CentraleSupélec

DECISION MODELING PRESENTATION

ELABORATION OF SOME DECISION MODELS FOR THE NUTRI-SCORE LABEL

GROUP:
MELISSA MEDJAHED
SIHEM BOUTEBAL
DILBAR ISAKOVA

1. PROBLEM STATEMENT

1. PROBLEM STATEMENT

- The Nutri-Score is a nutrition label that converts nutritional value of products into a simple code “A, B, C, D or E”.
- Building a database of at least 200 food products, including their Nutri-Score and Eco-Score.
- Developing and testing decision models: Sorting and ML models that determine the Nutri-Score of a food given its varous characteristics.
- Comparing the new models' classification accuracy with the actual Nutri-Score values.

Figure 1: Nutri-Score logo

2. DATABASE

Link to the database

2. CHOICE OF THE DATABASE

- The database was created via the website: https://fr-en.openfoodfacts.org/

- Categories = [Breakfast cereals, Snacks, Cookies, Industrial ready-made dishes, Beverages, Dairy products]

	Product Name	Brand	Category	Nutri-Score	Eco-Score	Energy (kcal)	Sugars (g)	Saturated Fat (g)	Salt (g)	Fiber (g)	Proteins (g)	Fruits/Vegetables (%)
0	cruesly mélange de noix	Quaker	Aliments et boissons à base de végétaux,Alimen...	a	b	462.0	12.0	2.0	0.000	10.0	8.5	10.700000
1	Céréales Chocapic	Nestlé	Plant-based foods and beverages,Plant-based fo...	a	e	388.0	22.4	2.0	0.220	7.5	8.8	0.000000
2	Flocons d'avoine	Bjorg	Aliments et boissons à base de végétaux,Alimen...	a	a	362.0	1.7	1.3	0.020	11.0	11.0	0.000000
3	Haferflocken	Crownfield	Plantaardige levensmiddelen en dranken,Plantaa...	a	a	372.0	0.7	1.3	0.030	10.0	13.5	0.000000
4	Weetabix	Weetabix	Plant-based foods and beverages, Plant-based f...	a	e	362.0	4.4	0.6	0.275	10.0	12.0	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...
938	All-bran	Kellogg's	Aliments et boissons à base de végétaux, Alime...	a	e	334.0	18.0	0.7	0.000	27.0	14.0	0.000000
939	All-Bran Original	Kellogg's	Plant-based foods and beverages, Plant-based f...	b	e	0.0	15.7	0.9	0.825	28.0	14.1	0.000000
940	Clusters	Nestlé	Aliments et boissons à base de végétaux, Alime...	c	e	392.0	19.9	1.4	0.910	9.6	10.3	0.000000
941	Bamboo	Bamboo	Aliments et boissons à base de végétaux, Alime...	a	b	452.0	5.8	2.6	0.110	7.6	12.0	25.500000
942	Müesli croccante all'avena	Conad	Cibi e bevande a base vegetale,Cibi a base veg...	c	a	481.0	20.0	4.5	0.380	6.0	8.0	0.642535

943 rows × 12 columns

3. PESSIMISTIC MAJORITY SORTING

Link to the code

3. PESSIMISTIC MAJORITY SORTING

a. Defining limiting profiles manually:

```
# Define limiting profiles for Nutri-Score categories (A to E)
limiting_profiles = {
    "A": {"Energy (kcal)": 335, "Sugars (g)": 5, "Saturated Fat (g)": 1, "Salt (g)": 0.1, "Proteins (g)": 8, "Fiber (g)": 10, "Fruits/Vegetables (%)": 40},
    "B": {"Energy (kcal)": 400, "Sugars (g)": 10, "Saturated Fat (g)": 2, "Salt (g)": 0.3, "Proteins (g)": 6, "Fiber (g)": 8, "Fruits/Vegetables (%)": 20},
    "C": {"Energy (kcal)": 500, "Sugars (g)": 15, "Saturated Fat (g)": 3, "Salt (g)": 0.5, "Proteins (g)": 4, "Fiber (g)": 5, "Fruits/Vegetables (%)": 10},
    "D": {"Energy (kcal)": 700, "Sugars (g)": 20, "Saturated Fat (g)": 5, "Salt (g)": 1.0, "Proteins (g)": 2, "Fiber (g)": 3, "Fruits/Vegetables (%)": 5},
    "E": {"Energy (kcal)": 900, "Sugars (g)": 30, "Saturated Fat (g)": 8, "Salt (g)": 2.0, "Proteins (g)": 0, "Fiber (g)": 0, "Fruits/Vegetables (%)": 0},
}
```

Using these weights:

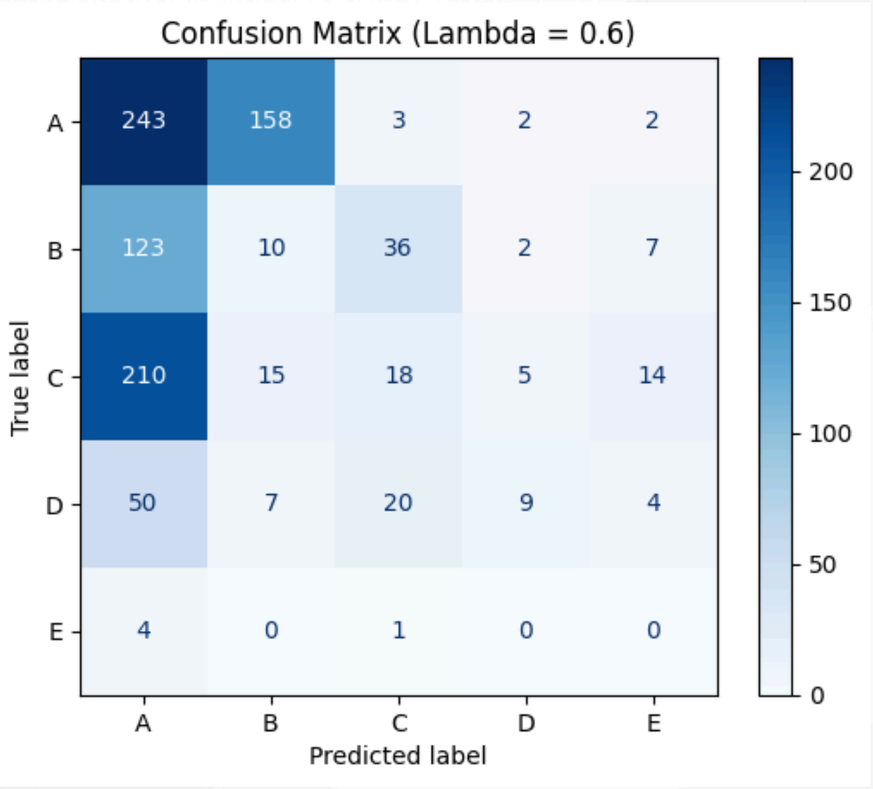
	1- Energy	2-Sugars	3-Satu. fat.	4- Salt	5- Proteins	6- Fiber	7-Fruits & vegetables
Weights w_i	4	3	3	3	2	2	1

Table 1: An example of the weights associated to the criteria

3. PESSIMISTIC MAJORITY SORTING

Results:

Product Name	Nutri score	Lambda	Pessimistic Category
cruesly mélange de noix	A	0.5	A
cruesly mélange de noix	A	0.6	A
Céréales Chocapic	A	0.5	A
Céréales Chocapic	A	0.6	A
Flocons d'avoine	A	0.5	B
Flocons d'avoine	A	0.6	B
Haferflocken	A	0.5	B
Haferflocken	A	0.6	B
Weetabix	A	0.5	B
Weetabix	A	0.6	B
Muesli Superfruits	A	0.5	A
Muesli Superfruits	A	0.6	A
Flocons d'avoine complète	A	0.5	B
Flocons d'avoine complète	A	0.6	B
Weetabix produit à base de blé complet 100%	A	0.5	B
Weetabix produit à base de blé complet 100%	A	0.6	B
Muesli Raisin, Figue, Abricot	A	0.5	A
Muesli Raisin, Figue, Abricot	A	0.6	A
Copos de avena	A	0.5	B
Copos de avena	A	0.6	B
NESTLE CHOCAPIC Céréales 750g	A	0.5	A
NESTLE CHOCAPIC Céréales 750g	A	0.6	A
Corn Flakes	B	0.5	A
Corn Flakes	B	0.6	C
Croustillant Chocolat	C	0.5	A



- Strong Performance for Category “A”
- 158 items with true label “A” were predicted as “B”
- Only 10 items labaled as “B” are correctly predicted
- Category “C” overlaps with other categories
- Small numbers of items labeled as “D” and “E” are correctly predicted

3. PESSIMISTIC MAJORITY SORTING

b. Defining limiting profiles using Quintile approach:

index	A	B	C	D	E
Energy (kcal)	369.0	380.0	401.19999999999993	443.0	601.0
Sugars (g)	5.620000000000002	13.58	17.673333333333595	22.0	48.0
Saturated Fat (g)	0.6	1.2	2.0	3.8	20.7
Salt (g)	0.02	0.07660000000000002	0.28	0.66	8.4
Proteins (g)	80.0	12.0	10.0	9.0	8.0
Fiber (g)	33.0	10.0	8.5	7.0	5.5
Fruits/Vegetables (%)	99.796875	12.98	1.5874999999999995	0.0	0.0

Using these weights:

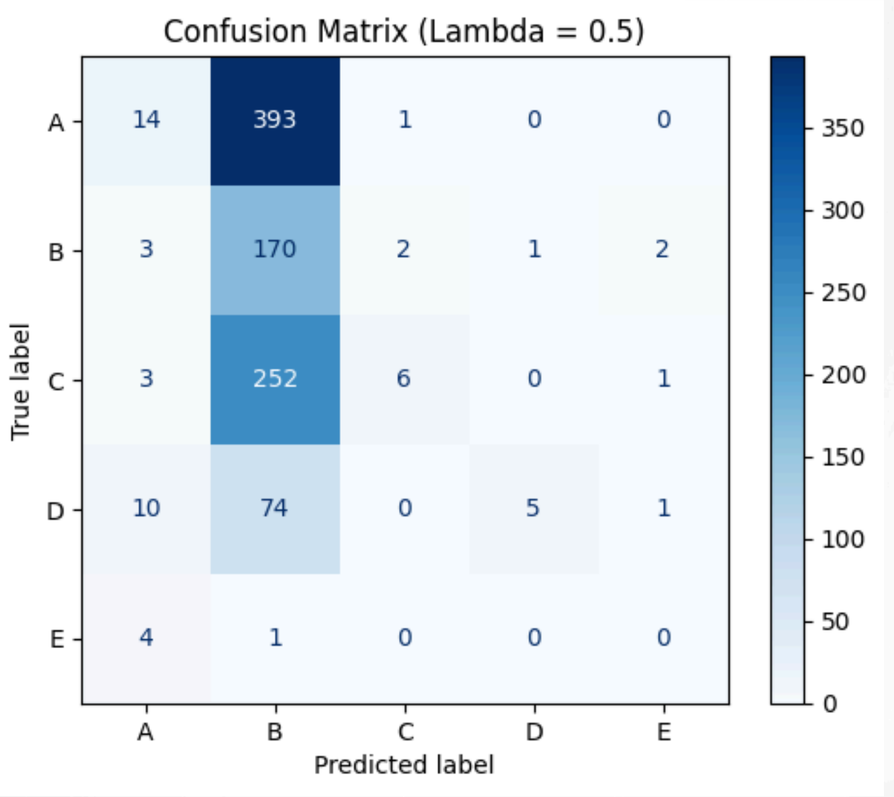
	1- Energy	2-Sugars	3-Satu. fat.	4- Salt	5- Proteins	6- Fiber	7-Fruits & vegetables
Weights w_i	4	3	3	3	2	2	1

Table 1: An example of the weights associated to the criteria

3. PESSIMISTIC MAJORITY SORTING

Results:

Product Name	Nutri score	Lambda	Pessimistic Category
cruesly mélange de noix	A	0.5	B
cruesly mélange de noix	A	0.6	B
Céréales Chocapic	A	0.5	B
Céréales Chocapic	A	0.6	B
Flocons d'avoine	A	0.5	B
Flocons d'avoine	A	0.6	B
Haferflocken	A	0.5	B
Haferflocken	A	0.6	B
Weetabix	A	0.5	B
Weetabix	A	0.6	B
Muesli Superfruits	A	0.5	B
Muesli Superfruits	A	0.6	B
Flocons d'avoine complète	A	0.5	B
Flocons d'avoine complète	A	0.6	B
Weetabix produit à base de blé complet 100%	A	0.5	B
Weetabix produit à base de blé complet 100%	A	0.6	B
Muesli Raisin, Figue, Abricot	A	0.5	B
Muesli Raisin, Figue, Abricot	A	0.6	B
Copos de avena	A	0.5	B
Copos de avena	A	0.6	B
NESTLE CHOCAPIC Céréales 750g	A	0.5	B
NESTLE CHOCAPIC Céréales 750g	A	0.6	B
Corn Flakes	B	0.5	B
Corn Flakes	B	0.6	E
Croustillant Chocolat	C	0.5	B



- Strong Confusion for Category “A”: Only 14 items with the true label “A” were correctly predicted
- Good Performance for Category “B”
- Significant Overlap for Category “C”: Only 6 items with the true label “C” were correctly predicted
- Category “C” overlaps with other categories
- Limited Accuracy for Categories “D” and “E”: Only 5 items with the true label “D” and 0 items with the true label “E” were correctly predicted

4. OPTIMISTIC MAJORITY SORTING

Link to the code

4. OPTIMISTIC MAJORITY SORTING

a. Defining limiting profiles manually:

```
# Define limiting profiles for Nutri-Score categories (A to E)
limiting_profiles = {
    "A": {"Energy (kcal)": 335, "Sugars (g)": 5, "Saturated Fat (g)": 1, "Salt (g)": 0.1, "Proteins (g)": 8, "Fiber (g)": 10, "Fruits/Vegetables (%)": 40},
    "B": {"Energy (kcal)": 400, "Sugars (g)": 10, "Saturated Fat (g)": 2, "Salt (g)": 0.3, "Proteins (g)": 6, "Fiber (g)": 8, "Fruits/Vegetables (%)": 20},
    "C": {"Energy (kcal)": 500, "Sugars (g)": 15, "Saturated Fat (g)": 3, "Salt (g)": 0.5, "Proteins (g)": 4, "Fiber (g)": 5, "Fruits/Vegetables (%)": 10},
    "D": {"Energy (kcal)": 700, "Sugars (g)": 20, "Saturated Fat (g)": 5, "Salt (g)": 1.0, "Proteins (g)": 2, "Fiber (g)": 3, "Fruits/Vegetables (%)": 5},
    "E": {"Energy (kcal)": 900, "Sugars (g)": 30, "Saturated Fat (g)": 8, "Salt (g)": 2.0, "Proteins (g)": 0, "Fiber (g)": 0, "Fruits/Vegetables (%)": 0},
}
```

Using these weights:

	1- Energy	2-Sugars	3-Satu. fat.	4- Salt	5- Proteins	6- Fiber	7-Fruits & vegetables
Weights w_i	4	3	3	3	2	2	1

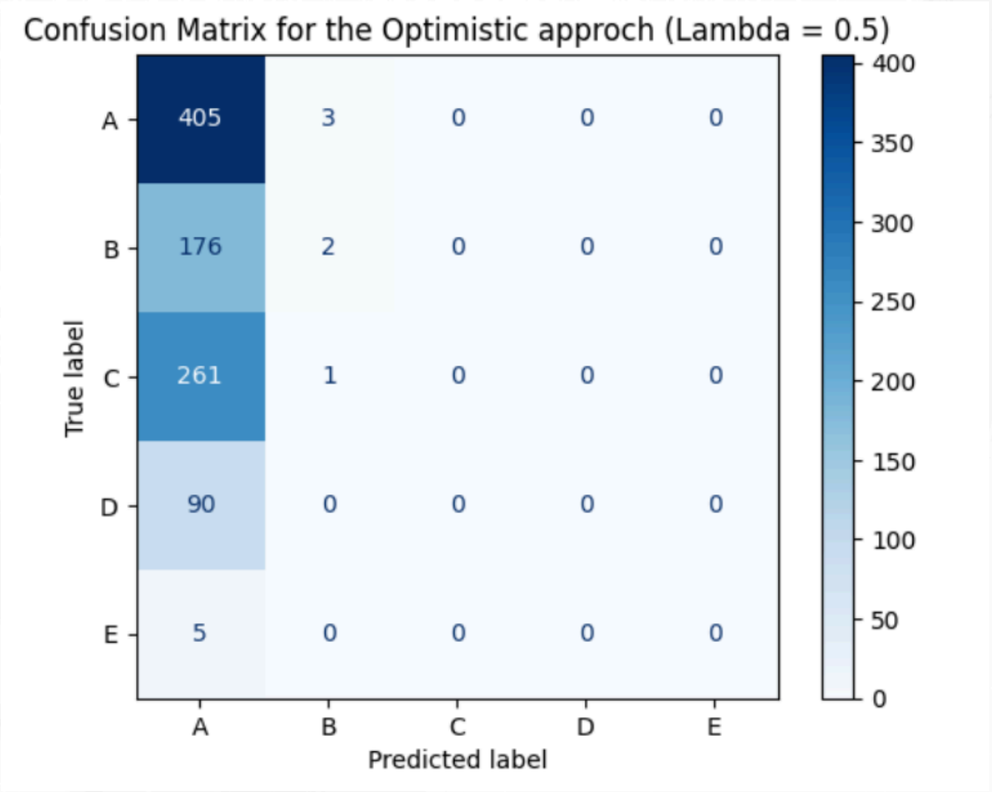
Table 1: An example of the weights associated to the criteria

4. OPTIMISTIC MAJORITY SORTING

Results:

	Product Name	Nutri-Score	Category	Lambda
0	cruesly mélange de noix	A	A	0.5
1	Céréales Chocapic	A	A	0.5
2	Flocons d'avoine	A	A	0.5
3	Haferflocken	A	A	0.5
4	Weetabix	A	A	0.5
...	...	...	...	...
2824	All-bran	A	A	0.7
2825	All-Bran Original	B	A	0.7
2826	Clusters	C	A	0.7
2827	Bamboo	A	A	0.7
2828	Müesli croccante all'avena	C	A	0.7

2829 rows × 4 columns



- Most predictions fall into class "A," with other classes heavily misclassified.
- Classes "B," "C," "D," and "E" are rarely predicted correctly.

4. OPTIMISTIC MAJORITY SORTING

b. Defining limiting profiles using Quintile approach:

index	A	B	C	D	E
Energy (kcal)	369.0	380.0	401.19999999999993	443.0	601.0
Sugars (g)	5.620000000000002	13.58	17.673333333333595	22.0	48.0
Saturated Fat (g)	0.6	1.2	2.0	3.8	20.7
Salt (g)	0.02	0.07660000000000002	0.28	0.66	8.4
Proteins (g)	80.0	12.0	10.0	9.0	8.0
Fiber (g)	33.0	10.0	8.5	7.0	5.5
Fruits/Vegetables (%)	99.796875	12.98	1.5874999999999995	0.0	0.0

Using these weights:

	1- Energy	2-Sugars	3-Satu. fat.	4- Salt	5- Proteins	6- Fiber	7-Fruits & vegetables
Weights w_i	4	3	3	3	2	2	1

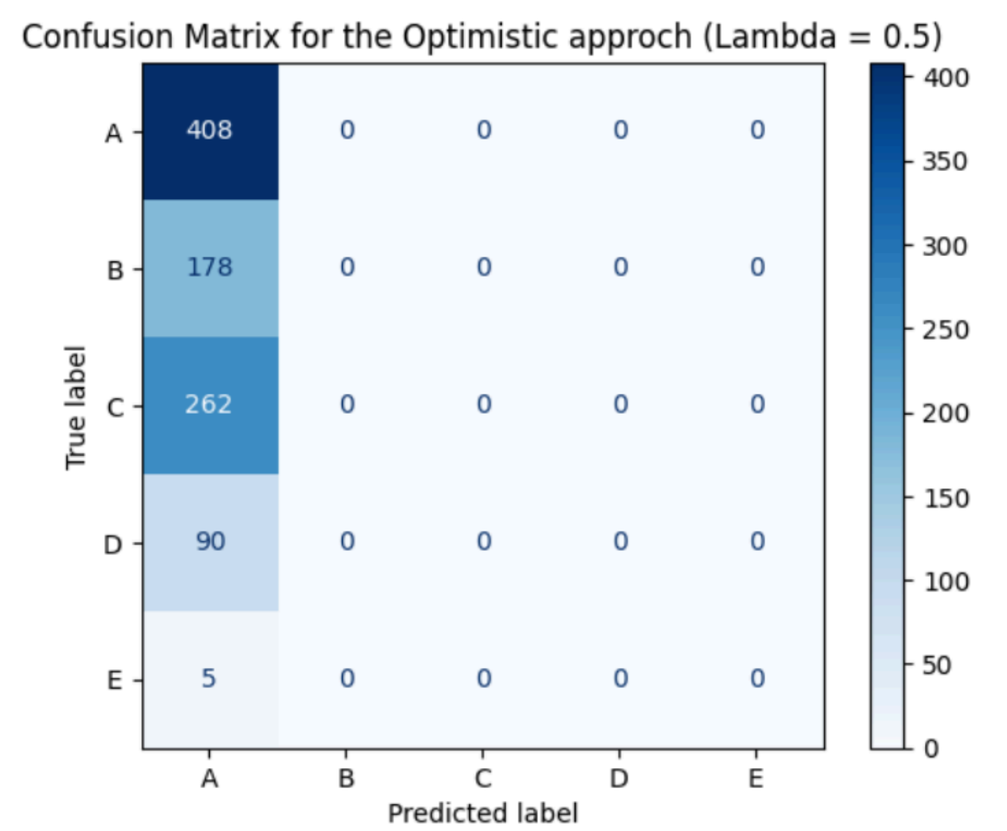
Table 1: An example of the weights associated to the criteria

4. OPTIMISTIC MAJORITY SORTING

Results:

	Product Name	Nutri-Score	Category	Lambda
0	cruesly mélange de noix	A	A	0.5
1	Céréales Chocapic	A	A	0.5
2	Flocons d'avoine	A	A	0.5
3	Haferflocken	A	A	0.5
4	Weetabix	A	A	0.5
...	...	...	...	...
2824	All-bran	A	A	0.7
2825	All-Bran Original	B	A	0.7
2826	Clusters	C	A	0.7
2827	Bamboo	A	A	0.7
2828	Müesli croccante all'avena	C	A	0.7

2829 rows × 4 columns



- All instances are predicted as class "A," regardless of their true labels.
- No predictions are made for classes "B," "C," "D," or "E."

4. OPTIMISTIC MAJORITY SORTING

WITH CLUSTERING ANALYSIS APPROACH

Link to the code

4. OPTIMISTIC MAJORITY SORTING WITH CLUSTERING ANALYSIS APPROACH

Why K-means?

- **Effective for segmenting data into distinct groups based on their attributes.**
- **Identifies centroids, which are representative of the average characteristics within each cluster.**
- **By using k-means, the limiting profiles derived are statistically grounded in the data, providing an objective basis for categorization.**

4. OPTIMISTIC MAJORITY SORTING WITH CLUSTERING ANALYSIS APPROACH

Comparison with other methods:

- ***Manual Setting:*** **can have biases and dependent on the decision-maker's perspective**

=> not accurately reflect the underlying data distribution.

- ***Quintile approach:*** **segment the data into equally sized bins, it might not accurately capture the natural groupings or relationships between variables.**

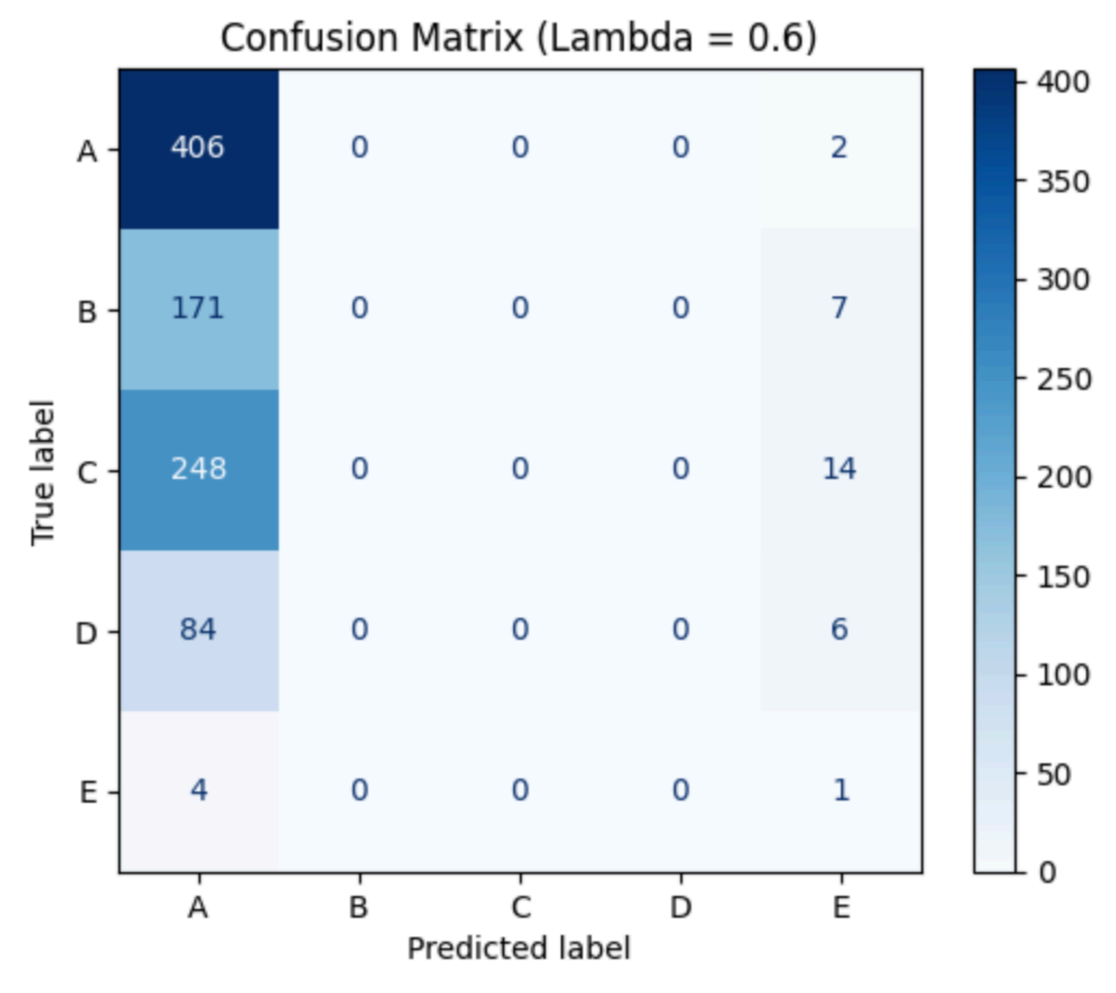
4. OPTIMISTIC MAJORITY SORTING WITH CLUSTER ANALYSIS APPROACH

Limiting profiles derived from clusters:

	Energy (kcal)	Sugars (g)	Saturated Fat (g)	Salt (g)	Proteins (g)	Fiber (g)	Fruits/Vegetables (%)	Nutri-Score
0	374.109269	12.247907	1.173311	0.362684	10.233188	8.249171	5.256934	E
1	213.000000	7.866667	1.966667	0.260000	6.500000	5.766667	7.518136	D
2	466.003286	17.544085	4.853713	0.277018	9.734885	7.494956	9.344656	C
3	426.291030	19.060652	3.043512	0.325902	9.870209	7.200275	8.114808	B
4	0.000000	10.040000	1.050000	0.250500	6.700000	9.020000	6.359783	A

- **In our case, we used k-means to cluster our dataset into five groups to mimic the Nutri-Score categories from A to E.**
- **Each cluster's centroid provides a limiting profile, representing average nutritional values for that category.**
- **This method ensures that each profile is closely aligned with actual data patterns, making our Nutri-Score categorization both reliable and meaningful.**

4. OPTIMISTIC MAJORITY SORTING WITH CLUSTER ANALYSIS APPROACH



Results:

- A: 406 items correctly classified, indicating a strong match between model predictions and actual data for the highest Nutri score.
- E: Only 4 items are in this category, but 1 is correctly predicted, which suggests some effectiveness but possibly indicates a rare category or insufficient data.
- Misclassified: 14 items as 'E', indicating a tendency of the model to falsely predict items in this category as being worse in nutritional value.

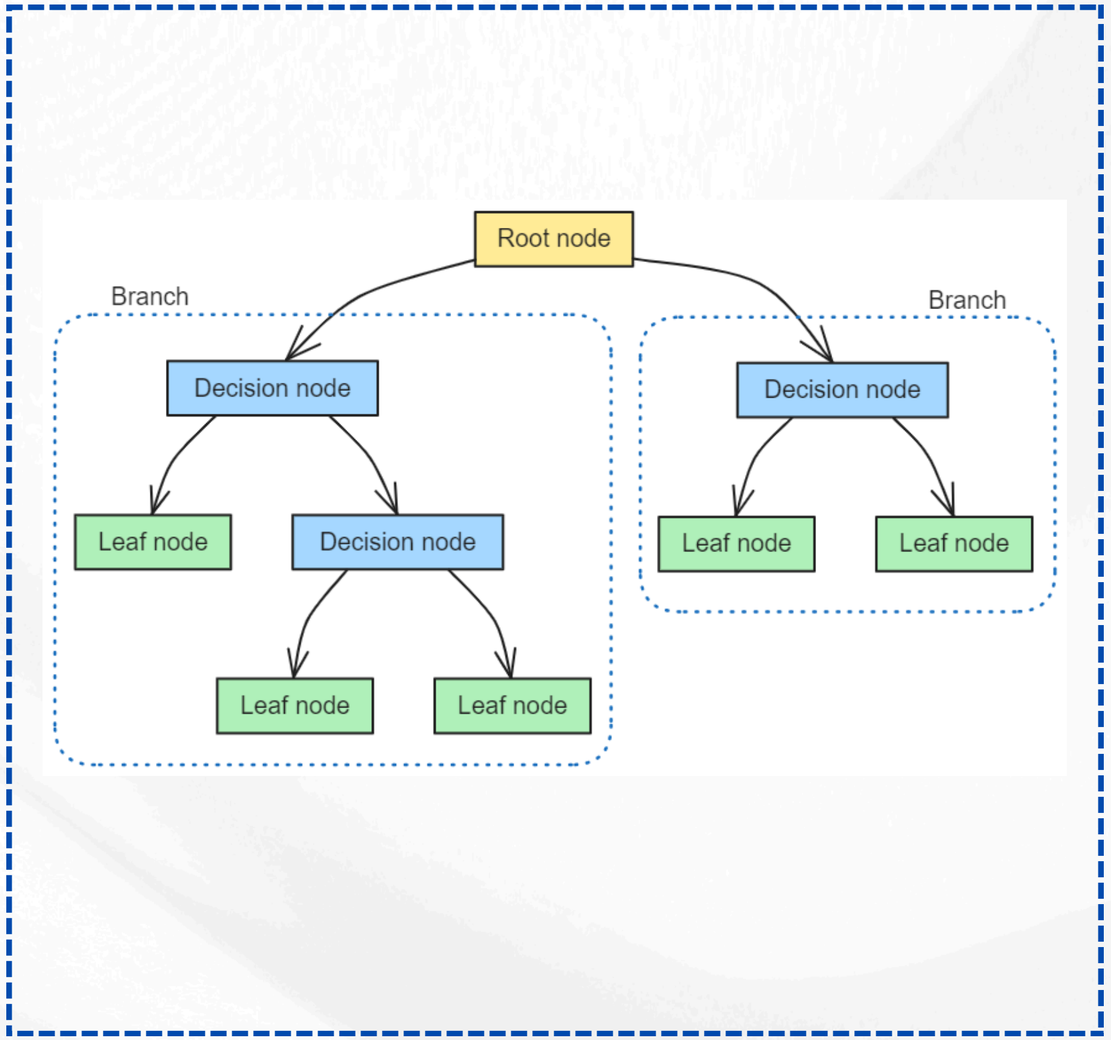
5. MACHINE LEARNING APPROACH

Link to the code

5. USING DECISION TREE ALGORITHM

Definition:

- A supervised learning algorithm used for classification and regression, structured like a tree where each internal node represents a decision rule on a feature, and each leaf represents a class or outcome.



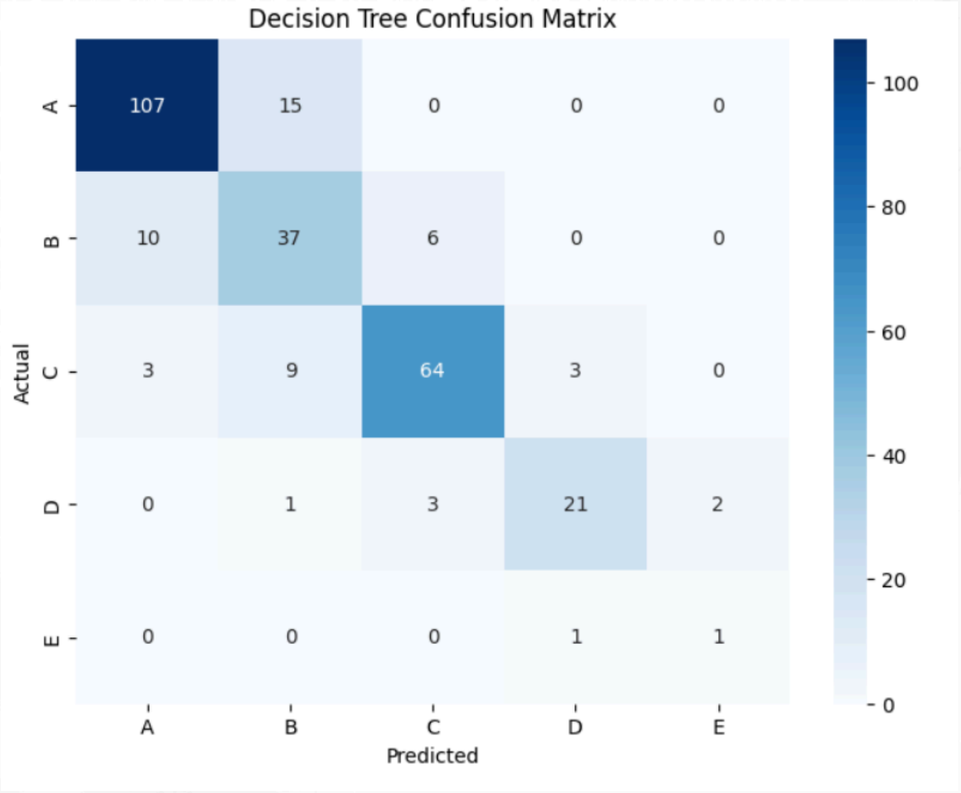
Steps:

- **Recursive Splitting:** Divides data into subsets based on feature thresholds.
- **Splitting Criteria:** Uses metrics like Gini impurity or entropy for optimal splits.
- **Stopping Conditions:** Stops when max depth or minimum samples per leaf is reached.

5. USING DECISION TREE ALGORITHM

Results

	precision	recall	f1-score	support
A	0.891667	0.877049	0.884298	122.000000
B	0.596774	0.698113	0.643478	53.000000
C	0.876712	0.810127	0.842105	79.000000
D	0.840000	0.777778	0.807692	27.000000
E	0.333333	0.500000	0.400000	2.000000
accuracy	0.812721	0.812721	0.812721	0.812721
macro avg	0.707697	0.732613	0.715515	283.000000
weighted avg	0.823390	0.812721	0.816688	283.000000

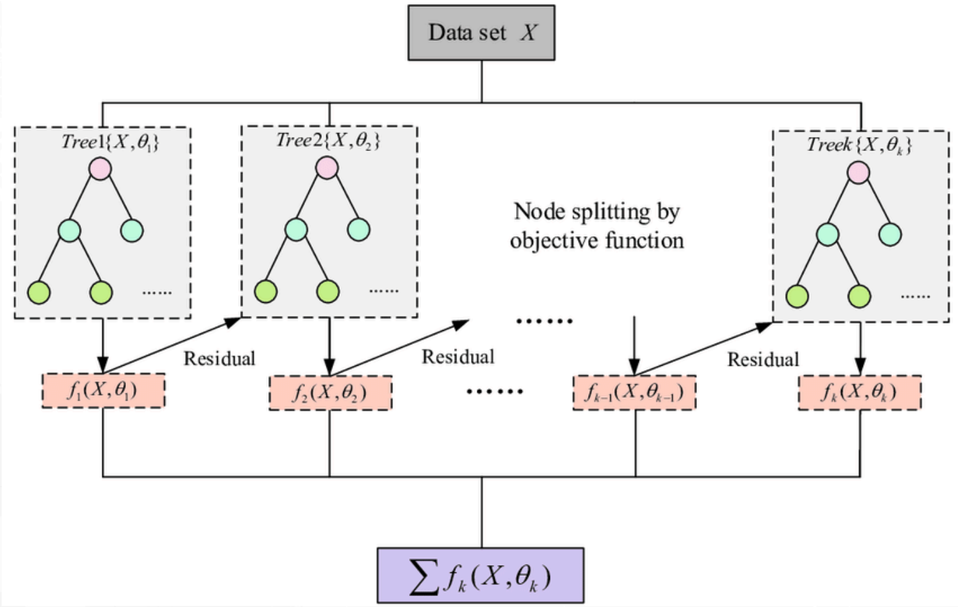


- Unlike the optimistic approach, the decision tree achieves classification for all classes, showing a more balanced performance.
- High precision and recall for classes "A" and "C" (F1-scores ~0.88, 0.84).
- Struggles with "B" (F1 = 0.64) and performs decently for "D" (F1 = 0.81)
- Poor performance for "E" (F1 = 0.40) due to limited data.

5. USING XGBOOST ALGORITHM

Definition:

- **XGBoost** (eXtreme Gradient Boosting) is a powerful machine learning algorithm based on gradient boosting, designed for speed and performance. It builds an ensemble of decision trees iteratively, where each tree corrects the errors of the previous ones.



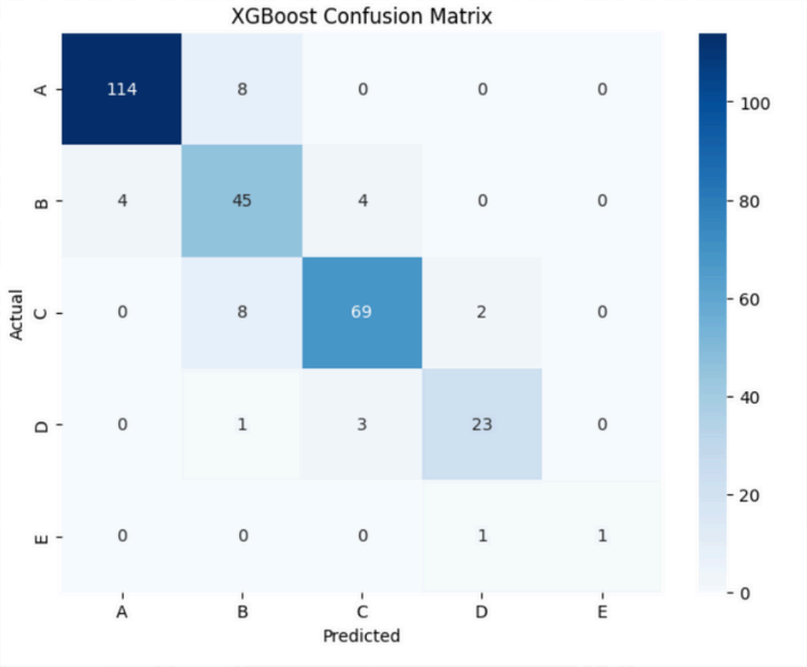
Steps:

- **Build Trees Iteratively:** Train decision trees sequentially, minimizing a loss function.
- **Update Predictions:** Adjust predictions by combining the previous predictions with the outputs of the new tree.
- **Regularization:** Apply techniques like shrinkage, column sampling, and pruning to prevent overfitting and improve generalization.

5. USING DECISION TREE ALGORITHM

Results

	precision	recall	f1-score	support
A	0.966102	0.934426	0.950000	122.000000
B	0.725806	0.849057	0.782609	53.000000
C	0.907895	0.873418	0.890323	79.000000
D	0.884615	0.851852	0.867925	27.000000
E	1.000000	0.500000	0.666667	2.000000
accuracy	0.890459	0.890459	0.890459	0.890459
macro avg	0.896884	0.801750	0.831504	283.000000
weighted avg	0.897316	0.890459	0.892159	283.000000



- High F1-scores for "A" (0.95) and "C" (0.89).
- Better F1-scores for "B" (0.78) and "D" (0.87) compared to the decision tree.
- Higher precision (1.0) and F1-score (0.67) than both the decision tree and the optimistic approach.
- XGBoost outperforms the decision tree (accuracy 0.89 vs. 0.81) and provides more balanced results than the optimistic approach.

THANK YOU !