

# International Journal Research Publication Analysis

Page: 01-11

---

## DIABETES PREDICTION USING MACHINE LEARNING: AN INTEGRATED ANALYTICS PIPELINE AND INTERACTIVE STREAMLIT APPLICATION

---

**\*Mopuri Shivananda Chary**

---

Under the guidance of MR. P. KRISHNA PRASAD, Internship Guide Chaitanya Bharathi  
Institute of Technology (Autonomous) Gandipet, Hyderabad, Telangana, India.

---

**Article Received: 06 September 2026**

**Article Revised: 26 September 2026**

**Published on: 16 October 2026**

**\*Corresponding Author: Mopuri Shivananda Chary**

Under the guidance of MR. P. KRISHNA PRASAD, Internship Guide  
Chaitanya Bharathi Institute of Technology (Autonomous) Gandipet,  
Hyderabad, Telangana, India.

DOI: <https://doi.org/10.1555/ijrpa.3472>

---

### ABSTRACT

Diabetes is a common condition whose early identification depends on interpreting several health parameters together, yet manual examination of those parameters is slow and offers no automated, data-driven estimate of diabetes risk. This motivates automated, comparable, and accessible prediction tools for healthcare data. This paper presents an end-to-end, layered machine-learning framework and interactive web application for diabetes prediction. The pipeline ingests the Healthcare-Diabetes dataset (2,768 records, 10 columns; 952 diabetic and 1,816 non-diabetic), performs data validation and cleaning in Python and Pandas (duplicate removal, replacement of physiologically invalid zeros with medians, and IQR-based outlier removal), conducts exploratory analysis with Matplotlib and Seaborn, and trains and compares Logistic Regression, Decision Tree, and Random Forest classifiers on an 80/20 split (2,214 training and 554 test samples). The selected Random Forest is deployed in a Streamlit application that accepts eight health parameters and returns a diabetes label with an estimated probability. The empirical results reveal three observations: (i) plasma glucose is the attribute most strongly correlated with the outcome (0.49), ahead of insulin and BMI (0.29 each); (ii) tree-based models substantially outperform Logistic Regression on the held-out fold, with accuracy of 0.996 for Random Forest and 0.989 for Decision Tree against 0.758 for Logistic Regression, whose recall is only 0.539; and (iii) the near-perfect tree-model scores (Random Forest: precision 1.000, recall 0.990, ROC-AUC 0.995; two errors in 554 test cases) are unusually high for this problem and warrant leakage and generalization checks before any

real-world reliance. The framework offers a reproducible, low-cost template for academic healthcare analytics; it is a predictive-support demonstration and not a clinical diagnostic tool.

**INDEX TERMS:** Diabetes Prediction, Machine Learning, Logistic Regression, Decision Tree, Random Forest, Exploratory Data Analysis, Streamlit Web Application, Healthcare Analytics.

## I. INTRODUCTION

### *A. Background of Diabetes Risk Assessment and Health Analytics*

Diabetes is among the most prevalent chronic conditions worldwide [2], [3], and its early identification depends on interpreting several measurements together, such as plasma glucose, blood pressure, skin-fold thickness, insulin, body mass index, family-history indicators, and age. Clinical and research repositories record these parameters for each patient, producing multi-attribute health footprints that machine-learning methods can convert into predicted outcomes [4], [5]. Such predictions, when delivered through an accessible interface, support early awareness of potential diabetes risk.

### *B. Limitations of Manual Assessment*

Traditional diabetes assessment relies on medical examination and manual analysis of individual parameters. Used alone on stored records, this approach shows recurring limitations:

- **Time-Consuming Review:** Analysing several health parameters by hand is slow and does not scale with the number of records.
- **No Automated Prediction:** Manual analysis does not learn from historical healthcare data to estimate the outcome for a new patient.
- **Hidden Feature Interactions:** Relationships among glucose, insulin, BMI, and age are hard to see in raw rows.
- **Limited Accessibility:** Analytical output kept inside notebooks cannot be used by people without programming skills.

### *C. Transition to Layered Analytics Frameworks*

Modern practice separates data loading, cleaning, exploration, modelling, and presentation into distinct, testable stages. Python with Pandas [11] and NumPy [12] provides the data-handling substrate; Matplotlib [13] and Seaborn [14] support exploration; scikit-learn [10] supplies classifiers and evaluation metrics; and Streamlit [15] converts a trained model into an interactive browser application without dedicated front-end engineering.

#### ***D. Research Challenges in Diabetes Prediction***

- **Invalid Zero Values:** Attributes such as glucose, blood pressure, skin thickness, insulin, and BMI cannot physiologically be zero, so recorded zeros behave as hidden missing values.
- **Class Imbalance:** Diabetic cases form a minority (34.39%), so accuracy alone can hide poor detection of diabetic patients and recall must be examined [16].
- **Outliers and Skew:** Extreme values in insulin and other measurements distort distributions and can mislead distance-or linear-model fitting.
- **Evaluation Leakage:** Overlap between training and test records can inflate reported scores [9].
- **Deployment Friction:** Predictions confined to code cannot be used by non-technical users.

#### ***E. Overview of Proposed Solution and Key Contributions***

This research presents an end-to-end framework that cleans and explores a diabetes dataset, trains and compares three classifiers, and exposes the selected model through a Streamlit web application. Contributions are:

- **Reproducible Layered Pipeline:** A decoupled workflow (data, preprocessing, model, and presentation layers) documented step by step.
- **Domain-Aware Cleaning:** Duplicate removal, conversion of invalid zeros to missing values with median imputation, and IQR outlier treatment.
- **Comparative Model Selection:** Logistic Regression, Decision Tree, and Random Forest compared on identical data with five metrics.
- **Empirical Findings:** Glucose, insulin, and BMI identified as the leading correlates of diabetes outcome, with candid analysis of the unusually high tree-model scores.
- **Interactive Prediction Application:** A Streamlit interface returning a label and estimated probability from eight input parameters.

**TABLE I: SUMMARY OF REPRESENTATIVE RELATED WORK.**

| Ref.       | Study Focus                           | Data Modality         | Reported Gap / Limitation                                   |
|------------|---------------------------------------|-----------------------|-------------------------------------------------------------|
| [4]        | Adaptive learning for diabetes onset  | Pima clinical records | Model-centred; no interactive prediction application.       |
| [5]        | ML methods in diabetes research       | Literature survey     | Survey without an end-to-end reproducible pipeline.         |
| [6]<br>[8] | Logistic, tree and forest theory      | General-purpose ML    | Algorithm theory without a healthcare application.          |
| [9]        | Leakage in data mining                | General-purpose ML    | Methodological guidance without a diabetes pipeline.        |
| [1]        | Diabetes prediction internship system | 2,768 patient records | Engineering report; lacked formal peer-reviewed evaluation. |

## II. LITERATURE SURVEY

### *A. Evolution of Diabetes Prediction*

Early computational work by Smith et al. [4] applied an adaptive learning algorithm to forecast the onset of diabetes in a Pima Indian population, establishing the eight-attribute feature set that most diabetes benchmarks still use. Kavakiotis et al. [5] surveyed machine-learning and data-mining methods in diabetes research, showing that classification, particularly with supervised learners, is the dominant approach and that data quality and feature preparation strongly affect results. These studies concentrate on the model rather than on an integrated, accessible application.

### *B. Classification Algorithms*

Logistic Regression [6] models the log-odds of the outcome as a linear function of the inputs and is a standard, interpretable baseline. Decision trees, formalized by Quinlan [7], split data recursively on feature conditions and capture non-linear interactions but may overfit. Random forests, introduced by Breiman [8], average many randomized trees to reduce variance. These families were selected because they represent linear, rule-based, and ensemble learning respectively.

### *C. Evaluation Metrics and Leakage*

Under class imbalance, accuracy alone is insufficient; Saito and Rehmsmeier [16] show that precision-recall measures are more informative than ROC analysis alone. Kaufman et al. [9] warn that information shared between training and evaluation data, for example through repeated or near-identical records, yields optimistic results. These cautions shape the evaluation and limitations in Sections IV and V.

### *D. Research Gap Analysis*

Three gaps recur: (1) **Model-Only Focus**: classifiers are reported without a usable end-user interface; (2) **Documentation Gap**: cleaning and modelling steps are seldom documented reproducibly for learners; and (3) **Evaluation Gap**: very high scores are reported without leakage analysis. This work addresses the first two and explicitly flags the third.

## III. SYSTEM METHODOLOGY & ARCHITECTURE

### *A. Proposed System Overview and Feasibility*

The proposed system replaces manual inspection of health parameters with a structured, layered pipeline. Feasibility was verified across three dimensions. **Economic**: the implementation relies entirely on open-source tooling (Python, scikit-learn, Streamlit) and freely available datasets, incurring no licensing cost. **Technical**: with 2,768 records, training

and evaluation complete quickly on a standard workstation. **Operational:** the Streamlit interface hides the model code, so users only enter values and read a result.

### B. Layered Data Processing Pipeline

The framework has four cooperating layers: (1) *Data Layer*: loads Healthcare-Diabetes.csv, checks missing values and duplicates, and performs exploratory analysis; (2) *Preprocessing Layer*: cleans invalid zeros, imputes medians, removes outliers, and splits the data; (3) *Analytics and Model Layer*: trains and evaluates the three classifiers; and (4) *Presentation Layer*: serves the selected model through a Streamlit application.

#### Algorithm 1: Data Cleansing, Model Training & Evaluation Pipeline

**Input:** Raw CSV Draw(2,768 rows, 10 columns)

**Output:** Cleaned data, trained models, metrics per model

```

1: Load Draw into DataFrame df; check nulls and duplicates
2: df ← df.drop_duplicates()
3: cols ← [Glucose, BloodPressure, SkinThickness, Insulin, BMI]
4: for each c in cols do
5:   replace 0 with NaN in df[c]
6:   df[c].fillna(df[c].median())
7:   remove outliers from df[c] by the IQR rule
8: end for
9: X ← 8 input features; y ← Outcome (Id excluded)
10: split (X, y) into train (2,214) and test (554)
11: for each M in {LogReg, DecisionTree, RandomForest} do
12:   fit M on train; predict on test
13:   compute accuracy, precision, recall, F1, ROC-AUC
14: end for; return metrics; deploy best model M*
  
```

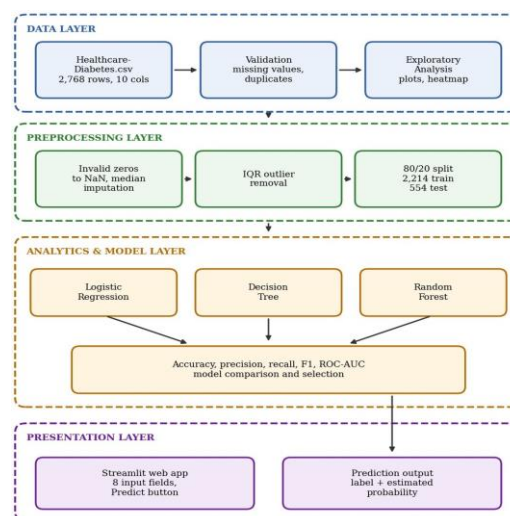


Fig. 1. Layered system architecture of the diabetes prediction framework.

**TABLE II: CORE DATASET SCHEMA AND ANALYTICAL ROLES.**

| Attribute                | Description                   | Role                  |
|--------------------------|-------------------------------|-----------------------|
| Id                       | Record identifier             | Not used for training |
| Pregnancies              | Number of pregnancies         | Input feature         |
| Glucose                  | Plasma glucose level          | Input feature         |
| BloodPressure            | Blood pressure                | Input feature         |
| SkinThickness            | Skin-fold thickness           | Input feature         |
| Insulin                  | Insulin level                 | Input feature         |
| BMI                      | Body mass index               | Input feature         |
| DiabetesPedigreeFunction | Hereditary risk score         | Input feature         |
| Age                      | Age of the patient            | Input feature         |
| Outcome                  | 0 = no diabetes, 1 = diabetes | Target variable       |

### ***C. Data Preprocessing and Feature Preparation***

The Healthcare-Diabetes dataset [17] contains 2,768 records and 10 columns; 1,816 records have Outcome 0 and 952 have Outcome 1. No column contains missing values, and duplicate rows are removed with `drop_duplicates()`. Because zero is not a meaningful value for Glucose, BloodPressure, SkinThickness, Insulin, and BMI, zeros in these columns are replaced with missing values and filled with each column's median, which avoids discarding records and is robust to skew. Box plots identify unusually high or low values, which are removed with the interquartile-range (IQR) rule and re-plotted to verify. The Id column is excluded, and the eight remaining attributes form the input features. Algorithm 1 summarizes the procedure.

### ***D. Classification Modeling***

Three supervised classifiers are trained on the same prepared data, which is divided into 2,214 training and 554 testing samples (about 80/20). **Logistic Regression** serves as a basic linear baseline, the **Decision Tree** predicts through a tree of feature conditions, and the **Random Forest** combines many decision trees for a more robust prediction. Each model is fitted on the training portion and evaluated on unseen test data using accuracy, precision, recall, F1-score, and ROC-AUC.

### ***E. Streamlit Web Application***

The selected Random Forest is integrated into a Streamlit application with input fields for the eight parameters (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age) and a *Predict* button. The entered values are passed to the trained model, and the interface displays either “No Diabetes Predicted” or “Diabetes

Predicted” together with the estimated probability of diabetes.

#### IV. EXPERIMENTAL EVALUATION & RESULTS

##### A. Analytical Baseline Comparison

Benchmarking the proposed pipeline against manual review (Table III) shows that automated, multi-attribute modelling with an interactive interface handles prediction, repeatability, and accessibility that manual examination of parameters cannot.

**TABLE III: FRAMEWORK BENCHMARK VS. MANUAL REVIEW.**

| Dimension          | Manual Review           | Proposed Framework                    |
|--------------------|-------------------------|---------------------------------------|
| Attribute handling | One parameter at a time | All eight features jointly            |
| Prediction         | None automated          | Label plus estimated probability      |
| Data quality       | Ad hoc                  | Zero handling, imputation, IQR filter |
| Evaluation         | None                    | Five metrics on a held-out split      |
| Accessibility      | Requires analyst        | Streamlit form for any user           |

##### B. Dataset Overview and Feature Correlations

The dataset holds 2,768 records, of which 952 (34.39%) are diabetic and 1,816 (65.61%) are not, a moderate class imbalance.

Fig. 2 ranks the correlation of each input attribute with the outcome. Glucose is the strongest correlate (0.49), followed by insulin and BMI (0.29 each), skin thickness (0.25), age (0.24), pregnancies (0.22), blood pressure (0.18), and the pedigree function (0.16); the identifier Id shows essentially none ( $-0.01$ ), as expected. Inter-feature correlations are also notable: skin thickness with BMI (0.64), glucose with insulin (0.57), and pregnancies with age (0.54).

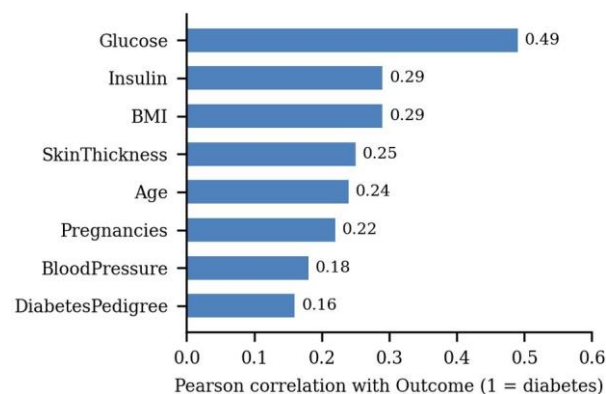


**Fig. 2. Correlation of input attributes with the diabetes outcome.**

##### C. Model Comparison

On the 554-sample test set, the Logistic Regression baseline reaches an accuracy of 0.758 with a recall of only 0.539, meaning it misses nearly half of the diabetic cases. The Decision

Tree and Random Forest score far higher (Table IV, Fig. 3): the Random Forest attains accuracy 0.996, precision 1.000, recall 0.990, F1-score 0.995, and ROC-AUC 0.995, and was chosen for deployment. The Decision Tree is close behind, with accuracy 0.989 and the highest recall (0.995) but lower precision (0.974).

**TABLE IV: MODEL COMPARISON ON THE HELD-OUT TEST SET.**

| Model               | Acc.             | Prec. | Recall | F1    | AUC   |
|---------------------|------------------|-------|--------|-------|-------|
| Logistic Regression | 0.758            | 0.691 | 0.539  | 0.606 | 0.706 |
| Decision Tree       | 0.989            | 0.974 | 0.995  | 0.984 | 0.990 |
| Random Forest       | 0.996 (selected) | 1.000 | 0.990  | 0.995 | 0.995 |

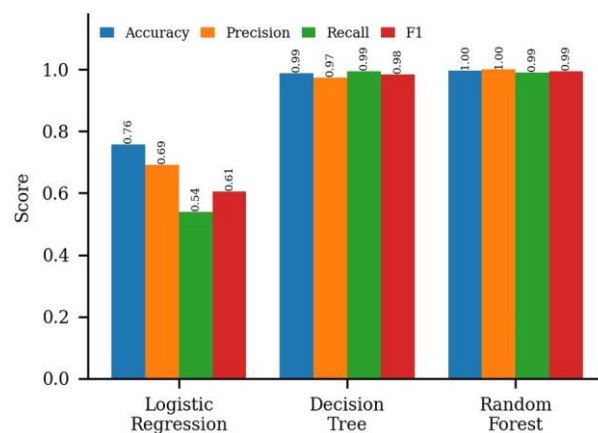


**Fig. 3. Accuracy, precision, recall and F1-score of the three classifiers.**

#### ***D. Random Forest Prediction Results***

The Random Forest classified 363 non-diabetic and 189 diabetic test patients correctly, produced no false positives, and made two false negatives (diabetic patients predicted as non-diabetic), as summarized in Table V. These counts are consistent with the reported precision of 1.000 and recall of 0.990 (189 of 191 diabetic patients).

**TABLE V: RANDOM FOREST TEST-SET CONFUSION COUNTS**

| Actual \ Predicted    | No Diabetes | Diabetes |
|-----------------------|-------------|----------|
| No Diabetes (363 + 0) | 363         | 0        |
| Diabetes (2 + 189)    | 2           | 189      |

#### ***E. Application Output and Verification***

The Streamlit application was exercised with sample inputs. One input (Pregnancies 1, Glucose 120, Blood Pressure 70, Skin

Thickness 20, Insulin 80, BMI 25, Pedigree Function 0.47, Age

30) and other test cases produced outputs of both kinds: a “No Diabetes Predicted” result with

an estimated diabetes probability of 8.50%, and a “Diabetes Predicted” result with an estimated probability of 59.00%. The interface therefore connects the trained model, input form, prediction step, and result display into a single working system.

## V. DISCUSSION & ANALYTICAL IMPLICATIONS

### A. *Practical Operational Contrast*

Compared with manual examination, the framework analyses all eight parameters jointly, produces a quantified prediction with an estimated probability, and exposes it through an interface that non-programmers can use. Comparing three algorithms, rather than relying on a single one, allowed selection on measured evidence.

### B. *Four-Point Analytical Takeaways*

- 1. Clean Domain-Invalid Values:** treating zeros in glucose, blood pressure, skin thickness, insulin, and BMI as missing, and imputing medians, preserves records while removing impossible values.
- 2. Glucose Leads the Signal:** glucose has the highest correlation with the outcome (0.49), followed by insulin and BMI.
- 3. Non-Linear Models Fit Better:** the tree-based models far exceed the linear baseline, whose recall of 0.539 would miss many diabetic patients.
- 4. Treat Near-Perfect Scores with Caution:** accuracy of 0.996 with two errors in 554 cases is unusually high and must be verified before the model is trusted.

### C. *Limitations*

Limitations include: (i) the tree-model scores are near-perfect, which may indicate that near-identical patient records appear in both training and test sets; because each record carries a unique Id, `drop_duplicates()` may not remove repeated clinical records, and the split of 2,214 plus 554 samples equals the original 2,768 records, so the effect of cleaning on the modelled row count should be confirmed; (ii) evaluation uses a single hold-out split with no cross-validation, and hyperparameters and random seeds are not reported; (iii) the dataset is a single, small, static snapshot that may not generalize to other populations; (iv) the application predicts from eight parameters and is a student project and predictive-support tool, not a substitute for professional medical diagnosis; and (v) no formal software test suite beyond sample-input checks is reported.

## VI. CONCLUSION AND FUTURE WORK

This paper presented an end-to-end, layered framework for diabetes prediction. By combining validation, domain-aware cleaning, exploratory analysis, comparison of three classifiers, and a Streamlit application, the workflow bridges raw health records and an accessible prediction tool. On 2,768 records, the Random Forest reached accuracy 0.996 and ROC-AUC 0.995 on the 554-sample test set and was deployed, while Logistic Regression lagged at accuracy 0.758; glucose emerged as the leading correlate of diabetes. Because the tree-model scores are unusually high, verifying that the train and test sets share no repeated records is the first priority. Future extensions will incorporate: (1) leakage checks, record-level de-duplication before splitting, and k-fold cross-validation with reported hyperparameters; (2) larger and more diverse healthcare datasets; (3) additional machine-learning and deep-learning models with systematic tuning; (4) patient history, graphical health analysis, and personalized recommendations in the application; and (5) integration with secure healthcare platforms or databases for more comprehensive analysis.

## ACKNOWLEDGMENT

The author expresses sincere gratitude to the MCA Department, Chaitanya Bharathi Institute of Technology (Autonomous), Gandipet, Hyderabad, to internship guide Mr. P. Krishna Prasad for guidance and academic support, and to Eduexpose and its team, under whose internship program this work was carried out.

## REFERENCES

1. M. S. Chary and P. K. Prasad, "Diabetes Prediction System Using Machine Learning," *Technical Internship Report*, MCA Department, Chaitanya Bharathi Institute of Technology (Autonomous), Gandipet, Hyderabad, Telangana, India, 2026.
2. World Health Organization, "Diabetes," fact sheet. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/diabetes>
3. International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels, Belgium: IDF, 2021.
4. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Annu. Symp. Comput. Appl. Med. Care*, 1988, pp. 261–265.
5. I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct.*

- Biotechnol. J.*, vol. 15, pp. 104–116, 2017.
6. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ: John Wiley & Sons, 2013.
  7. J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.
  8. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
  9. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, art. 15, 2012.
  10. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
  11. W. McKinney, "Data structures for statistical computing in Python," in *Proc. 9th Python in Science Conf.*, 2010, pp. 56–61.
  12. C. R. Harris et al., "Array programming with NumPy," *Nature*, vol. 585, pp. 357–362, 2020.
  13. J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.
  14. M. L. Waskom, "seaborn: statistical data visualization," *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, 2021.
  15. Streamlit Inc., "Streamlit: The fastest way to build and share data apps." [Online]. Available: <https://streamlit.io/>
  16. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.
  17. Kaggle, "Healthcare Diabetes dataset (Healthcare-Diabetes.csv)." [Online]. Available: <https://www.kaggle.com/>