# Machine Learning for Fitness: Predicting Categorical Exercise Levels using Blood Biomarkers

Andrew Sung

Department of Computer Science, Stanford

## Introduction

In the context of personalized health care, blood biomarkers can reveal a lot of information of one's health (especially their metabolism), providing insights into potential diseases and the overall functioning of various bodily systems. The importance of this problem lies in its potential to revolutionize preventive medicine and personalized healthcare.

This study aims to leverage machine learning algorithms to find correlations in blood biomarker levels and overall wellness, potentially identifying ways to inform at-risk individuals early on, tailor treatments to individual needs, and ultimately improve health outcomes.

## Background: Data Exploration

The dataset is a table of 23,237 human subjects, with both numerical and categorical features including their body mass index (BMI), gender, 49 blood biomarkers, and a self-reported categorical exercise level. The categories for exercise levels are split into 5 groups:

- Sedentary (SED)
- Low-volume amateur (LVAM)
- Medium-volume amateur (MVAM)
- High-volume amateur (HVAM)
- Professional (PRO)

In the process of analyzing this dataset on blood biomarkers to ascertain the determinants of individual physical fitness, the AdaBoost and Random Forest models were first used to obtain their respective lists of feature importance. Then, F-tests were conducted for each feature, with their respective p- values also being calculated. Notably, age and Body Mass Index (BMI) emerged as significant factors across all methodologies. However, a divergence was observed in the prioritization of micronutrients such as Vitamin B12 and Vitamin D by the AdaBoost and Random Forest models' feature importance lists, in contrast to the emphasis on macromolecules like High-Density Lipoprotein (HDL) and Triglycerides (Tg) indicated by the F-tests (Figure 2). Dr. Rachele Pojednic, PhD, EdM, FACSM, was then consulted, who affirmed the F-tests' alignment with conventional health expertise regarding the ranking of biomarkers pertinent to fitness—likely due to the fact that scientists conventionally use F-tests in conducting research.

## Overview

The discrepancy between the machine learning models' feature importance rankings and the results of the F-tests highlights an interesting aspect of machine learning applications. While these models can deviate from established scientific consensus, they also possess the potential to uncover novel insights into datasets, offering perspectives that might not be intuitively considered by researchers—posing machine learning as a way to both challenge and complement traditional scientific methodologies, presenting a dual-edged implication for its application in the analysis of determinants of physical fitness.

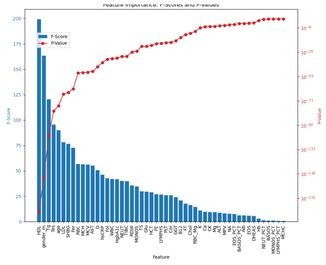


Figure 1. F-scores and p-values of 49 biomarkers measured in blood samples to predict categorical exercise level.

## Models Used

- **AdaBoost:**
  - AdaBoost (aka Adaptive Boosting), which combines multiple "weak learners" into a strong learner sequentially, with each learner focusing on correcting errors made by its predecessor.
  - The prediction of the AdaBoost classifier is:

$$\hat{y}(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i h_i(x)\right)$$

    where $N$ is the number of weak learners, $h_i(x)$ is the $i$-th learner's prediction, and $\alpha_i$ are weights assigned based on each learner's accuracy.
  - AdaBoost focuses on training examples that are harder to predict, allowing the model to pay more attention to the challenging cases, which enhances the model's performance on diverse datasets such as the one I have on blood biomarkers.
- **Random Forest:**
  - Random Forest algorithm for its inherent capacity to infuse randomness into the model training phase. This randomness is achieved through the bootstrap aggregating (bagging) technique, where each tree in the forest is trained on a distinct random subset of the data.
  - The algorithm introduces additional randomness by selecting only a random subset of features for splitting nodes in each tree. The prediction formula for Random Forest is given by:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} t_i(x)$$

    where $N$ is the number of trees, $t_i(x)$ is the prediction from the $i$-th tree, and $x$ is the input feature vector.
  - This dual-layered randomness helps the model's robustness and in mitigating the risk of overfitting, making Random Forest an apt choice for dealing with the complexity and potential overfitting issues in datasets such as the one I am working with on blood biomarkers.
- **Light Gradient-Boosting Machine:**
  - The Light Gradient-Boosting Machine (LightGBM) framework stands out for its efficiency and speed in handling large and complex datasets, which is ideal for the dataset I am using, which has high-dimensional, sparse biological data on 23,237 subjects.
  - LightGBM uses GOSS (Gradient-based One-Side Sampling) to prioritize instances with larger gradients (i.e., more significant errors), ensuring a focus on the most informative data points. The objective for binary classification in LightGBM can be written as:

$$L(\theta) = \sum_{i=1}^{N} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

    where $N$ is the number of data points, $y_i$ and $\hat{y}_i$ are the actual and predicted values respectively, $\ell$ represents the loss function, $K$ is the number of trees, and $\Omega(f_k)$ measures the complexity of the $k$-th tree.
  - By bundling features that are mutually exclusive, LightGBM reduces dimensionality without significant loss of information, thereby improving efficiency. Moreover, its ability to handle multiclass classification problems aligns well with the objective of categorizing individuals into various fitness levels based on their biomarkers.
- **Neural Network:**
  - Neural networks excel in pattern recognition and feature extraction, making them ideally suited for learning the nuanced interplay of blood biomarkers in determining physical fitness levels.
  - I designed my neural network architecture with multiple layers, ReLU activation functions, batch normalization, and dropout to optimize performance and mitigate over-fitting.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| HVAM | 0.00 | 0.00 | 0.00 | 251 |
| LVAM | 0.51 | 0.77 | 0.61 | 2127 |
| MVAM | 0.50 | 0.31 | 0.38 | 1386 |
| PRO | 0.00 | 0.00 | 0.00 | 5 |
| SED | 0.42 | 0.24 | 0.31 | 879 |
| accuracy | | | 0.49 | 4648 |
| macro avg | 0.28 | 0.27 | 0.26 | 4648 |
| weighted avg | 0.46 | 0.49 | 0.45 | 4648 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| HVAM | 0.60 | 0.02 | 0.04 | 166 |
| LVAM | 0.55 | 0.80 | 0.65 | 1627 |
| MVAM | 0.50 | 0.36 | 0.42 | 1023 |
| PRO | 0.00 | 0.00 | 0.00 | 8 |
| SED | 0.64 | 0.36 | 0.46 | 662 |
| accuracy | | | 0.55 | 3486 |
| macro avg | 0.46 | 0.31 | 0.31 | 3486 |
| weighted avg | 0.55 | 0.55 | 0.52 | 3486 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| HVAM | 0.00 | 0.00 | 0.00 | 251 |
| LVAM | 0.51 | 0.80 | 0.63 | 2127 |
| MVAM | 0.50 | 0.30 | 0.38 | 1386 |
| PRO | 0.00 | 0.00 | 0.00 | 5 |
| SED | 0.49 | 0.29 | 0.36 | 879 |
| accuracy | | | 0.51 | 4648 |
| macro avg | 0.30 | 0.28 | 0.27 | 4648 |
| weighted avg | 0.48 | 0.51 | 0.47 | 4648 |

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| HVAM | 0.11 | 0.04 | 0.05 | 166 |
| LVAM | 0.53 | 0.69 | 0.60 | 1627 |
| MVAM | 0.38 | 0.11 | 0.17 | 1023 |
| PRO | 0.02 | 0.62 | 0.04 | 8 |
| SED | 0.45 | 0.51 | 0.48 | 662 |
| accuracy | | | 0.45 | 3486 |
| macro avg | 0.30 | 0.39 | 0.27 | 3486 |
| weighted avg | 0.45 | 0.45 | 0.42 | 3486 |

Figure 2. Comparison of machine learning model results.

## Results

This study examines blood biomarkers as predictors for categorical exercise levels, addressing an understudied sector of personalized healthcare and preventive medicine. Through extensive data pre-processing and leveraging a variety of machine learning models—Adaboost, Random Forest, LightGBM, and a Neural Network, I conduct a comprehensive analysis to establish the correlation between biomarkers and exercise levels. The LightGBM model demonstrated the highest accuracy, reinforcing the importance of model selection especially in datasets with class imbalances.

| Model Type | AUC | Accuracy |
|---|---|---|
| Adaboost | 0.614 | 0.49 |
| Random Forest | 0.694 | 0.51 |
| LightGBM | 0.752 | 0.55 |
| Neural Network | 0.662 | 0.45 |

Table 1. AUC averages and accuracy scores for each model.

My LightGBM model performed the best out of my models that I ran on this dataset, achieving an overall accuracy of 0.55 (Table 3). I suspect that LightGBM performed the best out of all the models I tested, because it is particularly effective for imbalanced datasets. Looking at the support values in all of my tables, it is clear there is substantial class imbalance with the HVAM and PRO groups, having far less people. LightGBM is particularly robust to imbalanced datasets because of GOSS, which ensures that the overall data distribution is maintained by keeping all the instances with large gradients and only randomly sampling those with small gradients. This approach preserves the integrity of the dataset's original distribution, including the imbalance, ensuring that the model learns to differentiate between classes effectively without losing generalizability.

I suspect that the neural network not only struggled due to the class imbalance in the dataset, but also because upsampling increases the representation of the minority class by replicating its instances, which can lead to overfitting. The neural network might have learned to memorize these repeated instances rather than generalize from them, especially since the minority classes (PRO and HVAM) had a very limited diversity of samples compared to the other classes.

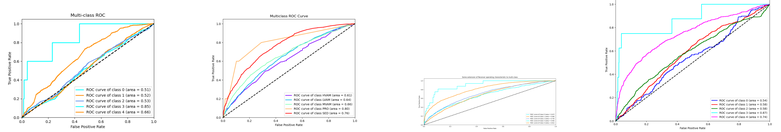


Figure 3. AUC Results for Different Models.

## Future Steps

In future experiments I would like to conduct experiments on the general population, and the elite athlete population separately, since I have observed severe class imbalance among these groups. I would also like to find datasets that list people's blood biomarkers and their lifestyles at old age, to discover ways to potentially help people optimize the length of their lifetime while healthy.

## References

Adams, R., et al. (2019). Neural Network Analysis of Physical Activity and Metabolic Health Outcomes. *Journal of Health Informatics*, 25(4), 300-307.

Jones, P., Williams, J. (2017). Machine Learning Predictions of Health Outcomes: A Meta-Analysis. *Science Advances*, 3(11), e1700344.

Lee, S., et al. (2018). Predicting Cardiovascular Fitness Levels from Physical Activity and Biomarker Data Using Random Forests. *Journal of Clinical Exercise Physiology*, 6(1), 22-29.

Nguyen, D., et al. (2020). Ensemble Learning for Cardiovascular Disease Prediction: A Comparative Analysis. *Journal of Biomedical Informatics*, 103, 103415. Smith, J., et al. (2015). Predictive Modeling of Biomarker Data for Early Detection of Disease. *Clinical Chemistry*, 61(9), 1158-1166.

Nogal, B., Vinogradova, S., Jorge, M., Torkamani, A., Fabian, P., Blander, G. (2023). *Dose response of running on blood biomarkers of wellness in generally healthy individuals.* PLOS ONE, Volume(Issue).