PUBLIC HEALTH
OpenJournal
Openventio PUBLISHERS

Original Research Article

Statistical Appraisal of the Pima Indians Diabetes Dataset: A Review and Empirical Re-analysis of Risk Factors and Predictive Models

Sujith Arokia Swamy B[1,2], Bhavana Kabber[3], Raj Shekar M[1,2], Harish Nayak[4*]

[1]The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi, India – 110012
[2]ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India – 110012
[3]University of Agricultural Sciences, Raichur, Karnataka, India – 584104
[4]University of Agricultural Sciences, Dharwad, Karnataka, India – 580005

Corresponding Author
Harish Nayak
Email: harishnayak626@gmail.com

Article Information:
Received: 05 December 2025 | Revised: 07 January 2026 | Accepted: 10 January 2026 | Published: January 21, 2026

Cite this article:
Sujith Arokia Swamy B, Raj Shekar M, Naveen Kumar P, Bhavana Kabber and Harish Nayak (2025). Beyond Simultaneous Waves: A Critical Review of COVID-19 Epidemic Synchronization and Different Country Transmission Dynamics. *Public Health Open Journal*. 10(2):24–33. https://doi.org/10.17140/PHOJ.10.02.24

Abstract

Diabetes mellitus is a major chronic disease, and statistical risk modelling of its onset has been studied for several decades. The Pima Indians Diabetes Database, comprising 768 women aged 21 years and above, has become one of the most frequently reused benchmark datasets in the biomedical statistics and machine learning (ML) literature, yet the statistical properties of the data and the validity of results derived from it are rarely examined together. This review combines a narrative synthesis of the literature with a full statistical re-analysis of the supplied dataset (268 diabetic, 500 non-diabetic; prevalence 34.9%). We examine data quality, descriptive and distributional properties, univariate group comparisons with effect sizes, categorical association tests, correlation structure, multivariable logistic regression and a ten-fold cross-validated comparison of seven classifiers. Our synthesis finds that (i) five physiological variables encode missing measurements as zero, so that only 392 of 768 records (51.0%) are complete; (ii) all eight predictors differ significantly between groups (Benjamini-Hochberg adjusted $p < 0.001$), with plasma glucose showing by far the largest standardised effect (Cohen's $d = 1.19$); (iii) after mutual adjustment, only glucose, BMI, number of pregnancies and the diabetes pedigree function remain independent predictors (McFadden $R^2 = 0.282$); and (iv) seven classifiers achieve cross-validated AUCs of 0.813-0.837 and accuracies of 74.6%-77.2%, with no model significantly better than plain logistic regression. We propose a short reporting checklist for studies using this dataset and identify six priority research directions.

Keywords: Diabetes mellitus, Pima Indians Diabetes Database, logistic regression, hypothesis testing, effect size, missing data, machine learning, narrative review

ISSN: 2472-3878 | Vol 10 | Issue 2 | 2025

1. Introduction

1.1 Motivation

Type 2 diabetes mellitus results from a combination of insulin resistance and progressive beta-cell dysfunction, and its risk is shaped by adiposity, age, reproductive history, glycaemic status and heredity. Early identification of high-risk individuals is a standing public-health objective, and statistical models that translate routine clinical measurements into a risk estimate have a long history in this field (Hosmer and Lemeshow, 2000; Kavakiotis et al., 2017). The Pima Indian population of Arizona was studied intensively because of its exceptionally high diabetes incidence (Knowler et al., 1978), and a subset of those survey data was released to the research community and later deposited in the UCI Machine Learning Repository (Smith et al., 1988; Dua and Graff, 2019).

The resulting table of 768 records and eight predictors is small, clean in appearance and freely available, which explains its popularity as a teaching and benchmarking dataset. These same properties are also its danger. The data hide a substantial missing-value problem, the sample is restricted to one demographic group, and accuracy figures reported in the literature depend heavily on how data were pre-processed and validated (Temurtas et al., 2009; Sisodia and Sisodia, 2018). A statistical appraisal that examines these issues before, and alongside, predictive modelling is therefore useful.

1.2 Scope and Contribution

This article is a narrative review supported by a complete statistical re-analysis of the supplied dataset. Our contributions are fourfold:

A structured description of the dataset, including provenance, variable definitions and a quantitative audit of implausible and missing values.

A battery of classical inferential tools (normality tests, Welch's t-test, the Mann-Whitney test, effect sizes, chi-square tests and correlation analysis) with multiplicity control.

Multivariable logistic regression with diagnostics and a sensitivity analysis of the conclusions to the handling of missing values, followed by a leakage-free cross-validated comparison of seven classifiers.

A reporting checklist and six priority research directions for future work that uses this dataset.

1.3 Review Questions

RQ1: What is the structure and quality of the dataset, and how should implausible values be handled?

RQ2: Which measured characteristics differ between diabetic and non-diabetic women, and how large are the differences?

RQ3: Which characteristics are independently associated with diabetes after mutual adjustment?

RQ4: Do flexible ML classifiers improve on logistic regression when validation is done correctly?

RQ5: What limitations remain and what are the priority directions for future research?

2. Background and Theoretical Foundations

2.1 Origin and Content of the Dataset

The dataset analysed here has the structure of the Pima Indians Diabetes Database originally compiled by the National Institute of Diabetes and Digestive and Kidney Diseases (Smith et al., 1988): 768 records of women of Pima Indian heritage aged at least 21 years, eight numeric predictors and a binary outcome. Predictors are the number of pregnancies, plasma glucose concentration two hours into an oral glucose tolerance test (mg/dL), diastolic blood pressure (mm Hg), triceps skinfold thickness (mm), two-hour serum insulin (µU/mL), body mass index (BMI, kg/m^2), the diabetes pedigree function (a score summarising family history of diabetes) and age (years). Descriptions of the exact outcome definition differ between secondary sources, so users should confirm it against the primary documentation before making clinical claims.

2.2 Statistical Tools for Group Comparison

Distributional shape was assessed with skewness, excess kurtosis and the Shapiro-Wilk test (Shapiro and Wilk, 1965). Because variances differed between groups for several variables, mean differences were tested with Welch's unequal-variance t-test (Welch, 1947), and results were cross-checked with the rank-based Mann-Whitney test (Mann and Whitney, 1947), which does not assume normality. Statistical significance alone says little about practical importance, so standardised mean differences (Cohen's d, interpreted with the conventional 0.2/0.5/0.8 benchmarks; Cohen, 1988) and rank-biserial correlations are reported alongside p-values. Because eight variables are tested simultaneously, the false discovery rate was controlled with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995).

2.3 Association and Regression Methods

Associations between categorised risk factors and diabetes status were tested with Pearson's chi-square test and summarised with Cramér's V. Monotone relationships were measured with Spearman's rank corre-

lation. Multivariable relationships were modelled with binary logistic regression, whose exponentiated coefficients are odds ratios (OR) (Hosmer and Lemeshow, 2000). Multicollinearity was examined with variance inflation factors (VIF), goodness of fit with the Hosmer-Lemeshow test and McFadden's pseudo-R^2, and discrimination with the area under the receiver operating characteristic curve (AUC) (Hanley and McNeil, 1982).

2.4 Predictive Model Families

Predictive studies on this dataset use three broad families: (1) linear and probabilistic models (logistic regression, linear discriminant analysis, Gaussian naive Bayes); (2) distance- and margin-based learners (k-nearest neighbours, support vector machines; Cortes and Vapnik, 1995); and (3) tree ensembles (random forests, Breiman, 2001; gradient boosting, Friedman, 2001). Each family offers a different trade-off between flexibility, interpretability and data requirements.

2.5 Handling Missing and Implausible Values

Little and Rubin (2019) distinguish data that are missing completely at random, at random, or not at random, and show that simple strategies such as complete-case analysis or single imputation are valid only under restrictive assumptions. In the Pima data, physiologically impossible zeros (for example a diastolic blood pressure of 0 mm Hg) act as missing-value codes. Treating them as genuine measurements distorts means, variances and correlations, so the choice of strategy is itself a modelling decision that should be reported and stress-tested.

3. Methodology of the Review

3.1 Review Design

The study has two components. The literature component is a selective narrative review of foundational and widely cited work on the dataset and on the statistical and ML methods applied to it; it does not claim the exhaustive search and screening counts of a formal PRISMA systematic review. The quantitative component is a full re-analysis of the supplied file, in which every figure and table in Sections 4 and 5 was computed directly from the data.

3.2 Data and Pre-processing

The file contains 768 rows and nine columns (eight predictors and the binary Outcome), with no duplicate rows and no empty cells. Zero values in glucose, blood pressure, skin thickness, insulin and BMI were treated as missing, whereas zero pregnancies (111 women) is a valid value. Three data sets were derived: (a) the raw data, with zeros left as recorded; (b) a median-imputed data set, used for inference; and (c) a complete-case data set (n = 392). For prediction, median imputation and standardisation were fitted inside each cross-validation training fold so that no information from the test fold could leak into pre-processing.

3.3 Statistical Analysis Plan

The analysis proceeded in six stages: (i) data-quality audit; (ii) descriptive statistics and distributional tests; (iii) univariate group comparisons with effect sizes and multiplicity control; (iv) chi-square tests on clinically categorised variables (age bands, WHO BMI classes, parity bands and a two-hour glucose threshold of 140 mg/dL); (v) correlation, VIF and multivariable logistic regression, with Hosmer-Lemeshow and Cook's distance diagnostics; and (vi) predictive comparison. With 268 events and eight candidate predictors there are about 33 events per variable, comfortably above the customary minimum of 10 (Vittinghoff and McCulloch, 2007).

3.4 Validation of Predictive Models

Seven classifiers (logistic regression, LDA, Gaussian naive Bayes, k-NN with k = 15, RBF-kernel SVM, random forest with 500 trees and gradient boosting) were evaluated with stratified ten-fold cross-validation (seed 42) using default hyperparameters. Accuracy, AUC, sensitivity, specificity, precision and F1 are reported at a 0.5 probability threshold. To compare models formally, ten-fold cross-validation was repeated five times (50 folds), and AUC differences from logistic regression were tested with the Nadeau-Bengio corrected resampled t-test, which corrects the over-optimism of naive paired tests on overlapping training sets (Nadeau and Bengio, 2003). Analyses used Python with SciPy, statsmodels and scikit-learn (Seabold and Perktold, 2010; Pedregosa et al., 2011). Reporting follows the spirit of the TRIPOD statement (Collins et al., 2015).

4. Structure and Quality of the Dataset

4.1 Variables

4.2 Class Distribution

Of 768 women, 268 (34.9%) are labelled diabetic and 500 (65.1%) non-diabetic, a class ratio of about 1.9:1. A classifier that always predicts "non-diabetic" would therefore be correct 65.1% of the time, and this majority-class rate is the correct reference point for judging any reported accuracy. Ages range from 21 to 81 years (median 29), and the maximum number of pregnancies is 17.

Table 1. Variables in the dataset

Variable	Description (unit)	Type
Pregnancies	Number of times pregnant	Count
Glucose	Plasma glucose, 2 h oral glucose tolerance test (mg/dL)	Continuous
BloodPressure	Diastolic blood pressure (mm Hg)	Continuous
SkinThickness	Triceps skinfold thickness (mm)	Continuous
Insulin	2-hour serum insulin (μU/mL)	Continuous
BMI	Body mass index (kg/m^2)	Continuous
DiabetesPedigreeFunction	Score summarising family history of diabetes	Continuous
Age	Age (years)	Integer
Outcome	Diabetes status (1 = diabetic, 0 = non-diabetic)	Binary

Table 2. Zero values recorded in physiological variables

Variable	Zeros (n)	% of records	Valid (n)
Glucose	5	0.7%	763
Blood pressure	35	4.6%	733
Skin thickness	227	29.6%	541
Insulin	374	48.7%	394
BMI	11	1.4%	757

4.3 Audit of Implausible Zero Values

Insulin (48.7%) and skin thickness (29.6%) are the most affected. Only 392 records (51.0%) are complete across all five variables, so complete-case analysis would discard 376 women (49.0%). Because the mechanism that generated the zeros is undocumented, assuming that they are missing completely at random is unsafe, and Section 5.5 tests how much the conclusions depend on this assumption.

4.4 Descriptive Statistics and Distributional Shape

The Shapiro-Wilk test rejected normality for every variable (all $p < 0.001$), though with n above 700 the test detects even trivial departures. The departures are practically meaningful for insulin (skewness 2.16, excess kurtosis 6.27) and the pedigree function (skewness 1.92, excess kurtosis 5.55), both of which are strongly right-skewed (Figure 1). Age is also right-skewed, reflecting a sample concentrated in the early reproductive and middle years (median 29 years). These shapes motivate the use of rank-based tests alongside t-tests.

5. Empirical Evidence from the Dataset

5.1 Univariate Group Comparisons

Zeros in physiological variables excluded. Welch degrees of freedom vary by variable. BH = Benjamini-Hochberg adjusted across eight tests.

Every predictor differed significantly between the groups, and the Welch and Mann-Whitney tests agreed in all cases. Effect sizes, however, vary widely. Plasma glucose shows a very large difference (diabetic mean 142.3 vs 110.6 mg/dL, d = 1.19), followed by moderate effects for insulin (d = 0.68), BMI (d = 0.69) and skin thickness (d = 0.57). Age, pregnancies, blood pressure and the pedigree function show small-to-medium effects (d between 0.36 and 0.48). The pedigree function and blood pressure are the weakest discriminators, with differences of 0.121 units and 4.4 mm Hg respectively.

5.2 Associations with Categorised Risk Factors

Zeros in glucose and BMI excluded. The underweight BMI class contains only 4 women, so the BMI chi-square approximation is imperfect. The 140 mg/dL cut-off is the conventional lower bound of impaired glucose tolerance for a two-hour post-load value (American Diabetes Association, 2024).

All four factors were strongly associated with diabetes (all $p < 0.001$). The gradients are steep: prevalence rises from 21.2% among women aged 21-29 to 55.1% at ages 40-49, and from 6.9% in the normal BMI class to 46.4% in the obese class, where 472 of 757 women (62.4%) are found. A two-hour glucose of at least 140 mg/dL is associated with an unadjusted odds ratio of 7.23 (95% CI 5.05-10.35) for the diabetic label. Parity showed a non-monotonic pattern (20.2% for 1-2 pregnancies against 34.2% for none), which probably reflects confounding with age rather than a protective effect of a first pregnancy; this cannot be resolved without adjustment (Section 5.4). Note also that the association of glucose with the outcome is partly definitional, since glucose tolerance is itself a diagnostic criterion for diabetes.

Table 3. Descriptive statistics (zeros in physiological variables treated as missing)

Variable	n	Mean	SD	Median	Q1-Q3	Range	Skew	Kurt.
Pregnancies	768	3.85	3.37	3.00	1.00-6.00	0.00-17.00	0.90	0.15
Glucose	763	121.7	30.5	117.0	99.0-141.0	44.0-199.0	0.53	-0.28
Blood pressure	733	72.4	12.4	72.0	64.0-80.0	24.0-122.0	0.13	0.90
Skin thickness	541	29.2	10.5	29.0	22.0-36.0	7.0-99.0	0.69	2.90
Insulin	394	155.5	118.8	125.0	76.3-190.0	14.0-846.0	2.16	6.27
BMI	757	32.5	6.9	32.3	27.5-36.6	18.2-67.1	0.59	0.85
Pedigree function	768	0.472	0.331	0.372	0.244-0.626	0.078-2.420	1.92	5.55
Age	768	33.2	11.8	29.0	24.0-41.0	21.0-81.0	1.13	0.63

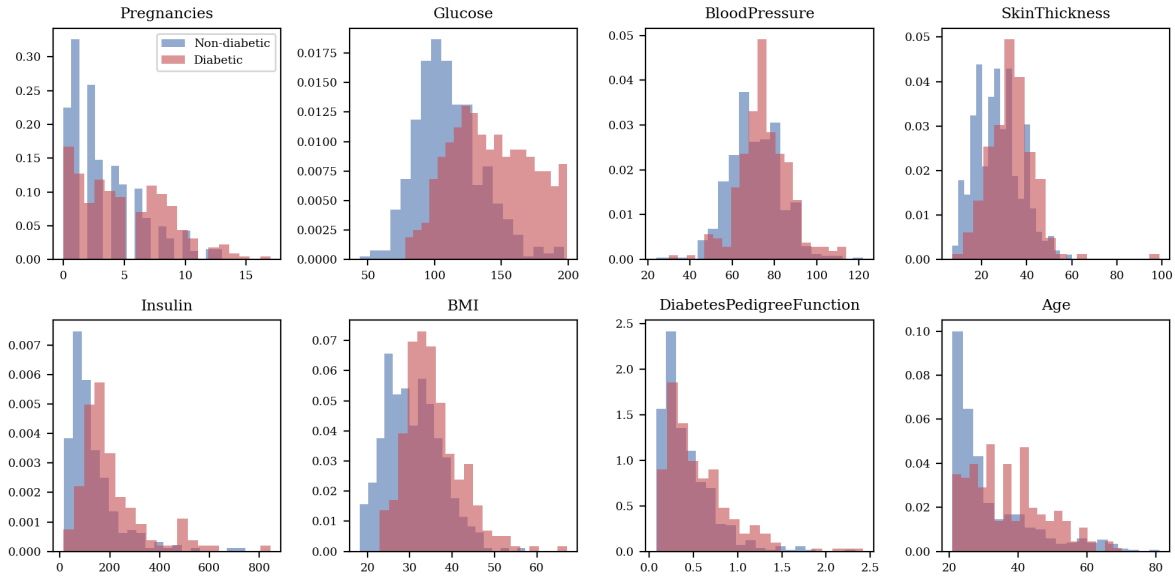


Figure 1. Distributions of the eight predictors by diabetes status

5.3 Correlation Structure and Multicollinearity

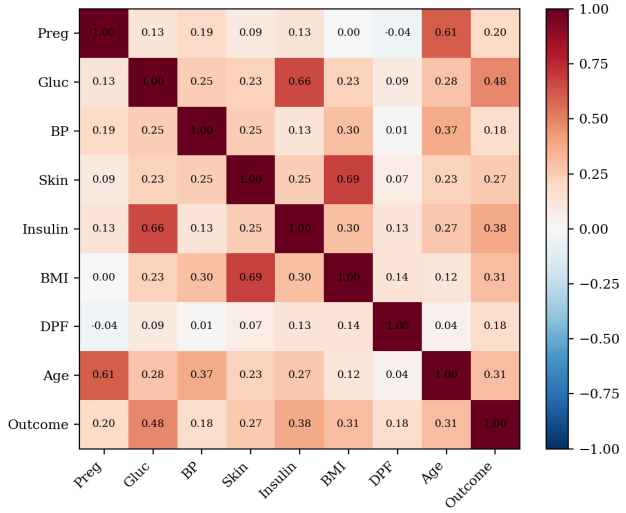


Figure 2. Spearman correlation matrix (pairwise complete observations; zeros treated as missing)

Among predictors, the strongest correlations are between BMI and skin thickness (Spearman $\rho = 0.68$), glucose and insulin ($\rho = 0.66$) and age and pregnancies ($\rho = 0.61$), as expected from body composition, glucose-insulin physiology and the mechanical link between age and parity. With the outcome, glucose has the strongest correlation (Pearson r = 0.49, Spearman ρ = 0.48), followed by BMI (r = 0.31) and insulin (r = 0.30). Despite these correlations, multicollinearity is not a serious problem for regression: the maximum VIF was 1.62 (age), far below the usual warning thresholds of 5 or 10.

5.4 Multivariable Logistic Regression

Intercept = −9.109. Likelihood-ratio $\chi^2 = 280.6$ (8 df), p < 0.001; McFadden $R^2 = 0.282$; AIC = 730.8; BIC = 772.6; Hosmer-Lemeshow $\chi^2 = 12.17$ (8 df), p = 0.144.

Four predictors are independently associated with diabetes. Each additional mg/dL of glucose raises the odds by 3.9% (OR 1.039), and one standard deviation of glucose is associated with roughly a threefold increase in odds (OR 3.17), the largest standardised effect. BMI (OR per SD 1.91), the number of pregnancies (OR per SD 1.52) and the pedigree function (OR per SD 1.34; OR 2.40 per unit) follow. In contrast, blood pressure, skin thickness, insulin and age are not significant once the other variables are in the model, although several of

Table 4. Comparison of diabetic and non-diabetic women

Variable	Diabetic mean (SD)	Non-diabetic mean (SD)	Difference [95% CI]	Welch t	Cohen's d	Mann-Whitney p (BH)
Pregnancies	4.87 (3.74)	3.30 (3.02)	1.57 [1.05, 2.09]	5.91	0.48	<0.001
Glucose	142.3 (29.6)	110.6 (24.8)	31.7 [27.5, 35.9]	14.88	1.19	<0.001
Blood pressure	75.3 (12.3)	70.9 (12.2)	4.4 [2.6, 6.3]	4.66	0.36	<0.001
Skin thickness	33.0 (10.3)	27.2 (10.0)	5.8 [3.9, 7.6]	6.18	0.57	<0.001
Insulin	206.8 (132.7)	130.3 (102.5)	76.6 [50.5, 102.7]	5.78	0.68	<0.001
BMI	35.4 (6.6)	30.9 (6.6)	4.5 [3.6, 5.5]	9.06	0.69	<0.001
Pedigree function	0.550 (0.372)	0.430 (0.299)	0.121 [0.069, 0.173]	4.58	0.37	<0.001
Age	37.1 (11.0)	31.2 (11.7)	5.9 [4.2, 7.5]	6.92	0.51	<0.001

Table 5. Prevalence of diabetes across clinically defined categories

Factor	Category	n	% diabetic	χ^2 (df)	p	Cramér's V
Age (years)	21-29	396	21.2%	69.9 (3)	<0.001	0.30
Age (years)	30-39	165	46.1%	69.9 (3)	<0.001	0.30
Age (years)	40-49	118	55.1%	69.9 (3)	<0.001	0.30
Age (years)	50+	89	48.3%	69.9 (3)	<0.001	0.30
BMI class (WHO)	Underweight (<18.5)	4	0.0%	77.1 (3)	<0.001	0.32
BMI class (WHO)	Normal (18.5-24.9)	102	6.9%	77.1 (3)	<0.001	0.32
BMI class (WHO)	Overweight (25-29.9)	179	22.3%	77.1 (3)	<0.001	0.32
BMI class (WHO)	Obese (≥30)	472	46.4%	77.1 (3)	<0.001	0.32
Pregnancies	0	111	34.2%	46.8 (3)	<0.001	0.25
Pregnancies	1-2	238	20.2%	46.8 (3)	<0.001	0.25
Pregnancies	3-5	200	35.5%	46.8 (3)	<0.001	0.25
Pregnancies	6+	219	50.7%	46.8 (3)	<0.001	0.25
2-h glucose (mg/dL)	<140	566	23.1%	132.5 (1)	<0.001	0.42
2-h glucose (mg/dL)	≥140	197	68.5%	132.5 (1)	<0.001	0.42

them were highly significant in the univariate analysis. This is the pattern expected under confounding and shared variance: age is correlated with pregnancies, skin thickness with BMI, and insulin with glucose.

The model fits adequately (Hosmer-Lemeshow p = 0.144, which gives no evidence of poor calibration, although this test has limited power) and its in-sample AUC is 0.845. A total of 49 observations had Cook's distance above 4/n (maximum 0.072); none is extreme enough to warrant removal, and no observations were deleted.

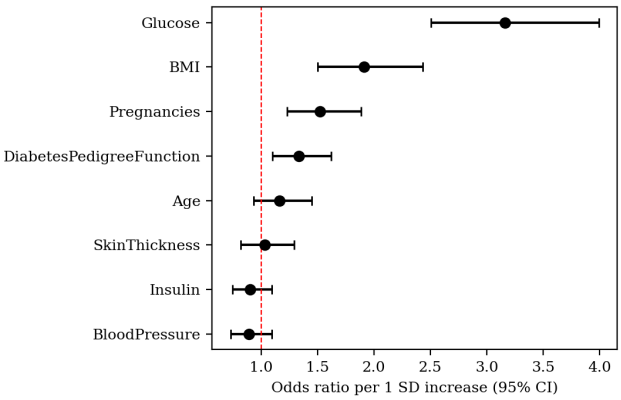


Figure 3. Standardised odds ratios (per one standard deviation) with 95% confidence intervals.

5.5 Robustness to Missing-Data Handling

Glucose, BMI and the pedigree function remain significant in both analyses, so those conclusions are robust. The pregnancy effect, however, is significant after imputation (p < 0.001) but not in the complete-case analysis (OR 1.086, p = 0.138), and age moves from clearly non-significant (p = 0.171) to borderline (p = 0.065). The complete-case model also has a higher pseudo-R^2 (0.309 vs 0.282), but it is based on half the sample and a possibly unrepresentative subset. A model fitted to the raw data with zeros treated as true values has the lowest pseudo-R^2 (0.272). These shifts show that conclusions about secondary predictors depend on how the zeros are handled and should not be reported without that caveat. Median imputation is a deliberately simple choice that shrinks variance and attenuates associations; multiple imputation by chained equations would be the more defensible approach (Little and Rubin, 2019).

Table 6. Multivariable logistic regression (median-imputed data, n = 768)

Predictor		SE	OR (95% CI), per unit	p	OR per 1 SD (95% CI)
Pregnancies	0.125	0.032	1.133 (1.063-1.207)	<0.001	1.52 (1.23-1.89)
Glucose	0.038	0.004	1.039 (1.031-1.047)	<0.001	3.17 (2.51-3.99)
Blood pressure	-0.009	0.009	0.991 (0.974-1.007)	0.275	0.89 (0.73-1.09)
Skin thickness	0.003	0.013	1.003 (0.978-1.030)	0.793	1.03 (0.82-1.29)
Insulin	-0.001	0.001	0.999 (0.997-1.001)	0.301	0.90 (0.75-1.09)
BMI	0.094	0.018	1.099 (1.061-1.138)	<0.001	1.91 (1.50-2.43)
Pedigree function	0.876	0.297	2.401 (1.342-4.295)	0.003	1.34 (1.10-1.62)
Age	0.013	0.010	1.013 (0.994-1.032)	0.171	1.17 (0.94-1.45)

Table 7. Odds ratios under alternative handling of zero values

Predictor	Median-imputed (n = 768)	p	Complete-case (n = 392)	p
Pregnancies	1.133 (1.063-1.207)	<0.001	1.086 (0.974-1.210)	0.138
Glucose	1.039 (1.031-1.047)	<0.001	1.039 (1.027-1.051)	<0.001
Blood pressure	0.991 (0.974-1.007)	0.275	0.999 (0.976-1.022)	0.904
Skin thickness	1.003 (0.978-1.030)	0.793	1.011 (0.978-1.046)	0.511
Insulin	0.999 (0.997-1.001)	0.301	0.999 (0.997-1.002)	0.528
BMI	1.099 (1.061-1.138)	<0.001	1.073 (1.017-1.132)	0.010
Pedigree function	2.401 (1.342-4.295)	0.003	3.130 (1.354-7.233)	0.008
Age	1.013 (0.994-1.032)	0.171	1.035 (0.998-1.072)	0.065

5.6 Predictive Performance of Seven Classifiers

Threshold 0.5. Majority-class accuracy = 0.651. Default hyperparameters were used for all models.

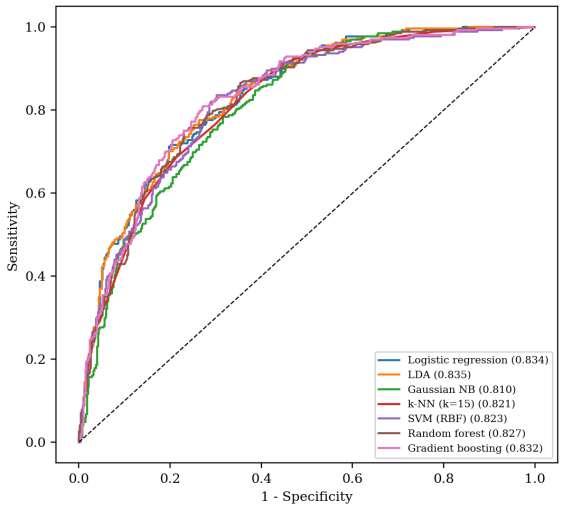


Figure 4. Cross-validated ROC curves of the seven classifiers (AUC in parentheses)

All classifiers perform similarly. Cross-validated accuracy ranges from 74.6% to 77.2%, only about 12 percentage points above the majority-class rate of 65.1%, and AUCs lie within a band of 0.024. At the default 0.5 threshold, sensitivity is modest (0.54-0.63) while specificity is higher (0.82-0.88), as expected with a 35% positive class; a screening application would need to move the threshold to favour sensitivity.

With the Nadeau-Bengio correction, none of the more flexible models (kNN, SVM, random forest, gradient boosting) is significantly different from logistic regression, and LDA is indistinguishable from it. Only Gaussian naive Bayes is significantly worse (mean AUC difference -0.025, p = 0.003), most plausibly because its conditional-independence assumption is violated by the correlated predictors. These results are for default hyperparameters; tuning inside a nested cross-validation loop might narrow or widen the gaps, but it is unlikely to overturn the main conclusion.

5.7 Single-Variable Discrimination, Variable Importance and Dimensionality

Glucose alone achieves an AUC of 0.792, which is within about 0.05 of the full multivariable models, whereas BMI and age alone give only 0.686 and 0.687. The random forest impurity importances rank glucose first (0.258), followed by BMI (0.163), age (0.130) and the pedigree function (0.123), which agrees with the regression ranking for glucose and BMI. A principal component analysis of the standardised predictors shows no low-dimensional structure: the first three components explain only 61.5% of the variance and five are needed to pass 80%. The first component is a broad size-adiposity-age contrast with loadings of similar magnitude on most variables (28.5% of variance).

6. Comparative Analysis and Discussion

6.1 RQ1: Structure and quality of the dataset

The dataset is small, has no duplicates and no blank cells, but it is not clean. Five variables use zero as a missing-value code, and the affected fraction is large for insulin (48.7%) and skin thickness (29.6%). Treating the zeros as real values (as many applied studies implicitly do) is statistically unjustified, and discard-

Table 8. Ten-fold cross-validated performance (imputation and scaling fitted within each training fold)

Model	Accuracy (SD)	AUC (SD)	Sensitivity	Specificity	Precision	F1
Logistic regression	0.767 (0.025)	0.837 (0.049)	0.556	0.880	0.726	0.623
LDA	0.764 (0.025)	0.837 (0.049)	0.556	0.876	0.723	0.622
Gaussian NB	0.746 (0.061)	0.813 (0.057)	0.601	0.824	0.661	0.622
k-NN (k=15)	0.762 (0.057)	0.821 (0.059)	0.586	0.856	0.687	0.631
SVM (RBF)	0.747 (0.039)	0.826 (0.039)	0.537	0.866	0.680	0.597
Random forest	0.772 (0.040)	0.828 (0.049)	0.634	0.844	0.689	0.659
Gradient boosting	0.772 (0.053)	0.834 (0.054)	0.627	0.850	0.697	0.655

Table 9. AUC differences relative to logistic regression (10×5-fold repeated cross-validation)

Model	Mean AUC difference	Corrected p
LDA	0.000	0.923
Gaussian NB	-0.025	0.003
k-NN (k=15)	-0.013	0.231
SVM (RBF)	-0.011	0.185
Random forest	-0.009	0.294
Gradient boosting	-0.007	0.471

ing incomplete rows removes 49.0% of the sample. A defensible workflow recodes zeros as missing, documents the amount and pattern of missingness, imputes within the modelling pipeline, and reports a sensitivity analysis.

6.2 RQ2: Which characteristics differ between groups

All eight predictors differ between groups, but the size of the difference matters. Glucose stands apart (d = 1.19); insulin, BMI and skin thickness have moderate effects; age, pregnancies, blood pressure and family history have small-to-medium effects. In line with the definitional role of glucose tolerance in diagnosing diabetes, the glucose finding is expected, and it should be interpreted as a marker of the condition rather than as an independent causal risk factor.

6.3 RQ3: Independent predictors

After mutual adjustment, glucose, BMI, number of pregnancies and the pedigree function are independently associated with the outcome; blood pressure, skin thickness, insulin and age are not. The first three conclusions are stable under complete-case analysis (apart from pregnancies), whereas the roles of age and pregnancies are sensitive to missing-data handling. Because the data are cross-sectional and from one population, associations should not be read as causal.

6.4 RQ4: Statistical versus machine-learning predictors

Logistic regression, the simplest and most interpretable model, matches the best-performing alternatives (AUC 0.837 against 0.834 for gradient boosting and 0.828 for random forest). This agrees with the observation by Kavakiotis et al. (2017) that reported performance in diabetes studies is strongly influenced by data preparation and validation rather than by the algorithm alone. On a dataset of this size, with roughly eight weakly non-linear predictors, there is little signal left for a flexible learner to capture beyond what a linear logit already provides. Studies that report markedly higher accuracy on this dataset should therefore be scrutinised for pre-processing leakage, for instance imputing or scaling before the train-test split, or for selective choice of test sets. In the present analysis, pre-processing was kept inside each training fold specifically to prevent this.

6.5 RQ5: Trade-offs and practical implications

Interpretability vs complexity: logistic regression offers transparent odds ratios with confidence intervals at no loss of accuracy. Sensitivity vs specificity: default thresholds give modest sensitivity, so clinical use would require threshold tuning against the cost of a missed case. Representativeness vs size: complete-case analysis keeps measured values but halves the sample, whereas imputation retains the sample but imposes assumptions. Generalisability: all conclusions are conditional on a single population of adult women and should be externally validated before any wider use.

7. Challenges and Limitations

7.1 Methodological Challenges

Missing-data handling: median imputation is simple and attenuates associations; multiple imputation was not used and would be preferable.

Untuned models: default hyperparameters were used, so the comparison is between default configurations and not between optimised ones.

Table 10. Discrimination of single predictors (median-imputed data)

Predictor	AUC	Youden cut-off	Sensitivity	Specificity
Glucose (mg/dL)	0.792	124	0.701	0.732
BMI (kg/m^2)	0.686	29.9	0.832	0.470
Age (years)	0.687	29	0.735	0.592

Table 11. Summary of evidence by review question

Review question	Principal finding	Evidence
RQ1 Data quality	49.0% of records have at least one zero-coded missing value	Table 4.2
RQ2 Group differences	All 8 predictors differ (BH $p < 0.001$); glucose $d = 1.19$	Table 5.1
RQ3 Independent predictors	Glucose, BMI, pregnancies, pedigree function	Table 5.3, 5.4
RQ4 ML vs statistical	AUC 0.813-0.837; no model beats logistic regression	Table 5.5, 5.6
RQ5 Limits	Single population, simple imputation, default hyperparameters	Section 7

Threshold dependence: sensitivity, specificity and F1 refer to a fixed 0.5 probability threshold and would change with the threshold.

Multiple testing and diagnostics: the Hosmer-Lemeshow test has limited power, and 49 observations had Cook's distance above 4/n though none was removed.

7.2 Data Challenges

Single population: only women of Pima heritage aged 21 and over are included, which limits generalisability to men, other ethnic groups and younger people.

Small sample: 768 records (and 392 complete ones) restrict the precision of subgroup estimates and the power to detect interactions.

Cross-sectional design: temporal ordering of risk factors and outcome cannot be established, and the exact outcome definition varies across secondary descriptions.

7.3 Reporting Challenges

Inconsistent treatment of zero values across published studies.

Accuracy reported without the majority-class baseline, calibration or confidence intervals.

Pre-processing steps and cross-validation seeds often undocumented.

The literature component of this review is selective and narrative, and not an exhaustive systematic search.

8. Priority Research Directions

Multiple imputation and missingness mechanisms. Use chained-equation imputation, with pooled estimates, and test explicitly whether the zeros are missing at random.

Calibration and decision curves. Report calibration plots and net benefit in addition to AUC, so that models can be judged by clinical usefulness.

Uncertainty quantification. Provide confidence intervals for performance metrics through bootstrapping or nested cross-validation.

External validation. Test models trained on this dataset in other, larger and more diverse cohorts.

Interaction and non-linearity. Examine splines and interactions (for example age $\times$ BMI) within a regression framework before turning to black-box models.

Reporting checklist. Adopt a minimal checklist for studies on this dataset: (i) state the treatment of zero values; (ii) fit pre-processing inside cross-validation folds; (iii) report the majority-class baseline; (iv) report AUC with an interval; (v) compare against logistic regression; (vi) state the threshold and the metrics at that threshold.

9. Conclusion

The Pima Indians Diabetes dataset remains a convenient resource, but this re-analysis shows that its apparent simplicity masks several statistical pitfalls. Three conclusions are robust. First, 49.0% of records contain zero-coded missing values, and the handling of these zeros materially affects the conclusions about secondary predictors. Second, plasma glucose is by far the dominant correlate of diabetes status ($d = 1.19$; OR per SD 3.17), followed by BMI, the pedigree function and the number of pregnancies, while blood pressure, skin thickness, insulin and age add little after adjustment. Third, flexible ML classifiers provide no significant improvement over logistic regression under leakage-free validation (cross-validated AUC about 0.84). Future studies should therefore justify added model complexity with properly validated gains, report

uncertainty and baselines, and evaluate transportability beyond this single population before drawing clinical conclusions.

References

American Diabetes Association Professional Practice Committee (2024). Diagnosis and classification of diabetes: Standards of Care in Diabetes-2024. Diabetes Care, 47(1).

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289-300.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates, Hillsdale, NJ.

Collins, G. S., Reitsma, J. B., Altman, D. G. and Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ, 350, g7594.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Hosmer, D. W. and Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.). Wiley, New York.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

Knowler, W. C., Bennett, P. H., Hamman, R. F. and Miller, M. (1978). Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. American Journal of Epidemiology, 108(6), 497-505.

Little, R. J. A. and Rubin, D. B. (2019). Statistical Analysis with Missing Data (3rd ed.). Wiley, Hoboken, NJ.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18(1), 50-60.

Nadeau, C. and Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239-281.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Seabold, S. and Perktold, J. (2010). statsmodels: econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference, 92-96.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3-4), 591-611.

Sisodia, D. and Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578-1585.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261-265.

Temurtas, H., Yumusak, N. and Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with Applications, 36(4), 8610-8615.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology, 165(6), 710-718.

Welch, B. L. (1947). The generalization of “Student’s” problem when several different population variances are involved. Biometrika, 34(1-2), 28-35.

World Health Organization (2000). Obesity: Preventing and Managing the Global Epidemic. WHO Technical Report Series 894, Geneva.