Integrated Analysis of Fitness Performance and Income Demographics

Instructor: Professor Joseph Taylor

MSBA 500: Culminating Experience and Advanced Topics in Business Analytics

Author: Selena Buzinky

Report: Integrated Analysis of Fitness Performance and Income Demographics

Author: Selena Buzinky

Abstract

This project examines how physical fitness behaviors and socioeconomic factors interact by integrating two datasets: the Gym Members Exercise Tracking (Gym) dataset and the Income dataset. The analysis seeks to understand how workout and demographic attributes influence both calories burned during exercise and the likelihood of earning above or below $50K annually. By combining fitness metrics and income indicators into a consolidated framework, the project highlights lifestyle and career trends that can impact long-term health and financial well-being.

Multiple tools and analytical techniques were used, including MySQL for data processing, Python for machine learning modeling, and Tableau for visualization. Age buckets and gender segmentation were applied consistently across datasets to enable direct comparison of trends. Additional calculated measures such as heart rate intensity percentage, calories burned per kilogram of body weight, and income percentages further enhanced interpretability.

The Gym dataset analysis showed that key behavioral predictors, including workout frequency, experience level, hydration, and cardiovascular effort, strongly influenced calories burned. Conversely, higher body fat percentage was negatively associated with performance, reinforcing the importance of healthy body composition. The Income dataset results revealed that education level and weekly work hours were among the strongest predictors of earning greater than $50K, along with employment in private or self-employed roles.

Modeling methods such as Multiple Linear Regression, Logistic Regression, and Decision Tree–based ensemble approaches were applied to assess predictive relationships. Machine learning performance metrics demonstrated that both fitness and income outcomes can be partially predicted by demographic and behavioral factors, with decision tree models identifying the most important predictors of calories burned and income.

Findings from this integrated analysis highlight modifiable actions that support improved health and financial success. Increasing exercise frequency and intensity, staying hydrated, reducing excess body fat, pursuing continued education, increasing hours worked per week, and engaging in career pathways with greater earning potential are among the most effective strategies indicated by the data. Overall, this research demonstrates that while factors such as age and gender cannot be changed, individuals have meaningful control over many predictors that shape both physical performance and income outcomes.

Introduction

The relationship between physical health and financial well-being is an increasingly important area of interest in data analytics. Understanding how exercise habits vary across demographic groups and how those trends relate to income can offer meaningful insight into lifestyle behaviors, access to resources, and long-term well-being. This project integrates information from a fitness-focused dataset and a socioeconomic dataset to explore how demographic characteristics relate to both health outcomes and financial outcomes.

Two datasets are used for this analysis: the Gym Members Exercise Tracking (Gym) dataset and the Income dataset. The Gym dataset provides detailed workout information such as calories burned, heart rate, hydration, and experience level, while the Income dataset contains demographic and employment attributes including education level, work class, and hours worked per week. Calories Burned is the primary outcome of interest in the Gym dataset, while Income (categorized as earning ≤50K or >50K) is the outcome variable in the Income dataset.

This study applies a combination of SQL-based data transformations, Python statistical modeling, and Tableau visualization techniques to analyze trends and uncover predictive relationships. Key steps include grouping both datasets into common age and gender segments, calculating a variety of metrics, and applying machine learning regression and classification models. The goal of this integrated approach is to identify actionable and data-driven insights that individuals can use to improve both their fitness performance and financial success.

Methods

Two datasets are used in this project: the Gym Members Exercise Tracking (Gym) dataset and the Income dataset. The data used in this project is downloaded in CSV format from Kaggle. Kaggle is an online platform that hosts a wide variety of public datasets and provides a collaborative environment for learning, exploring, and practicing data science and machine learning techniques. The chosen Outcome variables used in this analysis are Calories Burned for the Gym dataset and Income for the Income dataset. The analysis will focus on metrics, machine learning regression and other methods, charts, and dashboards that involve the Calories Burned and Income fields.

The first set of data transformations is in MySQL using MySQL Workbench. The Gym and Income datasets are loaded as tables into a MySQL schema. An Age Bucket field is added to both tables with 4 categories: 18-25, 26-35, 36-50, and 50+. Relevant metrics grouped by Age Bucket and Gender from both tables are calculated using MySQL. In addition, the Correlation between Calories Burned and Body Mass Index (BMI) and Correlation between Calories Burned and Fat Percentage metrics are calculated using Python.

Next, data transformations are done to create the Summary Table that includes calculations from both the Gym and Income datasets. The two tables are joined on Age Bucket and Gender, and then averages and percentages of other fields are calculated using MySQL. The Summary Table is grouped by Age Bucket and Gender. The Summary Table is then exported as a CSV in preparation for loading into Tableau.

The next set of data transformations is in Python and is comprised of Multiple Linear Regression, Logistic Regression, and Decision Tree analysis for both the Gym and Income datasets. First, the Gym dataset is loaded into a Jupyter Notebook. Numerical predictors are selected in order to create a Correlation Matrix. The Session Duration field was removed from this machine learning analysis because it is too obvious of a predictor for Calories Burned and it obscured the effect of the other predictor variables. The Outcome variable is Calories Burned. It is a numeric field, making this dataset suitable for Multiple Linear Regression and Decision Tree Regression analysis. Next, all predictors are selected and dummy variables are created for the categorical variables. The dataset is split into 60% training and 40% validation datasets. A Multiple Linear Regression model is then trained on the training dataset. The coefficients are printed and accuracy measures computed using the validation dataset. The primary accuracy measures examined are Root Mean Squared Error and R^2.

Next, Decision Tree Regression models are trained on the training dataset. First, a single tree is trained using Decision Tree Regressor and Root Mean Squared Error and R^2 accuracy measures are calculated. A Bagging Regressor is then trained. GridSearchCV is

used to find the optimal hyperparamters for the Bagging Regressor. Accuracy measures are calculated. The Bagging Regressor is made up of multiple decision trees, and each feature's importance is calculated by averaging the importance scores across all trees in the ensemble. The top 15 most important features are selected and plotted. Then, a Random Forest Regressor and XGBoost Regressor are trained and evaluated using the same steps as the Bagging Regressor.

The Income dataset is then loaded into the Jupyter Notebook to be analyzed with Python. The Outcome variable is Income. It is a binary variable (<=50K or >50K), making this dataset suitable for Logistic Regression and Decision Tree Classification analysis. Logistic Regression is well suited for a binary outcome variable because it models the probability of an event occurring using the logistic (sigmoid) function, which naturally outputs values between 0 and 1. All predictors are selected and dummy variables are created for the categorical variables. The dataset is split into 60% training and 40% validation datasets and the features are standardized. A Logistic Regression model is then trained on the training dataset. The coefficients are printed and accuracy measures computed using the validation dataset. The primary accuracy measures examined are the Accuracy Score and the Confusion Matrix.

Next, Decision Tree Classification models are trained on the training dataset. First, a single tree is trained using Decision Tree Classifier and Accuracy Score and Confusion Matrix accuracy measures are calculated. A Bagging Classifier is then trained. GridSearchCV is used to find the optimal hyperparamters for the Bagging Classifier. Accuracy measures are calculated. The Bagging Classifier is made up of multiple decision trees, and each feature's importance is calculated by averaging the importance scores across all trees in the ensemble. The top 15 most important features are selected and plotted. Then, a Random Forest Classifier and XGBoost Classifier are trained and evaluated using the same steps as the Bagging Classifier.

Then, analysis is performed in Tableau. The Gym, Income, and Summary Table datasets are loaded as CSVs into 3 separate Tableau Packaged Workbooks. Calculated fields are created, charts are created, dashboards are built, dashboard-wide filters are added, and data validation is performed. For the Gym dataset, the Health and Exercise Metrics dashboard is created. For the Income dataset, the Income Metrics dashboard is created. For the Summary Table dataset, the Exercise and Income Summary Metrics dashboard and Exercise and Income Summary Bubble Charts dashboard are created.

Results

Metrics

The first set of results will be metrics calculated using MySQL and Python.

Average Session Duration (hours):

age_bucket	gender	round(AVG(session_duration_
36–50	Male	1.281
26–35	Female	1.272
36–50	Female	1.269
26–35	Male	1.260
18–25	Male	1.258
50+	Female	1.256
18–25	Female	1.242
50+	Male	1.189

Average Calories Burned per Hour:

age_bucket	gender	round(AVG(calories_burned / session_duration..
18–25	Male	792.02
26–35	Male	778.57
36–50	Male	742.50
18–25	Female	725.01
50+	Male	714.04
26–35	Female	706.72
36–50	Female	667.09
50+	Female	649.30

Average Heart Rate Intensity Percentage. This shows effort level relative to maximum heart rate.

age_bucket	gender	round(AVG((avg_BPM / max_BPM) * 100..
36–50	Male	81.06
50+	Female	81.03
36–50	Female	80.60
18–25	Female	80.53
18–25	Male	80.45
50+	Male	80.14
26–35	Male	78.95
26–35	Female	78.62

Average delta between average heart rate and resting heart rate:

age_bucket	gender	round(AVG(avg_BPM - resting_BPM...
18–25	Female	83.33
50+	Male	82.73
36–50	Male	82.37
36–50	Female	81.91
50+	Female	81.66
18–25	Male	81.34
26–35	Male	79.26
26–35	Female	78.80

Average calories burned per kg of body weight:

age_bucket	gender	round(AVG(Calories_Burned / Weight_kg..
18–25	Female	15.35
26–35	Female	14.67
36–50	Female	14.52
50+	Female	13.69
36–50	Male	12.51
26–35	Male	11.69
18–25	Male	11.61
50+	Male	10.57

Average hydration ratio (liters/hour):

age_bucket	gender	ROUND(AVG(Water_Intake_liters / Session_Duration..
50+	Male	2.70
26–35	Male	2.57
36–50	Male	2.56
18–25	Male	2.54
36–50	Female	1.90
50+	Female	1.87
18–25	Female	1.87
26–35	Female	1.86

Average intensity-adjusted calories:

age_bucket	gender	ROUND(AVG(Calories_Burned * (Avg_BPM / Max_BPM)...
18–25	Male	811.95
26–35	Male	783.97
36–50	Male	779.56
18–25	Female	739.31
26–35	Female	708.72
36–50	Female	687.59
50+	Male	687.22
50+	Female	665.94

Average hydration per calorie (milliliters/calorie):

age_bucket	gender	ROUND(AVG((Water_Intake_liters * 1000) / Calories_Burned)...
50+	Male	3.82
36–50	Male	3.50
26–35	Male	3.34
18–25	Male	3.24
50+	Female	2.91
36–50	Female	2.89
26–35	Female	2.64
18–25	Female	2.62

Correlation between Body Mass Index (BMI) and Calories Burned:

	age_bucket	Gender	corr_BMI_Calories
0	18–25	Female	0.155111
1	18–25	Male	0.043814
2	26–35	Female	0.036323
3	26–35	Male	0.063294
4	36–50	Female	-0.030290
5	36–50	Male	-0.018215
6	50+	Female	-0.105039
7	50+	Male	0.014171

Correlation between Fat Percentage and Calories Burned:

	age_bucket	Gender	corr_fat_pct_Calories
0	18–25	Female	-0.531478
1	18–25	Male	-0.632553
2	26–35	Female	-0.713204
3	26–35	Male	-0.587073
4	36–50	Female	-0.584739
5	36–50	Male	-0.661897
6	50+	Female	-0.483384
7	50+	Male	-0.545659

Income percentages by age bucket and gender:

age_bucket	sex	income	percentage
18–25	Female	<=50K	98.64
18–25	Female	>50K	1.36
18–25	Male	<=50K	97.64
18–25	Male	>50K	2.36
26–35	Female	<=50K	88.30
26–35	Female	>50K	11.70
26–35	Male	<=50K	77.90
26–35	Male	>50K	22.10
36–50	Female	<=50K	82.45
36–50	Female	>50K	17.55
36–50	Male	<=50K	55.75
36–50	Male	>50K	44.25
50+	Female	<=50K	88.00
50+	Female	>50K	12.00
50+	Male	<=50K	61.01
50+	Male	>50K	38.99

Average Education Level by income percentages, age bucket, and gender:

age_bucket	sex	income	income_percentage	ROUND(c.avg_education_level..
18–25	Female	<=50K	98.64	9.82
18–25	Female	>50K	1.36	11.15
18–25	Male	<=50K	97.64	9.34
18–25	Male	>50K	2.36	10.39
26–35	Female	<=50K	88.30	10.23
26–35	Female	>50K	11.70	11.96
26–35	Male	<=50K	77.90	9.76
26–35	Male	>50K	22.10	11.62
36–50	Female	<=50K	82.45	10.03
36–50	Female	>50K	17.55	11.88
36–50	Male	<=50K	55.75	9.66
36–50	Male	>50K	44.25	11.70
50+	Female	<=50K	88.00	9.03
50+	Female	>50K	12.00	11.56
50+	Male	<=50K	61.01	8.66
50+	Male	>50K	38.99	11.40

Average Hours Worked per Week by income percentages, age bucket, and gender:

age_bucket	sex	income	income_percentage	avg_hours_per_week
18–25	Female	<=50K	98.64	31.7199
18–25	Female	>50K	1.36	40.7037
18–25	Male	<=50K	97.64	36.5121
18–25	Male	>50K	2.36	46.5082
26–35	Female	<=50K	88.30	38.9707
26–35	Female	>50K	11.70	41.2851
26–35	Male	<=50K	77.90	43.2032
26–35	Male	>50K	22.10	47.2747
36–50	Female	<=50K	82.45	39.0391
36–50	Female	>50K	17.55	40.6076
36–50	Male	<=50K	55.75	43.6595
36–50	Male	>50K	44.25	47.2468
50+	Female	<=50K	88.00	32.8720
50+	Female	>50K	12.00	38.8359
50+	Male	<=50K	61.01	37.3542
50+	Male	>50K	38.99	43.9518

Summary Table

The next result is the Summary Table created using MySQL by joining the Gym and Income tables on Age Bucket and Gender. It is grouped by Age Bucket and Gender and shows average and percentage metrics from both tables.

age_bucket	gender	pct_income_le_50k	pct_income_gt_50k	avg_education_level	avg_hours_worked_	avg_calories_burned	avg_session_duration	a
18-25	Female	98.64	1.36	9.84	31.84	903.62	1.24	
18-25	Male	97.64	2.36	9.37	36.75	998.94	1.26	
26-35	Female	88.3	11.7	10.43	39.24	894.8	1.27	
26-35	Male	77.9	22.1	10.17	44.1	981.88	1.26	
36-50	Female	82.45	17.55	10.36	39.31	846.1	1.27	
36-50	Male	55.75	44.25	10.56	45.25	951.4	1.28	
50+	Female	88	12	9.33	33.59	814.88	1.26	
50+	Male	61.01	38.99	9.73	39.93	848.93	1.19	

age_bucket	gender	avg_bpm	avg_max_bpm	avg_resting_bpm	avg_intensity_pct	avg_heart_rate_delta	avg_calories_per_kg
18-25	Female	145	180.83	61.67	80.53	83.33	15.35
18-25	Male	144	179.63	62.66	80.45	81.34	11.61
26-35	Female	141.35	180.57	62.55	78.62	78.8	14.67
26-35	Male	141.55	180.02	62.29	78.95	79.26	11.69
36-50	Female	143.73	179.24	61.82	80.6	81.91	14.52
36-50	Male	145	179.66	62.63	81.06	82.37	12.51
50+	Female	144.28	178.73	62.62	81.03	81.66	13.69
50+	Male	144.25	180.88	61.52	80.14	82.73	10.57

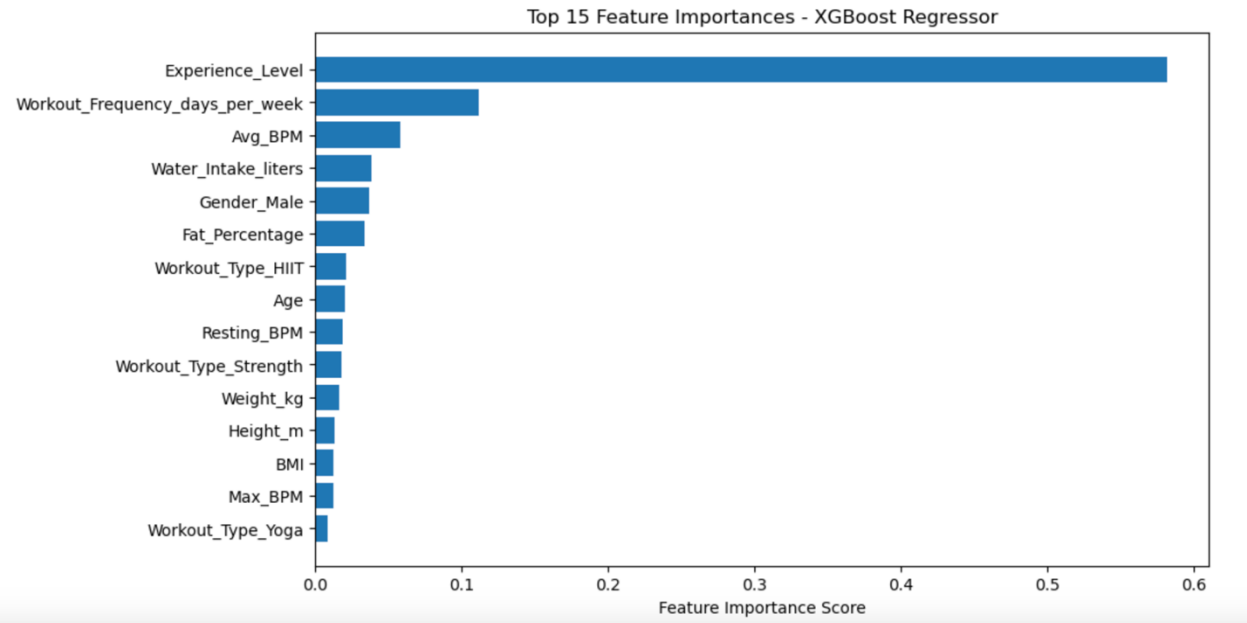
Machine Learning

The next set of results is machine learning analysis using Python. Calories Burned is the outcome variable for the Gym dataset and Income is the outcome variable for the Income dataset.

Multiple Linear Regression for the Gym dataset. Root Mean Squared Error = 160.05. R^2 = 0.62. Intercept and coefficients:

```
intercept  -310.30444138976964
                          Predictor  coefficient
0                               Age    -3.565539
1                         Weight_kg     0.399263
2                          Height_m    28.276786
3                           Max_BPM     0.083474
4                           Avg_BPM     6.583892
5                       Resting_BPM    -0.235657
6                  Fat_Percentage      -6.943042
7              Water_Intake_liters     32.109737
8   Workout_Frequency_days_per_week    16.334381
9                  Experience_Level   197.419351
10                              BMI    -0.880191
11                    Gender_Male     19.315890
12              Workout_Type_HIIT     41.716954
13          Workout_Type_Strength     30.076774
14              Workout_Type_Yoga     18.135953
```

Best-performing Decision Tree Regressor for the Gym dataset: XGBoost Regressor. Root Mean Squared Error = 163.69. R^2 = 0.60. The XGBoost Regressor is made up of multiple decision trees, and each feature's importance is calculated by averaging the importance scores across all trees in the ensemble. The top 15 most important features are selected and plotted:



Logistic Regression for the Income dataset. Accuracy and Confusion Matrix are provided below. Intercept and coefficients:

```
Confusion Matrix (Accuracy 0.8026)

       Prediction
Actual    0    1
     0 7087  468
     1 1506  939
```

```
intercept  [-1.63706744]
                     Predictor  Coefficient
0                          age     0.641437
1                education_num     0.946835
2                hours_per_week     0.418171
3         workclass_Federal-gov     0.207099
4           workclass_Local-gov     0.149988
5        workclass_Never-worked    -0.245494
6             workclass_Private     0.346458
7        workclass_Self-emp-inc     0.250292
8    workclass_Self-emp-not-inc     0.059689
9           workclass_State-gov     0.116270
10        workclass_Without-pay    -0.434194
11      race_Asian-Pac-Islander     0.024273
12                   race_Black     0.051431
13                   race_Other     0.000389
14                   race_White     0.190109
15                     sex_Male     0.542161
```

Best-performing Decision Tree Classifier for the Income dataset: Random Forest Classifier. Accuracy and Confusion Matrix are provided below. The Random Forest Classifier is made up of multiple decision trees, and each feature's importance is calculated by averaging the importance scores across all trees in the ensemble. The top 15 most important features are selected and plotted:

```
Confusion Matrix (Accuracy 0.8062)

       Prediction
Actual    0    1
     0 7030  525
     1 1413 1032
```

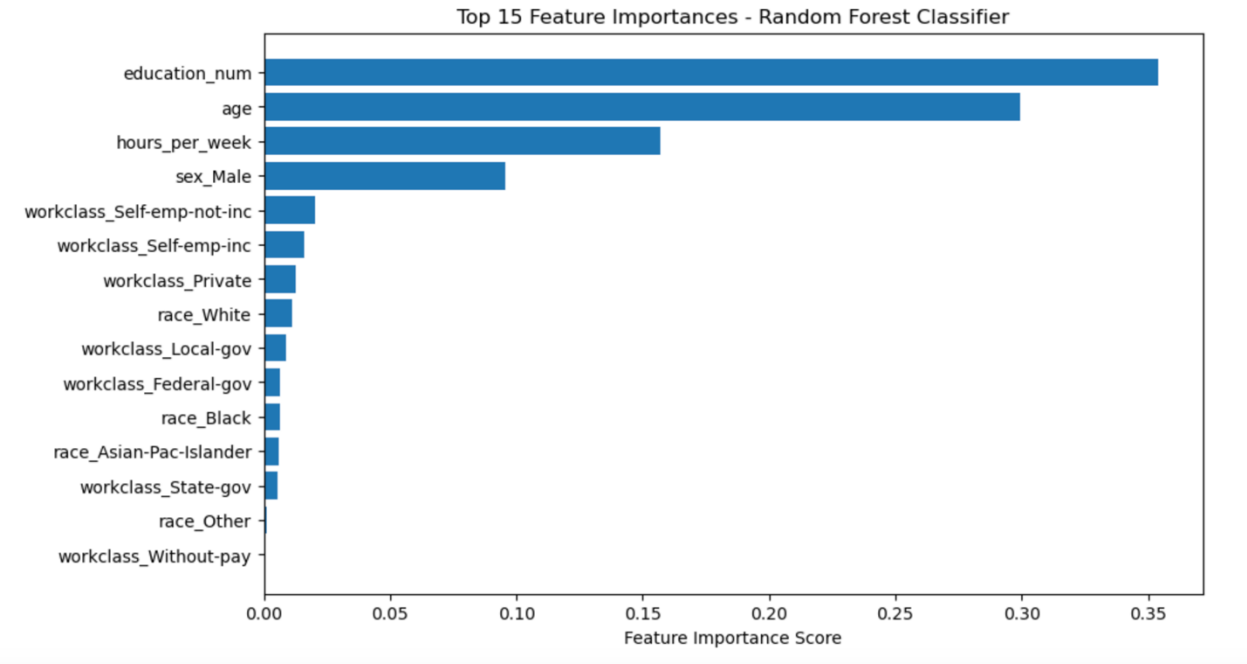
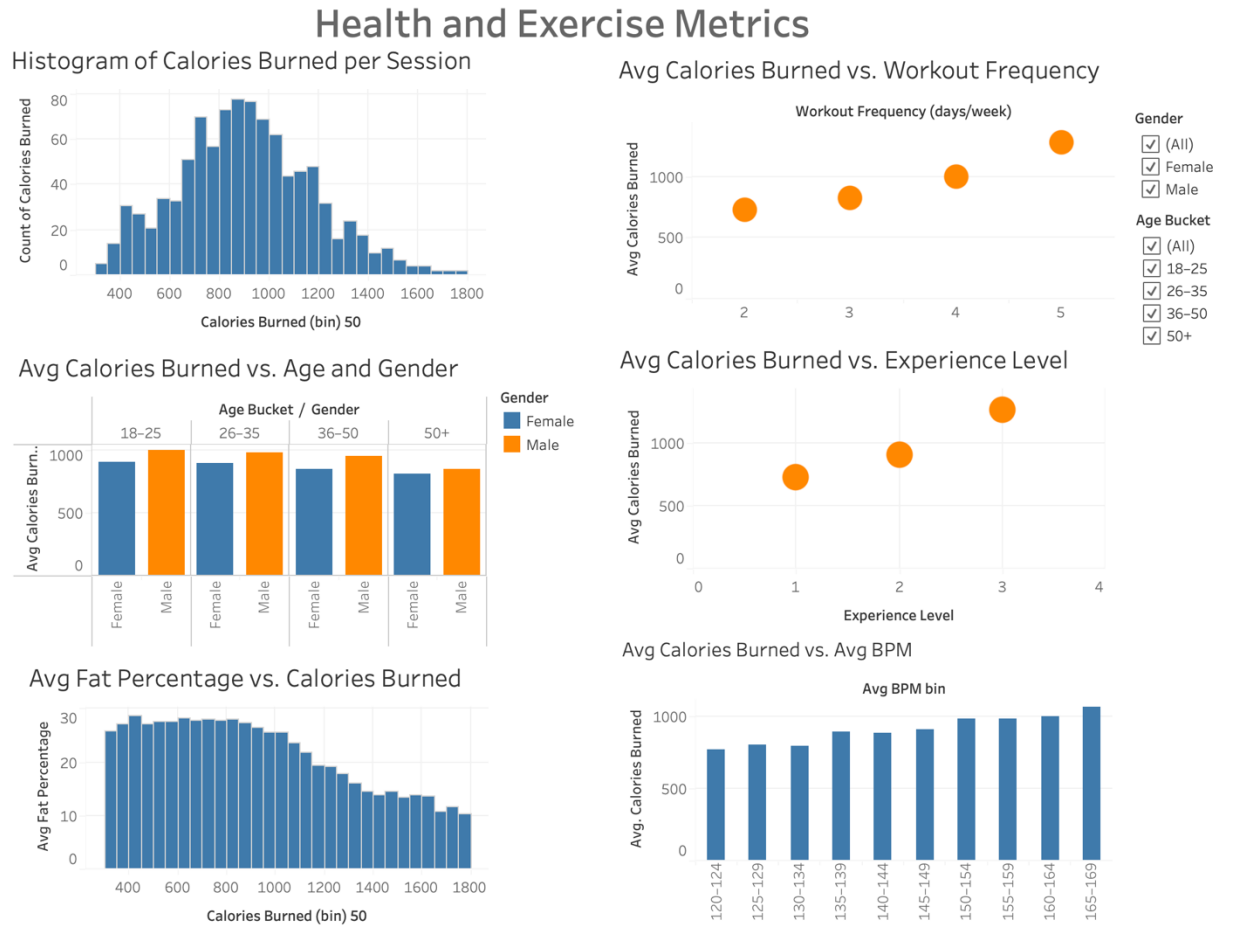
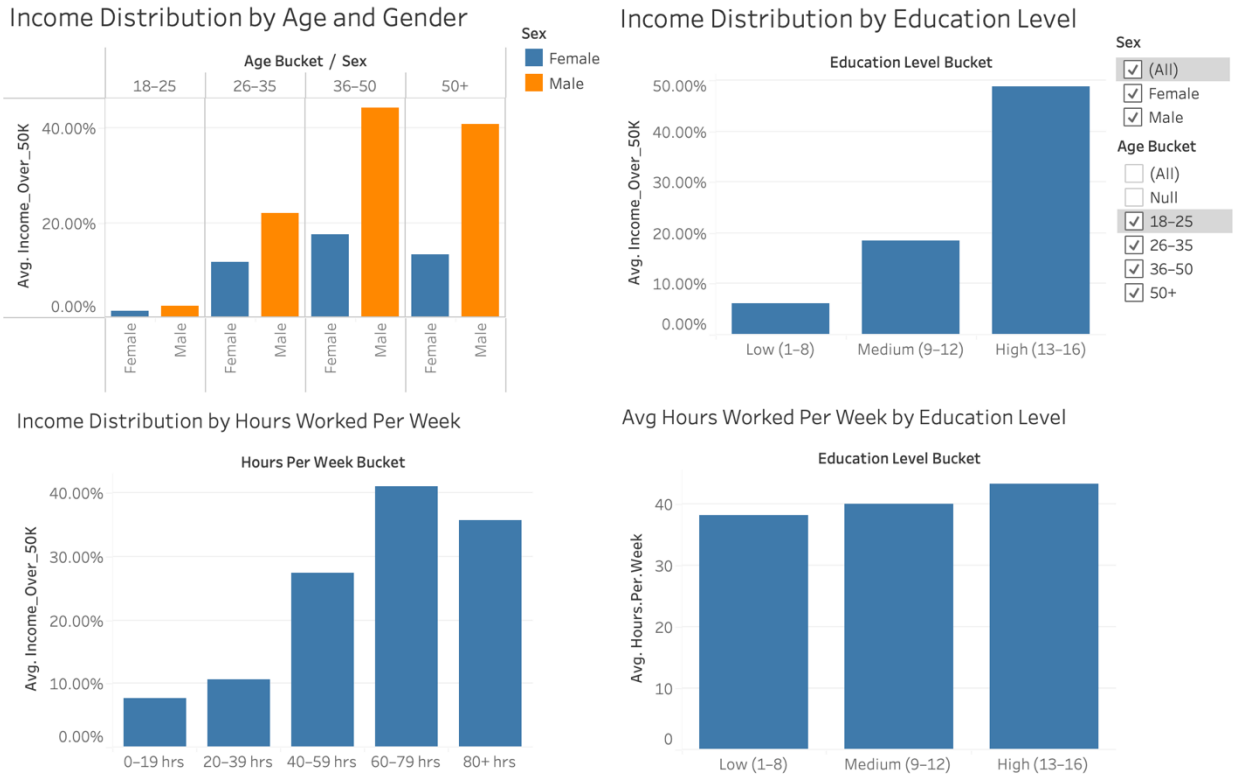

Top 15 Feature Importances - Random Forest Classifier
education_num
age
hours_per_week
sex_Male
workclass_Self-emp-not-inc
workclass_Self-emp-inc
workclass_Private
race_White
workclass_Local-gov
workclass_Federal-gov
race_Black
race_Asian-Pac-Islander
workclass_State-gov
race_Other
workclass_Without-pay
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Feature Importance Score

Tableau Dashboards

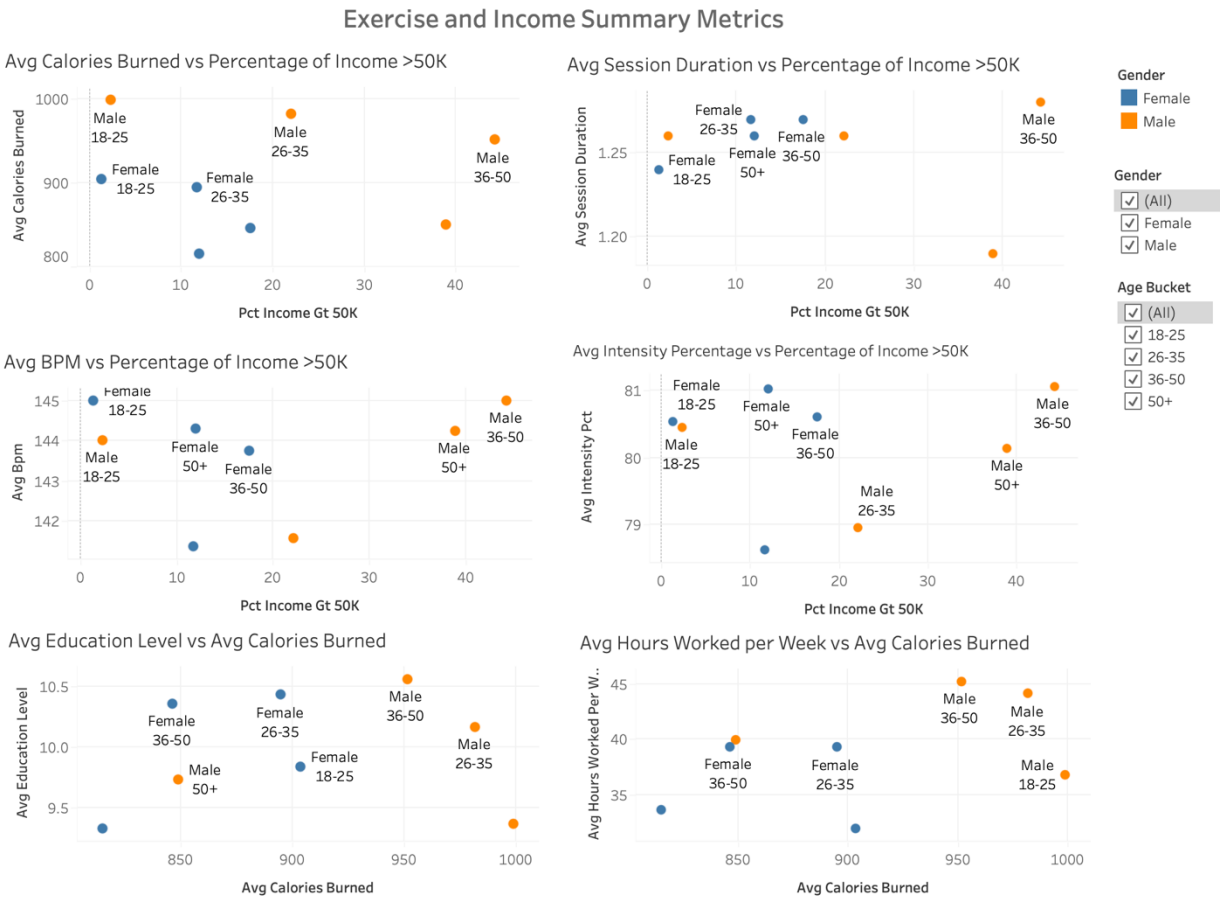


Description: This dashboard provides an overview of workout performance trends across different demographic groups. It visualizes key relationships such as calories burned versus workout frequency, age, gender, experience level, fat percentage, and heart rate. Together, these charts highlight how exercise intensity, frequency, and demographic factors contribute to overall fitness outcomes.

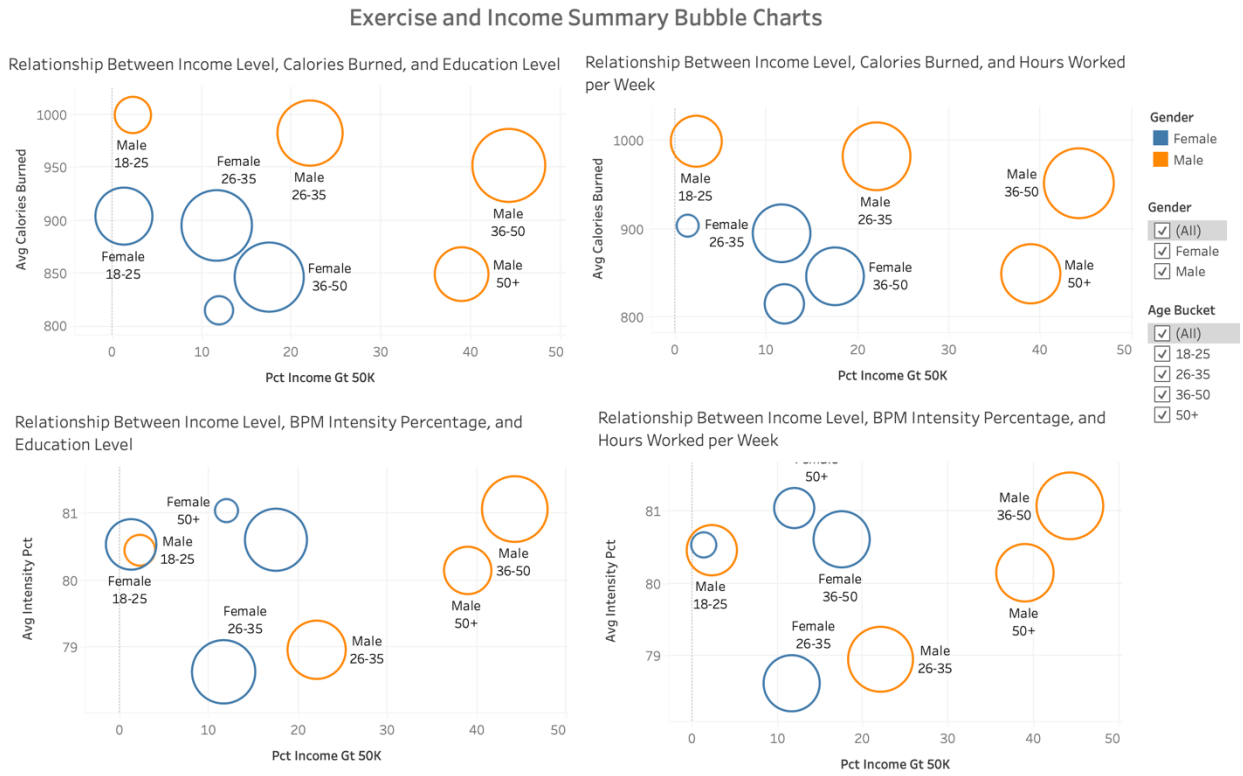
Income Metrics



Description: The purpose of this dashboard is to analyze how demographic and lifestyle factors influence income distribution. It aims to identify patterns in earnings across different age groups, genders, education levels, and work hour ranges. By visualizing these relationships, the dashboard provides insights into which factors most strongly correlate with higher income levels.



Description: The purpose of this dashboard is to analyze how physical fitness metrics relate to socioeconomic and lifestyle factors. It explores the relationships between income level, education level, and hours worked per week alongside workout performance measures such as calories burned, session duration, heart rate (BPM), and BPM intensity percentage. By combining exercise and income data across age and gender groups, the dashboard reveals how demographic and economic factors may influence overall fitness behavior and outcomes.



Description: The purpose of this dashboard is to visualize the multidimensional relationships between income, fitness performance, and lifestyle factors. Using bubble charts, the dashboard displays three variables simultaneously, where the x-axis represents income level, the y-axis shows workout performance metrics such as calories burned or BPM intensity percentage, and bubble size reflects education level or hours worked per week. This visualization helps reveal how demographic and socioeconomic factors interact with physical activity outcomes across different age and gender groups.

Discussion

Metric Analysis

The first part of the discussion section is derived from the Metrics and the Summary Table. Across all age groups, females tend to have slightly longer session durations than males, suggesting that women may spend more time per workout despite burning fewer total calories on average. The longest workout sessions appear in the 36–50 male group

(1.281 hours) and the 26–35 female group (1.272 hours), indicating that these mid-career adults may prioritize longer, possibly more structured workouts. Session duration declines in the 50+ group, particularly for males (1.189 hours), which may reflect age-related physical limitations.

Males outperform females in every age bucket in terms of calories burned per hour, reflecting typical physiological differences such as greater muscle mass and higher average body weight contributing to higher caloric expenditure. There is a noticeable gradual decline in calories burned per hour as age increases, with the lowest values appearing in the 50+ group (714.04 for males and 649.30 for females), which aligns with natural decreases in metabolism and cardiovascular output over time.

The highest heart rate intensity percentage values are observed in the 36–50 male group (81.06%) and the 50+ female group (81.03%), suggesting that individuals in older adulthood may prioritize maintaining strong exertion levels even if overall calories burned decline with age. The lowest intensity percentages appear in the 26–35 group for both males (78.95%) and females (78.62%), possibly reflecting increased professional and personal responsibilities in early career stages that could reduce energy available for peak effort in the gym.

The highest heart rate deltas appear in younger adults, females aged 18–25 (83.33 BPM), indicating a strong cardiovascular response to exercise during early adulthood when heart function, endurance, and recovery tend to peak. The smallest deltas occur in the 26–35 group (78.80–79.26 BPM), paralleling lower heart rate intensity percentages observed in this age range and suggesting that career and personal life demands at this stage may limit the energy available for reaching peak cardiovascular effort.

Females consistently burn more calories per kilogram of body weight than males across all age groups, suggesting higher relative workout efficiency even though males typically burn more total calories. The highest relative burn occurs in the youngest females (18–25, 15.35 calories/kg), indicating very strong metabolic output and workout effectiveness early in adulthood when lean mass and cardiovascular capacity peak. In the 50+ male group, the relative calorie burn drops to 10.57 calories per kilogram, the lowest among all demographics, indicating that aging effects such as reduced muscle mass and metabolic slowdown significantly diminish workout efficiency for older men.

Males show significantly higher hydration ratios than females across all age groups, with values above 2.5 liters/hour in several male groups compared to below 2.0 liters/hour for females, suggesting men may require or choose to consume more water during workouts due to higher sweat rates and muscle mass. Female hydration ratios remain consistent and lower across all age buckets (1.86–1.90 liters/hour), indicating either lower water needs due to physiology or possible under-hydration relative to workload.

The highest values for intensity-adjusted calories are in males 18-25 and males 26-35. Young adult males achieve the strongest combination of calorie burn and high cardiovascular effort, reflecting peak performance and physical capacity. Across all age buckets, males exhibit higher intensity-adjusted calorie output than females, reinforcing the pattern that men generate more total workout energy even when effort levels are normalized by heart-rate intensity. Performance declines gradually with age for both genders, with the 50+ group showing the lowest intensity-adjusted calories (687.22 for males and 665.94 for females), consistent with reductions in muscle mass, endurance, and metabolic rate over time.

Males consistently have higher hydration per calorie rates than females across every age group, suggesting higher perceived hydration needs or greater awareness of fluid replacement relative to exertion. Hydration per calorie increases with age for both men and women, suggesting that older adults consume more water relative to energy expenditure. Physiological efficiency declines with age and hydration becomes increasingly important for sustaining exercise performance.

In the 18–25 age group, correlation between BMI and calories burned values are positive for both females (0.155) and males (0.044), indicating that higher BMI is slightly associated with more calories burned in younger adults. As age increases, correlations weaken considerably and even become slightly negative in the 36–50 and 50+ groups for both genders, suggesting that increased BMI in midlife and older adults does not necessarily translate to improved calorie output during exercise. The most negative correlations occur for females in the 36–50 and 50+ groups (−0.038 and −0.106), implying that higher BMI may reflect lower physical efficiency or decreased workout intensity among older women.

All correlations between fat percentage and calories burned are negative across every age bucket and gender group (ranging from approximately −0.48 to −0.71), indicating a consistent trend that higher body fat percentage is associated with lower calories burned during exercise. Across most age groups, males tend to exhibit slightly stronger negative correlations than females, indicating that higher fat levels may interfere more with exercise efficiency for men. The results highlight that individuals with lower body fat percentage tend to use more energy (burn more calories) during exercise, reinforcing the importance of maintaining lean muscle and healthy weight to maximize workout effectiveness.

From the Income dataset, income percentages by age and gender are calculated. In the 18–25 age group, the vast majority of individuals earn ≤50K (98.64% of females and 97.64% of males), which reflects early career stages where earning potential is still developing. The 36–50 age group shows the strongest earning power, particularly for males: 17.55% of females earn >50K, compared to 44.25% of males earning >50K. This reflects mid-career peak earning years and a widening gender income divide. Across all

age groups, males consistently exhibit higher >50K earning percentages than females, and the gap grows sharply beginning in the 36–50 range, highlighting persistent gender-based disparities in economic opportunity and advancement.

Average education level by income percentages, age, and gender is calculated. For the 26–35 group, individuals earning >50K have noticeably higher education levels (11.96 for females and 11.62 for males) than those earning ≤50K (10.23 for females and 9.76 for males), indicating that early career income gains are closely tied to educational attainment. Across all age groups and income brackets, education levels are consistently lower for individuals earning ≤50K, reinforcing the strong connection between advanced education and economic success. The gender gap is also evident: within the same income bracket, women tend to have equal or higher education levels than men, yet still earn less, supporting evidence of structural wage inequality.

Average Hours Worked per Week by income percentages, age bucket, and gender is calculated. In the 26–35 group, a key career growth stage, hours worked increase across all workers from the 18-25 group, but the difference between income groups becomes even more pronounced: high earners work 41.29–47.27 hours per week, while lower earners work 38.97–43.20 hours, reinforcing the relationship between workload and income acceleration. At every age level, males work more hours than females within the same income bracket, which may reflect occupational differences such as greater representation in roles requiring long work schedules or societal career pressures. Across the full dataset, high-income individuals consistently work longer weeks than low-income individuals, showing a strong link between professional time commitment and earnings potential throughout the lifespan.

Summary Table Analysis

Across all age groups, males earn a higher percentage of >50K income and have higher average calories burned. However, females show higher calories burned per kilogram, meaning women may be exercising more efficiently relative to body weight. The 26–35 male group stands out with a notably high percentage earning >50K (22.10%) and also having among the highest average hours worked per week (44.10) and strong calorie burn (981.88), potentially reflecting a high-performance lifestyle driven by career and physical fitness. In contrast, younger groups (18–25) show very low percentages of >50K income and lower hours worked per week, but have the strongest overall workout performance, especially males (998.94 average calories burned), indicating that free time and energy, rather than income, may be the dominant driver of fitness at earlier life stages.

Individuals in the 26–35 and 36–50 age groups work the most hours per week and also fall in the middle range for calories burned and heart rate intensity percentage, suggesting that increased job demands may limit the time or energy available for high-output exercise even when income is higher. The youngest adults (18–25) work the fewest hours per week yet demonstrate the highest calories burned and heart rate deltas, indicating that more flexible schedules and lower work stress may allow for stronger physical effort in the gym. Males tend to work more hours on average than females in every age bucket, but this does not always translate to stronger fitness outcomes; for example, women consistently burn more calories relative to body weight despite generally working fewer hours.

The 36–50 male group, with the highest proportion earning >50K (44.25%), does not have the highest calories burned. Performance begins to decline slightly as income peaks, suggesting time or physical demands of professional life may reduce peak fitness levels. The 50+ age bucket shows increasing income stability in males (38.99% earning >50K) but reduced calories burned and lower session duration, pointing to age-related limitations despite financial improvements.

Education level rises with income across the age groups, and both factors show a moderate relationship with calories burned, particularly among individuals aged 26–35 and 36–50. This pattern suggests that midlife individuals with higher socioeconomic status may be more likely to engage in healthier lifestyle behaviors, such as maintaining higher workout intensity or consistently participating in exercise programs. Greater educational attainment can also provide increased health awareness, access to fitness resources, and motivation to invest in personal well-being.

Average heart rate remains relatively stable across all age groups and genders, ranging from approximately 141 to 145 BPM, suggesting that most adults working out are maintaining a consistent cardiovascular effort regardless of demographic differences. Heart rate delta (difference between average and resting BPM) shows the highest values in the youngest age bucket (18–25), especially among females (83.33 BPM), which aligns with greater energy reserves and stronger cardiovascular response earlier in adulthood.

Males show slightly higher heart rate intensity percentages (Avg BPM/Max BPM) in midlife, particularly in the 36–50 group (81.06%), suggesting that this demographic may still be pushing themselves at competitive effort levels despite small reductions in calories burned. Females generally show a greater heart rate delta than males within the same age bucket, indicating that women may experience a proportionally stronger heart-rate response to exercise.

Machine Learning Analysis

Multiple Linear Regression with Calories Burned as the outcome variable is performed on the Gym dataset. The coefficients show that Experience_Level has the strongest positive coefficient (197.42), suggesting that more experienced gym users burn substantially more calories per session. This is likely due to higher skill, workout structure, and physical conditioning. Higher Height_m (28.28) and Male gender (19.32) also significantly increase calories burned, consistent with known physiological relationships where larger body size and male muscle composition raise exertion levels. Workout type greatly influences output: HIIT has the highest added effect (41.72), followed by Strength training (30.08) and Yoga (18.14), confirming that high-intensity workouts drive greater caloric expenditure than low-intensity forms. Water_Intake_liters (32.19) and Workout_frequency_days_per_week (16.33) also show positive associations. Conversely, Fat_Percentage has a strong negative coefficient (–6.94), meaning individuals with higher body fat burn fewer calories, aligning with physiological evidence that muscle mass is more metabolically active than fat mass.

Decision Tree Regression analysis is performed on the Gym dataset. XGBoost Regressor is the best-performing model. Experience_Level is by far the most influential predictor, contributing nearly 60% of the total importance score. This reinforces that experienced gym users achieve significantly greater calorie burn due to better-trained movement efficiency, pacing, and workout quality. Workout_Frequency_days_per_week is the second-strongest predictor, showing that consistent exercise habits directly translate to higher caloric output. Avg_BPM and Water_Intake_liters rank next, demonstrating that cardiovascular effort and hydration support increased physical performance and calorie expenditure. Gender_Male continues to have a positive influence on calories burned, highlighting physiological differences such as greater lean mass and higher maximal power output among males. Fat_Percentage appears as a moderately important negative factor, aligning with earlier results: greater body fat reduces metabolic efficiency during exercise.

Logistic Regression with Income as the outcome variable is performed on the Income dataset. The coefficients show that education_num (0.95) and hours_per_week (0.42) are two of the strongest predictors, reinforcing that higher education and longer work hours are key drivers of higher income levels in the dataset. Sex_Male has a substantial positive coefficient (0.54), meaning men are significantly more likely than women to earn >50K. Individuals working for private companies (0.35) and self-employed incorporated (0.25) have higher odds of earning >50K. Age has a positive coefficient (0.64), indicating that earning >50K becomes more likely with career progression and accumulated experience.

Decision Tree Classification analysis is performed on the Income dataset. Random Forest Classifier is the best-performing model. Education_num is the most influential predictor by a wide margin, reflecting that higher levels of formal education are the strongest driver of high-income outcomes in this dataset. Age is the second-most important predictor, showing that income tends to increase as individuals advance in their careers. Hours_per_week is also a major contributor, demonstrating that time spent working plays a critical role in determining whether a person surpasses the 50K income threshold. Sex_Male remains a strong predictor, confirming the gender income disparity observed across the dataset. The work class categories self-employed (both incorporated and non-incorporated) and private are the next most important predictors of high income.

Conclusion

For fitness outcomes, experience level and workout frequency were the strongest predictors of calories burned, indicating that consistent training and skill development play a major role in improving physical performance. Higher average heart rate during exercise and increased water intake were also associated with greater calorie burn, reinforcing the benefits of higher-intensity training and proper hydration. Additionally, fat percentage showed a negative correlation with calories burned, suggesting that reducing excess body fat through balanced nutrition and regular activity may further enhance workout efficiency. These results point to clear and achievable strategies for improvement: exercising more frequently, gradually increasing intensity, staying hydrated, and maintaining healthy body composition.

On the income side, education level emerged as the most important driver of higher earnings. Individuals with more years of schooling were significantly more likely to fall into the >50K income category, highlighting the value of continued learning, certifications, or technical training. Hours worked per week also positively influenced income, reflecting the financial impact of increased workforce participation. Employment in private and self-employed-incorporated roles was linked with stronger earnings outcomes as well, suggesting that pursuing positions in higher-growth industries or entrepreneurial opportunities can provide meaningful economic benefits.

Together, the fitness and income findings show that while demographics such as age and gender cannot be changed, many key predictors of success are within personal

control. By focusing on skill development in the gym, working out more frequently, increasing workout intensity, staying hydrated, investing in education, and increasing professional engagement, individuals can take actionable steps to improve both their physical health and financial stability over time.