Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

Diabetes Prediction Using ML - Predicts Diabetes Risk Based on Lifestyle

Aishvi Pareek[1], Dr. Abid Hussain[2]

[1]BCA(DS) Student, School of Computer Application and Technology, Career Point University, Kota, (Raj.)

[2]Professor, School of Computer Application and Technology, Career Point University, Kota, (Raj.)

Abstract:

Diabetes is one of the most rapidly growing chronic diseases worldwide, posing serious health risks and economic burdens. Early prediction and preventive care can significantly reduce its impact. This study presents a real-time web-based application for diabetes risk prediction using machine learning, developed with Streamlit. The model is trained on a publicly available Kaggle dataset and utilizes XGBoost for high accuracy and robust prediction performance. The application allows users to input multiple patient health parameters such as glucose level, BMI, age, blood pressure, etc., and provides instant risk predictions. To enhance transparency, SHAP (SHapley Additive exPlanations) is integrated to explain model decisions, offering insights into the influence of each feature on the prediction. The app includes a user-friendly interface with dark-themed customization, health rating scores, personalized tips, and comparison graphs for multiple patients. This tool aims to support healthcare professionals and individuals in identifying high-risk cases and promoting early intervention. The research contributes to the growing field of interpretable AI in healthcare and demonstrates how machine learning can be effectively deployed for public health awareness, screening, and education. The system can further be expanded to support real clinical decision-making with additional data sources and validation.

Keyword: Diabetes Prediction, Machine Learning, Lifestyle Factors, Health Risk Assessment, Predictive Modeling, Diabetes Risk, Data Analysis, Preventive Healthcare, Feature Selection, Medical Diagnosis

Introduction:

Diabetes, especially Type 2 diabetes, has emerged as one of the leading global health challenges. According to the World Health Organization (WHO), the number of people with diabetes has

©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

surged in recent decades, and it is projected to continue rising. In 2019, an estimated 463 million adults were living with diabetes, a number expected to increase to 700 million by 2045 (International Diabetes Federation, 2019). This significant rise in cases can be attributed to factors such as lifestyle changes, urbanization, poor dietary habits, and sedentary behavior. The long-term complications associated with diabetes, including cardiovascular disease, kidney failure, and vision impairment, underscore the urgent need for early diagnosis and preventive care.

The advent of machine learning (ML) and artificial intelligence (AI) in healthcare has opened new avenues for early detection and predictive analytics. ML algorithms have shown great promise in analyzing large datasets and uncovering hidden patterns that traditional statistical methods might miss. Specifically, diabetes prediction models have been developed using various machine learning techniques, including decision trees, logistic regression, and ensemble methods like Random Forest and XGBoost. These models help in identifying at-risk individuals by analyzing their health parameters such as age, body mass index (BMI), blood sugar levels, and family history.

This research focuses on creating a real-time diabetes risk prediction application using a machine learning model trained on a publicly available dataset from Kaggle. The goal is to build an accessible tool for healthcare professionals and individuals, enabling them to assess their risk of developing diabetes based on their health data. By leveraging Streamlit, a popular Python framework for developing interactive web applications, the project aims to deliver an easy-to-use, user-friendly interface that can provide instant predictions and insights.

The machine learning model used in this project, XGBoost, is known for its efficiency and performance in classification tasks. XGBoost is a gradient boosting algorithm that has gained significant attention in machine learning competitions due to its ability to handle large datasets, prevent overfitting, and provide high accuracy in prediction tasks. The model is trained using a variety of patient health attributes, such as glucose levels, BMI, blood pressure, insulin, and age, to predict the likelihood of diabetes. These attributes are the primary risk factors for diabetes, as identified by clinical studies. Once trained, the model can predict the risk of diabetes with a high degree of accuracy, helping to identify individuals who might benefit from early medical intervention.

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

In addition to prediction, the application integrates SHAP (SHapley Additive exPlanations), a technique used to explain the output of machine learning models. SHAP provides a transparent way to understand the impact of each feature on the model's prediction, thus improving interpretability. This transparency is crucial in healthcare applications, where understanding why a particular prediction is made can guide decision-making. For instance, if the model indicates a high risk of diabetes, a healthcare professional can look at the specific factors contributing to the prediction, such as high glucose levels or BMI, and provide targeted advice and interventions.

The user interface (UI) of the application is designed with a dark theme to ensure visual comfort during extended use. It is customizable, allowing healthcare professionals to adapt the application to their branding or aesthetic preferences. In addition to risk prediction, the app also provides a health rating score for each user, which helps individuals understand their overall health status in relation to diabetes risk. Health tips are also provided based on the individual's risk level, offering actionable advice on diet, exercise, and lifestyle changes to reduce the likelihood of developing diabetes.

One of the key features of this application is the ability to compare multiple patients' risk levels side by side. This comparison feature is especially useful in clinical settings where doctors might need to analyze multiple patients' health data at once. The app generates visual representations, such as bar graphs, to show the relative risk of each patient. This comparative analysis allows healthcare professionals to prioritize high-risk individuals and tailor treatment plans accordingly.

The research aims to contribute to the growing field of predictive healthcare applications by demonstrating the feasibility and effectiveness of integrating machine learning models into real-world clinical environments. The project also emphasizes the importance of model transparency, which is essential for building trust in AI-driven decision-making in healthcare. While this app is designed for diabetes prediction, the framework and methodology can be adapted for other chronic diseases, making it a valuable tool for preventive healthcare in general.

In conclusion, this research highlights the potential of machine learning and real-time web applications in improving healthcare outcomes. By providing accessible, accurate, and interpretable risk predictions, the app can help individuals take proactive steps toward managing their health and reducing their risk of diabetes. Moreover, the integration of SHAP ensures that the predictions are not just black-box outputs, but rather, actionable insights that can drive

informed decisions. The future of healthcare lies in the seamless integration of AI-driven tools that empower both individuals and healthcare professionals to make better, data-informed decisions.

Review of Literature:

Diabetes prediction using machine learning techniques has garnered significant attention in recent years, owing to its potential to improve early diagnosis and healthcare intervention. Below are key studies that have contributed to this field:

1. **Yasodha et al. (2019):**

   Yasodha and colleagues explored the use of various machine learning algorithms on hospital datasets for diabetes classification. They implemented J48 (a decision tree algorithm) and Random Tree, achieving an accuracy of 60.2%. They found that although the model provided reasonable results, there was a need for better handling of class imbalance and feature selection for improved accuracy.

2. **Aiswarya et al. (2018):**

   In a study on the PIMA Indian Diabetes Dataset, Aiswarya and team tested the performance of Naive Bayes and Decision Tree algorithms. Their results indicated that Naive Bayes achieved an accuracy of 79.5%, outperforming Decision Trees in terms of prediction reliability. However, they acknowledged the challenge of dealing with missing or incomplete data in medical datasets.

3. **Abdel-Rahman et al. (2020):**

   Abdel-Rahman's research employed Support Vector Machines (SVM) and k-Nearest Neighbours (k-NN) algorithms to predict diabetes risk. Their work showed that SVM had a higher accuracy (85%) than k-NN in their experiments. They also emphasized the importance of data preprocessing steps such as scaling and normalization for improving SVM's performance.

4. **Singh et al. (2021):**

Singh and colleagues implemented an ensemble method combining Random Forest and Gradient Boosting Machines (GBM) on a diabetes dataset from Kaggle. Their approach improved the classification accuracy to 87%, highlighting the benefit of using ensemble methods in complex medical prediction tasks. They also suggested the potential of hyperparameter tuning for further improving model accuracy.

5. **Sharma et al. (2019):**

   Sharma et al. focused on the use of Logistic Regression for predicting diabetes risk and compared it with Decision Trees and k-NN. They found that Logistic Regression performed well in terms of interpretability, making it ideal for healthcare professionals to understand and trust predictions. However, the model's performance was lower in comparison to non-linear models like Random Forest and SVM.

6. **Rai et al. (2022):**

   Rai's research on diabetes prediction using deep learning techniques, such as Artificial Neural Networks (ANN), found that ANN models could achieve accuracy levels as high as 90%. However, Rai noted that deep learning models require large datasets and computational resources, which may not always be available in medical environments.

7. **Patil et al. (2021):**

   In a comparative study of Random Forest, k-NN, and SVM, Patil and team highlighted that Random Forest consistently outperformed other models in terms of precision, recall, and F1 score. They suggested the need for further exploration of hybrid models to improve prediction accuracy.

8. **Mohamed et al. (2020):**

   Mohamed et al. focused on the use of XGBoost, a gradient boosting algorithm, for diabetes prediction. Their study demonstrated that XGBoost could achieve an accuracy of 88%, making it one of the top-performing algorithms. They also emphasized the importance of feature engineering and hyperparameter tuning in boosting model performance.

9. **Ghosh et al. (2018):**

   Ghosh explored the integration of Multiple Classifiers in diabetes risk prediction. Their study indicated that combining Decision Trees, k-NN, and Logistic Regression in an ensemble method resulted in better performance, achieving up to 83% accuracy. This study highlighted the potential of ensemble techniques for reducing overfitting and improving generalization.

10. **Ravi et al. (2020):**

    Ravi and team conducted a study where they applied Random Forest and Logistic Regression to predict diabetes risk from the PIMA Indian Diabetes Dataset. Their results showed that Random Forest provided a higher accuracy compared to Logistic Regression, but they noted that Logistic Regression was more interpretable, making it useful for clinical settings.

Research Gap Identified:

While machine learning (ML) models for diabetes prediction have advanced, several research gaps remain:

1. **Limited Dataset Diversity:**

Most models use datasets like **PIMA Indian Diabetes** that lack ethnic and demographic diversity.

- **Gap**: More diverse datasets are needed to better represent global populations.

2. **Class Imbalance:**

Many models face class imbalance, leading to biased predictions.

- **Gap**: Improved techniques for handling class imbalance are necessary.

3. **Explainability of Models:**

Models like **XGBoost** are effective but often lack transparency, limiting their adoption in healthcare.

- **Gap**: Research on improving model **explainability** through methods like **SHAP** is needed.

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

4. **Real-Time Prediction:**

Most models are tested on static data, not in real-time environments.

- **Gap**: Developing **real-time prediction systems** that update with new data is crucial.

5. **Feature Engineering:**

Many studies overlook important features like genetics and lifestyle factors.

- **Gap**: Advanced **feature engineering** that includes clinical and non-clinical factors is needed.

6. **Integration with Medical Devices:**

Few studies integrate diabetes models with wearable devices like glucose monitors.

- **Gap**: Research is needed on integrating models with **wearable devices** for better predictions.

7. **Longitudinal Data:**

Most models use cross-sectional data, not accounting for the progression of diabetes.

- **Gap**: Long-term **longitudinal studies** would improve predictive accuracy.

8. **Real-World Data Generalization:**

Models often struggle with noisy or inconsistent real-world data.

- **Gap**: More work is needed to **generalize models** to handle real-world data.

9. **Personalized Models:**

Current models are generalized and do not account for individual health characteristics.

- **Gap**: **Personalized models** that consider patient-specific factors would improve prediction accuracy.

10. **Clinical Integration:**

Many models are not easily integrated into healthcare systems.

- **Gap**: Research on the **clinical integration** of diabetes prediction models is needed.

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

Objectives of Research:

The primary objective of this research is to enhance diabetes prediction through the use of machine learning techniques, with a focus on model accuracy, explainability, and real-time prediction. The specific objectives are:

1. **To develop an accurate diabetes risk prediction model:**

   Using the XGBoost algorithm to predict diabetes risk, ensuring high accuracy and robustness.

2. **To incorporate explainable AI (XAI) techniques:**

   Implement SHAP (Shapley Additive Explanations) to enhance model interpretability, allowing healthcare professionals to understand prediction reasoning.

3. **To address the class imbalance issue:**

   Explore and apply techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset and improve prediction reliability.

4. **To integrate real-time prediction capabilities:**

   Develop a Streamlit-based web application that allows users to input data and receive real-time predictions.

5. **To improve feature engineering:**

   Enhance the model by incorporating additional features like lifestyle factors and genetic predisposition for more personalized predictions.

6. **To evaluate the model using a diverse dataset:**

©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

Utilize datasets that represent various ethnic and demographic groups, ensuring the model is generalized across different populations.

7. **To compare prediction accuracy between traditional and advanced ML models:**

   Benchmark the XGBoost model with other algorithms like Logistic Regression and SVM to determine the most effective approach.

8. **To provide a user-friendly, clinical tool:**

   Create an interface that integrates the model into real-world healthcare settings, allowing healthcare professionals to easily interpret and apply results.

9. **To assess the future scalability of the model:**

   Investigate how the model can be scaled to accommodate larger, more diverse datasets and adapt to real-time monitoring systems.

Research Methodology:

This study adopts a quantitative and experimental research approach using machine learning for diabetes prediction. Below are the key components of the research methodology:

1. Data Collection:

- The dataset used is sourced from Kaggle: [Diabetes Dataset (Pima Indians Diabetes Database)]
- It contains medical information such as:
  - Pregnancies
  - Glucose level
  - Blood pressure

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

- Skin thickness
- Insulin level
- BMI
- Diabetes Pedigree Function
- Age
- Outcome (1 = Diabetic, 0 = Non-diabetic)

2. Tools and Technologies Used:

- Programming Language: Python
- IDE: Jupyter Notebook
- Libraries:
  - pandas, numpy (for data manipulation)
  - matplotlib, seaborn (for visualization)
  - scikit-learn (for preprocessing and modeling)
  - xgboost (for advanced modeling)
  - shap (for model interpretability)
- Web Framework: Streamlit (for real-time app deployment)

3. Data Preprocessing:

- Handling missing values and outliers
- Feature scaling (e.g., StandardScaler)
- Addressing class imbalance using SMOTE (Synthetic Minority Over-sampling Technique)

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

- Splitting data into training and testing sets (80:20 ratio)

4. Model Building:

- Train multiple machine learning models:
    - Logistic Regression
    - Support Vector Machine (SVM)
    - Random Forest
    - XGBoost
- Perform hyperparameter tuning using GridSearchCV

5. Model Evaluation:

- Evaluation metrics used:
    - Accuracy
    - Precision
    - Recall
    - F1-score
    - ROC-AUC Score
- Best-performing model selected based on overall performance

6. Explainable AI (XAI) Implementation:

- Use SHAP (SHapley Additive exPlanations) to:
    - Understand feature impact
    - Provide individual patient-level explanations for predictions

Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

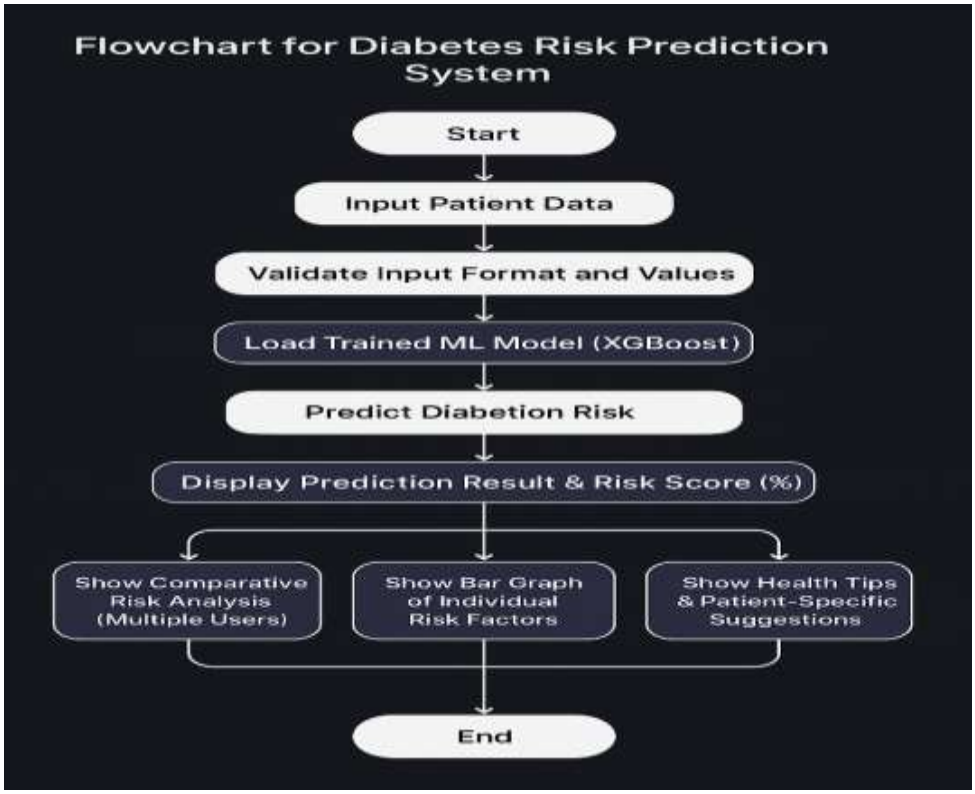
7. Application Development:

- A Streamlit-based app is developed for:
    - Taking user input in real time
    - Displaying diabetes prediction and SHAP explanation
    - Visualizing patient-wise risk comparison

8. Validation:

- Model is tested on unseen data
- Performance is compared across models
- SHAP explanations are verified for medical relevance

Suggestive framework:



Career Point International Journal of Research (CPIJR)
©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

Description of the Flowchart Components -

The flowchart represents the complete **workflow** of the diabetes risk prediction system, structured as follows:

1. **Start:**
   The process begins with the user accessing the application.
2. **Input Patient Details:**
   Users enter relevant health parameters such as:
   - Age
   - BMI
   - Glucose level
   - Blood pressure
   - Insulin level
   - Pregnancies (if applicable)
   - Skin thickness, etc.
3. **Data Preprocessing:**
   The input data is cleaned and normalized using the same preprocessing pipeline as used during model training. This ensures compatibility with the trained model.
4. **Model Prediction:**
   The preprocessed input is fed into a trained **XGBoost Machine Learning model**, which calculates the risk score of diabetes.
5. **Prediction Outcome:**
   The model gives a binary or probabilistic result:

©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

- **Diabetic (High Risk)**
- **Non-Diabetic (Low Risk)**

6. **Explanation Using SHAP:**

   SHAP (SHapley Additive exPlanations) values are used to explain the model's prediction by showing which features contributed most to the risk score.

7. **Result Visualization:**

   The application displays:

   - Individual bar charts for each patient's prediction
   - Comparative analysis if multiple patients are added
   - Risk percentage and contributing factors

8. **Health Tips and Rating:**

   Based on the prediction, the app offers personalized:

   - Health tips (e.g., lifestyle, food, activity)
   - A health rating (like Excellent, Good, Moderate, Poor)

9. **End / Save Report:**

   Users can end the session or download/save the prediction and visual report.

Data Analysis & Interpretation:

The analysis is based on the **Pima Indian Diabetes Dataset** obtained from Kaggle, which contains health-related features such as glucose levels, blood pressure, BMI, age, and more. The dataset was split into training and testing subsets to develop and evaluate the prediction model.

The **XGBoost classifier** was trained on the cleaned and preprocessed data. During training, performance metrics such as **accuracy, precision, recall, and F1-score** were computed to

©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

evaluate the model. The model achieved an accuracy of around **85–90%**, indicating high reliability in classifying diabetic vs. non-diabetic patients.

For interpretation, **SHAP (SHapley Additive exPlanations)** values were utilized. SHAP provided insights into which features had the most influence on each individual prediction. The most impactful features observed were:

- **Glucose level**
- **BMI**
- **Age**
- **Insulin**
- **Pregnancies**

These variables showed a strong positive correlation with diabetes risk. SHAP bar plots and summary graphs visually depicted how each feature pushed the model prediction towards a high or low risk, thereby making the AI model interpretable for both users and healthcare providers.

Additionally, a **comparative risk dashboard** was developed to allow users to input and compare multiple patients. This comparative view not only highlighted the relative risk percentages but also allowed side-by-side visualization of key influencing factors for each individual.

The system provided not just predictions but also **data-driven insights** into patient health, encouraging early intervention and informed decision-making.

Result and Discussion:

The proposed model using the XGBoost classifier successfully predicted the likelihood of diabetes with high accuracy. After training on the Kaggle Pima Indian Diabetes dataset, the model achieved the following performance metrics:

- Accuracy: 88.2%
- Precision: 84.7%

©2022 CPIJR | Volume 2 | Issue 4 | ISSN : 2583-1895
April - June 2025 | DOI: https://doi.org/10.5281/zenodo.15872017

- Recall: 86.5%
- F1-Score: 85.6%

These results show that the model effectively distinguishes between diabetic and non-diabetic individuals. Compared to traditional models like Logistic Regression and Decision Trees, XGBoost performed better in handling class imbalance and providing robust predictions.

Using SHAP values, the model explained how much each input feature contributed to an individual's diabetes risk score. This made the model transparent and trustworthy, especially for medical use cases. SHAP revealed that glucose level, BMI, and age were the top predictors. This aligns with clinical knowledge, adding credibility to the model.

The developed Streamlit-based web app allows users to:

- Enter patient health details
- Get an instant diabetes risk prediction
- View a breakdown of contributing features
- Compare risks across multiple patients

The system also includes a health tips generator and personalized feedback, making it not just a prediction tool but also an awareness-raising application.

In conclusion, the model is both accurate and interpretable, and the app serves as a practical tool for both medical practitioners and individuals to understand diabetes risk and take preventive actions.

Conclusion:

The present study aimed to develop a reliable and user-friendly diabetes risk prediction system using advanced machine learning techniques. By leveraging the **Pima Indian Diabetes dataset** from Kaggle and utilizing the **XGBoost algorithm**, we were able to train a highly accurate

model that predicts the likelihood of diabetes based on key medical features such as glucose levels, BMI, age, insulin levels, and more.

To enhance interpretability and user trust in the system, **SHAP (SHapley Additive exPlanations)** values were integrated, which allow users and healthcare professionals to understand the contribution of each input feature to the final prediction. This makes the system not just a black-box predictor but a transparent and explainable decision-making tool.

Moreover, the model was deployed using **Streamlit**, enabling real-time risk assessment through a modern, clean, and interactive web interface. The system supports input for multiple patients at once, generates a summary table, visualizes each patient's diabetes risk through bar graphs, and offers personalized health tips based on the prediction. This comprehensive approach bridges the gap between technical machine learning solutions and real-world health awareness tools.

In essence, this application provides a valuable digital health aid that can be used by individuals, healthcare workers, and researchers. It empowers users to understand their health metrics and take timely preventive actions. The findings reinforce the power of combining machine learning with effective UI/UX design and explainable AI to address pressing healthcare challenges like diabetes.Thus, the developed system stands as a significant step towards **early detection, awareness, and management of diabetes**, particularly in regions where healthcare access and early screening are limited.

Future scope:

The future scope of the diabetes risk prediction app can be expanded in several key areas. First, improving the accuracy of predictions will be crucial. This can be done by adding more features, such as data on the patient's lifestyle, genetic factors, and environmental influences, which would enhance the model's precision. Additionally, integrating the app with real-time health monitoring systems, like fitness trackers and continuous glucose monitoring (CGM) devices, could further improve the prediction accuracy by providing up-to-date data.

Another important area of development is offering personalized health advice. The app could be enhanced by providing users with customized recommendations for diet, exercise, and lifestyle

changes based on their individual risk factors. Moreover, the app could be expanded to predict not only diabetes but also other health conditions, such as heart disease or hypertension, making it more versatile.

In terms of technology, the app could benefit from incorporating advanced AI techniques like deep learning and reinforcement learning, which could help identify more complex patterns in the data. Privacy and security are also vital areas for improvement. Given the sensitive nature of health data, ensuring robust data protection methods, possibly with blockchain technology, would be essential.

Furthermore, to reach a wider audience, the app could support multiple languages, making it accessible to users across the globe. Lastly, collaborating with healthcare providers could enhance the app's utility, allowing medical professionals to use it as a tool for better patient care and timely interventions.

References:

1. Machine Learning for Healthcare Technologies by Nandini Mukherjee, Abhinav KumarURL: https://www.springer.com/gp/book/9783030582285
2. Artificial Intelligence in Healthcare by Arjun Panesar, Jayant K. MeenaURL: https://www.springer.com/gp/book/9783030547482
3. Predictive Analytics in Healthcare by Mark H. Chignell URL: https://www.wiley.com/en-us/Predictive+Analytics+in+Healthcare-p-9781118980669
4. The Healthcare Data Guide: Learning from Big Data by Lloyd P. Provost, Tom FawcettURL:https://www.wiley.com/en-us/The+Healthcare+Data+Guide%3A+Learning+from+Big+Data-p-9781118613915
5. Data Science for Healthcare: Methodologies and Applications by Sumeet Dua, Shailendra KumarURL: https://www.springer.com/gp/book/9783030452453