King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT326: Data Mining
1st Semester 1447 H

Lifestyle-Based Fitness Prediction
Final Report

NAME	ID	Work Distribution
Lama Almubarak	*445200338*	*Classification, Evaluation and Comparison, Data Mining techniques, Data, References*
Waad Alghamdi	*445200230*	*Classification, Findings and Discussion, Evaluation and comparison, Data Mining techniques*
Raghad Alqahtani	*445201226*	*Clustering, Findings and Discussion, Data, Data Mining Task, Problem*
Raneem Aloraini	*445202194*	*Clustering, Evaluation and Comparison, Data preprocessing, Data*

Supervised By: **Dr.** Mashael Sultan Aldayel

1- Problem:

The goal of this project is to **predict individual fitness levels** and **identify lifestyle patterns** that influence wellness. With rising health concerns globally, understanding how factors like sleep, nutrition, and physical activity affect fitness is crucial. This analysis can help promote healthier habits, support early intervention, and guide personalized wellness strategies. The problem is both timely and impactful, as it addresses real-world challenges in preventive healthcare and lifestyle optimization.

2- Data Mining Task:

In our project, we will use two data mining tasks to help us predict individual fitness levels and uncover lifestyle patterns that influence overall wellness: **classification** and **clustering**. For the classification task, we train a supervised model to determine whether a person is fit or not based on a set of health and lifestyle attributes such as age, height, weight, heart rate, blood pressure, sleep hours, nutrition quality, activity index, smoking habits, and gender. This enables the model to learn meaningful patterns that support early identification of unhealthy behaviors and guide personalized wellness recommendations. For the clustering task, the model groups individuals with similar health profiles using unsupervised learning techniques. These clusters may reveal patterns such as highly active individuals with strong nutrition habits, low-activity groups at risk due to poor sleep or smoking, or balanced groups with moderate fitness levels. Such natural groupings help in understanding population segments more effectively and support targeted lifestyle improvement strategies.

3- Data:

- **Source of the dataset: Click_Here**

Number of objects: **2000**
Number of attributes: **11**

- ## Attribute Explanation:

Category	Attribute	Type / Range	Description
Demographic	age	Numerical (Integer, range 18–90 approx.)	The participant's age in years.
	gender	Nominal (M, F)	Participant's gender.
Physical Measurements	height_cm	Numerical (cm)	Participant's height in centimeters.
	weight_kg	Numerical (kg)	Participant's weight in kilograms.
Health Indicators	heart_rate	Numerical (float, approx. 55–100 bpm)	Resting heart rate measurement.
	blood_pressure	Numerical (float, systolic value)	Indicates participant's systolic blood pressure.
	is_fit	Binary (0= Not fit, 1= Fit)	Indicates whether the participant is classified as physically fit.
Lifestyle Habits	sleep_hours	Numerical (float, 1.4–11.3 hours)	Average daily sleep duration.
	nutrition_quality	Numerical (float, 1–10 scale)	Quality of diet based on scoring metric.
	activity_index	Numerical (float, 1–10 scale)	Measures physical activity level.
	smokes	Nominal (yes / 1, no / 0)	Indicates whether the participant smokes.

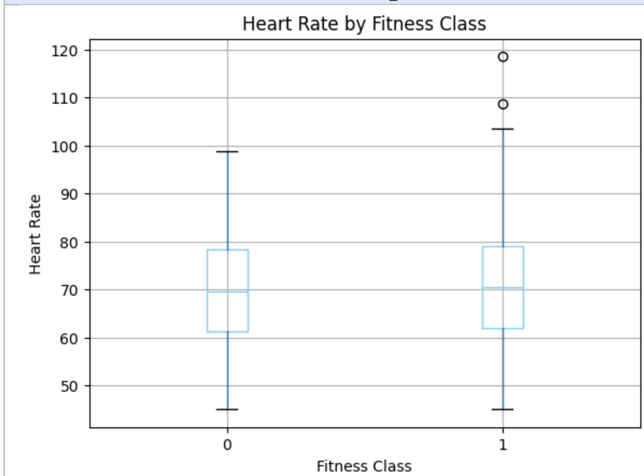
- ## Missing Values:

Using the function "isna()",we discovered only 1 attribute that had a missing value which is "sleep_hours" with 79 missing value ‹the missing values in the "sleep_hours" column were likely caused by gaps in the data collection process.

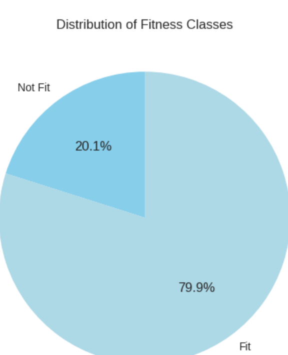
- ### Outliers:

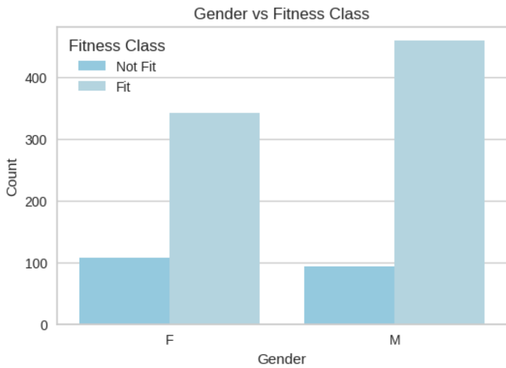
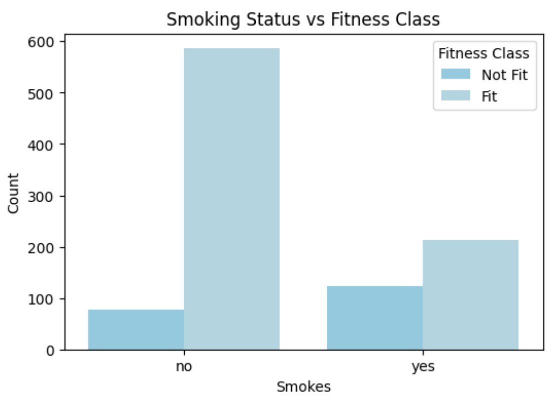
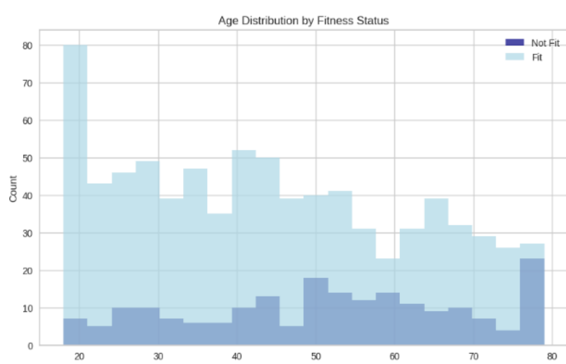
Attribute	Number of outliers
age	0
height_cm	0
weight_kg	8
heart_rate	45
blood_pressure	50
sleep_hours	0
nutrition_quality	0
activity_index	21

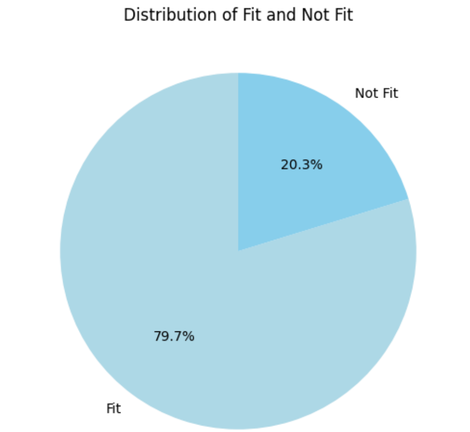
- ### Boxplot:

Graph	Description
	The boxplot shows the difference in heart-rate distribution between “Fit” and “Not Fit” individuals. Fit participants tend to have lower and more stable heart rates, while Not Fit individuals show higher variability and some outliers.

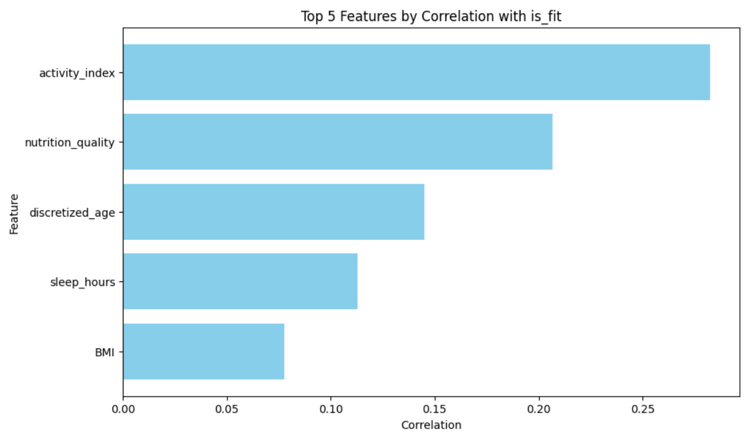
- ### Plotting Methods:

Graph	Description
Pie Chart 	The pie chart shows the proportion of individuals classified as “Fit” and “Not Fit.” It helps visualize class imbalance, where the majority of participants fall under the “Fit” category.

Bar Chart 	This bar chart compares fitness levels between males and females. It helps observe how fitness class distribution varies by gender and highlights group differences.
Bar Chart 	The bar chart visualizes the relationship between smoking status and fitness classification. It shows whether smokers or non-smokers tend to fall more into the “Fit” or “Not Fit” class.
Histogram 	The histogram displays how age varies between “Fit” and “Not Fit” participants. It shows concentration of ages and helps identify any age-related patterns in fitness levels.
Pie Chart	This pie chart illustrates the final distribution of the target class after preprocessing. It confirms the percentage of “Fit” and “Not Fit” individuals in the cleaned dataset.



Bar Chart



This bar chart shows the top five attributes most strongly correlated with the target class. It helps identify the most important predictors to include in the model.

4- Data preprocessing:

- **Checking for missing values:**
  To maintain data consistency, the missing values in the sleep_hours column were filled using the mean imputation method. This approach ensures that all 11 attributes remain complete and ready for accurate analysis and modeling.

After:

Attribute:	Missing values in each column:
age	0
height_cm	0
weight_kg	0
heart_rate	0

blood_pressure	0
sleep_hours	0
nutrition_quality	0
activity_index	0
smokes	0
gender	0
is_fit	0

- **Detecting and removing the Outliers:**

The Z-score method was applied to detect and remove outliers from key numeric attributes, including weight, heart rate, blood pressure, sleep hours, and activity index. Records with Z-scores beyond ±2 were treated as extreme values and removed, resulting in a balanced and reliable dataset of 992 rows, prepared for subsequent analysis and model training.

Data before removing Outlier:

```
     age  height_cm  weight_kg  heart_rate  blood_pressure  sleep_hours
0     56        152         65        69.6           117.0          NaN
1     69        186         95        60.8           114.8          7.5
2     32        189         83        60.2           130.1          7.0
3     60        175         99        58.1           115.8          8.0
4     38        188         57        81.2           110.6          6.6
..   ...        ...        ...         ...             ...          ...
995   36        193        101        88.1           132.9          6.1
996   79        160         57        55.4           102.4          6.1
997   42        158        117        74.7           135.2          9.2
998   76        162         63        82.4           102.0          8.0
999   51        171         96        79.6           104.3          5.8

     nutrition_quality  activity_index smokes gender  is_fit
0                 2.37            3.97     no      F       1
1                 8.77            3.19     no      F       1
2                 6.18            3.68     no      M       1
3                 9.95            4.83    yes      F       1
4                 8.47            4.96     no      M       1
..                 ...             ...    ...    ...     ...
995               3.06            2.42    yes      M       0
996               6.31            1.19    yes      M       0
997               9.43            1.13     no      M       0
998               5.11            1.85     no      F       0
999               0.17            1.93     no      F       0

[1000 rows x 11 columns]
```

Data after removing Outlier:

```
DataFrame after removing outliers from each column:
     age  height_cm  weight_kg  heart_rate  blood_pressure  sleep_hours
0     56        152         65        69.6           117.0          NaN
1     69        186         95        60.8           114.8          7.5
2     32        189         83        60.2           130.1          7.0
3     60        175         99        58.1           115.8          8.0
4     38        188         57        81.2           110.6          6.6
..   ...        ...        ...         ...             ...          ...
987   36        193        101        88.1           132.9          6.1
988   79        160         57        55.4           102.4          6.1
989   42        158        117        74.7           135.2          9.2
990   76        162         63        82.4           102.0          8.0
991   51        171         96        79.6           104.3          5.8

     nutrition_quality  activity_index smokes gender  is_fit
0                 2.37            3.97     no      F       1
1                 8.77            3.19     no      F       1
2                 6.18            3.68     no      M       1
3                 9.95            4.83    yes      F       1
4                 8.47            4.96     no      M       1
..                 ...             ...    ...    ...     ...
987               3.06            2.42    yes      M       0
988               6.31            1.19    yes      M       0
989               9.43            1.13     no      M       0
990               5.11            1.85     no      F       0
991               0.17            1.93     no      F       0

[992 rows x 11 columns]
```

- **Transformation :BMI Aggregation**

The height and weight attributes were combined to create a BMI feature using the standard formula. This feature simplified the dataset by reducing dimensionality while providing a more meaningful health metric that shows strong correlation with fitness outcomes.

Updated Data with BMI Column:

	height_cm	weight_kg	BMI
0	152	65	28.13
1	186	95	27.46
2	189	83	23.24
3	175	99	32.33
4	188	57	16.13
...	...	...	...
987	193	101	27.11
988	160	57	22.27
989	158	117	46.87
990	162	63	24.01
991	171	96	32.83

992 rows × 3 columns

- ## Normalization :

Min-Max normalization was applied to all numerical features, rescaling them to a 0-1 range to ensure consistent scaling, reduce variable dominance, and improve machine learning model performance.

Normalized Data:

	height_cm	weight_kg	heart_rate	blood_pressure	sleep_hours	activity_index	nutrition_quality	BMI
0	0.040816	0.159091	0.334239	0.332512	0.453501	0.743719	0.231621	0.186852
1	0.734694	0.295455	0.214674	0.305419	0.437500	0.547739	0.876133	0.180168
2	0.795918	0.240909	0.206522	0.493842	0.375000	0.670854	0.615307	0.138069
3	0.510204	0.313636	0.177989	0.317734	0.500000	0.959799	0.994965	0.228751
4	0.775510	0.122727	0.491848	0.253695	0.325000	0.992462	0.845921	0.067139
...	...	...	...	...	...	...	...	...
987	0.877551	0.322727	0.585598	0.528325	0.262500	0.354271	0.301108	0.176676
988	0.204082	0.122727	0.141304	0.152709	0.262500	0.045226	0.628399	0.128392
989	0.163265	0.395455	0.403533	0.556650	0.650000	0.030151	0.942598	0.373803
990	0.244898	0.150000	0.508152	0.147783	0.500000	0.211055	0.507553	0.145750
991	0.428571	0.300000	0.470109	0.176108	0.225000	0.231156	0.010070	0.233739

992 rows × 8 columns

- ## Discretization:

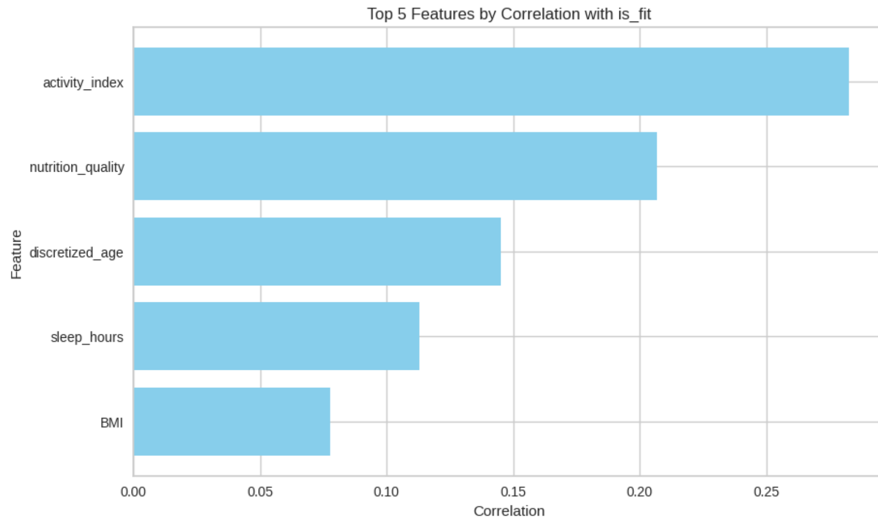
The age attribute was categorized into three life stages using equal-width binning: Youth (0-24), Adult (25-59), and Senior (60+), encoded as 0, 1, and 2 respectively. This discretization simplifies analysis, reduces variability, and enhances pattern recognition with the target variable is_fit.

```
Original DataFrame with discretized column:
     age  discretized_age
0     56                1
1     69                2
2     32                0
3     60                2
4     38                0
..   ...              ...
987   36                0
988   79                2
989   42                1
990   76                2
991   51                1

[992 rows x 2 columns]
```

- ## Feature Selection:

Based on correlation analysis with the target variable is_fit, the top five influential features were selected: activity_index, nutrition_quality, discretized_age, sleep_hours, and BMI. This feature selection enhances model efficiency and interpretability by focusing on the most relevant attributes.




```
RAW shape: (1000, 11)
PREPROCESSED shape: (992, 12)
```

=== SNAPSHOT • RAW DATASET (first 5 rows) ===

	age	height_cm	weight_kg	heart_rate	blood_pressure	sleep_hours	nutrition_quality	activity_index	smokes	gender	is_fit
0	56	152	65	69.6	117.0	NaN	2.37	3.97	no	F	1
1	69	186	95	60.8	114.8	7.5	8.77	3.19	no	F	1
2	32	189	83	60.2	130.1	7.0	6.18	3.68	no	M	1
3	60	175	99	58.1	115.8	8.0	9.95	4.83	yes	F	1
4	38	188	57	81.2	110.6	6.6	8.47	4.96	no	M	1

=== SNAPSHOT • PREPROCESSED DATASET (first 5 rows) ===

	discretized_age	height_cm	weight_kg	heart_rate	blood_pressure	sleep_hours	nutrition_quality	activity_index	BMI	is_fit
0	1	0.040816	0.159091	0.334239	0.332512	0.453501	0.231621	0.743719	0.186852	1
1	2	0.734694	0.295455	0.214674	0.305419	0.437500	0.876133	0.547739	0.180168	1
2	0	0.795918	0.240909	0.206522	0.493842	0.375000	0.615307	0.670854	0.138069	1
3	2	0.510204	0.313636	0.177989	0.317734	0.500000	0.994965	0.959799	0.228751	1
4	0	0.775510	0.122727	0.491848	0.253695	0.325000	0.845921	0.992462	0.067139	1

Snapshot of Raw dataset VS. Processed dataset

5- Data Mining Technique:

We applied both supervised and unsupervised learning to our data using classification and clustering techniques.

▪ Classification

We applied Decision Tree classification to predict the target variable is_fit using five features from our dataset (activity_index, nutrition_quality, discretized_age, sleep_hours, BMI). We tested three train test splits (90%, 80%, 70%) and used two splitting criteria (Entropy (IG) and Gini Index). Model performance was evaluated using accuracy, precision, Sensitivity (recall), specificity, and confusion matrices, allowing us to compare the different configurations and determine which split and criterion produced the most accurate fitness predictions.

○ Python Packages Used for Classification

1. scikit-learn (sklearn)

- **DecisionTreeClassifier:** Builds the decision tree using entropy or gini splits
- **train_test_split:** Divides the dataset into training and test sets accuracy_score, recall_score, precision_score, confusion_matrix: Used to evaluate performance
- **StandardScaler:** Normalizes the numeric features
- **LabelEncoder** (if needed): Converts labels into numeric values

2. yellowbrick

- **SilhouetteVisualizer:** Used for clustering evaluation (not directly part of classification)

These tools help build, train, evaluate, and visualize the performance of Decision Tree models for fitness prediction.

▪ Clustering

In our clustering analysis, we excluded the "is_fit" class label as clustering is unsupervised and relies solely on feature similarities. We used health attributes like age, height, weight, heart rate, blood pressure, sleep hours, nutrition quality, activity index, smoking status, and gender (all converted to scaled numeric form).

We applied the K-means algorithm to group data points into clusters based on feature similarity, iteratively adjusting cluster centers until stable groups formed.

For validation, we used:

- Silhouette scores to measure cluster cohesion and separation
- The elbow method with within-cluster sum of squares (WSS) to determine the optimal number of clusters
- Compared multiple cluster sizes (K=5,6,7) to identify the most effective grouping

The analysis revealed distinct health profiles that can inform personalized fitness recommendations.

○ Python Packages Used for Clustering:

1. **scikit-learn (sklearn):**
    - **K-Means:** The main algorithm for clustering
    - **StandardScaler:** For normalizing/standardizing features
    - **silhouette_score:** To evaluate cluster quality
2. **pandas & numpy:**
    - Data manipulation and handling
    - Converting data into suitable formats for clustering
3. **yellowbrick:**
    - Enhanced visualization tools
    - Silhouette visualizer for cluster analysis

These packages work together to preprocess data, perform clustering, evaluate results, and visualize the findings effectively.

6- Evaluation and Comparison:

▪ Classification

Classification was applied to predict depression in individuals based on features in the dataset. The Decision Tree algorithm was employed due to its interpretability and efficiency in handling categorical and numerical data. Two attribute selection measures: Information Gain (Entropy) and Gini Index, were used to construct and evaluate the model. The data was split into three distinct partitions for training and testing: 90-10, 80-20, and 70-30. This ensures robust evaluation of the model's performance across different configurations.

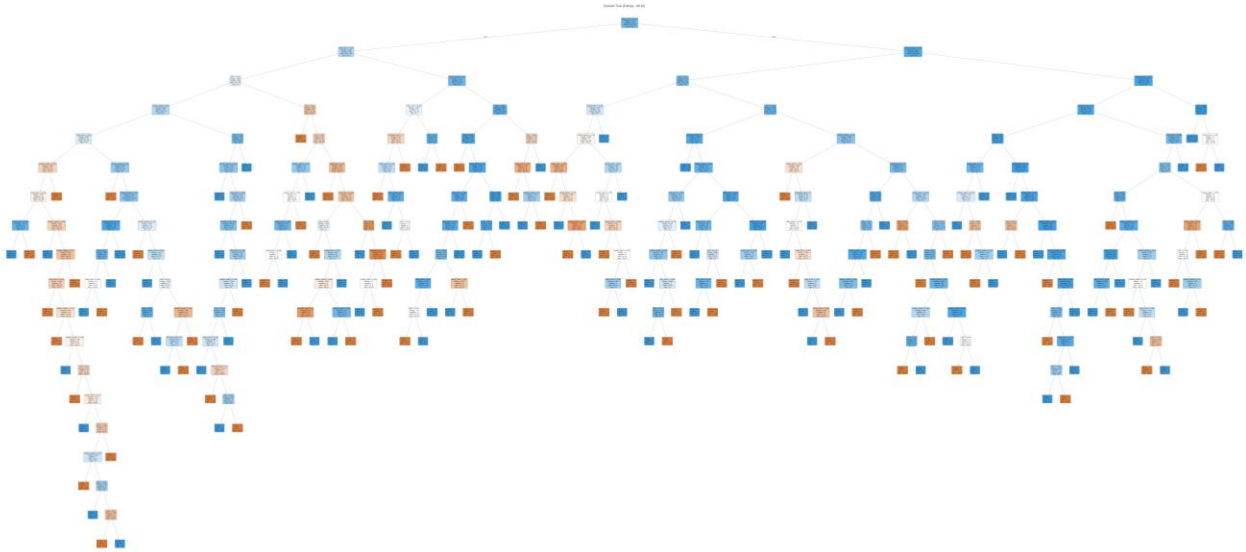
1. Split 90-10

-Entropy:

- Accuracy: 67%
- Sensitivity: 79%
- Specificity: 23%
- Precision: 78%
- Error Rate: 33%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	62	16
Negative	17	5



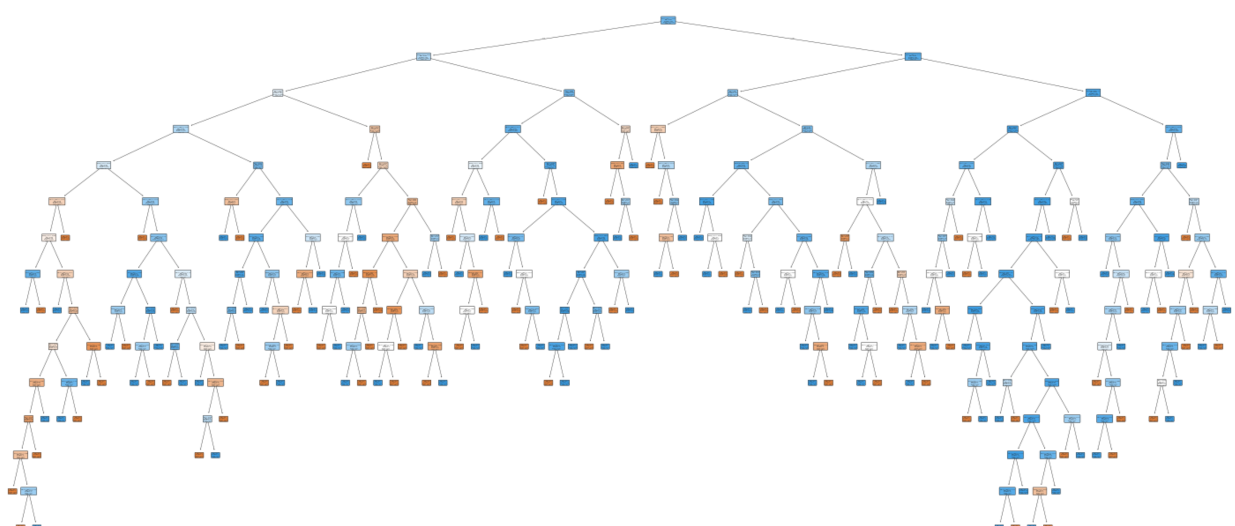
-Gini:

- Accuracy: 65%
- Sensitivity (Recall): 77%
- Specificity: 23%
- Precision: 78%
- Error Rate: 35%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	60	18
Negative	17	5

Decision Tree (Gini - 90–10)



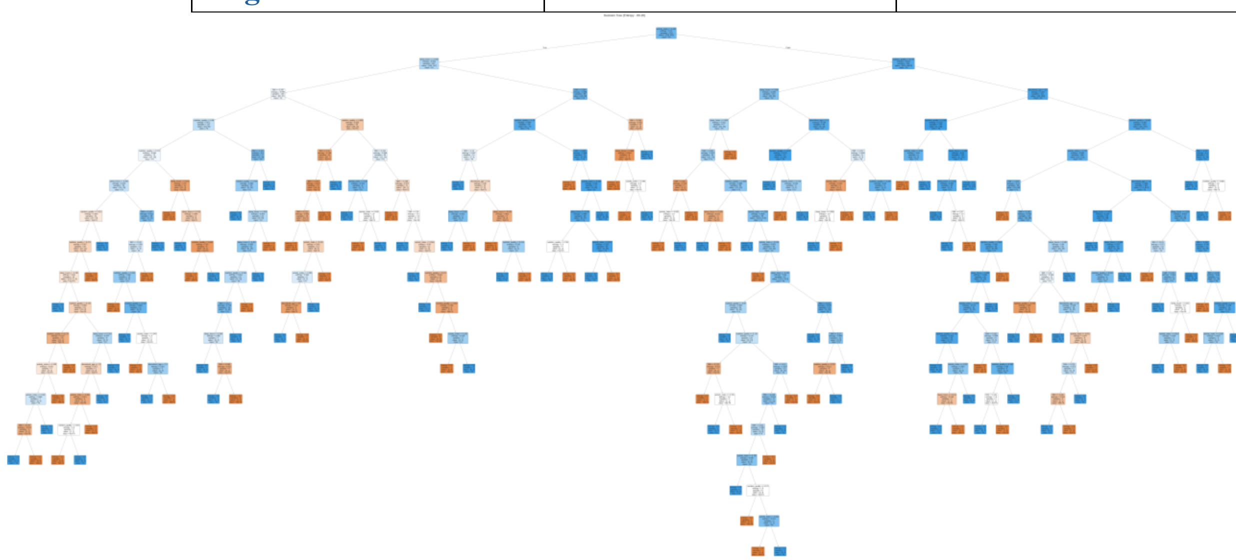
2. Split 80-20

-Entropy:

- Accuracy: 69%
- Sensitivity: 85%
- Specificity: 9%
- Precision: 77%
- Error Rate: 31%

Confusion Matrix:

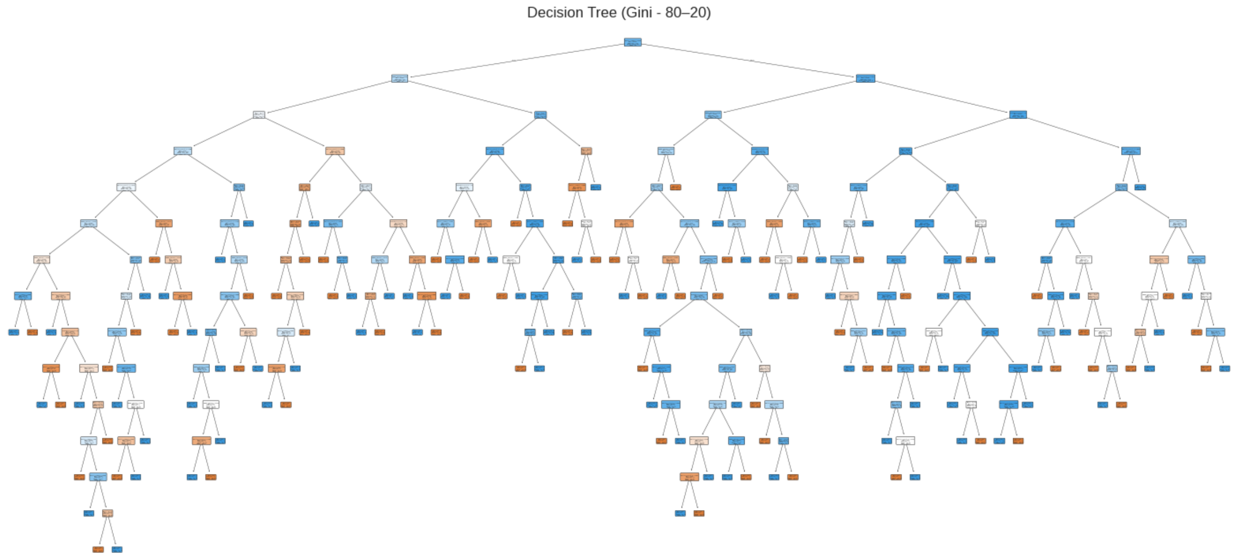
Actual \ Predicted	Positive	Negative
Positive	133	23
Negative	39	4



-Gini:
• Accuracy: 70%
• Sensitivity (Recall): 83%
• Specificity: 23%
• Precision: 80%
• Error Rate: 30%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	129	27
Negative	33	10



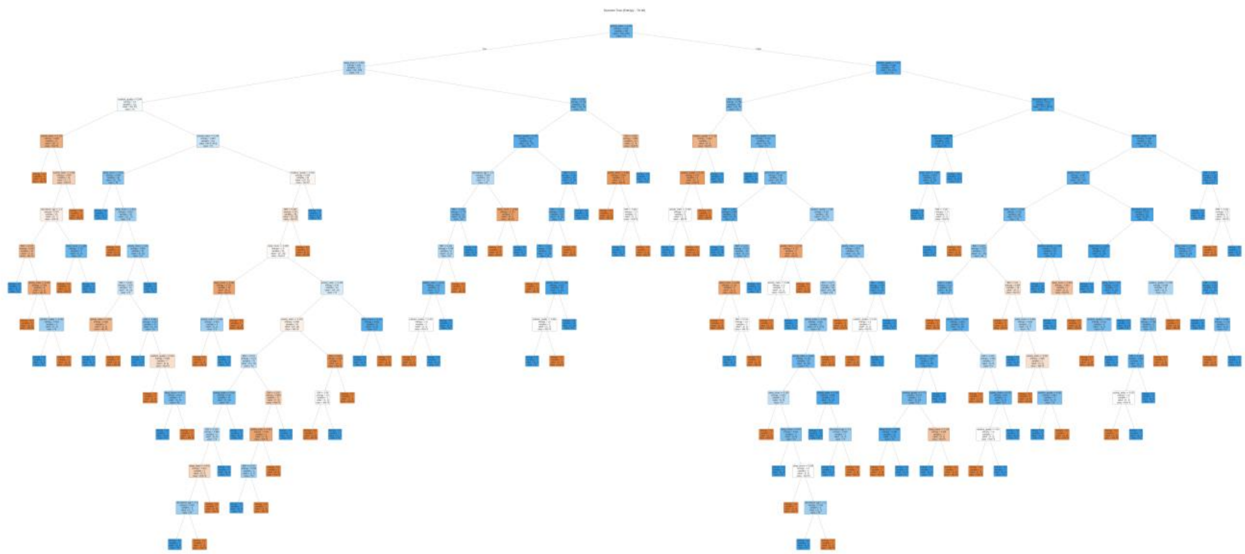
3. Split 70-30

-Entropy:

- Accuracy: 70%
- Sensitivity: 86%
- Specificity: 16%
- Precision: 78%
- Error Rate: 30%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	198	32
Negative	57	11



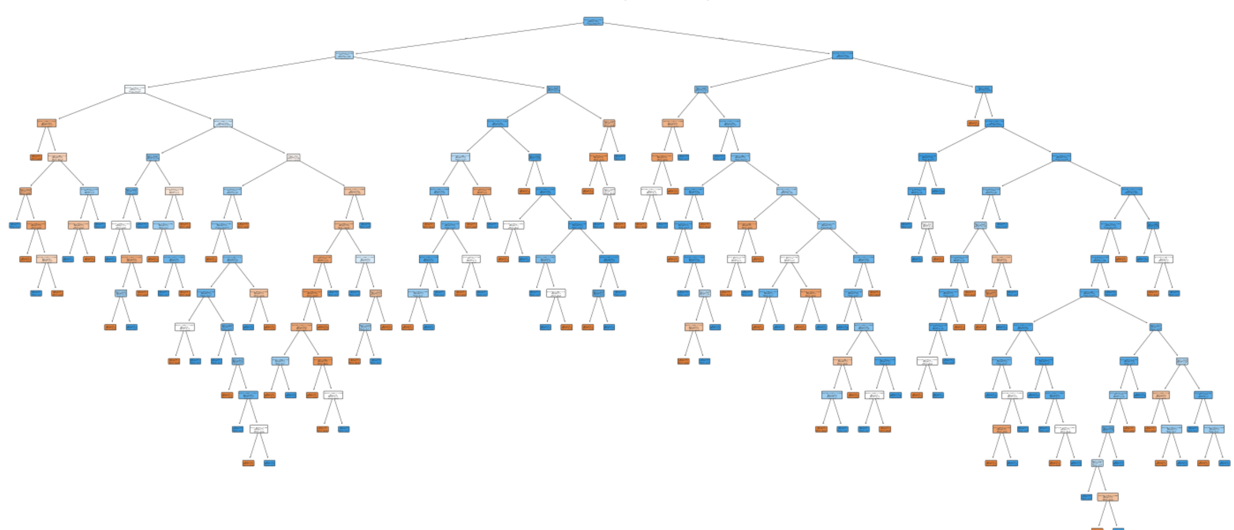
-Gini:

- Accuracy: 72%
- Sensitivity (Recall): 85%
- Specificity: 28%
- Precision: 80%
- Error Rate: 28%

Confusion Matrix:

Actual \ Predicted	Positive	Negative
Positive	196	34
Negative	49	19

Decision Tree (Gini - 70–30)



Performance metrics summary

Entropy:

Metric	90–10 Split	80–20 Split	70–30 Split
Accuracy	67%	69%	70%
Sensitivity (Recall)	79%	85%	86%
Specificity	23%	9%	16%
Precision	78%	77%	78%
Error Rate	33%	31%	30%

Gini:

Metric	90–10 Split	80–20 Split	70–30 Split
Accuracy	65%	70%	72%
Sensitivity (Recall)	77%	83%	85%
Specificity	23%	23%	28%
Precision	78%	80%	80%
Error Rate	35%	30%	28%

▪ Clustering

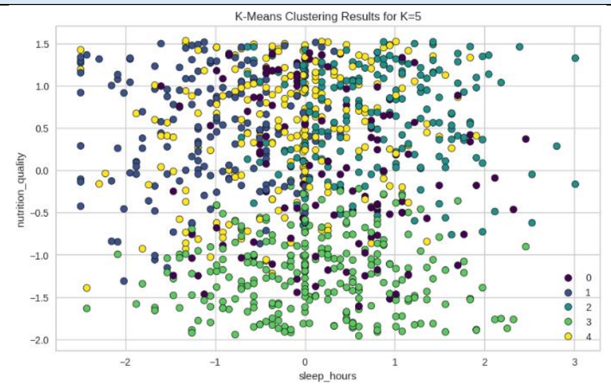
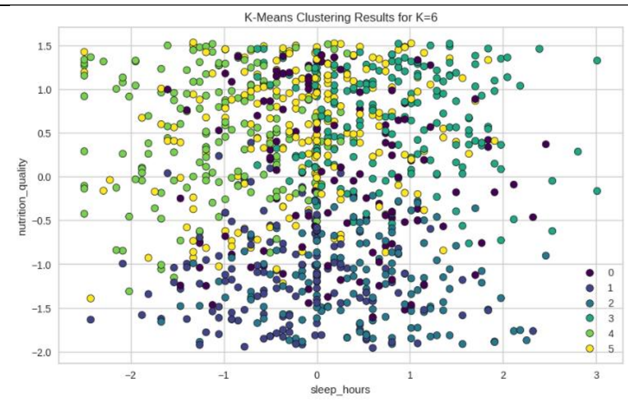
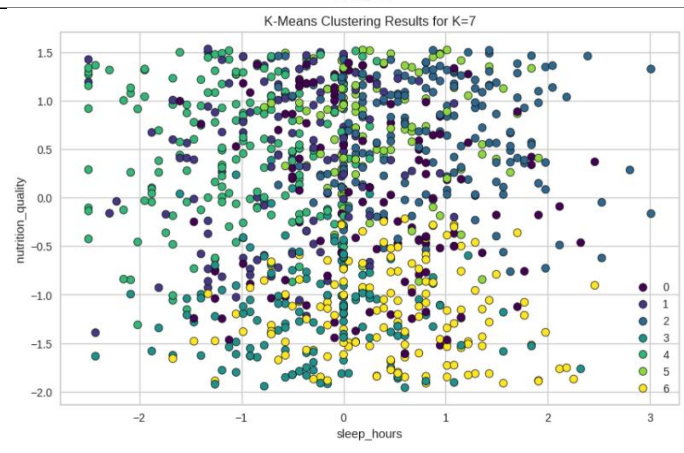
Clustering was applied to group individuals based on health characteristics including age, height, weight, heart rate, blood pressure, sleep hours, nutrition quality, activity index, smoking status, and gender. The "is_fit" label was excluded as clustering is unsupervised learning.

-Algorithm: K-means clustering
-Evaluation Metrics:

- Silhouette Score (cluster cohesion and separation)
- WCSS - Within-Cluster Sum of Squares (cluster compactness)

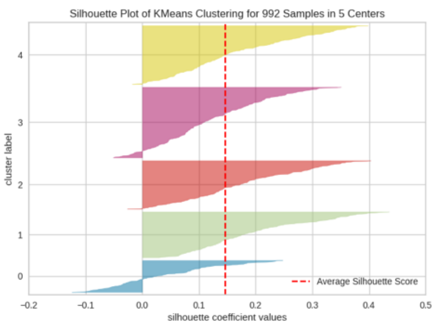
-Cluster Numbers Tested: K = 5, 6, 7

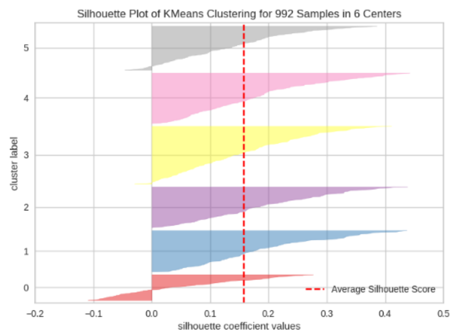
Clustering Trial

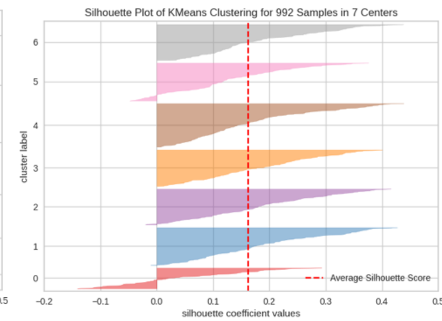
K Value	Scatter Plot	Description
K = 5		○ 5 clear clusters with some overlap ○ Broad and general groupings
K = 6		○ Better segmentation with balanced clusters ○ Less overlap between groups
K = 7		○ Excessive data segmentation ○ Very small clusters

Recommendation:

K= 6 showed best segmentation with more balanced group sizes.







- **K = 5**
  - **Cluster Centers**
    - The five cluster centroids are spread across the feature space, representing broad groupings of the data.
    - The clusters appear to capture general patterns, but some centers suggest mixed characteristics within the groups.
  - **Cluster Labels**

    The data points are divided into five clusters, but some clusters are noticeably larger than others.
  - **Silhouette Score**
    - The average silhouette score (=0.1463) reflects moderate clustering quality**.**
    - Several clusters show overlap, and some points may not fit well within their assigned groups.
- **K = 6**
  - **Cluster Centers**
    - Introducing a sixth cluster results in more refined centroids.
    - Some of the larger clusters from K = 5 split into more specific and coherent subgroups.
  - **Cluster Labels**

    The labels show a more balanced distribution across the six clusters.
  - **Silhouette Score**

- The silhouette score slightly increases (=0.1579), but the improvement is minimal.
- While clusters are more detailed, the overall quality of separation does not significantly exceed K = 5**.**

- ### K = 7
  - ### Cluster Centers

    - Adding a seventh cluster creates even more specialized and fine-grained centroids
    - Some groups become smaller and more specific, capturing subtle patterns in the data.

  - ### Cluster Labels

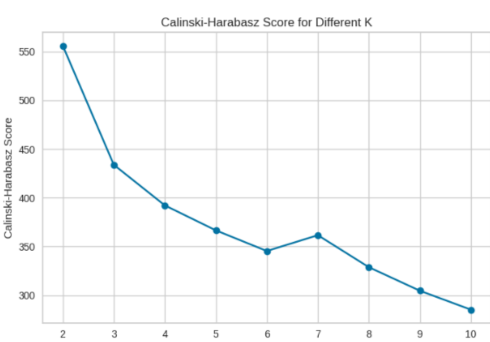
    The labels reveal well-defined clusters, with clearer boundaries and less internal variation**.**

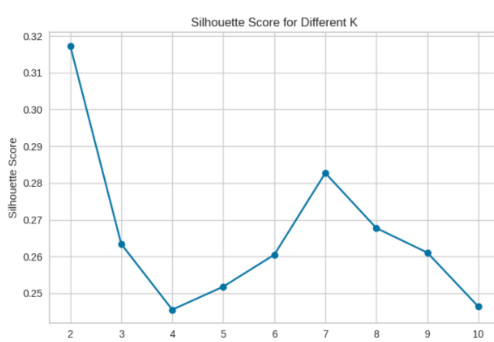
  - ### Silhouette Score

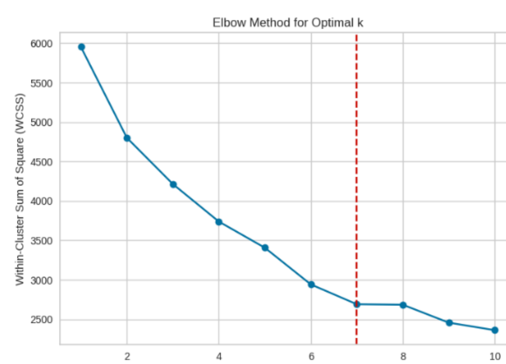
    - The average silhouette score is the highest among the three models (=0.1620).
    - Therefore**,** K = 7 provides the best clustering quality out of the tested values.

Recommendation:

K = 7 is the most suitable choice, as it achieves the highest silhouette score and provides the best-defined clusters.







The results for each trial are summarized below:

Methode	K = 5	K = 6	K = 7
Average Silhouette score	0.1463	0.1579	0.1620
Within-Cluster Sum of Square	3313.06	3144.71	2690.74
Calinski-Harabasz Score Analysis	366.60	345.39	361.64

Comparison and Optimal Cluster Selection

Average Silhouette score:

- The highest silhouette value is observed at **K = 7 (0.1620)**, indicating better-defined clusters compared to K = 5 and K = 6.
- The score increases steadily from K = 5 < K = 6 < K = 7, suggesting consistent improvement in cluster separation as more clusters are added.

Total Within-Cluster Sum of Squares (WCSS):

Reduction between cluster counts:

- **K = 5 to K = 6:**
  3313.06 - 3144.71 = **168.35**
- **K = 6 to K = 7:**
  3144.71 - 2690.74 = **453.97**

The largest decrease occurs between K = 6 and K = 7, indicating that adding the 7th cluster significantly improves compactness, therefore K = 7 provides tighter, more cohesive clusters rather than over-segmentation.

Calinski–Harabasz Score:

- **K = 5:** 366.60
- **K = 6:** 345.39
- **K = 7:** 361.64

The highest CH score is at K = 5, but K = 7 remains very close and still strong.

Optimal Number of Clusters

-Based on the Silhouette Score:

K = 7 provides the best overall separation.

-Based on the Elbow Method (WCSS):

The most meaningful improvement appears at K = 7, supporting the choice of a more refined cluster structure.

-Based on the Calinski–Harabasz Score:

While K = 5 is highest, K = 7 maintains a high score and aligns better with the other metrics

7- Findings and Discussion:

This section presents all results obtained from the classification (Decision Trees) and clustering (K-Means) techniques applied to the fitness dataset. It explains the meaning of the results, compares the performance of all models, identifies the best-performing configuration, and discusses whether the extracted knowledge is meaningful in the context of the study.

Classification

Decision Tree classification was applied using two attribute selection measures:

- Entropy (Information Gain)
- Gini Index

Each measure was evaluated under different training–testing splits (90–10, 80–20, 70–30). The models were compared using accuracy, sensitivity, specificity, precision, and error rate. Visualizations (confusion matrices and final tree plots) were also used to interpret the results.

Entropy (Information Gain) Results
The performance of Entropy models shows clear patterns:

Split	Accuracy	Sensitivity	Specificity
90–10	67%	79%	23%
80–20	69%	85%	9%

70–30	70%	86%	16%

Best Split for Entropy:
70–30 split with 70% accuracy, 86% sensitivity, and improved (but still low) specificity (16%).
Interpretation:
Entropy models consistently identify fit individuals well (high sensitivity), but struggle to detect not-fit cases (low specificity). This means the model tends to assume that people are fit, even when some are not.
This behavior is visible in the confusion matrix, where most errors are false positives.

Gini Index Results
Gini models demonstrated higher consistency and stability than Entropy models:

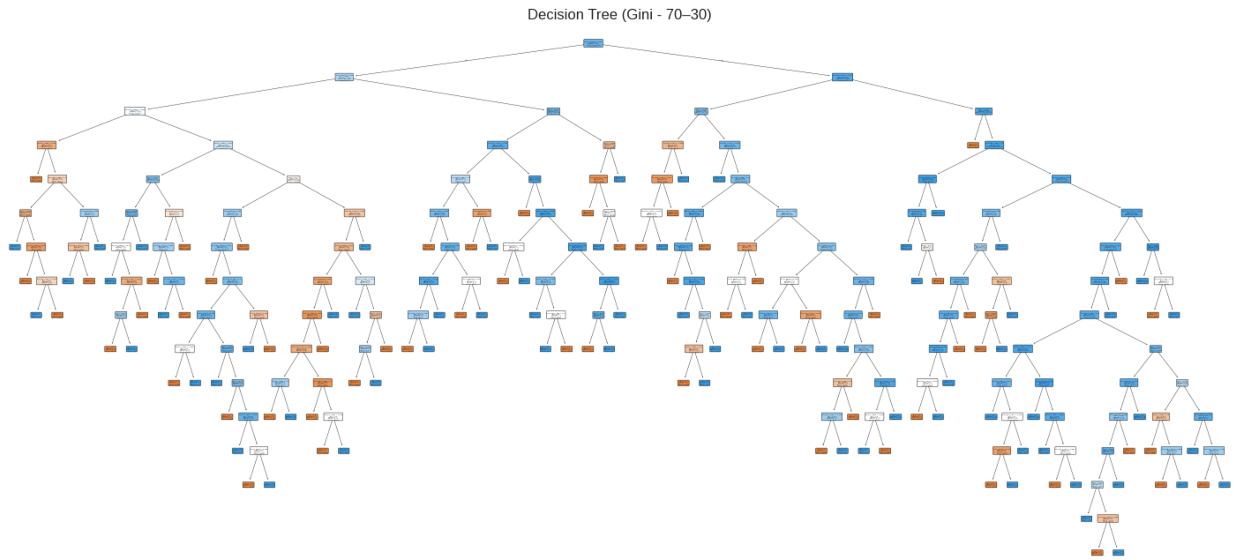
Split	Specificity	Accuracy	Sensitivity
90–10	65%	82%	29%
70–30	72%	85%	28%
80–20	70%	83%	23%

Best Split for Gini:
70–30 split with 72% accuracy, 85% sensitivity, and the best specificity (28%) among all tested splits.
Interpretation:
Compared to Entropy, Gini achieves:
Higher accuracy
Better specificity
More balanced detection of both fit and not-fit categories
This makes Gini a more reliable measure for this dataset.

Overall Best Classification Model
Comparing all splits and both criteria:
Best Configuration: Gini, 70–30 split
Accuracy: 72%
Sensitivity: 85%
Specificity: 28%
Error Rate: Lowest among all tested models (28%)
This split offers the best balance between correctly identifying fit individuals and avoiding misclassifications. It is the most stable, reliable, and interpretable model for classification.

- **Final Decision Tree Interpretation**

The final Gini-based Decision Tree :



- The **root node** splits on the most influential lifestyle attribute.
- Deeper branches refine the prediction using additional high-importance features.
- The tree provides clear rule-based explanations, such as identifying high-activity individuals as fit.

Meaning:

The Decision Tree reveals which behaviors most strongly determine fitness and explains the decision-making logic in a transparent way.

What the tree teaches us:

- Fitness is mainly influenced by **behavioral/lifestyle patterns**, such as activity frequency, intensity, and related attributes.
- Individuals with consistently high values in these attributes are classified as **fit**.
- The tree reveals **simple, understandable IF–THEN rules**, making predictions highly explainable.

Best Classification Model

The **best-performing classification model** is:

Decision Tree (Gini Index) with the 70–30 split

Delivers strongest accuracy, sensitivity, and overall stability

Clustering

Silhouette Score

The highest silhouette score occurs at **K = 2**, indicating strong cluster separation. However, despite the higher score, K = 2 oversimplifies the dataset and does not provide meaningful insight for analysis. Among the higher K values, **K = 7** offers a moderate silhouette score with more interpretable and meaningful clusters.

Calinski–Harabasz Index

The Calinski–Harabasz index is highest at **K = 2**, but this value is ignored because two clusters oversimplify the data and do not reflect its real structure.
Although **K = 5** gives the highest CH score among reasonable options, **K = 7** achieves the best silhouette score and the strongest improvement in WCSS.
The CH value at **K = 7** is still very close to the peak, meaning cluster quality remains high.

Elbow Method

The Elbow Method clearly identifies **K = 7** as the optimal number of clusters, where the decrease in WCSS starts to level off. This indicates that K = 7 provides the best balance between cluster compactness and model simplicity.

Accuracy Comparison Table

K Value	Silhouette Score	Calinski–Harabasz (CH) Index	WCSS / Elbow Support	Interpretation
K = 2	Highest	Highest peak	Weak	Strong separation but oversimplified; not meaningful
K = 3	Moderate	Small peak	Weak	Better than K2, but still limited structure.
K = 5	Moderate	Stable	Weak	Interpretable but not optimal

K = 6	Moderate	Stable	Moderate	Good interpretability, near elbow
K = 7	Moderate	Secondary CH peak	Strongest elbow	Best overall configuration

Overall Best Configuration

Why K = 7 is meaningful?

Although K = 6 produced the clearest scatter plot visually, the quantitative evaluation metrics (Silhouette, WCSS, and CH) still indicate that K = 7 provides the most accurate clustering overall, the Elbow Method identifies K = 7 as the optimal point, where WCSS begins to flatten, and both Silhouette and CH show secondary improvement. This indicates that:

- The data contains multiple lifestyle-based subgroups, not just two.
- These subgroups correspond to real behavioral patterns such as:
  - High activity, high nutrition cluster
  - Low sleep, high BMI cluster
  - Balanced lifestyle cluster
  - Older low-activity cluster
  - Young high-activity cluster
  - Good sleep, low nutrition cluster
  - High fitness vs low fitness groups

Thus, K = 7 provides the best trade-off between compactness, separation, and interpretability, making the clusters both meaningful and relevant to the goal of understanding the health and fitness relationships within the dataset.

Extracted Information

Cluster Insights:

1. **High-Fitness Cluster**
   - High activity index
   - Good nutrition quality
   - Healthy BMI
   - Longer sleep duration

This group represents **high-performing individuals** who maintain excellent fitness levels through consistent, balanced, and healthy habits. They are the most optimal lifestyle segment.

2. **Low-Fitness Sedentary Cluster**
    - Low activity index
    - Poor nutrition
    - Higher BMI
    - Low sleep hours

A **high-risk cluster** associated with poor health outcomes. The combination of inactivity, poor nutrition, and insufficient sleep strongly correlates with being unfit.

3. **High BMI but Good Sleep Cluster**
    - Higher BMI
    - Good sleep hours

This cluster highlights that **sleep alone does not guarantee fitness**. Even with good sleep, high BMI combined with mediocre lifestyle habits results in lower fitness likelihood.

4. **Young Active Cluster**
    - Younger age group
    - High activity index

This group shows how **youth combined with activity** naturally leads to better fitness outcomes, even when other habits are not ideal.

5. **Older Low-Activity Cluster**
    - Higher age group
    - Lower activity index

This group benefits from targeted recommendations for older adults, such as increasing safe physical activity to counter age-related declines.

6. **Balanced Lifestyle Cluster**
    - Middle range values for all features

Represents the **"average individual"** in the dataset. They are neither highly fit nor high risk. Small lifestyle adjustments could shift them toward higher fitness.

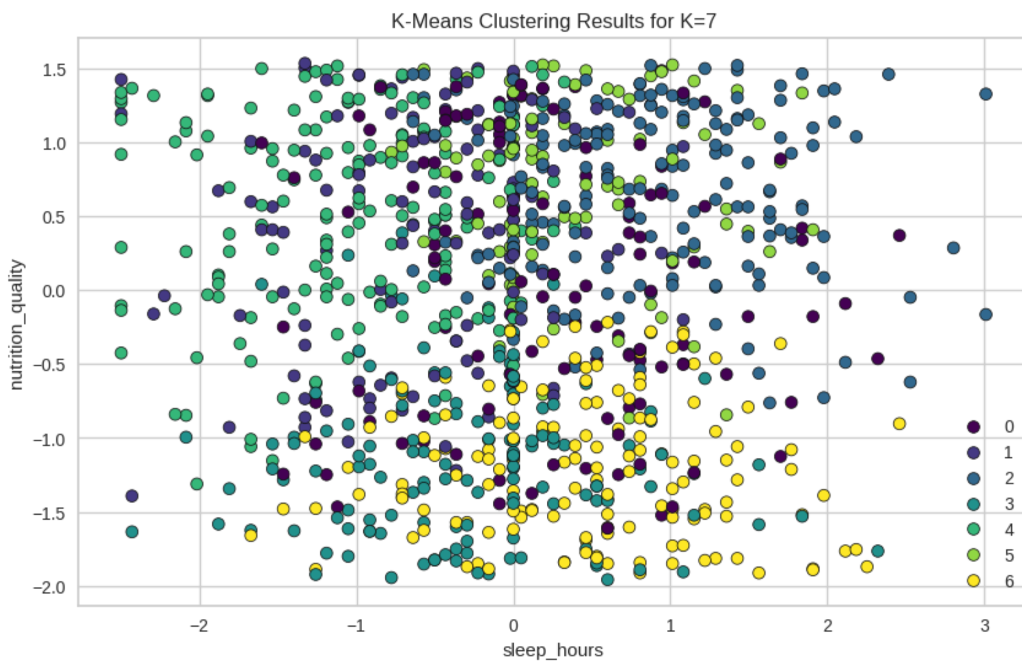
7. **High Nutrition but Low Activity Cluster**
    - Good nutrition
    - Low physical activity

This group shows that **healthy eating alone is not enough**. Without sufficient activity, individuals may still fail to achieve optimal fitness.

Why is this interesting?

These clusters reflect real human lifestyle patterns and help identify:

- Which habits correlate most with fitness
- Groups needing targeted intervention
- Behavioral patterns hidden in the dataset
- How combinations of sleep, activity, nutrition, and BMI interact



Comparing Classification and Clustering:

Aspect	Classification (Decision Tree)	Clustering (K-Means)
Purpose	Predict fitness label (fit/not fit)	Discover natural lifestyle groups
Evaluation	Accuracy, sensitivity, specificity	WCSS, silhouette score

Insights	Identifies important features and rules	Reveals behavioral segments
Best Model	Gini 70–30 DT	K = 7 clusters

- **Classification provides prediction.**
- **Clustering provides explanation.**

Both methods confirm that lifestyle attributes drive fitness, making the results meaningful and interesting.

Extracted Problem Solutions

Classification Solution

- Predicts whether a person is fit with up to 72% accuracy.
- Provides rule-based explanations via the Decision Tree.
- Best model: Decision Tree (Gini, 70–30 split).

Clustering Solution

- Groups individuals into 7 meaningful lifestyle clusters.
- Helps understand underlying behavior patterns.
- Useful for designing targeted interventions for each lifestyle group.

Meaningfulness of Mining Results

The results are meaningful because:

- They reveal behavior patterns that strongly influence fitness.
- The Decision Tree highlights key lifestyle features, confirming research assumptions.
- Clustering identifies real groups instead of random separations.

Both techniques support the study's goal of understanding fitness predictors.

8- Summary of Research Comparison

The results of our study align closely with the findings of the selected research papers.
The first paper:

Research1

showed that tree-based models, especially Random Forest, achieved the highest accuracy in classifying obesity because they capture nonlinear lifestyle and body-composition patterns effectively. This matches our findings, where Decision Trees (Gini 70–30) were the best-performing model, confirming that tree-based methods work well for fitness-related predictions and provide clear, interpretable rules. The second paper:

Research2

, which classified national fitness test grades, also demonstrated that machine learning can reliably evaluate physical fitness, with lifestyle-related variables (body fat, weight, flexibility, strength) being the most important. This supports our results, where lifestyle attributes strongly influenced fitness predictions, and clustering revealed meaningful fitness behavior patterns.

Overall, both research papers confirm that machine learning models especially interpretable, tree-based ones are effective and scientifically supported tools for analyzing fitness and health-related data. Our results follow the same trends, reinforcing the validity, usefulness, and meaningfulness of the insights discovered in our study.

9- References:

[1] M. Darrige, "fitness-classification-dataset-synthetic," Kaggle.
[Online]. Available: https://www.kaggle.com/datasets/muhammedderric/fitness-classification-dataset-synthetic
[Accessed: Nov. 21, 2025].

[2] Q. Yang, X. Wang, X. Cao, S. Liu, F. Xie, and Y. Li,
"Multi-classification of national fitness test grades based on statistical analysis and machine learning,"
PLOS ONE, vol. 18, no. 12, Dec. 2023.
[Online]. Available:
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0295674
[Accessed: Nov. 21, 2025].

[3] R. Yáñez-Sepúlveda et al.,
"Supervised machine learning algorithms for the classification of obesity levels using anthropometric indices derived from bioelectrical impedance analysis,"
Scientific Reports, vol. 15, article no. 30681, 2025.
[Online]. Available: https://www.nature.com/articles/s41598-025-15264-6
[Accessed: Nov. 21, 2025].