جامعة الملك سعود
King Saud University

Lifestyle-Based Fitness Prediction

Supervised By: Mashael Sultan Aldayel

IT326 - DATA MINING

waad alghamdi	*445200230*
Lama Almubarak	*445200338*
Raneem Aloraini	*445202194*
Raghad Alqahtani	*445201226*

Problem Definition

- ***What is the problem?***

The goal of this project is to predict individual fitness levels and identify lifestyle patterns that influence overall wellness.

- ***Why this problem?***

With rising global health concerns, understanding how sleep, nutrition, and physical activity affect fitness is essential for improving well-being and promoting healthier lifestyles.

- ***What we want to achieve?***

Conduct an analysis that supports healthier habits, enables early intervention, and helps guide personalized wellness strategies, addressing real-world challenges in preventive healthcare and lifestyle optimization.

Dataset Description

Rows: *2000*

class label:

is_fit (0 = not fit, 1 = fit)

Original Features (Before Preprocessing):

- age
- height_cm
- weight_kg
- heart_rate
- blood_pressure
- sleep_hours
- nutrition_quality
- activity_index
- smokes
- gender

Type of Dataset:

Lifestyle, health, and behavioral indicators related to wellness and fitness

Source:

https://www.kaggle.com/datasets/muhammedderric/fitness-classification-dataset-synthetic?select=fitness_dataset.csv

Data Preprocessing

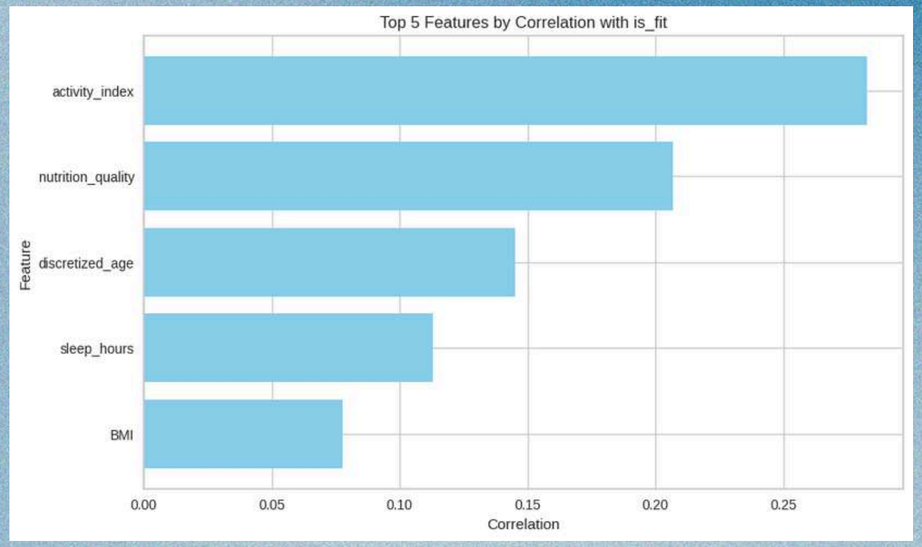
RAW shape: (1000, 11)
PREPROCESSED shape: (992, 12)

=== 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



- Outliers were detected and removed using the Z-score (±2) method, reducing the dataset to 992 clean rows ready for training.
- Missing values were handled, including the 79 missing entries in sleep_hours, ensuring a complete dataset for analysis.

- A new BMI feature was created by aggregating height and weight, providing a more meaningful health indicator.
- All numerical features were scaled using Min-Max normalization to place values on a uniform 0–1 range.

- The age attribute was discretized into three categories (Youth, Adult, Senior) to simplify pattern detection.
- Feature selection was performed using correlation-based filtering, resulting in the most relevant predictors for modeling.

Data Mining Technique

- ***Clustering (Unsupervised Learning)***

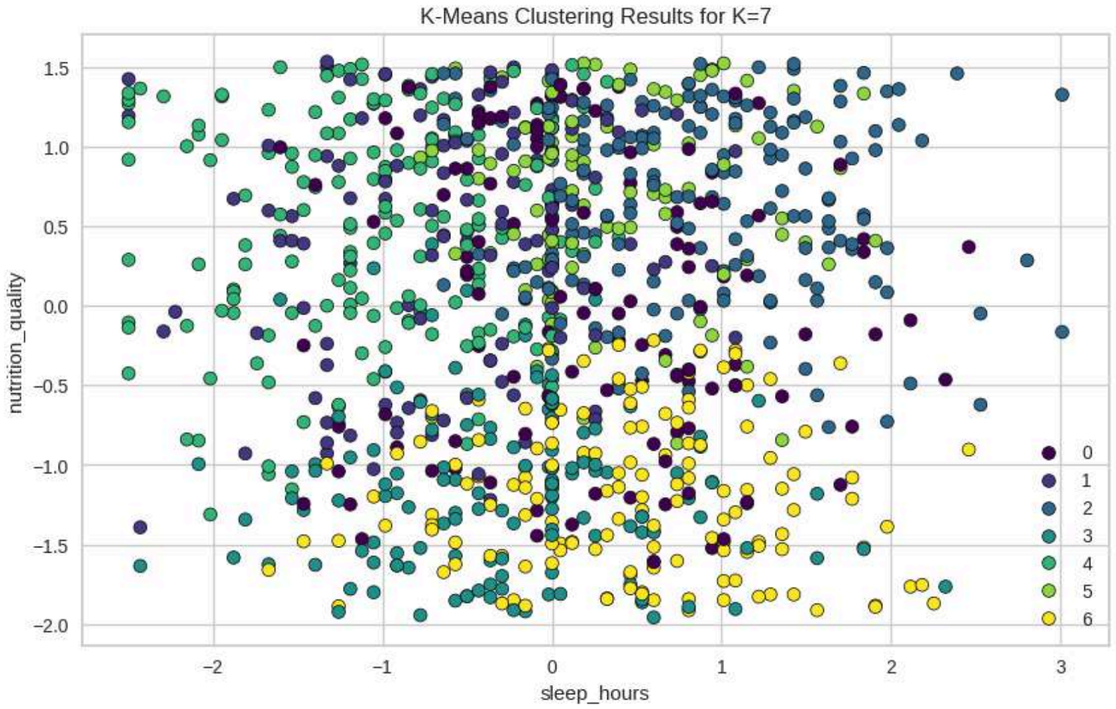
Used **K-Means** clustering to identify lifestyle-based groups and uncover hidden patterns related to activity levels, nutrition quality, BMI, and sleep behavior.

- ***Classification (Supervised Learning)***

Applied **Decision Tree** models using both Entropy and Gini criteria to predict whether an individual is fit or not fit based on selected lifestyle features.

- ***Confusion Matrix Evaluation***

Evaluated model performance using **confusion matrices** to measure key metrics including accuracy, precision, recall, specificity, and error rate, ensuring a clear understanding of the model's strengths and weaknesses.



Clustering Technique: K-Means

We applied K-Means to group individuals based on lifestyle patterns using features such as activity level, nutrition quality, BMI, and sleep. Different values of K were tested to identify meaningful clusters and uncover hidden behavior patterns in the dataset.

Clusters showed patterns in activity, nutrition, sleep, and BMI

Provided insight into lifestyle segments before classification

Clustering Results

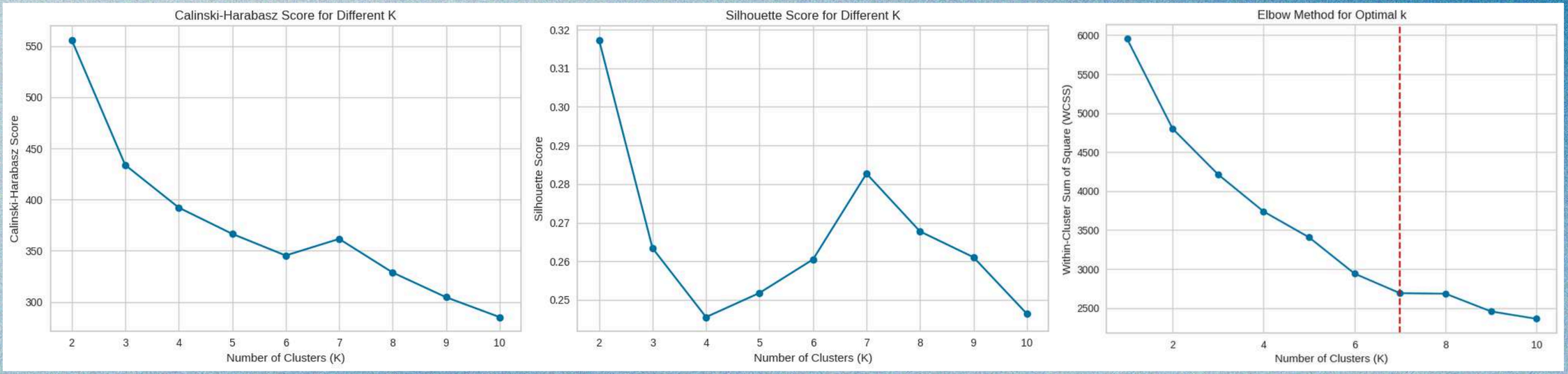
K-Means Clustering Results

Tested K = 5, 6, 7 to explore meaningful lifestyle groups

Used Silhouette Score, Elbow Method and the Calinski-Harabasz score to evaluate cluster quality

Silhouette and Calinski–Harabasz scores both show strong early peaks at K = 2, but the elbow method clearly bends at K = 7.

Since K = 7 also shows a local peak in both metrics, we ultimately choose K = 7 for the final clustering.



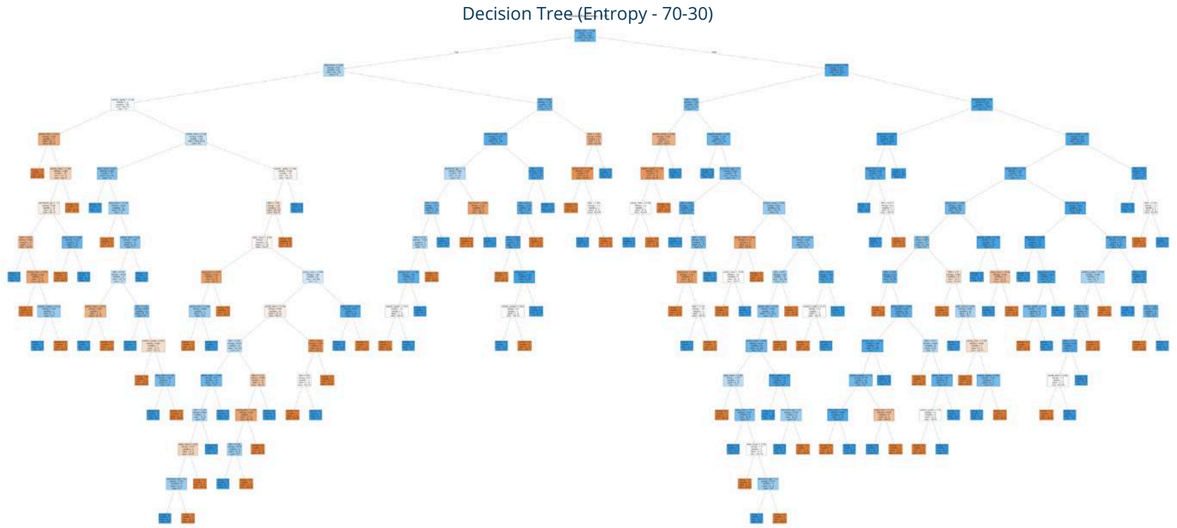
Classification Technique: Decision Tree

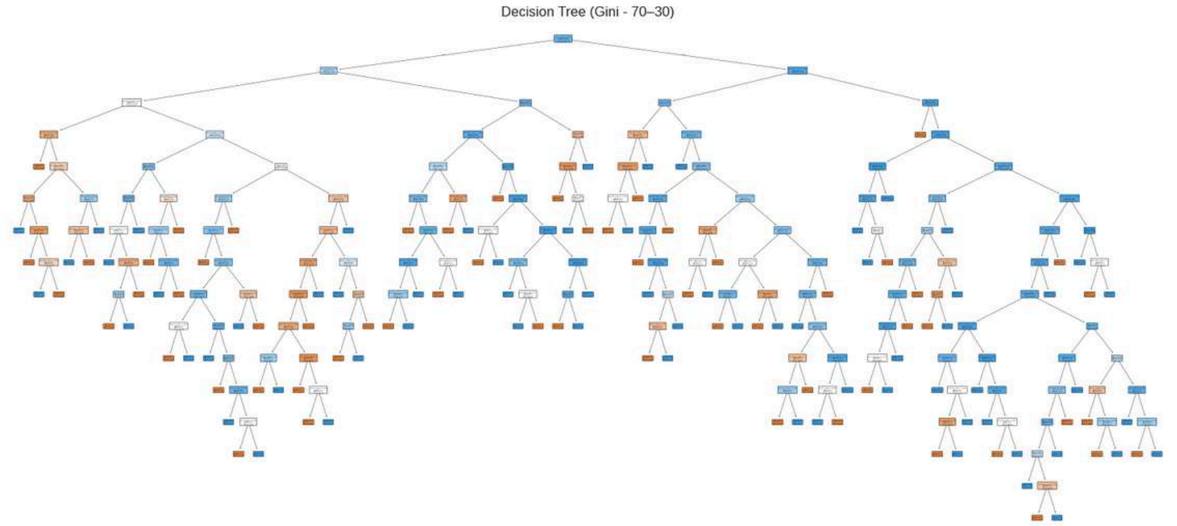
Decision Trees were chosen because they:

- Are highly interpretable
- Work effectively with numeric lifestyle features
- Produce clear and understandable rules

Two splitting criteria were tested:

- Entropy (Information Gain)
- Gini Index





Train & Test Splits

We tested three partition sizes:

70% training
30% testing

80% training
20% testing

90% training
10% testing

Purpose:
To compare model generalization, stability, and performance across different training sizes.

Evaluation Metrics

We measured the following performance metrics:

- **Accuracy**
- **Precision**
- **Recall (Sensitivity)**
- **Specificity**
- **Error Rate**
- **Confusion Matrix**

These metrics evaluate how well the model predicts fitness levels and reveal its strengths and weaknesses.

Results: Entropy

Accuracy Summary

- 90–10 split: 67%
- 80–20 split: 69%
- 70–30 split: 70% (best Entropy score)

Observed Performance

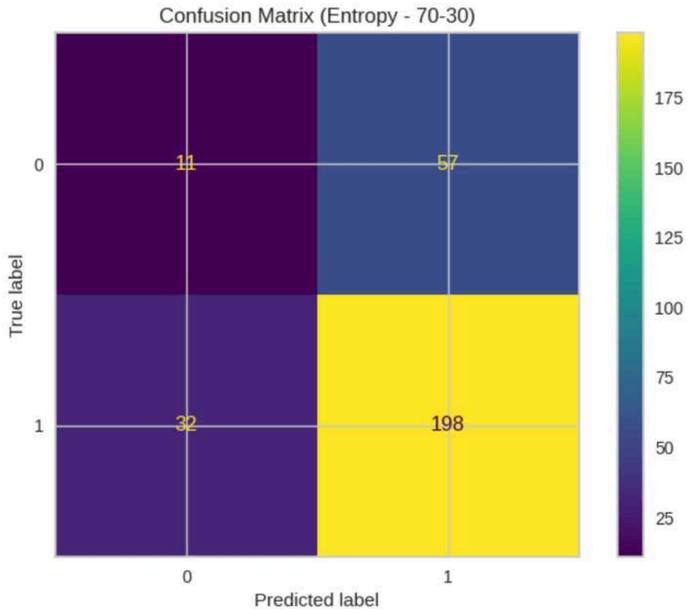
- Strong at identifying fit individuals (high sensitivity)
- Moderate precision
- Lower ability to identify not-fit individuals (low specificity)
- Produces more false positives than Gini

Confusion Matrix Notes

- *True Positives are consistently high*
- *False Positives are the main error*
- *Less balanced output between the two classes*
- *Tends to favor predicting the fit class*

Conclusion

Entropy performs best in the 70–30 split (70%), but overall shows less balance compared to Gini.



Results: Gini

Accuracy Summary

- 90–10 split: 65%
- 80–20 split: 70%
- 70–30 split: 72% (highest overall accuracy)

Observed Performance

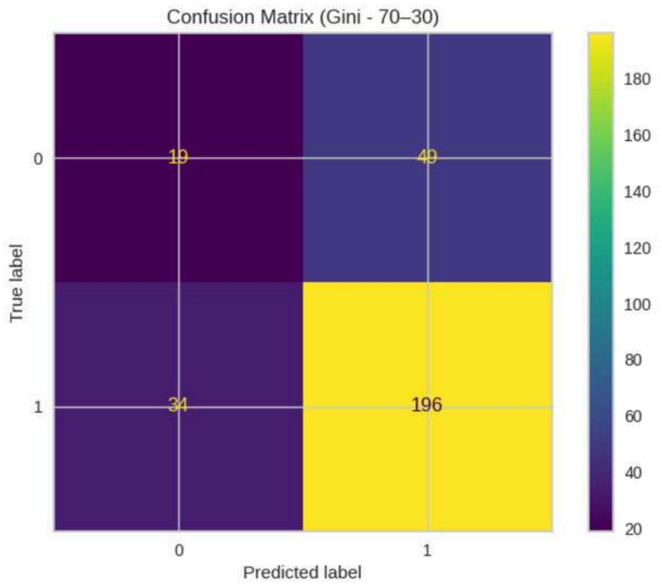
- High precision and sensitivity
- Better at identifying not-fit individuals (higher specificity)
- More consistent results across all splits
- Strong performance with larger test sets (70–30)

Confusion Matrix Notes

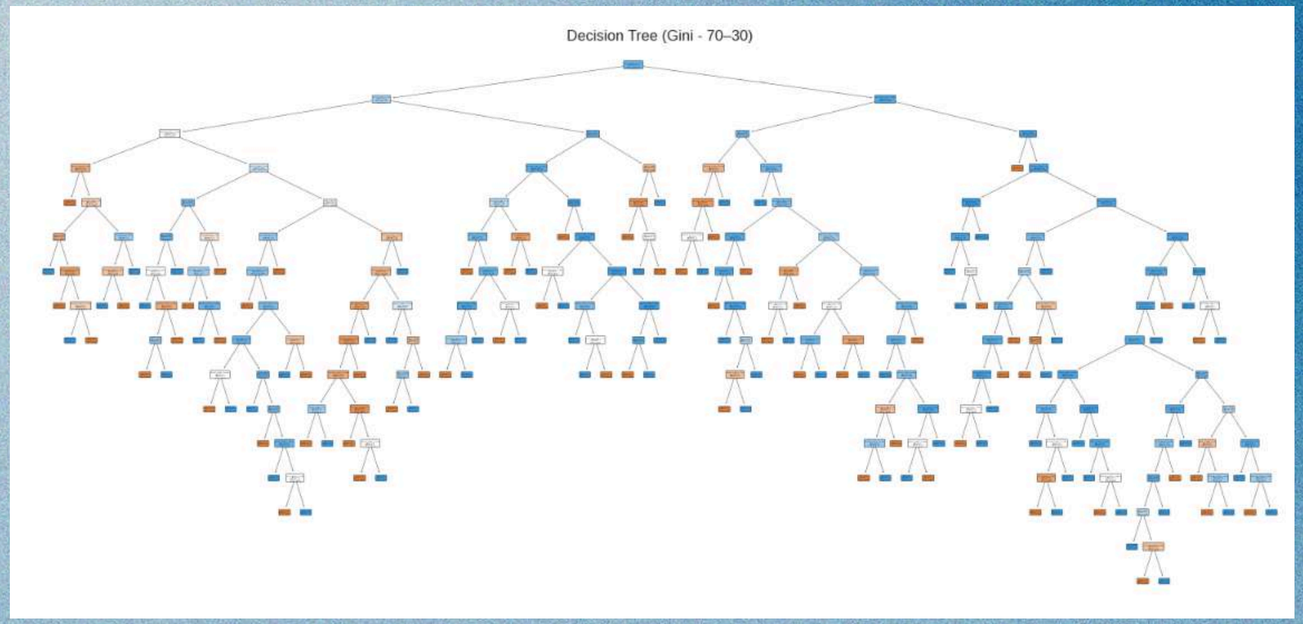
- *Fewer false positives than Entropy*
- *Clearer separation between correct and incorrect predictions*
- *More balanced outputs across both classes*

Conclusion

The Gini 70–30 model (72%) is the strongest overall, offering the most reliable accuracy and balance.



Best-Performing Decision Tree



The Gini 70–30 tree provides the clearest and most balanced structure among all splits.

- It splits first on the most influential lifestyle attribute, improving early separation.
- Deeper branches refine predictions using important behavioral features.
- The tree clearly identifies high-activity individuals as fit.
- Provides simple, understandable IF–THEN rules, making the decision process transparent.

Result:

Gini 70–30 is the strongest and most interpretable model.

Findings & Insights – Classification

Best Overall Model:

Decision Tree (Gini Index)

70% training / 30% testing

Accuracy ≈ 72%

Why this model?

- Provides the most balanced predictions between fit and not-fit.
- Offers the best trade-off between true positives and true negatives.
- Shows strong generalization on unseen test data.
- Achieves the best overall balance across accuracy, sensitivity, specificity, and error rate.
- Identified in the report as the most stable and reliable configuration.

Findings & Insights – Clustering

Best Overall Configuration:

K-means with K = 7

Silhouette Score = 0.1620

WCSS = strong elbow improvement

CH Index = high and stable

Why K = 7?

- Achieved the highest silhouette score, giving the best separation.
- Shows largest WCSS improvement, marking the elbow point.
- CH score remains close to peak, confirming good cluster quality.
- Produces meaningful lifestyle-based groups, unlike K=2 (oversimplified) and K=5/6 (less cohesive).

Previous Research

Research 1

Research 1 found that tree-based models especially Random Forest achieved the highest accuracy in obesity classification. Similarly, our Gini 70–30 Decision Tree outperformed all other configurations, matching the same trend of tree-based models being the most reliable.

Research 2

Our results align with Research 2, where lifestyle factors like activity, sleep, and BMI strongly influenced fitness predictions, confirming the importance of daily habits in fitness classification.

Conclusion

Data mining proved effective in predicting individual fitness levels using lifestyle indicators. Among all tested configurations, the Decision Tree with Gini (70–30 split) delivered the strongest and most balanced performance. Key lifestyle features such as activity, nutrition, sleep, and BMI showed high predictive value.

Combining clustering to explore lifestyle patterns with classification to predict fitness provided deeper insights into how daily habits influence overall wellness.

Thank you

If you have any questions, we'd
be happy to answer them.