Statement of Work

By Ashna Ali, Sreevardhan Reddy Soma, Venkata Sai Veeramalla, Sai Kiran Basetty

09/03/2024

SOW #001

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

**Team Members:**

* Ashna Ali, Project Lead
  + Role: Back-end for application, preprocessing data
  + GitHub Link: <https://github.com/aa57c>
* Venkata Sai Veeramalla
  + Role: Preprocessing data and analysis, model building
  + GitHub Link: <https://github.com/99vvs>
* Sai Kiran
  + Role: model training, UI framework (frontend)
  + GitHub Link: <https://github.com/BasettySaikiran>.
* Sreevardhan Reddy Soma
  + Role: Model training, UI framework (frontend)
  + GitHub Link: <https://github.com/vardhan-soma>

**Motivation and Significance**

The prevalence of diabetes, particularly Type-2 diabetes, pre-diabetes, and gestational diabetes, has seen a significant rise globally, affecting millions of individuals. Early detection and diagnosis of these conditions are crucial to managing and preventing serious complications such as cardiovascular diseases, neuropathy, and kidney failure. Despite advancements in medical diagnostics, many individuals remain undiagnosed or are diagnosed too late, leading to preventable health deterioration. This project aims to develop an AI-driven model that can predict the likelihood of an individual having pre-diabetes, Type-2 diabetes, or gestational diabetes using easily accessible user data, including demographics, lifestyle factors, and biometric measurements. It will also be able to give the appropriate guidance based on that prediction on how to improve the user’s health. The impact of solving this problem could lead to earlier interventions, improved patient outcomes, and reduced healthcare costs, contributing significantly to public health.

**Objectives**

The primary objectives of this project are as follows:

1. Develop a machine learning model capable of predicting the likelihood of pre-diabetes type 2 diabetes, or gestational diabetes based on user data.
2. Gather and preprocess a comprehensive dataset that includes demographic information, lifestyle factors, and biometric measurements relevant to diabetes prediction (i.e. age, ethnicity or demographic, weight, glucose level, gender, etc)
3. Train and validate the model using state-of-the-art machine learning techniques, ensuring high accuracy, sensitivity, and specificity.
4. Implement the model into a user-friendly application that can be easily accessed by healthcare providers or directly by individuals.
5. Evaluate the model's performance against existing diagnostic methods to determine its effectiveness and reliability.

**Existing Works**

Several research articles and studies have explored the use of machine learning for diabetes prediction. Some notable works include:

1. “Machine Learning and Deep Learning Predictive Models for Type 2 Diabetes: A Systematic Review” (Fregoso-Aparicio et al., 2021): provides a comprehensive review of the application of machine learning (ML) and deep learning (DL) models in predicting type 2 diabetes. The review addresses key challenges in the field, such as the heterogeneity of techniques and the lack of transparency in feature selection, which complicates model interpretability. The authors systematically analyze 90 studies, comparing 18 different ML and DL models, with tree-based algorithms like Random Forests emerging as top performers in terms of accuracy and interpretability. The review highlights that while Deep Neural Networks (DNNs) can handle large and complex datasets, they often fall short in terms of transparency and require further refinement for clinical applicability. The paper concludes by identifying opportunities for future research, particularly in improving the transparency and interpretability of predictive models, and in optimizing the balance between model complexity and practical utility in clinical settings. This review is a valuable resource for researchers aiming to develop more effective and transparent predictive models for type 2 diabetes.
2. “Machine and deep learning techniques for the prediction of diabetics: a review” (Kumar & Vijay Kumar Jha, 2024): provides a comprehensive overview of the application of machine learning (ML) and deep learning (DL) in the prediction of diabetes. The review covers the increasing global prevalence of diabetes and the urgent need for early diagnosis to mitigate its severe health impacts. The authors explore various ML and DL methodologies used over the past two decades, including supervised, unsupervised, and semi-supervised learning, emphasizing their roles in identifying risk factors and improving prediction accuracy. The paper discusses key challenges, such as class imbalance, model interpretability, and data privacy, which remain obstacles in clinical implementation. Through a detailed examination of techniques like Support Vector Machines (SVM), Random Forest, Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), the review highlights the successes and limitations of existing models, providing insights into future research directions. This article serves as a valuable resource for researchers and practitioners seeking to enhance diabetes prediction models using advanced computational techniques.
3. “Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers” (Hasan et al., 2020): Addresses the challenge of predicting diabetes using machine learning techniques. The authors propose a robust framework that incorporates various preprocessing steps, including outlier rejection, filling missing values, and data standardization, to enhance the quality of the dataset, specifically the Pima Indian Diabetes Dataset. The study applies several machine learning classifiers such as k-Nearest Neighbors, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost, and integrates them through a weighted ensembling technique based on the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). The results demonstrate that the proposed ensemble classifier outperforms other methods, achieving an AUC of 0.950, indicating its superiority in predicting diabetes with high sensitivity and specificity. This research contributes significantly to the field by addressing common issues like missing data and class imbalance while proposing a robust predictive model that surpasses state-of-the-art techniques in accuracy and reliability.
4. “Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning” (Wu et al., 2021): explores the use of machine learning (ML) techniques to predict gestational diabetes mellitus (GDM) during the first trimester of pregnancy. The study leverages a large dataset from the Chinese population, applying ML-driven feature selection methods to identify key variables associated with GDM. The authors develop several predictive models, including logistic regression (LR), support vector machines (SVM), k-nearest neighbor (KNN), and deep neural networks (DNN), using both a comprehensive 73-variable dataset and a more concise 7-variable dataset. The results indicate that the DNN model using the full dataset achieved the highest discriminative power (AUC = 0.80), while the simplified 7-variable LR model also performed well (AUC = 0.77). The study highlights the potential of early GDM prediction using easily accessible clinical data, offering valuable insights for early intervention strategies. The research also emphasizes the significance of thyroid function and lipid metabolism in GDM prediction, suggesting new avenues for future studies. The article is a significant contribution to the field, proposing a clinically viable model that could enhance early diagnosis and management of GDM in at-risk populations.

While previous research has focused on specific types of diabetes or has utilized a limited set of variables, our project aims to integrate predictions for pre-diabetes, Type-2 diabetes, and gestational diabetes into a single, unified model. Additionally, our project emphasizes real-time health coaching, providing personalized guidance based on predictions. This combination of prediction and intervention is unique, addressing not only the detection of diabetes but also offering actionable steps to prevent or manage the condition. Our application of the model will also prioritize user accessibility and ease of use, ensuring that the application can be adopted widely, whether by healthcare providers or individual users. By leveraging a comprehensive dataset that includes demographics, lifestyle factors, and biometric measurements, we aim to deliver a more holistic and personalized approach to diabetes management that extends beyond prediction to active health coaching, filling a critical gap in the current landscape of diabetes-related AI applications.

**Existing Relevant Projects**

In addition to academic literature, there are several open-source projects and competitions that address diabetes prediction using machine learning:

1. **Kaggle: “Diabetes Prediction using Machine Learning”**
   1. This project is a part of a Kaggle competition where participants developed machine learning models to predict diabetes occurrence. The project uses the Pima Indians Diabetes Database and explores various algorithms such as Logistic Regression, Random Forest, and XGBoost. The project focuses on improving prediction accuracy through comprehensive data preprocessing and feature engineering techniques.
2. **GitHub**: **Diabetes Health Prediction and Analysis**
   1. This GitHub repository presents a comprehensive project that predicts and analyzes diabetes health data using advanced machine learning models, including Logistic Regression, Random Forest, and XGBoost. The project includes detailed scripts for data preprocessing, feature engineering, model training, and evaluation, making it a well-rounded resource for diabetes prediction using machine learning.
3. **Papers With Code: HealthEdge – A Machine Learning-Based Smart Healthcare Framework for Prediction of Type 2 Diabetes**
   1. This project involves a smart healthcare framework for predicting type 2 diabetes using an integrated IoT, Edge, and Cloud Computing system. It aims to deliver real-time health monitoring and diabetes prediction, which aligns with the goals of our project.
4. **GitHub: Diabetes Prediction Using Machine Learning**
   1. This repository features a project from a Data Science Academy competition on Kaggle, focused on predicting diabetes using machine learning. The project utilizes datasets from the National Institute of Diabetes and Digestive and Kidney Diseases and provides a kernel for model development.

These projects provide valuable insights and resources that you can leverage in our own work, particularly in terms of model implementation, data preprocessing, and feature engineering. Our project can distinguish itself by integrating real-time health coaching and focusing on a broader range of diabetes types, such as pre-diabetes and gestational diabetes, rather than just type 2 diabetes.

**GitHub Project Link:** [**https://github.com/aa57c/DS\_Capstone\_Project**](https://github.com/aa57c/DS_Capstone_Project)

**Report**

* **Ashna Ali, Project Lead**
  + As the project lead for this assignment, I coordinated team members on what they needed to research in terms of already existing works and datasets. I was responsible for the literature review, finding journal articles and such that aligned with our project. I also examined the datasets we gathered to see which one would work best for validating, testing, and training our model.
* **Sreevardhan Reddy Soma**
* For this assignment, I was responsible for drafting and preparing an Ignite talk PowerPoint presentation to effectively communicate our project’s key points. Additionally, I conducted a search for relevant datasets and worked closely with the team to identify and coordinate the remaining tasks needed to advance our project.
* **Sai Kiran Basetty**
* I collaborated with my teammate in drafting and refining the Ignite talk PowerPoint presentation to clearly convey our project’s main ideas. In addition to this, I took the initiative to research and identify the most suitable machine learning models for our project. I also conducted a thorough search for datasets related to our research, ensuring we have the necessary resources to move forward effectively.
* **Venkata Sai Veeramalla:**
  + As a member of this project, I contributed the relevant datasets. Involved in Discussion and contributed to the approach of the project. Coordinated in preparing the Ignite talk (PPT). Joined and given insights in progressive discussions for project like model building, fixing attributes or features in project with teammates. responsible for data preprocessing and Analysis. Given my insights on final features selection for model building.

Works Cited

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & Metabolic Syndrome*, *13*(1). <https://doi.org/10.1186/s13098-021-00767-9>

Kumar, S., & Vijay Kumar Jha. (2024). Machine and deep learning techniques for the prediction of diabetics: a review. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-024-19766-9>

Hasan, Md. K., Alam, Md. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. *IEEE Access*, *8*, 76516–76531. <https://doi.org/10.1109/access.2020.2989857>

Wu, Y.-T., Zhang, C., Ben W.J. Mol, Kawai, A., Li, C., Chen, L., Wang, Y., Sheng, J.-Z., Fan, J., Shi, Y., & Huang, H.-F. (2021). Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning. *The Journal of Clinical Endocrinology and Metabolism*, *106*(3), e1191–e1205. <https://doi.org/10.1210/clinem/dgaa899>

*Papers with Code - Diabetes Prediction*. (2022). Paperswithcode.com. <https://paperswithcode.com/task/diabetes-prediction>

*Diabetes Prediction using Machine Learning*. (n.d.). Kaggle.com. <https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning>

ThecoderPinar. (2024). *GitHub - ThecoderPinar/Diabetes\_Health\_Prediction\_and\_Analysis: A comprehensive project to predict and analyze diabetes health data using advanced machine learning models, including Logistic Regression, Random Forest, and XGBoost. 📊🔍*. GitHub. <https://github.com/ThecoderPinar/Diabetes_Health_Prediction_and_Analysis>

cpatrickalves. (2019). *GitHub - cpatrickalves/kaggle-diabetes-prediction: A Kaggle competition promoted by Data Science Academy aimed to create a Machine Learning model to predict the occurrence of diabetes.* GitHub. <https://github.com/cpatrickalves/kaggle-diabetes-prediction>