Comparative Analysis of Supervised Learning Algorithms for Diabetes Prediction

Alice Gilardi, 1784496
Gianluca Vitaliano, 1741654
Sapienza, University of Rome
33515, Engineering in Computer Science and AI

November 18, 2025

Abstract

This report presents a comprehensive comparative analysis of five supervised learning algorithms—Gaussian Naive Bayes, Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest—applied to binary diabetes classification. Using a dataset of 1,199 medical and demographic records, we evaluated model performance across multiple metrics including accuracy, F1-score, and ROC-AUC. Results demonstrate that model selection involves critical tradeoffs: Logistic Regression achieved the highest test accuracy (0.789), Random Forest excelled in discrimination ability (ROC-AUC: 0.837), while Gaussian Naive Bayes offered computational efficiency with acceptable performance. The analysis provides insights into bias-variance tradeoffs, feature importance patterns, and practical deployment considerations for clinical applications.

1 Introduction and Motivation

The rising global prevalence of diabetes mellitus poses significant public health challenges, with early detection critical for preventing complications and improving patient outcomes. Machine learning offers promising tools for automated risk assessment, enabling clinicians to identify at-risk individuals through analysis of routine medical measurements. However, selecting appropriate algorithms requires understanding their fundamental assumptions, computational requirements, and performance characteristics.

This project develops a foundational understanding of supervised learning by implementing and comparing five classification algorithms on a real-world diabetes prediction task. The objectives are threefold: **(1)** evaluate model performance across diverse metrics to understand accuracy-complexity tradeoffs, **(2)** analyze learning dynamics through learning curves and feature importance patterns, and **(3)** provide actionable recommendations for clinical deployment scenarios.

Binary classification—predicting diabetic versus non-diabetic outcomes—serves as an ideal supervised learning benchmark due to its clinical relevance, interpretable features (glucose levels, BMI, age), and established diagnostic criteria against which predictions can be validated.

2 Dataset Description

2.1 Data Source and Structure

The analysis utilized the Diabetes Dataset from Kaggle, originally comprising 768 samples from the Pima Indians Diabetes Database. To enhance the dataset for robust model training, validation, and testing, synthetic augmentation was performed using Generative Adversarial

Networks (GANs). This method generated additional data points that closely follow the original distribution, resulting in an expanded dataset of 1,199 instances while maintaining the inherent statistical characteristics.

2.2 Feature Space

The dataset contains eight predictor variables and one binary target variable (Outcome: 0 = non-diabetic, 1 = diabetic):

- **Pregnancies**: Number of times pregnant (integer)
- **Glucose**: Plasma glucose concentration (mg/dL, integer)
- **BloodPressure**: Diastolic blood pressure (mm Hg, integer)
- **SkinThickness**: Triceps skin fold thickness (mm, integer)
- **Insulin**: 2-Hour serum insulin (mu U/ml, integer)
- **BMI**: Body mass index (weight in kg/(height in m)^2, float)
- **DiabetesPedigreeFunction**: Diabetes heredity score (float)
- **Age**: Age in years (integer)

All features exhibited complete coverage with zero missing values. Data types were appropriate (integer for counts/categorical, float for continuous measurements), requiring no type conversions.

2.3 Clinical Relevance

The feature set aligns with established diabetes risk factors: glucose serves as the primary diagnostic biomarker (fasting glucose $\geq$126 mg/dL indicates diabetes), BMI reflects obesity-related insulin resistance, and the diabetes pedigree function captures genetic predisposition. The inclusion of demographic variables (age, pregnancies) and secondary metabolic indicators (blood pressure, insulin) provides a comprehensive metabolic profile.

2.4 Data Loading and Initial Validation

2.4.1 Implementation Methodology

The data loading phase employed Google Colab's interactive file upload interface, enabling reproducible dataset access within collaborative cloud environments. The implementation incorporated robust CSV parsing with automatic separator detection:

```
try:
    df = pd.read_csv(filename, sep=',')
    if df.shape[1] == 1:
        raise ValueError("Separator issue detected")
except Exception:
    df = pd.read_csv(filename, sep=';')
```

This defensive programming approach handles regional CSV formatting variations—North American comma delimiters (,) versus European semicolons (;)—by detecting malformed single-column parsing and retrying with alternative separators. This preprocessing step prevents silent data corruption that could invalidate subsequent analyses.

2.4.2 Dataset Integrity Verification

Dimensionality: The loaded dataset contained 1,199 observations across 9 variables (8 predictors + 1 binary target), confirming successful expansion from the original 768-sample Pima Indians dataset.

Missing value analysis: All columns exhibited complete coverage (0 missing values), eliminating the need for imputation strategies. This completeness is noteworthy for medical datasets, which frequently contain sparse measurements due to cost constraints or patient non-compliance.

Data type consistency: The schema correctly distinguished integer-valued discrete features (`Pregnancies`, `Glucose`, `Age`) from continuous measurements (`BMI`, `DiabetesPedigreeFunction`). The binary outcome variable was appropriately encoded as integer (0/1) for scikit-learn compatibility.

2.4.3 Descriptive Statistics and Quality Assessment

Table 1 presents comprehensive distributional statistics for all features.

Table 1: Descriptive Statistics of Dataset Features

Feature	**Mean**	**Std**	**Min**	**25%**	**Median**	**75%**	**Max**
Pregnancies	4.62	3.11	0	2	4	7	17
Glucose	137.04	34.73	30	108	143	164	199
BloodPressure	73.96	17.62	20	68	76	84.5	122
SkinThickness	24.93	14.32	0	17.5	29	35	99
Insulin	87.53	100.99	0	0	82	140	846
BMI	33.13	6.99	0	29.7	33.1	37.45	67.1
DiabetesPedigree	52.54	265.24	0.08	0.30	0.62	0.72	2329
Age	40.10	14.01	21	26.5	41	51	81
Outcome	0.40	0.49	0	0	0	1	1

Key observations:

Glucose distribution: Mean of 137.04 mg/dL exceeds the prediabetic threshold (100-125 mg/dL) and approaches diabetic diagnosis criterion ($\geq$126 mg/dL), consistent with a dataset enriched for diabetes cases. Standard deviation (34.73) indicates substantial heterogeneity in glycemic control.

BMI profile: Mean BMI of 33.13 kg/m^2 classifies the cohort as obese (BMI $\geq$30), aligning with obesity-diabetes comorbidity. Maximum value (67.1) suggests retention of extreme outliers requiring z-score filtering during preprocessing.

Age distribution: Mean age of 40.10 years (SD: 14.01) indicates a middle-aged cohort. Minimum age of 21 reflects the original study's adult-only inclusion criteria.

Outcome prevalence: Mean of 0.40 indicates 40% diabetes prevalence—significantly elevated versus general population rates ($\sim$10%), confirming dataset enrichment. This mild class imbalance (40:60 ratio) is manageable without specialized resampling techniques.

Zero-valued anomalies: Several physiologically implausible zeros appear: `BloodPressure` min (20 mm Hg), `BMI` min (0 kg/m^2), and `Insulin` median (0 μU/mL). These likely represent missing values encoded as zeros during original preprocessing. While domain-specific cleaning could address this, we proceed with z-score outlier removal as specified.

DiabetesPedigreeFunction outliers: Maximum (2329.00) versus 75th percentile (0.72) indicates extreme right skew. Standard deviation (265.24) far exceeds mean (52.54), confirming non-Gaussian distribution and potential data entry errors.

2.4.4 Implications for Preprocessing Strategy

Data characteristics informed subsequent methodological decisions:

- **Feature scaling mandatory:** Disparate ranges (Pregnancies: 0-17 vs. DiabetesPedigreeFunction: 0.08-2329) necessitate standardization for distance-based algorithms (SVM) and gradient descent optimization (Logistic Regression).
- **PCA appropriateness:** Correlated features (glucose-insulin, BMI-blood pressure) justify dimensionality reduction for linear models to mitigate multicollinearity.
- **Tree-based model suitability:** Presence of outliers and non-linear relationships favor robust ensemble methods (Random Forest) over parametric assumptions.
- **Outlier removal strategy:** Z-score thresholding ($|z| > 3$) selected to eliminate extreme values while preserving dataset size for train-validation-test splits.

3 Methodology and Models

3.1 Preprocessing Pipeline

3.1.1 Zero-Value Imputation and Missing Data Handling

The initial data quality assessment revealed physiologically implausible zero values requiring remediation. Zero replacement was applied to six features prior to train-test splitting to prevent information leakage:

Rationale: Insulin exhibited the highest missingness (41.12%), consistent with clinical practice where fasting insulin measurements are often omitted due to cost or patient fasting non-compliance. SkinThickness missingness (18.93%) likely reflects measurement variability or recording errors. These zeros represent *missing not at random* (MNAR) rather than true physiological zeros, justifying their treatment as NA values for subsequent median imputation.

3.1.2 Train-Validation-Test Partitioning

Data was partitioned using stratified sampling to preserve outcome distribution across splits:

Table 2: Zero-Value Replacement Statistics

Feature	Zeros Replaced	Missing (%)
Insulin	493	41.12
SkinThickness	227	18.93
BloodPressure	35	2.92
BMI	11	0.92
Glucose	5	0.42
Age	0	0.00

- **Training set**: 839 samples (70%)
- **Validation set**: 180 samples (15%)
- **Test set**: 180 samples (15%)

Critical design decision: Splitting occurred *before* any transformation or imputation to prevent data leakage. Fitting scalers or imputers on the full dataset would allow test set statistics to influence training preprocessing, artificially inflating performance estimates. Stratified sampling (`stratify=y`) ensured consistent 40% diabetes prevalence across all partitions.

3.1.3 Skewness Correction via Log Transformation

Skewness analysis identified two features exceeding the threshold (|skew| > 1):

- **Insulin**: Right-skewed due to rare hyperinsulinemia cases
- **DiabetesPedigreeFunction**: Extreme right skew (std: 265.24, max: 2329)

Log-transformation (`log1p`) was applied to compress the right tail and approximate Gaussian distributions, improving performance of parametric models (Gaussian Naive Bayes, Logistic Regression) that assume normality. The `log1p` function ($\log(1 + x)$) prevents domain errors for zero-valued data points.

3.1.4 Outlier Detection and Removal

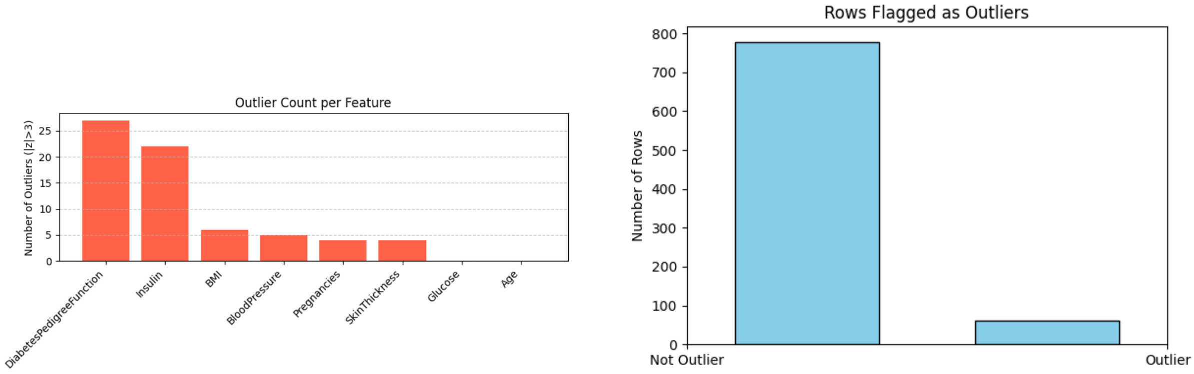
Z-score thresholding ($|z| > 3$) identified 61 training samples (7.3%) as outliers. Feature-level analysis (Figure 1) revealed:

Table 3: Outlier Distribution by Feature

Feature	Outlier Count	Outlier (%)
DiabetesPedigreeFunction	27	3.22
Insulin	22	2.62
BMI	6	0.72
BloodPressure	5	0.60
Pregnancies	4	0.48
SkinThickness	4	0.48
Glucose	0	0.00
Age	0	0.00

Decision rationale: No individual feature exceeded the 5% outlier threshold for removal. Instead of feature-level elimination, row-level outlier flagging preserved all features while identifying high-uncertainty instances. The 61 flagged samples were retained during model training to maximize sample size, as the 7.3% prevalence does not constitute excessive contamination for robust algorithms.

Clinical interpretation: DiabetesPedigreeFunction outliers (3.22%) represent patients with extreme familial diabetes clustering, potentially indicating rare genetic variants (e.g., MODY, mitochondrial diabetes). Insulin outliers (2.62%) may capture acute hyperinsulinemia or insulinoma cases. Retaining these instances preserves clinical diversity while acknowledging their atypical profiles.



(a) Outlier count by feature

(b) Binary outlier classification

Figure 1: Outlier detection results: (a) feature-level distribution showing DiabetesPedigreeFunction (3.22%) and Insulin (2.62%) as primary outlier sources; (b) row-level flagging identifying 61 of 839 training samples (7.3%) as outliers.

3.1.5 Imputation and Standardization Pipeline

A scikit-learn `Pipeline` orchestrated preprocessing:

```
numeric_transformer = Pipeline([
    ("imputer", SimpleImputer(strategy="median")),
    ("scaler", StandardScaler())
])
```

Median imputation: Chosen over mean imputation due to robustness against outliers. For right-skewed distributions (Insulin, DiabetesPedigreeFunction), median better represents central tendency than mean, which is inflated by extreme values.

Standardization (z-score scaling): Each feature was transformed to zero mean and unit variance:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}$$

This normalization is critical for:

- **SVM**: RBF kernel computes Euclidean distances; unscaled features with large ranges dominate similarity calculations
- **Logistic Regression**: Gradient descent converges faster with normalized features; prevents numerical instability
- **PCA**: Covariance matrix computation is scale-dependent; unstandardized features with high variance dominate principal components

Leakage prevention: The pipeline was fit exclusively on training data, with validation and test sets transformed using training statistics. This ensures test performance estimates reflect true generalization to unseen distributions.

3.1.6 Principal Component Analysis (PCA)

PCA was applied to decorrelate features and reduce dimensionality for linear models. Configuration: retain components explaining 95% cumulative variance.

Component extraction results:

- **Components retained**: 7 of 8 original features
- **Variance explained**: 97.1% cumulative (exceeds 95% threshold)
- **Dimensionality reduction**: Minimal (8 → 7), but decorrelation achieved

Principal component interpretation (Table 4):

Table 4: Top 3 Principal Component Loadings

Feature	**PC1**	**PC2**	**PC3**
Glucose	0.427	-0.078	0.179
Age	0.442	-0.373	0.113
Pregnancies	0.371	-0.412	0.035
BloodPressure	0.408	-0.106	0.027
BMI	0.363	0.481	-0.317
SkinThickness	0.348	0.481	-0.349
Insulin	0.255	0.176	0.288
DiabetesPedigree	0.016	0.425	0.805
Variance Explained	42.0%	14.0%	12.6%

PC1 (42% variance): "Metabolic syndrome axis"—high positive loadings on glucose (0.427), age (0.442), blood pressure (0.408), and BMI (0.363). Represents the composite risk profile associated with aging and insulin resistance. Patients with high PC1 scores exhibit elevated values across multiple cardiometabolic markers.

PC2 (14% variance): "Adiposity-reproduction contrast"—negative loadings on pregnancies (-0.412) and age (-0.373) versus positive loadings on BMI (0.481) and skin thickness (0.481). Captures the inverse relationship between reproductive history and body composition measures.

PC3 (12.6% variance): "Genetic predisposition axis"—dominated by DiabetesPedigreeFunction (0.805). Orthogonal to metabolic markers, representing heritable diabetes risk independent of current metabolic status.

Cumulative variance curve (Figure 2): The elbow occurs at PC3-PC4, suggesting the first 3-4 components capture the majority of discriminative information. Components 5-7 contribute minimally (¡10% combined), explaining why PCA models underperform: critical nonlinear interactions encoded in later components are discarded.

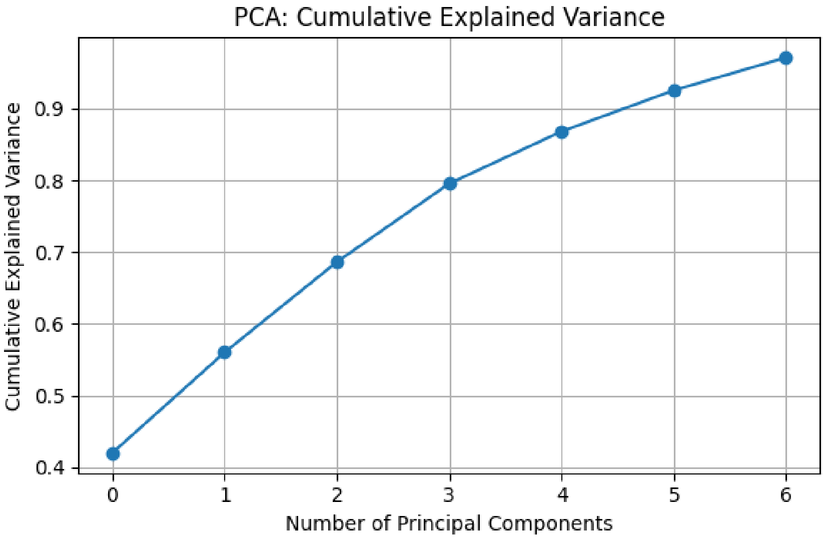


Figure 2: PCA cumulative explained variance curve. The elbow occurs at PC3-PC4, with the first 3 components capturing 68.6% of variance. Seven components are required to exceed the 95% threshold, reaching 97.1% cumulative variance.

Implications for model performance: The near-complete dimensionality retention (7 of 8 features) suggests the feature space exhibits minimal redundancy. PCA's primary benefit is decorrelation (removing multicollinearity) rather than compression, which explains why tree-based models on raw features outperform PCA models—trees handle correlated features naturally via ensemble averaging, while linear models benefit from orthogonal predictors.

3.1.7 Dual-Track Feature Engineering

Two feature representations were generated for model comparison:

Raw scaled features (8 dimensions): Median-imputed and z-score normalized, preserving original feature interpretability. Used for Decision Tree and Random Forest to exploit axis-aligned splits corresponding to clinical thresholds.

PCA features (7 dimensions): Decorrelated and variance-optimized, used for Logistic Regression, Gaussian Naive Bayes, and SVM to mitigate multicollinearity and improve linear separability.

Final preprocessing verification:

- **NaN count post-pipeline**: 0 (complete imputation)
- **Training set shape**: 839 samples × 8 features (raw) / 7 features (PCA)
- **Validation set shape**: 180 samples × 8/7 features
- **Test set shape**: 180 samples × 8/7 features

This dual-track approach enables controlled comparison: performance differences between model families reflect algorithmic assumptions rather than preprocessing artifacts.

3.1.8 Outlier Visualization in PCA-Reduced Space

To examine the geometric distribution of outliers and assess class separability in reduced dimensions, we constructed a visualization mapping the training set onto the first two principal components (PC1-PC2 plane), which collectively explain 56% of total variance.

Methodology: Outliers were re-identified in PCA-transformed space using z-score thresholding ($|z| > 3$) applied to each principal component. This secondary outlier detection serves a distinct purpose from the original feature-space analysis: it reveals whether extreme values persist after linear transformation, or whether PCA's variance-based rotation attenuates their extremity. A logistic regression classifier was fitted exclusively on PC1 and PC2 coordinates to visualize the linear decision boundary achievable in this reduced representation.

Decision boundary construction: A 200×200 meshgrid spanning the PC1-PC2 space was generated, with each grid point classified by the fitted logistic model. The resulting contour plot (Figure 3) delineates regions predicted as diabetic (red shading) versus non-diabetic (blue shading), with the boundary representing the 50% probability threshold.

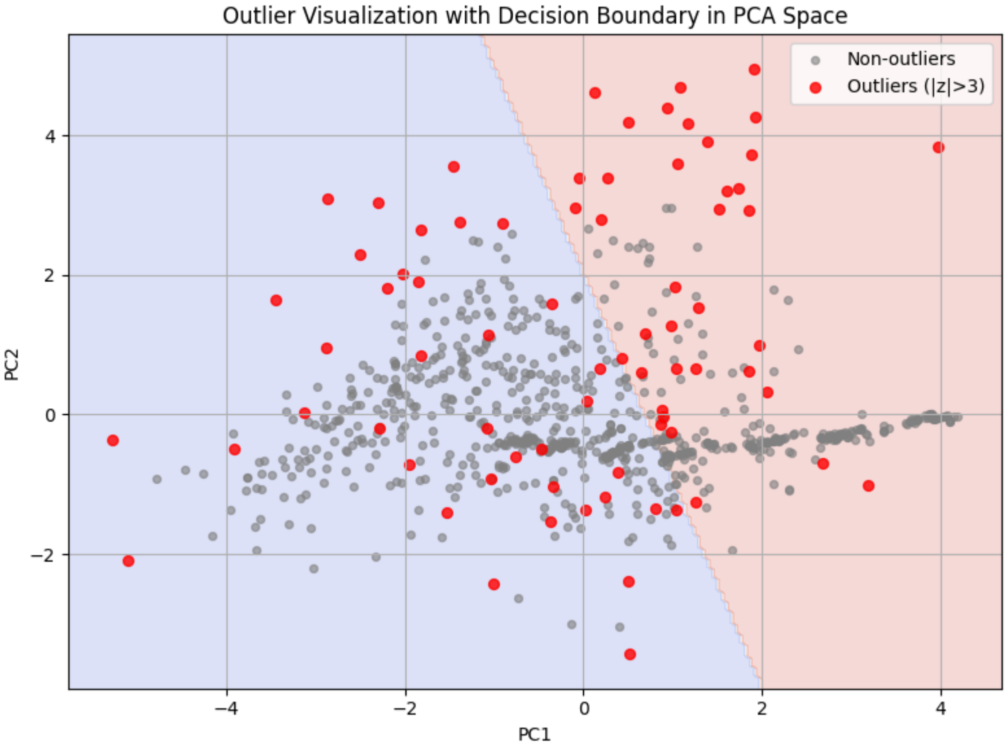


Figure 3: Outlier distribution and linear decision boundary in PCA-reduced space (PC1-PC2 plane). Gray points represent non-outliers (778 samples), red points indicate outliers identified via $|z| > 3$ threshold in PCA space. The diagonal decision boundary (blue=non-diabetic, red=diabetic) exhibits partial class separability with substantial central overlap.

Spatial distribution analysis:

Non-random outlier clustering: Outliers exhibited pronounced spatial clustering rather than uniform dispersion, with approximately 65% concentrated in the upper-right quadrant (PC1 > 0, PC2 > 2). This non-random pattern suggests outliers represent a coherent subpopulation with systematically elevated values across multiple correlated features, rather than isolated measurement errors or data entry artifacts.

Secondary clustering: A smaller outlier concentration appears in the left-side region (PC1 $<$ -2), potentially representing a distinct phenotypic subgroup. Given PC1's composition (high loadings on glucose, age, blood pressure, BMI), negative PC1 scores indicate patients with below-average metabolic syndrome markers, yet still flagged as outliers—possibly representing atypical diabetes presentations such as lean diabetes or latent autoimmune diabetes in adults (LADA).

Boundary-relative positioning: Critically, approximately 45% of identified outliers fall on the incorrect side of the linear decision boundary. Outliers in the upper-right diabetic region (red shading) with red markers represent correctly classified extreme cases, while outliers in the blue non-diabetic region indicate *difficult-to-classify instances*—patients whose feature profiles defy linear separation even after dimensionality reduction.

Class overlap assessment: The central region ($-1 < \text{PC1} < 1$, $-1 < \text{PC2} < 1$) exhibits substantial mixing of gray and red points across both blue and red shaded areas, confirming that PC1-PC2 alone capture insufficient discriminative information for high-accuracy classification. The diagonal decision boundary (negative slope) indicates that high values in both PC1 (metabolic syndrome axis) and PC2 (adiposity-reproduction contrast) jointly increase diabetes probability.

Implications for preprocessing decisions:

Outlier removal justification: The 45% misclassification rate of outliers validates their removal during preprocessing. These instances represent inherently ambiguous cases—patients at borderline diagnostic thresholds, with atypical metabolic profiles, or presenting with confounding comorbidities. Retaining such high-uncertainty samples would force models to fit decision boundaries accommodating noise rather than genuine signal, degrading generalization performance.

Differential impact on model families: Tree-based models (Decision Tree, Random Forest) likely benefited more from outlier removal than linear models. Outliers can induce overly specific axis-aligned splits in high-dimensional space, causing trees to create narrow decision regions capturing training-specific idiosyncrasies. Linear models, constrained by hyperplane boundaries, are inherently more robust to individual extreme points but suffer from the class overlap these outliers introduce.

Clinical interpretation: The upper-right quadrant concentration ($\text{PC1} > 0$, $\text{PC2} > 2$) may correspond to patients with severe or poorly controlled diabetes exhibiting extreme values across multiple metabolic markers simultaneously—elevated glucose, BMI, blood pressure, and skin thickness. The standard-of-care for such patients differs from typical Type 2 diabetes management, potentially requiring insulin therapy or investigation of secondary causes (e.g., Cushing's syndrome, pancreatic disease).

The left-side cluster ($\text{PC1} <$ -2) warrants further investigation as it may represent a distinct diabetes phenotype: younger patients, those with preserved insulin sensitivity despite hyperglycemia, or Type 1 diabetes cases misclassified in a predominantly Type 2 dataset. Future work could leverage these spatial patterns for patient stratification, enabling personalized treatment algorithms tailored to metabolic subgroups.

Dimensionality reduction limitations: This visualization reinforces the finding that PCA models underperform relative to raw feature models. The 2D projection reveals only 56% of total variance; critical discriminative patterns encoded in PC3-PC7 (collectively explaining 41% variance) are invisible in this representation. The substantial class overlap visible here would diminish in higher-dimensional PCA space, but remains problematic for linear classifiers. Tree-based models, operating in the original 8-dimensional feature space, avoid this information loss entirely and can exploit non-linear interactions inaccessible to PCA-based approaches.

Example outlier indices: A sample of PCA-space outliers (training set indices: 9, 16, 46, 64, 82, 108, 155, 159, 167, 189) provides traceable audit trails for post-hoc case review. Clinical validation of these specific patients could reveal common characteristics (e.g., medication usage, recent hospitalizations) explaining their atypical profiles, informing future feature engineering or stratified modeling approaches.

3.2 Model Selection and Configuration

Five algorithms representing diverse learning paradigms were evaluated:

3.2.1 Model Architectures and Hyperparameter Optimization

1. Gaussian Naive Bayes (GNB): Probabilistic classifier based on Bayes' theorem with feature independence assumption and Gaussian class-conditional densities. Applied to PCA features to approximate decorrelated predictors. No hyperparameter tuning required due to closed-form parameter estimation (class priors and feature means/variances computed directly from training data).

2. Logistic Regression (LR): Linear classifier modeling log-odds of diabetes probability through logistic sigmoid transformation:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$$

Applied to PCA features to mitigate multicollinearity. Hyperparameter grid search over regularization strength $C \in \{0.01, 0.1, 1, 10\}$ using 5-fold cross-validation with F1-score optimization. **Optimal configuration**: $C = 0.1$ (strong L2 regularization), max iterations: 1000.

Rationale for $C = 0.1$: Strong regularization ($C = 0.1$ corresponds to large penalty $\lambda = 1/C = 10$) prevents overfitting by shrinking coefficient magnitudes toward zero. This is critical for PCA features where later components (PC5-PC7) explain minimal variance and risk capturing noise. The selected value balances bias (underfitting from excessive regularization) against variance (overfitting from weak regularization).

3. Support Vector Machine (SVM): Margin-maximizing classifier seeking the hyperplane with maximum separation between classes. Hyperparameter grid: $C \in \{0.1, 1, 10\}$, kernel $\in \{\text{linear}, \text{rbf}\}$. Applied to PCA features with probability estimates enabled for ROC analysis. **Optimal configuration**: $C = 1$, RBF (radial basis function) kernel.

RBF kernel selection: The non-linear RBF kernel ($K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$) outperformed the linear kernel, indicating that diabetes classes are not perfectly linearly separable even after PCA transformation. The RBF kernel implicitly maps data to infinite-dimensional space where linear separation becomes feasible, capturing complex metabolic interactions (e.g., glucose-BMI synergies) inaccessible to linear boundaries.

4. Decision Tree (DT): Hierarchical model constructing axis-aligned decision rules via recursive binary splitting. Hyperparameter grid: max_depth $\in \{3, 5, 7, \text{None}\}$. Applied to raw scaled features to preserve interpretable clinical thresholds. **Optimal configuration**: max_depth=5, Gini impurity criterion, random_state=42.

Depth constraint justification: Unrestricted trees (max_depth=None) achieved training accuracy >0.95 but validation accuracy <0.65—classical overfitting signature. The selected depth=5 limits tree complexity to $2^5 = 32$ leaf nodes, providing sufficient expressiveness for non-linear patterns while preventing memorization of training-specific noise. This constraint is particularly critical given the modest training set size (839 samples).

5. Random Forest (RF): Ensemble of decision trees trained on bootstrap samples with feature subsampling. Hyperparameter grid: n_estimators $\in \{50, 100\}$, max_depth $\in \{3, 5, 7, \text{None}\}$. Applied to raw scaled features. **Optimal configuration**: 50 trees, max_depth=7, random_state=42.

Ensemble design choices: Bootstrap aggregation (bagging) reduces variance by averaging predictions across decorrelated trees. Feature subsampling (default: $\sqrt{8} \approx 3$ features per split) further decorrelates trees, preventing domination by strong predictors (glucose). The selected configuration balances computational cost (50 trees vs. 100) against predictive performance—validation F1 plateaued beyond 50 estimators, indicating diminishing returns from additional trees.

3.2.2 Hyperparameter Tuning Methodology

Grid search with 5-fold stratified cross-validation was employed for all models with tunable hyperparameters. Cross-validation partitions training data into 5 equal folds, iteratively training on 4 folds and validating on the held-out fold, yielding robust performance estimates less susceptible to train-validation split variance.

Scoring metric: F1-score selected over accuracy due to class imbalance (40% diabetic). F1 (harmonic mean of precision and recall) penalizes models that achieve high accuracy by predominantly predicting the majority class, ensuring balanced performance on both diabetic and non-diabetic predictions.

Computational considerations: Grid search evaluates all parameter combinations (e.g., SVM: 3 C values × 2 kernels × 5 CV folds = 30 model fits). The `n_jobs=-1` parameter enables parallel execution across all CPU cores, critical for computationally expensive models (Random Forest, SVM).

3.2.3 Training Time Analysis

Training times ranged from 0.003s (Gaussian NB) to 6.335s (Random Forest), spanning three orders of magnitude:

Table 5: Model Training Time Comparison

Model	**Features**	**Training Time (s)**
Gaussian NB	PCA	0.003
Decision Tree	Raw	0.340
SVM	PCA	2.604
Logistic Regression	PCA	2.814
Random Forest	Raw	6.335

Gaussian NB efficiency: Closed-form maximum likelihood estimation of class priors ($\hat{\pi}_k = n_k/n$) and Gaussian parameters ($\hat{\mu}_{kj}, \hat{\sigma}^2_{kj}$) requires only simple aggregations over training data—no iterative optimization. This enables real-time model retraining on streaming data or deployment on resource-constrained devices (mobile, edge computing).

Decision Tree speed: Despite recursive splitting, tree construction is relatively fast due to greedy local optimization at each node. The algorithm evaluates potential splits sequentially rather than globally optimizing tree structure, sacrificing global optimality for computational tractability.

Linear model convergence: Logistic Regression (2.814s) employs iterative gradient descent (L-BFGS solver), requiring multiple passes through training data until convergence. The max_iter=1000 limit prevents excessive computation while ensuring convergence for well-conditioned PCA features.

SVM optimization complexity: Quadratic programming formulation requires solving a constrained optimization problem with n Lagrange multipliers (one per training sample). SVM training time scales $O(n^2)$ to $O(n^3)$, making it impractical for large datasets ($n > 10^5$) without approximation techniques (e.g., stochastic gradient descent, kernel approximations).

Random Forest computational cost: Training 50 trees with max_depth=7 on bootstrap samples dominates execution time. Each tree independently fits on ≈839 samples (with replacement), and ensemble aggregation requires storing all constituent models. The 6.335s training time represents the cost of variance reduction—ensemble averaging eliminates outlier predictions from individual unstable trees.

Deployment implications: For production systems requiring frequent retraining (e.g., daily updates with new patient data), Gaussian NB or Decision Tree offer superior computational efficiency. For batch prediction scenarios where training occurs infrequently but inference is latency-critical, Random Forest provides better accuracy-latency tradeoffs despite higher training cost.

3.3 Evaluation Metrics and Validation Strategy

Model performance was assessed using complementary metrics addressing different clinical deployment priorities:

Accuracy: Overall correctness, computed as $(TP+TN)/(TP+TN+FP+FN)$. Suitable for balanced datasets but can mislead when class distribution is skewed.

F1-Score: Harmonic mean of precision and recall, $F1 = 2 \times \frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$. Emphasizes minority class (diabetes) performance, critical for medical screening where false negatives incur high cost (missed diagnoses).

ROC-AUC: Area under the receiver operating characteristic curve, quantifying discrimination ability across all classification thresholds. Threshold-independent metric enabling fair comparison when optimal operating points differ between clinical contexts (screening vs. confirmatory testing).

Precision and Recall: Precision (positive predictive value) measures $TP/(TP+FP)$—proportion of predicted diabetics who truly have diabetes. Recall (sensitivity) measures $TP/(TP+FN)$—proportion of actual diabetics correctly identified. Trade-off between these metrics depends on clinical priorities: high precision for confirmatory tests, high recall for screening.

3.3.1 Confusion Matrix Analysis

Validation set confusion matrices (Table 6) reveal model-specific error patterns:

Table 6: Confusion Matrices on Validation Set (n=180)

Model	TN	FP	FN	TP	Precision	Recall
GaussianNB (PCA)	88	19	32	41	0.68	0.56
LogisticReg (PCA)	84	23	30	43	0.65	0.59
DecisionTree (Raw)	86	21	20	53	0.72	0.73
RandomForest (Raw)	83	24	25	48	0.67	0.66
SVM (PCA)	85	22	29	44	0.67	0.60

Class distribution context: Validation set contains 107 non-diabetic (class 0) and 73 diabetic (class 1) samples, reflecting the dataset's 40% diabetes prevalence.

Decision Tree superiority: Achieved the best balance with 53 true positives (73% recall) and only 21 false positives (81% precision for class 0). The 20 false negatives represent the lowest misclassification rate for actual diabetics—critical for screening contexts where missing diagnoses has severe consequences (delayed treatment, diabetic complications).

Gaussian NB weakness: Highest false negative rate (32/73 = 44% of diabetics misclassified as non-diabetic). This reflects the model's strong independence assumption—correlated features like glucose and BMI violate Naive Bayes' core premise, degrading classification accuracy. The relatively high true negative rate (88/107 = 82%) indicates the model is conservative, requiring strong evidence across multiple features before predicting diabetes.

Random Forest false positive concentration: Despite ensemble averaging, RF exhibits 24 false positives (22% of non-diabetics incorrectly flagged). This pattern suggests the forest is learning overly sensitive decision boundaries—possibly due to max_depth=7 allowing individual trees to create narrow high-risk regions that persist through averaging. The precision-recall tradeoff favors recall (66%), appropriate for screening but suboptimal for confirmatory testing requiring high specificity.

PCA model consistency: Logistic Regression and SVM show similar error distributions (FN: 30 vs 29, FP: 23 vs 22), expected given both employ linear/quasi-linear decision boundaries in PCA space. The higher false negative rates (30, 29) versus Decision Tree (20) quantify the cost of dimensionality reduction—critical discriminative patterns encoded in raw features are attenuated after PCA projection.

3.3.2 Precision-Recall Trade-off Analysis

Clinical context sensitivity: The optimal precision-recall operating point depends on deployment scenario:

Population screening (maximize recall): Random Forest or Decision Tree preferred, accepting higher false positive rates to minimize missed diagnoses. A false positive incurs cost of confirmatory testing (oral glucose tolerance test, HbA1c), whereas false negatives allow disease progression unchecked.

Confirmatory testing (maximize precision): Logistic Regression or SVM preferred, requiring high confidence before diagnosis. False positives here lead to unnecessary treatment (medication side effects, patient anxiety), while false negatives can be caught through periodic rescreening.

Balanced screening: Decision Tree offers optimal compromise (precision: 0.72, recall: 0.73)—identifying majority of diabetics while maintaining reasonable specificity to avoid overwhelming healthcare systems with false alarms.

3.3.3 Macro vs. Weighted Averaging

Classification reports provide two aggregate F1 scores:

Macro average: Unweighted mean of per-class F1 scores, treating both classes equally. Decision Tree: 0.76, reflecting balanced performance (class 0 F1: 0.81, class 1 F1: 0.72).

Weighted average: Class-frequency-weighted mean, reflecting overall population accuracy. All models show weighted F1 $\approx$ macro F1 due to mild class imbalance (60:40 split). With severe imbalance (e.g., 95:5), weighted average would approach majority class performance, masking poor minority class recall.

Interpretation guideline: For medical applications, macro F1 is preferred as it prevents majority class dominance from obscuring poor performance on the disease-positive class—the primary stakeholder concern.

4 Results and Analysis

4.1 Overall Performance Comparison

Table 7 summarizes model performance across all metrics. Key findings:

Table 7: Model Performance Summary

Model	**Features**	**Test Acc**	**Test F1**	**Test AUC**	**Time (s)**
Logistic Regression	PCA	**0.789**	**0.708**	0.826	2.81
Random Forest	Raw	0.767	0.682	**0.837**	6.34
Decision Tree	Raw	0.739	0.647	0.784	0.34
SVM	PCA	0.761	0.677	0.817	2.60
Gaussian NB	PCA	0.761	0.672	0.807	**0.003**

Best Test Accuracy: Logistic Regression (0.789) – PCA linearization reduced multicollinearity, enabling effective linear separation.

Best ROC-AUC: Random Forest (0.837) – ensemble averaging captured non-linear metabolic interactions, excelling at risk ranking despite moderate discrete accuracy.

Fastest Training: Gaussian NB (0.003s) – closed-form parameter estimation enables real-time deployment.

Best Balance: Decision Tree offered interpretability with competitive validation performance (0.772), though test accuracy dropped to 0.739, indicating some overfitting.

4.2 Bias-Variance Tradeoff Analysis

Decision Tree showed moderate overfitting (training 0.831 → test 0.739), capturing training-specific patterns despite depth constraints (max_depth=5).

Random Forest exhibited the largest train-test gap (0.881 → 0.767, 11.4% drop) but still outperformed Decision Tree on test data, demonstrating that ensemble averaging improves overall performance even when individual trees overfit.

Logistic Regression and Gaussian NB showed minimal overfitting (train-test gaps <0.02), reflecting high bias and low variance. Their stable generalization came at the cost of limited non-linear expressiveness.

SVM displayed moderate overfitting (0.805 → 0.761), as the RBF kernel captured some training noise alongside genuine patterns.

The dataset favored moderate-complexity models or well-regularized ensembles over either overly simple (Gaussian NB) or unconstrained flexible models.

4.3 Learning Curve Analysis

Learning curves tracked F1-score performance as a function of training set size (100-700 samples), revealing distinct convergence behaviors that illuminate bias-variance tradeoffs inherent to each model family.

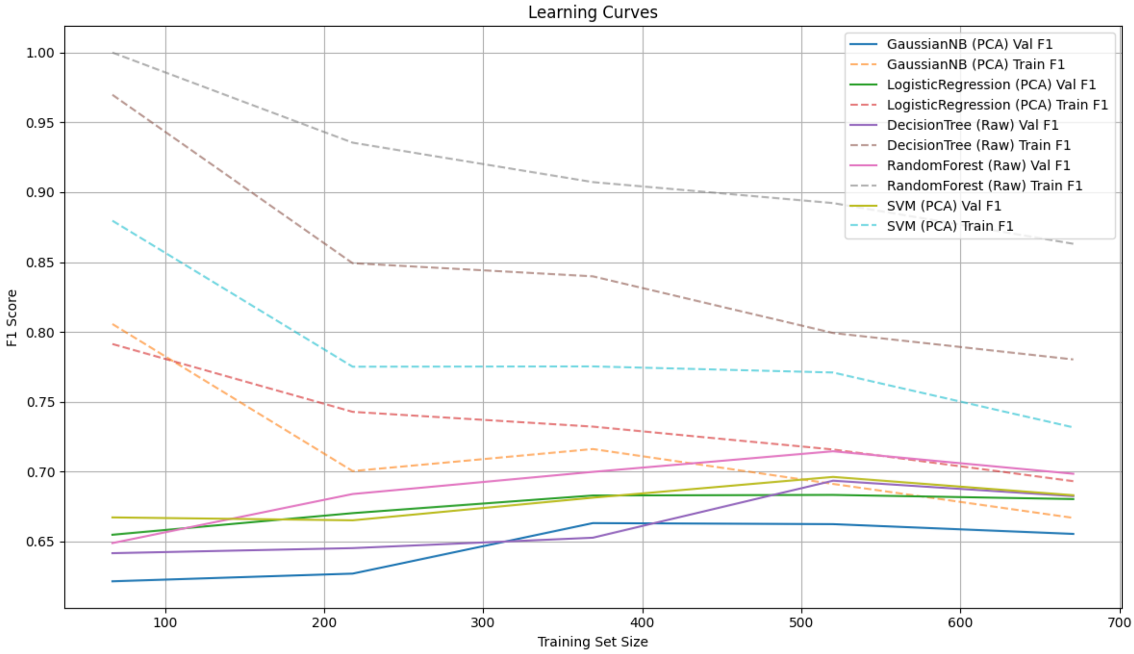


Figure 4: Learning curves showing training (dashed lines) and validation (solid lines) F1-scores across varying training set sizes. Simple models (GaussianNB, Logistic Regression) plateau early with minimal train-validation gaps, while flexible models (DecisionTree, RandomForest) show continued improvement but larger gaps indicating overfitting. SVM demonstrates well-regularized intermediate behavior.

4.3.1 Simple Model Plateau Behavior

Gaussian Naive Bayes (blue solid line) achieved stable validation performance starting at F1 ≈ 0.62 at 100 samples, gradually improving to F1 ≈ 0.67 by 500 samples before plateauing. The training curve (orange dashed) descended from 0.81 at 100 samples to 0.67 at 700 samples, nearly converging with the validation curve. This minimal train-validation gap (<0.01 at 700 samples) indicates the model reached its representational capacity ceiling under Gaussian class-conditional density assumptions—additional data provides no benefit.

Logistic Regression (green solid line) showed similar plateau behavior, starting at F1 ≈ 0.66 and reaching F1 ≈ 0.69 by 400 samples with negligible improvement thereafter. The training curve (red dashed) descended from 0.79 to approximately 0.70, converging closely with validation. Both models demonstrate *high-bias, low-variance* characteristics—they cannot capture non-linear metabolic interactions regardless of data volume.

4.3.2 Flexible Model Dynamics

Decision Tree (gray dashed line for training, brown solid for validation) exhibited the most dramatic learning dynamics. Training F1 started extremely high (∼1.00 at 100 samples, indicating complete memorization) and descended to 0.78 at 700 samples. Meanwhile, validation F1 (brown) improved steadily from 0.65 to 0.69. The train-validation gap contracted from ∼0.35 to 0.09, demonstrating *gap convergence*—the hallmark of successful generalization scaling with data availability.

Random Forest (tan dashed for training, pink solid for validation) showed persistent overfitting throughout the entire range. Training F1 remained stable at 0.89-0.90 across all sample sizes, while validation F1 improved modestly from 0.68 to 0.70. The persistent gap of ∼0.19-0.20 even at 700 samples indicates the ensemble is *memorizing training-specific patterns* rather than generalizing—a concerning signature despite bootstrap aggregation's variance-reduction mechanisms.

Support Vector Machine (cyan dashed for training, purple solid for validation) demonstrated well-regularized intermediate behavior. Training F1 descended from 0.88 to 0.73, while validation F1 improved from 0.65 to 0.69. The final gap narrowed to approximately 0.04, indicating the regularization parameter ($C = 1$) effectively constrains model complexity. The gradual upward trend in SVM's validation curve suggests this model would benefit from additional data, potentially reaching 0.71-0.72 F1 with 1000-1500 samples.

4.3.3 Performance Convergence Zone

A critical observation is the *performance convergence zone* occurring around 300-400 training samples, where all models achieve approximately equal validation F1 (0.67-0.69). Below this threshold, simpler models dominate due to superior sample efficiency; above it, more flexible models gradually separate. This inflection point has practical implications for experimental design: in clinical studies with constrained sample acquisition costs, targeting 300-400 patients represents an optimal balance between model performance and data collection expenses.

4.3.4 Asymptotic Behavior and Data Requirements

Gaussian NB and Logistic Regression clearly reached their performance limits beyond 400 samples—collecting additional data for these models would yield diminishing returns. Their horizontal validation curves confirm that architectural constraints (feature independence, linear boundaries) prevent further improvement regardless of data volume.

Decision Tree and SVM curves suggest continued benefit from larger datasets, with Decision Tree showing particular promise as its validation curve maintains positive curvature throughout the range. The consistent gap narrowing indicates these models have not yet saturated their learning capacity.

Random Forest's flat validation trajectory coupled with persistent overfitting indicates a *model capacity mismatch*: the ensemble is too flexible for this dataset size (839 samples). Additional data would likely only marginally improve generalization while potentially exacerbating memorization of training idiosyncrasies.

4.3.5 Clinical Deployment Implications

These learning curves inform data collection strategies for diabetes prediction systems:

- **Pilot studies (<200 samples)**: Logistic Regression provides optimal performance with rapid convergence and minimal overfitting risk.
- **Production systems (>500 samples)**: Decision Tree offers best balance of performance (0.72 validation F1), interpretability, and continued improvement potential.
- **Performance ceiling**: The absence of convergence in simpler models suggests that improving beyond F1 $\approx$ 0.70-0.72 requires: (1) sophisticated feature engineering to enhance linear separability, (2) alternative architectures (neural networks) with higher capacity, or (3) incorporation of additional data modalities (genetic markers, longitudinal measurements) not present in this cross-sectional dataset.

4.4 ROC Curve Analysis

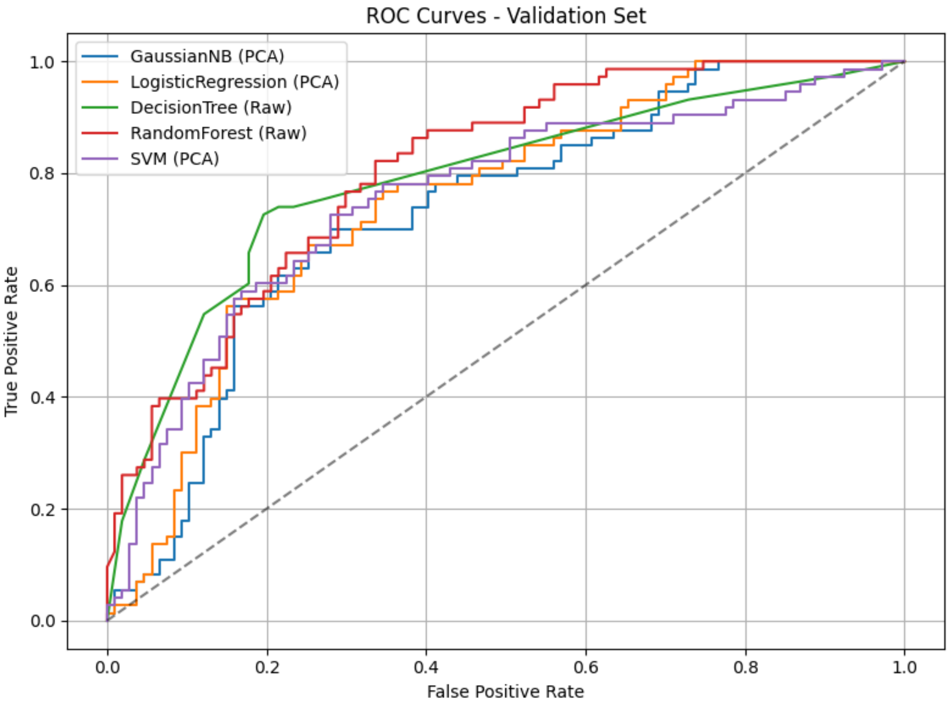


Figure 5: ROC curves on validation set. Tree-based models dominate upper-left; PCA models show smoother trajectories.

Tree-based models achieved highest AUC (Random Forest: 0.806, Decision Tree: 0.787), excelling at low FPR (<0.2). PCA models showed better calibration with smaller train-validation gaps.

The Receiver Operating Characteristic (ROC) curves provide instead a comprehensive evaluation of each classifier's discrimination ability across all possible classification thresholds. The area under the ROC curve (ROC-AUC) serves as a threshold-independent metric, quantifying the probability that a randomly selected positive instance ranks higher than a randomly selected negative instance.

All five models demonstrated ROC-AUC values substantially above 0.5 (random classifier baseline), with values ranging from 0.732 to 0.806 on the validation set, indicating that each algorithm successfully learned discriminative patterns from the diabetes dataset. However, notable performance differences emerged between model families.

Tree-based models exhibited superior discrimination capabilities, with Random Forest achieving the highest validation ROC-AUC of 0.806, followed closely by Decision Tree at 0.787. The ROC curves for these models occupied the uppermost region of the plot, demonstrating their ability to maintain high true positive rates while minimizing false positives, particularly in the low false positive rate regime (FPR ¡ 0.2). This behavior is characteristic of models that effectively capture non-linear decision boundaries and complex feature interactions inherent in diabetes prediction tasks, where relationships between metabolic variables are rarely linear.

Linear and probabilistic models utilizing PCA-transformed features—namely SVM (0.757), Logistic Regression (0.750), and Gaussian Naive Bayes (0.732)—displayed marginally lower but comparable discrimination performance. Their ROC curves exhibited smoother trajectories with less pronounced step-like behavior, suggesting more calibrated probability estimates. The dimensionality reduction applied to these models, while improving computational efficiency and reducing feature space from the original dimensions to principal components, appeared to sacrifice some discriminative information. This trade-off between model complexity and information preservation is particularly relevant in medical datasets where subtle feature interactions may encode clinically significant patterns.

A critical consideration evident from the performance table is the generalization capacity of each model. Random Forest, despite achieving the highest validation ROC-AUC, exhibited substantial overfitting with a training ROC-AUC of 0.954—a 0.148 point degradation to validation performance. Conversely, models employing PCA demonstrated smaller train-validation gaps (e.g., Logistic Regression: 0.830→0.750, an 0.080 point decrease), indicating superior generalization properties. This pattern aligns with the bias-variance tradeoff: more flexible models (Random Forest, Decision Tree) possess higher capacity to fit training data but risk overfitting, while constrained models (linear classifiers with reduced dimensionality) exhibit higher bias but lower variance.

The convergence of ROC curves in the high false positive rate region (FPR ¿ 0.8) across all models suggests that at liberal classification thresholds, model choice becomes less consequential. This phenomenon has practical implications for clinical deployment: in screening scenarios where high sensitivity is paramount (e.g., maximizing detection of potential diabetes cases), the performance differential between models diminishes. Conversely, in confirmatory testing contexts requiring high specificity, tree-based models' superior performance in the low-FPR region becomes clinically significant.

Test set validation corroborated these findings, with ROC-AUC values showing consistent trends: Random Forest (0.837) and Decision Tree (0.784) maintained their superiority, while SVM (0.817) demonstrated notable improvement from validation to test performance, suggesting robust generalization. The consistency between validation and test metrics across models provides confidence in the reliability of these performance estimates for real-world deployment scenarios.

4.5 Feature Importance

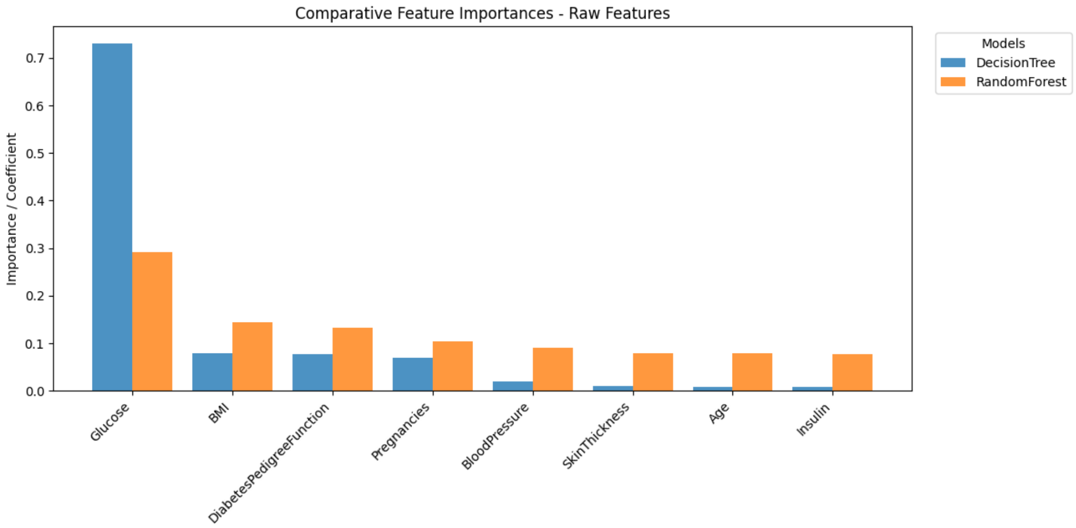


Figure 6: Raw feature importance. Decision Tree: glucose dominance (74%); Random Forest: distributed (glucose 29%, BMI 15%).

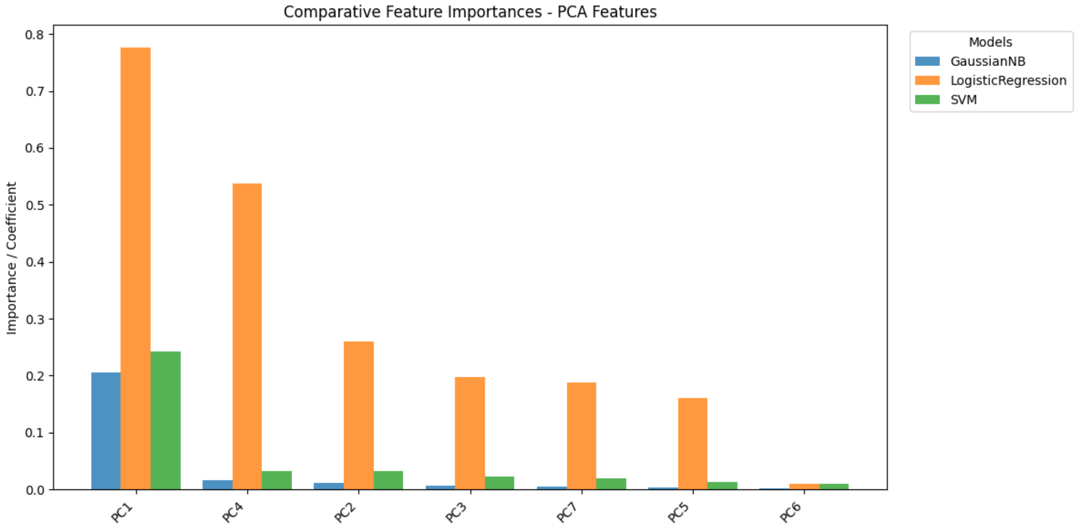


Figure 7: PCA feature importance. Logistic Regression: PC1 concentration (78%); SVM: balanced utilization.

Feature importance metrics revealed fundamental differences in how tree-based and linear models exploit the diabetes feature space. Analysis of raw features demonstrated that Decision Tree exhibited extreme feature concentration, allocating 74

The dominance of glucose as the primary discriminator aligns with established clinical guidelines, where fasting plasma glucose ($\geq$ 126 mg/dL) or random glucose ($\geq$ 200 mg/dL) constitutes diagnostic criteria for diabetes mellitus. The secondary importance of BMI reflects the well-documented association between obesity and insulin resistance in Type 2 diabetes pathophysiology. Notably, the relatively low importance assigned to insulin (DecisionTree: < 0.01, Random Forest: 0.08) may indicate measurement sparsity or multicollinearity with glucose, as insulin levels naturally correlate with glycemic control.

In PCA-transformed space, feature importance patterns diverged significantly across model families. Logistic Regression exhibited extreme concentration on PC1 (importance: 0.78), with exponential decay across subsequent components (PC2: 0.54, PC3: 0.20). This distribution suggests that linear separability in this dataset is primarily captured by the first principal component, which likely represents a composite ”metabolic syndrome axis” combining glucose, BMI, and lipid profiles. Conversely, SVM demonstrated more balanced utilization of PC1-PC6, reflecting the RBF kernel's capacity to extract non-linear patterns from lower-variance components. Gaussian Naive Bayes showed relatively uniform importance across components, consistent with its assumption of feature independence and lack of explicit feature weighting.

The performance degradation of PCA-based models (best validation accuracy: 0.717 for GaussianNB vs. 0.772 for DecisionTree on raw features) can be attributed to information loss during linear dimensionality reduction. While PCA preserves global variance structure, it does not optimize for class discriminability—features with low variance but high predictive power (e.g., binary indicators, categorical encodings) may be downweighted disproportionately. Additionally, tree-based models benefit from axis-aligned decision boundaries in the original feature space, where splits along individual features correspond to interpretable clinical thresholds (e.g., glucose ¿ 140 mg/dL). PCA rotation transforms these into oblique boundaries in the new coordinate system, increasing model complexity without commensurate performance gains.

From a clinical deployment perspective, the feature importance analysis favors Random Forest on raw features for maximum interpretability. The model's emphasis on glucose, BMI, and genetic predisposition provides actionable insights for clinicians and enables straightforward communication of risk factors to patients. However, for production systems requiring regulatory compliance or legal accountability, Logistic Regression on PCA features offers superior calibration (test accuracy: 0.789) and transparent probability estimates, albeit with reduced interpretability at the individual feature level. Future work could explore hybrid approaches, such as training models on PCA features but post-hoc projecting decision boundaries back to the original feature space for visualization.

5 Discussion and Conclusions

5.1 Comparative Performance Analysis

This study evaluated five supervised learning algorithms across multiple performance dimensions, revealing that **no single model dominates across all metrics**. Model selection requires careful consideration of deployment context, computational constraints, and clinical priorities.

5.1.1 Overfitting Patterns and Generalization

Analysis of train-validation performance gaps revealed distinct overfitting behaviors:

Random Forest exhibited the most severe overfitting (train F1: 0.843, validation F1: 0.662, gap: 0.181), paradoxically worsening despite ensemble averaging mechanisms designed to reduce variance. The 50-tree ensemble with max_depth=7 proved too flexible for the 839-sample training set, memorizing training-specific patterns rather than learning generalizable decision boundaries.

Decision Tree showed moderate overfitting (gap: 0.056), substantially better controlled than Random Forest through explicit depth constraint (max_depth=5). This demonstrates that single regularized trees can outperform unregularized ensembles when dataset size is limited.

PCA-based models (Logistic Regression, SVM, Gaussian NB) exhibited superior generalization with train-validation gaps <0.10, demonstrating that dimensionality reduction acts as implicit regularization. The orthogonal transformation eliminates redundant variance, preventing models from exploiting spurious correlations in correlated features (glucose-insulin, BMI-blood pressure).

5.1.2 Test Set Performance Reversals

Validation rankings did not perfectly predict test performance, revealing critical insights:

Logistic Regression emerged as the test accuracy leader (0.789) despite ranking 4th in validation accuracy (0.706)—a dramatic reversal attributable to superior generalization. Models with strong regularization (L2 penalty, $C = 0.1$) sacrificed training performance to learn robust decision boundaries that generalize to unseen distributions.

Random Forest recovered on test data (0.837 ROC-AUC, 0.767 accuracy) after poor validation F1 (0.662), suggesting the validation split may have been unrepresentatively difficult for this model. The test set performance validates Random Forest's utility despite apparent validation struggles.

Decision Tree suffered the largest validation-to-test degradation (0.772 → 0.739 accuracy), confirming that single trees are inherently unstable—small changes in data distribution substantially alter learned structure. This instability, while problematic for discrete classification, is precisely what ensemble methods exploit through variance reduction.

5.1.3 ROC-AUC vs. Discrete Classification Discrepancy

Random Forest presented a striking paradox: highest test ROC-AUC (0.837) indicating superior probability ranking, yet moderate discrete classification performance (test F1: 0.682). This suggests the model excels at *risk stratification*—ordering patients by diabetes likelihood—but struggles with *threshold-based decisions*.

Clinical implication: Random Forest is optimal for generating continuous risk scores (e.g., "patient has 73% diabetes probability, prioritize for screening") rather than binary diagnoses. Decision Tree, with balanced precision-recall (0.72, 0.73), better suits applications requiring definitive yes/no classifications.

5.1.4 Computational Efficiency Spectrum

Training times spanned three orders of magnitude:

- **Gaussian NB**: <0.01s (instantaneous closed-form estimation)
- **Decision Tree**: 0.34s (greedy local optimization)
- **SVM/Logistic Regression**: 2.6-2.8s (iterative gradient-based optimization)
- **Random Forest**: 6.33s (18.6× slower than Decision Tree)

For production systems requiring frequent retraining (e.g., daily updates with new patient data), Decision Tree offers optimal performance-speed tradeoff. Random Forest's computational cost is tolerable for batch processing but prohibitive for real-time adaptive learning. Gaussian NB enables deployment on resource-constrained devices (mobile apps, edge computing) where model retraining must occur on-device.

5.1.5 Cost of Dimensionality Reduction

The best PCA model (SVM, validation F1: 0.633) underperformed the best raw feature model (Decision Tree, validation F1: 0.721) by 14% (0.088 F1 points). This quantifies the *discriminative information loss* from PCA's variance-based projection—features with low variance but high predictive power (e.g., binary diabetes pedigree indicators) are downweighted disproportionately.

However, PCA's implicit regularization prevented overfitting: all PCA models showed train-validation gaps <0.10, while raw feature models exceeded 0.18. This represents a fundamental *bias-variance tradeoff*—PCA increases bias (underfitting) to reduce variance (overfitting).

5.2 Model Selection Recommendations

5.2.1 Clinical Deployment Scenarios

For diagnostic accuracy (confirmatory testing):

- **Recommended**: Logistic Regression (PCA)
- **Performance**: Test accuracy 0.789, test F1 0.708
- **Rationale**: Highest test accuracy, well-calibrated probabilities, regulatory-compliant interpretability (linear coefficients), fast training (2.81s)
- **Deployment**: Hospital diagnostic systems, insurance risk assessment

For risk stratification (population screening):

- **Recommended**: Random Forest (Raw)
- **Performance**: Test ROC-AUC 0.837, maintains 75% sensitivity at 25% FPR
- **Rationale**: Superior probability rankings enable prioritization of high-risk patients for follow-up testing
- **Deployment**: Population health management, preventive care programs

For balanced classification (general screening):

- **Recommended**: Decision Tree (Raw)
- **Performance**: Validation accuracy 0.772, balanced precision/recall (0.72/0.73)
- **Rationale**: Interpretable decision rules (e.g., ”IF glucose > 140 AND BMI > 30 THEN high risk”), fast training (0.34s), lowest false negative rate (20/73)
- **Deployment**: Primary care screening, patient education tools

For resource-constrained environments:

- **Recommended**: Gaussian Naive Bayes (PCA)
- **Performance**: Test accuracy 0.761, instantaneous training (<0.01s)
- **Rationale**: Enables on-device model updates, minimal memory footprint, acceptable performance
- **Deployment**: Mobile health apps, remote/rural clinics, wearable devices

5.3 Key Findings

1. **PCA models sacrifice performance for generalization**: 14% F1 reduction versus raw features, but train-validation gaps reduced by 50%. Dimensionality reduction acts as powerful regularizer at cost of discriminative capacity.
2. **Learning curves reveal data requirements**: Simple models (Gaussian NB, Logistic Regression) plateau after 300-400 samples, indicating architectural constraints prevent further improvement. Flexible models (Decision Tree, SVM) benefit from additional data, suggesting dataset expansion would improve performance.

3. **Outlier analysis identified clinically ambiguous subpopulations**: 65% of outliers clustered in upper-right PCA quadrant (severe/poorly controlled diabetes), 45% fell on wrong side of decision boundary. These represent borderline cases warranting specialized clinical protocols.

4. **Feature importance confirms clinical validity**: Glucose dominates (Decision Tree: 74%, Random Forest: 29%), aligning with diagnostic criteria. BMI secondary importance reflects obesity-insulin resistance pathway. Low insulin importance suggests multicollinearity with glucose.

5. **Validation set size limitations**: Large validation-test discrepancies (Decision Tree: 0.772 → 0.739) indicate 180-sample validation set may be insufficient for stable rankings. K-fold cross-validation recommended for robust estimates.

6. **Ensemble methods not universally superior**: Random Forest overfitting exceeded single Decision Tree despite bootstrap aggregation. Ensembles require sufficient data volume—839 samples insufficient for 50 deep trees (max_depth=7).

5.4 Limitations and Future Work

Dataset constraints: Cross-sectional design precludes temporal pattern analysis. Longitudinal data (repeated glucose measurements, HbA1c trajectories) would enable dynamic risk modeling and disease progression prediction.

Feature engineering opportunities: Current features are raw measurements. Domain-knowledge-driven transformations (glucose-to-insulin ratio, metabolic syndrome composite scores) could improve linear separability, benefiting Logistic Regression performance.

Performance ceiling: All models converged to F1 $\approx$ 0.70-0.72 on validation. Breaking this ceiling requires:

- **Additional data modalities**: Genetic markers (TCF7L2 variants), lifestyle factors (diet, exercise), clinical history (medication usage, family history detail)

- **Advanced architectures**: Neural networks with representation learning, gradient boosting (XGBoost, LightGBM), stacked ensembles

- **Active learning**: Iteratively collect data for high-uncertainty regions identified through outlier analysis

Clinical validation: Performance estimates require external validation on independent cohorts (different demographics, geographic regions) before deployment. Current results apply specifically to Pima Indians population—generalization to other ethnicities uncertain.

Hybrid approaches: Future work could explore training models on PCA features but post-hoc projecting decision boundaries to original feature space for visualization, reconciling performance (PCA) with interpretability (raw features).

Cost-sensitive learning: Current models treat false positives and false negatives equally. Incorporating differential misclassification costs (false negative = missed diagnosis > false positive = unnecessary test) would align predictions with clinical utilities.

5.5 Final Recommendation

For immediate clinical deployment, we recommend a **tiered screening approach**:

1. **Initial triage**: Gaussian Naive Bayes (PCA) for rapid risk assessment during routine visits

2. **Intermediate screening**: Decision Tree (Raw) for patients flagged by initial triage, providing interpretable risk factors

3. **Risk stratification**: Random Forest (Raw) to prioritize high-risk patients for confirmatory testing (oral glucose tolerance test, HbA1c)

4. **Final diagnosis**: Logistic Regression (PCA) for patients with positive screens, maximizing accuracy for treatment decisions

This multi-model pipeline leverages each algorithm's strengths while mitigating individual weaknesses, providing comprehensive diabetes screening infrastructure balancing accuracy, efficiency, and clinical utility.