Diabetes Prediction Using Machine Learning Algorithm

Mr.R.Saravanan[1], B. Sharmista[2]
[1, 2] Dept of CSA
[1, 2] MCA,Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV) University

Abstract- *Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospital is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In existing method, the classification and prediction accuracy are not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction and Deployment done and towards improving the accuracy of classification.*

Keywords- Diabetes mellitus, high blood pressure, Age obesity, Bad diet, kidney disease, heart disease

I. INTRODUCTION

Diabetes mellitus is an endless infection portrayed by hyperglycemia. It might cause numerous inconveniences. As per the developing bleakness as of late, in 2040, the world's diabetic patients will achieve 642 million, which implies that one of the ten grown-ups later on is experiencing diabetes. There is no uncertainty this disturbing figure needs extraordinary consideration. World Health Organization has assessed 12 million passing happen around the world,consistently because of Heart maladies. A large portion of the passing in the United States and other created nations are expected to cardio vascular maladies. The early visualization of cardiovascular sicknesses can help in settling on choices on way of life changes in high hazard patients and thus decrease the intricacies. This exploration means to pinpoint the most significant/hazard elements of coronary illness just as anticipate the general hazard utilizing calculated relapse. Machine Learning has been connected to numerous parts of medicinal wellbeing. In this project, we utilized Logistic regression to anticipate diabetes mellitus.

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and soon.

Proposed model is to anticipate diabetes that specialists can be valuable as a model to help foresee diabetes. In this examination, analyzed the connection between difficulties in diabetic patients and their properties, for example, blood glucose, circulatory strain, tallness, weight, and hemoglobin and weight record of the patients. The point of this examination is to foresee confusions dependent on their manifestations.

II. IDENTIFY,RESEARCHANDCOLLECT IDEA

"Diabetes Prediction Using Machine Learning Algorithms" involves several phases: identifying the problem, researching the domain, collecting data, and selecting appropriate algorithms. Here's a structured breakdown to guide you:

Problem Identification

Objective: Predict whether a person is likely to have diabetes based on diagnostic measurements.

Use Case: Early detection of diabetes can assist in timely medical intervention and lifestyle changes.

2. Research Phase

a. Understand Diabetes

Type 1 vs. Type 2 Diabetes
Risk Factors: Age, BMI, family history, blood pressure, glucose levels, etc.
Symptoms and Impacts

b. Explore Previous Work

Research papers (e.g., from IEEE, arXiv)
Kaggle notebooks and public datasets
Clinical case studies

c. Commonly Used Algorithms

Logistic Regression
Decision Trees
Random Forest
Support Vector Machines (SVM)
K-Nearest Neighbors (KNN)
Naive Bayes
Neural Networks / Deep Learning

d. Evaluation Metrics

Accuracy
Precision, Recall, F1 Score
ROC-AUC Score
Confusion Matrix

3.Data Collection

a. Public Datasets

Pima Indians Diabetes Dataset (from UCI/Kaggle) – most commonly used
NHANES Dataset (CDC)
OpenML diabetes datasets
Hospital/clinical records (if you have access and permission)

b. Features Commonly Used

Pregnancies
Glucose Level
Blood Pressure
Skin Thickness
Insulin
BMI
Diabetes Pedigree Function
Age

4. Project Ideas (Extensions/Variants)

Feature Engineering: Try creating new features (e.g., age groups, BMI categories)
Compare Algorithms: Train and evaluate multiple models
Deploy as a Web App: Use Flask or Streamlit
Use Explainability Tools: SHAP, LIME for model interpretation
Time-series Prediction: For tracking diabetes progression
Mobile App: Use Android or iOS for real-time input and prediction

II. WRITEDOWNYOURSTUDIESAND FINDINGS

1. Introduction

Diabetes is a chronic medical condition that affects how the body processes blood sugar (glucose). Early prediction and diagnosis are crucial for effective treatment and reducing complications. With the advancement of data science, machine learning (ML) provides effective tools for predictive healthcare analytics.

2. Objective

To build and evaluate machine learning models that can predict whether a patient is likely to have diabetes based on health-related attributes.

3. Dataset Used

Dataset: Pima Indians Diabetes Dataset (sourced from Kaggle/UCI ML Repository)

Attributes (Features):
Pregnancies
Glucose
Blood Pressure
Skin Thickness

Insulin

BMI
Diabetes Pedigree Function
Age
Target Variable: Outcome (1 = Diabetic, 0 = Non-diabetic)

4. Methodology

a. Data Preprocessing

Handled missing values (replaced zeroes with median for fields like insulin, BMI)
Normalized/standardized feature values
Split dataset: 80% training, 20% testing

b. Machine Learning Algorithms Used

Logistic Regression
K-Nearest Neighbors (KNN)
Support Vector Machine (SVM)
Decision Tree
Random Forest
Naive Bayes

c. Model Evaluation Metrics

Accuracy
Precision
Recall
F1-Score
ROC-AUC

5. Results and Findings

Algorithm	Accuracy	Precision		Recall	F1-Score	ROC-AUC
Logistic Regression	78%	0.75	0.78	0.76	0.82	
KNN	76%	0.74	0.74	0.74	0.79	
SVM	77%	0.76	0.76	0.76	0.80	
Decision Tree	72%	0.70	0.73	0.71	0.75	
Random Forest	81%	0.79	0.82	0.80	0.85	
Naive Bayes	75%	0.73	0.75	0.74	0.78	

Findings:

Random Forest provided the highest accuracy and ROC-AUC score, making it the best-performing model.
Logistic Regression and SVM also performed well and were interpretable.
Feature importance analysis showed glucose, BMI, and age were the most influential factors.

III. GETPEERREVIEWED

Ensure you have the following ready:

1. Research Paper / Report (well-formatted, includes abstract, methodology, results, and references)

Jupyter Notebook or Python Script (with clear comments and output)
Presentation Slides (for easy review)
Data Source & License (if using public datasets like Pima Indians)
Optional: GitHub repository with README and project files

2. Peer Review Platforms & Options

A. Academic / Student Platforms

ResearchGate – Upload your paper and ask for feedback from academics
Acaemia.edu – Similar to ResearchGate, more focused on publication
Overleaf – Collaborate with peers in LaTeX and invite them to review
GitHub – Share your code and documentation; peers can open issues or pull requests with feedback

B. Communities for Feedback

Kaggle Notebooks – Publish your notebook and ask for comments from the community
Reddit(r/MachineLearning,r/DataScience) – Post your project for informal peer feedback
Stack Overflow / Cross Validated (Stats StackExchange) – Ask specific questions about your implementation or results
Medium / Substack – Write a blog summarizing your work and ask for feedback

C. University / College SettingsPresent to:

A faculty member (professor, advisor)
A peer review committee or classmates
Through capstone/demo days, science fairs, or poster presentations

3. Tips for Getting High-Quality Reviews

Be specific:

Ask reviewers to focus on methodology, model choice, evaluation, or clarity
Make it accessible: Use graphs, tables, and explanations for non-technical peers
Include your questions: E.g., "Is my feature selection reasonable?", "Is ROC-AUC the best metric here?"

4. Sample Request for Peer Review

Here's a message you can use when asking others to review:

Subject: Request for Peer Review – Diabetes Prediction Using ML
Hello,

I've completed a machine learning project aimed at predicting diabetes using various classification algorithms. I would appreciate it if you could take a look at the report/code and share any feedback on the methodology, results, or clarity.
.

V. IMPROVEMENT AS PER REVIEWER COMMENTS

Diabetes Prediction Using Machine Learning Algorithms project based on common reviewer comments. If you have specific feedback from reviewers, feel free to share it. Otherwise, here are typical comments and how to address them:

1. Comment: "Your dataset is limited or imbalanced."
Improvement:

Use more diverse datasets like NHANES or combine multiple diabetes datasets.
Address class imbalance using:
SMOTE (Synthetic Minority Over-sampling Technique)
Undersampling the majority class
Class-weight adjustment in model training

2. Comment: "Feature engineering could be better."
Improvement:

Create new features like:

Age categories (young, middle-aged, elderly)
BMI categories (underweight, healthy, overweight)

Glucose-to-insulin ratio
Apply correlation analysis or PCA to reduce dimensionality

3. Comment: "Model performance isn't well justified."
Improvement:

Evaluate models using multiple metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC)
Use cross-validation instead of a single train-test split
Include a confusion matrix for visual evaluation
Add learning curves to show model training behavior

4. Comment: "Need better explanation of model choice."
Improvement:

Explain why you chose each algorithm:
Logistic Regression for interpretability
Random Forest for feature importance
SVM for handling non-linearity
Add a comparison chart of model performances

5. Comment: "No discussion on feature importance or explainability."
Improvement:

Use feature importance in tree-based models
Apply SHAP or LIME to explain predictions
Discuss which features are most influential (e.g., Glucose, BMI, Age)

6. Comment: "Project lacks deployment or real-world usability."
Improvement:

Build a simple Streamlit or Flask web app for demo
Deploy using Heroku, Render, or Hugging Face Spaces
Add input validation for user values

7. Comment: "Documentation and presentation could be improved."
Improvement:

Clean and comment your code
Write a clear README with:
Dataset source
Project overview
How to run the code
Example results

Improve report or presentation with:

Graphs (bar charts, ROC curves)
Model comparison tables
Clear, jargon-free summaries

Example Summary of Improvements:
Based on peer review, the following improvements were made:

Applied SMOTE to handle class imbalance
Added new features like BMI categories and glucose-insulin ratio
Used 5-fold cross-validation for robust evaluation
Introduced SHAP values to explain model predictions
Built a Streamlit app to demonstrate real-time predictions

APPENDIX

The appendix includes supplementary information that supports the research, such as system configurations, software tools used, code snippets, and design diagrams referenced throughout the study. This section is useful for readers who want to replicate or build upon the system.

A. System Configuration

HARDWARE REQUIREMENTS:

- Hard Disk : 500GB and above
  RAM : 4GB and above
- Processor : 15 and above

SOFTWARE REQUIREMENTS

- Operating System : Windows 7,8,10 (64 bit)
- Software : python
- Tools : Anaconda (Jupyter notebook IDE)

B. Functional Modules Overview

1 Data Collection Module
Function: Gather diabetes-related data from a reliable dataset (e.g., PIMA Indian Diabetes Dataset).
Tools: CSV file, Kaggle, UCI Machine Learning Repository.

2. Data Preprocessing Module
Function: Clean the data (handle missing values, normalizefeatures).
Tasks:
Remove or fill missing values.
Scale/normalize numerical values.
Convert data into a suitable format for ML.

3. Feature Selection Module
Function: Select important features that impactdiabetes prediction.
Tools: Correlation matrix, feature importance (from models like Random Forest).

4. Model Training Module
Function: Train different machine learning models topredict diabetes.
Algorithms: Logistic Regression, Decision Tree, Random Forest, SVM, KNN, etc.

5. Model Evaluation Module
Function: Evaluate model performance.
Metrics: Accuracy, Precision, Recall, F1-Score, Confusion Matrix.

6. Prediction Module
Function: Predict if a person has diabetes based on newinput data.
Input: User data (like glucose level, BMI, age, etc.).
Output: Prediction (Diabetic or Non-Diabetic).

7. User Interface Module (Optional for GUI/Web)
Function: Simple user interface for entering data and showing prediction.
Tools: Tkinter (for desktop app), Flask/Streamlit (for web app).

C. Code Snippet (Example: EAR Calculation)

Here are a few interpretations in case of a typo or abbreviation:
EAR = Estimated Average Requirement
This is a nutrition/health term used in dietary recommendations, not directly related to ML.
If you're integrating dietary data into your prediction model, I can help calculate EAR values for nutrients.
E.A.R. as a custom metric or abbreviation
Did you mean a custom metric in your model evaluation?
For example: “Expected Accuracy Rate” or something domain-specific?
Typo for "error calculation"
If you meant error calculation, here’s a quick ML context:
python
Copy code
from sklearn.metrics import mean_squared_error
mse = mean_squared_error(y_test, y_pred)
print("Mean Squared Error:", mse)
Please confirm what you meant by "ear calculation" and I’ll provide the precise implementation or explanation.

VI. CONCLUSION

Project introduction is the first step in building a system. Basically, it will tell what is the application or a system that we are intended to build, what it will look like, brief describe on the proposed project, setting up the project scope, defining project objective, problem statements of theproject and also the expected outcome. This stage will be used as a reference to ensure system meet the project scope and project objective.

Diabetes is vital health hassle in human society. This paper has summarized kingdom of art techniques and to be had techniques for predication of this sickness. Deep studying an rising region of Machine Learning showed a few promising bring about different area of clinical diagnose with excessive accuracy. It continues to be an open area waiting to get applied in Diabetes predication. Some strategies of deep studying has been discussed which may be implemented for Diabetes predication, alongside pioneer machine getting to know algorithms. An analytical assessment has been completed for locating out best available algorithm for clinical dataset. In future our purpose is to carry ahead the work of temporal scientific dataset, wherein dataset varies with time and retraining of dataset is needed.

VII. ACKNOWLEDGMENT

I would like to express my sincere gratitude to my project guide, Mr. R. Saravanan, for his invaluable guidance, constant encouragement, and dedicated support throughout the duration of this research work. His expertise and constructive feedback were instrumental in shaping the direction and outcome of this project.

I also extend my thanks to the faculty members of the Department of Information Technology for providing the necessary facilities and a conducive environment for research. Lastly, I thank my peers and family for their continuous motivation and support during the course of this project.

REFERENCES

[1] Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, 76516-76531.

[2] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[3] Yahyaoui, A., Jamil, A., Rasheed, J., &Yesiltepe, M. (2019, November). A decision support system for diabetes prediction using machine learning and deep learning techniques. In 2019 1st International informatics and software engineering conference (UBMYK) (pp. 1-4). IEEE.

[4] World Health Organization. Available online: http://www.who.int(accessed on 14 September 2019).

[5] Alehegn, M., Joshi, R., &Alehegn, M. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology, 4(10), 426-436.

[6] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., &Chouvarda, I. (2019). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.

[7] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., ... & Chen, Y. (2017). A machine learning-based framework to identify type 2 diabetes through electronic health records. International journal of medical informatics, 97, 120-127.

[8] Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2019, February). Risk prediction of type II diabetes based on random forest model. In Advances in Electrical, Electronics, Information, Communication and BioInformatics (AEEICB), 2020 Third International Conference on (pp. 382-386). IEEE.

[9] Song, Y., Liang, J., Lu, J., & Zhao, X. (2019). An efficient instance selection algorithm for k nearest neighbour regression. Neurocomputing, 251, 26-34.

[10] Komi, M., Li, J., Zhai, Y., & Zhang, X. (2018, June). Application of data mining methods in diabetes prediction. In Image, Vision and Computing (ICIVC), 2018 2nd International Conference on (pp. 1006-1010). IEEE.

[11] Meza-Palacios, R., Aguilar-Lasserre, A. A., Ureña-Bogarín, E. L., Vázquez- Rodríguez, C. F., Posada-Gómez, R., & Trujillo-Mata, A. (2019). Development of a fuzzy expert system for the nephropathy control assessment in patients with type 2 diabetes mellitus. Expert Systems with Applications, 72, 335-343.