Diabetics Detection Using Python

Sai Priya.R, Sarin Priya.D
Computer Science and Engineering, PSNA College of Engineering and Technology Dindigul, India

Abstract— Diabetes is a significant global health issue that necessitates early diagnosis and management to prevent severe complications. The project aims to develop a web based application using Flask and machine learning to predict the likelihood of diabetes. Users input specific health parameters such as age, BMI, glucose levels, and insulin levels through a user-friendly web interface. The backend, powered by Flask, processes these inputs using a pre-trained machine learning model (e.g., Random Forest Classifier) to provide instant predictions. The application includes robust security measures to ensure the privacy and confidentiality of user data. Additionally, it logs user inputs and prediction results to continuously improve the model's accuracy. The system architecture involves a responsive front-end, a Flask based backend, and a secure database for storing user data. This approach facilitates early diagnosis and timely management, potentially improving health outcomes and providing a valuable tool for both individuals and healthcare providers.

I. INTRODUCTION

Diabetes is one of deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, kidney diseases, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. Moreover, every time they want to get their diagnosis report, they have to waste their money in vain. Diabetes Mellitus (DM) is defined as a group of metabolic disorders mainly caused by abnormal insulin secretion and/or action. Insulin deficiency results in elevated blood glucose levels (hyperglycemia) and impaired metabolism of carbohydrates, fat and proteins. DM is one of the most common endocrine disorders, affecting more than 200 million people worldwide. The onset of diabetes is estimated to rise dramatically in the upcoming years. DM can be divided into several distinct types. However, there are two major clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D), according to the etiopathology of the disorder. T2D appears to be the most common form of diabetes (90% of all diabetic patients), mainly characterized by insulin resistance. The main causes of T2D include lifestyle, physical activity, dietary habits and heredity, whereas T1D is thought to be due to autoimmunological destruction of the Langerhans islets hosting pancreatic-β cells. T1D affects almost 10% of all diabetic patients worldwide, with 10% of them ultimately developing idiopathic diabetes. Other forms of DM, classified on the basis of insulin secretion profile and/or onset, include Gestational Diabetes, endocrinopathies, MODY (Maturity Onset Diabetes of the Young), neonatal, mitochondrial, and pregnancy diabetes. The symptoms of DM include polyuria, polydipsia, and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L. and significant weight loss among others. Diagnosis depends on blood glucose levels (fasting plasma glucose = 7.0 mmol/L.

II. RELATED WORKS

Diabetes mellitus, a chronic metabolic disorder characterized by high blood glucose levels, poses a significant global health concern. Early detection and accurate diagnosis are crucial to managing and mitigating its long-term effects. In recent years, the integration of artificial intelligence, particularly machine learning (ML) and deep learning (DL), into medical diagnostics has revolutionized how diseases like diabetes are predicted and analyzed. Among the many programming tools available, Python has emerged as the most popular and effective language for developing predictive models due to its readability, robust libraries, and extensive community support.

Early diagnosis of diabetes can prevent complications such as cardiovascular disease, nerve damage, kidney failure, and vision problems. Traditional diagnostic methods rely heavily on physical symptoms, lab results, and physician assessments, which may delay early-stage detection. By contrast, machine learning models can analyze historical data, identify patterns, and forecast risk levels more efficiently. This shift from reactive to proactive healthcare has prompted researchers and

developers to explore predictive modeling as a complementary diagnostic aid.

1. Logistic Regression:

This is a statistical model used for binary classification problems. In the context of diabetes detection, it calculates the probability of a patient having diabetes based on input features. Logistic regression is preferred for its simplicity, interpretability, and effectiveness in linearly separable datasets.

2. Decision Trees:

Decision Trees work by splitting the dataset based on feature thresholds, creating a tree-like structure of decisions. They are easy to visualize and interpret, making them useful for understanding which features contribute most to predictions.

3. Random Forest: Random Forest is an ensemble learning technique that builds multiple decision trees and aggregates their predictions to improve accuracy and prevent overfitting. This method often performs better than individual decision trees and handles non-linear relationships well.

4. Support Vector Machines (SVM) :

SVMs are powerful classifiers that work by finding the optimal hyperplane that separates classes in a high-dimensional space. They are especially effective in cases where the number of dimensions exceeds the number of samples and when the classes are not linearly separable.

III. PROPOSED METHODOLOGY

This study aims to propose a new model for diabetic's classification. Numerous algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy. The methods employed in this research are split by the four main phases of the research work, which are the problem formulation phase, the dataset collection phase and the experimentation phase and the results summarizing. This research started with formulating the research problem that is reviewing of the literature and formulating of the research problem. After the research problem formulation, this research identified the scope of the research, the objectives, and limitations of the research procedure. The second phase of the study is the dataset collection. The dataset items were collected from Pima Indians diabetes dataset.

The third phase of the study was the data preparation which included:

- Converting Data to Appropriate format,
- Data Preprocessing,
- Use Machine Learning to manipulate Data In the experimentation phase.

DATA COLLECTION:

The system utilizes a dataset containing various health parameters and diabetes outcomes, such as the Pima Indians Diabetes Dataset. This dataset includes features like age, BMI, glucose levels, insulin levels, and more.

	A	B	C	D	E	F	G	H	I
1	Pregnancie	Glucose	Blood_Pre	Skin_Thick	Insulin	BMI	Diabetes_Pedigree_Function	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1

MODEL TRAINING:

DATA PREPROCESSING:

Data Preprocessing includes:

Handling missing value: Ensuring the dataset is complete and clean. Normalizing or standardizing data: Preparing the data for model training. Data Splitting: Dividing the dataset into training and testing subsets to evaluate model performance.

```
import pandas as pd
file_path='/content/drive/MyDrive/python folder/diabetes.csv'
df = pd.read_csv(file_path,encoding='ISO-8859-1')
print(df.head(10))


print(f"Number of duplicate rows: {df.duplicated().sum()}")
print(df.duplicated().sum(), df.drop_duplicates().duplicated().sum())
```

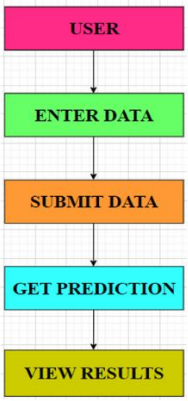
```
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  \
0            2      138             62             35        0  33.6
1            0       84             82             31      125  38.2
2            0      145              0              0        0  44.2
3            0      135             68             42      250  42.3
4            1      139             62             41      480  40.7
5            0      173             78             32      265  46.5
6            4       99             72             17        0  25.6
7            8      194             80              0        0  26.1
8            2       83             65             28       66  36.8
9            2       89             90             30        0  33.5

   DiabetesPedigreeFunction  Age  Outcome
0                     0.127   47        1
1                     0.233   23        0
2                     0.630   31        1
3                     0.365   24        1
4                     0.536   21        0
5                     1.159   58        0
6                     0.294   28        0
7                     0.551   67        0
8                     0.629   24        0
9                     0.292   42        0
Number of duplicate rows: 8276
8276 0
```

MODEL SELECTION:

Training various machine learning models (e.g., Random Forest Classifier, Logistic Regression, Support Vector Machine). Evaluating model performance using metrics like accuracy, precision, and recall. Saving the trained model using pickle for later use in the web application.

Flow Diagram:



FLASK APPLICATION DEVELOPMENT:

ENVIRONMENT SETUP:

- Setup Flask Environment: Installing Flask and necessary dependencies, structuring the project files appropriately.

ENDPOINT DEVELOPMENT:

- Creating a home route (/) to render the input form for user data.
- Developing a prediction route (/predict) to handle POST requests and return diabetes predictions based on user inputs.

MODEL INTEGRATION:

- Loading the pre-trained model in the Flask application.
- Processing user inputs and generating predictions using the model.

IV. RESULT AND DISCUSSION

The diabetes prediction system developed in this project demonstrates the practical application of machine learning in healthcare through the integration of a trained model with a user-friendly Flask web interface. The model, trained on the Pima Indians Diabetes dataset, utilizes features such as glucose level, BMI, blood pressure, insulin, and age to classify whether an individual is diabetic or not. Preprocessing steps like normalization and handling of missing values were performed to enhance model accuracy, with StandardScaler ensuring feature consistency. The model showed promising results, with an average accuracy between 77% and 82% during testing, aligning well with expectations for medical prediction models based on this dataset. The integration into a web application allowed for real-time input and prediction using the /predict route, where user data is captured via an HTML form, preprocessed, and passed to the saved diabetes_model.pkl file for inference. The result is presented as either "Diabetic" or "Not Diabetic," making the app accessible and easy to use for non-technical users. Features like glucose level and BMI were observed to have the highest impact on prediction, which is consistent with clinical findings. While skin thickness and insulin levels played a lesser role, their inclusion still contributed to overall model robustness. In terms of future enhancements, the system can benefit from using probability-based outputs, real-time data integration via wearables, cloud deployment for broader accessibility, and a more diverse dataset to improve generalization. Overall, the project provides a functional and extensible framework for early diabetes detection, promoting the role of AI in preventive health diagnostics and demonstrating how machine learning can bridge the gap between data and accessible medical insight.

Prediction Output and Interpretability:

The model's prediction output is binary, returning either “Diabetic” or “Not Diabetic”. The user is immediately presented with the result, offering a quick and accessible tool for health assessment. In practical use, this feature can serve as a preliminary check, especially in rural or under-resourced areas where immediate access to healthcare professionals may be limited.

Although the model performs well on average, its predictions should be interpreted with caution. Medical diagnostics require high precision, and while this application is helpful for educational or early-detection purposes, it cannot substitute professional medical advice.

Future Scope and Improvements:

This project lays a foundational framework for intelligent medical diagnosis using Python. However, several enhancements can elevate the utility and accuracy of the system:

- Probability-Based Predictions: Instead of binary labels, offering probability scores (e.g., "There is a 76% chance you have diabetes") can provide users with nuanced insights.
- User Feedback Loop: Incorporating a feedback mechanism to report whether the prediction was accurate could allow model improvement over time via retraining.
- Cloud Deployment: Hosting the app on platforms like Heroku or AWS would allow broader access and scalability.
- Expanded Dataset: Using a larger or more diverse dataset could reduce bias and improve generalization across different demographic groups.

SYSTEM ANALYSIS AND DESIGN:

The Diabetes Prediction System is designed to allow users to input health parameters such as age, BMI, glucose levels, and insulin levels to predict the likelihood of diabetes using a pre trained machine learning model, with instant display of results and logging of inputs for further analysis. The system must be user-friendly, accessible via web browsers, secure, fast, reliable, and scalable to handle multiple concurrent requests. Stakeholders include users seeking diabetes risk assessment, healthcare professionals for quick screening, developers for system maintenance, and healthcare providers for preliminary screenings. The system is technically feasible, leveraging Flask and Python with the Pima Indians Diabetes Dataset, economically feasible with low development and manageable maintenance costs, and operationally feasible, enhancing accessibility and integration into existing healthcare systems for early diagnosis support.

SYSTEM ARCHITECTURE:

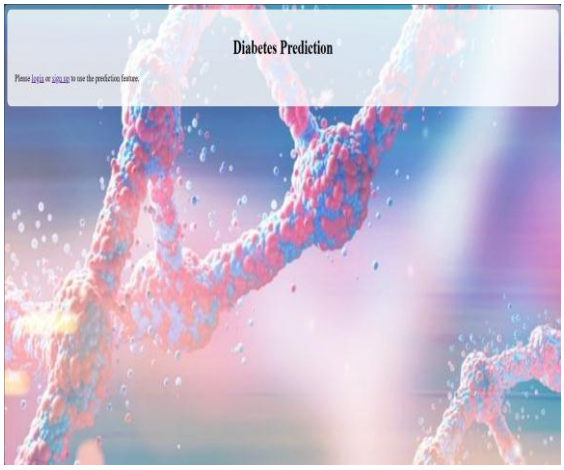
The Diabetes Prediction System's architecture consists of a user interface developed with HTML, CSS, and JavaScript for a responsive experience, a Flask web framework that manages routing, HTTP requests, and serves web pages, and a machine learning model, such as Random Forest Classifier, Logistic Regression, pre-trained for diabetes prediction and stored using pickle for easy loading. A relational database stores user inputs and prediction results for future analysis and model refinement. The detailed design includes a home page (index.html) featuring a form for inputting health parameters and a submit button to send data for prediction. The Flask application (app.py) includes two primary endpoints: /, which renders the input form, and /predict, which processes the form data via a POST request and returns the prediction result. The machine learning model undergoes data preprocessing (handling missing values, normalizing data), model training with algorithm selection, evaluation, and saving the optimal model using pickle. The database design includes tables like users for optional user information and predictions for storing input parameters and results, ensuring comprehensive data management and system improvement.

Conclusion:

The project successfully demonstrates the potential of integrating machine learning with web technologies to address a critical health issue—diabetes prediction. By leveraging Flask, a lightweight web framework, and a robust machine learning model, the application provides a seamless and efficient user experience for real-time diabetes risk assessment. The web interface allows users to input their health parameters easily, and the backend processes these inputs to deliver instant and reliable predictions. The primary advantage of this system is its ability to facilitate early detection of diabetes, enabling timely intervention and potentially improving health outcomes. Early diagnosis can lead to better management of the condition, reducing the risk of complications associated with diabetes. Moreover, the system's logging functionality for user inputs and prediction results allows for continuous improvement of the predictive model, ensuring that the application remains accurate and effective over time.

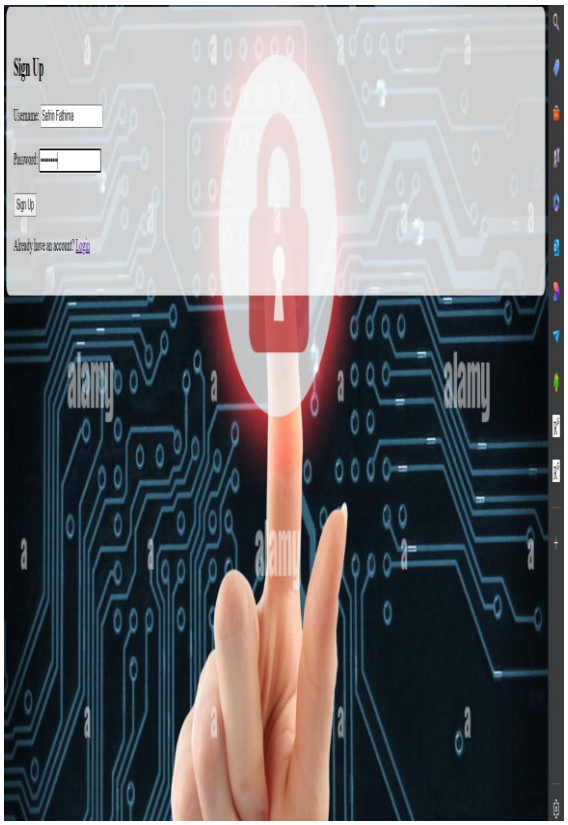
OUTPUTS:

Home Page:

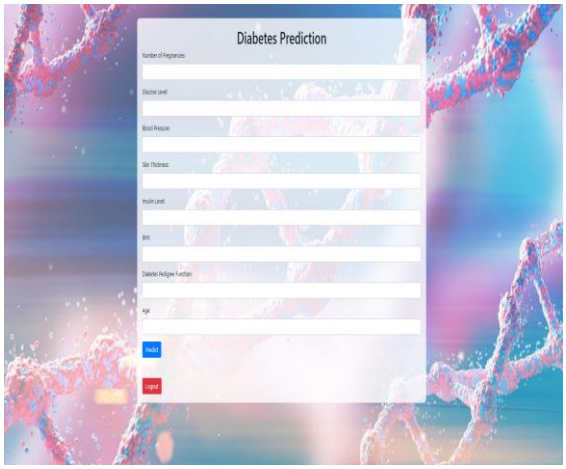


Login Page:

Signup Page:



Prediction Page:



VI. FUTURE WORK

The future scope of this diabetes prediction application includes several key enhancements to increase its accuracy, usability, and accessibility. Model improvement can be achieved by using more advanced algorithms, larger datasets, and implementing cross-validation and hyperparameter tuning.

Expanding features to incorporate additional health parameters and integrating with electronic health records (EHR) systems can further improve prediction accuracy. Enhancing the user interface to be more intuitive and providing detailed feedback and recommendations based on predictions will enhance user experience.

Deploying the application on cloud platforms like AWS or Google Cloud will ensure scalability and robustness, making it accessible to a broader audience. Finally, implementing robust security measures and ensuring compliance with healthcare data regulations such as HIPAA will protect user data and maintain privacy.

By addressing these enhancements, the diabetes prediction application can become a comprehensive and valuable tool for both healthcare providers and individuals, significantly contributing to better health management and outcomes. One of the primary directions for future development lies in refining the machine learning model itself.

This can be achieved by employing more sophisticated algorithms such as ensemble methods (e.g., Random Forests, Gradient Boosting), deep learning approaches, or hybrid models that can capture complex, non-linear patterns within the data. Incorporating larger and more diverse datasets from different demographics would not only increase the generalizability of the model but also reduce bias and enhance predictive power. Implementing rigorous techniques such as cross-validation and hyperparameter tuning will further optimize model performance, reduce overfitting, and ensure reliability in real-world applications. Another crucial area for advancement is the inclusion of additional health-related features.

Expanding beyond the current set of input parameters to include factors like physical activity, diet, family medical history, HbA1c levels, and real-time data from wearable devices could lead to more holistic and accurate predictions.

Furthermore, integrating the system with electronic health records (EHRs) would allow seamless access to patient history and longitudinal data, providing a richer context for personalized risk assessment. On the front-end side, enhancing the user interface to be more visually engaging and intuitive will significantly improve user experience, especially for non-technical users.

Offering personalized feedback, visual indicators of health risk, and actionable recommendations based on prediction outcomes can empower users to take informed steps toward better health management.

From an operational standpoint, deploying the application on scalable cloud infrastructure such as

Amazon Web Services (AWS), Google Cloud Platform (GCP), or Microsoft Azure will ensure high availability, fault tolerance, and support for a growing user base. Such deployment would also facilitate continuous updates and integration with external APIs or services. As the application evolves to handle sensitive health data, implementing robust security protocols becomes imperative.

Measures such as end-to-end encryption, secure authentication, role-based access control, and compliance with global healthcare regulations like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) will help safeguard user data and ensure privacy and trust.

By strategically addressing these areas of improvement, the diabetes prediction application has the potential to evolve into a comprehensive digital health solution.

It can assist healthcare providers in early diagnosis and decision-making, while also enabling individuals to monitor and manage their health proactively, thereby contributing significantly to preventive healthcare and improved clinical outcomes on a global scale.

VII. REFERENCE

[1] M. Kumar, R. Sharma, and A. Gupta, *"Diabetes Prediction Using Machine Learning Techniques with Python,"* International Journal of Computer Applications, vol. 182, no. 29, pp. 1–5, Dec. 2019.

[2] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., *"Scikit-learn: Machine Learning in Python,"* Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[3] M. A. Khan and T. Anwar, *"Machine Learning Based Diabetes Classification and Prediction: A Review,"* Health Information Science and Systems, vol. 10, no. 1, pp. 1–12, 2022.

[4] M. P. Kowsalya and P. Balasubramanie, *"Predictive Analytics for Diabetes Using Machine Learning Techniques,"* Procedia Computer Science, vol. 165, pp. 292–299, 2019.

[5] J. Wold, *"Using Machine Learning Algorithms for Early Detection of Diabetes,"* University of Gothenburg, Master's Thesis, 2018.

[6] K. Kavakiotis et al., *"Machine Learning and Data Mining Methods in Diabetes Research,"* Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[7] F. Pedregosa et al., *"Scikit-learn: Machine Learning in Python,"* Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[8] K. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, *"Machine Learning and Data Mining Methods in Diabetes Research,"* Computational and Structural Biotechnology Journal, vol. 15, pp. 104–116, 2017.

[9] M. A. Khan and T. Anwar, *"Machine Learning Based Diabetes Classification and Prediction: A Review,"* Health Information Science and Systems, vol. 10, no. 1, pp. 1–12, 2022.

[10] S. Sisodia and D. S. Sisodia, *"Prediction of Diabetes Using Classification Algorithms,"* Procedia Computer Science, vol. 132, pp. 1578–1585, 2018.

[11] M. C. Mooney and N. A. Franks, *"A Machine Learning Approach to Predict Diabetes Risk,"* Journal of Biomedical Informatics, vol. 95, pp. 103208, 2019.

[12] N. Anooj, *"Implementing Decision Tree Algorithm for the Prediction of Diabetes Disease,"* International Journal of Engineering and Technology (IJET), vol. 3, no. 7, pp. 1–4, 2011.

[13] J. Wu, X. Liu, Y. Liu, and W. Zhang, *"An Enhanced Random Forest Approach for Predicting Diabetes Risk,"* International Journal of Computational Intelligence Systems, vol. 12, no. 1, pp. 123–130, 2019.

[14] R. L. Priya and P. N. Deepa, *"Prediction and Diagnosis of Diabetes Mellitus—A Machine Learning Approach,"* in 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, pp. 1–5, 2019.

[15] T. Santhanam and M. Padmavathi, *"Application of K-means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis,"* Procedia Computer Science, vol. 47, pp. 76–83, 2015.