ANALYZING LIFESTYLE AND DEMOGRAPHIC RISK FACTORS OF DIABETES WITH BRFSS DATA

Arnav Singh (U19589314)
Jyoti Shree (U74678990)
Prem Rajendran (U99248729)
Saneeya Vichare (U75237907)
Sara Alsowaimel (U86273437)

INTRODUCTION

PREDICTING DIABETES USING BRFSS SURVEY DATA

- Diabetes is a major U.S. public health challenge
- BRFSS: largest behavioral risk survey (CDC)
- Goal: classify individuals as
  - Diabetes / No Diabetes / Prediabetes
- **Challenges:**
  - class imbalance
  - self-reported noise
  - hundreds of variables

BRFSS 2024 DATASET OVERVIEW

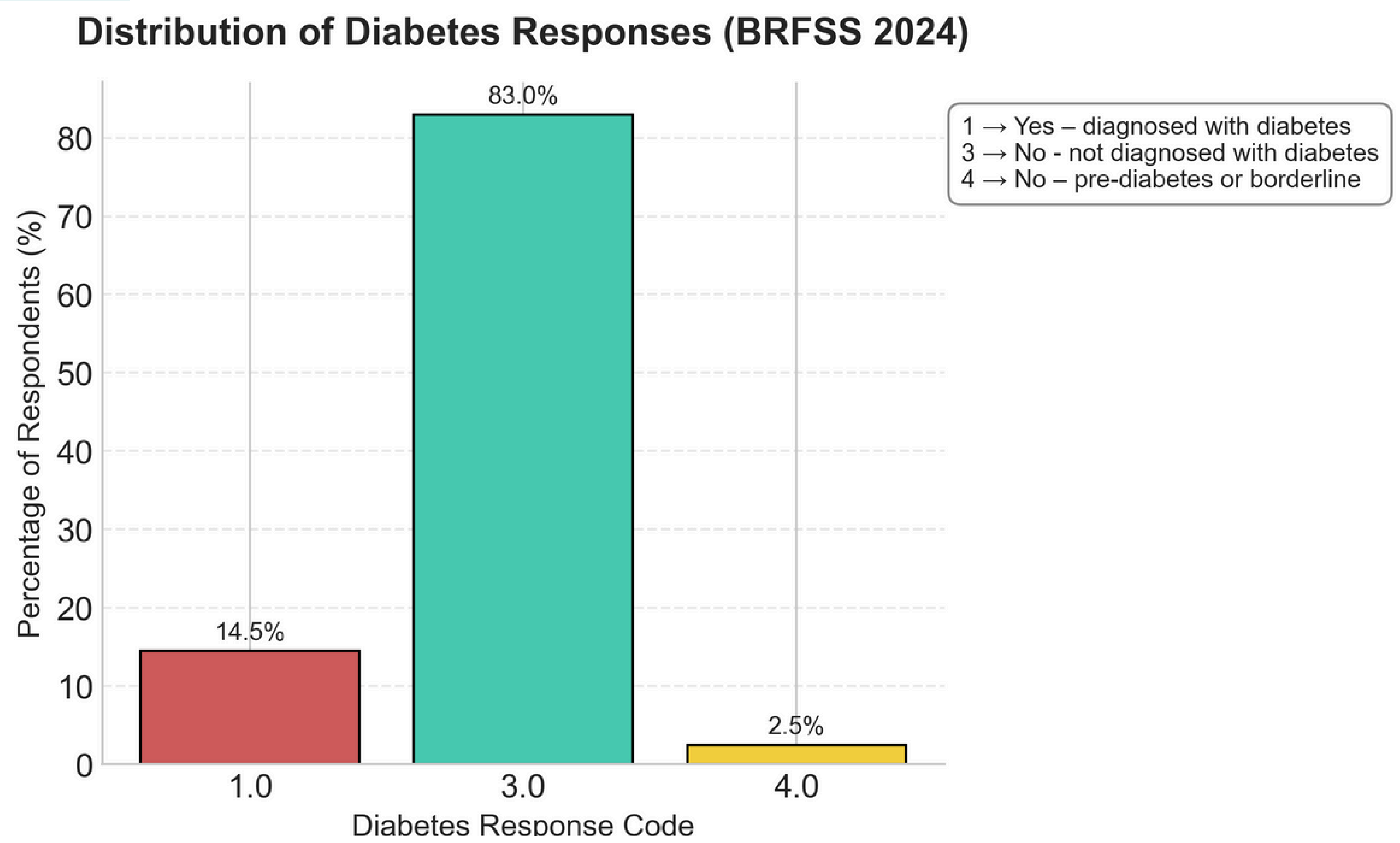
- ~450,000 cleaned samples
- Target variable (DIABETE4) with 3 classes
- 198 engineered + cleaned features
- Categories included:
  - Demographics
  - Lifestyle behaviors (exercise, smoking)
  - Health indicators (BMI, blood pressure, cholesterol)
  - Socioeconomic factors (income, education)

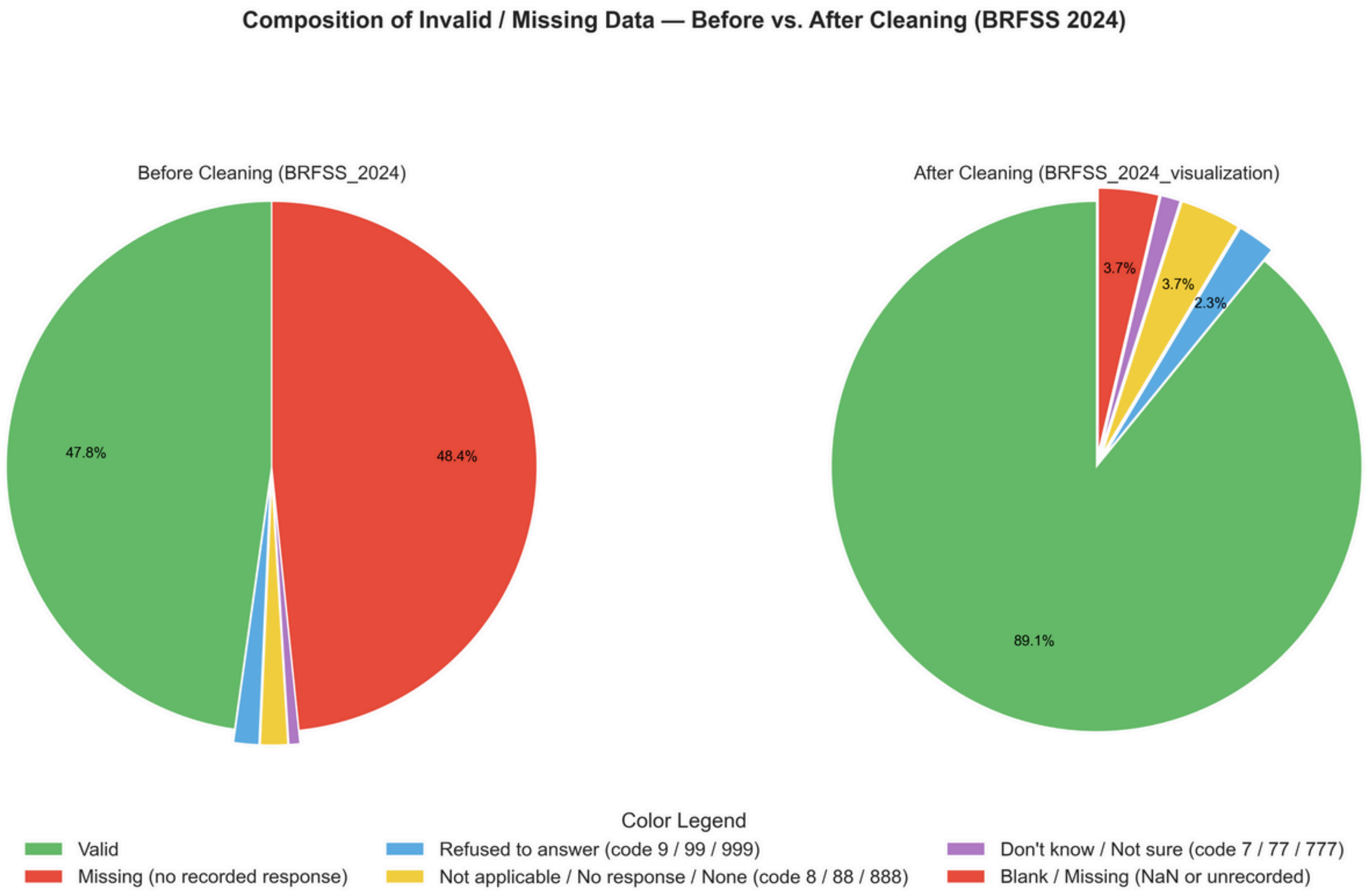
DATA CLEANING & FEATURE ENGINEERING PIPELINE

- Removed invalid BRFSS codes (7, 9, 77, 99 → NaN)
- Dropped features with >30% missing
- Encoded categorical variables
- Handled missing values
- Kept clinically meaningful variables
- Final shape: 198 features × 450K samples

EXPLORATORY DATA ANALYSIS

DATA DISTRIBUTION

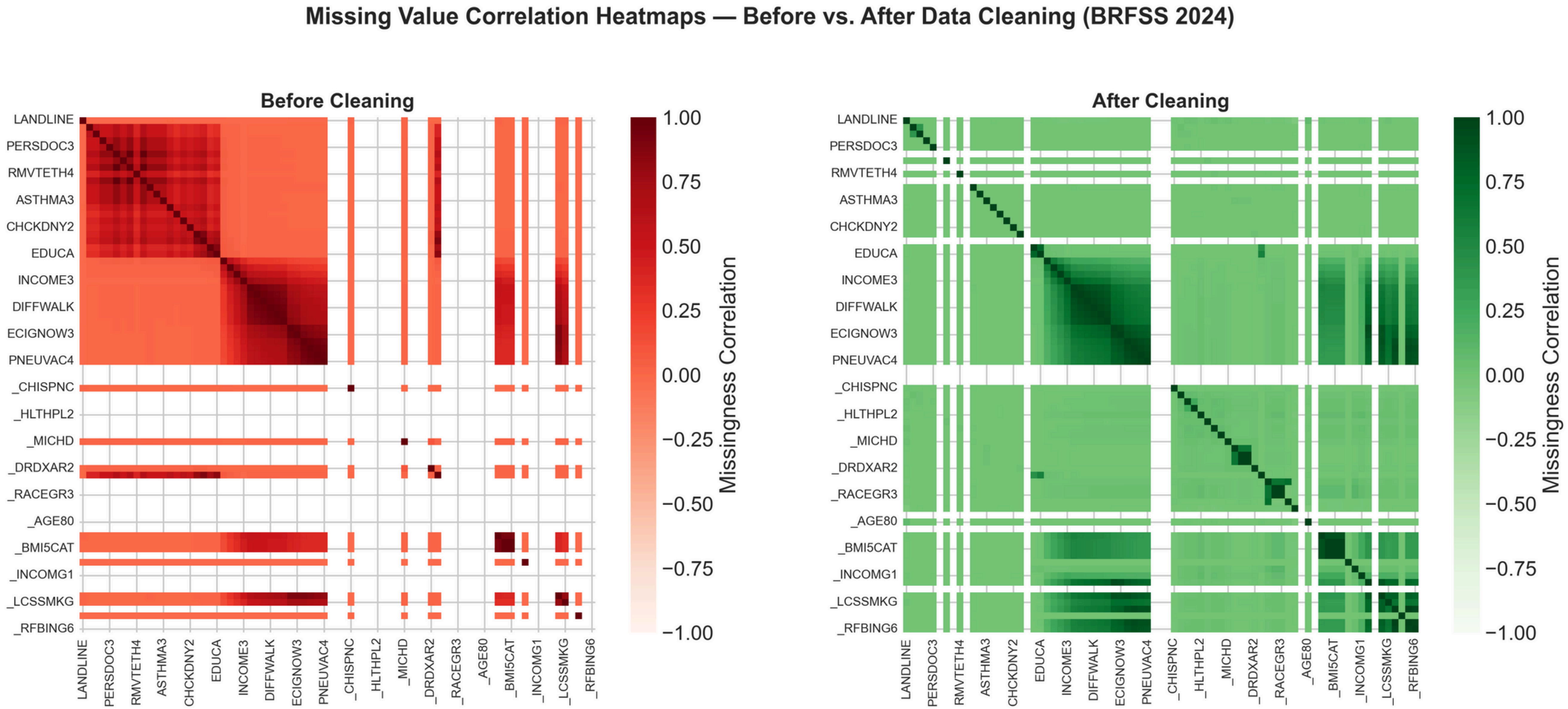




- Majority of respondents report no diabetes diagnosis.
- Only a small fraction fall into diabetic or pre-diabetic categories.
- Strong class imbalance impacts model learning and evaluation.

- High volume of invalid/missing responses before cleaning.
- Cleaning greatly increases the percentage of valid entries.
- Reduces noise and improves dataset reliability.

DATA DISTRIBUTION

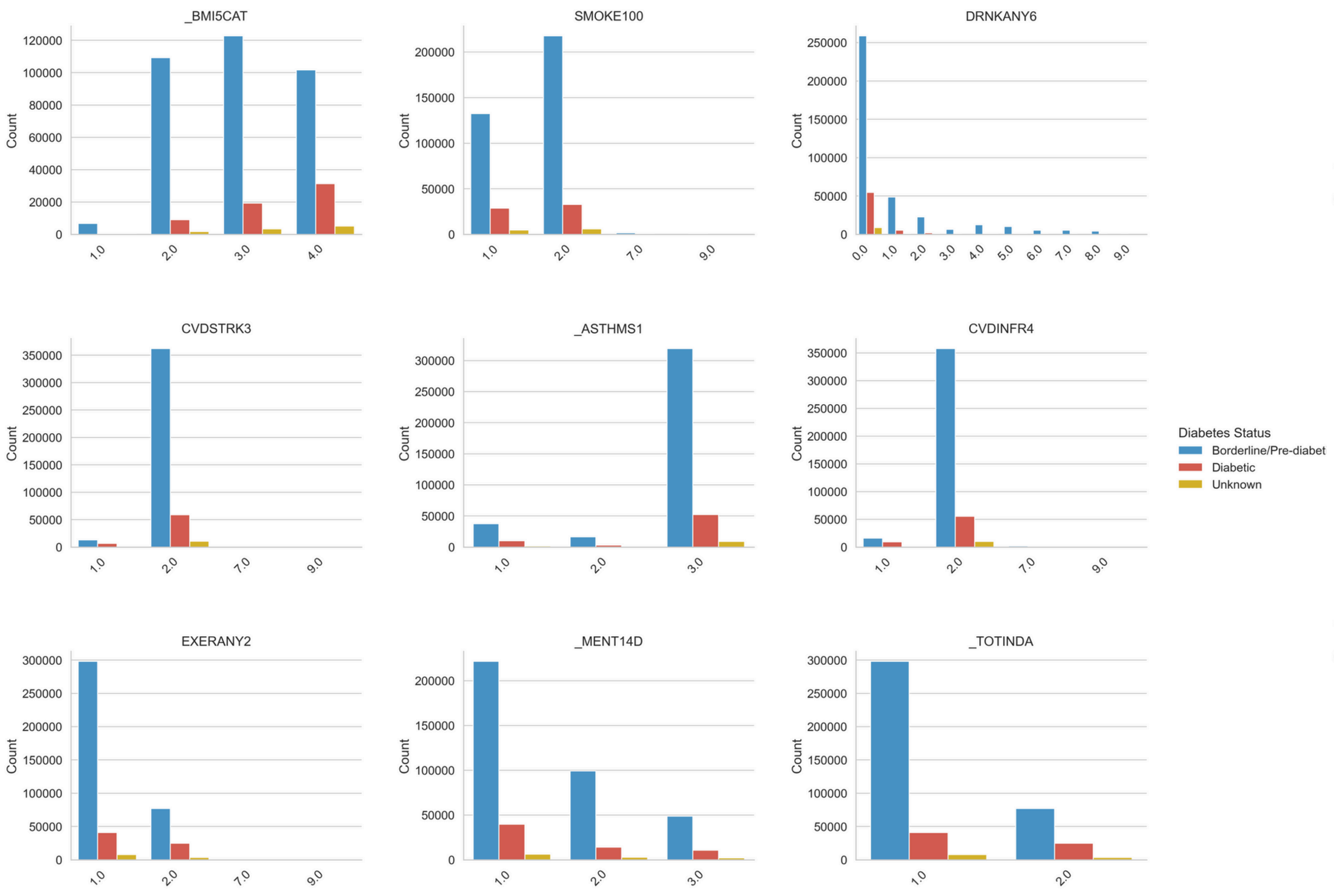


Red / Green Intensity → Strength of correlation in missingness
1.0 → Variables tend to be missing together
0.0 → Independent missingness (no relationship)
-1.0 → Opposite missing patterns

- Before cleaning, several variables show **correlated missingness**.
- After cleaning, missingness patterns shrink or disappear.
- Results in more robust and higher-quality features for modeling.

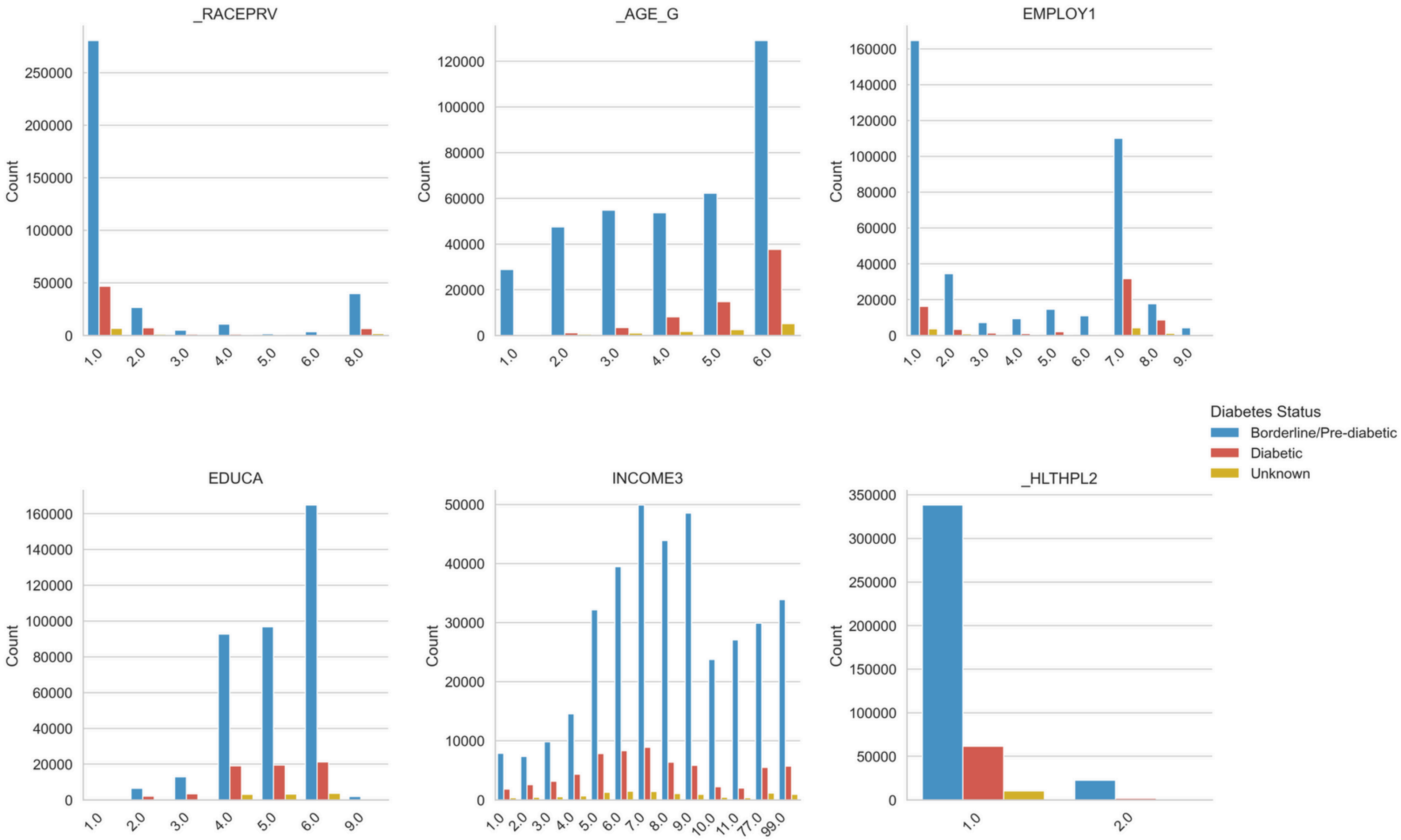
DATA DISTRIBUTION

Distribution of Health & Behavioral Variables by Diabetes Status (BRFSS 2024)



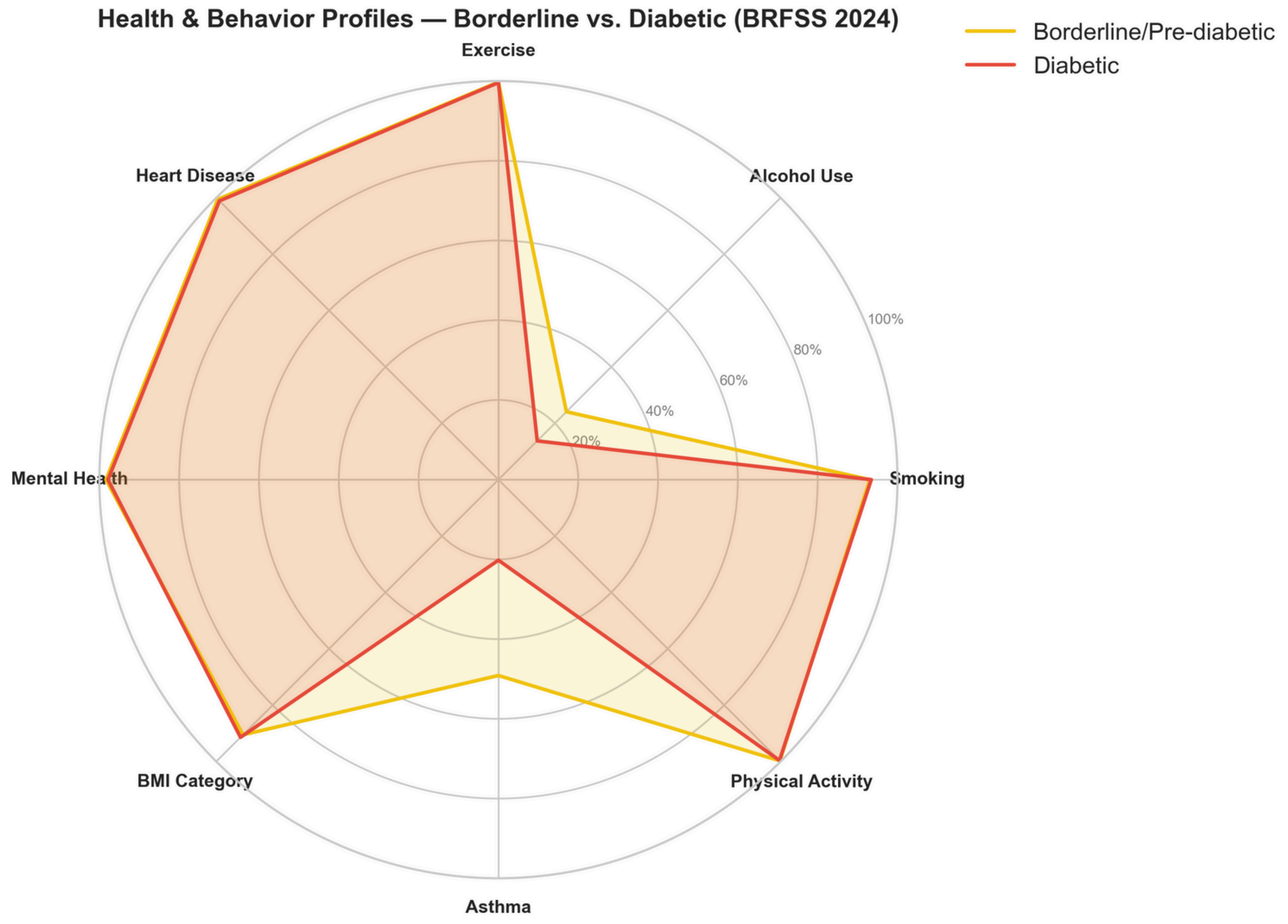
Each facet shows how diabetes prevalence differs across key health and behavior variables.
Variables: BMI Category, Smoking, Alcohol Use, Stroke, Asthma, Heart Disease, Exercise, Mental Health, and Overall Physical Activity.

Distribution of Demographics & Lifestyle Factors by Diabetes Status (BRFSS 2024)



Each facet shows how diabetes prevalence differs across key demographic and lifestyle variables.
Variables: Age, Race, Education, Income, Employment, and Healthcare Access.

DATA DISTRIBUTION



Radar plot comparing proportions of key health and lifestyle behaviors across Borderline/Pre-diabetic and Diabetic groups.
Variables: Smoking, Alcohol, Exercise, Heart Disease, Mental Health, BMI, Asthma, and Physical Activity.

- Compares key health & lifestyle profiles for borderline/pre-diabetic vs diabetic groups.
- Diabetic respondents show a **higher burden of comorbid conditions** (e.g., heart disease, mental health issues, asthma).
- Differences in smoking, BMI, and physical activity highlight behaviors linked to progression from borderline to diabetes.
- Helps identify targets for **intervention beyond blood sugar alone** (lifestyle + overall health).

MODELING

MODELING SETUP & EVALUATION

- 3-class diabetes prediction (No, Pre-, Yes)
- 80/20 train–test split (~360k / 90k)
- ADASYN oversampling + class-weighted loss
- Main loss: weighted cross-entropy (log-loss)
- Metrics: Accuracy, Precision, Recall, F1, ROC-AUC, Log-loss
- Primary metric: macro F1 (imbalanced data)

BASELINE MODELS

- All trained on ADASYN-balanced, class-weighted data
- Naïve Bayes: weakest, strong majority-class bias
- Decision Tree: higher accuracy, overfit, poor on minorities
- kNN: Manhattan > Euclidean, but many minority errors
- Logistic Regression: best macro F1 + log-loss among baselines, interpretable

Model	**Accuracy**	**Precision (Macro)**	**Recall (Macro)**	**F1 Score (Macro)**	**Log Loss**
Naïve Bayes	0.4594	0.4129	0.4754	0.342	8.1898
Decision Tree	0.7772	0.4087	0.4072	0.4057	-
kNN (Euclidean)	0.3921	0.4223	0.4706	0.3142	1.1728
kNN (Manhattan)	0.6171	0.427	0.4999	0.4103	0.8575
Logistic Regression	0.6043	0.4395	0.5303	0.416	0.9061

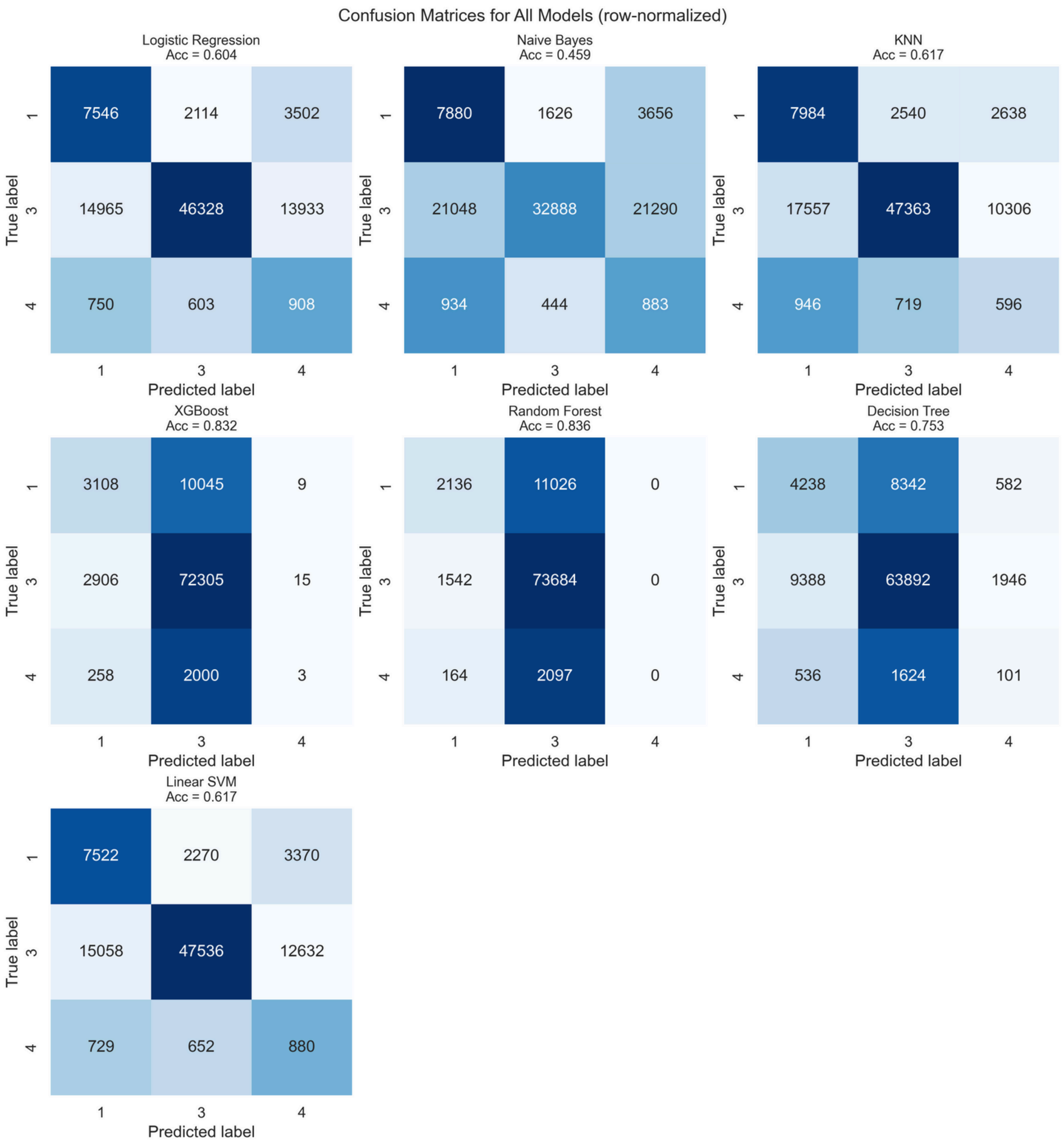
ADVANCED MODELS & TAKEAWAYS

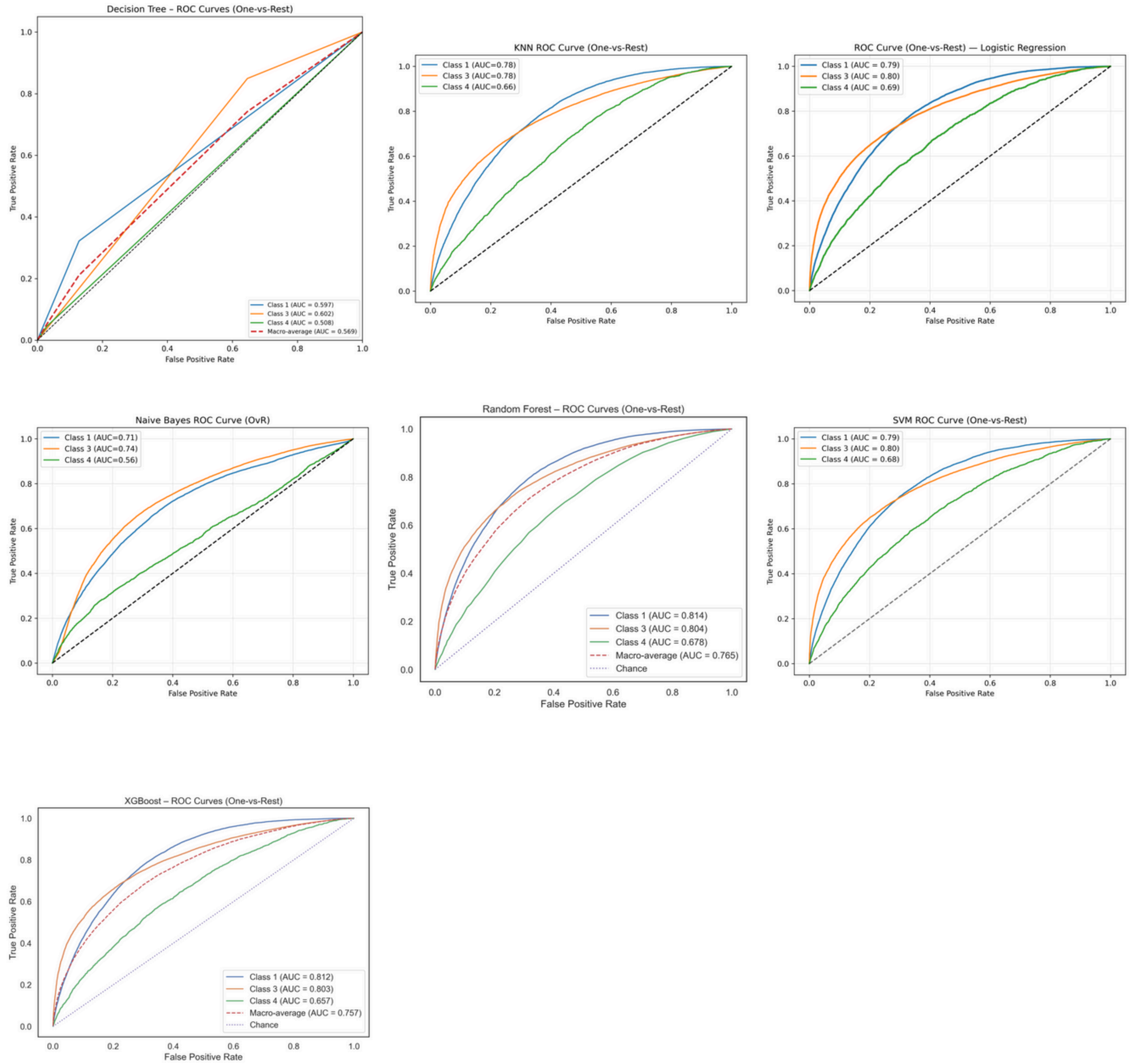
Model	Accuracy	Precision (Macro)	Recall (Macro)	F1 Score (Macro)	Log Loss
Random Forest	0.8364	0.4683	0.3806	0.3869	0.4508
Linear SVM	0.6171	0.439	0.5309	0.4203	-
XGBoost	0.832	0.4879	0.3995	0.4096	0.4458

- Only linear SVM (RBF too expensive on full dataset)
- RF & XGBoost: highest accuracy (~0.83), low log-loss
- XGBoost: best macro precision, strong overall balance
- Linear SVM: highest macro recall & F1; best at minority detection, lower accuracy
- Overall:
    - For minority sensitivity → LogReg / Linear SVM
    - For calibrated risk scores → XGBoost / Random Forest

POST-MODELING VISUALIZATIONS

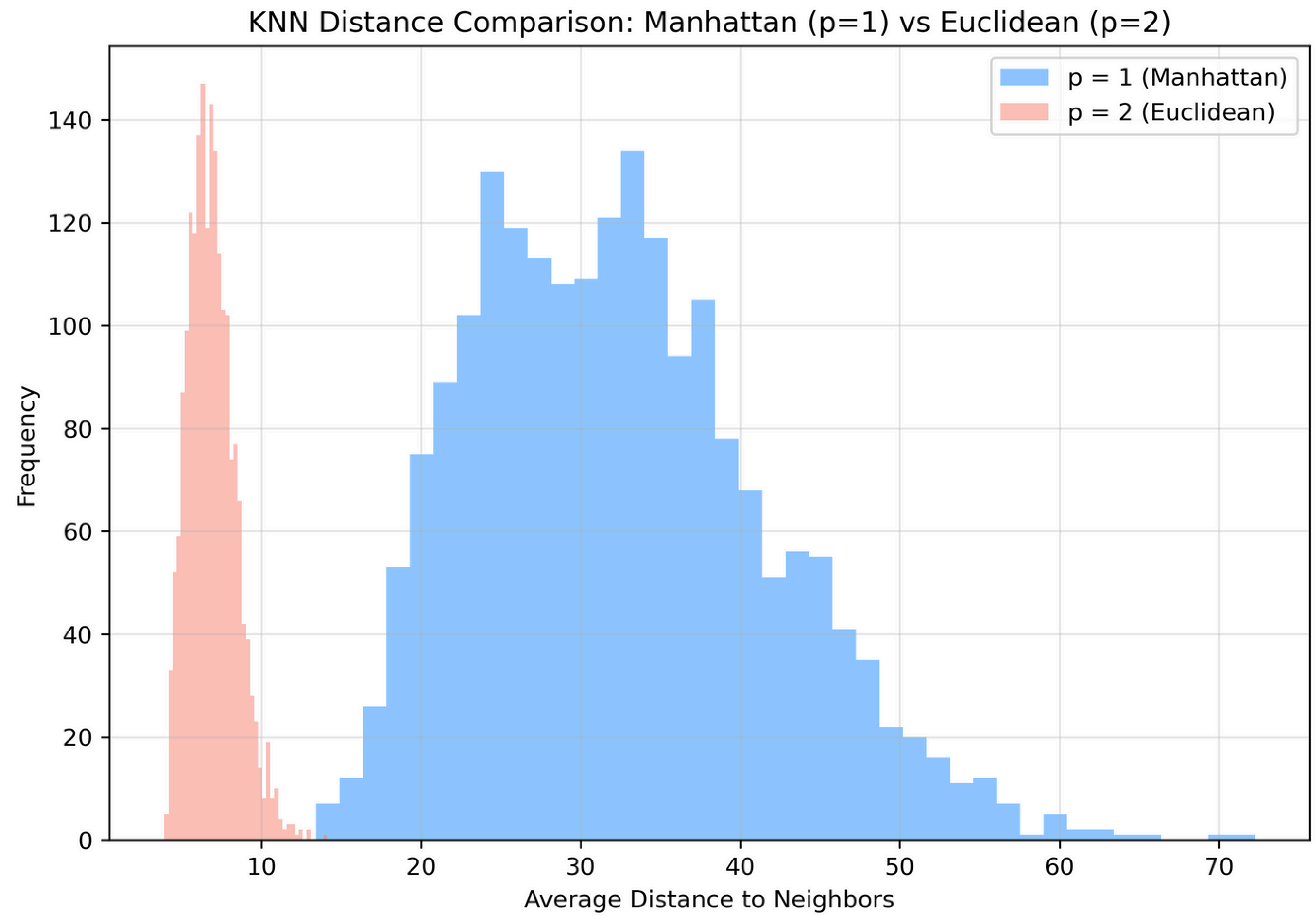
CONFUSION MATRIX & ROC CURVES

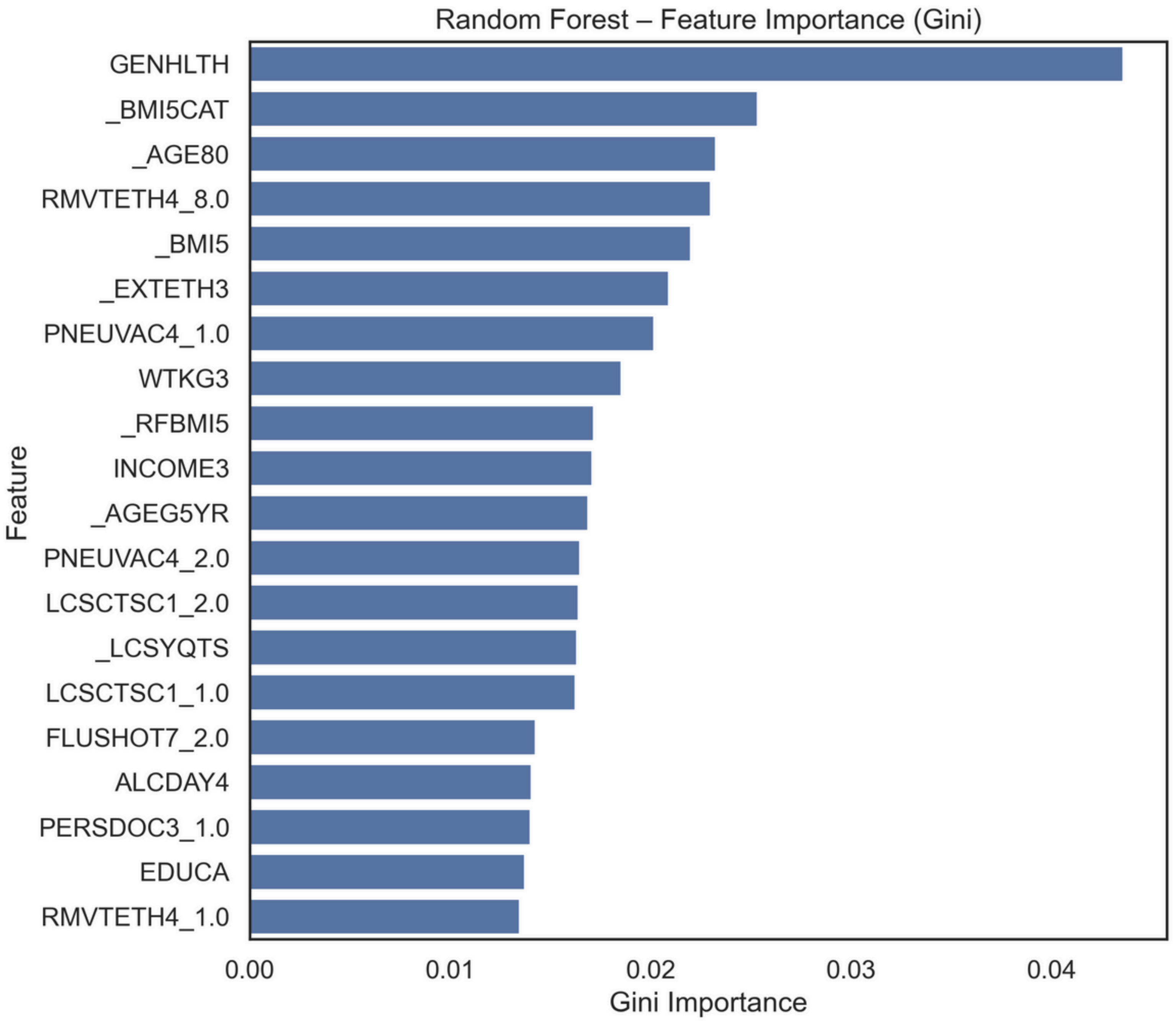




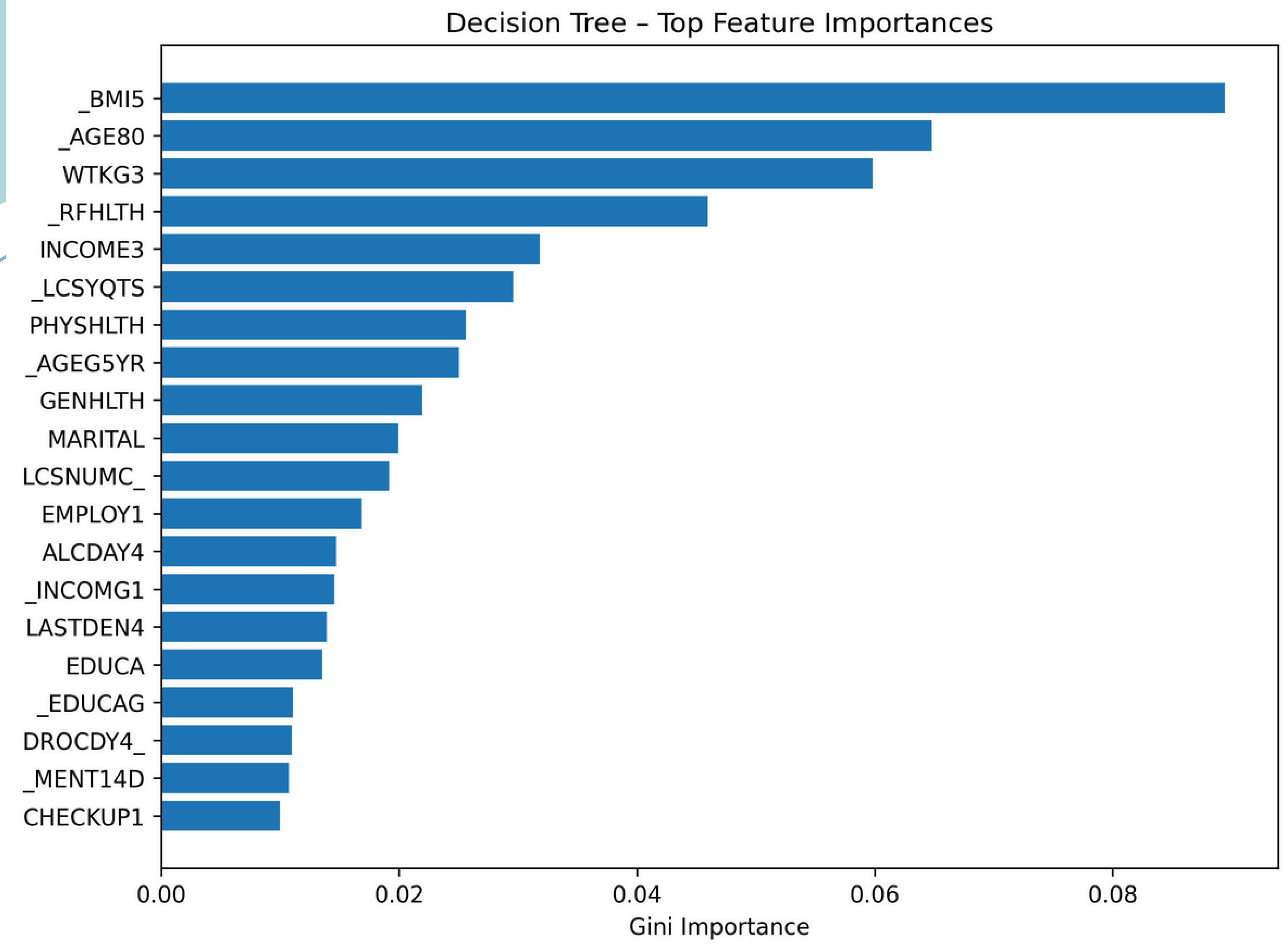
Interactive Confusion Matrix here

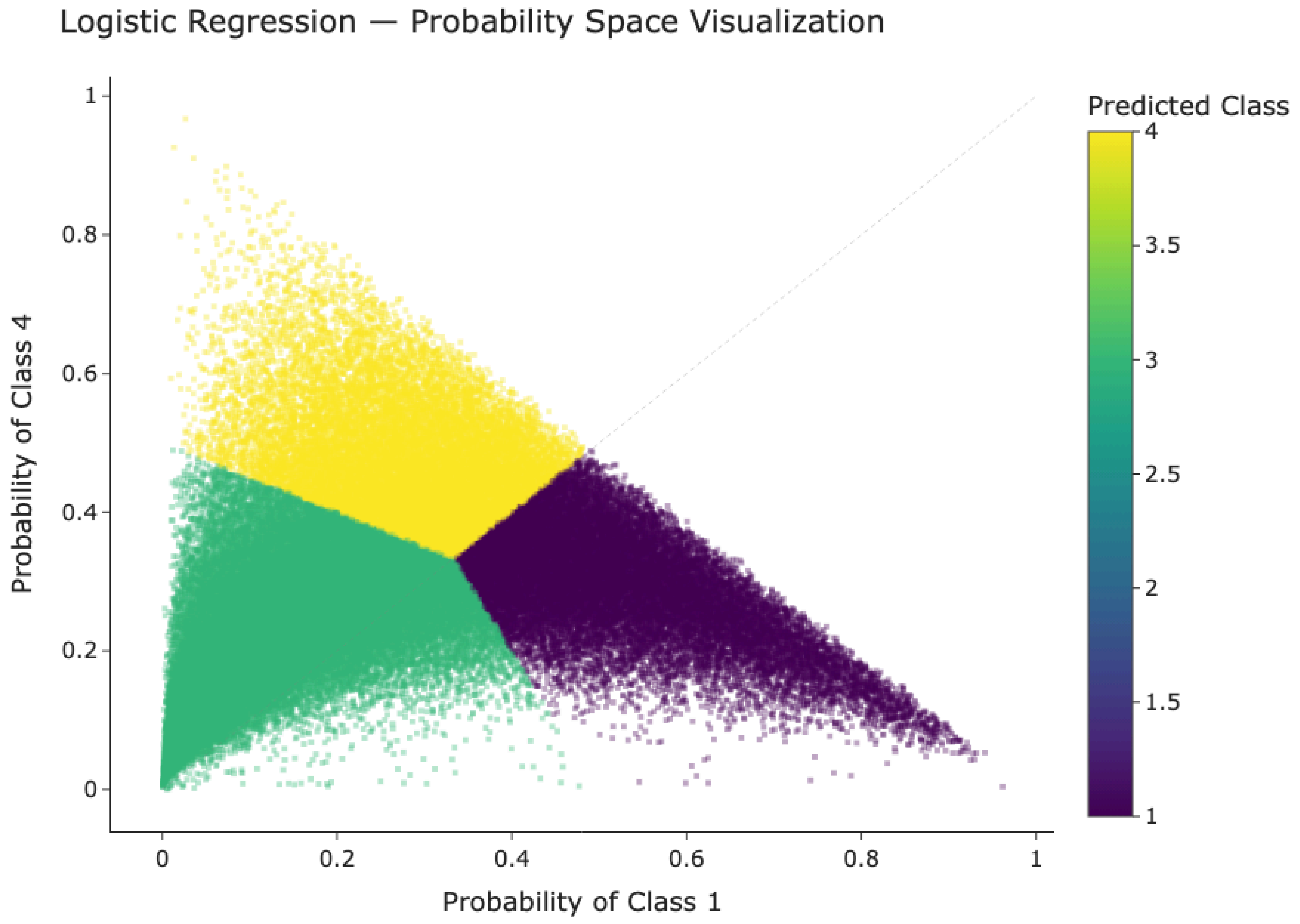
MODEL VISUALIZATIONS





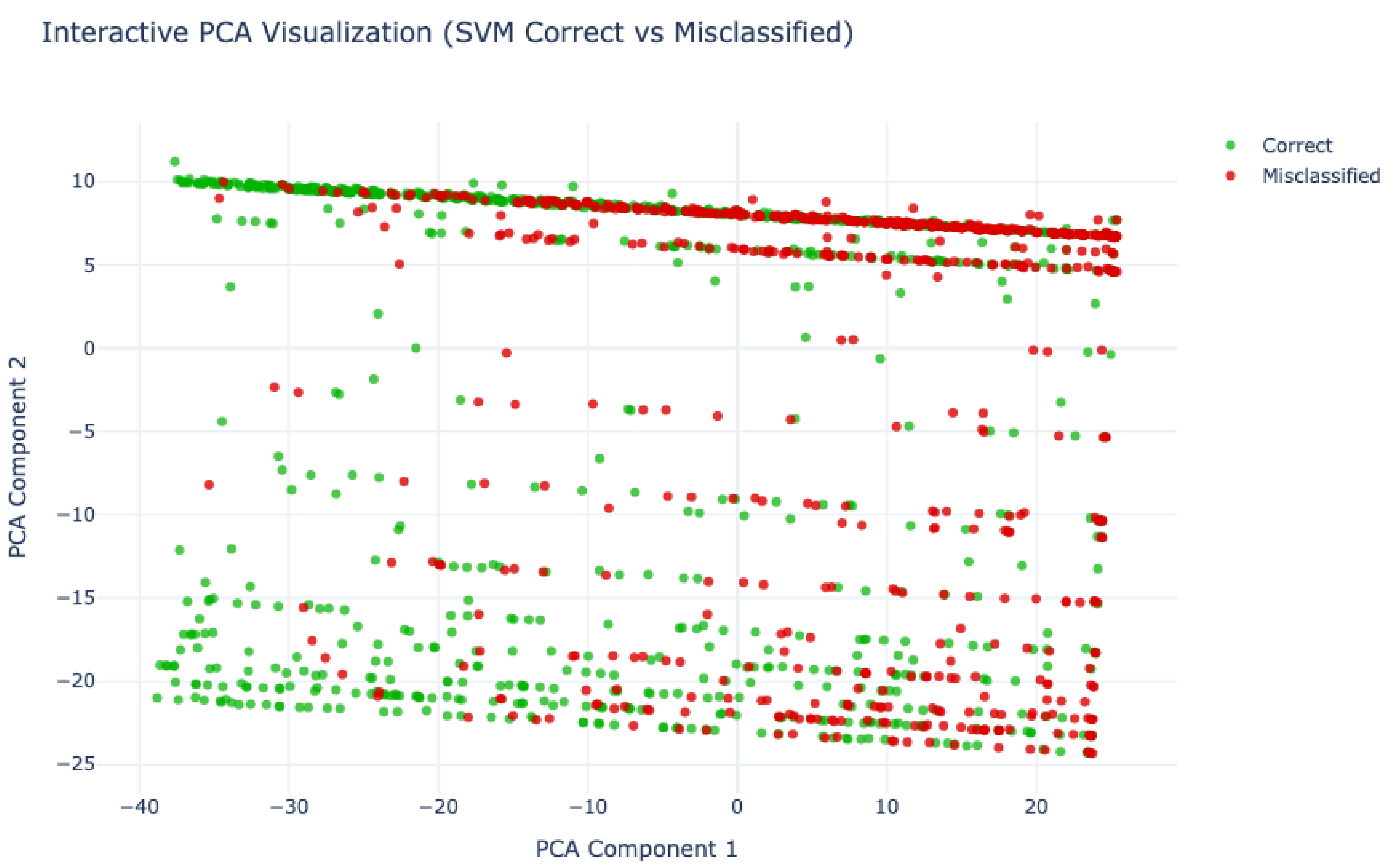
MODEL VISUALIZATIONS

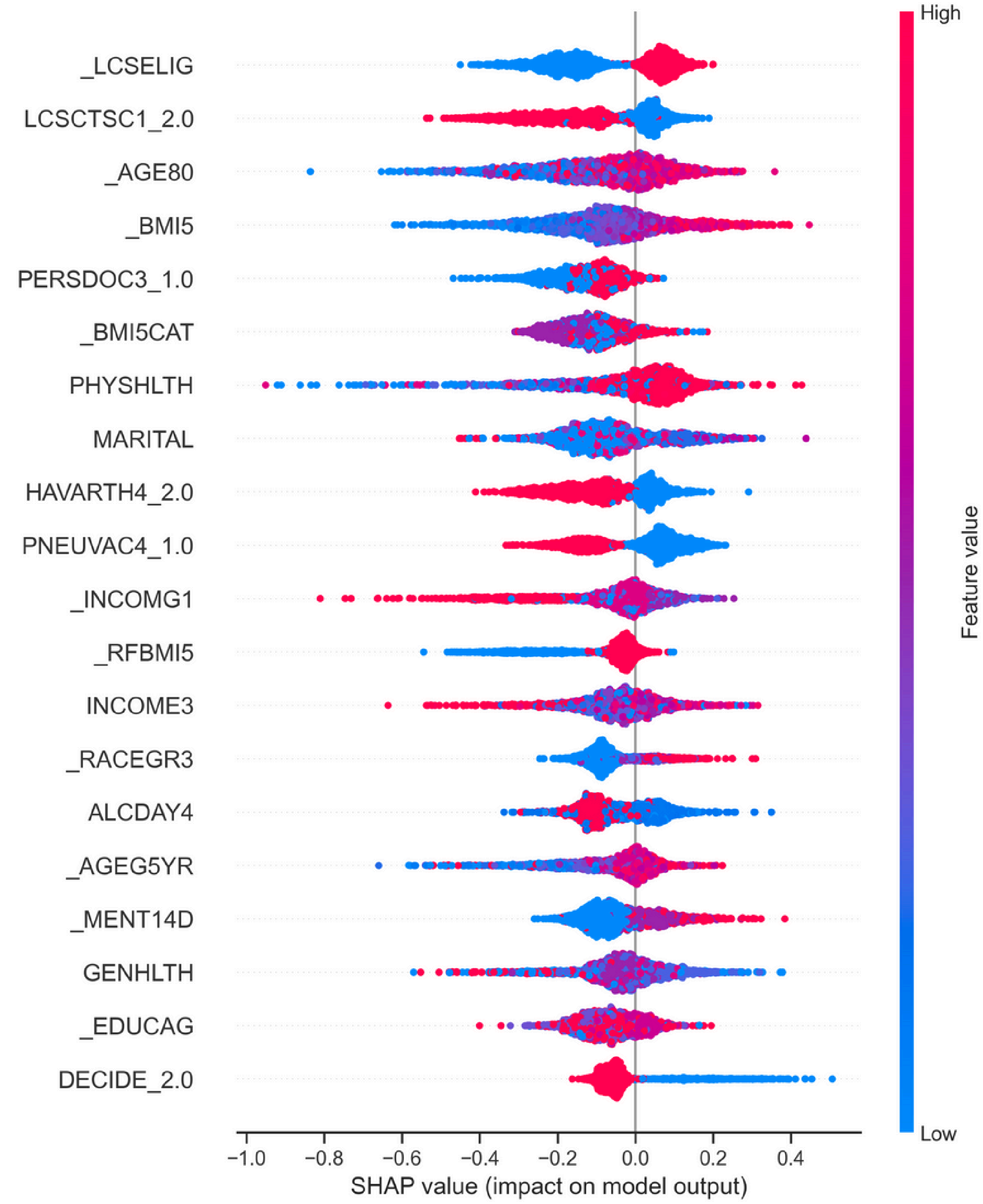




[Interactive LR Probability Visualization]

MODEL VISUALIZATIONS

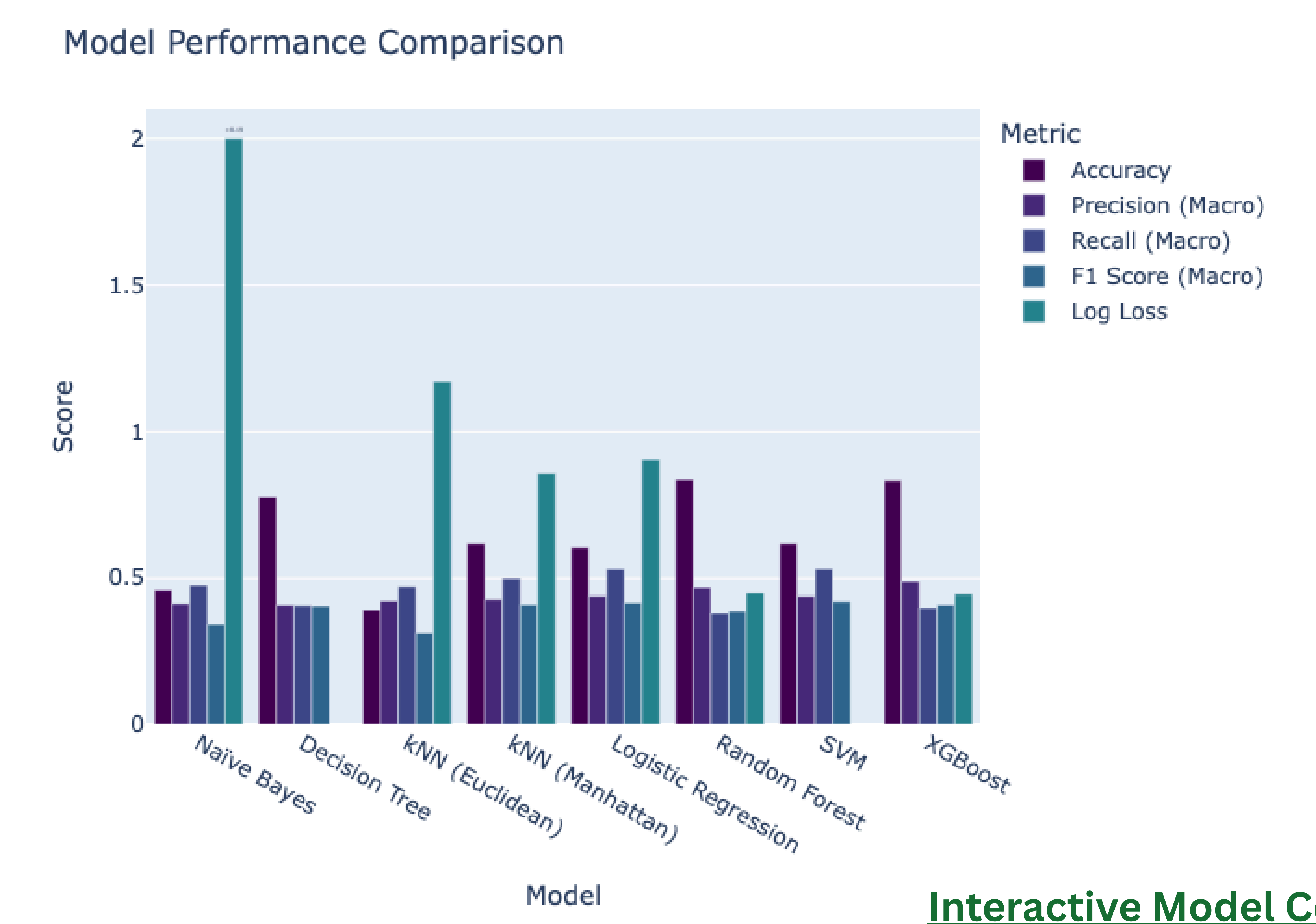




Interactive PCA

Interactive SHAP

MODEL PERFORMANCE COMPARISON



- Random Forest and XGBoost deliver the highest overall performance across most metrics.
- Logistic Regression and SVM achieve strong, balanced results, especially in precision and recall.
- Naïve Bayes performs weaker overall, with higher log loss due to modeling assumptions.
- KNN models show moderate performance, with Manhattan distance slightly outperforming Euclidean.
- Tree-based methods show better robustness on imbalanced classes, reflected in higher F1 and accuracy.

Interactive Model Comparison here

CONCLUSION

END-TO-END PIPELINE & KEY MODELING FINDINGS

- Built a full, reproducible pipeline from raw ASCII BRFSS data → cleaned, engineered dataset ready for large-scale modeling.
- Reduced 300+ raw variables to a focused set of meaningful predictors; raised valid response rate from ~48% → ~89%.
- Exploratory analysis showed diabetes clusters around: older age, higher BMI, lower income/education, limited physical activity, and multiple chronic conditions.
- Severe class imbalance (≈83% non-diabetic) required ADASYN oversampling and class weighting to train stable models.
- Logistic Regression, Manhattan kNN, and Linear SVM provided the strongest macro F1 scores (~0.41–0.42), balancing minority-class detection with interpretability.
- Random Forest and XGBoost delivered the highest accuracy (~0.83–0.84) and best probability calibration, especially for distinguishing diabetic vs. non-diabetic respondents.

PRACTICAL INSIGHTS & FUTURE IMPLICATIONS

- Reinforces public-health knowledge: diabetes risk is shaped by behavioral, metabolic, and socioeconomic factors together, not in isolation.
- Highlights the challenge of reliably identifying pre-diabetic individuals, even with oversampling—important for early intervention programs.
- Demonstrates which features consistently matter: general health, BMI, age, preventive-care indicators, and income/education.
- Shows that ensemble models excel in accuracy and risk ranking, while linear models excel in minority-class sensitivity.
- Provides a reusable ML pipeline for future BRFSS analyses, policy evaluation, and population-level risk modeling.
- Opens doors for next steps: cost-sensitive learning, fairness evaluations across demographic groups, and larger hyperparameter searches.

THANK YOU FOR YOUR ATTENTION