Diabetes Risk Prediction Using Machine Learning

Group Members: Aarna Shah, Jordyn Richardson, Kiana Tse, Katherine Phy, Kayla Tywoniuk, Raaghuv Nandur

Github Repository Link : https://github.com/kianatse/Diabetes-Predictor---Final-Project

Problem Statement

Diabetes mellitus is a chronic disease that affects about 12 % of U.S. adults and is a leading cause of cardiovascular diseases, kidney failure, blindness, and premature mortality. Early identification of high-risk individuals can help initiate preventative lifestyle changes and targeted clinical follow-up, mitigating long-term complications and cost. However, broad screening with laboratory tests is expensive and logistically difficult to implement on large populations. We therefore address the following biomedical problem: *Can self-reported health-behavior and demographic variables be utilized to predict whether an adult currently has—or is very likely to soon develop—diabetes?*

Literature Review:

Recent advancements in machine learning (ML) have demonstrated superior performance over traditional statistical methods for predicting diabetes risk.(Kokhar, 2025) Ensemble methods such as Random Forest classifiers have emerged as highly effective due to their accuracy, interpretability, and adaptability to a wide range of datasets.(Ooka, 2021) Multiple studies utilizing data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) Diabetes Health Indicators dataset—a comprehensive resource consisting of self-reported demographic and health-related factors—is used widely in data analysis, more specifically data mining methods to build classification models.(Sunmoo, 2015) Commonly identified predictors include hypertension, body mass index (BMI), age, history of heart disease or stroke, and physical activity levels, all of which are seen to be correlated with higher risk of diabetes. Researchers emphasize the importance of these features, targeting interventions focusing on lifestyle modification and early clinical management. (Kaliappan, 2024)

While alternative models like neural networks have shown similar predictive strength, their complexity, computational demands, and lack of interpretability create barriers to practical public health implementation.(Romano, 2022) Given these considerations, Random Forest classifiers remain a preferred choice due to their balance between predictive capability, interpretability, and deployment feasibility in clinical and public health settings. The practical implications of employing Random Forest models in clinical and public health settings are significant, offering a viable method for early identification and intervention in diabetes management and prevention strategies.

A study by Ooka et al. (2021) applied Random Forest models to large-scale health check-up data in Japan to predict changes in glycated hemoglobin (HbA1c) levels, a key marker for diabetes. The Random Forest approach outperformed traditional regression models, effectively identifying early indicators of type 2 diabetes. Notably, the model highlighted clinically significant predictors that were not evident in conventional analyses, thereby emphasizing the utility of ensemble methods in uncovering nuanced risk factors.

Similarly, a study by Iparraguirre-Villanueva et al. (2024) explored the use of machine learning models, including decision trees, K-nearest neighbor (K-NN), and Naïve Bayes, to predict the risk of type 2 diabetes. The authors focused on patient data, including demographics and clinical factors like BMI, blood glucose levels, and age, to classify individuals as diabetic or non-diabetic. The research demonstrated that machine learning models could effectively predict diabetes, with K-NN and Naïve Bayes showing the highest accuracy. This aligns with our focus on using self-reported health-behavior and demographic variables for diabetes prediction and highlights the importance of leveraging accessible health data for early diabetes detection. Their findings also support the notion that machine learning techniques can uncover subtle, critical risk factors that might be overlooked by traditional statistical methods

Tasin et al. (2025) further illustrates the utility of machine learning in diabetes prediction by developing a model using both the Pima Indian dataset and a private dataset from Bangladeshi female patients. By employing feature selection algorithms and ensemble techniques such as XGBoost, the authors achieved an 81% accuracy rate in predicting diabetes risk. The use of explainable AI methods like SHAP and LIME ensured that the model's predictions were not only accurate but also transparent. Therefore, making the system both reliable and interpretable for public health applications. This research adds to the growing body of evidence supporting the use of machine learning models for practical, real-time diabetes prediction in clinical settings.

Proposed Solution:

To address the outlined biomedical challenge, this project uses the Diabetes Health Indicators dataset, which includes health-related and demographic data from a large number of individuals. Key predictive factors from the dataset were high blood pressure, high cholesterol, cholesterol check history, BMI, smoking habits, history of stroke or heart disease, physical activity, dietary habits (fruit and vegetable intake), heavy alcohol consumption, access to healthcare, gender, age, and education level.

Before training the model, the data underwent basic preprocessing. Missing values in the dataset were handled appropriately to maintain data quality. Given the type of data used (mostly binary and ordinal variables) and the classifier, additional encoding and scaling features were not needed.

The dataset was then split into two groups for training and testing (20%) and class imbalance was considered to ensure the model learned effectively from both groups.

Results and Evaluation:

To assess the performance of our diabetes risk prediction model, we conducted a thorough evaluation using a test set derived from an online dataset. Our model utilizes a subset of the available health indicators, focusing only on the most impactful predictors: High Blood Pressure (HighBP), Body Mass Index (BMI), General Health (GenHlth), Mental Health (MentHlth), Physical Health (PhysHlth), Age, and Education. These variables were selected based on their correlation with diabetes and their ability to contribute meaningful signals for classification. The user would provide their answers in binary where 1 means yes and 0 means no. However, for the Age and BMI variables the user could insert the actual numerical value. The confusion matrix below summarizes how well our trained model predicted diabetes outcomes compared to the actual labels in our test set (20% of the dataset).

Confusion Matrix		
	Predicted Diabetes	Predicted No Diabetes
Actual Yes	30,446	4,900
Actual No	63,434	154,900

Confusion Matrix Breakdown:

- True Positive (TP) = 30,446
  - This is the number of people who had diabetes, and the model correctly predicted it.
- False Negative (FN) = 4,900
  - This is the number of people who had diabetes, but the model incorrectly predicted that they did not.
- False Positive (FP) = 63,434
  - This is the number of people who did not have diabetes, but the model incorrectly predicted that they did.
- True Negative (TN) = 154,900
  - This is the number of people who did not have diabetes, and the model correctly predicted it.

The model makes predictions on each instance in the test set. Each prediction is compared with the actual label. Every time the prediction and actual label match or don't match, it counted in one of the four categories (TP, FN, FP, or TN). After going through all test instances, the counts are compiled into the confusion matrix. The values from the confusion matrix were then used to calculate each of the model performance metrics below.

Model Performance Metrics	
Name	**Value**
Accuracy	0.7306
Precision	0.3243

Recall	0.8614
F1 Score	0.4712

The high recall (86.14%) indicates that the model is highly effective at identifying individuals who are at risk of diabetes. However, the precision value (32.43%) was relatively low. This means a significant number of individuals predicted to be diabetic were not actually diabetic. The accuracy (73.06%) and F1 score (47.12%) values reflect the balance between detecting true cases and managing false alarms.

One major issue was due to a mismatch between the features used during model training and those present in the evaluation dataset. The model had been trained on a specific subset of seven features, but the evaluation dataset initially included all original variables from the source CSV file. This mismatch caused the evaluation script to give an error, as the model expected a specific input structure. The issue was resolved by programmatically extracting the feature names the model had been trained on and filtering the test data to match those exact variables.

Another significant challenge was developing the model itself. Up until this point in the course, most of the projects and assignments had involved guided instructions or starter code. However, this project required writing and debugging a relatively complex script independently. For many of us, this was the first time writing a machine learning script from scratch. This process involved a lot of trial and error, researching documentation, and learning how to interpret error messages effectively. While it took quite a bit of time, we were eventually able to get the code to run successfully and extract meaningful results from the model.

Practical Application:

The machine learning model developed in this project can play a vital role in shifting diabetes care from reactive treatment to proactive prevention. One of the most practical and immediate applications of this model is in primary care clinics. Often, primary care physicians rely on periodic lab testing to identify diabetes, which may miss early-stage or high-risk individuals due to factors such as infrequent screening or cost. By integrating our solution into electronic health record systems, providers could run real-time risk assessments during routine visits using self-reported information that patients already provide. For example, weight, physical activity, smoking status, and family history. When a patient's risk exceeds a certain threshold, the system could prompt the clinician to order confirmatory lab tests or initiate a conversation about lifestyle interventions.

Beyond individual clinics, this tool could be implemented at the public health level. To illustrate, state or county health departments often collect data through large-scale surveys like the Behavioral Risk Factor Surveillance System (BRFSS). This model can be applied directly to such datasets to identify high-risk populations in specific geographic regions. Public health officials could then use these insights to prioritize community outreach efforts, create targeted educational campaigns, or allocate resources for diabetes prevention programs.

Another promising application is in digital health platforms or telemedicine apps. Many consumers now use apps to track their health behaviors and get personalized wellness advice. Embedding the model within such platforms could provide users with an evidence-based diabetes risk score generated from their daily input data (e.g., diet, exercise, weight updates). Unlike clinical tools that require lab values, this model's reliance on self-reported and behavioral data makes it well-suited for consumer health applications. A user flagged as high-risk could then be offered tailored advice or even a virtual consultation.

References

1. "National Diabetes Statistics Report." Centers for Disease Control. https://www.cdc.gov/diabetes/php/data-research/index.html

2. T. Pannmie. "🍖 Diabetes : EDA | 🎄 Random Forest 🎄 + HP." Kaggle. https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp

3. https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1557467/full

   M. Kiran et. al, "Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis," Frontiers in Digital Health, vol. 7, March 2025, doi: 10.3389/fdgth.2025.1557467

4. **https://www.sciencedirect.com/science/article/pii/S0933365725000673**

   P. B. Khokhar, C. Gravino, & F. Palomba, "Advances in artificial intelligence for diabetes prediction: insights from a systematic literature review," *Artificial Intelligence in Medicine,* vol. 164, June 2025, https://doi.org/10.1016/j.artmed.2025.103132.

5. "Behavioral Risk Factor Surveillance System." Centers for Disease Control and Prevention. https://www.cdc.gov/brfss/index.html

6. **https://pmc.ncbi.nlm.nih.gov/articles/PMC4580372/**

   S. Yoon, B. Taha, & S. Bakken, "Using a Data Mining Approach to Discover Behavior Correlates of Chronic Disease: A Case Study of Depression, Stud Health Technol Inform, vol. 201, pp. 71-78, 2014.

7. **https://pmc.ncbi.nlm.nih.gov/articles/PMC11371799**

J. Kaliappan et. al, "Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets," Frontiers in Artificial Intelligence, Aug. 2024, doi: 10.3389/frai.2024.1421751.

8. **https://pmc.ncbi.nlm.nih.gov/articles/PMC9313085/**

R. Weiss, S. Karimijafarbigloo, D. Roggenbuck, & S. Rödiger, “Applications of Neural Networks in Biomedical Data Analysis,” Biomedicines, vol. 10, no. 7, June 2022, doi: 10.3390/biomedicines10071469.

9. **https://nutrition.bmj.com/content/early/2021/03/09/bmjnph-2020-000200**

   Ooka, Tadao et al. “Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: large-scale health check-up data in Japan.” *BMJ Nutrition, Prevention & Health* 2021;:bmjnph-2020-000200. https://doi.org/10.1136/bmjnph-2020-000200

10. **https://pmc.ncbi.nlm.nih.gov/articles/PMC10378239/**

    Iparraguirre-Villanueva, Orlando et al. “Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes.” *Diagnostics (Basel, Switzerland)* vol. 13,14 2383. 15 Jul. 2023, doi:10.3390/diagnostics13142383

11. **https://pmc.ncbi.nlm.nih.gov/articles/PMC10107388/#htl212039-sec-0070**

    Tasin, Isfafuzzaman et al. “Diabetes prediction using machine learning and explainable AI techniques.” *Healthcare technology letters* vol. 10,1-2 1-10. 14 Dec. 2022, doi:10.1049/htl2.12039