2025

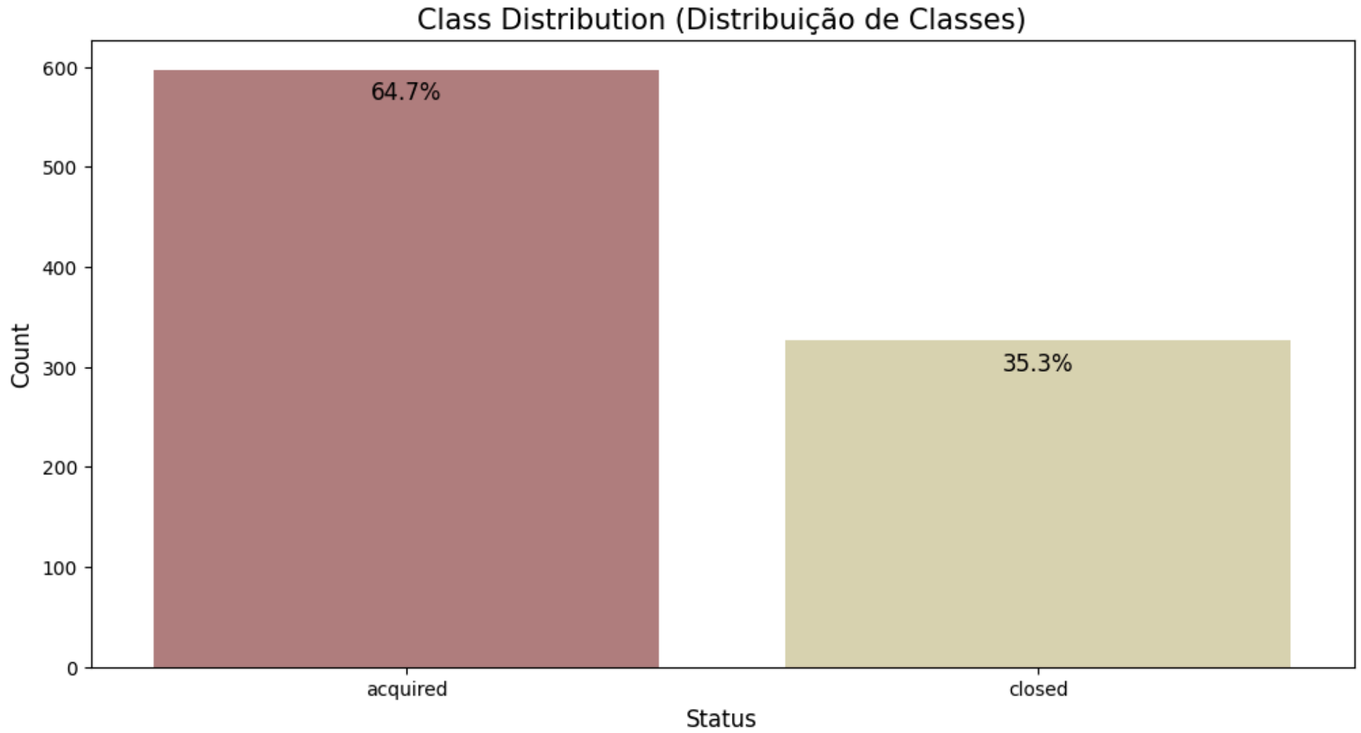
STARTUP SUCCESS Prediction

JULIA MOREIRA, GUILHERME SCHWARZ,
RENAM BIAVATI, MATHEUS FRANCISCO

STATISTICAL

DESCRIPTION

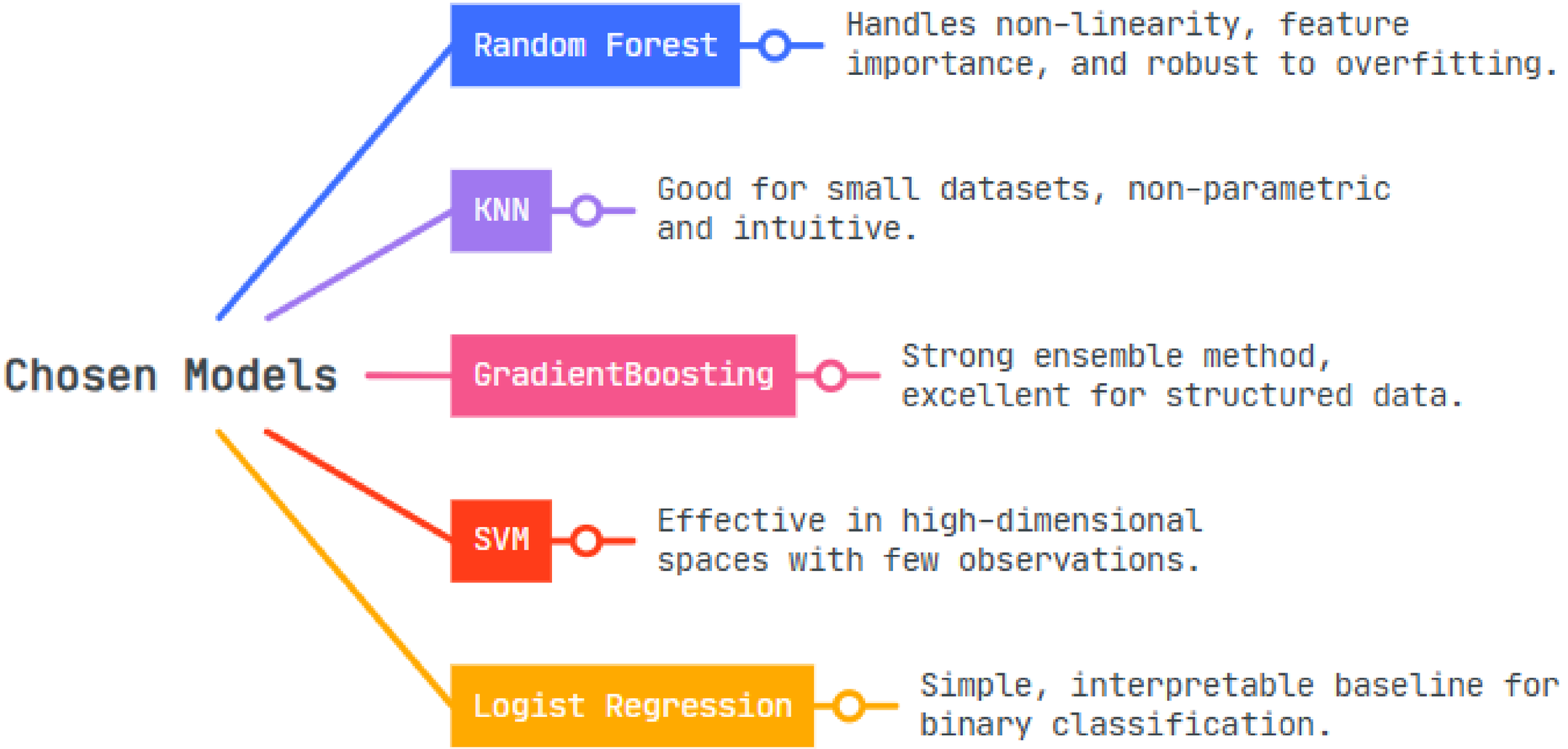
Number of instances: 923 | Number of features: 49 | Number of classes: 2



MACHINE LEARNING

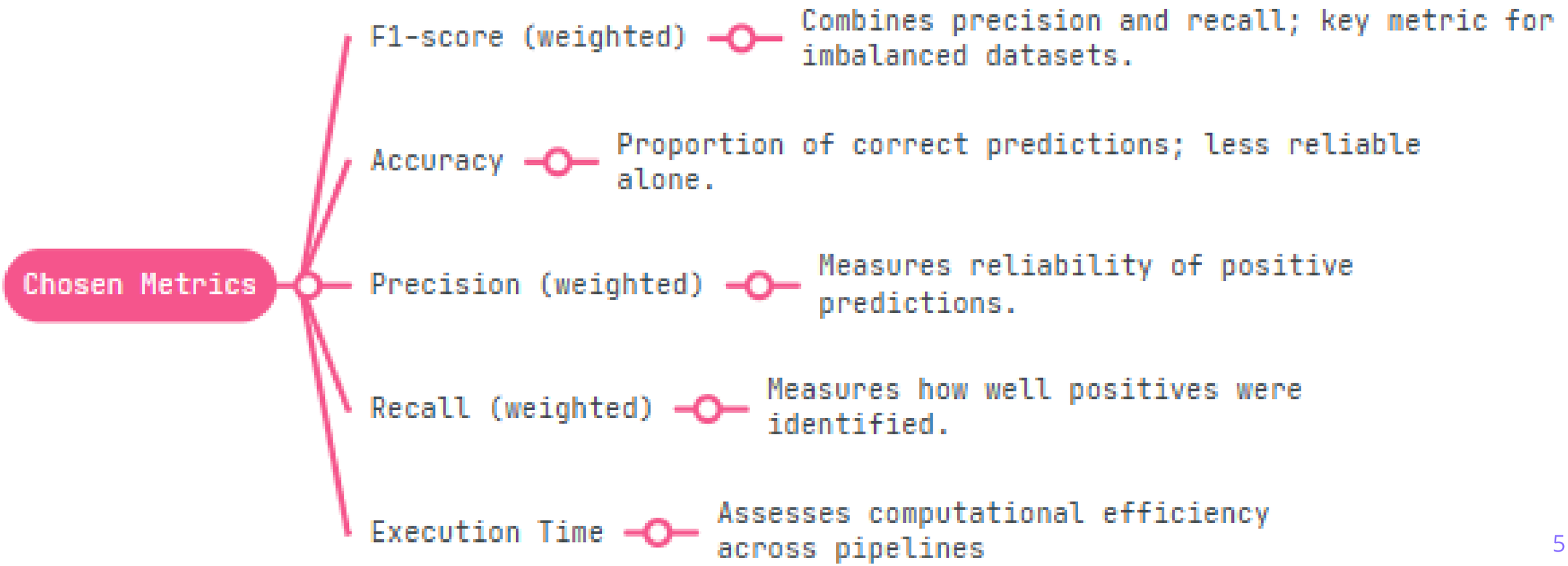
CLASSIFICATION PROBLEM

WHAT'S THE STATUS?

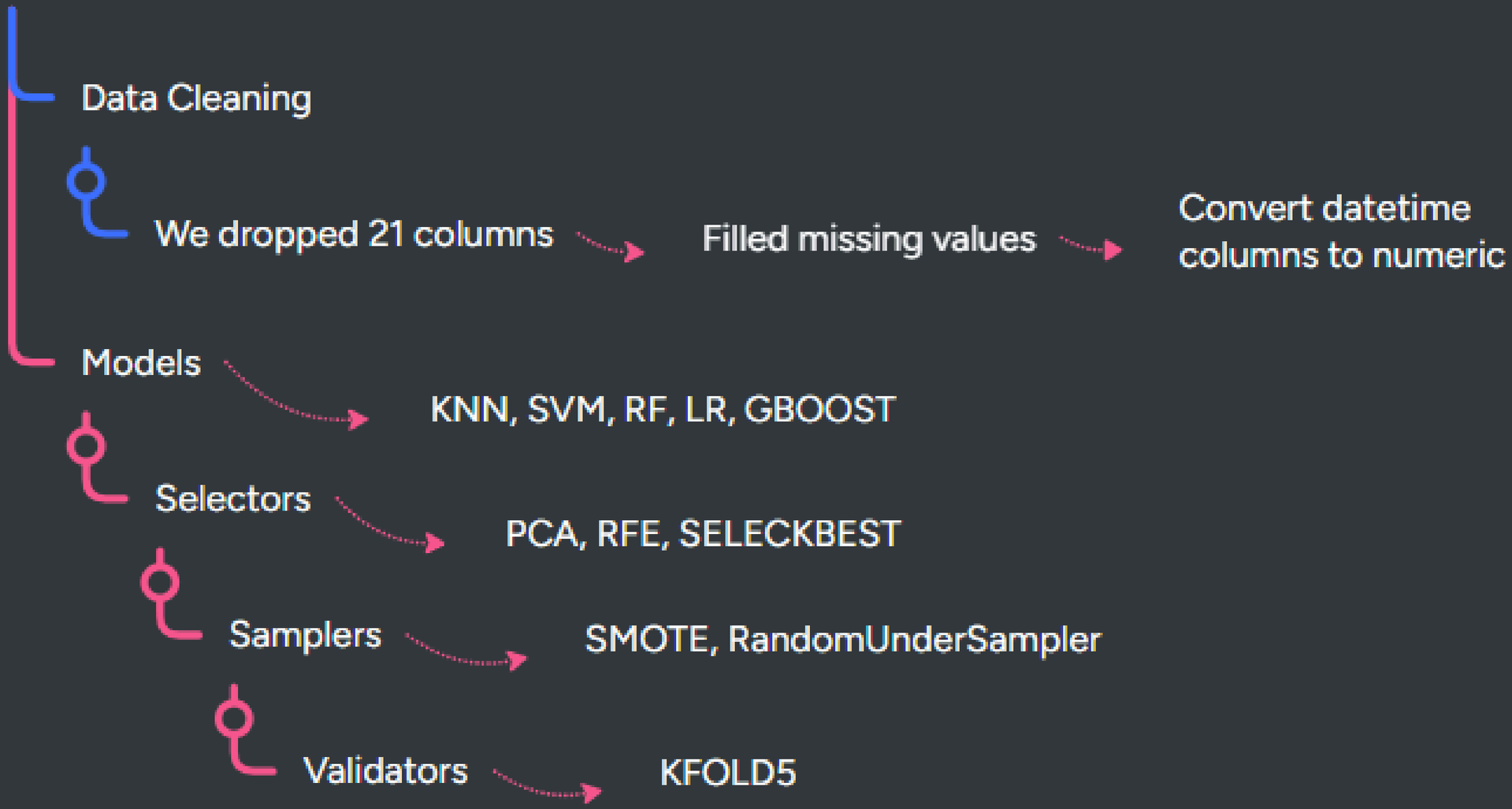

Chosen Models
Random Forest
Handles non-linearity, feature importance, and robust to overfitting.
KNN
Good for small datasets, non-parametric and intuitive.
GradientBoosting
Strong ensemble method, excellent for structured data.
SVM
Effective in high-dimensional spaces with few observations.
Logist Regression
Simple, interpretable baseline for binary classification.

EVALUATION METRICS

We used five metrics to ensure balanced and informative evaluation, especially given the class imbalance (64% vs 36%):



Data Pipeline



We did a loop to test the 19 different combinations in the data pipeline

VALIDATION AND THE OVERFITTING ISSUE

To reduce overfitting and ensure generalization:

- We applied **Stratified K-Fold (5-fold) cross-validation.**
- Class balance was maintained in each fold.
- **Leave-One-Out was considered but discarded due to inefficiency on this dataset.**
- Plans include testing RepeatedStratifiedKFold in future work.

This validation ensured reliable performance estimation across all 19 model configurations that we testet.

RFE was skipped for KNN, SVM and Logistic Regression because these models lack the attributes required for feature elimination (coef_ or feature_importances_).

PCA was avoided with tree-based models (RandomForest, GradientBoost) since it removes the original feature structure they rely on for performance.

RESULTS ANALYSIS

WORST COMBINATIONS

COMPUTATIONAL EFFICIENCY

RandomForest + RandomUnderSampler + RFE → F1 = 0.99, Time = 632.37s

GradientBoost + RandomUnderSampler + RFE → F1 = 0.99, Time = 802.90s

RandomForest + SMOTE + RFE → F1 = 0.99, Time = 805.28s

GradientBoost + SMOTE + RFE → F1 = 0.99, Time = 1534.00s

RFE WITH A HAVIER MODEL IS TERRIBLE

WORST COMBINATIONS

CLASSIFICATION METRICS

LogisticRegression + SMOTE + PCA → F1 = 0.83, Time = 0.98s

KNN + SMOTE + PCA → F1 = 0.78, Time = 1.06s

SVM + RandomUnderSampler + SelectKBest → F1 = 0.99, Time = 1.12s

SVM + RandomUnderSampler + PCA → F1 = 0.81, Time = 1.16s

SVM + SMOTE + PCA → F1 = 0.83, Time = 1.10s

PCA MAYBE REMOVES IMPORTANT NON-LINEAR RELATIONS

WHY PCA UNDERPERFORMED?

PCA REDUCED MODEL PERFORMANCE ACROSS THE BOARD

It performs a linear projection, which discards interactions and nonlinearities.

This affected distance-based models like KNN and linear models like Logistic Regression, which rely heavily on the original feature space.

Result: **The 3 lowest-performing combinations all used PCA.**

Conclusion: PCA should be avoided in this context — better alternatives like SelectKBest preserve informative structure.

BEST COMBINATION

LOGISTICREGRESSION + RANDOMUNDERSAMPLER + SELECTKBEST

→ F1 = 0.99, Accuracy = 0.99, Precision = 0.99, Recall = 0.99, Time = 0.63s

KNN + RANDOMUNDERSAMPLER + SELECTKBEST

→ F1 = 0.98, Accuracy = 0.98, Precision = 0.98, Recall = 0.98, Time = 0.72s

THANKS!