2026 ASA Student Data Challenge

Predictive Modeling of HDL Cholesterol Levels

A Triple-Stacked Ensemble Approach using NHANES 2024 Data

Matheus Gomes, Satvik Hulikere, and Joaquin Hidalgo

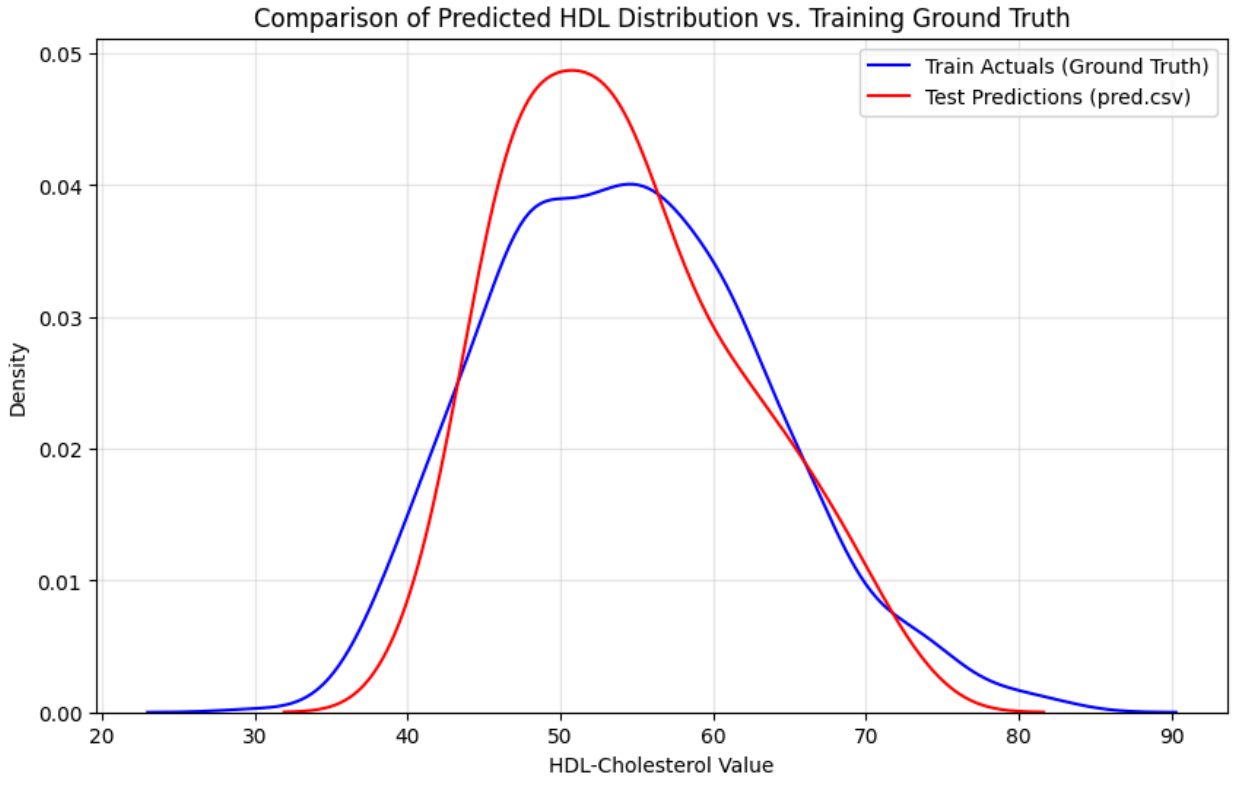

Figure 1: *Comparison of Predicted HDL Distribution vs. Training Ground Truth. The strong overlap indicates the Triple-Stack model successfully captured the underlying statistical structure.*

Introduction

This report details the development of a machine learning pipeline to predict High-Density Lipoprotein (HDL) cholesterol levels using a curated dataset of 1,200 individuals from the 2024 NHANES survey. HDL is a critical biomarker for cardiovascular health, and identifying its dietary and physiological predictors is essential for public health interventions. To address the high dimensionality and structural complexity of the data, we employ a stacked ensemble learning approach that integrates linear regularization and nonlinear machine learning models. Specifically, the framework combines Lasso regression, Random Forests, and gradient boosting, with a Ridge regression meta-model trained on cross-validated predictions to optimally weight each base learner. This approach aims to improve predictive accuracy, reduce overfitting, and enhance generalization by leveraging complementary modeling assumptions within a unified ensemble framework.

Data Preprocessing

The NHANES dataset contains a diverse set of variables, many of which are highly correlated or exhibit low variability. To prepare the data for modeling, we applied several preprocessing steps aimed at reducing noise and improving model stability:

- **Feature Engineering:** We renamed technical NHANES codes to human-readable labels to improve interpretability. This step was particularly important for model diagnostics and feature-level reasoning.
- **Noise Reduction:** We identified and removed variables with near-zero variance and handled multicollinearity by dropping features with a correlation coefficient greater than 0.9 strictly for linear models. This step was especially important for models such as Lasso and Ridge, which can be sensitive to highly correlated predictors.
- **Final Feature Set:** After cleaning, 77 features were retained for training, including dietary intake (e.g., Energy, Protein), body measures (e.g., Waist Circumference, BMI), and demographic data (e.g., Age, Gender).

Modeling Strategy

Given the complexity of the relationships governing HDL cholesterol, no single modeling approach is expected to perform optimally across all regions of the feature space. For our final model we adopted a **Triple-Stacked Regressor** that combines three distinct algorithms:

1. **Lasso Regression:** Acts as a linear stabilizer, penalizing irrelevant features to prevent overfitting. Lasso shrinks irrelevant coefficients toward zero, reducing variance and improving interpretability.
2. **Random Forest:** An ensemble of 600 decision trees (max depth 10). This model captures nonlinear relationships and higher-order interactions between dietary, physiological, and demographic variables
3. **XGBoost:** A gradient boosting model (learning rate 0.01) to incrementally correct residual errors from previous iterations. This model excels at capturing subtle nonlinear patterns and reducing systematic bias left by other models.

Stacked Ensemble and Meta-Learning

In order to capture the complex structure underlying HDL cholesterol levels, we employed a stacked ensemble learning framework that combines multiple complementary regression models. Lasso regression was used to provide a stable linear baseline and implicit feature selection through regularization, while Random Forest and XGBoost models were incorporated to capture nonlinear relationships and interaction effects among dietary, anthropometric, and demographic variables. Each base model was trained independently, allowing the ensemble to leverage diverse modeling assumptions rather than relying on a single predictive structure.

The base model predictions were combined using a stacking approach, in which a Ridge regression meta-model learned the optimal weighting of each learner's output. To prevent information leakage and overfitting, the meta-model was trained on out-of-fold predictions generated via cross-validation on the training set. This design ensures that the final model

balances bias and variance effectively, producing predictions that are more robust and generalizable than those from any individual model alone.

Results & Validation

Model performance was evaluated using a 20% hold-out validation set that was not used during training or stacking. Two primary metrics were considered:

Root Mean Squared Error (RMSE): Measuring average prediction error magnitude

R^2 Score: Measuring the proportion of variance explained relative to a mean-only baseline

The final stacked ensemble achieved:

- **Validation RMSE:** 4.9871
- **Validation R^2 Score:** 0.7016
- **Overfitting Check:** To assess overfitting, training and validation RMSE values were compared. While the training RMSE (2.60) was lower than the validation RMSE, the observed gap is consistent with expectations for ensemble models and does not suggest severe overfitting.