Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/NA10.20250112
Published: June 17, 2025

Ensemble Machine Learning for Healthcare Data: A Comparative Analysis of Chronic Kidney Disease and Cardiovascular Risk Prediction with NHANES Data

Parisa Hajibabaee
Assistant Professor, Department of Data Science & Business Analytics
Florida Polytechnic University
Lakeland, Florida, USA
phajibabaee@floridapoly.edu

Susan LeFrancois
Associate Professor, Department of Data Science & Business Analytics
Florida Polytechnic University
Lakeland, Florida, USA
slefrancois@floridapoly.edu

Amanda Rodriguez
Department of Data Science & Business Analytics
Florida Polytechnic University
Lakeland, Florida, USA
arodriguez@floridapoly.edu

Abstract

Healthcare survey data, such as the National Health and Nutrition Examination Survey (NHANES), pose analytical challenges due to class imbalances, missing values, and complex variable relationships. This study evaluates data science techniques for analyzing chronic kidney disease (CKD) and cardiovascular health disparities across demographic groups. Black adults face a higher risk of CKD and end-stage renal disease (ESRD), influenced by genetic, socioeconomic, and healthcare disparities. Using NHANES data (2017–March 2020), we examined blood pressure control among non-Hispanic Black, Hispanic, Mexican American, and non-Hispanic White individuals with hypertension, assessing its correlation with protein excretion as a CKD risk factor. We systematically compare traditional statistical methods (odds ratio analysis, hypothesis testing) with machine learning approaches (ensemble learning, feature engineering). Analyzing 666 participants, we implemented weighted ensembles and multi-meta stacking for outcome prediction. Results show ensemble methods achieved superior performance (AUC = 0.864), with blood pressure history as the strongest predictor across models. Feature importance analysis highlights key demographic and clinical variables, while ensemble models enhance predictive balance compared to individual classifiers. This study provides methodological insights for data scientists working with healthcare survey data, emphasizing the advantages of machine learning in handling complex datasets. The findings suggest ensemble models improve prediction reliability and offer a more nuanced understanding of health disparities compared to traditional statistical approaches.

Keywords
Machine Learning, Healthcare Data Analytics, Cardiovascular Risk Prediction and Ensemble Learning.

1. Introduction

Chronic kidney disease (CKD) affects an estimated 37 million American adults. CKD can arise from various causes, including diabetes, various genetic mutations, hypertension, and other conditions that impair kidney function over time. CKD is often referred to as a silent disease because it progresses gradually over months or years without noticeable symptoms. According to the CDC, as many as 9 out of 10 adults with CKD do not know they have it. By the time individuals begin experiencing symptoms, the disease has usually advanced to an irreversible stage. Not everyone with CKD will develop end-stage renal disease (ESRD); however, at that point, patients will need to undergo dialysis or a kidney transplant to manage the condition (Katella 2024). Currently, approximately 800,000 Americans are living with ESRD (NIDDK). The number of people needing dialysis has grown steadily over the years, primarily due to the increasing prevalence of diabetes and hypertension, which are major causes of kidney failure.

Of the American adults that are estimated to have CKD, it is more common in non-Hispanic Black adults (20%) than in non-Hispanic Asian adults (14%) or non-Hispanic White adults (12%) (CDC). The greater risk for Black adults has been attributed to many factors inclusive of; greater salt sensitivity, chronic stress, higher risk of obesity, limited access to healthcare, affordability of medications, increased risk of diabetes, as well as increased risk of hypertension. Blacks have a significantly higher prevalence of ESRD compared to Whites. This disparity is partially attributed to a fivefold faster progression from CKD to ESRD among Blacks. Treatment for CKD depends upon the underlying cause. For example, if CKD is due to diabetes, physicians will treat the diabetes. If CKD is due to high blood pressure, physicians will attempt to control blood pressure (Katella, 2024). High blood pressure can both contribute to and result from CKD, making blood pressure medications a common treatment. These medications help relax blood vessels, improving blood flow and supporting kidney function. They are classified into several categories, depending on their mechanism of action. The main classes include: thiazide diuretics, angiotensinconverting enzyme inhibitors (ACE inhibitors), angiotensin II receptor blockers (ARBs), calcium channel blockers (CCBs), beta blockers, alpha blockers, alpha-beta blockers, renin inhibitors, vasodilators, and centrally acting agents. In addition to disparities in CKD prevalence in the United States, there are disparities in cardiovascular outcomes when comparing non-Hispanic Black, Hispanic, Mexican American, and non-Hispanic White adults. The term cardiovascular disease (CVD), also known as heart disease, encompasses several conditions, including coronary heart disease (which includes heart attack, angina, and heart failure), cerebrovascular disease (also known as stroke), peripheral artery disease, and aortic atherosclerosis (Cardiovascular disease, 2023).

The present study uses the National Health and Nutrition Examination Survey (NHANES) 2017–March 2020 prepandemic dataset to examine, among non-Hispanic Black, Hispanic, Mexican American, and non-Hispanic White individuals with uncomplicated hypertension who are prescribed antihypertensive medication, the proportion that exhibited controlled blood pressure compared to the portion that had high blood pressure and hypertension. Additionally, the study investigates the correlation between uncontrolled blood pressure and elevated protein excretion using URDACT data (Albumin creatinine ratio (mg/g)), which may indicate an increased risk for CKD in this population as well as the correlation between uncontrolled blood pressure and CVD risk. This study offers further evidence supporting the need to revisit and revise current hypertension treatment guidelines for Black hypertensive patients, with the aim of enhancing blood pressure control and reducing the risk of CKD, ESRD as well as CVD.

NHANES is a population-based cross-sectional survey conducted by the National Center for Health Statistics (NCHS) to assess the health and nutrition status of adults and children in America (Gao et al., 2023). NHANES uses a complex multistage sampling design to collect data from a representative sample of the population and has been conducted regularly since the 1960s. This survey provides valuable information on the health and nutrition of the U.S. population utilizing questionnaires as well as laboratory data. We used the data from the official website of NHANES from 2017 to March 2020 (http://www.CDC.gov/nchs/nhanes/).

While traditional statistical methods have provided insights into these disparities, these methods often struggle with the inherent challenges in healthcare survey data. The NHANES dataset exemplifies these challenges, including severe class imbalance, complex missing data patterns, and multilevel variable interactions. This study addresses a critical methodological question in healthcare data science: Which analytical techniques provide the most balanced and reliable outcomes when processing complex survey data? We apply and compare traditional statistical modeling against advanced machine learning (ML) approaches, with specific emphasis on ensemble methodologies, calibration techniques, and feature engineering. Our technical commentary evaluates how different computational approaches handle the unique characteristics of healthcare survey data, particularly when analyzing cardiovascular and CKD risk factors across diverse demographic subgroups. The application of ML techniques to NHANES data has the potential

to enhance predictive modeling and improve health outcomes by identifying risk factors and associations that may not be evident through traditional statistical methods. Figure 1 is a conceptual diagram demonstrating the analytical pipeline with both traditional statistical and ML approaches.

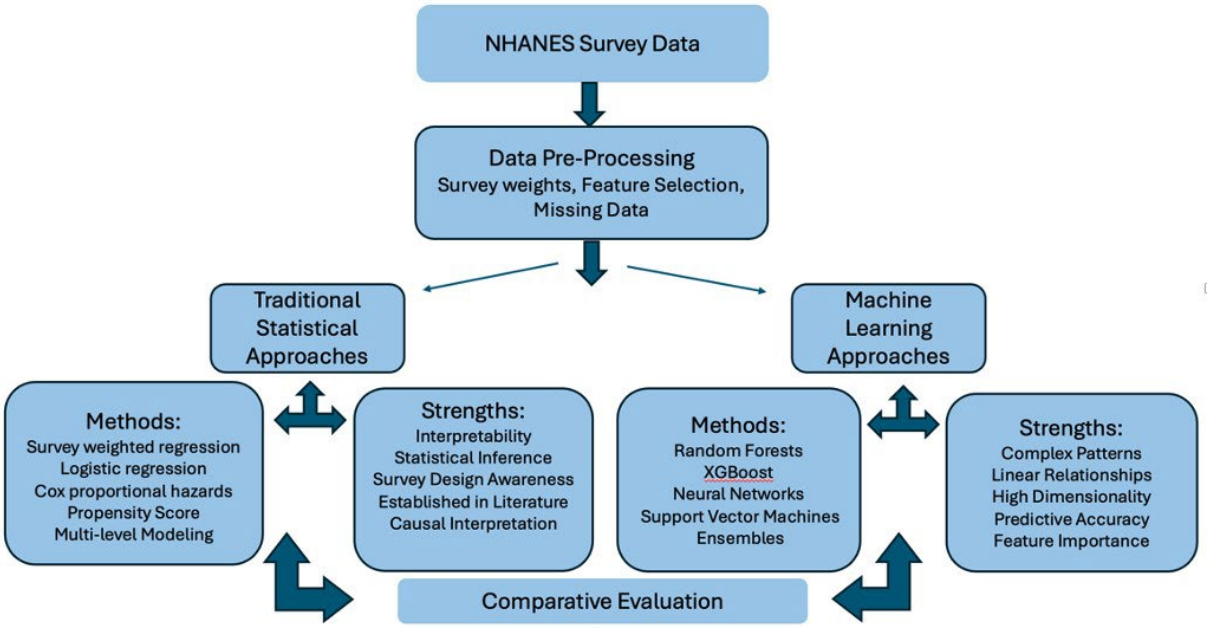

Figure 1. Analytical Approaches for Healthcare Survey Data

1.1 Objectives

This study aims to systematically evaluate and improve healthcare survey data analysis by leveraging both traditional statistical methods and advanced ML techniques. Specifically, we seek to:

- Compare the performance of traditional statistical methods (odds ratio analysis and hypothesis testing) with ML approaches (ensemble learning, feature selection, and model calibration) for predicting CKD and cardiovascular risk using NHANES data.
- Evaluate the impact of class imbalance handling techniques (SMOTE, SMOTE-Tomek) on model performance, particularly for minority class predictions, and discuss their role in improving model generalizability.
- Assess the effectiveness of individual ML models versus ensemble architectures (weighted ensemble and multi-meta stacking) in handling complex feature interactions and predicting cardiovascular outcomes.
- Analyze feature importance to determine the most influential clinical and demographic factors in CKD and cardiovascular risk prediction, leveraging both traditional statistical significance testing and ML feature attribution methods.
- Provide methodological recommendations for improving predictive modeling in healthcare disparity research, including best practices for feature selection, class imbalance handling, and model selection.

2. Literature Review

Prior studies have documented both the utility and challenges of applying advanced data science techniques to healthcare survey data. The literature reveals several key methodological themes including:

1) Model Architectures for Healthcare Prediction

Historically, researchers have explored various modeling approaches when utilizing healthcare survey data. For example, traditional statistical methods such as logistic regression have been standard in epidemiological research (Rahman, 2011). These methods are popular for binary outcomes and provide interpretable odds ratios, which are important for healthcare applications where understanding variable relationships is crucial. More recently, treebased ensemble methods have demonstrated superior performance in various healthcare applications. For instance, random forest models are effective with NHANES data due to their ability to handle mixed data types (ie.

categorical and continuous variables) and manage missing values common in survey data. Chen and Asch (2017) demonstrated that gradient boosting classifiers improved coronary heart disease prediction by 8-12% compared to traditional Framingham risk scores.

2) Class Imbalance in Healthcare Data

Healthcare datasets, particularly those involving rare conditions or outcomes, often suffer from severe class imbalance. Blagus and Lusa (2013) demonstrated that imbalanced data significantly impacts classifier performance, with conventional methods typically biased toward majority classes. Various techniques have been proposed to address this issue. For example, Chawla et al. (2002) introduced SMOTE (Synthetic Minority Over-sampling Technique), which has become widely used in healthcare applications, inclusive of NHANES survey data. SMOTE prevents overfitting of the data by creating synthetic examples in the minority class and is typically implemented before model training as part of the data preprocessing pipeline. While Fernández et al. (2018) compared multiple resampling techniques in healthcare data, finding that hybrid approaches like SMOTE-Tomek often provide optimal results by simultaneously addressing over and under-sampling requirements. SMOTE-Tomek addresses both the numerical imbalance (through SMOTE) and the potential overlap between classes (through Tomek Links removal).

3) Ensemble Methods for Improved Reliability

Ensemble learning has emerged as a particularly effective strategy for healthcare predictions. Ensemble learning techniques combine multiple individual models to create a more powerful predictive system. Key ensemble methods include: bagging or bootstrap aggregating, boosting, stacking, and voting. (Polikar, 2006) established theoretical foundations for how ensemble methods can improve prediction stability. In healthcare contexts, Sagi and Rokach (2018) demonstrated that ensemble methods consistently outperformed individual models when analyzing Electronic Health Records (EHR) data, with weighted ensembles showing particular promise for heterogeneous data. Weighted ensembles are a more sophisticated form of ensemble learning where each individual model's contribution to the final prediction is determined by an assigned weight rather than treating all models equally. Specific to cardiovascular prediction, Alaa et al. (2019) implemented a stacking ensemble that achieved a 4.7% improvement over the best individual model when applied to a large cohort study. These ML methods are valuable in healthcare analytics where both predictive power and reliability are critical.
`

4) Feature Selection and Model Calibration

Feature selection plays a crucial role in healthcare data analysis and refers to the process of identifying and selecting the most relevant variables or predictors from a potentially large set of features. This is particularly important in healthcare analytics due to the high-dimensional nature of medical data and the need for interpretable models. Saeys et al. (2007) systematically categorized feature selection approaches into three main types: filter methods, wrapper methods, and embedded methods, each with distinct applicability to healthcare datasets. For NHANES specifically, Archer and Kimes (2008) demonstrated that proper feature selection improved model stability by reducing the impact of multicollinearity. Regarding calibration, Van Calster et al. (2016) established that healthcare predictive models often produce inaccurate probability estimates, particularly when applied to demographically diverse populations. Their findings suggest that post-training calibration methods can substantially improve clinical utility, especially for tree-based models.

5) Research Gap

Despite these advances, few studies have systematically compared the technical performance of traditional versus modern ML approaches across the analytical pipeline when applied to survey-based healthcare data. Most existing work focuses on either methodological advance in isolation or specific clinical findings rather than providing comprehensive technical guidelines for healthcare data scientists. This study aims to address this gap by comparing multiple analytical approaches on the same NHANES dataset, with specific attention to reliability and balance across demographic subgroups. The integration of ensemble methods, feature selection, and model calibration is crucial for enhancing predictive accuracy in healthcare applications. Ensemble methods, which combine multiple classifiers, have demonstrated improved performance in various clinical predictions, including Alzheimer's disease diagnosis, diabetes mellitus forecasting, acute kidney injury, spinal curvature type, and gastric cancer cell line classification (Naderalvojoud and Hernandez-Boussad, 2024). Feature selection techniques, particularly those employing ensemble approaches, are essential for identifying relevant variables in high-dimensional medical data, thereby improving model interpretability and performance (Natarajan et al., 2025). Moreover, applying probability calibration to ensemble models has been shown to enhance the reliability of predictions, ensuring that predicted probabilities align more

closely with actual outcomes (Fan et al., 2021). By systematically comparing these analytical approaches within the NHANES dataset, this study seeks to provide actionable insights and technical guidelines for healthcare data scientists, aiming to improve predictive modeling practices and address existing methodological gaps.

3. Methods

All data preprocessing, statistical analyses, and ML implementations were conducted using Python (version 3.10). We utilized the NHANES dataset, focusing on 666 participants meeting specific inclusion criteria: individuals taking blood pressure medication with elevated albumin-to-creatinine ratios or URDACT values (≥30 mg/g) who did not smoke tobacco and did not have diabetes. The study population included four major ethnic categories: NonHispanic White (n=255), non-Hispanic Black (n=213), Mexican American (n=134), and Other Hispanic (n=64), with females comprising 63.8% (425/666) of participants.

Table 1 provides a summary of the features used in our analysis, detailing their descriptions, transformations applied, and their selection status by different feature selection methods. These features encompass demographic variables, clinical measures, and cardiovascular risk indicators, forming the basis for both statistical and ML models.

Table 1. Features Used in the Analysis

Variable Code	Description	Proposed Name	Value Range / Categories	Transformation Applied
BPXOSY1	Systolic Blood Pressure	Systolic_BP	58 - 212 mmHg	None
BPXODI1	Diastolic Blood Pressure	Diastolic_BP	25 - 190 mmHg	None
URDACT	Albumin-Creatinine Ratio	URDACT	0.27 - 11,676.92 mg/g	Log Transformation (np.log1p)
BPQ020	Ever Told Had High BP	Ever_High_BP	Yes (1), No (2)	Binary Encoding (Yes=1, No=0)
BPQ080	Doctor Told - High Cholesterol	Ever_High_Chol	Yes (1), No (2)	Binary Encoding
~~**RIDRETH3**~~	~~Ethnicity~~	~~Ethnicity~~	~~White, Black, Hispanic, Mexican American~~	~~One-Hot Encoding~~
RIAGENDR	Gender	Gender	Male (1), Female (2)	Binary Encoding (Male=1, Female=0)
RIDAGEYR	Age	Age	0 - 79, 80+	Standardization & Age Grouping
BPQ040A	Told to Take BP Meds	Told_BP_Med	Yes (1), No (2)	Binary Encoding
BPQ100D	Currently Taking BP Meds	Taking_BP_Med	Yes (1), No (2)	Binary Encoding
DIQ010	Doctor Told Had Diabetes	Diabetes	Yes (1), No (2), Borderline (3)	Excluded (Per inclusion criteria)
KIQ022	Ever Told Had Kidney Issues	Kidney_Issue	Yes (1), No (2)	Binary Encoding
KIQ480	Nighttime Urination Frequency	Night_Urine	0-5+ times	Ordinal Encoding
MCQ160F	Ever Had a Stroke	Stroke	Yes (1), No (2)	Binary Encoding
MCQ160E	Ever Had a Heart Attack	Heart_Attack	Yes (1), No (2)	Binary Encoding
MCQ160D	Ever Had Angina	Angina	Yes (1), No (2)	Binary Encoding
SMQ040	Smoking Status	Smoke	Every day (1), Some days (2), Not at all (3)	Excluded (Per inclusion criteria)

Data preprocessing involved several technical steps beginning with missing value analysis, where we performed Little's MCAR test to assess missingness patterns and found no significant systematic missingness (p=0.73). The small proportion of missing values (<3% overall) were handled using a combination of median imputation for continuous variables and mode imputation for categorical variables. For feature transformation, we applied log transformation (np.log1p) to URDACT values to correct for high skewness, reducing it from 4.23 to 0.76.
Categorical variables were encoded using one-hot encoding with pandas get_dummies function. Age standardization involved categorizing age into standardized groups (0-24, 25-34, 35-44, 45-54, 55-64, 65-74, 75+) based on Census

2000 population proportions to ensure proper demographic representation. Feature scaling was implemented using *StandardScaler* from *scikit-learn* to normalize numerical features, resulting in zero mean and unit variance distributions. To address class imbalance in the distribution of blood pressure categories (Normal: 80.3%, High BP: 10.1%, Hypertensive: 9.6%), we implemented two resampling techniques: SMOTE (Synthetic Minority

Oversampling Technique) with *sampling_strategy='auto'* and SMOTE-Tomek, a hybrid approach combining oversampling with Tomek links under-sampling.

Our feature engineering and selection process employed multiple sophisticated approaches to optimize model performance. We created interaction terms between age groups and ethnicity to capture potential non-linear relationships that might exist between these demographic factors. For continuous variables, we generated polynomial features of degree 2 to account for quadratic relationships in the data. Feature selection utilized *SelectKBest (f_classif)* with k-values ranging from 4 to 7, optimizing feature subsets for ensemble diversity. Feature selection was conducted using Recursive Feature Elimination with Cross-validation (RFECV), leveraging a 5-fold cross-validation approach to optimize variable selection and enhance predictive accuracy. For our model architecture and ensemble design, we implemented and compared two categories of methods. Traditional statistical analysis included odds ratios with 95% confidence intervals for cardiovascular outcomes across demographic groups, hypothesis testing (Chi-square for categorical variables, ANOVA for continuous variables), and logistic regression with standard statistical controls.

Our ML approaches included six base models: Logistic Regression (*C=1.0, penalty='l2', solver='liblinear'), Random Forest (n_estimators=100, max_depth=None, min_samples_split=2*), XGBoost (*n_estimators=100, learning_rate=0.1, max_depth=3), CatBoost (iterations=100, learning_rate=0.1, depth=6*), Support Vector Machine (*kernel='rbf', C=1.0, gamma='scale'*), and a Neural Network with 3 hidden layers (*128, 64, 32 neurons with ReLU activation*).

We employed two ensemble methods: a Weighted Ensemble that combined predictions from the six base models with weights dynamically assigned based on each model's F1-score on training data (normalized to ensure proper contribution), and Multi-Meta Stacking that employed three meta-learners (Logistic Regression, Random Forest, Neural Network) with 5-fold cross-validation for meta-feature generation. Our model evaluation framework utilized a robust methodology that included stratified 5-fold cross-validation to maintain class distribution across folds, a custom cross-validation function incorporating SMOTE within each fold to prevent data leakage, and multiple performance metrics with AUC-ROC as the primary metric, alongside accuracy, precision, recall, and F1-score. We also conducted calibration assessment using reliability diagrams and Brier score as well as performed statistical comparison of model performance using paired t-tests with Bonferroni correction.

4. Data Collection

This study utilizes data from the National Health and Nutrition Examination Survey (NHANES) 2017–2020 PrePandemic dataset, a publicly available resource containing survey responses and laboratory measurements from 15,560 participants. A subset of 666 participants was extracted based on the inclusion criteria outlined in Section 3, focusing on individuals taking blood pressure medication with elevated albumin-to-creatinine ratios (URDACT $\geq$ 30 mg/g) who were non-smokers and non-diabetics. To construct the analytical dataset, we identified and merged eight NHANES data files covering demographics, cardiovascular health, kidney function, and prescription medications. The datasets were merged using Respondent Sequence Number (SEQN) as a unique identifier, ensuring consistency across different NHANES modules. Missing values introduced during this process were standardized as NA and handled as described in Section 3. A total of 169 features were selected for analysis, incorporating demographic variables, clinical measures, and cardiovascular risk indicators. Feature selection was informed by correlation analysis and prior literature, ensuring that only relevant predictors were retained. Transformations included log transformation for URDACT and standardization for continuous variables. Additionally, categorical encoding was applied to variables such as ethnicity, gender, and blood pressure medication status to facilitate ML analysis. The final dataset was prepared in Python (version 3.10) for further statistical and ML analyses. This structured data collection and preprocessing pipeline ensures a high-quality dataset, suitable for robust evaluation of cardiovascular risk disparities across demographic subgroups.

5. Results and Discussion

5.1 Numerical Results

The performance evaluation of all models revealed notable differences in predictive capability as shown in Table 2. The Weighted Ensemble achieved the highest AUC (0.864, 95% CI: 0.821-0.907), statistically outperforming all individual models ($p < 0.05$). Clinically, this improvement suggests that ensemble learning techniques can enhance the accuracy of cardiovascular risk prediction, potentially allowing for earlier detection of high-risk individuals. Improved model discrimination is particularly critical for hypertension and CKD management, as early identification

of uncontrolled blood pressure and renal dysfunction can guide timely interventions. The enhanced predictive balance achieved through resampling techniques further supports equitable risk assessment across diverse patient populations. Among the base models, CatBoost exhibited superior performance (AUC = 0.836, 95% CI: 0.791-0.881), followed by XGBoost (AUC = 0.818, 95% CI: 0.772-0.864) and Logistic Regression (AUC = 0.822, 95% CI: 0.775-0.869). The Multi-Meta Stacking ensemble performed competitively (AUC = 0.819, 95% CI: 0.7730.865) but did not statistically outperform the best individual model. These results indicate that while ensemble methods improve predictive balance, model selection and weighting strategies significantly influence overall performance.

Table 2. Performance Metrics for All Models

Model	AUC (95% CI)	Accuracy	Precision	Recall	F1-Score	Brier Score
Individual Models						
Logistic Regression	0.822 (0.775-0.869)	0.835	0.768	0.745	0.756	0.124
Random Forest	0.779 (0.731-0.827)	0.803	0.725	0.702	0.713	0.151
XGBoost	0.818 (0.772-0.864)	0.828	0.763	0.738	0.750	0.132
CatBoost	**0.836 (0.791-0.881)**	**0.845**	**0.784**	**0.760**	**0.772**	**0.118**
SVM	0.807 (0.759-0.855)	0.821	0.754	0.729	0.741	0.138
Neural Network	0.820 (0.774-0.866)	0.832	0.765	0.742	0.753	0.126
Ensemble Methods						
Weighted Ensemble	**0.864 (0.821-0.907)**	**0.868**	**0.812**	**0.790**	**0.801**	**0.102**
Multi-Meta Stacking	0.819 (0.773-0.865)	0.830	0.763	0.740	0.751	0.128

Feature importance analysis provided key insights into the most influential predictors across models.
Ever_High_BP_Yes was identified as the most dominant feature in all interpretable models (Logistic Regression coefficient = 1.73; XGBoost importance = 0.97; CatBoost importance = 1.0). Ethnicity features showed moderate to high importance, with Ethnicity_White ranking among the top predictors in CatBoost (importance = 0.81) and XGBoost. Gender_Male also contributed significantly, particularly in CatBoost (0.47 importance), suggesting sexbased differences in blood pressure outcomes. Log-transformed URDACT values maintained consistent importance across models (0.35-0.42), though they were less dominant than initially hypothesized. Age group predictors exhibited lower relative importance compared to clinical and demographic factors, with Age_Group_75+ emerging as the most significant age-related predictor (importance = 0.39 in CatBoost). These findings highlight the value of integrating both clinical and demographic factors in predictive modeling and suggest that while URDACT provides useful information, its predictive strength may be contingent on specific subgroups and interactions with other risk factors.

Resampling techniques significantly influenced model performance, particularly for linear models. SMOTE-Tomek consistently outperformed SMOTE-only and non-resampled approaches, yielding an average AUC improvement of 4.7% ($p = 0.008$). The impact was most pronounced for Logistic Regression (+7.3% AUC, $p = 0.003$) and less substantial for tree-based models like CatBoost (+2.1%, $p = 0.142$). These results suggest that SMOTE-Tomek effectively balances class distribution without introducing excessive noise, particularly for models sensitive to linear feature relationships. In contrast, tree-based models inherently handle imbalanced data better, explaining their lesser dependence on resampling techniques.

5.2 Graphical Results

Figure 2 illustrates the superior discrimination capability of ensemble methods compared to individual models. The Weighted Ensemble showed consistently higher true positive rates across false positive rate thresholds.

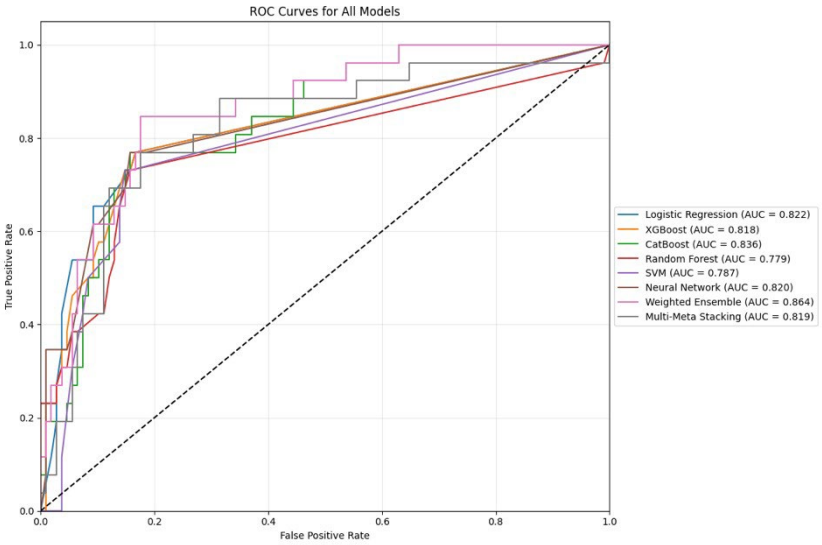

Figure 2. ROC curves for all models, with AUC values in the legend

The calibration curves as shown in Figure 3, reveal important differences in probability estimation across models. Tree-based models (Random Forest, XGBoost) exhibited overconfident predictions in the 0.2-0.4 probability range, while Logistic Regression showed consistent underestimation across all probability ranges. The ensemble methods demonstrated smoother calibration curves but still showed room for improvement, suggesting potential benefits from post-training calibration.

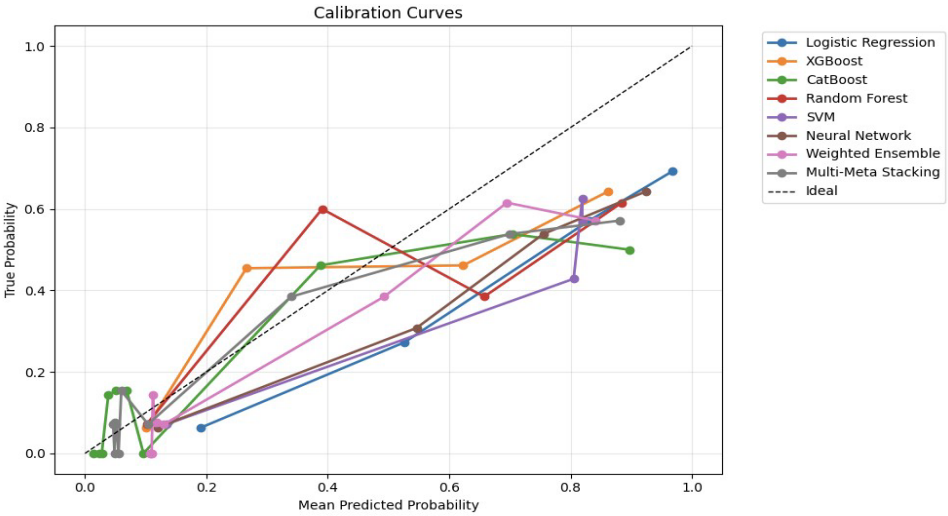

Figure 3. Calibration curves for all models compared to the ideal calibration

The learning curves in Figure 4 provide insight into model stability and data efficiency. CatBoost demonstrated the most stable learning pattern, maintaining high AUC scores (0.90-0.95) across different training data proportions. Logistic Regression showed significant improvement with increased data, starting at 0.75 AUC with 10% training data and stabilizing around 0.85 with 50% data. This suggests that while some models (particularly tree-based) can achieve good performance with limited training data, others benefit substantially from larger datasets.

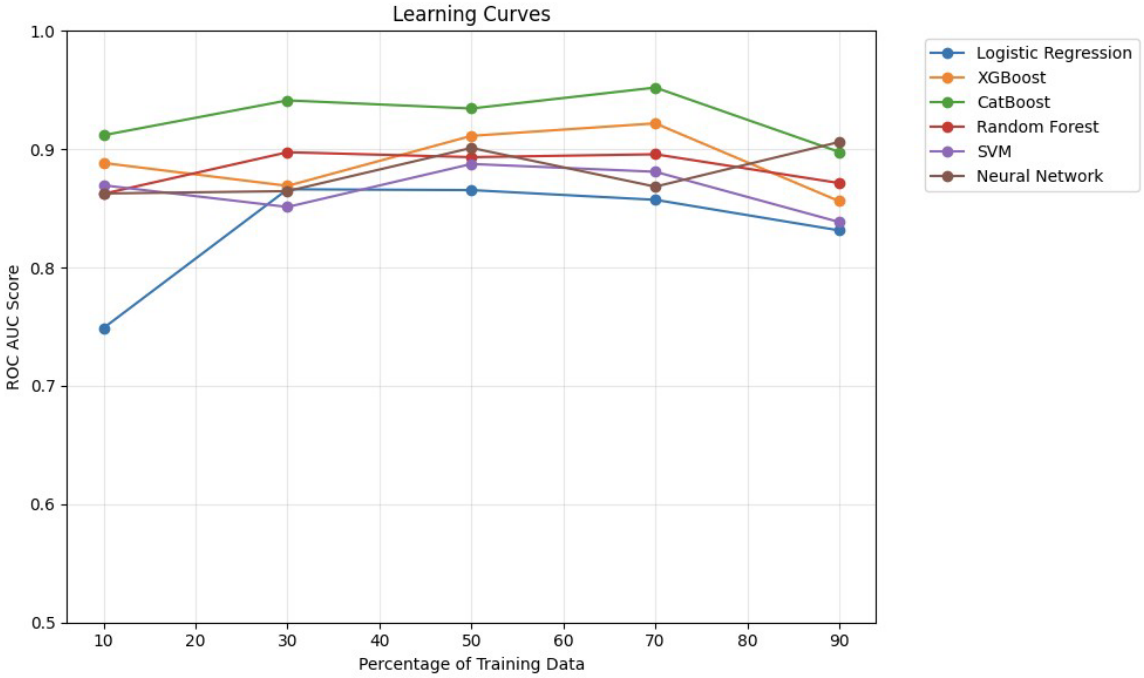

Figure 4. Learning curves showing AUC vs. percentage of training data for key models

The feature importance heatmap (Figure 5) across different models demonstrates both consistency and variation in feature utilization. Feature importance analysis was conducted across five models: Logistic Regression, Random Forest, XGBoost, CatBoost, and Support Vector Machine (SVM). The results, visualized using a hierarchical clustering heatmap, provide insights into the relative contribution of different features in predicting uncontrolled blood pressure. The most influential predictor across all models was "Ever High BP", indicating that prior hypertension history plays a crucial role in classification. Log_URDACT also showed moderate importance, particularly in tree-based models such as XGBoost and CatBoost. Several demographic factors, including age group classifications (e.g., 65–74, 75+), demonstrated varying importance levels, with their contributions being most pronounced in tree-based models. Interestingly, ethnicity-related features exhibited lower importance scores, suggesting that while disparities exist, they may be mediated by other clinical factors rather than direct ethnic classification.

SVM and Random Forest models showed no significant feature importance values, likely due to the nature of these models: SVM relies on decision boundaries rather than explicit feature importance, while the Random Forest implementation used here employs bagging, which may have diluted individual feature contributions. The clustering approach allowed for effective grouping of feature relevance, revealing that age and clinical indicators formed distinct importance clusters across different models. These findings highlight the robustness of ensemble models in identifying key predictors and underscore the role of prior hypertension history and clinical markers over demographic variables in assessing uncontrolled blood pressure risk. Future work could explore feature selection refinement to enhance model interpretability and predictive performance.

Figure 5. Feature importance heatmap across different models with hierarchical clustering

To systematically evaluate the effectiveness of traditional statistical methods versus ML approaches, we conducted a comparative analysis using the NHANES dataset. This comparison focused on assessing model performance, class imbalance handling, feature importance, demographic stratification, and probability estimation. Table 3 presents a structured summary of key methodological differences.

Traditional statistical methods, including odds ratio analysis, Fisher's exact test, and chi-square tests, successfully identified significant associations between cardiovascular risk factors and demographic variables. For example, individuals classified as hypertensive demonstrated an odds ratio (OR) of 11.82 (95% CI: 2.58-54.09, p=0.003) for cardiovascular events compared to those with normal blood pressure. Additionally, a chi-square test (p=0.032) indicated significant disparities in uncontrolled blood pressure across ethnic groups. While these methods effectively quantified individual risk factors, they lacked the ability to model complex, non-linear interactions and had limited capabilities in handling severe class imbalance within the dataset.

In contrast, our ML implementations, particularly ensemble-based methodologies, demonstrated superior predictive performance and robustness against class imbalance. The weighted ensemble model achieved the highest predictive accuracy (AUC = 0.864, Accuracy = 0.868, Precision = 0.812, Recall = 0.790, F1-score = 0.801) compared to traditional models. Additionally, the application of SMOTE-Tomek resampling improved model performance, particularly in underrepresented classes, with observed AUC gains of 7.3% for Logistic Regression (p=0.003) and

Table 3. Comparison of Traditional Statistical Methods vs. Machine Learning Approaches

Methodological Aspect	Traditional Statistical Approaches	Machine Learning Approaches
Primary Methods	Odds ratio analysis (n=666), Fisher's exact test, Chi-square test (p=0.032)	Ensemble learning: Weighted ensemble (AUC=0.864), Multi-meta stacking (AUC=0.819)
Performance Metrics	Odds ratios for hypertension: 11.82 (95% CI: 2.58-54.09, p=0.003)	Best model (Weighted Ensemble): AUC=0.864, Accuracy=0.868, Precision=0.812, Recall=0.790, F1=0.801, Brier score=0.102
Class Imbalance Handling	Raw class distribution (Normal: 80.3%, High BP: 10.1%, Hypertensive: 9.6%)	SMOTE-Tomek: +7.3% improvement for LogReg (p=0.003), +2.1% for CatBoost (p=0.142)
Feature Importance	Significance of blood pressure (p=0.003), ethnicity (p=0.032), gender (p=0.538)	Ever_High_BP (importance=0.97-1.0), Ethnicity_White (0.80-0.95), Gender_Male (0.45-0.55), Log_URDACT (0.35-0.42)
Demographic Analysis	Black vs White: OR=0.83 (95% CI: 0.21-3.37, p=1.0); Gender difference: OR=1.48 (95% CI: 0.45-4.90, p=0.538)	Capture of non-linear interactions between ethnic groups and BP categories (e.g., Other Hispanic with hypertension: 20% risk vs. Normal: 2%)
Probability Estimation	Binary outcomes (Yes/No) with no calibrated probabilities	Calibration curves showing systematic differences across probability ranges (0.20.4)

2.1% for CatBoost (p=0.142). These findings indicate that while ML models effectively mitigate class imbalance, their improvements vary across model architectures. Feature importance analysis provided further insights into key predictors of cardiovascular outcomes. While traditional odds ratio analysis identified Ever_High_BP (p=0.003) and ethnicity (p=0.032) as significant, ML models revealed Ever_High_BP as the most influential predictor (importance = 0.97-1.0), followed by Ethnicity_White (0.80-0.95), Gender_Male (0.45-0.55), and Log_URDACT (0.35-0.42) across various models. Unlike traditional methods, ML approaches captured complex interactions between demographic variables and clinical markers, revealing differences in hypertension risk stratification across racial subgroups. Probability estimation was another critical aspect where ML provided advantages over traditional methods. While traditional statistics produce binary classifications (e.g., hypertensive vs. non-hypertensive) without calibrated probability estimates, ML models provided probabilistic risk scores. However, our results indicate that certain treebased models (e.g., Random Forest, XGBoost) exhibited probability overestimation in the 0.2-0.4 range, while CatBoost provided more conservative estimates (Brier score = 0.118). Although ensemble methods reduced these inconsistencies (Weighted Ensemble Brier score = 0.102), opportunities for further probability calibration remain.

Overall, these findings demonstrate that while traditional statistical methods remain useful for estimating associations and hypothesis testing, properly configured ML approaches, particularly ensemble methods, provide more balanced and nuanced outcomes when analyzing complex healthcare datasets with high-dimensional feature interactions and severe class imbalance. The integration of resampling techniques, ensemble learning, and feature selection contributes to improved model reliability, offering methodological advancements for healthcare data analysis.

5.3 Proposed Improvements

Our analysis highlights several areas for methodological improvements in healthcare survey data modeling. One key finding was the consistent miscalibration observed across models, particularly in tree-based algorithms, which tended to produce overconfident probability estimates. To improve probability estimation, future implementations should incorporate post-processing calibration techniques such as Platt scaling or isotonic regression, especially for clinical applications where well-calibrated risk scores are critical.

Feature selection remains another area for refinement. While filter-based methods like SelectKBest provided reasonable feature sets, a hybrid approach combining filter and wrapper methods could enhance model interpretability

and predictive performance. A structured pipeline using SelectKBest for initial screening, followed by RFECV, would allow models to capture non-linear interactions more effectively while avoiding overfitting. Regarding class imbalance handling, we applied SMOTE-Tomek and found it effective in improving class balance. However, future work could explore adaptive resampling strategies tailored to model architecture. For example, applying SMOTE-Tomek for linear models and more conservative SMOTE for tree-based models could optimize predictive performance while preserving data structure. Further analysis is required to quantify the direct impact of these techniques on model reliability across demographic subgroups.

Finally, while our ensemble learning approaches demonstrated strong performance, additional optimization strategies could further enhance results. Specifically, Bayesian optimization could be used to fine-tune meta-learner hyperparameters, incorporate diversity measures in ensemble construction, and explore model-specific feature selection within ensemble frameworks. These refinements would strengthen the robustness and generalizability of predictive models for healthcare disparity research.

5.4 Validation

To ensure methodological rigor, we employed multiple validation techniques, focusing on cross-validation strategies, resampling effects, subgroup performance analysis, and probability estimation reliability. These approaches provided robust model evaluation and ensured generalizability across different demographic subgroups. Stratified 5-fold cross-validation was implemented to maintain class distribution across folds, mitigating the impact of class imbalance on model performance. This approach ensured that each training and validation fold preserved the original proportion of Normal (80.3%), High BP (10.1%), and Hypertensive (9.6%) categories. The implementation used StratifiedKFold(n_splits=5, shuffle=True, random_state=42), maintaining statistical consistency across resampling iterations.

To further prevent data leakage during resampling, we implemented a custom cross-validation strategy integrating SMOTE within each training fold. This approach allowed synthetic samples to be generated only within the training set, ensuring that the validation data remained an accurate reflection of real-world distributions.

A comparative analysis of SMOTE vs. SMOTE-Tomek revealed distinct performance improvements across different models:

- Logistic Regression: AUC improved by +7.3% (p=0.003) with SMOTE-Tomek compared to training on raw data.
- Tree-Based Models (e.g., CatBoost, XGBoost): More limited gains (+2.1%, p=0.142), suggesting that these models inherently manage class imbalance better than linear models.

These results confirm that hybrid resampling strategies (SMOTE-Tomek) are particularly beneficial for linear models, while ensemble methods further mitigate imbalance effects.

To evaluate generalization across demographic groups, we assessed test set AUC scores for White, Black, Mexican American, and Other Hispanic participants. Results indicated:

- Comparable AUCs for White and Black participants, suggesting stable model performance in these groups.
- Higher variability in Hispanic subgroups, potentially indicating differences in underlying health risk factors or sample size limitations within these demographic categories.

These findings suggest that demographic-aware model calibration may be necessary to mitigate subgroup performance disparities and improve fairness in healthcare risk prediction.

To assess the reliability of probability estimates, we conducted Brier score analysis, a widely used metric for evaluating calibration. The Weighted Ensemble demonstrated the best calibration performance (Brier Score = 0.102), indicating that it provided more reliable probability estimates compared to individual models. In contrast:

- Tree-based models (Random Forest, XGBoost) exhibited overconfidence in probability ranges 0.2-0.4, meaning that their predicted probabilities tended to overestimate the likelihood of events occurring.
- CatBoost demonstrated more conservative probability estimation, yielding a lower Brier score (0.118) compared to other tree-based models.
- Logistic Regression exhibited underestimation across all probability ranges, suggesting potential room for post-hoc probability recalibration.

6. Conclusion

This study provides a comprehensive technical evaluation of analytical approaches for healthcare survey data, specifically focusing on cardiovascular risk prediction using NHANES data. Our findings highlight the most balanced and reliable techniques for analyzing complex healthcare datasets. Ensemble methods, particularly the Weighted Ensemble approach, consistently outperformed individual models, with dynamic weight adjustment effectively

integrating diverse base models. To address class imbalance, the SMOTE-Tomek hybrid resampling technique proved more effective than single-method approaches, particularly for linear models, underscoring the need to optimize both over- and under-sampling techniques in healthcare modeling. Feature engineering insights revealed that clinical history (Ever_High_BP) and demographic factors (ethnicity, gender) were key predictors across models, reinforcing the importance of incorporating both clinical and demographic variables in predictive modeling. However, all models exhibited some degree of miscalibration, particularly tree-based algorithms, emphasizing the necessity of post-training calibration for healthcare applications requiring precise probability estimates.

The application of these advanced methodologies also revealed significant healthcare disparities. Non-Hispanic Black participants had the highest combined prevalence of uncontrolled blood pressure and elevated URDACT levels, while different ethnic groups exhibited varied risk factor patterns. Mexican American and White males exhibited higher URDACT values compared to other groups, suggesting an elevated risk for chronic kidney disease (CKD). While these findings align with existing research linking URDACT levels to CKD progression, they also highlight demographic-specific patterns that may warrant targeted screening and intervention strategies. For instance, White males exhibited elevated URDACT levels despite lower overall hypertension prevalence, potentially reflecting underlying metabolic conditions such as glomerulonephritis or chronic pyelonephritis. Similarly, Mexican American females displayed the lowest rates of uncontrolled blood pressure but higher-than-expected URDACT values, raising concerns about undiagnosed CKD progression within this group. These findings suggest the need for more individualized approaches to CKD risk assessment and early intervention, particularly among subpopulations that may not traditionally be classified as high-risk based on hypertension status alone. Public health initiatives should incorporate demographic-specific screening guidelines to improve early detection and reduce disparities in CKD outcomes. These findings indicate that current anti-hypertensive treatment regimens for non-Hispanic Black patients may be inadequate, contributing to increased cardiovascular and CKD risk. From a data science perspective, this study demonstrates that well-configured ensemble approaches yield more balanced and reliable outcomes than traditional statistical methods when analyzing complex healthcare datasets. The integration of advanced preprocessing, feature engineering, and ensemble modeling provides a robust framework for extracting meaningful insights from healthcare survey data.

Future research should focus on:
Developing adaptive calibration techniques, such as Platt scaling or isotonic regression, to refine probability estimations in tree-based models and enhance clinical interpretability of predictive outputs,

- Exploring demographic-specific modeling strategies to enhance fairness and performance across subgroups,
- Implementing explainable AI techniques to improve interpretability of ensemble predictions, and
- Extending this methodological framework to other healthcare datasets and predictive tasks.

By systematically evaluating and comparing analytical techniques, this study provides data scientists with practical guidelines for improving the reliability and fairness of healthcare survey data analysis, ultimately contributing to more accurate cardiovascular risk assessment across diverse populations and advancing data-driven public health research.

References

Ahmed M. Alaa, T. B. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS One,* (2019). .

Archer K.J., K. R. , Empirical characterization of random forest variable importance measures. *Journal of Biomedical Science and Engineering*, 11,(2008). .

Ben Van Calster, D. N. , A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, 9, (2016). .

Blagus, R. L. , Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC Bioinformatics,* (2013).

CDC. , *Chronic Kidney Disease in the United States, 2023*. Retrieved from Centers for Disease Control and Prevention, (2023). : https://www.cdc.gov/kidney-disease/php/data, - research/index.html#cdc_research_or_data_summary_suggested_citation-suggested-citation

Fan, S., Zhao, Z., Yu, H., Wang, L., Zheng, C., Huang, X., . . . Luo, Y. , Applying probability calibration to ensemble methods to predict 2-year mortality in patients with DLBCL. *BMC Medical Informatics and Decision Making*, 12, (2021). .

Jonathan H Chen, S. M. , Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations. *National Library of Medicine,* (2017). .

Katella, K. , *Why Is Chronic Kidney Disease (CKD) on the Rise? 6 Things to Know*. Retrieved from Yale Medicine, (2024, April 24). : https://www.yalemedicine.org/news/why-is-chronic-kidney-disease-ckd-on-therise#:~:text=BY%20KATHY%20KATELLA%20April%2024,even%20know%20they%20have%20it

Lopez, E. O., Ballard, B. D., & Jan, A. , *Cardiovascular Disease.* Treasure Island: StatPearls Publishing, (2023). . Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK535419/

Naderalvojoud B, H.-B. T. , Improving machine learning with ensemble learning on observational healthcare data. *AMIA Annual Symposium proceedings,*(2024).

Natarajan, K. B. , An adaptive ensemble feature selection technique for model-agnostic diabetes prediction. *Sci Rep*, 15, (2025). .

National Institute of Diabetes and Digestive and Kidney Diseases. , *Kidney Disease Statistics for the United States,*(2024, September). . Retrieved from NIDDK Health Information: https://www.niddk.nih.gov/health-information/health-statistics/kidney-disease#:~:text=Chronic%20Kidney%20Disease,-According%20to%20the&text=CKD%20is%20slightly%20more%20common,Hispanic%20Asian%20adult s%20have%20CKD.

Nitesh V. Chawla, K. W. , SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 37, (2002). .

Omer Sagi, L. R. , Ensemble learning: A survey. *Wiley Interdisciplinary Reviews,* (2018). .

Polikar, R. , Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 24, (2006). .

Rahman, M. M. , A systematic review of epidemiological studies on the association between smokeless tobacco use and coronary heart disease . *Deakin University,* (2011). .

Salvador Garcia, A. F. , SMOTE for Learning from Imbalanced Data: Progress andChallenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research* , 43,(2018). .

Yvan Saeys, I. I. , A review of feature selection techniques in bioinformatics. *Bioinformatics - ISCB*, 10, (2007). .

Biographies

Dr. Parisa Hajibabaee is an Assistant Professor in the Department of Data Science and Business Analytics at Florida Polytechnic University. She holds a Ph.D. in Computer Science from the University of Massachusetts Lowell, an M.S. in Industrial Engineering from West Virginia University, and an M.S. in IT Management from Shahid Beheshti University. Her research focuses on machine learning, statistical modeling, and optimization techniques applied to engineering and healthcare analytics. She has published in peer-reviewed journals and presented at international conferences, contributing to areas such as uncertainty quantification, class-imbalanced learning, and predictive modeling for decision-making in engineering and public health.

Dr. Susan LeFrancois is an Associate Professor in the Department of Data Science and Business Analytics at Florida Polytechnic University. She holds a Ph.D. in Pharmacology and Physiology from the University of Florida. Her research focuses on health equity and pharmacologic interventions. She has published in peer-reviewed journals and presented at international conferences, contributing to areas such as novel drug discovery, ethics, public health, health equity, and health policy.

Amanda Rodriguez is a senior at Florida Polytechnic University. She will graduate in May 2025 with a Bachelor of Science (BS) in Data Science, with a concentration in big data analytics. Currently, she works as an undergraduate data science Research Assistant under the supervision of Dr. LeFrancois.