



DIABETES / HEART DISEASE PREDICTION USING MACHINE LEARNING

Mrs.Gandhi.R.R¹, Taslima Shaikh²

¹ Faculty Guide & Supervisor | ² Dattakala Group Of Institutions Faculty Of Engineering Swami - Chincholi tal. Daund dist. Pune, India

Corresponding author: shaikhtaslima830@gmail.com

ARTICLE ID IJREK-2026-00141	ARTICLE TYPE Research Article	RESEARCH AREA Machine Learning & Deep Learning	PUBLICATION Volume 1, Issue Issue 1 (2026)
RECEIVED 21 October 2026	REVISED 25 October 2026	ACCEPTED 22 October 2026	PUBLISHED 28 October 2026

Abstract

Diabetes and heart disease are important health conditions for which assessment of risk factors may support timely discussion with healthcare professionals. This paper proposes a machine-learning-based prediction system that uses selected attributes from an appropriate, documented dataset to classify records into dataset-defined categories for diabetes or heart disease. Possible algorithms include Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine. The workflow includes data cleaning, preprocessing, model training, testing, and presentation through a simple interface. The system is intended for educational and research purposes only. Its output is not a diagnosis, does not replace clinical assessment, and must not be used to make treatment decisions. This is a proposed design and evaluation plan; no measured accuracy or clinical effectiveness is claimed.

Keywords: Machine Learning; Diabetes Prediction; Heart Disease Prediction; Classification; Healthcare; Risk Estimation; Data Preprocessing. Research-paper draft for academic learning. Dataset details; implementation; evaluation results; and references must be verified and updated to match the completed project before submission.

2

I. INTRODUCTION

A. Background / Context Diabetes affects how the body regulates blood glucose, while heart disease includes conditions affecting the heart and blood vessels. Health-related risk may be associated with factors such as age, family history, blood pressure, glucose measurements, cholesterol, and lifestyle. Machine learning can identify patterns in labelled datasets and estimate dataset-defined outcomes. A model's predictions depend on the quality and representativeness of its data and the validation method used. B. Problem Statement Assessing health risk requires multiple factors and professional judgement. A student project can demonstrate how machine-learning algorithms classify records using selected attributes, but an unvalidated model may make incorrect predictions or perform poorly for groups not represented in training data. A transparent prototype is needed that evaluates its model carefully and explains its limitations. C. Aim and Objectives • Develop a prototype for diabetes or heart-disease classification using a suitable labelled dataset. • Preprocess health data and handle missing or inconsistent values. • Train and compare appropriate classification algorithms. • Evaluate the model using precision, recall, F1-score, and a confusion matrix. • Display results with a clear warning that the output is not a medical diagnosis. • Document dataset limitations, privacy concerns, and possible sources of bias. D. Significance and Scope The project demonstrates machine learning for healthcare data analysis. Diabetes and heart disease should generally be implemented as separate prediction tasks because their datasets, features, and target labels differ. This is a student research prototype, not a clinically approved product. It should not advise users to start, stop, or change medication or treatment.

II. LITERATURE REVIEW / RELATED WORK

Healthcare prediction research has examined statistical methods and machine-learning classifiers for estimating disease-related outcomes from structured data. Logistic Regression provides an interpretable baseline, while Decision Trees and Random Forests can model non-linear relationships. Support Vector Machines may also be evaluated with suitable preprocessing. Performance varies with dataset source, population, feature definitions, missing data, and validation method. Responsible

evaluation should use held-out test data or suitable cross-validation, avoid data leakage, and report more than accuracy alone. False negatives and false positives may have different consequences in health-related classification, so precision, recall, and the confusion matrix should be examined. A high score on a small or unrepresentative dataset does not prove clinical suitability. Research Gap A student prototype should identify its dataset and target label, explain preprocessing and model selection, report actual test results, and present predictions as estimates rather than facts. Diabetes and heart disease require separate evaluation unless a justified multi-output design and suitable data are available.

III. PROPOSED SYSTEM

TABLE I

Component	Main responsibility
User interface	Collects supported attributes and displays the output and disclaimer.
Data validation	Checks required fields, ranges, and missing or inconsistent values.
Preprocessing	Handles missing data and scales or encodes features when required.
ML classifier	Predicts the dataset-defined target class using a trained model.
Evaluation module	Measures performance on data not used for training.
Results screen	Displays the output as an estimate and recommends professional advice for health concerns.

IV. METHODOLOGY

A. Dataset and Input Features Use a public, documented dataset and record its source, license, target label, population, and limitations. Example diabetes datasets may contain age, body-mass index, glucose-related measurements, blood pressure, or family-history indicators, depending on the dataset. Heart-disease datasets may contain age, sex, chest-pain category, resting blood pressure, cholesterol, and other recorded attributes. Use only fields actually present in the chosen dataset. These examples are not a clinical checklist. B. Data Preprocessing • Inspect data types, ranges, duplicates, and target labels. • Handle missing values using a documented

method; do not silently invent values. • Encode categorical variables and scale numerical variables when needed. • Split training and test data before fitting preprocessing steps to avoid data leakage. • Check class imbalance and select suitable evaluation methods. • Remove direct identifiers and follow dataset terms and privacy requirements. C. Machine Learning Algorithms Compare Logistic Regression as an interpretable baseline with Decision Tree, Random Forest, or Support Vector Machine models. Train and tune models only on training data. Diabetes and heart disease should be treated as separate tasks unless a justified multi-output design and suitable data are available. Consider performance, interpretability, and limitations rather than choosing a model based only on one score. D. Prediction Output The interface may show the predicted dataset-defined class with a notice that the result is not a diagnosis. If probabilities are displayed, explain that they are model outputs; do not present them as a medically confirmed personal probability unless the model has been appropriately validated and calibrated. E. Ethical and Safety Considerations • State that the prototype is for academic demonstration, not clinical decision-making. • Do not advise users to change medication, treatment, or care based on the output. • Protect health-related information and avoid collecting unnecessary personal data. • Evaluate potential bias and group-level performance where data permits. • Use qualified clinical oversight for future research involving real patients. • Document uncertainty, false positives, false negatives, and dataset limitations.

V. SYSTEM DESIGN AND IMPLEMENTATION PLAN

TABLE II	
Layer	Possible student-level choice
Programming language	Python.
Data processing	pandas and NumPy.
Machine learning	scikit-learn: Logistic Regression, Decision Tree, Random Forest, or SVM.
Visualization	Matplotlib or a simple dashboard.
Interface	Streamlit, Flask, or a simple web interface.
Dataset	A documented public dataset suitable for the chosen task.

VI. RESULTS AND EVALUATION PLAN

TABLE III		
Test ID	Test case	Expected result
T1	Valid record with all required attributes.	System validates input and returns a model output.
T2	Required value is missing.	System handles missing input safely or asks for correction.
T3	Value is outside allowed input range.	System shows a clear validation message.
T4	Model is evaluated on held-out test data.	Metrics are calculated from test predictions.
T5	Input is poorly represented in training data.	Limitations are explained; output is not presented as certain.
T6	User views the prediction screen.	Non-diagnostic disclaimer is visible.

VII. ADVANTAGES AND LIMITATIONS

TABLE IV	
Advantages	Limitations
Demonstrates practical ML on structured health data.	Predictions depend on data quality and representativeness.
Allows comparison of several algorithms.	Models can make false positive or false negative predictions.
Can provide quick outputs for educational test records.	A student prototype is not clinically validated.
Supports learning about preprocessing and evaluation.	Datasets may use different labels and input attributes.
Can be expanded with improved interface and validation.	Results may not generalize to other populations or settings.

VIII. APPLICATIONS

- Academic demonstrations of machine-learning classification.
- Research practice using public, de-identified health datasets.
- Educational dashboards for explaining model evaluation.
- Prototype comparisons on diabetes or heart-disease datasets.

IX. FUTURE SCOPE

- Evaluate models on larger and more representative datasets.
- Improve interpretability with clear feature explanations.
- Assess calibration and subgroup performance where appropriate.
- Develop secure interfaces and stronger data governance.
- Conduct formal validation with qualified healthcare researchers before any clinical use.
- Compare models across datasets with reproducible evaluation methods.

X. CONCLUSION

This paper proposes a machine-learning-based system for studying diabetes or heart-disease classification from structured data. The workflow includes dataset selection, preprocessing, model training, evaluation, and an interface that presents estimates with appropriate warnings. Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine models can be compared using precision, recall, F1-score, and confusion matrices. This proposed student project claims no measured performance or clinical benefit. Future development should prioritize validation, privacy, fairness, transparency, and qualified medical oversight.

REFERENCES

[1] World Health Organization, “Diabetes,” WHO fact sheet and topic resources. Verify the current title, update date, and access date before submission.

[2] World Health Organization, “Cardiovascular diseases (CVDs),” WHO fact sheet and topic resources. Verify current details before submission.

[3] Scikit-learn developers, “Supervised learning” and “Model evaluation: quantifying the quality of predictions,” Scikit-learn User Guide. Verify current documentation details.

[4] UCI Machine Learning Repository, documentation for the exact diabetes or heart-disease dataset selected. Cite its version, authorship, license, and access date.

[5] Add recent peer-reviewed studies on diabetes prediction, heart disease prediction, model validation, and responsible AI in healthcare. Verify all bibliographic details and DOI. Submission note: This is a research-paper draft for academic learning, not a completed clinical study. Match references to the actual dataset and implementation. Do not invent experimental results or use the model for diagnosis or treatment decisions.

Publication archival note: Official scholarly contribution published under CC BY 4.0 International License. Permanent record on <https://www.ijrek.org> | Article ID: IJREK-2026-00141