# Personalized Workout Recommendation Using Machine Learning Classification Techniques

Melanie Duong, Johnny Nguyen, Charlene Vu

DATA 300: Statistical and Machine Learning

## INTRODUCTION

In an era where fitness and wellness are becoming increasingly personalized, there is a growing demand for data-driven tools that can tailor workout plans to individual needs. This project explores the use of machine learning to build a personalized fitness recommendation system. By analyzing users' health indicators—such as age, BMI, hypertension, diabetes status, and fitness goals—we aim to predict the most suitable workout regimen for each individual. Our dataset includes various user profiles and corresponding recommended workout classes, allowing us to train and evaluate multiple classification models for accuracy and effectiveness.

## OBJECTIVES

The primary objective of this project is to develop a machine learning model that can accurately recommend personalized workout plans based on a user's health and demographic data. Specifically, we aim to:

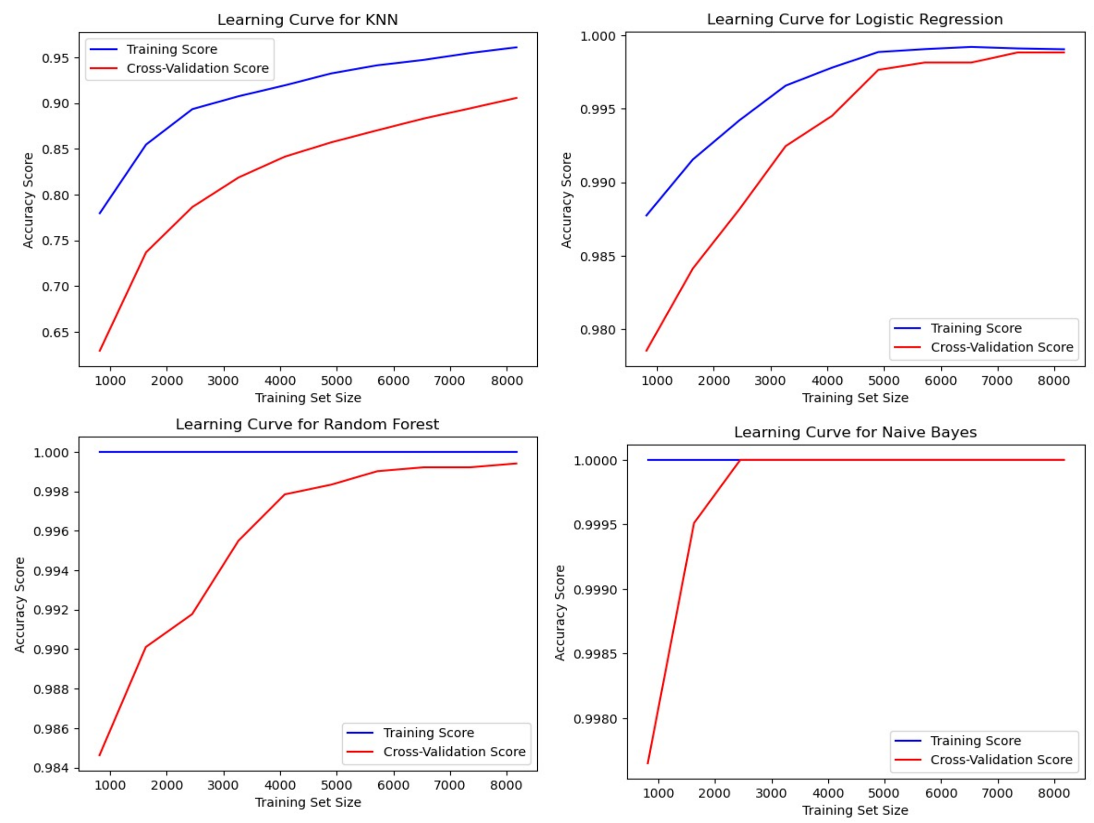
- Classify users into appropriate workout categories using predictive modeling.
- Compare the performance of multiple classification algorithms (e.g., Logistic Regression, Random Forest) to determine the most effective approach.
- Create an interpretable, scalable recommendation system that could be integrated into real-world fitness or health applications.

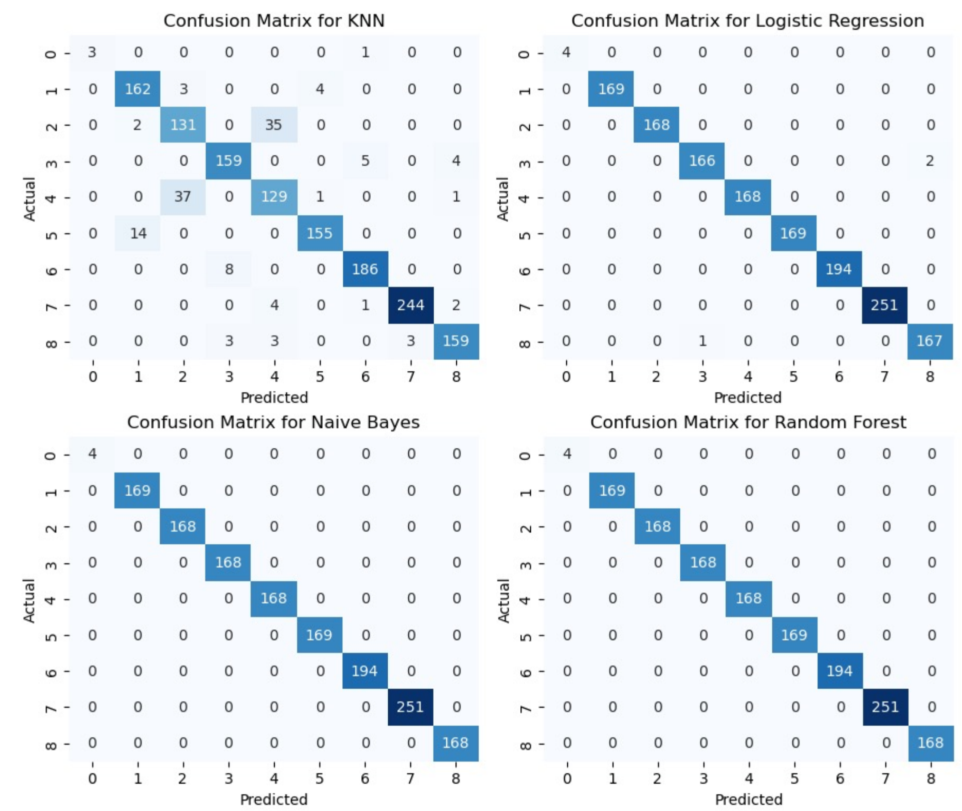
## DATA AND METHODOLOGY

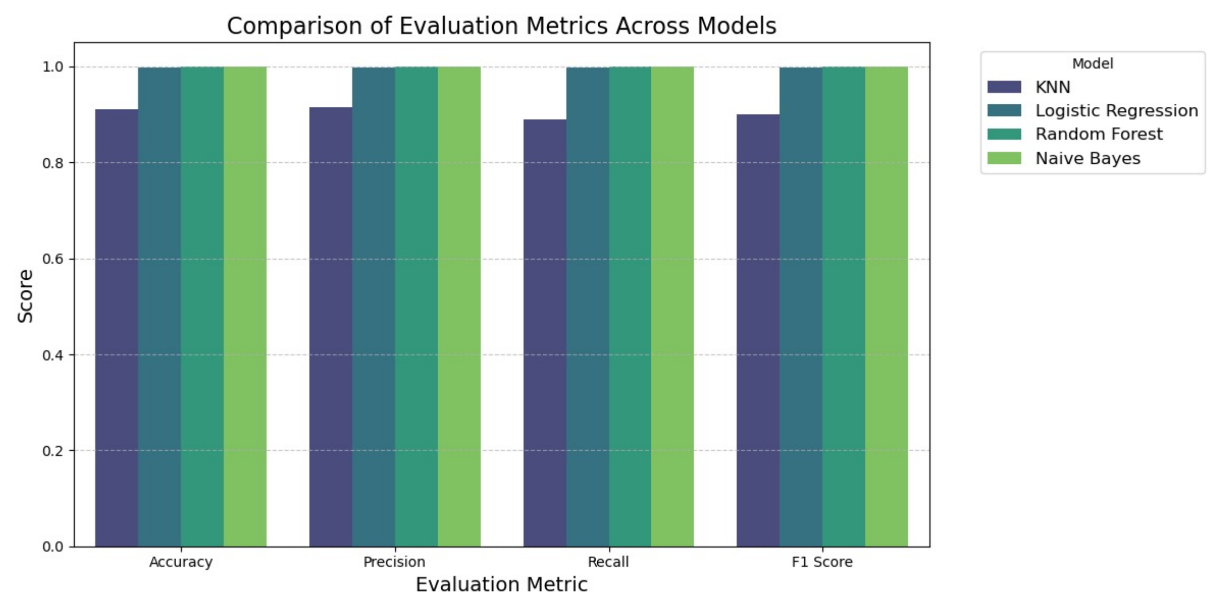
We built a machine learning classifier to generate personalized workout recommendations based on individual traits and fitness goals. We used a dataset of 14,589 entries and 15 variables from Mendeley Data, including demographics, physical metrics, and fitness-related info.

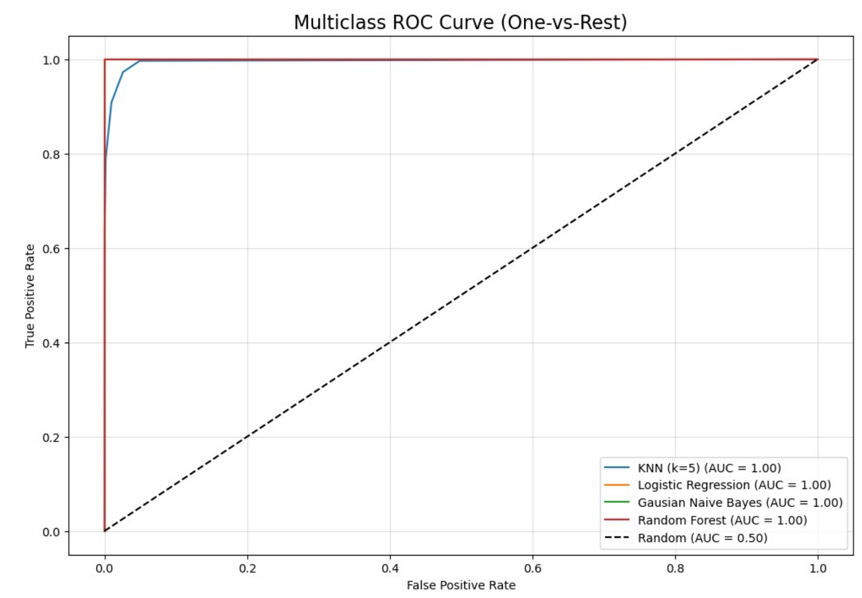
Since the dataset had no missing values, preprocessing was streamlined. We dropped the non-informative ID column, then label-encoded 10 categorical variables (e.g., Sex, Fitness Goal, Exercise Type) to make them model-ready. We skipped dimensionality reduction since the feature set was manageable. We separated the target variable, scaled the data to normalize feature ranges, and split it into training (70%), validation (20%), and testing (10%) sets.

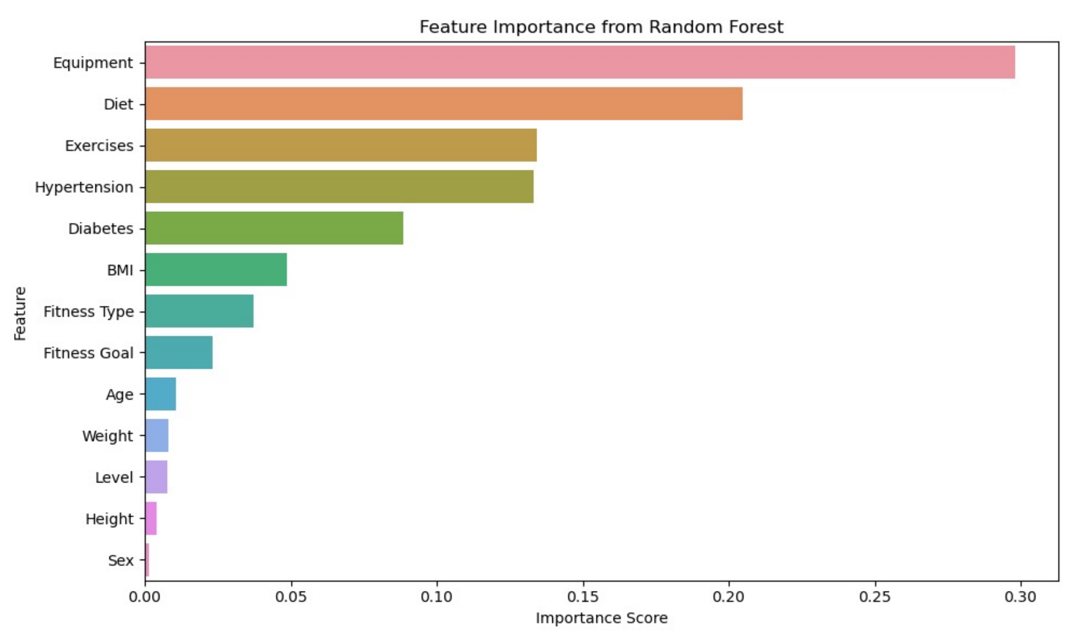
- **K-Nearest Neighbors (KNN):** Groups users based on distance; optimized k via cross-validation.
- **Logistic Regression:** Estimates probability for each workout class; coefficients used to interpret feature influence.
- **Naïve Bayes:** Fast, probabilistic model assuming independent features; suited for categorical targets.
- **Random Forest:** Ensemble of decision trees offering robust prediction and feature importance insights.

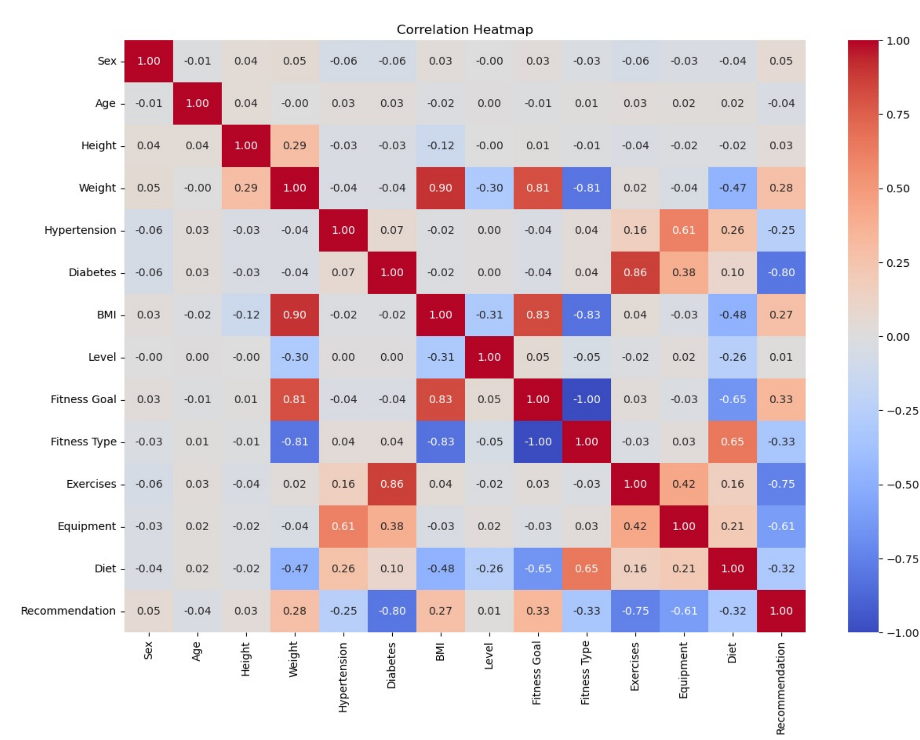



## DISCUSSION

- **Random Forest**:
  The Random Forest model achieved perfect accuracy, with "Equipment" being the most influential feature. It prioritized behavioral and health-related factors like diet and exercise over demographics. Feature correlation analysis led to the removal of "Fitness Goal" due to redundancy. The model demonstrates excellent performance and interpretability, making it well-suited for personalized recommendations.
- **Multinomial Logistic Regression**:
  This model delivered near-perfect performance, with validation and testing accuracy close to 100%. Its interpretability and probabilistic outputs make it ideal for multiclass classification. However, it assumes linear relationships and lacks built-in nonlinear modeling. While highly effective here, its success may reflect ideal data conditions rather than robust generalizability to noisier real-world environments.
- **K-Nearest Neighbors (KNN)**:
  KNN performed well with high accuracy and F1-scores, particularly at k=1. Its logic fits the recommendation task, grouping users by similarity. However, it's computationally expensive and sensitive to outliers or scale, limiting scalability. While ideal for benchmarking and exploratory analysis, KNN may not suit real-time systems without further optimization.
- **Gaussian Naive Bayes**:
  Achieved flawless accuracy on validation and test sets, likely due to the clean, highly separable dataset. It suits continuous data assuming normal distribution and independence between features. While efficient and easy to implement, these ideal results raise concerns about overfitting or artificial simplicity, making real-world generalization uncertain without further validation.

## CONCLUSION

The models performed nearly perfectly, raising concerns about the dataset being overly clean or easy to classify. While the results are exciting, they likely won't hold in real-world settings where data is messy and imperfect. To assess true reliability, the next step is testing on a more realistic, challenging dataset.

## REFERENCE

- Tavares, B. F., Pires, I. M., Marques, G., Garcia, N.M., Zdravevski, E., Lameski, P., Trajkovik, V., & Jevremovic, A. (2020). Mobile Applications for Training Plan Using Android Devices: A Systematic Review and a Taxonomy Proposal. Information, 11(7), 343.
- Alian, Shadi, Juan Li, and Vikram Pandey. "A Personalized Recommendation System to Support Diabetes Self-Management for American Indians." IEEE Access, vol. 6, 2018, pp. 73041–73051.