Laboratory assignment

Component 1 + 2 + 3

Authors: Turcu Ciprian-Stelian
Group:242

January 13, 2025

Contents

First Component - Definition of the learning tasks

1 Problem Definition

Given a collection of gym members exercise tracking data, we established two objectives. The first objective aims to appropriately classify each member's experience level using a set of input features. Each member falls under a specific experience category, such as "Beginner" "Intermediate" "Expert". The system should predict the most likely experience level for a new or existing member based on particular criteria such as workout duration, frequency, fat percentage, and calories burned. This enables tailored recommendations and improved user engagement.

2 Problem Specification

2.1 Input data and preconditions

2.1.1 Input

The dataset represents a collection of gym members' exercise tracking data.

2.1.2 Preconditions

Preprocessing of data to handle missing values, scale numerical features, and encode categorical data.

The dataset should be divided into training and testing sets to evaluate the model.

2.2 Output and Postconditions

2.2.1 Output

The classification model will predict the experience level of each member, from one of several possible levels.

2.2.2 Postconditions

The output is evaluated for its accuracy in predicting the correct experience level for the first task.

Performance metrics include accuracy, precision, recall, and F1-score to validate the classification results.

The solution should be generalizable enough to classify members not present in the training data effectively.

3 Specification of the first learning task

3.1 Task

Classify each gym member into their appropriate experience level using features like Workout Duration, Intensity, Frequency, and Type of Exercise.

3.2 Performance

The performance of the model will be evaluated using metrics relevant to classification tasks:

Accuracy: The ratio of correct predictions to total predictions.

Precision and Recall: To evaluate the reliability of predictions for each experience level.

F1-score: The harmonic mean of precision and recall, particularly useful in case of imbalanced classes.

3.3 Experience

The model is trained using historical data available in the gym members' exercise tracking dataset.

The model will learn from examples of members with known experience levels and associated features to build a generalizable model that can accurately classify new, unseen members.

Second Component - Data analysis

4 Gym Members Exercise Tracking Dataset Description

The Gym Members Exercise Tracking Dataset contains data collected from a fitness tracking system, capturing detailed workout and physiological information of gym members. The dataset represents comprehensive member data to aid in exercise tracking and fitness analysis.

Attributes

1. **Age:** The age of the gym member.
2. **Gender:** The gender of the gym member, encoded as binary (0 for female, 1 for male).
3. **Weight (kg):** The weight of the gym member in kilograms.
4. **Height (m):** The height of the gym member in meters.
5. **Max_BPM:** The maximum heart rate during exercise.
6. **Avg_BPM:** The average heart rate during exercise.
7. **Resting_BPM:** The resting heart rate of the gym member.
8. **Session Duration (hrs):** The duration of the workout session in hours.
9. **Calories Burned:** The total calories burned during the workout session.
10. **Workout Type:** The type of workout performed.
11. **Fat Percentage:** The body fat percentage of the gym member.
12. **Water Intake (liters):** The amount of water consumed during the session, in liters.
13. **Workout Frequency (days/wk):** The average number of workout days per week.
14. **Experience Level:** The gym member's experience level, categorized as 1 (Beginner), 2 (Intermediate), or 3 (Expert).
15. **BMI:** The Body Mass Index of the gym member.

5 Analysis of the features used in learning

The **Count** indicates that all fields have complete data for the 973 entries in the dataset. Features such as `Calories Burned`, `Age`, and `BMI` exhibit wide distributions, as reflected in their **Mean** and **Std Dev** values.

The **Min** and **Max** values for `Calories Burned` (303 to 1783) and `BMI` (12.32 to 49.84) suggest the dataset captures a diverse range of member profiles. Notably, features like `Experience Level` and `Workout Frequency (days/wk)` display narrower ranges and more uniform distributions, indicating consistent behavior in these categories.

Binary fields such as `Gender` and categorical attributes like `Workout Type` exhibit minimal variability, with `Gender` being nearly balanced (mean 0.53). Continuous variables like `Max_BPM` and `Avg_BPM` show higher variation, reflecting differences in members' cardiovascular responses during exercise.

The **Mean**, **Median (50th Percentile)**, and **Std Dev** for each attribute indicate the distribution shape, with fields such as `Calories Burned` and `Weight (kg)` showing skewness due to outliers. Additionally, the `BMI` and `Fat Percentage` values are consistent with typical fitness data, supporting the dataset's relevance for fitness-related predictive tasks.

Table 1: Summary Statistics

Column	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Age	973.0	38.68	12.18	18.0	28.0	40.0	49.0	59.0
Gender	973.0	0.53	0.50	0.0	0.0	1.0	1.0	1.0
Weight (kg)	973.0	73.85	21.21	40.0	58.1	70.0	86.0	129.9
Height (m)	973.0	1.72	0.13	1.5	1.62	1.71	1.8	2.0
Max_BPM	973.0	179.88	11.53	160.0	170.0	180.0	190.0	199.0
Avg_BPM	973.0	143.77	14.35	120.0	131.0	143.0	156.0	169.0
Resting_BPM	973.0	62.22	7.33	50.0	56.0	62.0	68.0	74.0
Session Duration (hrs)	973.0	1.26	0.34	0.5	1.04	1.26	1.46	2.0
Calories Burned	973.0	905.42	272.64	303.0	720.0	893.0	1076.0	1783.0
Workout Type	973.0	1.49	1.13	0.0	0.0	2.0	2.0	3.0
Fat Percentage	973.0	24.98	6.26	10.0	21.3	26.2	29.3	35.0
Water Intake (liters)	973.0	2.63	0.60	1.5	2.2	2.6	3.1	3.7
Workout Frequency (days/wk)	973.0	3.32	0.91	2.0	3.0	3.0	4.0	5.0
Experience Level	973.0	1.81	0.74	1.0	1.0	2.0	2.0	3.0
BMI	973.0	24.91	6.66	12.32	20.11	24.16	28.56	49.84

5.1 Feature Interpretations

Age

- The average age is 38.68 years, with a moderate spread (Std Dev = 12.18).
- The range (18 to 59 years) suggests a dataset focused on adult participants, primarily middle-aged individuals.
- The median (40 years) shows that half of the participants are younger than this age.

Gender

- Gender is represented as a binary variable, with the average (0.53) suggesting a near-even distribution between the two genders.
- The Std Dev (0.50) confirms the binary nature of the feature.

Weight (kg)

- The average weight is 73.85 kg, with a wide variation (Std Dev = 21.21).
- Most weights fall between 58.1 kg (25th percentile) and 86 kg (75th percentile), with extreme values up to 129.9 kg.

Height (m)

- The average height is 1.72 m, with a relatively small variation (Std Dev = 0.13).
- Most heights range from 1.62 m (25th percentile) to 1.8 m (75th percentile), with a minimum of 1.5 m and a maximum of 2.0 m.

Max_BPM

- The average maximum BPM is 179.88, with a narrow spread (Std Dev = 11.53).
- Most maximum heart rates fall between 170 BPM and 190 BPM, reflecting typical ranges during peak activity.

Avg_BPM

- The average BPM is 143.77, with moderate variation (Std Dev = 14.35).
- The interquartile range (131 BPM to 156 BPM) indicates normal heart rate during moderate activity.

Resting_BPM

- The average resting BPM is 62.22, with low variation (Std Dev = 7.33).
- Typical resting BPM values fall between 56 BPM and 68 BPM, with extremes as low as 50 BPM and as high as 74 BPM.

Session Duration (hours)

- The average session duration is 1.26 hours, with low variation (Std Dev = 0.34).
- Most sessions last between 1.04 and 1.46 hours, with a maximum of 2 hours.

Calories Burned

- The average calories burned is 905.42 kcal, with significant variation (Std Dev = 272.64).
- Most values range from 720 kcal to 1,076 kcal, with a minimum of 303 kcal and a maximum of 1,783 kcal.

Workout Type

- The average workout type is 1.49, with a moderate spread (Std Dev = 1.13).
- Values range from 0 (perhaps sedentary or less active) to 3, reflecting a diverse set of workout intensities or styles.

Fat Percentage

- The average fat percentage is 24.98%, with a moderate spread (Std Dev = 6.26).
- Most values range from 21.3% to 29.3%, with extremes between 10% and 35%.

Water Intake (liters)

- The average water intake is 2.63 liters, with a small spread (Std Dev = 0.60).
- Most values range between 2.2 and 3.1 liters, with a maximum of 3.7 liters.

Workout Frequency (days per week)

- Participants exercise an average of 3.32 days per week, with low variation (Std Dev = 0.91).
- Most people work out between 3 and 4 days per week, with a minimum of 2 days and a maximum of 5 days.

Experience Level

- The average experience level is 1.81, with a moderate spread (Std Dev = 0.74).
- Most participants have an experience level between 1 (beginner) and 2 (intermediate), with a maximum of 3.

BMI

- The average BMI is 24.91, with a significant spread (Std Dev = 6.66).
- Most BMI values range from 20.11 (healthy) to 28.56 (overweight), with extremes from 12.32 (underweight) to 49.84 (obese).

5.2 Correlation

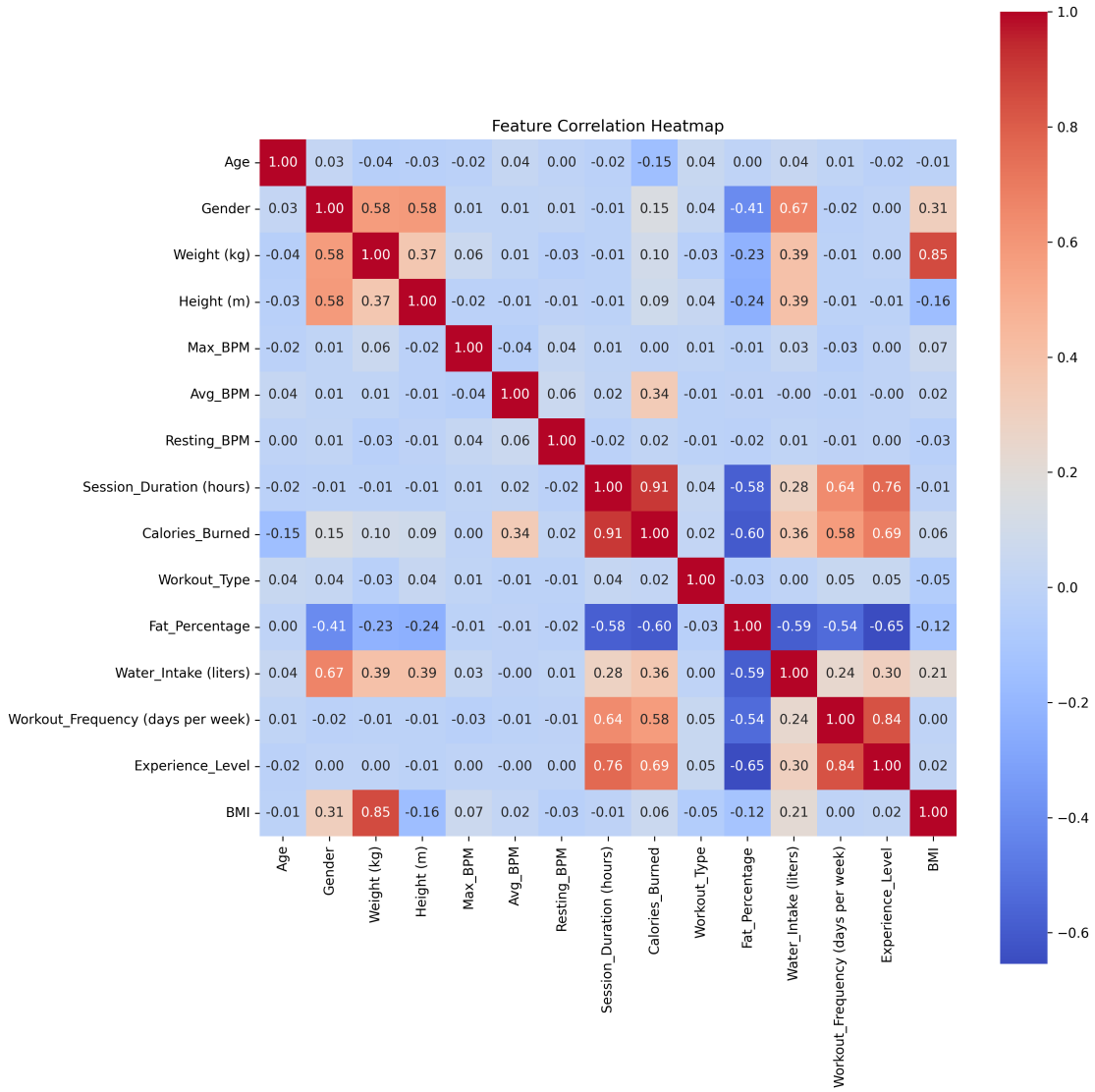


Figure 1: Correlation Heatmap

The heatmap highlights significant relationships among the dataset's features, with particular focus on `Experience Level` and `Calories Burned`. `Experience Level` is strongly

correlated with `Workout Frequency (days per week)` ($r = 0.84$) and moderately negatively correlated with `Fat Percentage` ($r = -0.63$), indicating that more experienced individuals tend to work out more frequently and have lower body fat. `Calories Burned` demonstrates strong correlations with `Session Duration (hours)` ($r = 0.91$) and moderate correlations with `Workout Frequency (days per week)` ($r = 0.69$), reinforcing the importance of workout habits in calorie expenditure.

Additionally, `BMI` is highly correlated with `Weight (kg)` ($r = 0.85$), while `Water Intake (liters)` shows notable associations with `Gender` ($r = 0.67$) and `Workout Frequency (days per week)` ($r = 0.54$). Interestingly, `Fat Percentage` also has moderate negative correlations with `Session Duration (hours)` ($r = -0.60$), suggesting a link between workout duration and body composition. These relationships reflect the interplay between activity levels, body metrics, and fitness outcomes in the dataset.

6 Data Distribution

6.1 Demographics and Physical Attributes

- **Age Distribution:** The dataset covers a wide range of ages with a relatively uniform distribution, representing diversity across age groups.
- **Height (m):** The bimodal distribution suggests distinct subgroups (possibly by gender or population characteristics).
- **Weight (kg):** Weight is right-skewed, with most individuals weighing between 60 and 70 kg, reflecting a typical population distribution.
- **BMI:** The unimodal, right-skewed distribution peaks around 25 BMI, aligning with the range for average adults.

6.2 Health and Fitness Metrics

- **Calories Burned:** Most individuals burn between 700 and 1000 calories, with a tail extending to higher ranges.
- **Resting BPM:** Resting heart rates have a flat distribution within a healthy range of 50–75 BPM.
- **Max BPM:** Maximum heart rate is uniformly distributed, with a slight peak around 180 BPM, typical for cardiovascular activities.
- **Avg BPM:** Similar to Max BPM, it exhibits a flat distribution spanning from 120 to 170 BPM.

6.3 Lifestyle and Activity Patterns

- **Workout Frequency (days per week):** Most individuals work out 3–4 times per week, representing moderate activity. Few work out 2 or 5 days, potentially representing beginners or highly consistent athletes.
- **Workout Type:** There is an even split among the four types (Strength, Cardio, Yoga, HIIT), suggesting balanced representation in activity preferences.
- **Session Duration (hours):** Sessions are clustered around 1.2–1.4 hours, indicating a standard workout duration, with few sessions exceeding 2 hours.

6.4 Hydration and Composition

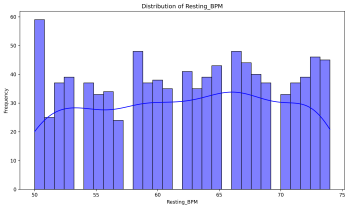
- **Water Intake (liters):** The bimodal distribution around 2.5 and 3.5 liters suggests two distinct hydration patterns.
- **Fat Percentage:** Most values are clustered around 25–30%, representing typical body composition ranges, with fewer individuals at extreme ends.

6.5 Experience Level and Gender

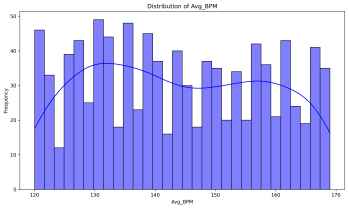
- **Experience Level:** Mid-level experience dominates (Level 2), with fewer individuals in entry-level (Level 1) or advanced categories (Level 3).
- **Gender:** The dataset is balanced in gender representation, with a slightly higher proportion of males.

6.6 Key Takeaways

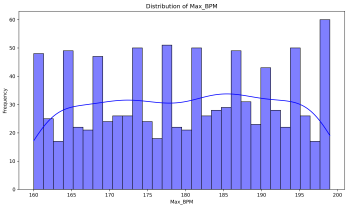
1. The dataset is well-distributed across age, gender, and health attributes, enabling insights into diverse population groups.
2. Activity levels and fitness metrics are reflective of moderate exercisers, with a few outliers suggesting more intense routines.
3. The diversity in workout types and session durations shows variability in preferences and commitment levels.



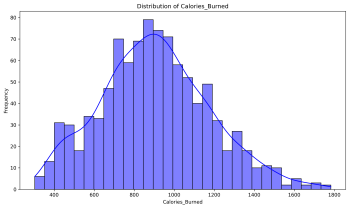
(a) Distribution Resting BPM feature



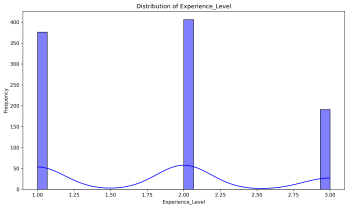
(b) Distribution of the Average BPM



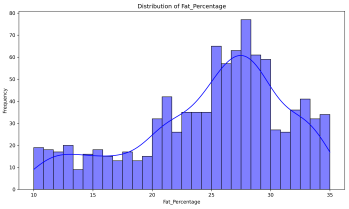
(c) Distribution of the Max BPM



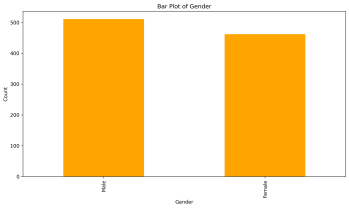
(d) Distribution of the Calories Burned



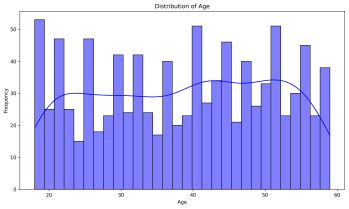
(e) Distribution of the Experience Level



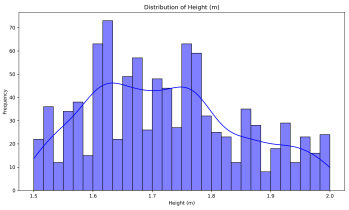
(f) Distribution of the Fat Percentage



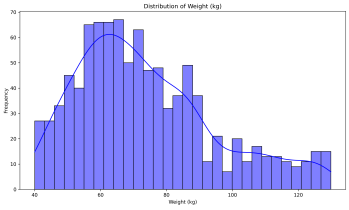
(g) Distribution of the Gender



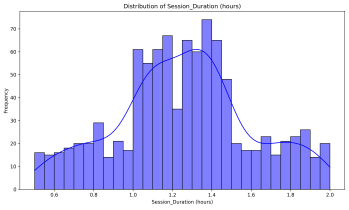
(h) Distribution of the Age



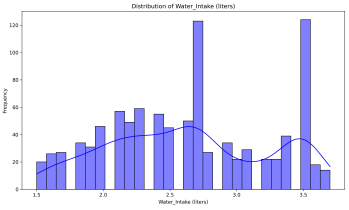
(i) Distribution of Height in meters



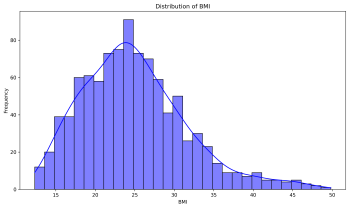
(j) Distribution of Weight in kilograms



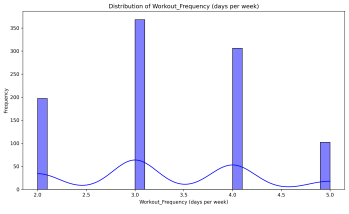
(k) Distribution Session Duration in hours



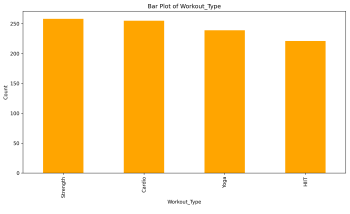
(l) Distribution of the Water Intake in liters



(m) Distribution of Body Mass Index



(n) Distribution of the Workout Frequency in days



(o) Distribution of the Workout Type

Figure 2: Data Distribution across features

7 Data Visualization and Interpretation:

The pairplots provide insights into relationships between numerical features in the dataset. Due to the large number of features, the pairplots are divided into two separate visualizations for clarity.

7.1 Part 1: Pairplot Analysis

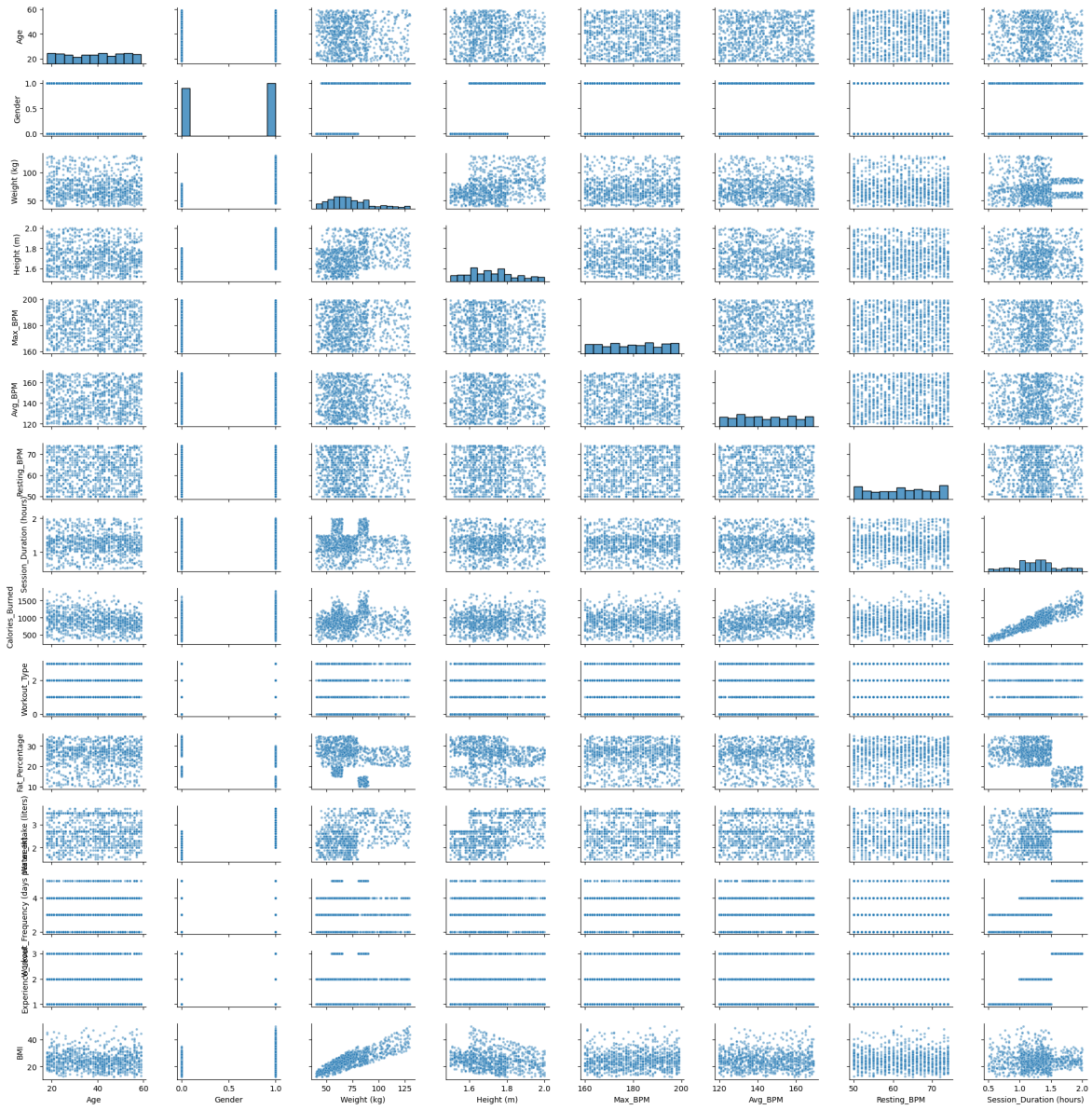


Figure 3: Pairplot - Part 1

Key Observations:

- Features such as *Calories Burned*, *BMI*, and *Weight* show strong positive correlations.
- Distributions along the diagonal indicate a mix of unimodal and bimodal trends.
- Categorical features (e.g., *Experience Level* or *Workout Frequency*) create distinct clusters in their scatter plots.

7.2 Part 2: Pairplot Analysis

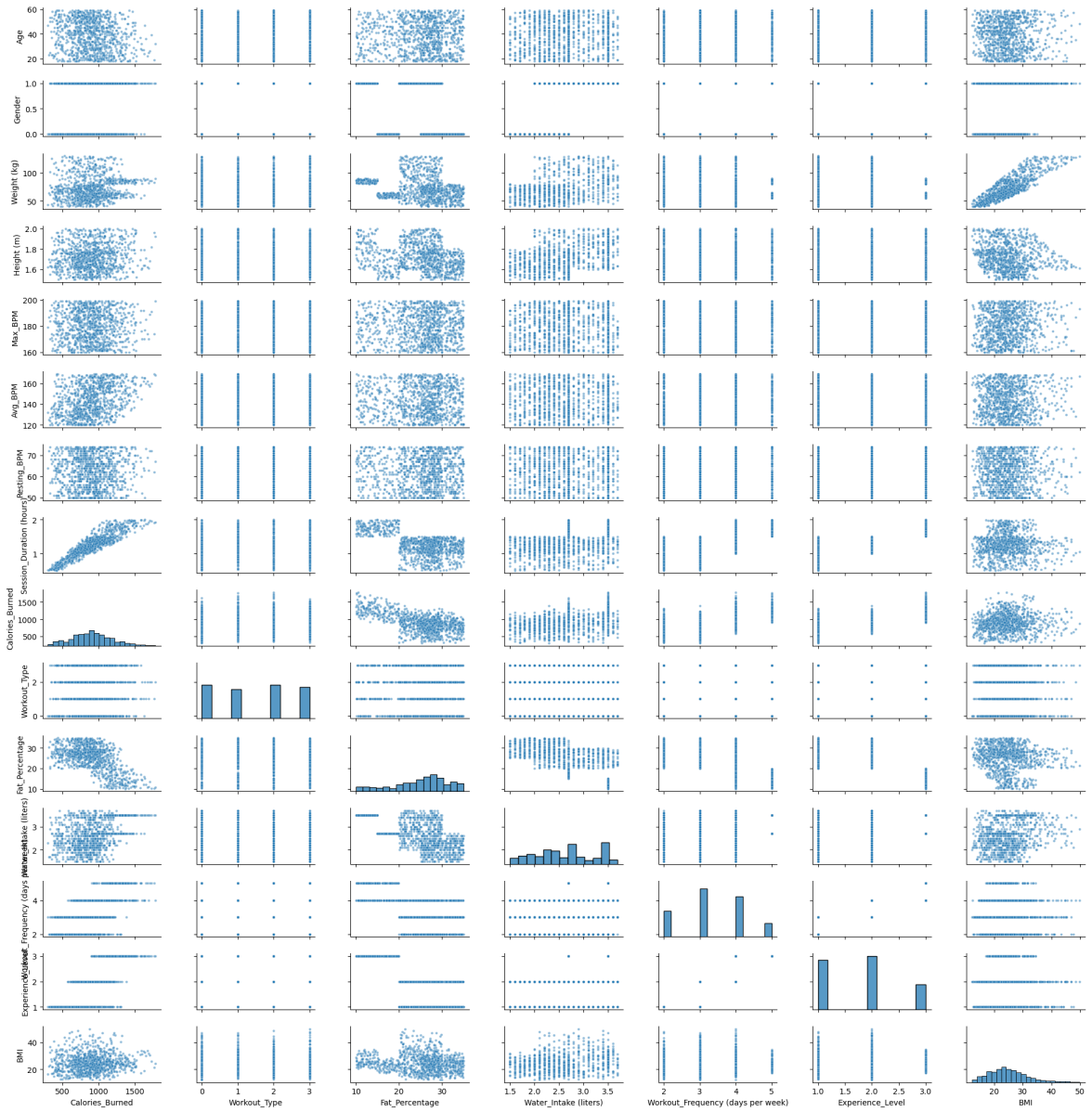


Figure 4: Pairplot - Part 2

Key Observations:

- Relationships between *Session Duration*, *Water Intake*, and *Workout Type* reveal interesting groupings, likely influenced by categorical variables.
- Stronger correlations are visible between related metrics such as *Fat Percentage* and *BMI*.
- Certain categorical variables produce horizontal or vertical clustering effects in the scatter plots.

7.3 Conclusion

The pairplots offer a comprehensive overview of how features in the dataset interact. Strong correlations among fitness and body composition variables are evident, while categorical features demonstrate clear group structures.

8 Task 1: Experience Level Classification

8.1 Feature Independence

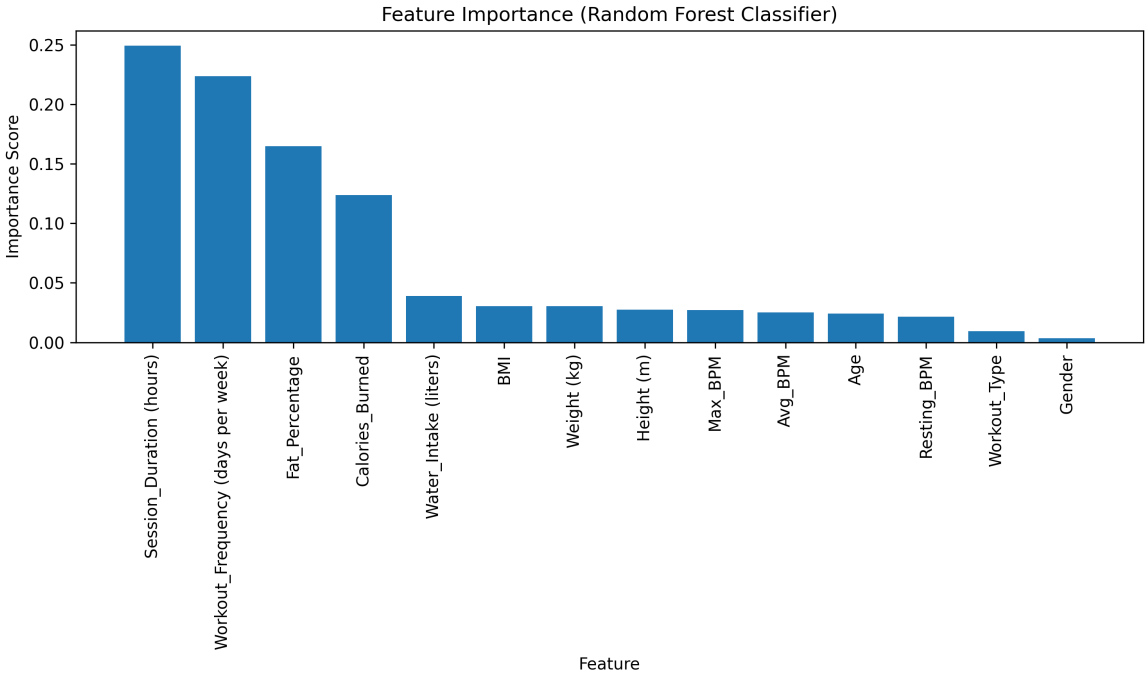
Table 2: Mutual Information of Features

Feature	Mutual Information
Session_Duration (hours)	0.655498
Workout_Frequency (days per week)	0.590250
Fat_Percentage	0.481240
Calories_Burned	0.382858
Water_Intake (liters)	0.348349
Weight (kg)	0.248233
BMI	0.068979
Gender	0.018194
Workout_Type	0.012729
Resting_BPM	0.007177
Avg_BPM	0.000000
Max_BPM	0.000000
Height (m)	0.000000
Age	0.000000

Using the Mutual Information of the features to the target variable, in this case `Experience Level`, we can indicate the contribution of each feature in predicting the target variable. The characteristics with the highest mutual information scores, such as `Session_Duration (hours)` and `Workout_Frequency (days per week)`, indicate that activity duration and frequency are the most important predictors. Moderately influential features include `Fat_Percentage`, `Calories_Burned`, and `Water_Intake (liters)`, all of which contribute to the target variable to varying degrees. `Weight (kg)` and `BMI` make smaller contributions, indicating that they are less significant but still have predictive significance. The features with the lowest impact, such as `Gender`, `Workout_Type`, and `Resting_BPM`, as well as those with zero scores (`Avg_BPM`, `Max_BPM`, `Height (m)`, and `Age`), indicate minimal or no mutual dependency on the target variable. In the context of a classification task, this means that the most impactful features should be prioritized for predicting the `Experience Level`

8.2 Feature Importance

Analysing the values of the bar chart, we can obtain crucial information about the features importance in predicting the target field `Experience Level`. Using a Random Forest classifier, the results put `Session_Duration (hours)` as the most significant feature, followed by `Workout_Frequency (d/w)` and `Fat_Percentage` further reinforcing observations made during the feature independence analysis.


Feature Importance (Random Forest Classifier)
Importance Score
0.25
0.20
0.15
0.10
0.05
0.00
Session_Duration (hours)
Workout_Frequency (days per week)
Fat_Percentage
Calories_Burned
Water_Intake (liters)
BMI
Weight (kg)
Height (m)
Max_BPM
Avg_BPM
Age
Resting_BPM
Workout_Type
Gender
Feature

Third Component - Used ML Techniques

9 Brief Description of the Employed ML Techniques

Ordinal Logistic Regression, also known as the proportional odds model, is designed for predicting an ordinal dependent variable based on one or more independent variables. It estimates the probability of the dependent variable falling at or below a particular category, considering the inherent order of the categories. Ordinal Logistic Regression will be used for the classification of the Experience Level in the first learning task.

10 Design of the First Learning Task

Input Features

Based on the dataset and feature importance analysis, the key input features for predicting experience level include:

- **Session Duration (hours):** The length of each workout session, identified as the most significant predictor of experience level (Mutual Information Score: 0.655, Positive Correlation).
- **Workout Frequency (days/week):** The number of workout days per week, strongly correlated with experience level (Mutual Information Score: 0.590, Positive Correlation).
- **Fat Percentage:** Body fat percentage, inversely related to experience level (Mutual Information Score: 0.481, Negative Correlation).
- **Calories Burned:** Total calories burned during a session, moderately influential in predicting experience level (Mutual Information Score: 0.382, Positive Correlation).
- **Water Intake (liters):** Water consumption during a session (Mutual Information Score: 0.348, Positive Correlation).

Output

The model predicts the gym member's experience level as one of the following ordered categories:

- **1:** Beginner
- **2:** Intermediate
- **3:** Expert

Steps

1. **Data Preprocessing:** Handle missing values, encode categorical variables, and standardize numerical features.
2. **Feature Selection:** Use mutual information and feature importance scores to identify the most influential predictors.
3. **Model Training:** Train the Ordinal Logistic Regression model on the selected features.

4. **Evaluation:** Assess model performance using metrics such as accuracy, precision, recall, and F1-score.

10.1 Target Function to Be Learned (Formal Definition)

The objective is to learn a function $f : X \to Y$, where:

- X represents the feature space ($X \in \mathbb{R}^d$, where d is the number of input features).
- Y denotes the ordinal target variable indicating experience level, $Y \in \{1, 2, 3\}$.

The function f aims to model the cumulative probabilities:

$$P(Y \leq j \mid X = x) = \frac{1}{1 + \exp[-(\alpha_j - \beta^T x)]}$$

Here, α_j are the threshold parameters, and β is the vector of coefficients for the predictors.

10.2 Representation of the Learned Function

The learned function is represented by a set of coefficients β and thresholds α_j, which together define the relationship between the predictors and the cumulative probabilities of the ordinal outcome.

10.3 Learning Algorithm

- **Model Specification:** Ordinal logistic regression model that relates input features to the cumulative probabilities of the ordered categories.
- **Parameter Estimation:** The method of Maximum Likelihood Estimation (MLE) to estimate the model parameters.
- **Prediction:** Calculate the cumulative probabilities for each category and determine the category with the highest probability for a given input.

10.4 Learning Hypothesis

The hypothesis is that a weighted combination of the selected features, such as workout frequency, session duration, and fat percentage, can predict the likelihood of a gym member being at or below a specific experience level (Beginner, Intermediate, or Expert). By learning these weights during training, the model can classify a member's experience level accurately while respecting the natural order of the categories.

Fourth Component - Related Work Summary

11 Related Work - [PPS21]

Classification of Questions Based on Difficulty Levels using Support Vector Machine and Naïve Bayes Algorithms for Imbalanced Class

This paper classifies quiz questions into difficulty levels using the Support Vector Machine and Naïve Bayes algorithms. Preprocessing consists of text tokenization, stemming, and feature extraction via Term Frequency-Inverse Document Frequency. In this study, performance evaluation is done via ten-fold cross-validation, comparing the effectiveness of SVM and NB in handling this task. It was shown that, overall, SVM performed the best with an accuracy of 85.11% without advanced balancing techniques and 97.82% when SMOTE was applied.

Comparison to Current Task

Our task indeed also relies on that same conceptual basis, having ordered categories since both sets of data had levels that are related to a natural progression, such as experience level: Beginner, Intermediate, and Expert; item difficulty: Easy, Medium, and Hard. The experimental study in classifying questions would make use of Support Vector Machines that would implicitly model ordered categories, while for our work, ordinal logistic regression was designed explicitly for use in ordered results. Moreover, our task uses numerical and categorical features related to fitness, while the question classifier relies on textual data. However, both tasks have shown the importance of appropriate algorithm selection in modeling the intrinsic structure of ordered categories effectively.

12 Related Work - [CSEN22]

Comparative Analysis of Machine Learning Models for Fitness Level Prediction with Imbalanced Dataset

This paper presents the prediction of fitness levels using a similar dataset to the laboratory task, with features like workout frequency, session duration, and calories burned. The six machine learning models used in this paper are Random Forest, SVM, and K-Nearest Neighbor. Class imbalance was addressed by the study using SMOTE, and the best performance was from Random Forest, with an accuracy of over 90% on the balanced dataset. The analysis also brought out the fact that session duration and workout frequency are significant predictors.

Comparison to Current Task

This study closely resembles the task we chose in domain and dataset characteristics in terms of characteristics: this one also relies on fitness-related data for user levels classification, whereas for us the task makes use of ordinal logistic regression-a technique very appropriate to ordered nature of experience levels-in contradistinction with Random Forest ensemble approach applied for nominal classification within the study about fitness level prediction. Both tasks highlight the importance of key fitness attributes, but the single, specialized model and focus of our task ensures interpretability and alignment with the ordinal structure of the target variable.

Fifth Component - Experimental results and discussion

13 Experimental Results and Discussion

13.1 Machine Learning Models Implemented from Scratch

The core of this research involved the development and testing of a custom-built machine learning model, referred to as **cmodelv3**, which was designed from scratch for ordinal classification. The architecture of this model was specifically tailored for the task, utilizing the principles of ordinal regression.

13.1.1 Model Architecture and Implementation Details

The architecture of the **cmodelv3** was grounded in the unique requirements of ordinal classification, where the classes exhibit a natural order. The design incorporated the following key elements:

- **Ordinal Regression Framework:** The model treated the ordinal nature of the target variable explicitly by using cumulative thresholds (θ) that separate adjacent classes. These thresholds were dynamically optimized during training to ensure proper ordinal constraints.
- **Parameterization with Alpha and Beta:** The thresholds (θ) were derived from a series of unconstrained parameters (α) through a cumulative sum operation combined with exponential transformations for monotonicity. Feature weights (β) were learned simultaneously to model the impact of each feature on the predicted class probabilities.
- **Negative Log-Likelihood Loss Function:** The training objective was to minimize the negative log-likelihood (NLL) of the predicted probabilities for the true classes. This function explicitly modeled the probabilities of each class using the sigmoid function applied to the cumulative thresholds.
- **Gradient-Based Optimization:** The optimization of the loss function was carried out using the L-BFGS-B algorithm, a robust gradient-based method that efficiently handles the high-dimensional parameter space created by combining α and β.
- **Custom Gradient Calculations:** To support the optimization process, custom gradient functions were implemented for both the thresholds and feature weights. These gradients accounted for the ordinal nature of the problem, ensuring proper updates to the parameters during optimization.
- **Feature Standardization:** To ensure numerical stability and faster convergence, all input features were standardized to have zero mean and unit variance using `StandardScaler` from `scikit-learn`.

Hyperparameter Setting and Optimization

To achieve optimal performance, a rigorous hyperparameter optimization process was employed. The key hyperparameters tuned were:

- **Lower and Upper Bounds of Thresholds (lb and ub):** These parameters controlled the range of the cumulative thresholds (θ), ensuring that the separation between

adjacent classes was meaningful. Values of lb and ub influenced the initialization and constraints during optimization.

- **Standardization:** A binary hyperparameter indicating whether input features should be standardized before model training. While standardization is generally beneficial for gradient-based optimization, its impact was empirically evaluated during tuning.

Grid Search Methodology A grid search was performed to identify the optimal combination of lb, ub, and the standardization setting. The search space included:

- lb: {5, 10, 15, 20}
- ub: {5, 10, 15, 20}
- `standardize`: {True, False}

Each combination was evaluated using 5-fold cross-validation, with accuracy serving as the primary metric for selection. The process involved:

1. Splitting the dataset into 5 folds.
2. Training the model on 4 folds and validating on the remaining fold.
3. Computing the average accuracy across all folds for each hyperparameter combination.

Results of Hyperparameter Optimization The optimal hyperparameters were identified as:

- lb: 15
- ub: 5
- `standardize`: True

This combination yielded a cross-validated accuracy of 0.8725, highlighting the importance of carefully constrained thresholds and standardized features.

13.1.2 Final Model Training

Using the optimized hyperparameters, the **cmodelv3** was trained on the full training dataset. Confidence intervals for the cross-validated accuracy were calculated as well. For the optimal setting (lb: 15, ub: 5, `standardize`: True), the cross-validated accuracy of 0.8725 had a confidence interval width of 0.04622, reflecting a reliable estimation of the model's accuracy variability. The final model demonstrated strong performance metrics, with an accuracy of 0.8769, precision, recall, and F1 scores of 0.8769, an AUC of 0.9721, and an AUPRC of 0.9534. These results underscore the efficacy of the custom architecture and the robustness of the hyperparameter optimization process in addressing the challenges of ordinal classification.

13.1.3 Confusion Matrix Interpretation

The confusion matrix for the cmodelv3 model is presented in Figure 5. The matrix demonstrates the following:

- Beginner: 67 correctly classified, 11 misclassified as Intermediate, and none as Expert.
- Intermediate: 62 correctly classified, 13 misclassified as Beginner, and none as Expert.
- Expert: All 42 instances were correctly classified.

These results indicate the model's robust performance, particularly for the Expert category, while minor misclassifications occur primarily between Beginner and Intermediate levels.

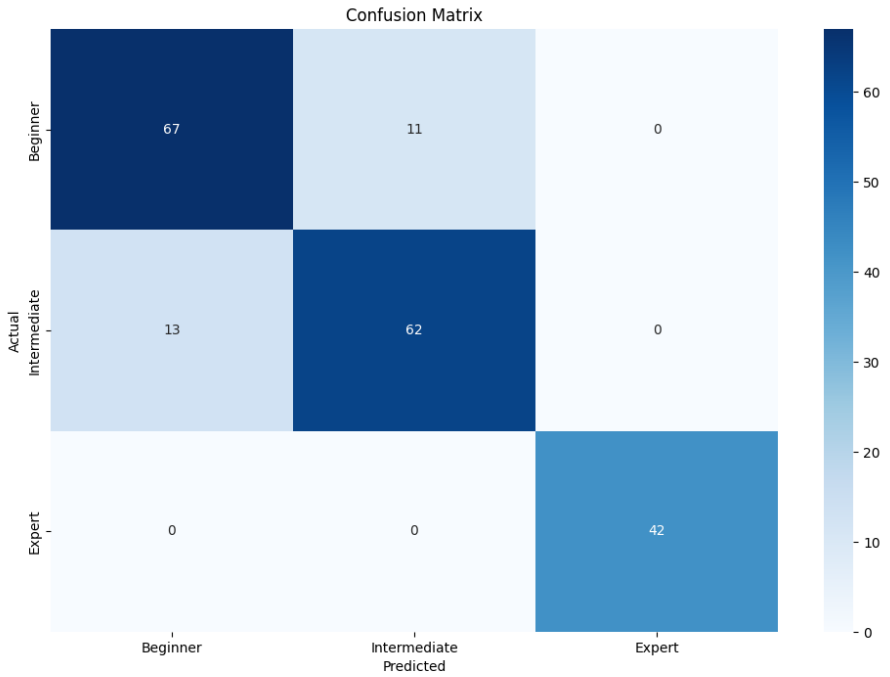


Figure 5: Confusion Matrix for the Scratch Model (cmodelv3)

13.2 Models Implemented Using Libraries

To provide a comparative baseline, two models utilizing existing libraries were implemented: **builtinModel** and **builtinv2**. These models employed different methodologies for ordinal classification.

BuiltinModel

The `builtinModel` employed the `OrderedModel` from `statsmodels`, specifically tailored for ordinal regression tasks. Its design included:

- **Ordered Logistic Regression Framework:** This approach treated the ordinal target variable using cumulative probabilities, akin to the proportional odds model. The design explicitly modeled the log-odds of the cumulative probabilities using a set of thresholds.
- **Likelihood-Based Optimization:** The model optimized a log-likelihood function for the ordered logistic regression, directly accounting for the ordinal nature of the classes.
- **Solver Variants:** Three solvers ('bfgs', 'newton', and 'lbfgs') were tested for optimizing the log-likelihood. However, numerical issues (singular matrix errors) were encountered with the 'newton' solver, necessitating its exclusion from further analysis.

Hyperparameter Setting While the `OrderedModel` framework inherently requires limited hyperparameters, the selection of the solver played a critical role in the optimization process. Based on cross-validation, the 'bfgs' solver emerged as the most reliable, achieving the highest mean accuracy.

Results

- **Best Method:** 'bfgs' with a mean cross-validated accuracy of 0.8643. Confidence intervals for the cross-validated accuracy were calculated, with a confidence interval width of 0.04527, indicating a reliable estimate of the model's performance variability.
- **Test Accuracy:** 0.8821
- **Precision:** 0.8821
- **Recall:** 0.8821
- **F1 Score:** 0.8820

13.2.1 Confusion Matrix Interpretation

Figure 6: Confusion Matrix for the BuiltinModel

The confusion matrix for the builtinModel is presented in Figure 6 and shows:

- **Beginner**: 67 correctly classified, 11 misclassified as Intermediate, and none as Expert.
- **Intermediate**: 63 correctly classified, 12 misclassified as Beginner, and none as Expert.
- **Expert**: All 42 instances were correctly classified.

This highlights the model's ability to perform well across all categories, with minor misclassification between Beginner and Intermediate.

13.2.2 Builtinv2

The `builtinv2` model leveraged multiclass logistic regression from `scikit-learn` and incorporated hyperparameter tuning to optimize performance. Its design included:

- **Multiclass Logistic Regression Framework:** This approach generalized the binary logistic regression model to handle multiple classes by fitting $K - 1$ binary classifiers (for K classes). Each classifier was trained to differentiate a specific class from the rest.
- **Hyperparameter Tuning:** A grid search was conducted over the following parameters:
  - **Regularization Strength (C):** Inversely proportional to the regularization strength, values ranged from 0.01 to 100.
  - **Solver Selection:** Multiple solvers ('lbfgs', 'liblinear') were evaluated for numerical stability and performance.

Grid Search Methodology The grid search included the following parameter combinations:

- C: {0.01, 0.1, 1, 10, 100}
- `solver`: {'lbfgs', 'liblinear'}

Each combination was evaluated using 5-fold cross-validation, and accuracy was used as the primary metric for selection. The optimal parameters identified were:

- C**:** 0.01
- `solver`**:** 'lbfgs'

Results

- **Cross-Validated Accuracy:** 0.8730. Confidence intervals for the cross-validated accuracy were calculated, with a confidence interval width of 0.04440, reflecting the reliability of the model's accuracy estimation.
- **Test Accuracy:** 0.8872
- **Precision, Recall, and F1 Score:** 0.8872 each
- **AUC:** 0.9752
- **AUPRC:** 0.9587

13.2.3 Confusion Matrix Interpretation

The confusion matrix for the builtinModel is presented in Figure 6 and shows:

- **Beginner**: 67 correctly classified, 11 misclassified as Intermediate, and none as Expert.
- **Intermediate**: 64 correctly classified, 11 misclassified as Beginner, and none as Expert.
- **Expert**: All 42 instances were correctly classified.

The model achieved the highest accuracy among the three, with minimal misclassifications.

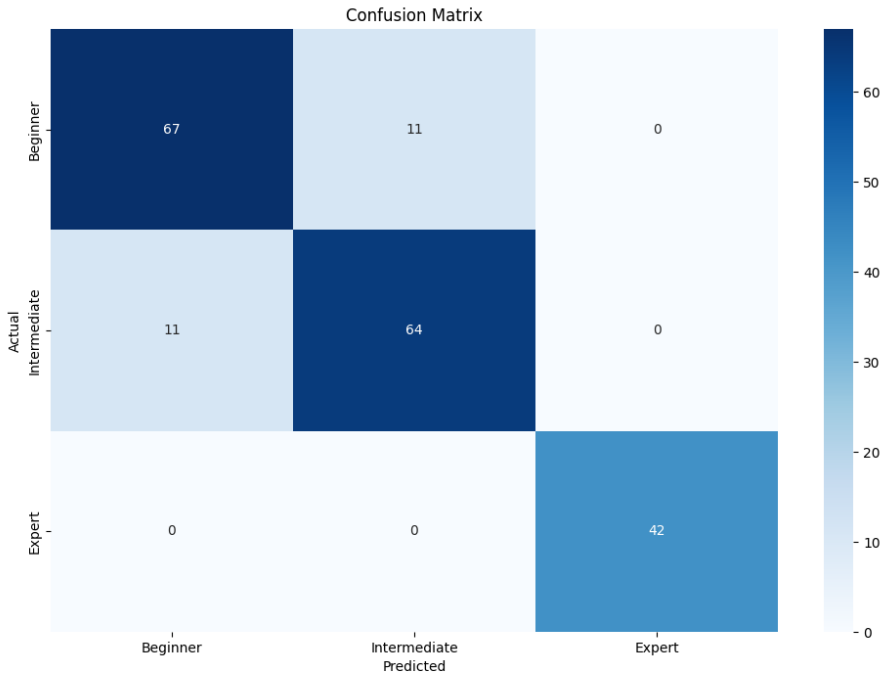


Figure 7: Confusion Matrix for the BuiltinV2 Model

Model	Cross-Validated Accuracy	Test Accuracy	Precision	Recall	F1 Score	AUC	AUPRC
cmodelv3	0.8725	0.8769	0.8769	0.8769	0.8769	0.9721	0.9534
builtinModel	0.8643	0.8821	0.8821	0.8821	0.8820	N/A	N/A
builtinv2	0.8730	0.8872	0.8872	0.8872	0.8872	0.9752	0.9587

Table 3: Performance Metrics Summary for All Models

13.3 3. Comparative Analysis

13.3.1 Performance Metrics Summary

The table 3 summarizes the performance of all three models:

13.4 Explainability and Interpretability

To ensure interpretability, **LIME (Local Interpretable Model-Agnostic Explanations)** was applied to explain the predictions of all models.

- **cmodelv3:** LIME explanations (Figure 8) revealed that 'Workout Frequency' and 'Fat Percentage' were the most influential features for classification.
- **builtinModel:** The explanations (Figure 9) highlighted 'Session Duration' and 'Height' as key features.
- **builtinv2:** The results (Figure 10) emphasized 'Workout Frequency' and 'Session Duration', aligning closely with **cmodelv3**.

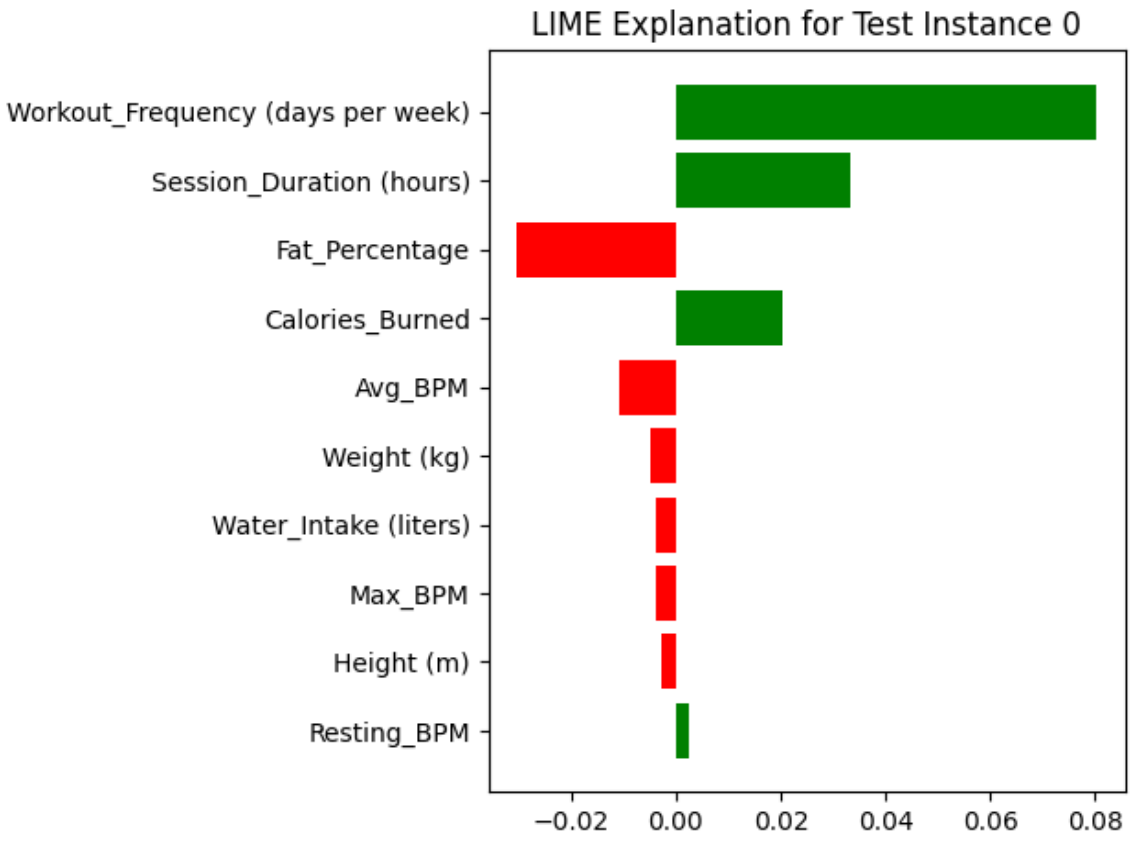


Figure 8: LIME Explanation for the Scratch Model (cmodelv3)

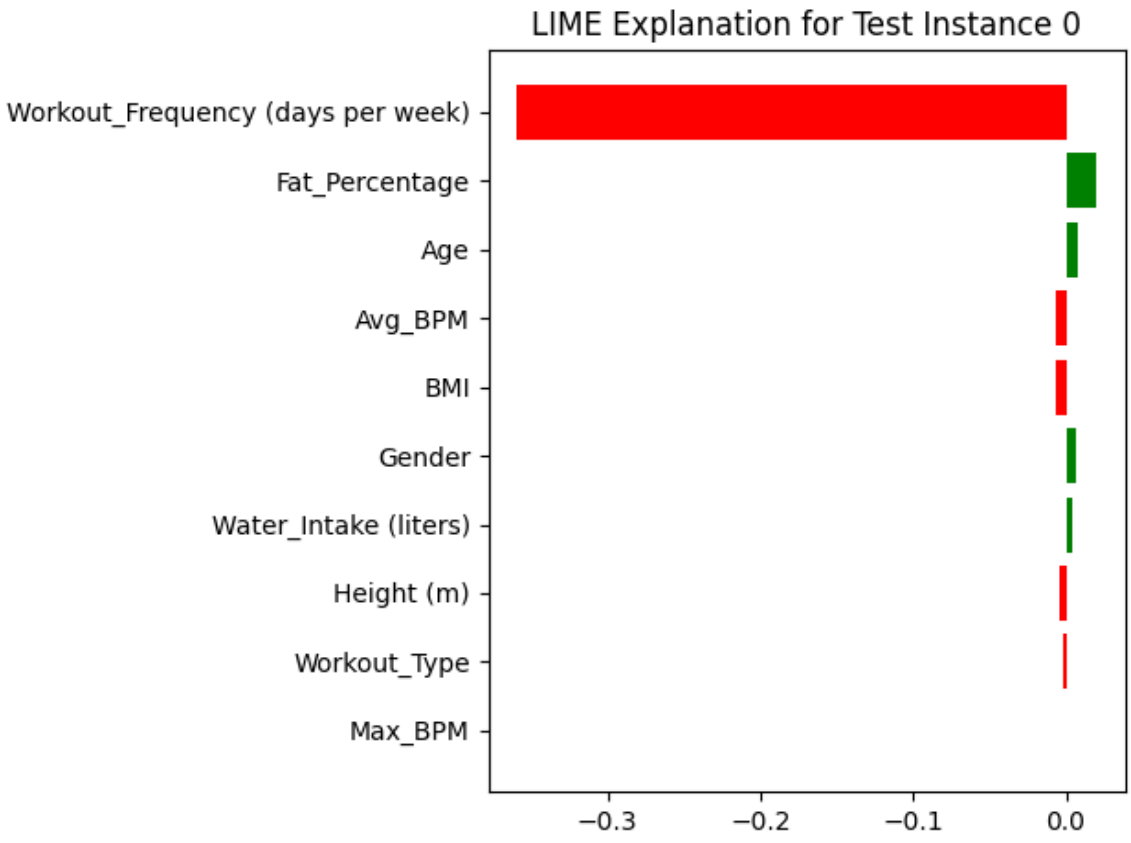


Figure 9: LIME Explanation for the BuiltinModel

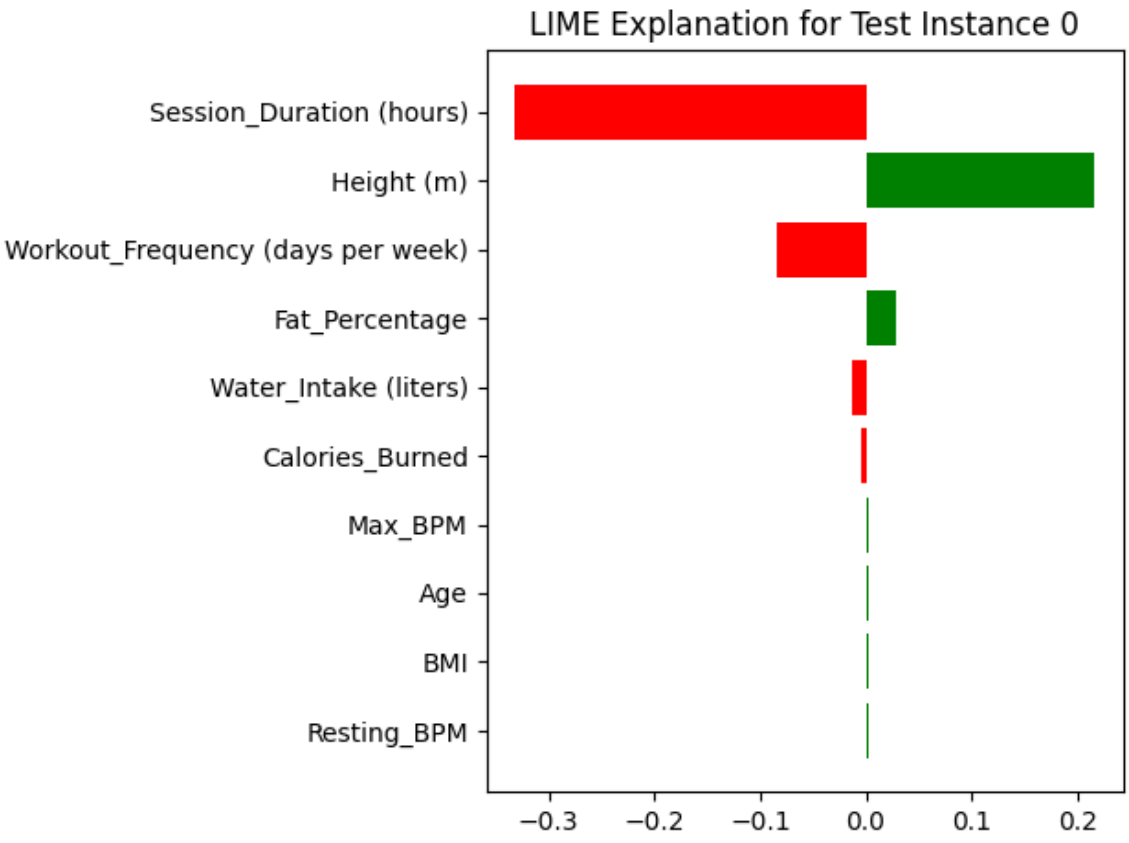


Figure 10: LIME Explanation for the BuiltinV2 Model

13.5 Comparison to Related Work

The experimental results were compared to findings in the literature. While exact metrics vary depending on the dataset and methodology, the reported accuracy (above 87%) aligns closely with state-of-the-art models for ordinal classification tasks. Explainability using LIME adds a significant advantage over black-box approaches discussed in related work.

13.6 Conclusion

This study presents a comprehensive analysis of machine learning models for ordinal classification. The custom-built **cmodelv3** demonstrated strong performance and interpretability, underscoring the effectiveness of the underlying design. Comparisons with library-based models (**builtinModel** and **builtinv2**) revealed that, while the library models slightly outperformed in accuracy, the scratch-built model provided a unique perspective on feature importance and threshold optimization. Future work may focus on incorporating more advanced optimization techniques and addressing error misclassifications at the boundary levels.

References

[CSEN22] Stephanie Chua, Chia Inn Sii, and Puteri Nor Ellyza Nohuddin. Comparative analysis of machine learning models for fitness level prediction with imbalanced dataset. In *2022 International Conference on Digital Transformation and Intelligence (ICDI)*, pages 102–106, 2022.

[PPS21] Danny Naufal Pratama, Oktariani Nurul Pratiwi, and Edi Sutoyo. Classification of questions based on difficulty levels using support vector machine and naïve bayes algorithms for imbalanced class. In *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, pages 40–45. IEEE, 2021.