Lifestyle Data Correlation Analysis & Calorie Burn Prediction

Group Members:

Anthony Henderson – anhenderson@ucsd.edu

Shunkai Yu – shyu@ucsd.edu

Riqian Hu – rih006@ucsd.edu

Yuzhou Ren – yur004@ucsd.edu

Huayang Yu – huy016@ucsd.edu

Outline

1. Motivation
2. Dataset Overview
3. Exploratory Data Visualization and Analysis
4. Putting a Predictive Model to the Test
5. Conclusion

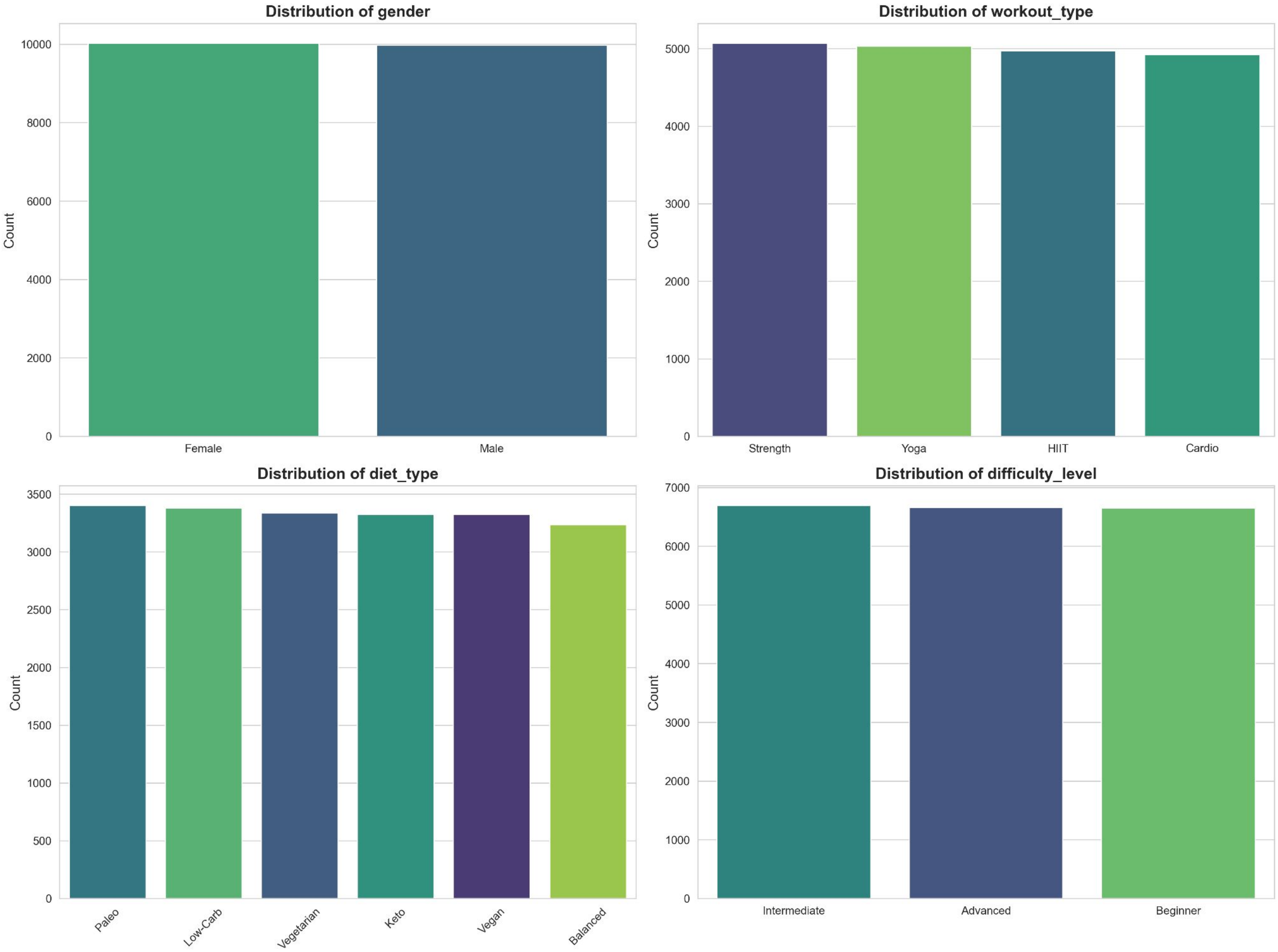
Motivation

- People track workouts and diet but often don't know which factors actually drive calorie burn.
- Large lifestyle datasets are available (i.e. Kaggle) but need proper analysis.
- Overarching Questions:
  - How do daily lifestyle habits influence health and calorie burn?
    - Which workout types burn the most calories?
    - Are calories burned more about behavior or demographics/intake?
    - Do different diet types lead to different mood ratings?
  - Can we accurately predict calorie-burn categories using machine learning?
    - Which features does the model rely on the most to predict calorie-burn categories?

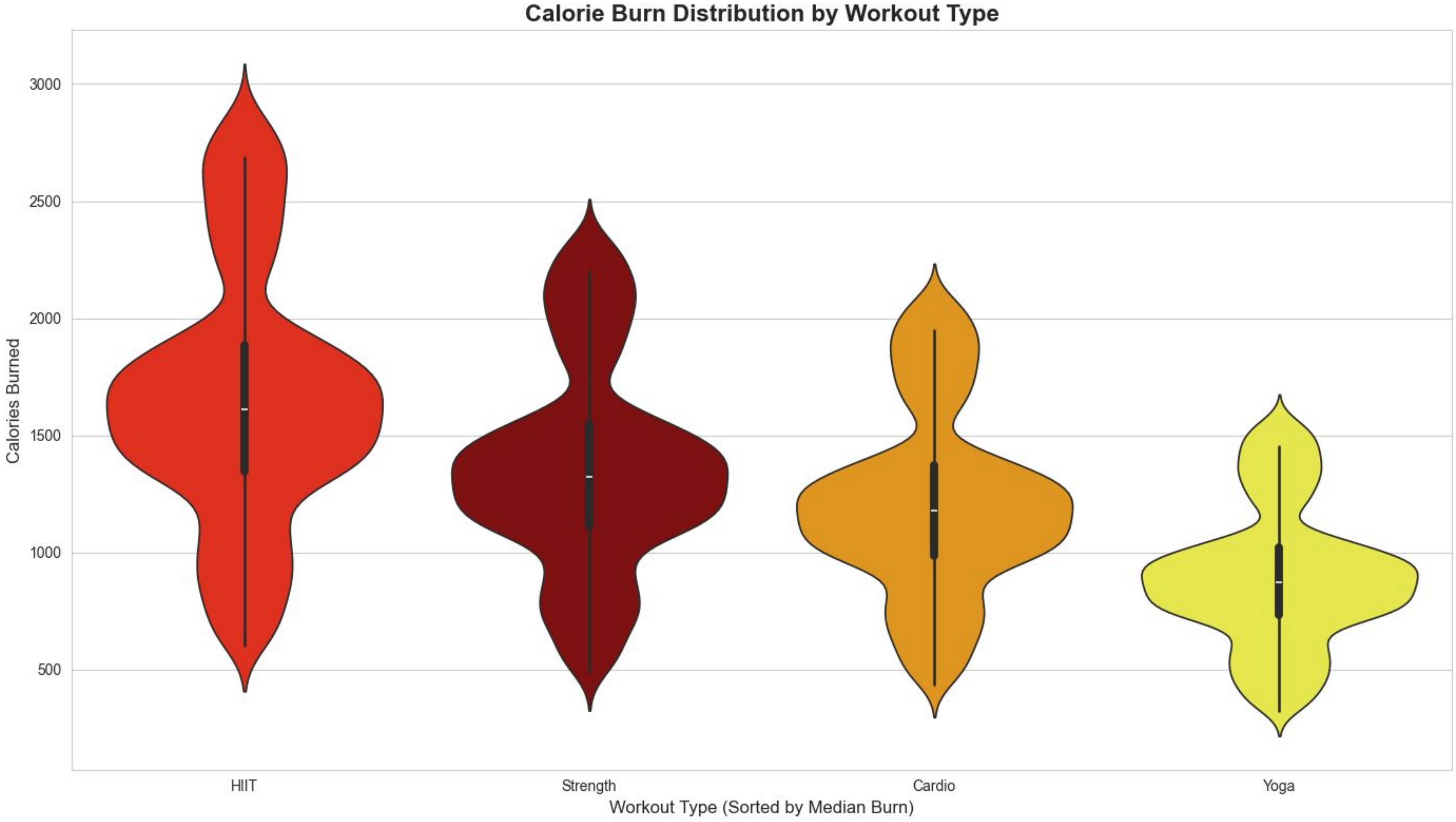
Dataset Overview

- Data Source Utilized: Kaggle – Lifestyle Data Analysis and Prediction (20,000 lifestyle and workout records)
- Main fields (examples):
  - Demographics: age, gender, maybe BMI if available.
  - Workout behavior: workout type, duration, intensity, frequency.
  - Diet: diet type, meal timing / composition.
  - Target variable: calorie burn category (Low / Medium / High / Very High).
- Process Flow:
  - Data Collection → Data Standardization → Data Cleaning → Data Analysis → Visualization

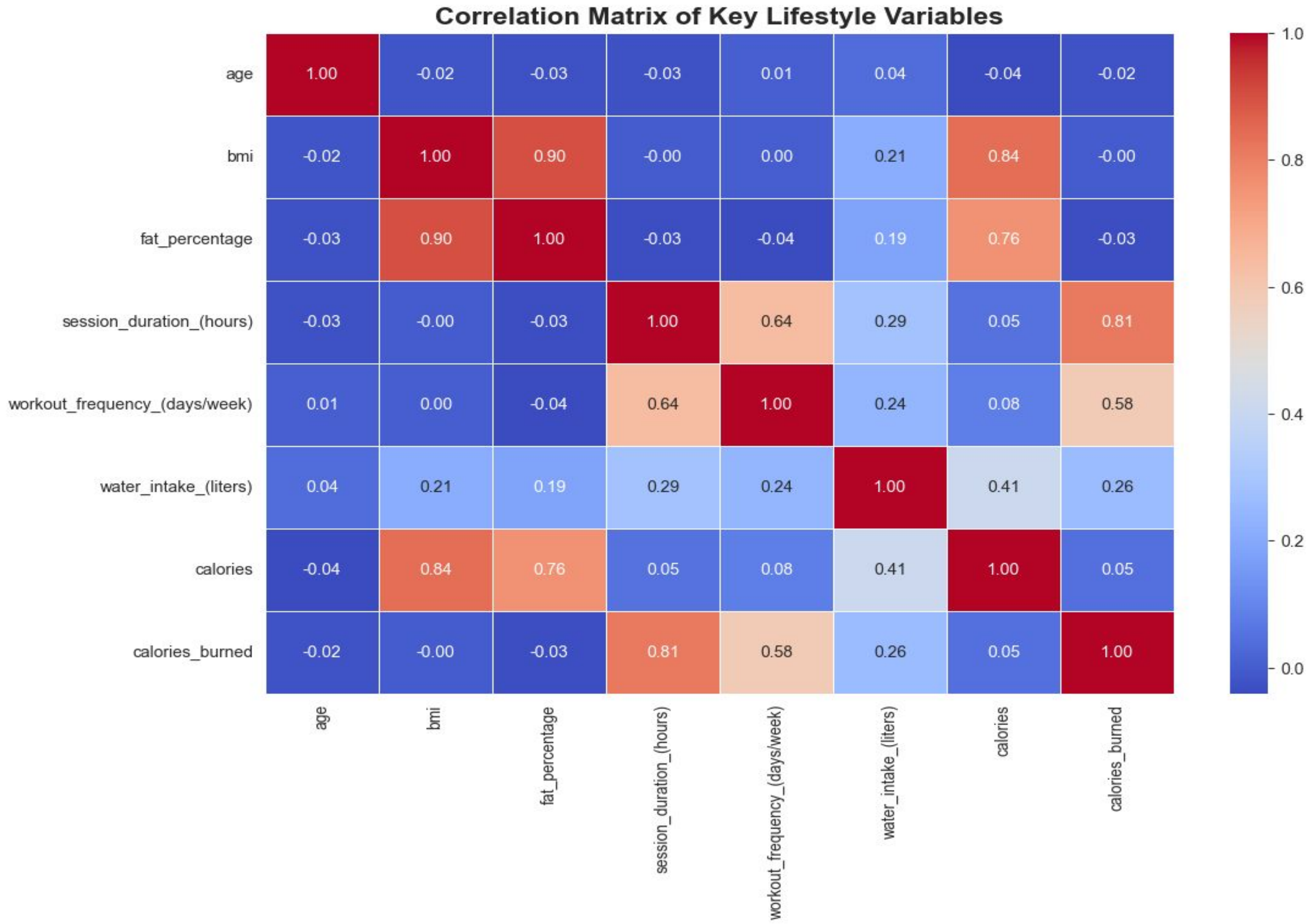
Important Categorical Distributions to Consider


Distribution of gender
Count
10000
8000
6000
4000
2000
0
Female
Male
Distribution of workout_type
Count
5000
4000
3000
2000
1000
0
Strength
Yoga
HIIT
Cardio
Distribution of diet_type
Count
3500
3000
2500
2000
1500
1000
500
0
Paleo
Low-Carb
Vegetarian
Keto
Vegan
Balanced
Distribution of difficulty_level
Count
7000
6000
5000
4000
3000
2000
1000
0
Intermediate
Advanced
Beginner

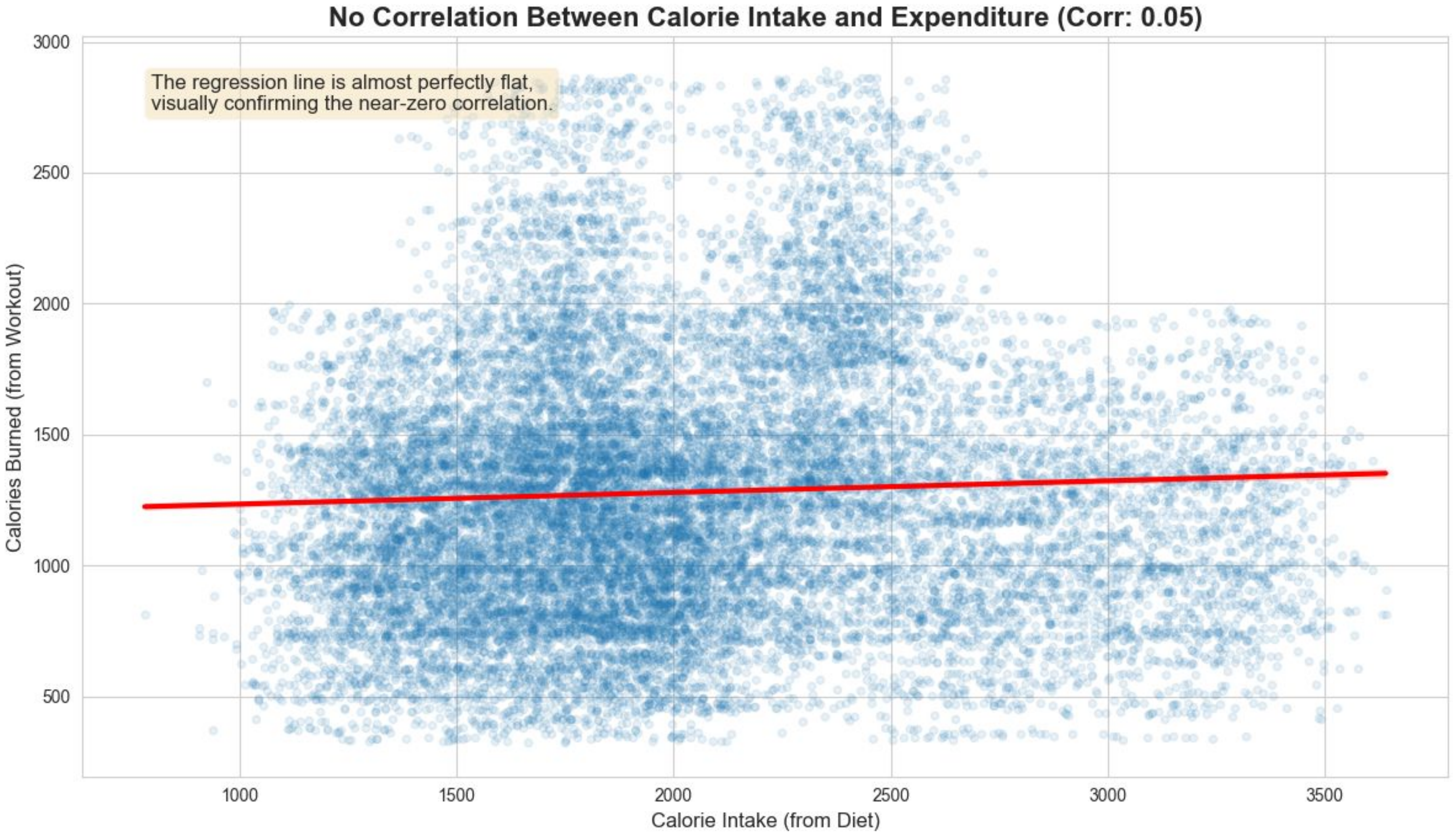
Which Workout Type Burns the Most Calories?


Calorie Burn Distribution by Workout Type
Calories Burned
3000
2500
2000
1500
1000
500
HIIT
Strength
Cardio
Yoga
Workout Type (Sorted by Median Burn)

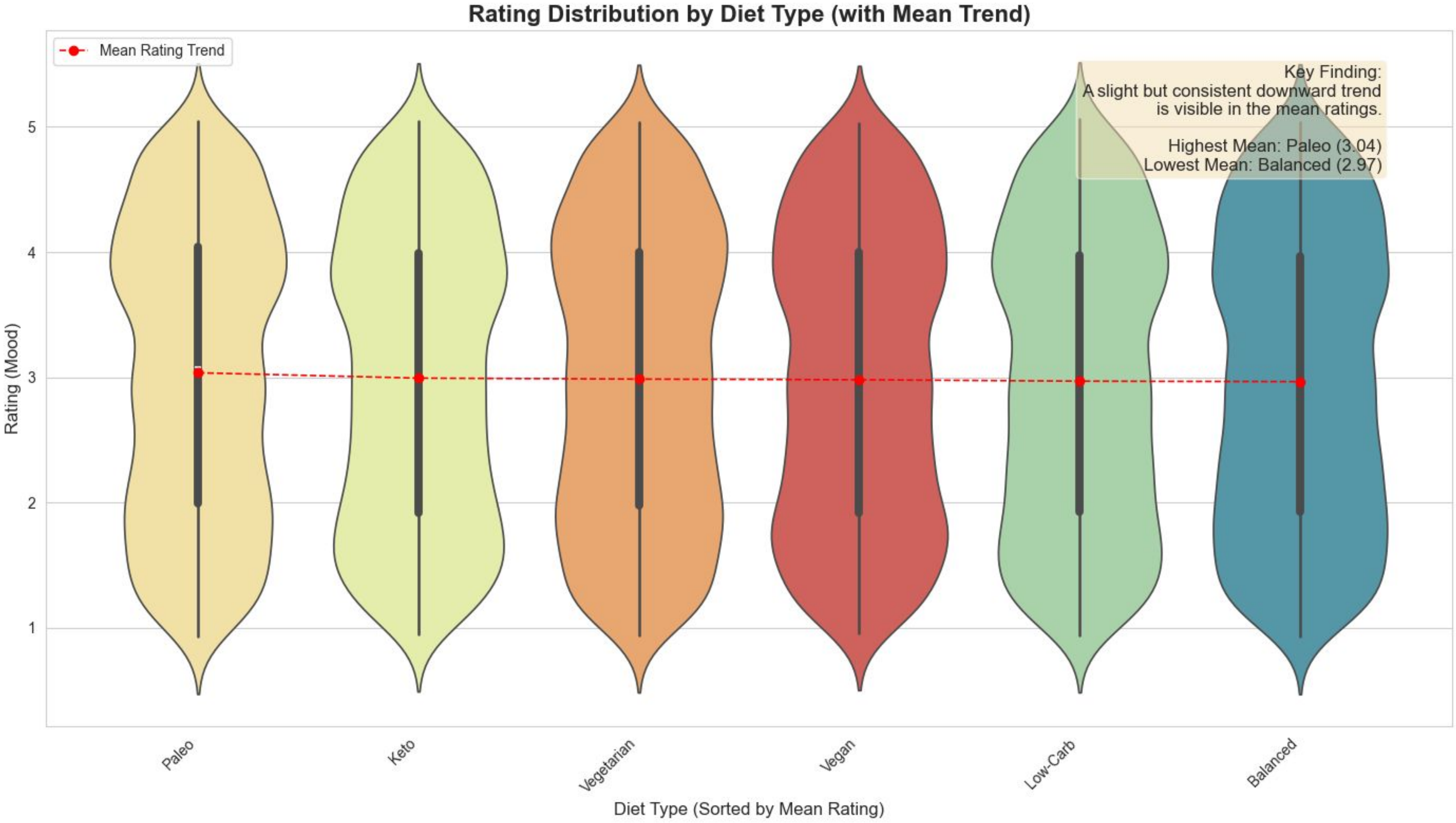
Which Key Lifestyle Habits and Health Outcomes are Related?


Correlation Matrix of Key Lifestyle Variables
age
bmi
fat_percentage
session_duration_(hours)
workout_frequency_(days/week)
water_intake_(liters)
calories
calories_burned
1.00 -0.02 -0.03 -0.03 0.01 0.04 -0.04 -0.02
-0.02 1.00 0.90 -0.00 0.00 0.21 0.84 -0.00
-0.03 0.90 1.00 -0.03 -0.04 0.19 0.76 -0.03
-0.03 -0.00 -0.03 1.00 0.64 0.29 0.05 0.81
0.01 0.00 -0.04 0.64 1.00 0.24 0.08 0.58
0.04 0.21 0.19 0.29 0.24 1.00 0.41 0.26
-0.04 0.84 0.76 0.05 0.08 0.41 1.00 0.05
-0.02 -0.00 -0.03 0.81 0.58 0.26 0.05 1.00
1.0
0.8
0.6
0.4
0.2
0.0

How are Dietary Habits and Exercise Habits Related?


No Correlation Between Calorie Intake and Expenditure (Corr: 0.05)
The regression line is almost perfectly flat,
visually confirming the near-zero correlation.
Calories Burned (from Workout)
Calorie Intake (from Diet)
3000
2500
2000
1500
1000
500
1000
1500
2000
2500
3000
3500

How Does Dieting Affect Mood?

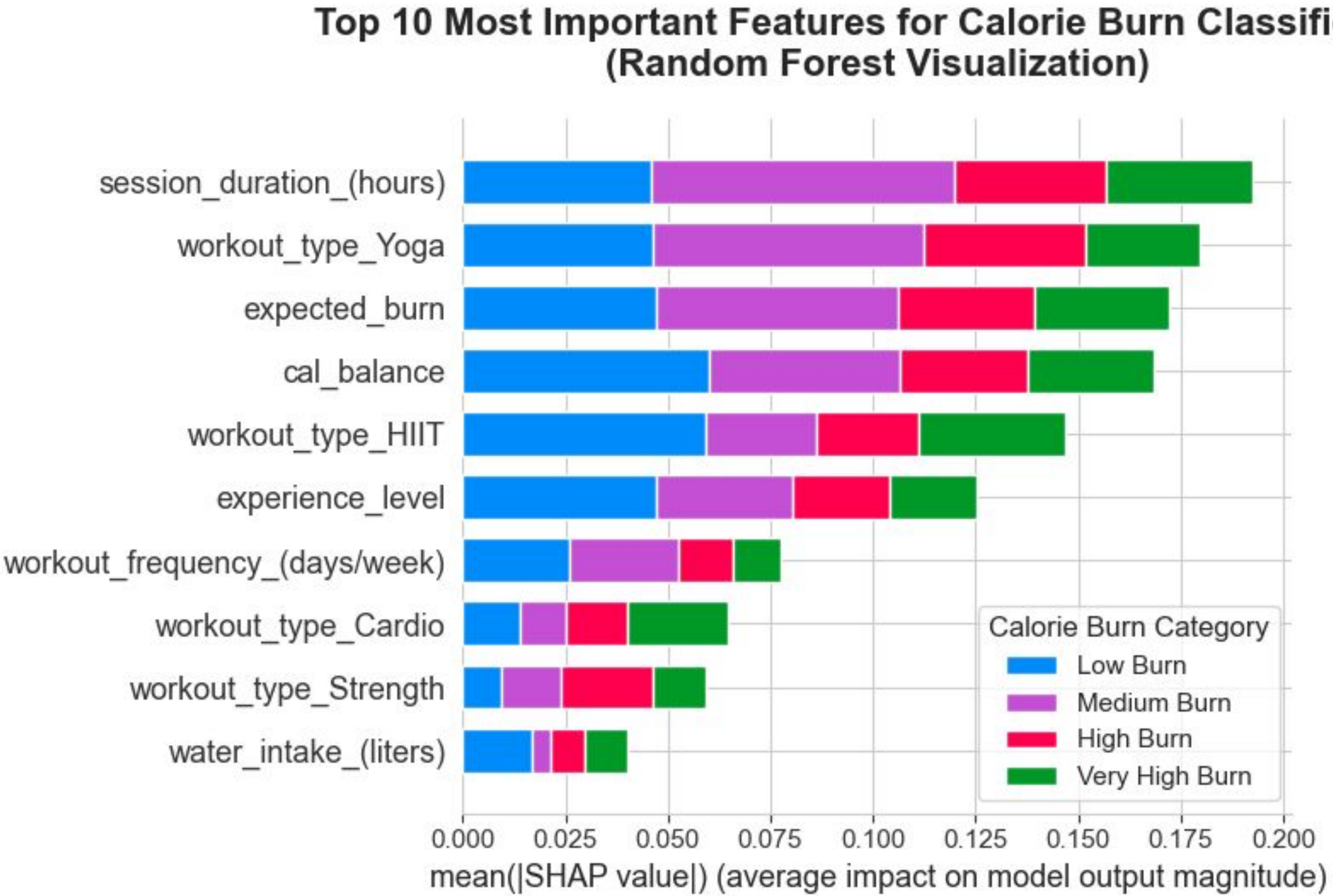

Rating Distribution by Diet Type (with Mean Trend)
Mean Rating Trend
Key Finding:
A slight but consistent downward trend
is visible in the mean ratings.
Highest Mean: Paleo (3.04)
Lowest Mean: Balanced (2.97)
Rating (Mood)
1
2
3
4
5
Paleo
Keto
Vegetarian
Vegan
Low-Carb
Balanced
Diet Type (Sorted by Mean Rating)

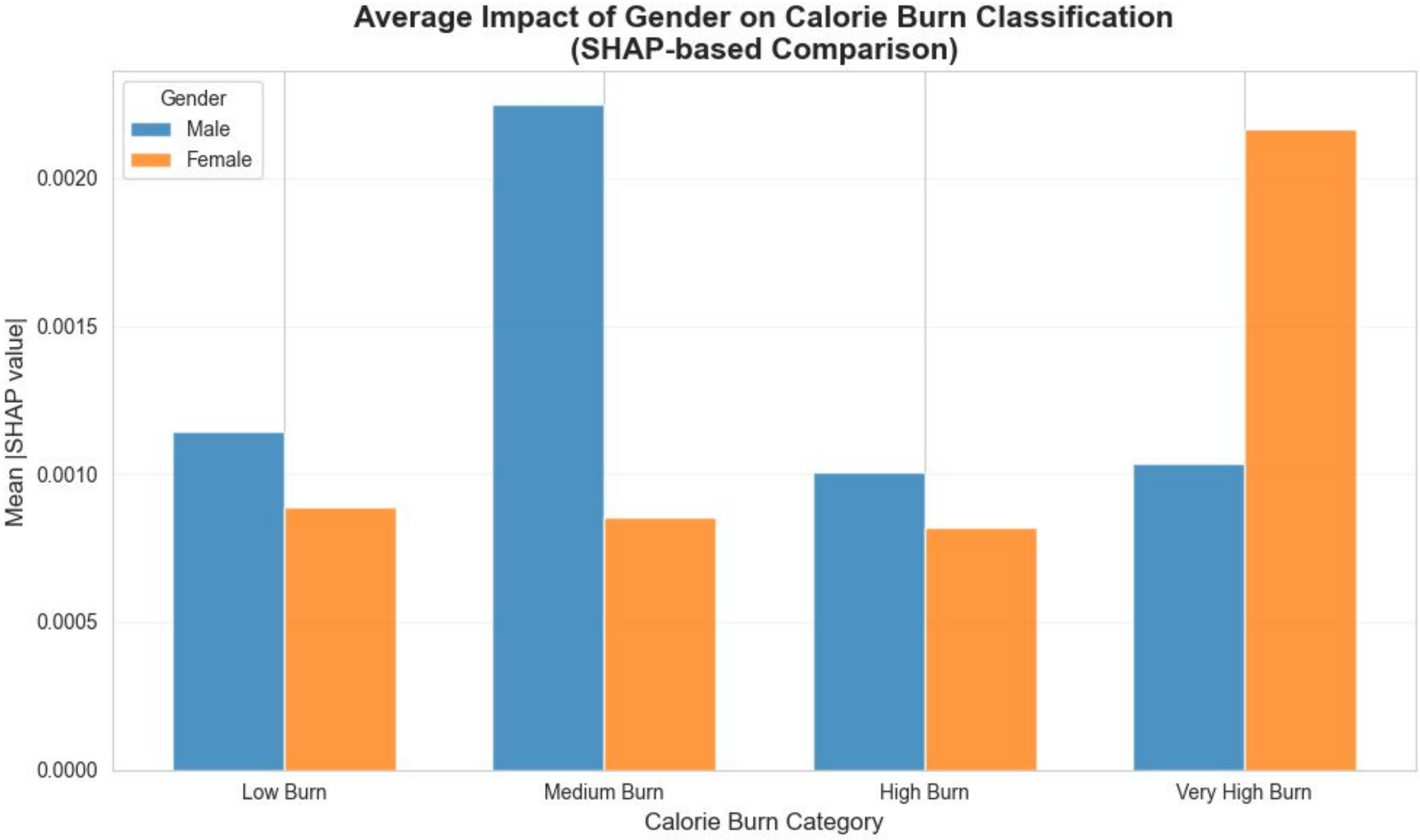
From Exploration to Prediction

- Our EDA showed that workout behavior (type, duration, frequency) is strongly related to calories burned, while demographics have little effect.
- We also saw almost no relationship between short-term calorie intake and workout calories.
- This suggests that we can reasonably predict calorie-burn level from lifestyle and workout features.
- Next, we formalize this as a classification problem and train a Random Forest model to make those predictions.

Random Forest Model and SHAP

- We use a Random Forest classifier to predict four calorie-burn levels (Low / Medium / High / Very High) from lifestyle and workout features, using a sklearn pipeline with encoded categorical and numeric inputs.
- Random Forests are ensembles of decision trees that capture non-linear patterns (e.g., long + high-intensity workouts → High burn) and work well on mixed data.
- We then apply SHAP (SHapley Additive exPlanations) to measure how much each feature contributes to the model's predictions.
- SHAP shows that the model relies mainly on controllable behaviors (duration, workout type, difficulty, frequency) and much less on demographics like gender.


Top 10 Most Important Features for Calorie Burn Classification
(Random Forest Visualization)
session_duration_(hours)
workout_type_Yoga
expected_burn
cal_balance
workout_type_HIIT
experience_level
workout_frequency_(days/week)
workout_type_Cardio
workout_type_Strength
water_intake_(liters)
Calorie Burn Category
Low Burn
Medium Burn
High Burn
Very High Burn
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
mean(|SHAP value|) (average impact on model output magnitude)


Average Impact of Gender on Calorie Burn Classification
(SHAP-based Comparison)
Gender
Male
Female
Mean |SHAP value|
0.0020
0.0015
0.0010
0.0005
0.0000
Low Burn
Medium Burn
High Burn
Very High Burn
Calorie Burn Category

Conclusion

1. How do daily lifestyle habits influence health and calorie burn?
Daily habits like session duration, workout intensity, workout type, and how often people exercise have a strong impact on calories burned, while demographics and short-term intake play a much smaller role.

2. Which workout types burn the most calories?
HIIT and strength sessions have the highest typical calorie burn, cardio is in the middle, and yoga consistently burns the least per session.

3. Are calories burned more about behavior or demographics/intake?
Calories burned are primarily about behavior—how long, how hard, and what kind of workouts people do. Age, gender, and even daily calorie intake show weak or near-zero relationships with exercise calories.

Conclusion-2

4. Do different diet types lead to different mood ratings?
 Different diet types have very similar average mood ratings (all around 3/5). Any differences are small, suggesting that mood in this dataset is only weakly related to declared diet patterns.

5. Can we accurately predict calorie-burn categories using machine learning?
 Yes. A Random Forest classifier can predict Low/Medium/High/Very High burn categories with good overall accuracy, mainly confusing neighboring classes (e.g., Medium vs High).

6. Which features does the model rely on the most to predict calorie-burn categories?
 The model relies most on controllable lifestyle features: session duration, workout type, workout difficulty/experience, expected burn, and frequency. Gender and other demographics have much smaller importance, which is encouraging from a fairness perspective.

Conclusion-3

Calorie burn depends on what you do, not who you are. Behavioral decisions like duration of workout, intensity, and consistency dominate over biology.

- What we can take away from our findings:
    - Focus on longer sessions, higher intensity, and training consistency.
    - Don't overestimate gender or body type differences
        1. Anyone can increase calorie burn with effective workout behaviors.
        2. We are NOT limited by genetics

Thank You !