Strategies for Mitigating Class Imbalance in Diabetes Prediction Models

STA 5353 Final Report

Pritom Roy, Nailah Rawnaq and Carson Slater

August 14, 2025

Abstract

This research investigates the effectiveness of various class imbalance correction strategies in improving diabetes prediction model performance. Using a diabetes dataset from Iraqi hospitals, we compared four different class imbalance mitigation techniques (undersampling, oversampling, SMOTE, and ADASYN) against a baseline approach across six different machine learning models. Our findings reveal that Random Forest models consistently achieved the highest performance across multiple metrics, with class imbalance strategies showing model-specific benefits rather than universal improvements. Random Forest with baseline configuration achieved 98.1% accuracy, while oversampling variants reached 99.1% balanced accuracy and 99.9% ROC AUC, establishing Random Forest as the optimal choice for similar diabetes prediction tasks.

1 Introduction

1.1 Background and Motivation

Class imbalance represents one of the most pervasive challenges in real-world machine learning applications, particularly in medical diagnosis where the condition of interest (disease) typically affects a minority of the population. In healthcare datasets, this imbalance can arise from two primary sources: the natural rarity of the condition in the population, or sampling bias that fails to adequately represent the true population distribution.

The implications of class imbalance extend beyond simple statistical concerns. In medical contexts, misclassifying minority class instances (failing to detect disease) can have severe consequences, making it crucial to develop models that maintain high sensitivity while preserving overall predictive accuracy. Traditional machine learning algorithms often exhibit bias toward the majority class, achieving high overall accuracy while performing poorly on the minority class that is often of greatest clinical interest.

1.2 Research Objectives

This study aims to comprehensively evaluate different strategies for addressing class imbalance in diabetes prediction models. Specifically, we seek to: (1) compare the effectiveness of four distinct class imbalance mitigation strategies, (2) assess performance across multiple machine learning algorithms, (3) evaluate models using a comprehensive set of performance metrics appropriate for imbalanced classification, and (4) provide evidence-based recommendations for practitioners working with similar medical datasets.

2 Literature Review

2.1 Class Imbalance in Medical Prediction

Class imbalance in medical prediction has been extensively studied, with researchers developing numerous approaches to address the fundamental challenge of learning from skewed data distributions. The problem is particularly acute in healthcare applications where the prevalence of disease is naturally low, yet accurate identification of positive cases is critically important.

2.2 Remedial Strategies

Oversampling addresses class imbalance by increasing the representation of minority classes. Simple random oversampling duplicates existing minority class instances, while more sophisticated approaches like SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic examples by interpolating between existing minority class instances and their k-nearest neighbors (Chawla et al. 2002). Undersampling reduces the size of the majority class to achieve better balance, though this approach risks losing potentially valuable information by discarding data points. Advanced techniques like ADASYN (Adaptive Synthetic Sampling Approach) improve upon SMOTE by generating synthetic samples adaptively, focusing more attention on minority class examples that are difficult to learn (He et al. 2008).

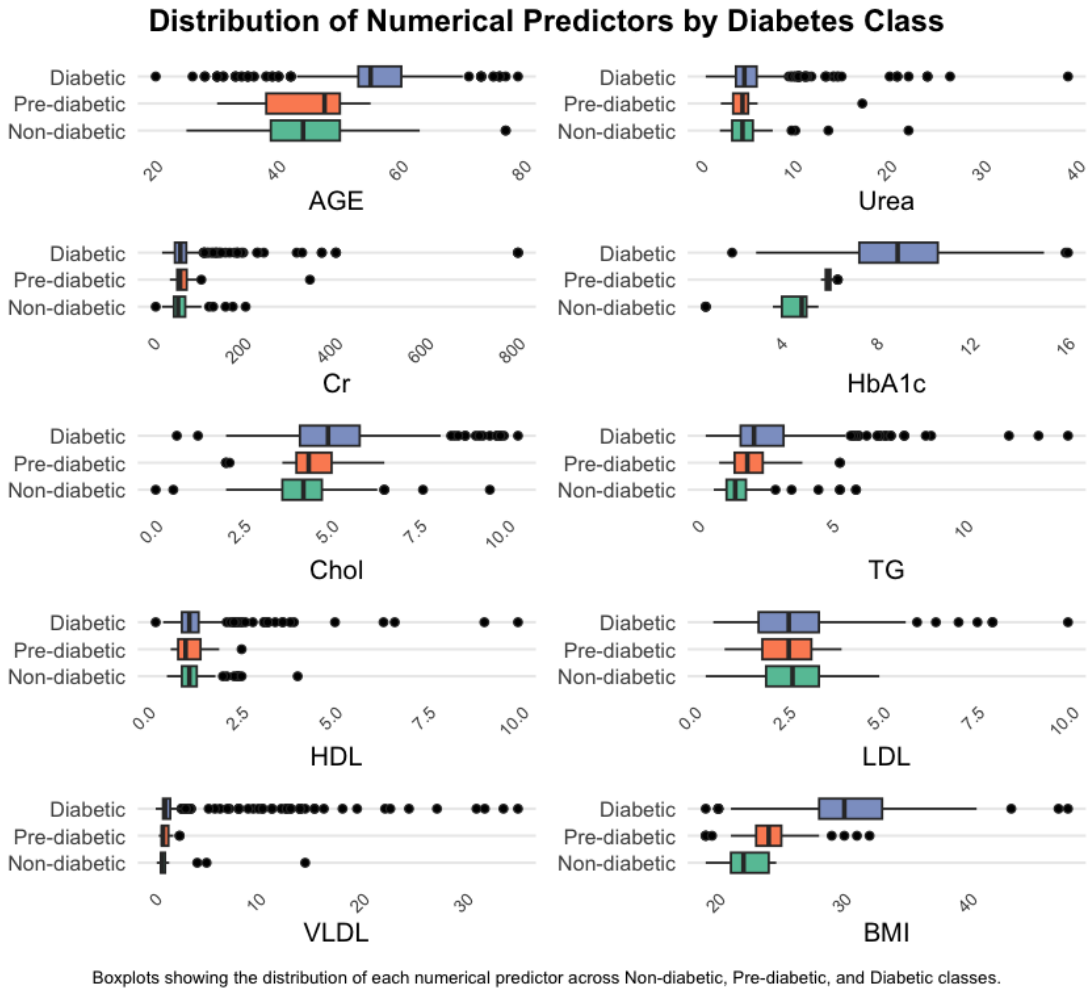

Figure 1: Relationships between the continuous variables and response classes.

3 Methodology

3.1 Dataset Description

Our analysis utilized a diabetes dataset originally used for a 2022 study by Olisah et al., which employed neural networks and various preprocessing methods for diabetes diagnosis. The dataset was sourced from the Medical City Hospital and the Specialized Center for Endocrinology and Diabetes in Iraq, providing a real-world clinical context for our analysis.

The dataset contained 1,000 original patient records, reduced to 826 unique records after duplicate removal. The response variable exhibited trichotomous classification with severe class imbalance: diabetic patients comprised 83.5% (689 patients), non-diabetic patients represented 11.6% (96 patients), and pre-diabetic patients constituted only 4.8% (40 patients) of the sample.

3.2 Feature Variables

The dataset includes 11 predictor variables encompassing demographic, metabolic, and biochemical markers commonly used in diabetes assessment. Table 1 describes all of these features in detail. Demographic variables included gender (biological sex coded as 0=Female, 1=Male) and age in years, with diabetes risk increasing significantly after age 45.

Figure 1 illustrates several interesting trends emerge regarding the distribution of numerical predictors across the three diabetes classes. For **Age**, there's a clear upward trend: the median age increases from non-diabetic to pre-diabetic to diabetic groups. This suggests that the risk of developing diabetes rises with age. The boxplots for **HbA1c** and **BMI** also show a consistent increase in median values as the diabetes class progresses from non-diabetic to pre-diabetic and finally to diabetic. This is particularly notable for **HbA1c**, as it's a key indicator of long-term blood glucose levels. Conversely, **HDL** (high-density lipoprotein) shows a downward trend; its median value is highest in the non-diabetic group and lowest in the diabetic group, indicating that lower levels of "good cholesterol" are associated with diabetes. Lastly, some variables like **Urea**, **Cr** (creatinine), and **VLDL** (very-low-density lipoprotein) appear to have a less clear distinction between the pre-diabetic and diabetic classes compared to the non-diabetic group, though outliers are present across all classes.

Table 1: Feature Summary for Diabetes Prediction

Feature	Summary Description
Gender	Biological sex (0=Female, 1=Male); can influence risk.
AGE	Age in years; risk increases significantly after 45.
Urea	Blood urea (mg/dL); high levels may indicate kidney issues. Normal: ~7–20 mg/dL.
Cr	Blood creatinine (mg/dL); a key marker for kidney function. Normal: ~0.6–1.3 mg/dL.
HbA1c	Average 2-3 month blood glucose (%); Normal: <5.7%, Diabetic: >6.5%.
Chol	Total blood cholesterol (mg/dL); high levels are a cardiovascular risk. Normal: <200 mg/dL.
TG	Blood triglycerides (mg/dL); high levels are linked to insulin resistance. Normal: <150 mg/dL.
HDL	"Good" cholesterol (mg/dL); higher is better. Ideal: >40 (men), >50 (women).
LDL	"Bad" cholesterol (mg/dL); contributes to arterial plaque. Optimal: <100 mg/dL.
VLDL	Another "bad" cholesterol (mg/dL) that carries triglycerides. Normal: 2–30 mg/dL.
BMI	Body fat measure (kg/m^2); obesity (BMI >30) is a major risk factor.
Class	**Target variable**: Diabetes status (0=Non-Diabetic, 1=Diabetic, 2=Pre-Diabetic).

3.3 Class Imbalance Mitigation Strategies

We implemented and compared five distinct approaches: (1) Baseline with no class imbalance correction representing standard machine learning practice, (2) Undersampling through systematic reduction of majority class instances, (3) Oversampling via replication of minority class instances, (4) SMOTE generating synthetic minority class examples through interpolation between existing instances and k-nearest neighbors, and (5) ADASYN providing adaptive synthetic sample generation focusing additional attention on difficult-to-learn minority class examples.

3.4 Machine Learning Models

We evaluated two statistical models and four different machine learning algorithms representing diverse modeling approaches. Tree-based methods included Decision Tree classifiers using recursive binary splits for interpretable rule-based classification, and Random Forest ensemble methods combining multiple decision trees with bootstrap aggregation. Linear methods encompassed Multinomial Logistic Regression extending logistic regression for multi-class problems, and Quadratic Discriminant Analysis (QDA) assuming quadratic decision boundaries between classes. Non-linear methods comprised Support Vector Machines with linear kernels for maximum margin classification and Neural Networks using single-layer perceptrons with hidden

layers for non-linear pattern recognition.[1]

3.5 Model Training and Validation

Table 2: Best hyperparameters for each model type and sampling strategy after cross validation.

	Decision Tree			Neural Network		Random Forest			Linear SVM
Model Name	Cost Complexity	Tree Depth	Min Node Size	Penalty	Hidden Units	Min Node Size	mtry	Number of Trees	SVM Cost
Baseline	0.000 002 83	5	8	0.0118	6	14	7	500	2.82
Undersampling	0.005 52	13	12	0.004 51	8	13	5	1500	12.4
Oversampling	0.0394	12	37	0.835	7	15	5	1500	2.82
Smote	0.001 70	5	22	0.0118	6	11	3	1500	1.39
Adasyn	0.000 007 85	8	29	0.004 51	8	2	3	1500	2.82

The modeling process was designed to ensure robust performance and generalizability across the evaluated class imbalance strategies and machine learning models. The dataset was initially split into training and test sets, with 75% of the data allocated for training and 25% reserved for final evaluation, maintaining the original class distribution to preserve the imbalance characteristics. The training set was further divided into 10 subsets for cross-validation, ensuring each subset reflected the class proportions to support reliable hyperparameter tuning.

For each combination of imbalance mitigation strategy—baseline, undersampling, oversampling, SMOTE, and ADASYN—and model type—Decision Tree, Random Forest, Multinomial Logistic Regression, Quadratic Discriminant Analysis, Linear Support Vector Machine, and Neural Network—a customized preprocessing pipeline was applied. This pipeline included converting categorical variables into a suitable format and, for balanced strategies, adjusting the class representation through replication of minority instances, reduction of majority instances, or generation of synthetic examples. For synthetic methods, additional steps normalized numeric variables and encoded categorical ones to enhance model performance, with specific adjustments tailored to the complexity of each approach.

Hyperparameter optimization was conducted systematically for each model-strategy pair, exploring a range of configurations to identify the best settings. This involved testing various levels of model complexity, such as tree depth and minimum node size for tree-based methods, ensemble size and feature selection for forest models, regularization strength for linear and neural models, and penalty terms for support vector methods. The optimization process used a random search over 10 to 15 candidate settings per model, prioritizing discriminative performance in the multi-class setting. A fixed seed was applied during sampling steps to ensure reproducibility across runs. The best configuration was selected based on cross-validation results and then applied to train the final model on the full training set, with performance assessed on the holdout test set. Parallel processing was utilized to accelerate the computationally intensive tuning phase, leveraging multiple processing units to handle the extensive grid search efficiently.

This approach ensured that each model was optimally configured for the specific data characteristics introduced by the imbalance correction methods, as summarized in Table 2. The decision to limit Neural Networks to a single layer was made to manage computational demands within the project timeline, though more complex architectures could be explored with additional resources.

3.6 Performance Evaluation Metrics

Given the multi-class nature of our problem and the importance of minority class detection in diabetes prediction, we employed seven comprehensive evaluation metrics. These metrics were selected to provide a balanced assessment of model performance, accounting for class imbalance and the need for reliable minority class identification. Overall performance was evaluated using Accuracy, representing the proportion of correctly classified instances, and Balanced Accuracy, which averages recall across all classes to give equal weight regardless of class frequency. Class-specific measures included Precision, focusing on avoiding false positives; Recall (or Sensitivity), emphasizing the capture of all positive instances; and F1 Score, the harmonic

[1]Using a single layer was a conscious decision for computation purposes. Had the timeline for this project been longer, we would have considered more complicated NN architechures.

mean of precision and recall for a balanced view. Advanced metrics comprised Cohen's Kappa, assessing agreement beyond chance, and ROC AUC, measuring discriminative ability across classification thresholds. For multi-class problems, precision, recall, and F1 scores were computed using macro-averaging to ensure equal weighting of each class, preventing bias toward the majority class in imbalanced datasets.

Balanced Accuracy is particularly valuable for imbalanced multi-class scenarios, as it mitigates the misleading high scores that overall accuracy might yield by prioritizing majority class performance. It is defined as:

$$\text{Balanced Accuracy} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i},$$

where C is the number of classes, TP_i is true positives for class i, and FN_i is false negatives for class i. This equal-weight averaging makes it robust for datasets like ours, where the diabetic class dominates.

Precision, Recall, and F1 Score offer class-specific insights, with macro-averaging ensuring fairness across classes in multi-class settings. Precision quantifies the avoidance of false positives:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad P_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FP_i},$$

where FP is false positives. Recall focuses on identifying all positives:

$$\text{Recall} = \frac{TP}{TP + FN}, \quad R_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i}.$$

The F1 Score balances these:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, \quad F1_{\text{macro}} = \frac{1}{C}\sum_{i=1}^{C}\frac{2TP_i}{2TP_i + FP_i + FN_i}.$$

Macro-averaging is ideal for multi-class imbalance, as it treats each class equally, highlighting performance on minority classes like pre-diabetic patients.

Cohen's Kappa evaluates overall agreement beyond what would be expected by chance, making it suitable for multi-class problems where random guessing could skew results due to imbalance:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement proportion, and p_e is the expected chance agreement, computed as $\sum_{i=1}^{C}(\text{row}_i \cdot \text{col}_i / N^2)$ from the confusion matrix, with N as total observations. Values range from less than 0 (worse than chance) to 1 (perfect agreement), providing a holistic measure that accounts for all classes.

ROC AUC assesses the model's ability to distinguish classes across thresholds. For binary cases, it is the area under the curve of true positive rate (TPR) versus false positive rate (FPR):

$$\text{AUC} = \int_0^1 \text{TPR}(t)\, d\text{FPR}(t).$$

In multi-class, we used macro-weighted one-vs-rest averaging:

$$\text{AUC}_{\text{macro-wt}} = \frac{\sum_{i=1}^{C} w_i \cdot \text{AUC}_i}{\sum_{i=1}^{C} w_i},$$

where w_i is the number of instances in class i. This weighted approach adapts well to imbalance, emphasizing overall separability while considering class prevalence.

4 Results

4.1 Hyperparameter Optimization Results

The hyperparameter optimization process revealed interesting patterns in how different models adapted to various class imbalance strategies. Decision Tree adaptations showed baseline models favoring minimal cost complexity (0.00000283) with moderate depth (5 levels), while undersampling led to more complex trees with higher cost complexity (0.00552) and deeper structures (13 levels). Synthetic methods (SMOTE, ADASYN) resulted in intermediate complexity configurations.

Neural Network configurations demonstrated regularization penalties varying significantly across strategies (0.00451 to 0.835), with hidden unit counts remaining relatively stable (6-8 units) across most approaches. Oversampling required the highest regularization (0.835), suggesting increased model complexity. Random Forest behavior showed all balanced sampling methods preferring larger ensembles (1,500 trees versus 500 for baseline), with SMOTE and ADASYN favoring fewer features per split (mtry=3), potentially indicating more careful feature selection.

4.2 Comparative Model Performance

The comprehensive performance evaluation revealed several key patterns across models and metrics. Random Forest models consistently achieved the highest performance across virtually all metrics, with particularly impressive results when combined with appropriate class imbalance strategies. Key achievements included 98.1% accuracy with baseline configuration, 99.1% balanced accuracy with oversampling, 96.4% F1 Score with baseline, 92.9% Cohen's Kappa with baseline, and 99.9% ROC AUC with oversampling.

Table 3 presents the optimal model-strategy combinations for each performance metric, demonstrating the model-dependent nature of class imbalance strategy effectiveness.

Analysis of model-specific strategy benefits revealed differential responses to class imbalance correction. Tree-based models (Decision Tree, Random Forest) showed mixed responses, often performing best with baseline approaches for some metrics while benefiting from oversampling for others. Linear models (Multinomial Logistic Regression, SVM) demonstrated consistent improvement with synthetic data generation methods, particularly SMOTE and ADASYN. Neural Networks showed strong performance with SMOTE across multiple metrics, suggesting synthetic data generation helped learn better decision boundaries. QDA generally performed best with baseline approaches, possibly due to assumptions about class-specific covariance structures.

5 Visual Analysis of Model Performance Patterns

The comprehensive performance visualization reveals several critical patterns across the seven evaluation metrics and five sampling strategies.

5.1 Performance Metric Patterns

The performance plot demonstrates clear model hierarchies that persist across different class imbalance strategies. Random Forest (orange line) consistently maintains the highest performance across nearly all metrics, showing remarkable stability with performance scores typically above 0.95. Decision Tree and QDA models show more variable performance, with Decision Trees particularly sensitive to the choice of sampling strategy.

5.2 Remedial Strategy Effects by Model

Figure 2 reveals distinct model-specific responses to class imbalance correction. Random Forest and Decision Tree, both tree-based models, show relatively stable performance across sampling strategies, with Random Forest maintaining superiority regardless of the approach used. Neural Network displays dramatic performance variations across sampling strategies, with particularly poor performance under certain configurations (notably

Table 3: Best Method per Model-Metric Combination

Model	**Method**	**Metric**	**Estimate**
Decision Tree	Baseline	Accuracy	0.971
Decision Tree	Oversampling	Balanced Accuracy	0.967
Decision Tree	Baseline	F1 Score	0.939
Decision Tree	Baseline	Cohen's Kappa	0.894
Decision Tree	Baseline	Precision	0.933
Decision Tree	Oversampling	Recall	0.966
Decision Tree	Baseline	ROC AUC	0.992
Multinomial Logistic Reg.	ADASYN	Accuracy	0.908
Multinomial Logistic Reg.	ADASYN	Balanced Accuracy	0.923
Multinomial Logistic Reg.	ADASYN	F1 Score	0.776
Multinomial Logistic Reg.	ADASYN	Cohen's Kappa	0.717
Multinomial Logistic Reg.	ADASYN	Precision	0.718
Multinomial Logistic Reg.	ADASYN	Recall	0.880
Multinomial Logistic Reg.	SMOTE	ROC AUC	0.973
Neural Network	SMOTE	Accuracy	0.947
Neural Network	SMOTE	Balanced Accuracy	0.967
Neural Network	SMOTE	F1 Score	0.878
Neural Network	SMOTE	Cohen's Kappa	0.842
Neural Network	ADASYN	Precision	0.843
Neural Network	SMOTE	Recall	0.953
Neural Network	SMOTE	ROC AUC	0.989
QDA	Baseline	Accuracy	0.942
QDA	Baseline	Balanced Accuracy	0.929
QDA	Baseline	F1 Score	0.877
QDA	Baseline	Cohen's Kappa	0.818
QDA	Baseline	Precision	0.858
QDA	Baseline	Recall	0.901
QDA	Oversampling	ROC AUC	0.961
Random Forest	Baseline	Accuracy	0.981
Random Forest	Oversampling	Balanced Accuracy	0.991
Random Forest	Baseline	F1 Score	0.964
Random Forest	Baseline	Cohen's Kappa	0.929
Random Forest	Baseline	Precision	0.964
Random Forest	Oversampling	Recall	0.990
Random Forest	Oversampling	ROC AUC	0.999
Linear SVM	SMOTE	Accuracy	0.932
Linear SVM	SMOTE	Balanced Accuracy	0.961
Linear SVM	SMOTE	F1 Score	0.859
Linear SVM	SMOTE	Cohen's Kappa	0.805
Linear SVM	Baseline	Precision	0.887
Linear SVM	SMOTE	Recall	0.947
Linear SVM	SMOTE	ROC AUC	0.988

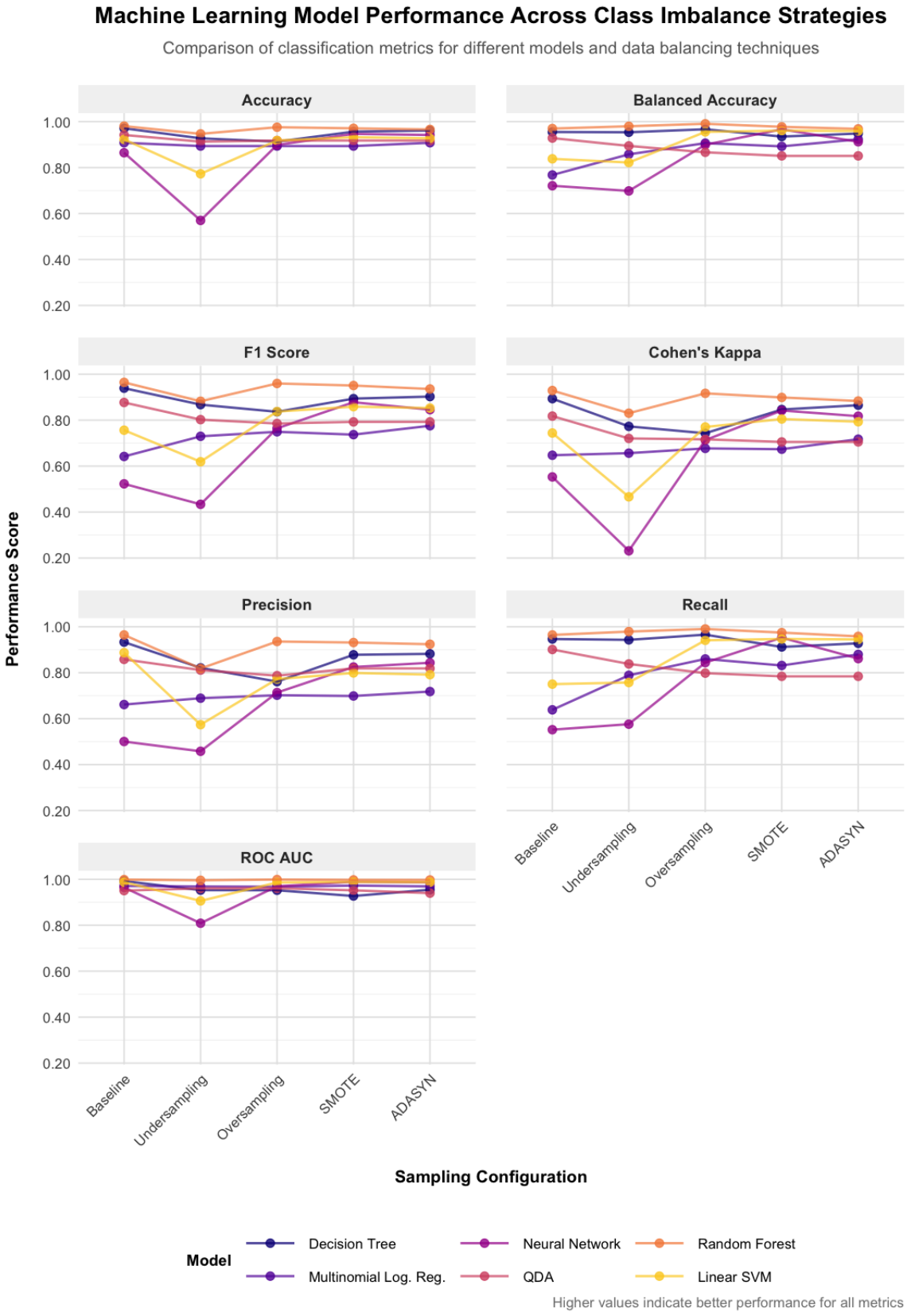

Figure 2: Model performance by remedial strategy.

dropping to around 0.4 for some metrics with undersampling) but achieving strong performance with SMOTE and ADASYN. Multinomial Logistic Regression shows consistent improvement with synthetic sampling methods (SMOTE and ADASYN) compared to baseline, particularly evident in precision and recall metrics. Linear SVM demonstrates moderate sensitivity to sampling strategy with generally improved performance using synthetic methods.

5.3 Metric-Specific Observations

Different metrics reveal varying aspects of model performance. Accuracy shows that most models maintain high accuracy (>0.8) across strategies, masking potential issues with minority class detection. Balanced Accuracy provides a clearer picture of true performance across all classes, showing more variation between sampling strategies. F1 Score and Cohen's Kappa highlight the challenge of minority class prediction, with several model-strategy combinations showing substantial drops in performance.

5.4 Class-Specific Performance Patterns

The confusion matrices in Figure 3 reveal consistent patterns in how different models handle the three diabetes categories. For the Diabetic Class (Y) Performance, all models show strong performance in correctly identifying diabetic patients, with most achieving high true positive rates (diagonal values typically >150 out of ~175 diabetic cases). This reflects both the majority class advantage and the clinical distinctiveness of diabetic patients in the dataset.

Non-Diabetic Class (N) Performance shows more variability across model-strategy combinations. Random Forest consistently achieves near-perfect classification of non-diabetic patients across all sampling strategies, while other models show varying degrees of misclassification, particularly with undersampling approaches.

The Pre-Diabetic Class (P) Performance, representing the smallest class (~40 patients), shows the most dramatic variation across approaches. Baseline models generally achieve moderate success in identifying pre-diabetic patients, while oversampling shows improved pre-diabetic detection across most models. Undersampling consistently demonstrates poor performance for pre-diabetic classification, and SMOTE and ADASYN show variable performance depending on the base model.

5.5 Model-Specific Confusion Matrix Insights

Random Forest Excellence is evident in the confusion matrices, which show consistently clean diagonal patterns with minimal off-diagonal errors across all sampling strategies, explaining its superior performance across metrics. Neural Network Volatility is revealed through the confusion matrices as the source of its variable performance. While it can achieve excellent results with appropriate sampling (SMOTE/ADASYN), it shows substantial misclassification under suboptimal conditions, particularly with undersampling.

Undersampling Problems are consistently visible in the undersampling columns, which show the most scattered confusion matrices across all models, with increased off-diagonal errors indicating poor class discrimination. This visual evidence strongly supports the quantitative finding that undersampling is consistently detrimental. Multi-Class Boundary Effects are observed in several models that show interesting boundary confusion patterns, particularly between diabetic and pre-diabetic classes, which is clinically relevant as these categories represent a natural progression continuum.

5.6 Metric-Specific Insights

While several models achieved high overall accuracy with baseline data, balanced accuracy often improved significantly with class imbalance correction, particularly oversampling. This pattern highlights the importance of using multiple metrics when evaluating models on imbalanced data. The F1 score results revealed interesting trade-offs between precision and recall. Models achieving high precision with baseline data sometimes sacrificed recall, while class imbalance methods generally improved recall at some cost to precision. ROC AUC scores were consistently high across most model-strategy combinations, with Random Forest achieving near-perfect discrimination (99.9%) when combined with oversampling.

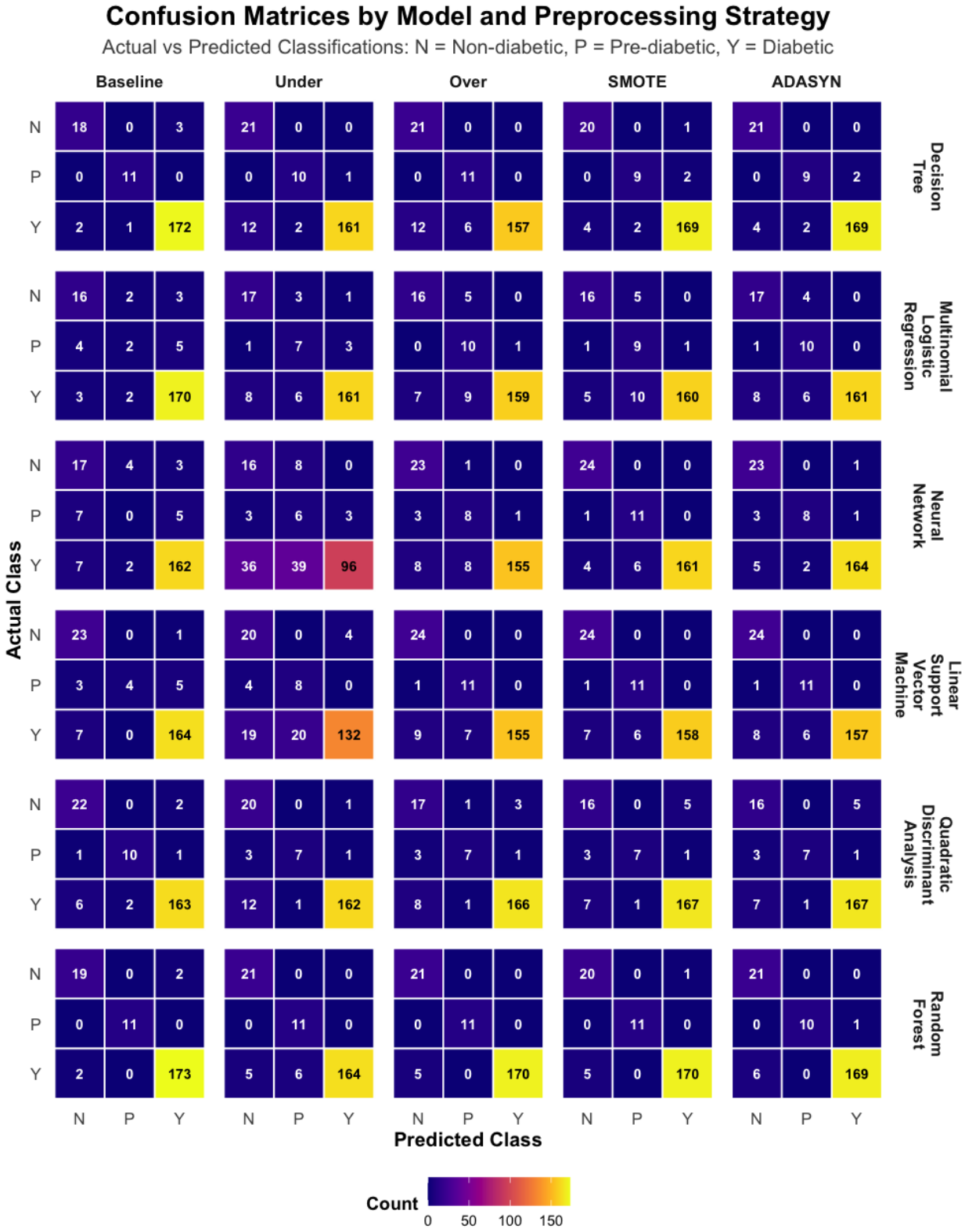

Figure 3: The confusion matrix heatmap provides detailed insight into classification patterns across all model-strategy combinations for the trichotomous diabetes classification problem.

5.7 Class Imbalance Strategy Effectiveness

Undersampling consistently emerged as the least effective strategy across all models and metrics, aligning with theoretical expectations as discarding majority class data reduces information available for learning while potentially eliminating important boundary cases. Simple oversampling showed particular strength in improving recall and balanced accuracy metrics, proving especially effective for Random Forest models when detecting minority class instances was the primary concern.

Synthetic methods (SMOTE and ADASYN) demonstrated model-dependent benefits, with Neural Networks and SVMs benefiting substantially from synthetic data generation. Linear models achieved best performance with ADASYN, while tree-based methods showed variable responses, sometimes performing better with baseline data.

5.8 Optimal Model-Strategy Combinations

Table 4 summarizes the preprocessing and model combinations that achieved optimal performance for each metric, highlighting Random Forest dominance across most evaluation criteria.

Table 4: Best Model-Strategy Combinations for Each Performance Metric

Model	**Class Imbalance Correction**	**Metric**	**Estimate**
Random Forest	Baseline	Accuracy	0.981
Random Forest	Oversampling	Balanced Accuracy	0.991
Random Forest	Baseline	F1 Score	0.964
Random Forest	Baseline	Cohen's Kappa	0.929
Random Forest	Baseline	Precision	0.964
Random Forest	Oversampling	Recall	0.990
Random Forest	Oversampling	ROC AUC	0.999

6 Discussion

6.1 Interpretation of Key Findings

The consistent superior performance of Random Forest models across metrics and class imbalance strategies can be attributed to several factors. Ensemble averaging reduces overfitting while maintaining model flexibility, bootstrap sampling naturally provides some protection against class imbalance, feature randomization helps the model learn robust decision boundaries, and tree-based splits can naturally adapt to complex class boundaries.

The finding that class imbalance strategies show model-dependent benefits rather than universal improvements has important practical implications. This suggests that practitioners should test multiple combinations of models and remedial strategies, avoid assuming that class imbalance correction will universally improve performance, and consider the specific characteristics of their data and problem requirements.

The consistent poor performance of undersampling reinforces theoretical concerns about information loss. In medical applications where data collection is often expensive and time-consuming, discarding potentially valuable information appears particularly problematic.

6.2 Clinical Implications

The choice between different models and class imbalance strategies should consider the intended clinical use. For screening applications where high sensitivity is crucial, oversampling combined with Random Forest models offers excellent recall. For diagnostic confirmation where specificity is important, baseline Random Forest models provide excellent overall performance.

The trichotomous nature of our diabetes classification (non-diabetic, pre-diabetic, diabetic) adds complexity to clinical interpretation. The strong performance on balanced accuracy suggests that our best models can effectively distinguish between all three categories, which is valuable for clinical decision-making.

This study provides a template for evaluating class imbalance strategies in medical prediction tasks by testing multiple remedial approaches across diverse model types, using comprehensive metric evaluation appropriate for imbalanced multi-class problems, and providing statistical rigor through cross-validation and hyperparameter optimization. Our findings offer evidence-based guidance for practitioners facing similar class imbalance challenges in medical prediction tasks.

7 Limitations

7.1 Dataset Limitations

The dataset originates from Iraqi hospitals, which may limit generalizability to other populations with different genetic backgrounds, dietary patterns, or healthcare systems. With 826 unique patient records, the dataset size, while adequate for initial analysis, may not capture the full complexity of diabetes prediction in larger populations. The cross-sectional nature of the data prevents assessment of prediction accuracy over time and doesn't capture disease progression dynamics.

7.2 Methodological Limitations

We focused on four primary remedial strategies, but numerous other approaches exist, including cost-sensitive learning, ensemble methods specifically designed for imbalanced data, and threshold optimization techniques. While we employed systematic grid search with cross-validation, the hyperparameter spaces explored may not have captured all optimal configurations, particularly for complex models like neural networks. The study utilized features as provided in the original dataset without extensive feature engineering, transformation, or selection processes that might improve model performance.

7.3 Evaluation Limitations

Despite using seven different evaluation metrics, other measures such as area under the precision-recall curve might provide additional insights, particularly for severely imbalanced datasets. The study lacks formal statistical testing to determine whether observed performance differences between strategies are statistically significant.

8 Conclusions and Recommendations

8.1 Primary Conclusions

Random Forest consistently outperformed all other algorithms across metrics and class imbalance strategies, making it the recommended choice for diabetes prediction tasks with similar data characteristics. Class imbalance correction strategies do not uniformly improve performance across all models, emphasizing the need for comprehensive evaluation rather than assuming universal benefit. Undersampling consistently yielded the poorest performance across all model-metric combinations, suggesting that information loss outweighs any benefit from class balance.

Simple oversampling, despite its theoretical limitations, proved effective in improving balanced accuracy and recall, particularly for ensemble methods. SMOTE and ADASYN showed particular benefits for neural networks and support vector machines, suggesting these approaches are valuable when using these algorithms.

8.2 Practical Recommendations

For practitioners, we recommend starting with Random Forest given its consistent superior performance for similar diabetes prediction tasks. Rather than assuming class imbalance correction will help, systematically

evaluate baseline and corrected approaches for your specific model and data combination. If implementing class imbalance correction, favor oversampling or synthetic methods over undersampling to preserve information. Evaluate models using multiple metrics, particularly balanced accuracy and F1 score, when working with imbalanced medical data.

For researchers, future work should investigate advanced techniques including cost-sensitive learning, ensemble methods designed for imbalanced data, and threshold optimization approaches. Studies using datasets from multiple healthcare systems and populations would improve generalizability. Investigating how class imbalance strategies perform in predicting disease progression over time would provide valuable clinical insights.

8.3 Future Directions

Investigation of more sophisticated approaches such as cost-sensitive learning, which assigns different misclassification costs to different classes, or ensemble methods specifically designed for imbalanced data like EasyEnsemble or BalanceCascade represents a promising direction. Exploration of deep learning architectures with built-in mechanisms for handling class imbalance, such as focal loss functions or attention mechanisms that focus on difficult examples, could provide additional insights.

Bibliography

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. "SMOTE: Synthetic Minority Over-Sampling Technique." *Journal of Artificial Intelligence Research* 16 (June):321–57. https://doi.org/10.1613/jair.953.

Elreedy, Dina, and Amir F. Atiya. 2019. "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for Handling Class Imbalance." *Information Sciences* 505 (December):32–64. https://doi.org/10.1016/j.ins.2019.07.070.

He, Haibo, Yang Bai, Edwardo Garcia, and Shutao Li. 2008. "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning." In, 1322–28. https://doi.org/10.1109/IJCNN.2008.4633969.

Olisah, Chollette C., Lyndon Smith, and Melvyn Smith. 2022. "Diabetes Mellitus Prediction and Diagnosis from a Data Preprocessing and Machine Learning Perspective." *Computer Methods and Programs in Biomedicine* 220 (June):106773. https://doi.org/10.1016/j.cmpb.2022.106773.