Technical Report

Real-Time AI-Powered Health Coaching for Diabetes: Prototype

Team Name: Data Divers (Team 1)

Team Members: Ashna Ali, Sreevardhan Reddy Soma, Venkata Sai Veeramalla, Sai Kiran Basetty

10/21/2024

**Project Title:** Real-Time AI-Powered Health Coaching for Diabetes

**Problem Statement**

Diabetes is a widespread chronic condition, with its forms—Type 2 diabetes, prediabetes, and gestational diabetes—posing significant health risks if not detected and managed early. Accurate prediction of diabetes risk remains a challenge due to the complexity of contributing factors, including demographic, biometric, and lifestyle variables, as well as real-time physiological data such as glucose levels. Existing prediction models often rely on structured health data but fail to incorporate continuous glucose monitoring (CGM) data, which can provide critical insights into real-time glucose fluctuations. Additionally, imbalanced datasets, particularly with underrepresented classes like gestational diabetes and prediabetes, compromise the reliability of predictive models.

This study aims to address these challenges by developing an advanced machine learning model that integrates both structured data (e.g., age, BMI, blood pressure) and time-series CGM data to improve the prediction accuracy of diabetes risk. The project leverages techniques such as Synthetic Minority Over-sampling (SMOTE) to handle data imbalance and Recursive Feature Elimination (RFE) to select the most relevant predictors. To ensure interpretability, Local Interpretable Model-agnostic Explanations (LIME) are employed to provide transparency into the model’s predictions. The model is deployed through a user-friendly Streamlit interface, allowing real-time interaction and prediction, with results stored in a secure MongoDB repository.

This research seeks to enhance early detection of diabetes and its subtypes by addressing key limitations in current predictive methodologies, thus contributing to more effective disease management and intervention.

**Objectives**

1. Develop a machine learning model capable of predicting the likelihood of prediabetes, Type 2 diabetes, or gestational diabetes using a combination of user demographic data, biometric measurements, and continuous glucose monitoring (CGM) time-series data.
2. Gather and preprocess a comprehensive dataset that includes demographic information, lifestyle factors, biometric measurements, and CGM data relevant to diabetes prediction (i.e., age, weight, glucose levels, blood pressure, gender, sleep patterns, etc.). Data preprocessing includes handling missing values, outliers, and applying normalization and encoding techniques.
3. Train, fine-tune, and validate the model using advanced machine learning techniques, including transfer learning and Recursive Feature Elimination (RFE), to improve accuracy, sensitivity, and specificity while addressing data imbalance through techniques such as SMOTE.
4. Integrate the model into a user-friendly Streamlit-based web application, allowing real-time user interaction and prediction submissions. The interface should be easily accessible by healthcare providers or individuals.
5. Implement a secure MongoDB database for storing user input, CGM data, and model predictions, enabling real-time updates and retrieval.
6. Evaluate the model’s performance against existing diagnostic methods, ensuring the model’s effectiveness, interpretability (via LIME), and reliability in predicting diabetes-related outcomes.

**Technical Details**

We employed a machine learning approach using a comprehensive dataset that incorporated user demographics and biometric data. The main features include age, BMI, sleep patterns, gender, and blood pressure levels. Preprocessing involved cleaning the data by filling missing values, handling outliers, and applying normalization and encoding techniques. The numerical features (age, BMI, blood pressure, etc.) were scaled using the standard scaler, and the categorical features (gender, family history of diabetes, high blood pressure diagnosis, etc.) were one-hot encoded.

In a significant development, we created two distinct models—one for female patients and one for male patients. Each model features an XGBoost model for structured data and a deep learning model trained on CGM data. These models are then combined in an ensemble to provide predictions for both male and female patients. For the XGBoost models, we applied SMOTE for the training data, while class weights were used for the deep learning models to handle data imbalance effectively.

1. **Model Building Methodology**

a) **Implementation of Dual Model:** In the current project, the application interface allows both male and female users to predict their diabetes risk by selecting their gender from a dropdown menu. This selection customizes the user experience by displaying input fields relevant to each gender, ensuring that data collection is personalized. Presently, a single unified model is used for predicting diabetes risk for both males and females, processing inputs uniformly despite gender-specific differences. However, our future approach aims to enhance prediction accuracy and personalization by implementing two distinct models: one specifically fine-tuned for male-related characteristics and another for female-specific factors.

b) **Implementation of Male Model:**

A Machine Learning model XGBoost model is used for male dataset with selected male features which is structured data that can fetch a high accuracy and efficiency in predicting diabetes. The Model tackles the class imbalance effectively with SMOTE.

c)**Implementation of Female Model:**

A Machine Learning model “XGBoost” model is used for Female dataset with selected female features which is also a structured data. By using techniques like SMOTE for class imbalance makes the model accuracy high and efficient one.

**Benefits:** The benefits of using two models are multifold. Gender-specific models enable more accurate and relevant predictions by focusing on the most impactful features for each gender, reducing noise and improving overall performance.

The selected features for each gender are as follows:

* **Male Selected Features:** ['Age', 'HighBP', 'PhysicallyActive', 'BMI', 'Smoking', 'Sleep', 'SoundSleep', 'RegularMedicine', 'Stress', 'BPLevel']
* **Female Selected Features:** ['Age', 'HighBP', 'PhysicallyActive', 'BMI', 'Sleep', 'SoundSleep', 'BPLevel', 'Pregnancies', 'Gestation in previous Pregnancy', 'PCOS']

The best parameters for the **Male** XGBoost model are:

* **Best Parameters:** {'colsample\_bytree': 1.0, 'learning\_rate': 0.3, 'max\_depth': 6, 'n\_estimators': 100, 'subsample': 1}
* **Accuracy for Male XGBoost Model:** 0.93

The best parameters for the **Female** XGBoost model are:

* **Best Parameters:** {'colsample\_bytree': 1.0, 'learning\_rate': 0.2, 'max\_depth': 7, 'n\_estimators': 300, 'subsample': 0.8}
* **Accuracy for Female XGBoost Model:** 0.8249960980177931

The structured data and CGM data were imputed for both male and female models to ensure that the patient samples align with each other.

The model itself was trained on several datasets, some of which were sourced from Kaggle and GitHub repositories. We have uploaded those datasets as a binary file release on our team GitHub repository. Our model was initially trained on a dataset that had mainly health indicators such as high blood pressure, high cholesterol, physical activity, etc. The target variable was for diagnosing Type 2 diabetes, prediabetes, or no diabetes. To include more samples of women diagnosed with gestational diabetes, we fine-tuned and transfer-learned on the PIMA Indians Dataset (labeled as gestational\_diabetes\_dataset on GitHub). We used the same technique to gather more samples of prediabetes and Type 2 diabetes with other datasets from additional sources.

Finally, for our time-series data, we fine-tuned and transfer-learned on a dataset found on the GitHub repository Awesome-CGM. This dataset contains continuous glucose monitoring data. Since this dataset is primarily used to classify patients with and without Type 2 diabetes, the other two classes of prediabetes and gestational diabetes were omitted. To train the model, we divided it into two sections: one section deals with the structured data of health indicators and symptoms mentioned before, and the other part of the model handles the time-series data. The model topology is shown below:

1. **Final Model Topology**

* **Structured Data Model (Sequential 6)**
* **Time Series Model (Sequential 44)**

The model was validated using cross-validation and hyperparameter tuning techniques, achieving high accuracy in predicting diabetes risk categories (prediabetes, Type 2 diabetes, or gestational diabetes). Model performance was evaluated using metrics such as ROC-AUC, F1-score, and confusion matrices.

1. **Triangle Model: Frontend, Data Repository, and AI/ML Component**
2. The implementation of the project can be broken down into three key components:
3. **Frontend:** A web-based user interface was developed using the Streamlit framework, providing an accessible interface for users to input their health data. Users can view the prediction after submitting their health data through a survey question-based frontend. The application features two distinct views, tailored for female and male users, with questions varying based on the selected features for each group:
4. **Female View:**
   1. Selected Features: Age, HighBP, PhysicallyActive, BMI, Sleep, SoundSleep, BPLevel, Pregnancies, Gestation in previous Pregnancy, PCOS.
   2. Questions include prompts like:
      1. "How many pregnancies have you had in total?"
      2. "What was the gestation in your previous pregnancy?"
      3. "Do you have Polycystic Ovary Syndrome (PCOS)?"
5. **Male View:**
   1. Selected Features: Age, HighBP, PhysicallyActive, BMI, Smoking, Sleep, SoundSleep, RegularMedicine, Stress, BPLevel.
   2. Questions include prompts like:
      1. "Do you smoke?"
      2. "Are you on regular medication for diabetes?"
      3. "What is your current stress level?"
6. Other common questions across both views include:
7. "Enter your height in inches."
8. "Enter your weight in pounds."
9. "How many hours of sleep were you staying still in bed (not interrupted sleep)?"
10. "How many hours of sleep per day do you get on average?"
11. "Enter your age."
12. "On average, how physically active are you per day?"
13. "What is your blood pressure level?"
14. "Are you diagnosed with high blood pressure?"
15. "Enter your CGM (Continuous Glucose Monitoring) data."
16. After the user clicks submit, the predictions are calculated and sent to the MongoDB database directly. Streamlit offers a “no API necessary” structure to send data to the database.
17. **Data Repository:** The dataset, hosted on MongoDB, includes user demographics, lifestyle information, and clinical metrics. This repository is securely integrated with the AI component, allowing real-time data access and updates. The predictions are sent to the dataset along with the user input. The database entry is represented as a JSON document, which includes various attributes related to input values, continuous glucose monitoring (CGM) data, prediction probabilities, and a diagnosis. The key components of this structure are:
18. **\_id:** A unique identifier for the database entry. It uses MongoDB’s ObjectId format, where each entry is assigned a unique identifier ($oid), e.g., "66edc8d89d71445bfcb518a4".
19. **input\_values:** A dictionary that contains the input values used for prediction. Each input is stored in either an integer ($numberInt) or floating-point ($numberDouble) format. The input values include:
    1. **pregnancies:** Number of pregnancies, stored as an integer.
    2. **bmi:** Body Mass Index (BMI), stored as a double.
    3. **sound\_sleep:** Hours of sound sleep, stored as a double.
    4. **sleep:** Total sleep duration in hours, stored as a double.
    5. **gender\_male** and **gender\_female:** Binary indicators (1 or 0) representing the gender of the individual.
    6. **age:** The age of the individual, stored as an integer.
    7. **physically\_active:** Binary indicators of physical activity levels.
    8. **bp\_level:** Blood pressure level, stored as an integer.
    9. **high\_bp\_diagnosis:** A binary indicator representing high blood pressure diagnosis (1 for yes, 0 for no).
    10. **Pcos:** diagnosis of PCOS, (1 for yes, 0 for no)
    11. **cgm\_data:** A list of dictionaries, where each dictionary represents a CGM data entry. Each entry has a timestamp (stored in ISO format) and the glucose level at that timestamp. The glucose levels are stored as double values.
    12. **prediction\_probability:** The predicted probability of each diabetes risk category (prediabetes, Type 2 diabetes, gestational diabetes), stored as an array of doubles.
    13. **diagnosis:** A string indicating the predicted diagnosis based on the probabilities, such as "Prediabetes" or "Type 2 Diabetes".
20. **AI/ML Component:** The AI component comprises the trained machine learning models, which process user input and generate predictions. Both the structured and CGM data models are used to make predictions. After prediction, the results are returned to the frontend for user display and stored in the database.

**Challenges and Solutions**

1. **Data Imbalance:** One of the major challenges was handling class imbalance in the dataset, where most of the datasets represented non-diabetic individuals. This was addressed by applying techniques such as Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic data for underrepresented classes, improving model performance.
2. **Feature Selection:** Selecting the most relevant features for the model was crucial. By applying Recursive Feature Elimination (RFE), we narrowed down the features to the top 15 most significant factors affecting diabetes risk, such as age, BMI, and family history.
3. **Model Interpretability:** The deep learning model’s black-box nature posed challenges in explaining its predictions. To address this, we used LIME (Local Interpretable Model-agnostic Explanations) to generate interpretable results, making the model’s predictions more transparent and understandable to end users.

**Next Steps for Model and Application Enhancements**

**1. Integration of CGM Data into the Model:**

Although our current CGM-based models demonstrate exceptional accuracy, achieving nearly **99%**, they are not yet included in the present application. This is because the existing models can only predict two categories: **No Diabetes or Type-2 Diabetes**. However, the project requires a more comprehensive classification:

-Three categories for male users

- Four categories for female users

Future updates will focus on expanding the model to handle these additional categories for both genders and integrating the enhanced CGM-based predictions.

**2. Implementing LLMs for Enhanced Recommendations:**

We attempted to incorporate **Large Language Models (LLMs)** into the current version to provide personalized recommendations. However, due to runtime errors and technical challenges, we temporarily shifted to standard recommendation logic based on user data. In the next release, we will revisit this implementation to ensure LLM-powered recommendations are functional and improve the user experience.

**3.Enhanced UI for Gender-Specific Interaction:**

The current version introduces dual models for male and female users, with a basic user interface for navigation between them. In future updates, we plan to create gender-specific UIs that are more engaging and tailored to user needs. This improvement aims to enhance the interactivity and user experience for each gender, offering a more intuitive and personalized interface.

**4.User Accounts for Personalized Recommendations and Guidance:**

Currently, the application supports diabetes prediction and classification based on the type of diabetes. To make the platform more engaging, the next update will introduce user account functionality. This feature will enable users to create individual profiles, allowing them to:

Receive personalized recommendations based on their health data.

Access tailored guidance and tips for better diabetes management.

This enhancement aims to make the application more interactive, user-focused, and supportive for long-term health management.

**Conclusion**

This project aims to revolutionize the early detection and management of diabetes by leveraging machine learning techniques, structured health data, and continuous glucose monitoring data. The deployment of dual models for males and females ensures that the application is tailored to the specific health profiles of users, enhancing prediction accuracy and relevance. By integrating advanced modeling techniques, we aspire to provide users with reliable and actionable insights, ultimately contributing to better health outcomes in the fight against diabetes.