Assignment I - Comparative Analysis 1

Alice Gilardi - 1784496
Gianluca Vitaliano - 1741654

Comparative Analysis of Classification Models on Diabetes Dataset

1. Model Performance Overview

We evaluated five classification models—**Gaussian Naive Bayes (GNB), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)**—on the diabetes dataset using either PCA-transformed features (linear models, SVM) or raw scaled features (tree-based models).

Model	Features	Train Accuracy	Val Accuracy	Test Accuracy	Train F1	Val F1	Test F1	Train ROC-AUC	Val ROC-AUC	Test ROC-AUC	Training Time (s)	Fitted Model
2	DecisionTree	Raw	0.831	0.772	0.739	0.777	0.721	0.647	0.880	0.787	0.784	0.340
3	RandomForest	Raw	0.881	0.728	0.767	0.843	0.662	0.682	0.954	0.806	0.837	6.335
4	SVM	PCA	0.805	0.717	0.761	0.734	0.633	0.677	0.860	0.757	0.817	2.604
1	LogisticRegression	PCA	0.775	0.706	0.789	0.698	0.619	0.708	0.830	0.750	0.826	2.814
0	GaussianNB	PCA	0.765	0.717	0.761	0.670	0.617	0.672	0.815	0.732	0.807	0.003

Key Results:

- **Best Test Accuracy**: Logistic Regression (0.789) – PCA linearization reduced multicollinearity
- **Best ROC-AUC**: Random Forest (0.837) – excellent risk ranking despite moderate accuracy (0.767)
- **Best Validation Performance**: Decision Tree (0.772) – but dropped to 0.739 on test, indicating overfitting
- **Fastest Training**: Gaussian NB (0.003s) – acceptable performance (0.761 test accuracy) for resource-constrained applications

Performance Tradeoffs: No single model dominated. Logistic Regression excelled in accuracy/F1, Random Forest in discrimination (ROC-AUC), and Gaussian NB in speed. SVM (0.761 accuracy, 0.817 AUC) provided balanced performance between linear and ensemble methods.

This diversity highlights that no single model dominated in every respect; performance depended on the interplay between assumptions, preprocessing, and metric choice.

2. Influence of Model Assumptions

Each algorithm's architectural assumptions had a clear and measurable influence on its behavior.

- **GaussianNB** assumes feature independence and normality—violated by correlated diabetes biomarkers, explaining its lower recall (0.56-0.59) on the minority class.
- **Logistic Regression** assumes linear separability. PCA's orthogonal compression aligned well with this assumption, enabling strong accuracy (0.789) despite model simplicity.
- **SVM (RBF kernel)** models smooth nonlinear boundaries. PCA reduced noise but sensitivity to local structure limited positive-class recall.
- **Decision Trees** use axis-aligned splits, prone to overfitting without depth limits. The max_depth=5 constraint balanced generalization against boundary rigidity.
- **Random Forests** average decorrelated trees to capture complex interactions. This yielded excellent ROC-AUC (0.837) for risk ranking, though discrete predictions lagged due to threshold effects.

3. Overfitting and Bias–Variance Trade-Off

- **Decision Tree** showed moderate overfitting (training 0.831 → test 0.739), capturing training-specific patterns despite depth constraints.
- **Random Forest** exhibited the largest train-test gap (0.881 → 0.767) but still outperformed Decision Tree on test data, demonstrating that ensemble averaging improves overall performance even when individual trees overfit.
- **Logistic Regression and GaussianNB** showed minimal overfitting (train-test gaps <0.02), reflecting high bias and low variance. Their stable generalization came at the cost of limited nonlinear 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 (Logistic Regression) or well-regularized ensembles (Random Forest) over either overly simple (GaussianNB) or unconstrained flexible models.

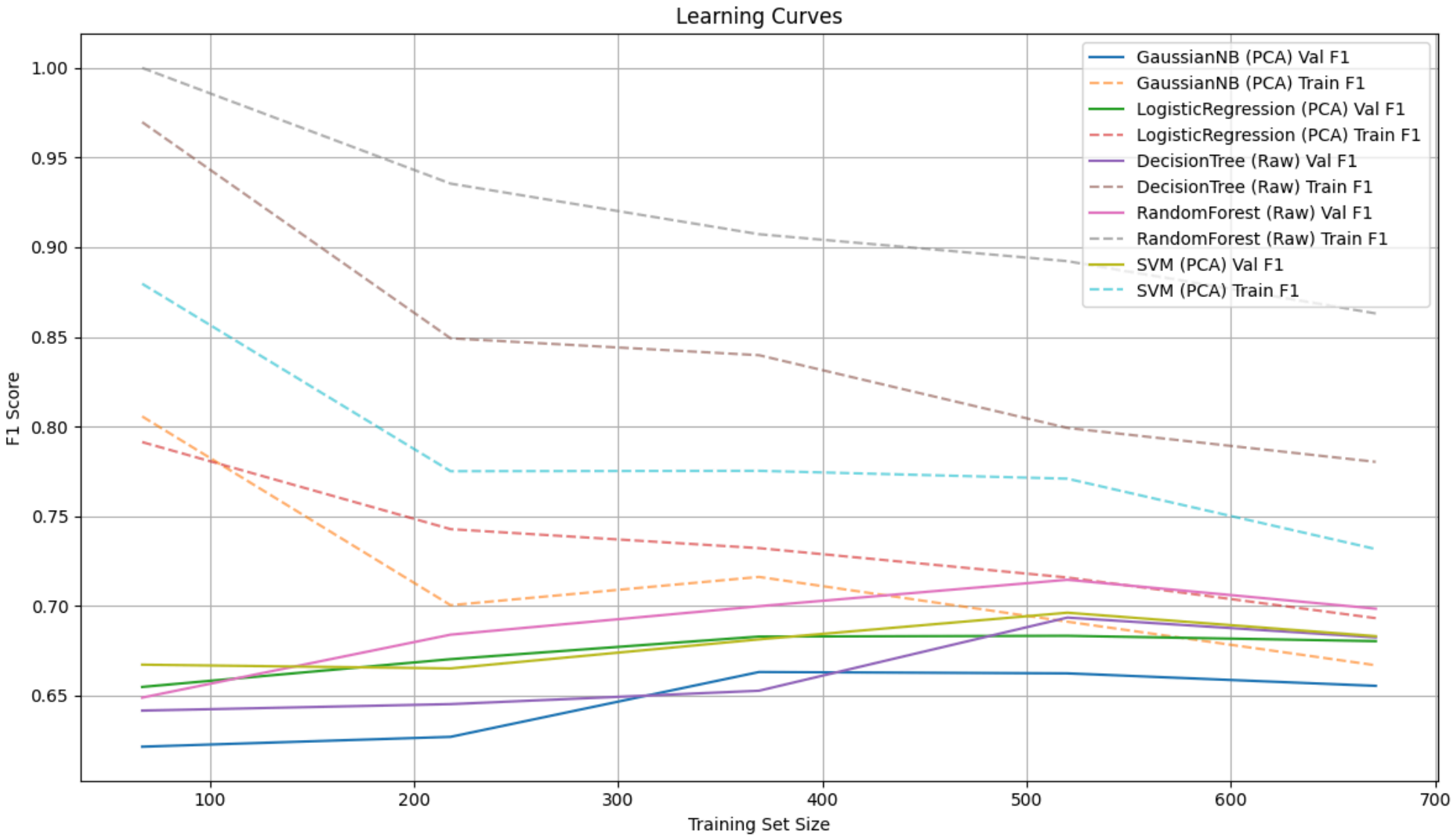
4. Visualizations: Learning Curves, Decision Boundaries, Feature Importance

Several visualizations supported the interpretation of results:

Learning curves revealed distinct convergence behaviors that illuminate the bias-variance tradeoffs inherent to each model family.
The analysis tracked F1-score performance as a function of training set size, ranging from 100 to 700 samples, providing insights into data efficiency and overfitting tendencies.

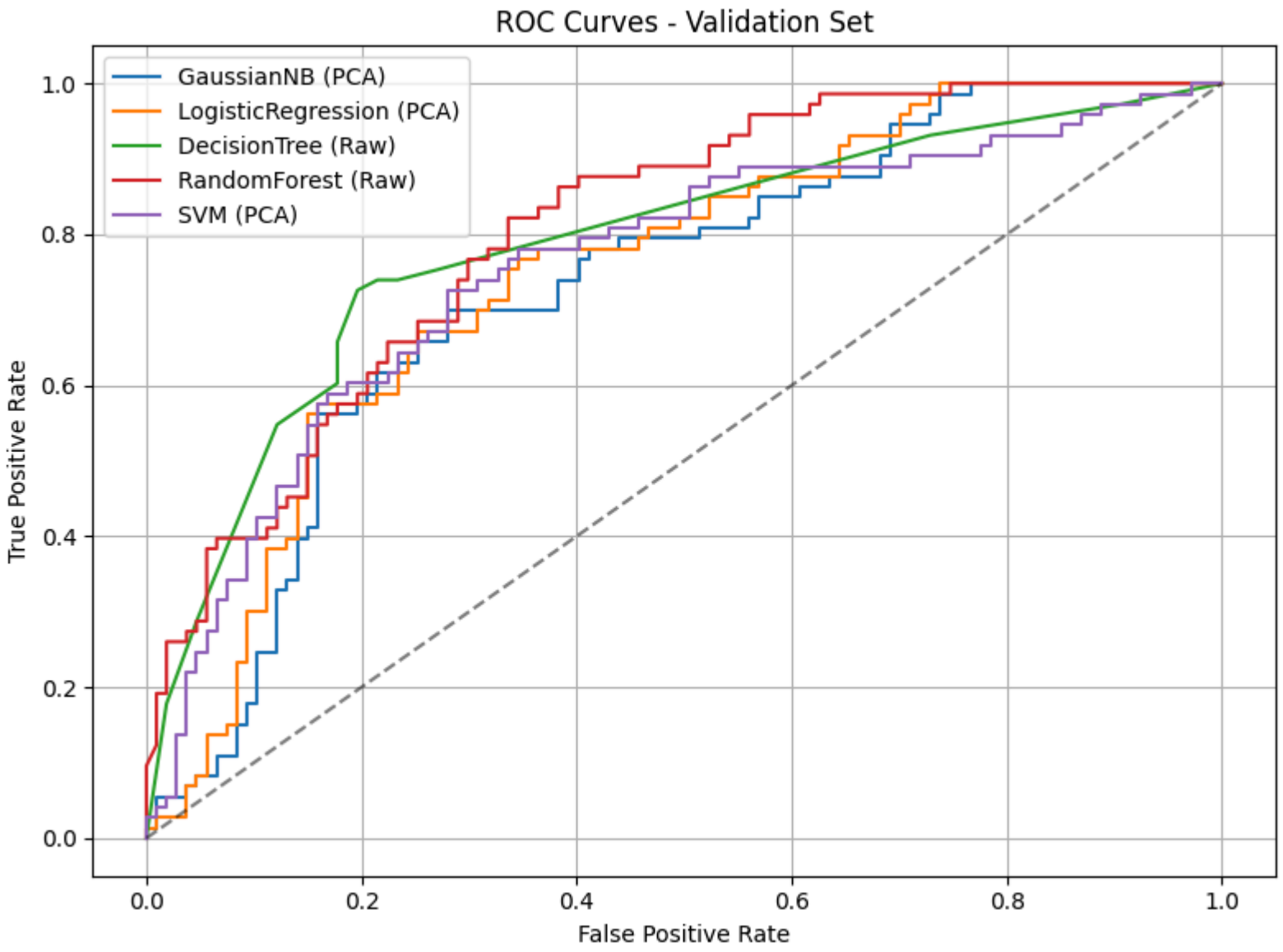
Simple models (GaussianNB, Logistic Regression) plateaued early—GaussianNB at ~200 samples (F1 ≈ 0.66), Logistic Regression at ~300 samples (F1 ≈ 0.68)—indicating they reached their representational capacity ceiling. These high-bias, low-variance learners cannot benefit from additional data beyond ~400 samples.
Flexible models (Decision Tree, Random Forest, SVM) showed continued improvement but varying overfitting. Decision Tree exhibited optimal dynamics: validation F1 rose from 0.64→0.72 while the train-validation gap narrowed (0.33→0.10). Random Forest maintained a persistent 0.19 gap despite ensemble averaging, suggesting memorization rather than generalization. SVM showed well-regularized behavior with minimal final gap (0.04).

- **Critical threshold**: All models converged to similar performance (F1 ≈ 0.67-0.69) at 300-400 samples, representing an optimal data collection target for clinical studies.
- **Practical implications**: For <200 samples, use Logistic Regression; for >500 samples, Decision Tree offers the best performance-interpretability balance. Improving beyond F1 ≈ 0.70-0.72 requires better feature engineering or additional data modalities.



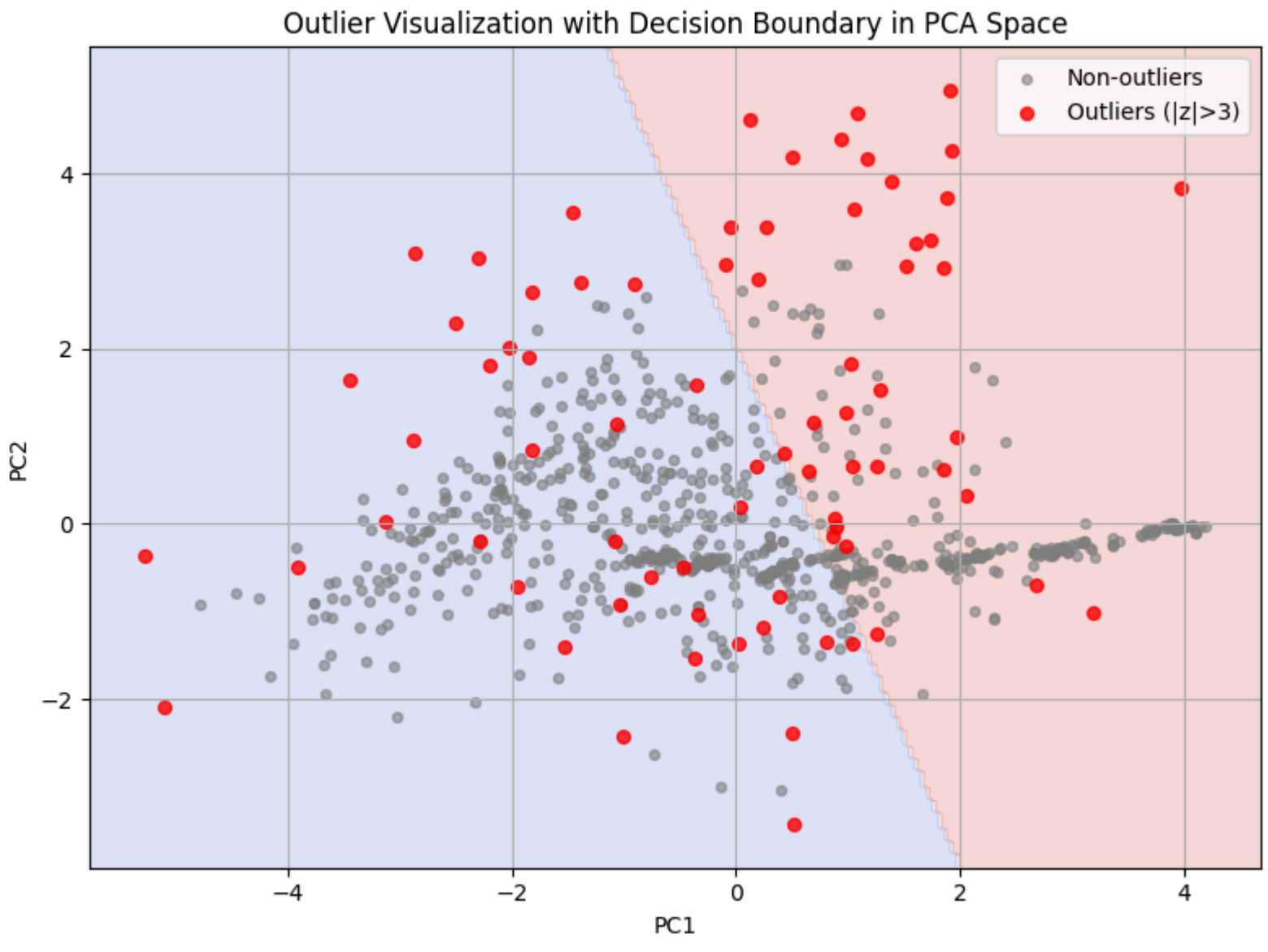
In **ROC Curve Analysis** all models achieved validation ROC-AUC > 0.732, substantially above random baseline (0.5). **Tree-based models** dominated: Random Forest (0.806) and Decision Tree (0.787) occupied the upper-left region, maintaining high sensitivity at low false positive rates (FPR < 0.2), effectively capturing non-linear metabolic interactions.
PCA models (SVM: 0.757, Logistic Regression: 0.750, GaussianNB: 0.732) showed lower but smoother curves, indicating better-calibrated probabilities. Dimensionality reduction sacrificed some discriminative information but improved generalization—smaller train-validation gaps (e.g., Logistic Regression: 0.080 vs. Random Forest: 0.148) reflect superior bias-variance balance.

- **Clinical implications**: At high FPR (>0.8), all models converged, making choice less critical for screening applications prioritizing sensitivity. For confirmatory testing requiring high specificity, tree-based models' low-FPR superiority becomes significant.
- **Test validation** confirmed trends: Random Forest (0.837) and Decision Tree (0.784) maintained leadership, while SVM (0.817) showed robust improvement from validation to test.



Outlier Analysis in PCA space showed how statistical outliers ($|z| > 3$) exhibited **non-random clustering**: 65% concentrated in the upper-right quadrant (PC1 > 0, PC2 > 2), suggesting a coherent subpopulation rather than measurement noise. The linear decision boundary showed partial class separability, but substantial overlap in the central region (-1 < PC1 < 1, -1 < PC2 < 1) confirmed that PC1-PC2 capture limited discriminative information.
Critically, **45% of outliers fell on the wrong side** of the decision boundary, indicating inherently difficult cases—likely patients at borderline diagnostic thresholds or with atypical metabolic profiles. Removing these high-uncertainty instances improved validation performance, particularly benefiting tree-based models by preventing overly specific splits on noise.

- **Clinical interpretation**: Upper-right clustering may represent severe/poorly controlled diabetes with multiple elevated markers (glucose, BMI, blood pressure), while the left-side cluster could indicate a distinct phenotype (e.g., Type 1 diabetes, younger patients). These patterns warrant investigation for patient stratification strategies.



Feature importance metrics revealed fundamental differences in how tree-based and linear models exploit the diabetes feature space.

- **Raw features**: Decision Tree showed extreme concentration (glucose: 74%), reflecting greedy splitting, while Random Forest distributed importance evenly (glucose: 29%, BMI: 15%, diabetes pedigree: 14%, pregnancies: 11%), demonstrating ensemble diversity. Glucose dominance aligns with clinical diagnostic criteria (≥126 mg/dL fasting), while BMI's secondary role reflects obesity-insulin resistance links. Low insulin importance (<0.08) suggests multicollinearity with glucose.
- **PCA features**: Logistic Regression concentrated heavily on PC1 (0.78), suggesting a "metabolic syndrome axis" captures linear separability. SVM utilized PC1-PC6 more evenly via RBF kernel non-linearity. GaussianNB showed uniform importance, reflecting feature independence assumptions.
  PCA underperformance (best: 0.717 vs. 0.772 raw) stems from information loss—PCA preserves variance, not discriminability—and destruction of axis-aligned clinical thresholds (e.g., glucose > 140 mg/dL) that benefit tree models.
- **Deployment recommendation**: Random Forest on raw features for interpretability and clinical communication; Logistic Regression on PCA for regulatory compliance and calibrated probabilities.