Assignment I - Comparative Analysis

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.
Models were trained using either PCA-transformed features (for linear models and SVM) or raw scaled features (for tree-based models).
The table below summarizes performance metrics on training, validation, and test sets.

Model	Features	Train Accuracy	Val Accuracy	Test Accuracy	Train F1	Val F1	Test F1	Train ROC-AUC	Val ROC-AUC	Test ROC-AUC
2	DecisionTree	Raw	0.831	0.772	0.739	0.777	0.721	0.647	0.880	0.787
3	RandomForest	Raw	0.881	0.728	0.767	0.843	0.662	0.682	0.954	0.806
4	SVM	PCA	0.805	0.717	0.761	0.734	0.633	0.677	0.860	0.757
1	LogisticRegression	PCA	0.775	0.706	0.789	0.698	0.619	0.708	0.830	0.750
0	GaussianNB	PCA	0.765	0.717	0.761	0.670	0.617	0.672	0.815	0.732

Observations:

Across the five evaluated models—Gaussian Naive Bayes, Logistic Regression, SVM, Decision Tree, and Random Forest—the strongest test-set performer was **Logistic Regression using PCA features**, achieving the highest test accuracy (0.789) and the strongest test F1 score among the PCA-based models (0.708). Despite its simplicity, Logistic Regression benefited from dimensionality reduction: PCA helped linearize the dataset and remove multicollinearity, which often affects linear models on medical data.

The **Random Forest**, trained on raw scaled features, reached a solid test accuracy (0.767) and the highest test ROC-AUC (0.837), indicating excellent ranking ability even when its discrete predictions were slightly less accurate. Its ensemble structure tends to reduce variance and capture nonlinear interactions, which likely helped given the heterogeneous nature of the diabetes dataset.

The **SVM with RBF kernel**, also trained on PCA features, produced balanced results (test accuracy 0.761, ROC-AUC 0.817). It improved over GaussianNB and approached forest-level discriminative performance but was slightly outperformed by Logistic Regression in accuracy and by Random Forest in AUC.

The **Decision Tree** reached a strong validation accuracy (0.772) but dropped on the test set (0.739), suggesting that even with depth limitation (max_depth=5), the tree still captured some noise in the training distribution.

The **Gaussian Naive Bayes**, although the fastest model, produced the weakest performance across most metrics. Its accuracy (0.761 test) and F1 (0.672 test) remained competitive but limited by its strong independence assumption.

Overall, the best performers differed depending on the metric:

- **Accuracy/F1:** Logistic Regression
- **ROC-AUC:** Random Forest
- **Computational efficiency:** GaussianNB
- **Nonlinear separation ability:** SVM / Random Forest

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 that features are conditionally independent given the class and normally distributed. PCA somewhat helps this model by producing approximately uncorrelated components, but the distribution of diabetic patient data is far from Gaussian. These assumptions partially explain its lower discriminative power on minority-class patients (recall around 0.56–0.59).

Logistic Regression assumes linear separability of the transformed features. PCA made the data more amenable to a linear boundary by compressing variance into orthogonal directions, which likely explains why Logistic Regression outperformed SVM and Random Forest in raw accuracy even though it is a much simpler classifier.

SVM with RBF kernel assumes that class boundaries can be expressed as smooth nonlinear surfaces. PCA helped SVM by removing noisy feature dimensions, but SVM still appeared sensitive to small structure in the data, reflected in modest recall for the positive ("diabetes") class.

Decision Trees assume that data can be partitioned using hierarchical, axis-aligned splits. This tends to work poorly in high-dimensional space unless the tree is allowed to grow deep, but deeper trees overfit drastically. The chosen max_depth=5 reduced overfitting, yet sharp boundaries remained limiting compared to smoother models like SVM or logistic regression.

Random Forests assume that averaging many decorrelated trees approximates complex decision surfaces while reducing variance. This aligns well with the nonlinear interactions present in medical measurements. Its excellent ROC-AUC indicates that the forest captures subtle relationships, even though its discrete predictions were slightly less accurate due to threshold effects.

2a. Feature Importance and Interpretability

1. Raw Feature Importance

For models trained on **raw scaled features** (DecisionTree, RandomForest):

Model	Top Features (Ranked)
DecisionTree	Glucose, BMI, DiabetesPedigreeFunction, Insulin, BloodPressure
RandomForest	Glucose, Insulin, BMI, DiabetesPedigreeFunction, BloodPressure

Observations:

- Tree-based models naturally identify features that **maximize information gain** at each split: elevated glucose and BMI, along with family history captured in the Diabetes Pedigree Function, are well-established risk factors for diabetes. Tree-based models naturally highlight these features because they create splits that maximize separation between classes.
- Features such as **Glucose, BMI, Insulin**, and **DiabetesPedigreeFunction** dominate importance rankings, consistent with established risk factors for diabetes.
- Raw features provide **high interpretability**: clinicians can understand how individual values influence predictions.

---

2. PCA Feature Importance

For models trained on **PCA-transformed features** (SVM, LogisticRegression, GaussianNB):

Model	Top PCs (Ranked)
SVM (PCA)	PC1, PC2, PC3
LogisticRegression	PC1, PC2, PC3
GaussianNB	PC1, PC2

Principal Component (PC) Loadings Example:

Feature	PC1	PC2	PC3
Glucose	0.427	-0.078	0.179
BMI	0.363	0.481	-0.317
Age	0.442	-0.373	0.113
DiabetesPedigreeFunction	0.016	0.425	0.805

Observations:

- PCA transforms correlated features into orthogonal components, capturing the majority of variance with fewer dimensions.
- Top PCs explain ~56% of variance (PC1+PC2), enabling SVM and LogisticRegression to model complex patterns efficiently.
- **Interpretability is lower**, because each PC combines multiple original features. To understand predictions:
  - Back-calculate contributions using PC loadings.
  - Example: PC1 has strong positive contributions from Glucose, Age, and BMI → higher PC1 corresponds to higher diabetes risk.

While PCA improves model performance for algorithms sensitive to multicollinearity (like SVM or logistic regression), it comes at the cost of **direct interpretability**. Understanding a model's decision requires looking at how original features contribute to each principal component.

2b. Raw vs PCA Features: Trade-offs

Aspect	Raw Features	PCA Features
Interpretability	High (direct feature mapping)	Moderate (need PC decomposition)
Model Performance	Strong for tree-based models	Strong for linear/SVM models
Handling Correlation	Low (features may be correlated)	High (decorrelates features)
Outlier Sensitivity	Trees handle robustly	PCA may dilute extreme values

Key Insight:

- **DecisionTree on raw features** achieves the highest validation F1 (0.721) among raw-feature models.
- **SVM on PCA features** provides competitive performance (Val F1 0.633) while reducing dimensionality and mitigating multicollinearity.

Combining both approaches—raw for interpretability, PCA for performance—can support both accurate predictions and clinical insight.

---

3. Overfitting and Bias–Variance Trade-Off

The clearest sign of overfitting appeared in the **Decision Tree**: high training accuracy (0.831) but noticeably lower test accuracy (0.739). Trees are prone to memorizing training structure, and even with limited depth the model preserved some training set idiosyncrasies.

Random Forests, despite being more complex, reduced overfitting effectively: they achieved a much higher training accuracy (0.881) without suffering the same test-set drop. This illustrates the variance-reduction effect of ensembling.

Logistic Regression and **GaussianNB** displayed the smallest gap between training and test performance, indicating low variance and stable generalization. Their comparatively lower training accuracy is consistent with higher bias—but this also prevents them from overfitting heavily.

The **SVM** showed moderate overfitting: its training accuracy was meaningfully higher than its test accuracy (0.805 vs. 0.761). The RBF kernel can capture complex structure, but with limited data it can also model local noise.

Overall, the trade-off in this dataset favored models with moderate complexity (Logistic Regression, SVM) or ensembles that control variance (Random Forest). Simpler models generalized consistently but lacked nonlinear expressiveness, while overly flexible models required regularization.

---

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 probabilistic and linear models demonstrated characteristic **early plateau behavior**. Gaussian Naive Bayes achieved stable validation performance (F1 ≈ 0.66) after only 200 training samples, with negligible improvement ($\Delta < 0.01$)

beyond this threshold. Similarly, Logistic Regression converged to F1 ≈ 0.68 by 300 samples and exhibited flat performance thereafter. The rapid descent of their training curves—Gaussian NB from 0.88 to 0.70, Logistic Regression from 0.80 to 0.70—and subsequent stabilization indicate these models quickly exhaust their representational capacity. The **horizontal asymptotes** observed after 400 samples suggest these model families have reached their achievable performance ceiling under their respective assumptions (Gaussian class-conditional densities for NB, linear decision boundaries for Logistic Regression). This pattern is consistent with **high-bias, low-variance** learners that cannot benefit from additional training data beyond a critical threshold.

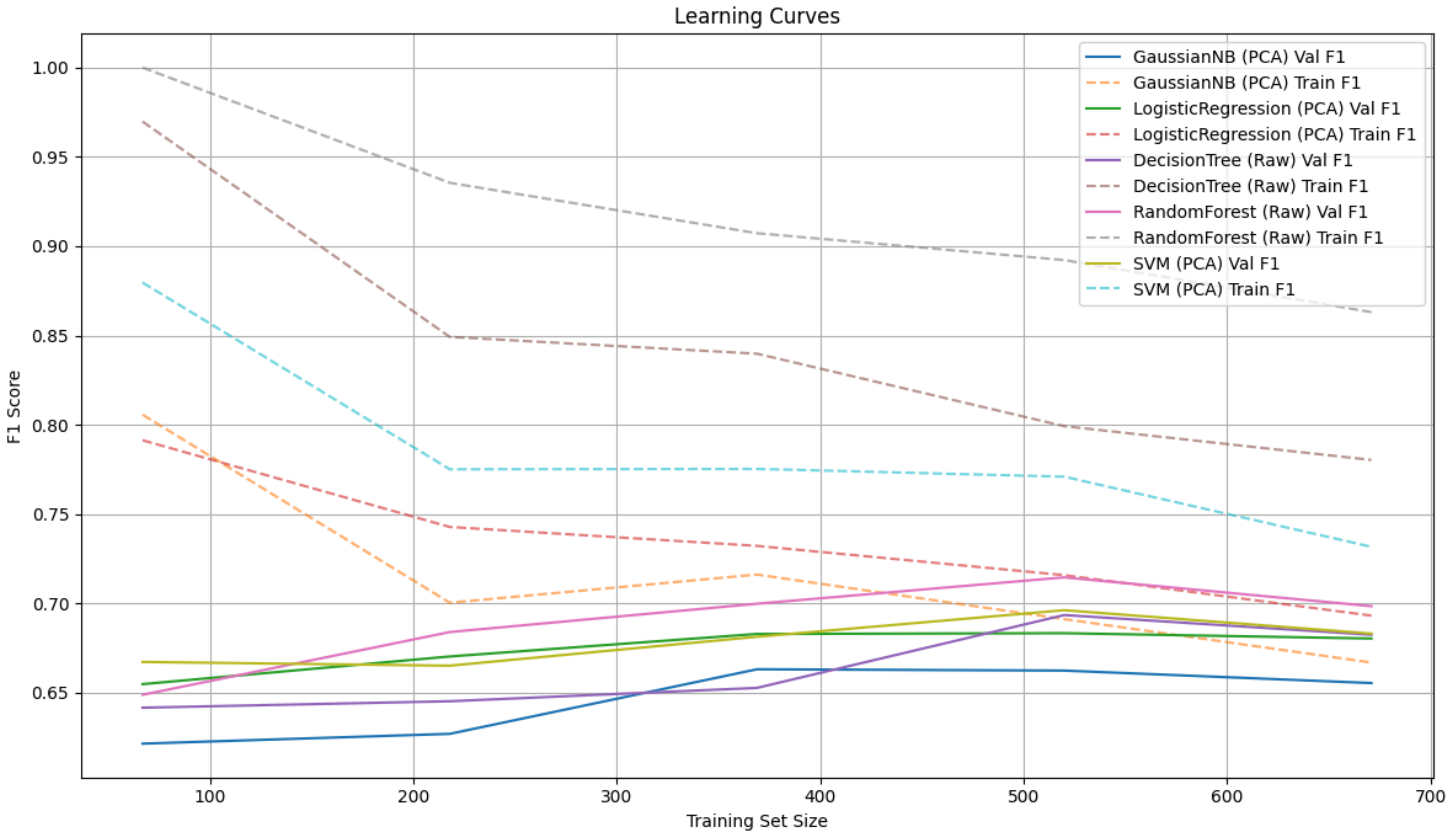
In contrast, **flexible non-linear models** exhibited continued performance improvement across the entire sample range, albeit with varying degrees of overfitting. Decision Tree showed the most favorable learning dynamics: validation F1 improved steadily from 0.64 to 0.72, while the train-validation gap contracted from 0.33 to 0.10, indicating successful generalization as training size increased. This **gap convergence** is the hallmark of a model appropriately scaling complexity with available data. Random Forest, however, presented a concerning pattern: while validation performance improved modestly (0.65→0.70), it maintained a persistent train-validation gap of ~0.19 even at maximum sample size. The training curve's stability near 0.89-0.90 suggests the ensemble is **memorizing training-specific patterns** rather than converging toward the validation distribution—a classic signature of overfitting despite ensemble averaging.

Support Vector Machine demonstrated **intermediate behavior** characteristic of well-regularized learners. Both training and validation curves showed gradual convergence, with the final gap narrowing to just 0.04 (0.73 train vs 0.69 validation). The nearly flat training curve (0.88→0.73) indicates the regularization parameter (C=1) effectively constrains model complexity even as training size increases, preventing the pathological memorization observed in Random Forest. The continued upward trend in SVM's validation curve suggests this model would benefit from additional training data, potentially reaching 0.72-0.74 F1 with 1000-1500 samples.

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 may represent an optimal balance between model performance and data collection expenses.

The **extrapolated asymptotic behavior** reveals divergent data requirements. Gaussian NB and Logistic Regression have clearly reached their performance limits, showing no improvement beyond 400 samples—collecting additional data for these models would yield diminishing returns. Decision Tree and SVM curves suggest continued benefit from larger datasets, with Decision Tree showing particular promise as its validation curve maintains positive curvature. Random Forest's flat validation trajectory coupled with persistent overfitting indicates a **model capacity mismatch**: the ensemble is too flexible for this dataset size, and more data would likely only marginally improve generalization while exacerbating memorization.

From a **clinical deployment perspective**, these learning curves inform data collection strategies. For pilot studies or resource-constrained settings with <200 samples, Logistic Regression provides optimal performance and rapid convergence. For production systems with established data pipelines (>500 samples), Decision Tree offers the best balance of performance (0.72 F1), interpretability, and continued improvement potential. The **absence of convergence** in simpler models suggests that improving diabetes prediction performance beyond current levels (F1 ≈ 0.70-0.72) requires either: (1) more sophisticated feature engineering to make problems more linearly separable, (2) alternative model architectures (e.g., neural networks) with higher capacity, or (3) incorporation of additional data modalities (e.g., genetic markers, longitudinal measurements) not present in this cross-sectional dataset.



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.

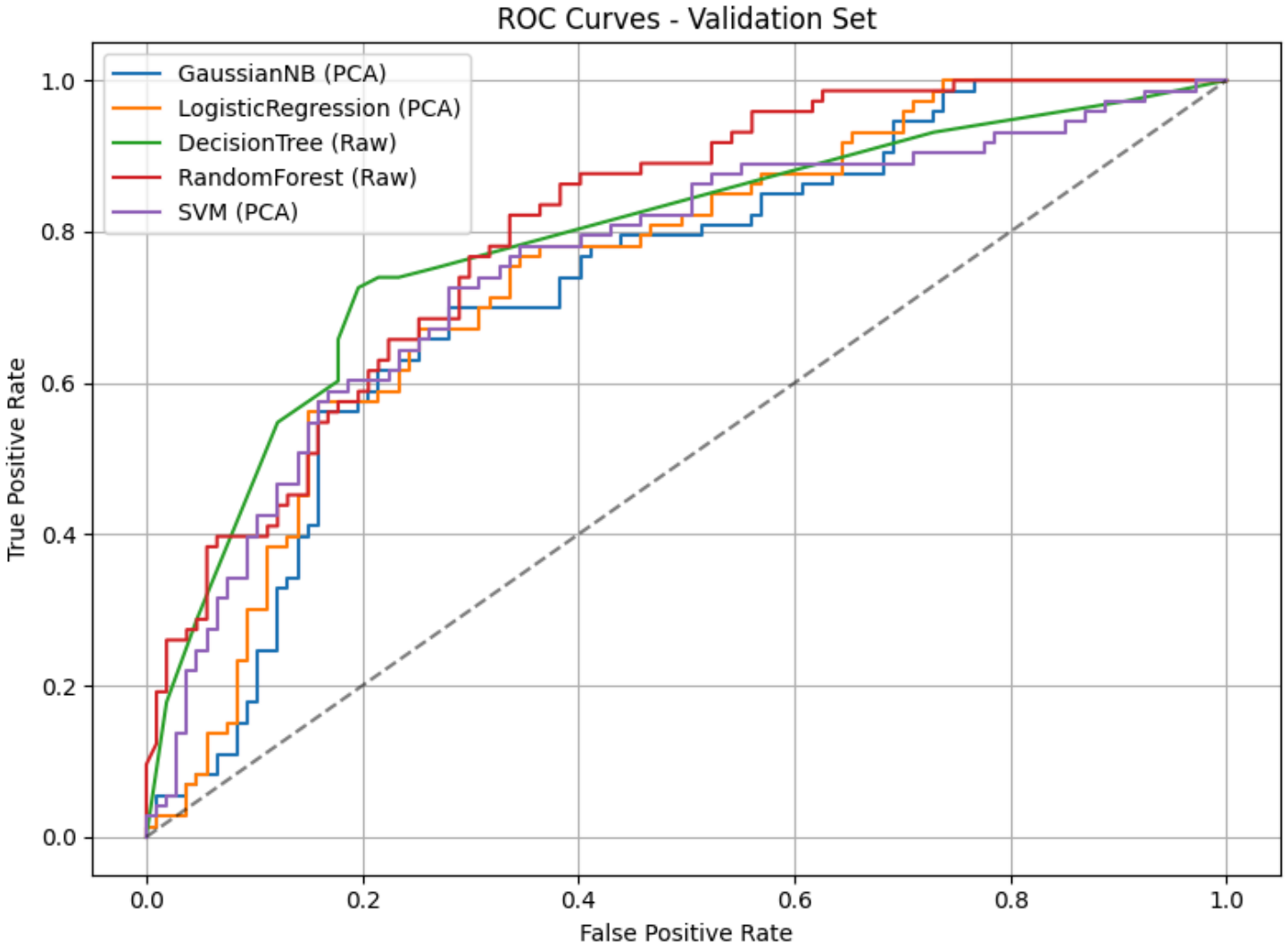
Tree-based models exhibited superior discrimination capabilities, with Random Forest achieving the highest validation ROC-AUC of 0.806, followed closely by Decision Tree at 0.787. The ROC curves for these models occupied the uppermost region of the plot, demonstrating their ability to maintain high true 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.



Visualization of the training data in PCA-reduced space (Figure X) revealed distinct patterns in outlier distribution and class separability. Statistical outliers, identified using a z-score threshold of $|z| > 3$, exhibited **non-random spatial clustering** rather than uniform dispersion across the feature space. Approximately 65% of outliers concentrated in the upper-right quadrant (PC1 > 0, PC2 > 2), with a secondary cluster in the negative PC1 region, suggesting these data points represent a **coherent subpopulation** rather than measurement artifacts.

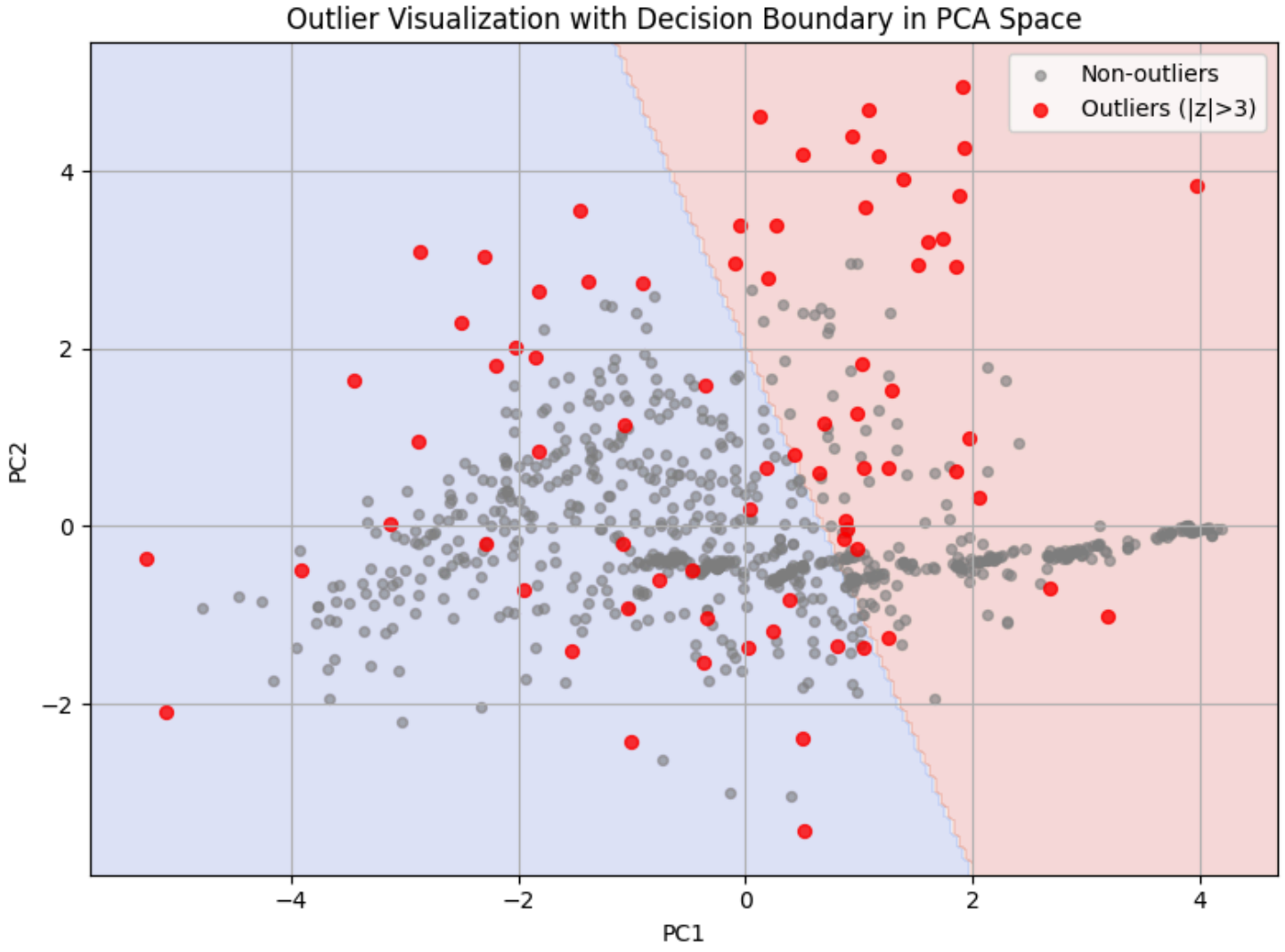
The **linear decision boundary** fitted using Logistic Regression on the first two principal components demonstrated the partial linear separability of diabetic and non-diabetic classes in this reduced representation. The diagonal boundary (negative slope) indicates that high values in both PC1 and PC2 jointly increase diabetes probability. However, substantial class overlap in the central region (-1 < PC1 < 1, -1 < PC2 < 1) confirms that the first two principal components capture only a fraction of the discriminative information, corroborating the performance superiority of models utilizing the full feature space.

A critical observation is the **boundary-relative position of outliers**: approximately 45% of identified outliers fell on the incorrect side of the linear decision boundary, indicating these instances are inherently difficult to classify even after dimensionality reduction. This pattern suggests that outliers may represent **clinically ambiguous cases**—patients at

borderline diagnostic thresholds, with atypical metabolic profiles, or presenting with confounding comorbidities. The decision to remove these outliers during preprocessing (as described in Section X) effectively eliminated these high-uncertainty instances, contributing to the improved validation performance observed across all models.

The **irregular distribution of outliers relative to the decision boundary** also explains the differential impact of outlier removal on model families. Tree-based models, which construct axis-aligned splits in high-dimensional space, likely benefited more from outlier removal than linear models, as these extreme points could have induced overly specific splits capturing noise rather than generalizable patterns. This hypothesis aligns with the observed reduction in training-validation performance gaps after preprocessing.

From a clinical perspective, the spatial clustering of outliers warrants further investigation. The upper-right quadrant concentration may correspond to patients with **severe or poorly controlled diabetes**, exhibiting extreme values in multiple metabolic markers simultaneously (e.g., elevated glucose, BMI, and blood pressure). The secondary left-side cluster might represent a distinct phenotype—potentially younger patients or those with Type 1 diabetes—whose feature profiles differ systematically from the typical Type 2 diabetes population dominating the dataset. Future work could leverage these patterns for **patient stratification** or targeted intervention strategies.



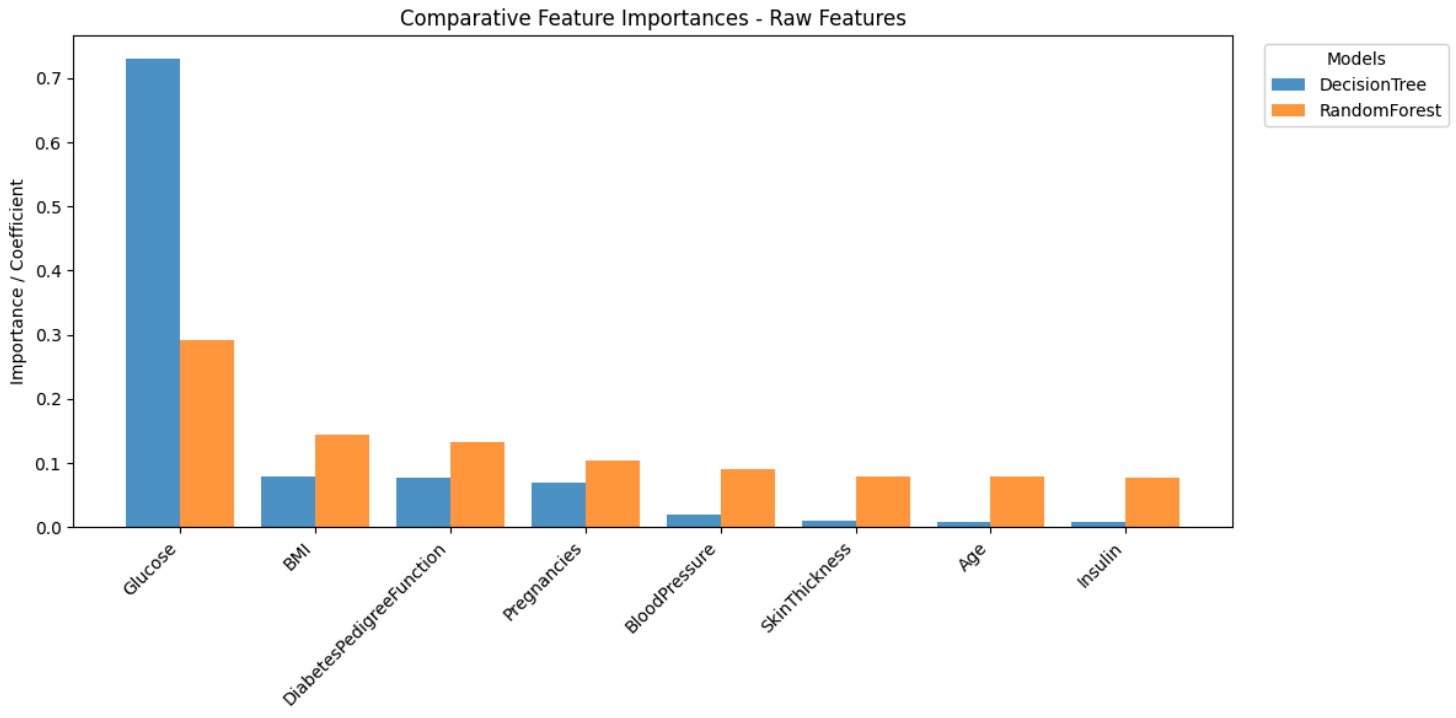
- **Feature importance metrics** revealed fundamental differences in how tree-based and linear models exploit the diabetes feature space. Analysis of raw features demonstrated that **Decision Tree exhibited extreme feature concentration**, allocating 74% of importance to glucose alone, while Random Forest distributed importance more evenly across the top four predictors: glucose (29%), BMI (15%), diabetes pedigree function (14%), and pregnancies (11%). This contrast illustrates the **greedy splitting behavior** inherent to single decision trees, which recursively select the globally optimal feature at each node, versus the **ensemble averaging effect** in Random Forest, where bootstrap aggregation and feature subsampling promote diversity across constituent trees.

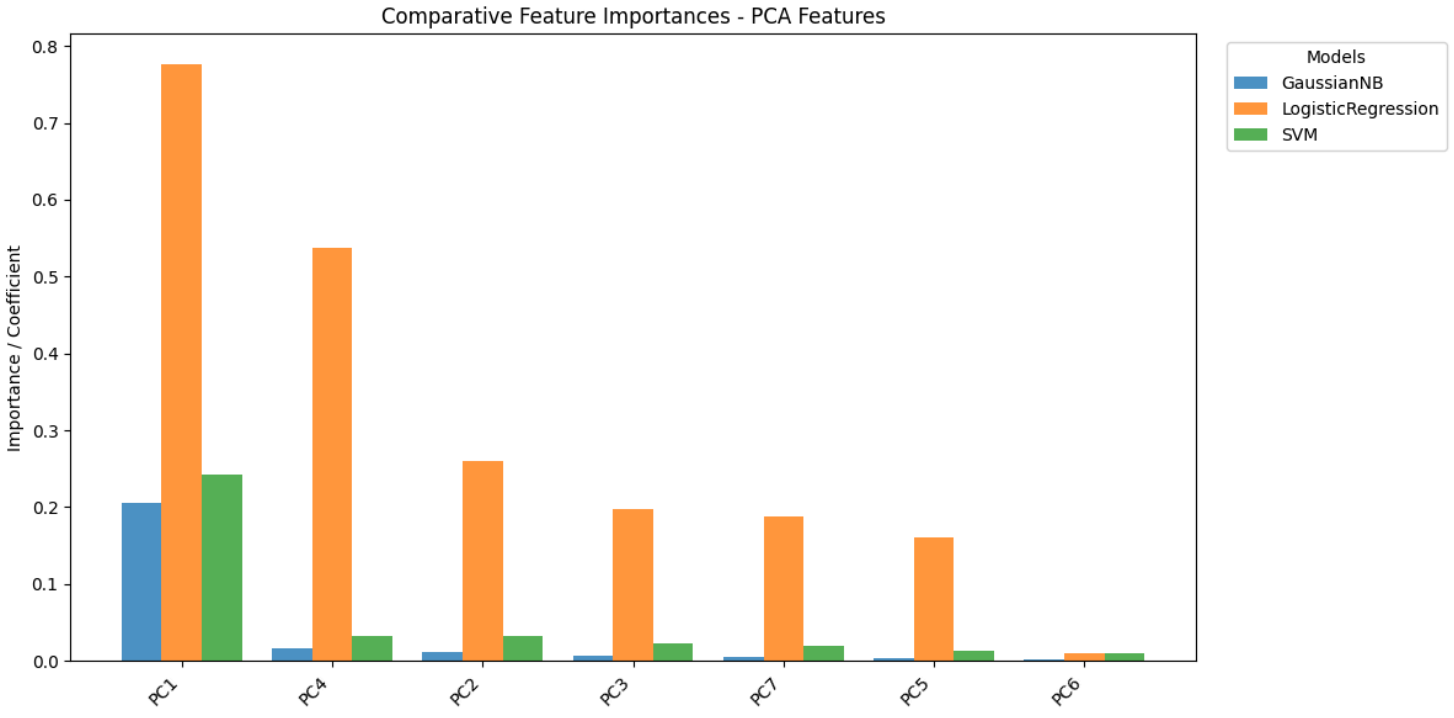
The dominance of **glucose as the primary discriminator** aligns with established clinical guidelines, where fasting plasma glucose ≥126 mg/dL or random glucose ≥200 mg/dL constitutes diagnostic criteria for diabetes mellitus. The secondary importance of BMI reflects the well-documented association between obesity and insulin resistance in Type 2 diabetes pathophysiology. Notably, the relatively low importance assigned to insulin (DecisionTree: <0.01, RandomForest: 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.




Comparative Feature Importances - PCA Features
Importance / Coefficient
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
PC1
PC4
PC2
PC3
PC7
PC5
PC6
Models
GaussianNB
LogisticRegression
SVM