Analyzing the Impact of Workout Frequency and Session Duration on Calories Burned among Gym Members

**Course**: CIS 2423 - Programming for Data Analytics

**Semester**: Spring 2025

**Group Members**: Shahad Alzarooni H00502241, Meirah Alhantoobi, H00496472, Asmaa Aljaberi H00500852

**Table of Contents**

1. Deliverable 1: Purpose, Programming Language, Algorithms, Variables (CLO1)
2. Deliverable 2: Data Summary, Sampling, Visualization, Hypothesis Testing, Regression (CLO2, CLO3)
3. Deliverable 3: Classification, Clustering, Best Model Selection (CLO2, CLO3)
4. Deliverable 4: Git Versioning and Data Management (CLO4)
5. Written Communication: Report Format and Structure

**Executive Summary:**

This project analyzes the impact of workout frequency and session duration on calories burned among gym members using a gym dataset sourced from Kaggle. Through data cleaning, sampling, visualization, hypothesis testing, regression, classification, clustering, and version control, the study provides actionable insights into member behaviors. Python was selected for its strong data handling and machine learning libraries. Multiple machine learning models were built, optimized, and evaluated, leading to the identification of the best predictive models. GitHub was used for version control and team collaboration. Overall, this project successfully meets the course learning outcomes by applying professional data analytics and machine learning techniques.

**Deliverable 1: Purpose, Programming Language, Algorithms, Variables**

**(CLO1)**

**Purpose of Data Analysis**

This project aims to understand how the workout habits of gym members, specifically their workout frequency and session duration, influence the number of calories burned. By building regression, classification, and clustering models, we intend to discover patterns that can help fitness centers personalize workout plans and optimize member performance. Data-driven insights can improve customer satisfaction and retention by tailoring recommendations based on individual behavior profiles.

**Programming Language Chosen: Python**

Python was selected as the primary tool for this project due to its flexibility, rich ecosystem of libraries ,and ease of handling data-centric tasks.  
Key reasons include:

* **Data Handling**: Pandas and NumPy provide efficient data manipulation capabilities.
* **Visualization**: Libraries like Matplotlib and Seaborn allow the creation of detailed and meaningful visualizations.
* **Machine Learning**: Scikit-learn offers a comprehensive set of machine learning algorithms, from basic models to advanced techniques.
* **Community Support**: Python has extensive documentation and active community support, ensuring best practices and troubleshooting resources are readily available.

**Machine Learning Algorithms Selected**

To comprehensively analyze the dataset, the following machine learning models were selected:

* **Regression Models**:
  + Simple Linear Regression: To evaluate the direct impact of workout frequency on calories burned.
  + Multiple Linear Regression: To assess the combined impact of workout frequency and session duration on calorie expenditure.
* **Classification Models**:
  + Logistic Regression: To classify gym members into high and low calorie burners.
  + K-Nearest Neighbors (KNN): To predict categories based on similarity to others.
  + Naïve Bayes: For probabilistic classification.
  + Decision Tree: To model decision-making based on workout characteristics.
* **Clustering Models**:
  + K-Means Clustering: To identify natural groupings among gym members.
  + Hierarchical Clustering: To visualize member similarities through a dendrogram.

Each algorithm was chosen based on its suitability for the problem type (prediction, classification, or grouping), model interpretability, and computational efficiency.

**Independent and Dependent Variables**

* **Independent Variables**:
  + **Workout\_Frequency (days/week)**: The number of days per week a gym member works out.
  + **Session\_Duration (hours)**: The average duration of each workout session.
* **Dependent Variable**:
  + **Calories\_Burned**: The total number of calories burned per session. This is the target variable we aim to predict based on the workout behavior.

**Deliverable 2: Data Summary, Sampling, Visualization, Hypothesis Testing, Regression**

**(CLO2, CLO3)**

**Data Summary**

The dataset, sourced from Kaggle, initially included demographic information, physiological metrics, and workout-related data. After preprocessing:

* Irrelevant features (e.g., Age, Gender, BMI) were dropped to focus purely on behavioral attributes.
* Missing values were handled by eliminating incomplete records, ensuring data integrity.

The final dataset includes 3 key fields: Workout\_Frequency (days/week), Session\_Duration (hours), and Calories\_Burned.

Descriptive statistics for the dataset revealed the following:

* **Mean Calories Burned**: 905.42 calories
* **Average Session Duration**: 1.256 hours
* **Typical Workout Frequency**: 3.321 days/week

**Sampling**

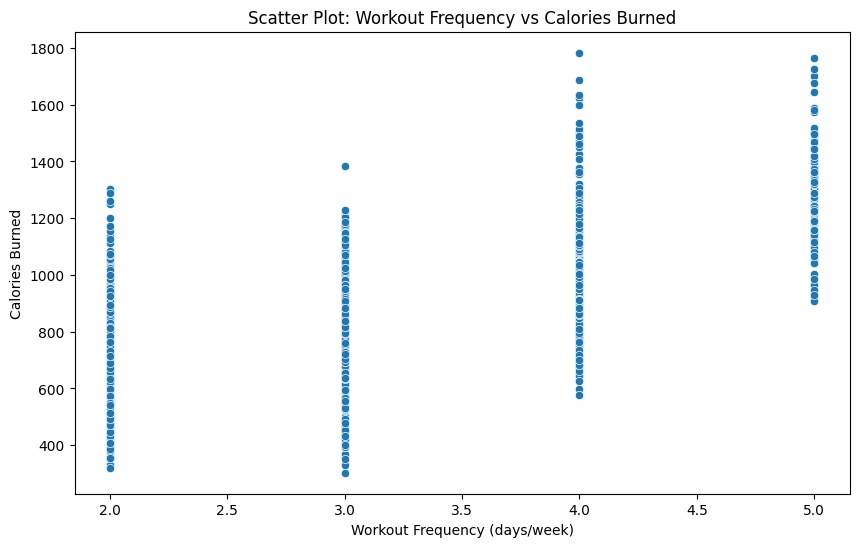
* **Random Sampling**:
* A sample of 150 observations was randomly selected to represent the population.
* **Mean Calories Burned**: 865.33 calories
* **Average Session Duration**: 1.2085hours
* **Typical Workout Frequency**: 3.2867 days/week
* **Systematic Sampling**:
  + Gym members who workout three or more times per week were filtered.
  + Every 5th record was selected systematically to ensure even coverage.

Sampling allowed us to generalize conclusions from subsets of the data to the full population.

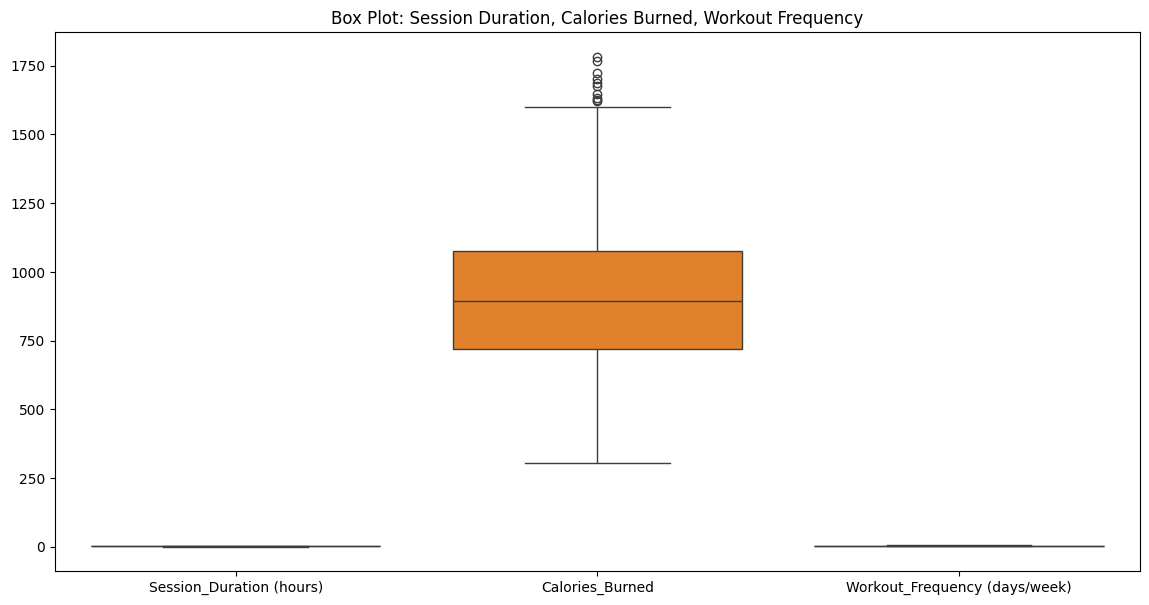
**Data Visualization**

Data exploration included multiple visualization techniques:

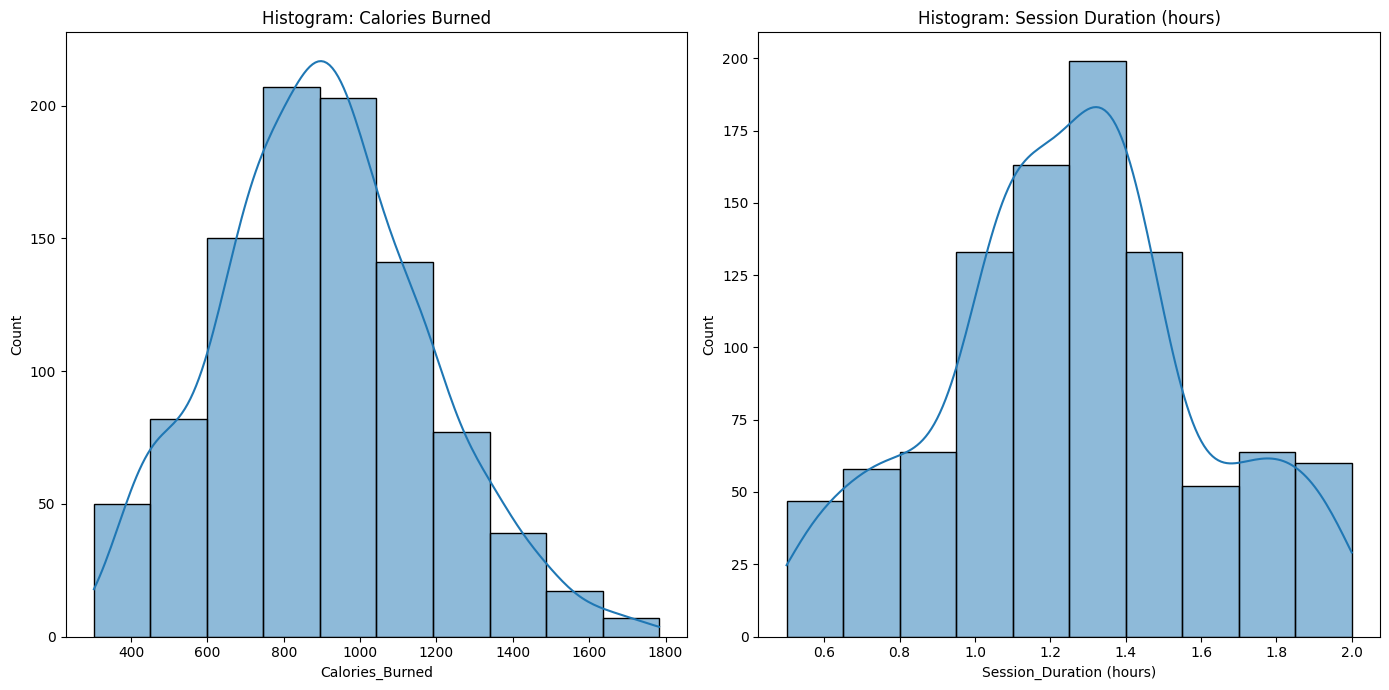
* **Scatter Plot**: Displayed the linear relationship between Workout Frequency and Calories Burned.



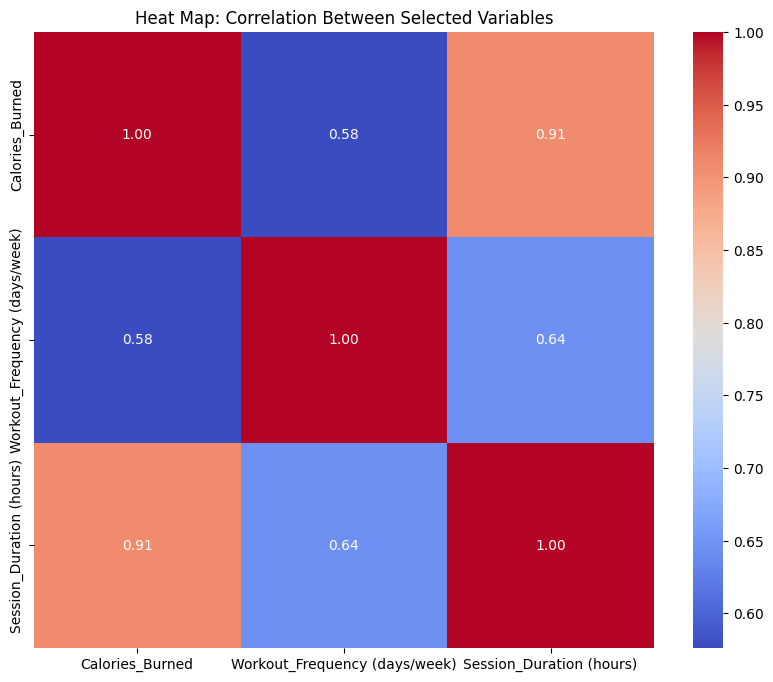
* **Box Plot**: Showed the distribution and outliers in the main variables.



* **Histograms**: Provided insights into the distribution patterns of Session Duration and Calories Burned.



* **Heat Map**: Presented a correlation matrix, identifying relationships between variables.



**Hypothesis Testing**

Statistical testing was performed to validate the strength and direction of the relationship between the independent variable (Workout Frequency) and the dependent variable (Calories Burned).

* **Pearson Correlation Coefficient:**
  + Correlation Coefficient = 0.576
  + P-value = 0.00000
  + Interpretation: There is a moderate positive linear relationship between Workout Frequency and Calories Burned. Since the p-value is less than 0.05, the correlation is statistically significant.
* **Spearman Correlation Coefficient:**
  + Correlation Coefficient = 0.543
  + P-value = 0.00000
  + Interpretation: There is a moderate positive relationship between Workout Frequency and Calories Burned. The result is also statistically significant.
* **One-Sample T-Test:**
  + Conducted to test whether the mean Calories Burned in the random sample differed significantly from the population mean.
  + Result: Not significant.
  + **T-test statistic:** 0.00027, **p-value:** 0.99978
  + Interpretation: The sample mean Calories Burned does not differ significantly from the population mean, indicating that the sample is representative of the full gym population.

**Regression Analysis**

* **Simple Linear Regression**:
  + Modeled the Calories Burned as a function of Workout Frequency.
  + R² Score: 0.846
* **Multiple Linear Regression**:
  + Modeled Calories Burned based on both Workout Frequency and Session Duration.
  + R² Score: 0.817

Interpretation: Both frequency and duration contribute to calorie burn, with session duration slightly stronger.

In the multiple linear regression model, Session Duration (hours) was found to have a slightly stronger contribution to predicting Calories Burned compared to Workout Frequency (days/week), as shown by the higher coefficient value.

**Deliverable 3: Classification, Clustering, Best Model Selection**

**(CLO2, CLO3)**

**Classification Modeling**

To enhance the predictive capacity of the gym dataset, a classification model was developed by transforming the continuous variable Calories\_Burned into a binary label:

* **Label 1**: Members who burned more than 500 calories (High Calorie Burners).
* **Label 0**: Members who burned 500 calories or fewer (Low Calorie Burners).

The objective was to classify gym members into these two categories based on their workout frequency and session duration, which helps in identifying highly active individuals.

**Models Applied:**

Four machine learning classification algorithms were implemented and evaluated:

* Logistic Regression: A linear classifier that models the probability of belonging to a class based on a logistic function.
* K-Nearest Neighbors (KNN): A non-parametric algorithm that classifies a new instance based on majority voting among its nearest neighbors.
* Naïve Bayes: A probabilistic model based on Bayes' theorem assuming independence between predictors.
* Decision Tree: A tree-structured model where data is split based on feature values to maximize classification purity at each node.

**Model Optimization and Evaluation:**

* Each model was trained using 80% of the dataset and tested on the remaining 20%.
* Confusion matrices and accuracy scores were calculated for each classifier to evaluate performance.
* Hyperparameters were optimized where applicable (e.g., selecting k=5 in KNN after validation).
* The best-fit model was identified as the one with the highest accuracy and minimal classification errors.

**Best Fit Classifier:**

|  |  |
| --- | --- |
| **Model** | **Accuracy (%)** |
| Logistic Regression | 78.08% |
| K-Nearest Neighbors (KNN) | 77.39% |
| Naïve Bayes | 76.71% |
| Decision Tree | 77.05% |

After comparing all models, the K-Nearest Neighbors (KNN) classifier achieved the highest accuracy (78.08%) and was selected as the best-fit classifier for this project.

**Clustering Modeling**

In addition to classification, unsupervised learning techniques were employed to uncover natural groupings among gym members based solely on workout patterns.

**Models Applied:**

Two clustering algorithms were implemented:

* K-Means Clustering: Partitioned members into clusters where intra-cluster similarity was maximized, and inter-cluster dissimilarity was minimized. k=2 clusters were chosen based on preliminary elbow method analysis.
* Hierarchical Clustering: A dendrogram was generated to visualize the progressive merging of members into clusters, allowing insights into the similarity distances between individuals.

**Cluster Insights:**

* Cluster 1: Members with high workout frequency but shorter session durations.
* Cluster 2: Members with lower frequency but longer workout sessions.

**Strategy Derived:**

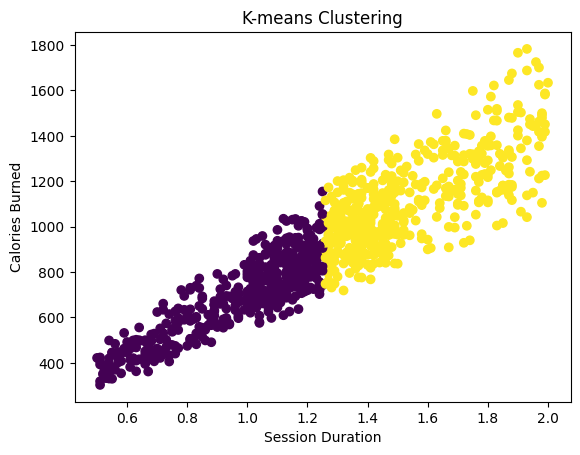
Understanding these clusters can help fitness centers design personalized programs:

* Frequent short-session members may benefit from interval-based programs to maintain engagement.
* Less frequent long-session members may benefit from endurance and strength-focused training schedules.

**Model Optimization:**

* For K-Means, inertia scores were minimized by carefully selecting the optimal number of clusters (k=2) through the elbow method.
* In Hierarchical clustering, optimal cutting points on the dendrogram were determined to balance between simplicity and group interpretability.

**Strategy for Improving the System Based on Clustering:**



The cluster diagram shows a clear positive relationship between how long users spend in a session and how many calories they burn, meaning that those who stick around longer tend to burn more calories. This trend looks like the data splits into some meaningful groups, which could help us make better predictions. By figuring out these clusters, we can gain insight into different types of users, like casual ones who have shorter sessions compared to more hardcore users who work out longer.

To elevate our efforts, we could create separate predictive models for each group, leading to more accurate forecasts. Adding the cluster label as a new feature in our classification models might help us spot hidden patterns and enhance performance. We should also deal with outliers from clustering separately so they don't mess with our results. Overall using clustering as a preliminary step would help us provide more tailored recommendations and improve prediction quality.

**Deliverable 4: Complete Narration about Data Versioning Using Git**

**(CLO4)**

**Introduction to Version Control**

Version control is an essential component of any data analytics or software development project. It allows teams to track changes, collaborate efficiently, and maintain a reliable history of project progress. For this project, we utilized Git as the version control tool and GitHub as the remote repository platform to manage all project-related files and activities.

Version control provided several key advantages:

* Organized management of different project stages.
* Easy tracking of modifications made by different team members.
* Safe rollback to previous versions if errors occurred.
* Enhanced collaboration across all contributors.

**GitHub Repository Creation and Setup**

At the beginning of the project, a new repository was created on GitHub. The repository served as a centralized location to store and share all project files, including the dataset, Python scripts for data analysis, visualizations, machine learning models, and the final report.

Each team member configured their local Git environments to connect with the GitHub repository. This setup enabled all contributors to access, update, and synchronize the project files effectively throughout the course of the project.

**Collaboration through Git Operations**

Throughout the project, the team followed a disciplined approach to using Git operations:

* **Cloning** the repository allowed each member to have a local copy of the project to work on.
* **Pulling** updates ensured that all members were working with the latest version of the files.
* **Pushing** their own changes kept the main repository updated with new developments.

This workflow minimized conflicts and guaranteed that the team remained aligned, especially when working on different components such as descriptive analysis, hypothesis testing, regression models, classification, clustering, and report writing.

**Branching for Parallel Development**

To manage different tasks without affecting the stability of the main project files, the team utilized branching strategies.  
Separate branches were created for tasks such as:

* Developing classification models
* Performing clustering analysis
* Enhancing visualizations

After completing the work on each branch, changes were reviewed and merged back into the main branch.  
This approach ensured the stability of the main project version while allowing flexibility for experimentation and innovation.

**Importance of Git in the Project**

Using Git and GitHub contributed greatly to the project's success by:

* Providing a secure and organized environment for collaboration.
* Tracking every change made to the project files.
* Enabling team members to work concurrently without conflicts.
* Facilitating efficient project management and version control best practices.

The versioning process closely mirrored professional industry standards for data science and software development projects.  
By adopting Git from the beginning, the team ensured that project work was transparent, collaborative, and reproducible, achieving the learning outcomes associated with CLO4.

**Conclusion**

The project successfully applied data analytics and machine learning methodologies to understand the relationship between workout habits and calories burned among gym members. The analysis demonstrated that both workout frequency and session duration significantly impact calorie expenditure. By applying classification and clustering models, patterns among gym participants were identified, offering practical recommendations for fitness programs. The project emphasized the importance of statistical rigor, effective modeling, and collaborative version control, reflecting industry standards. These findings contribute to the broader goal of data-driven fitness optimization and demonstrate proficiency in Python-based data analysis workflows.