaluate, and optimize a machine learning model capable of accurately predicting the risk of diabetes based on key health indicators.
2. **Create an Intuitive User Interface:** To build an interactive and user-friendly web application where non-technical users can easily input their data and understand the results.
3. **Provide Personalized and Actionable Guidance:** To generate tailored lifestyle and nutritional recommendations based on the user's predicted risk level.
4. **Empower Users:** To provide a tool that encourages proactive health management and raises awareness of personal risk factors.

**2. System Architecture and Methodology**

The project was executed in two distinct phases: an offline model training phase and an online real-time prediction phase, as illustrated in the system block diagram below.

![alt text](https://storage.googleapis.com/aai-web-images/assessed-knowledge/block\_diagram\_diabetes\_report\_v2.png)

**2.1. Dataset**

The model was trained on the diabetes\_prediction\_dataset.csv, a large-scale dataset containing 100,000 records. The key features include age, gender, BMI, hypertension, heart\_disease, smoking\_history, HbA1c\_level, blood\_glucose\_level, and the target variable, diabetes.

**2.2. Tools and Technologies**

* **Programming Language:** Python
* **Data Manipulation:** Pandas, NumPy
* **Machine Learning:** Scikit-learn
* **Web Framework:** Streamlit
* **Data Visualization:** Plotly
* **Model Persistence:** Joblib / Pickle

**3. Implementation Details**

**3.1. Data Preprocessing**

To prepare the data for the model, a rigorous preprocessing pipeline was established:

1. **Categorical Feature Encoding:** Text-based features (gender, smoking\_history) were converted into a numerical format using one-hot encoding (pd.get\_dummies). This creates binary columns for each category, making the data suitable for the algorithm.
2. **Feature Scaling:** All numerical input features were standardized using StandardScaler. This process scales the data to have a mean of 0 and a standard deviation of 1, preventing features with larger ranges (like blood\_glucose\_level) from disproportionately influencing the model.

**3.2. Model Training and Optimization**

The **Random Forest Classifier** was chosen for its high accuracy, robustness against overfitting, and its ability to handle complex, non-linear relationships in the data.

To achieve maximum performance, the model's hyperparameters were optimized using **GridSearchCV**. This technique systematically tested numerous combinations of parameters (e.g., n\_estimators, max\_depth) using 3-fold cross-validation to identify the configuration that yielded the highest accuracy. The best-performing model was then trained on the full training dataset.

**3.3. Web Application Development**

The front-end was developed using Streamlit. The application features a two-tab interface:

1. **Risk Assessment:** A user-friendly form for inputting health data. Upon submission, the app uses the saved model and scaler to generate a real-time prediction, which is displayed using a Plotly gauge chart.
2. **Action Plan:** Based on the predicted risk level, this tab provides personalized lifestyle advice and includes a nutrition logger for tracking daily sugar intake.

**4. Results and Evaluation**

The performance of the final, optimized model was evaluated on a test set of 20,000 unseen samples. The model achieved an outstanding overall **accuracy of 97.05%**.

The detailed Classification Report provides deeper insights:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |
| **0 (No Diabetes)** | 0.97 | 1.00 | 0.99 | 18,300 |
| **1 (Has Diabetes)** | 1.00 | 0.67 | 0.81 | 1,700 |
| **Weighted Avg** | 0.97 | 0.97 | 0.97 | 20,000 |

**4.1. Analysis of Results**

* **High Precision:** The model demonstrates exceptional precision. Notably, for the positive class (1), the precision is a perfect 1.00. This means that **every single individual the model identified as having diabetes was indeed diabetic**. This is a critical strength, as it ensures the application does not generate false alarms.
* **Excellent Performance on Negative Class:** The model correctly identified 100% of the non-diabetic individuals (Recall of 1.00 for class 0), meaning it gives a reliable "all-clear" to healthy users.
* **Key Limitation - Recall for Positive Class:** The model's primary weakness is its **recall of 0.67 for the diabetic class**. This indicates that the model **failed to identify 33% of the individuals who actually have diabetes** in the test set. These "false negatives" represent a significant risk in a real-world medical context, as these users would be incorrectly assured that they are low-risk.

**5. Conclusion**

This project successfully achieved its goal of developing a high-accuracy, AI-powered prototype for diabetes risk prediction. The Wellness Advisor application serves as a powerful proof-of-concept, demonstrating that machine learning can make personal health insights more accessible. The final model's **97% accuracy** is a testament to the effectiveness of the chosen methodology.

However, the most critical outcome of this project is the insight gained from the detailed evaluation. While highly accurate, the model's tendency to miss a third of diabetic cases (low recall) highlights the challenge of working with imbalanced medical datasets. This finding is crucial for guiding the responsible and ethical deployment of such a tool. The application is a valuable educational and awareness tool, but its limitation regarding false negatives must be clearly communicated.

**6. Future Work**

To build upon this project's success and address its primary limitation, the following steps are recommended:

* **Improve Model Recall:** Implement techniques specifically designed for imbalanced datasets, such as using the class\_weight='balanced' parameter in the model, or applying advanced data sampling methods like SMOTE (Synthetic Minority Over-sampling Technique).
* **Enhance Nutrition Module:** Replace the static dictionary with a more sophisticated regression model to predict the nutritional content of a wider variety of user-inputted food items.
* **Develop User Profiles:** Add functionality for user accounts to allow individuals to track their risk score, health metrics, and food logs over time.
* **Deployment:** Deploy the application to a cloud service (e.g., Streamlit Cloud, Azure) for broader accessibility and public use.