פרוייקט גמר בקורס:
מבוא למדעי הנתונים

מגיש: נועם דמארי

שלבי הפרוייקט:

1 - בחירת נושא ושאלת מחקר

2 - הרכשת הנתונים

3 - ניקוי וטיוב הנתונים

4 - EDA ויזואליזציות

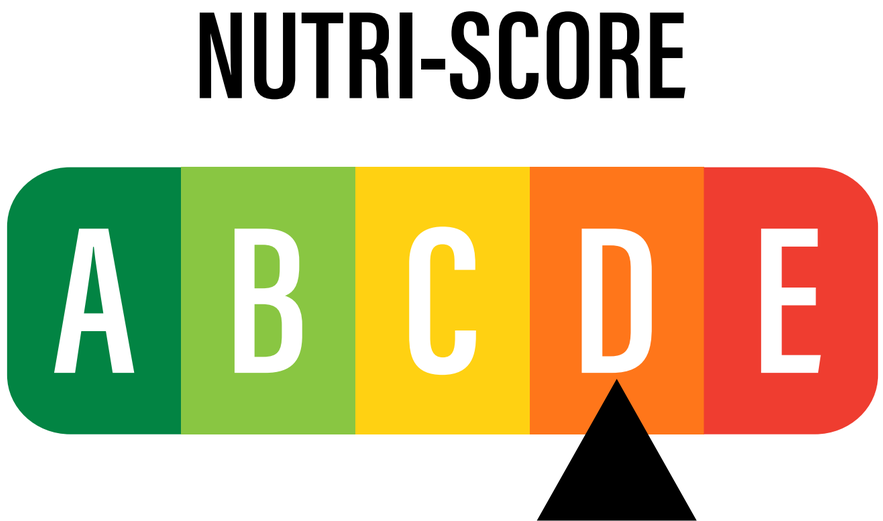
5 - למידת מכונה

6 - מסקנות וסיכום

נושא הפרוייקט – מוצרי מזון

האם ניתן לחזות ציון תזונתי של מוצרי מזון
על פי ערכיהם התזונתיים?

איזה ערך תזונתי משפיע הכי הרבה על הציון?
איזה ערך הכי פחות?



הרכשת הנתונים

https://world.openfoodfacts.org/

- open food facts
- crawling
- selenium

https://world.openfoodfacts.org/

open FOOD facts

Search for a product

Discover Contribute GET THE APP

Eau de source - Cristaline - 1,5 L

Nutella - Ferrero - 400 g

smbcmdns - Lu - 300 g

Coca-cola - 330 mL

Nutella - Ferrero - 1 kg

Coca Cola Zero - 330 ml

Sésame - Gerblé - 230g

Nutella biscuits - 304 g

Cruesli Mélange De Noix - Quaker - 450 g

Céréales Chocapic - Nestlé - 430 g

Pur beurre de cacahuète - Jardin Bio - 350 g

Flocons d'avoine céréale complète - Bjorg - 500 g

https://world.openfoodfacts.org/

Nutella - Ferrero - 400 g

NUTRI-SCORE A B C D E

NOVA 4

eco score D

Nutrition facts	As sold for 100 g / 100 ml	As sold per serving (15g)	Compared to: Cocoa and hazelnuts spreads
Energy	2,252 kj (539 kcal)	338 kj (80 kcal)	-1%
Fat	30.9 g	4.63 g	-9%
Saturated fat	10.6 g	1.59 g	+43%
Carbohydrates	57.5 g	8.62 g	+11%
Sugars	56.3 g	8.44 g	+20%
Fiber	0 g	0 g	-100%
Proteins	6.3 g	0.945 g	-3%
Salt	0.107 g	0.016 g	+3%
Alcohol	0 % vol	0 % vol	
Fruits, vegetables, nuts and rapeseed, walnut and olive oils (estimate from ingredients list analysis)	0 %	0 %	

	Items	Nutri-Score	Energy	Fat	Saturated fat	Carbohydrates	Sugars	Fiber	Proteins	Salt
0	Eau de source - Cristaline - 1,5 L	A	0	0	0	0	0	0	0	0.052
1	Nutella - Ferrero - 400 g	E	2,252	30.9	10.6	57.5	56.3	0	6.3	0.107
2	Prince - Lu - 300 g	E	1,962	17	5.6	69	32	0	6.3	0.49
3	Coca-cola - 330 mL	E	180	0	0	10.6	10.6	?	0	0
4	Nutella - Ferrero - 1 kg	E	2,252	30.9	10.6	57.5	56.3	0	6.3	0.107
...	...	...	...	...	...	...	...	...	...	...
5748	Lentilles vertes - Carrefour - 500 g	A	1,381	1.8	0.2	45	1.1	16	25	0.05
5749	Compote aromatisé à la pomme et la fraise - Ve...	B	195	0.2	0	10	10	1.6	0.3	0
5750	Pulco citronnade - Citron - Framboise	D	88	0	0	4.9	4.7	?	0	0
5751	Atún claro en aceite de oliva - Hacendado	D	1,063	19.3	3.3	0	0	0	20	3
5752	Flocons d'avoine bio - AUCHAN - 500 g	A	1,544	6.9	1.2	59	1	10	12	0.01

5753 rows × 10 columns

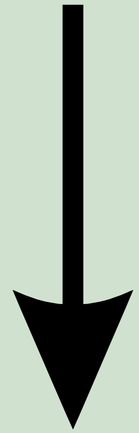
ניקוי וטיוב הנתונים

- הסרת כפילויות
- טיפול בערכים חסרים
- התאמת הטיפוס של הנתונים
- הסרת תווים מיותרים
- טיפול בחריגים

	Items	Nutri-Score	Energy	Fat	Saturated fat	Carbohydrates	Sugars	Fiber	Proteins	Salt
0	Eau de source - Cristaline - 1,5 L	A	0	0	0	0	0	0	0	0.052
1	Nutella - Ferrero - 400 g	E	2,252	30.9	10.6	57.5	56.3	0	6.3	0.107
2	Prince - Lu - 300 g	E	1,962	17	5.6	69	32	0	6.3	0.49
3	Coca-cola - 330 mL	E	180	0	0	10.6	10.6	?	0	0
4	Nutella - Ferrero - 1 kg	E	2,252	30.9	10.6	57.5	56.3	0	6.3	0.107
...	...	...	...	...	...	...	...	...	...	...
5748	Lentilles vertes - Carrefour - 500 g	A	1,381	1.8	0.2	45	1.1	16	25	0.05
5749	Compote aromatisé à la pomme et la fraise - Ve...	B	195	0.2	0	10	10	1.6	0.3	0
5750	Pulco citronnade - Citron - Framboise	D	88	0	0	4.9	4.7	?	0	0
5751	Atún claro en aceite de oliva - Hacendado	D	1,063	19.3	3.3	0	0	0	20	3
5752	Flocons d'avoine bio - AUCHAN - 500 g	A	1,544	6.9	1.2	59	1	10	12	0.01

5753 rows × 10 columns

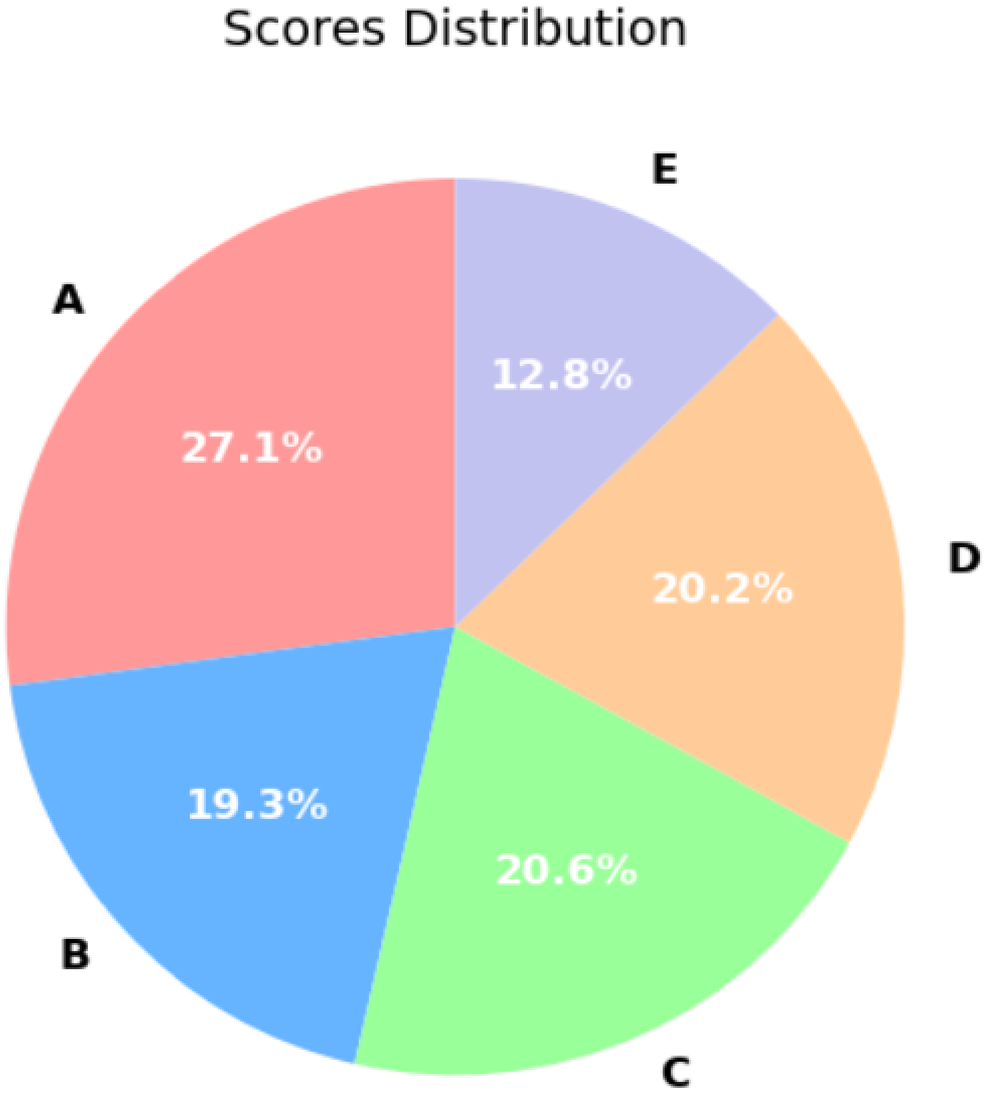
Before Cleaning



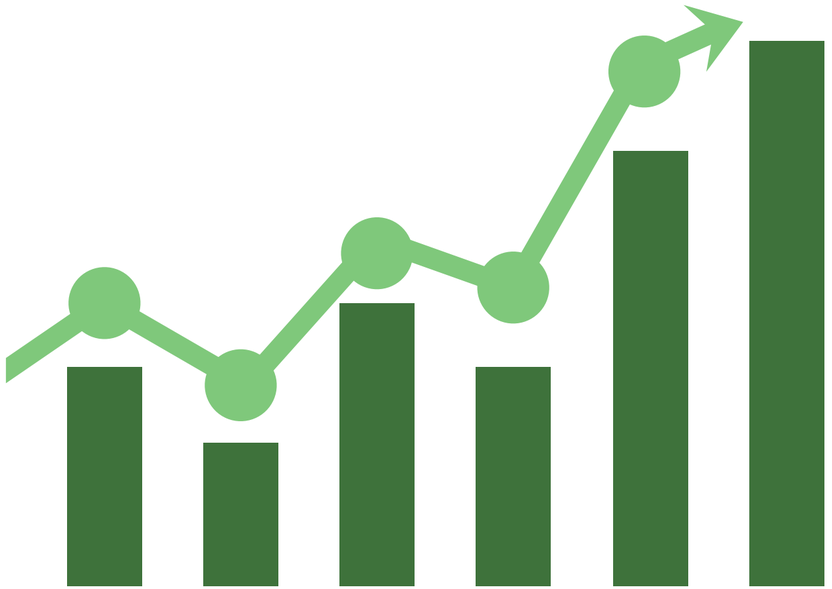
After Cleaning

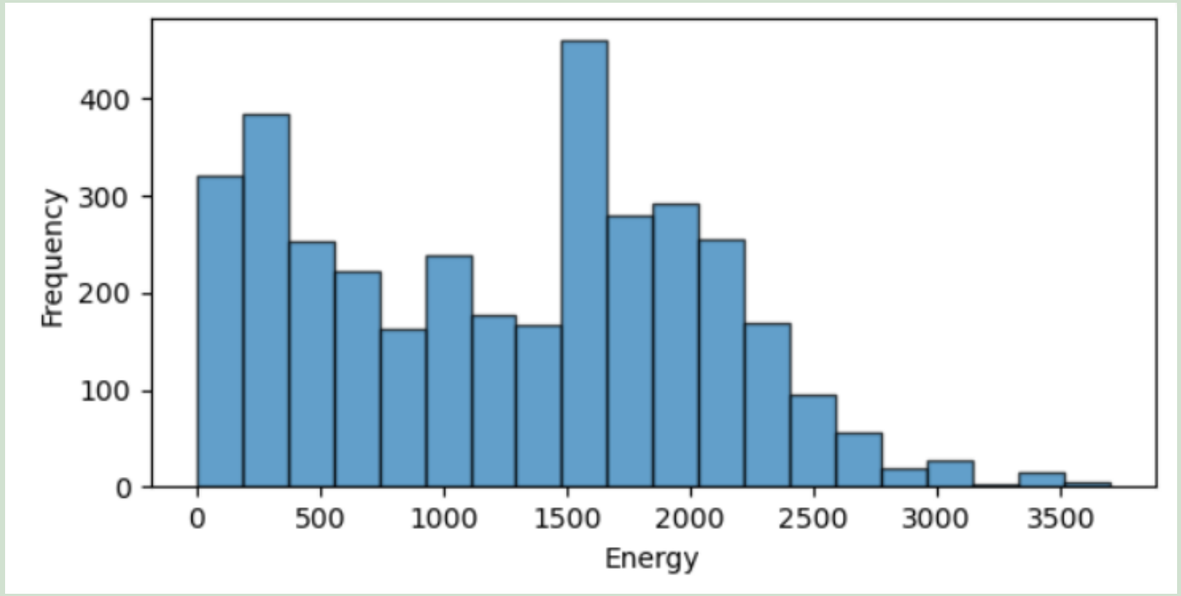
	Items	Nutri-Score	Energy	Fat	Saturated fat	Carbohydrates	Sugars	Fiber	Proteins	Salt
0	Eau de source - Cristaline - 1,5 L	A	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.052
1	Nutella - Ferrero - 400 g	E	2252.0	30.9	10.6	57.5	56.3	0.0	6.3	0.107
2	Prince - Lu - 300 g	E	1962.0	17.0	5.6	69.0	32.0	0.0	6.3	0.490
5	Coca Cola Zero - 330 ml	B	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.020
6	Sésame - Gerblé - 230g	B	1961.0	18.0	2.0	64.0	17.0	4.6	10.0	0.380
...	...	...	...	...	...	...	...	...	...	...
5747	Pringles Smokey Bacon Flavour - 175g	D	2166.0	29.0	6.3	56.0	3.0	3.6	6.6	1.600
5748	Lentilles vertes - Carrefour - 500 g	A	1381.0	1.8	0.2	45.0	1.1	16.0	25.0	0.050
5749	Compote aromatisé à la pomme et la fraise - Ve...	B	195.0	0.2	0.0	10.0	10.0	1.6	0.3	0.000
5751	Atún claro en aceite de oliva - Hacendado	D	1063.0	19.3	3.3	0.0	0.0	0.0	20.0	3.000
5752	Flocons d'avoine bio - AUCHAN - 500 g	A	1544.0	6.9	1.2	59.0	1.0	10.0	12.0	0.010

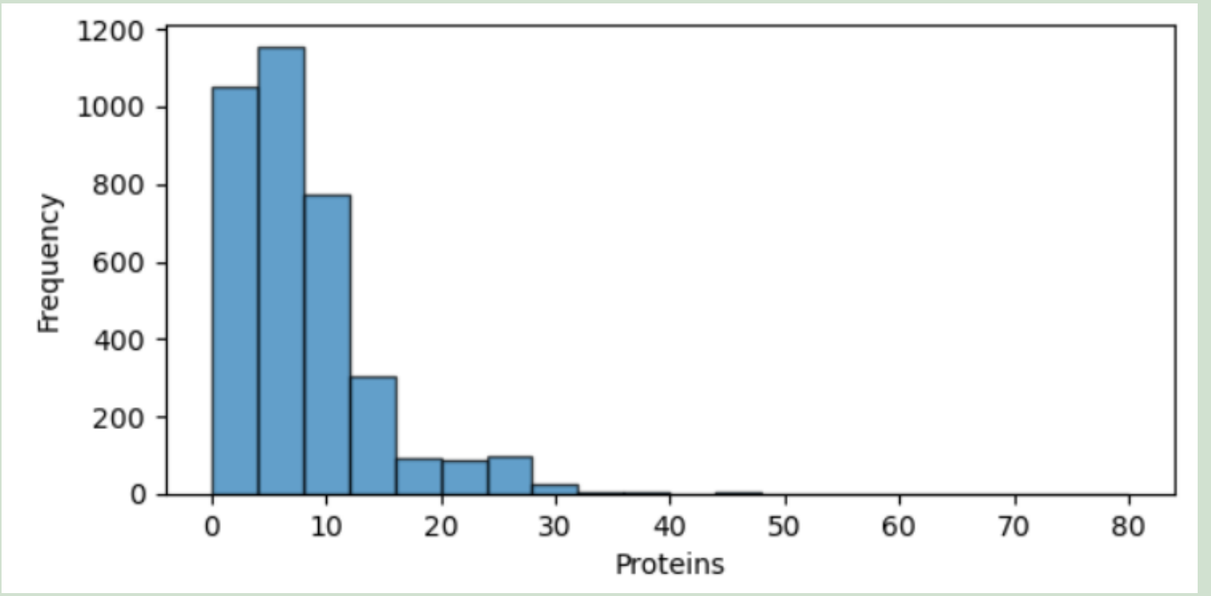
3601 rows × 10 columns

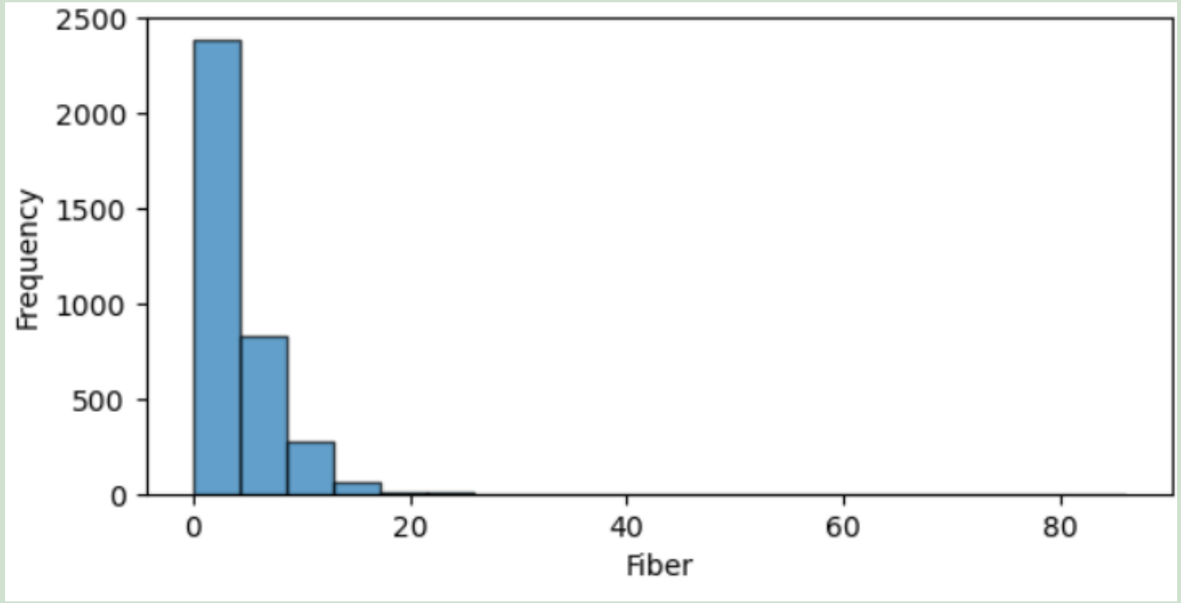


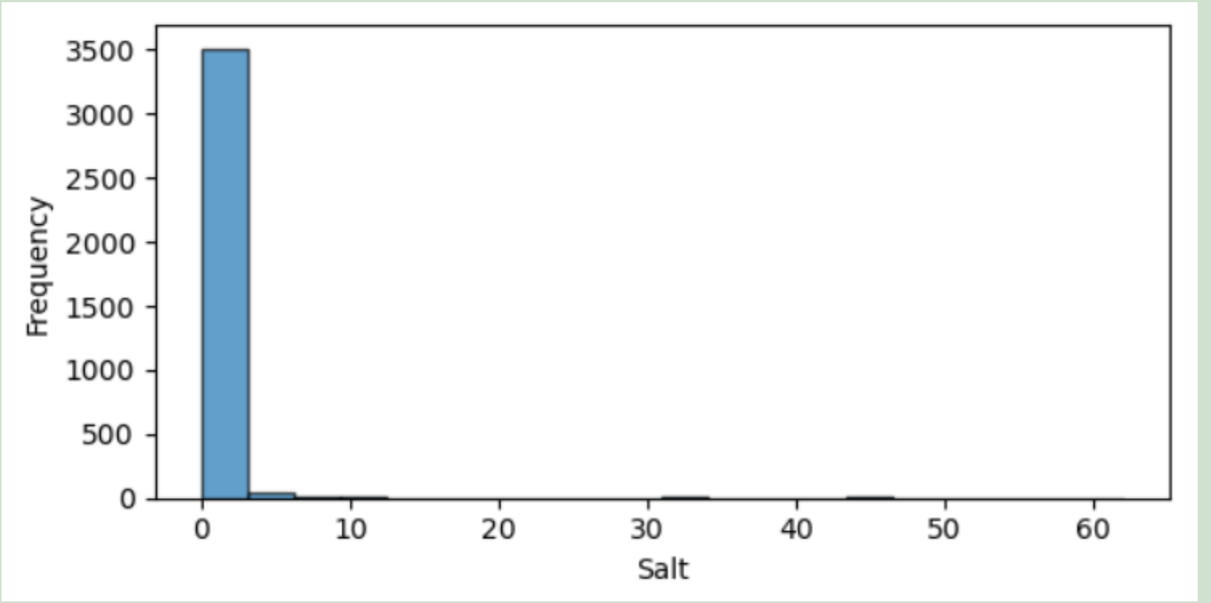
EDA

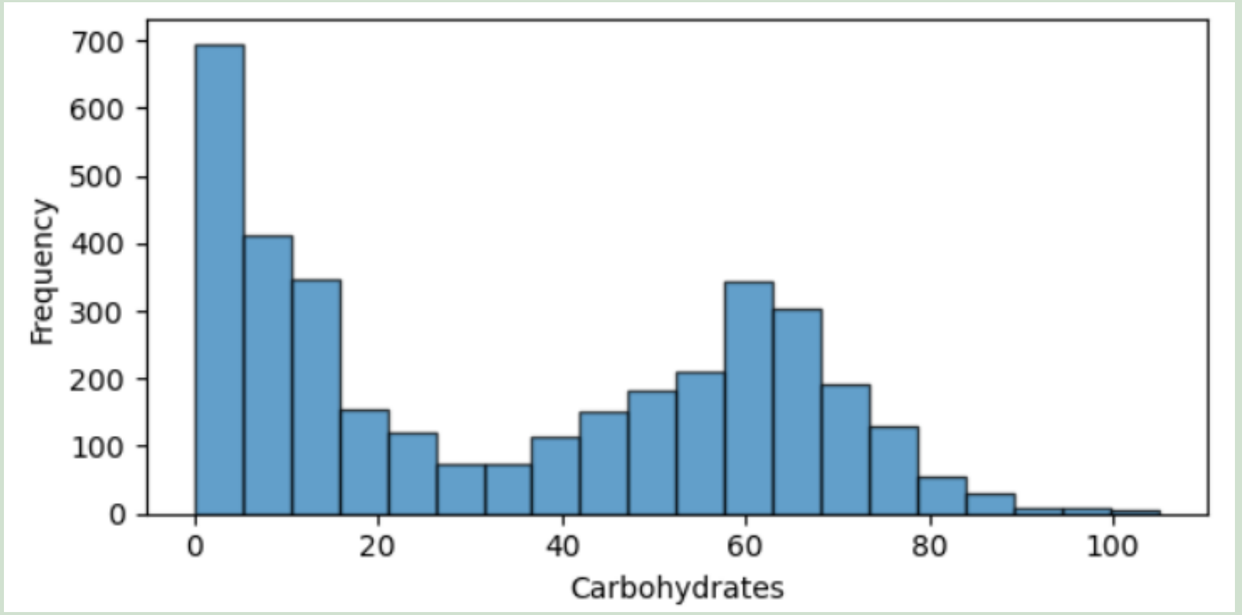


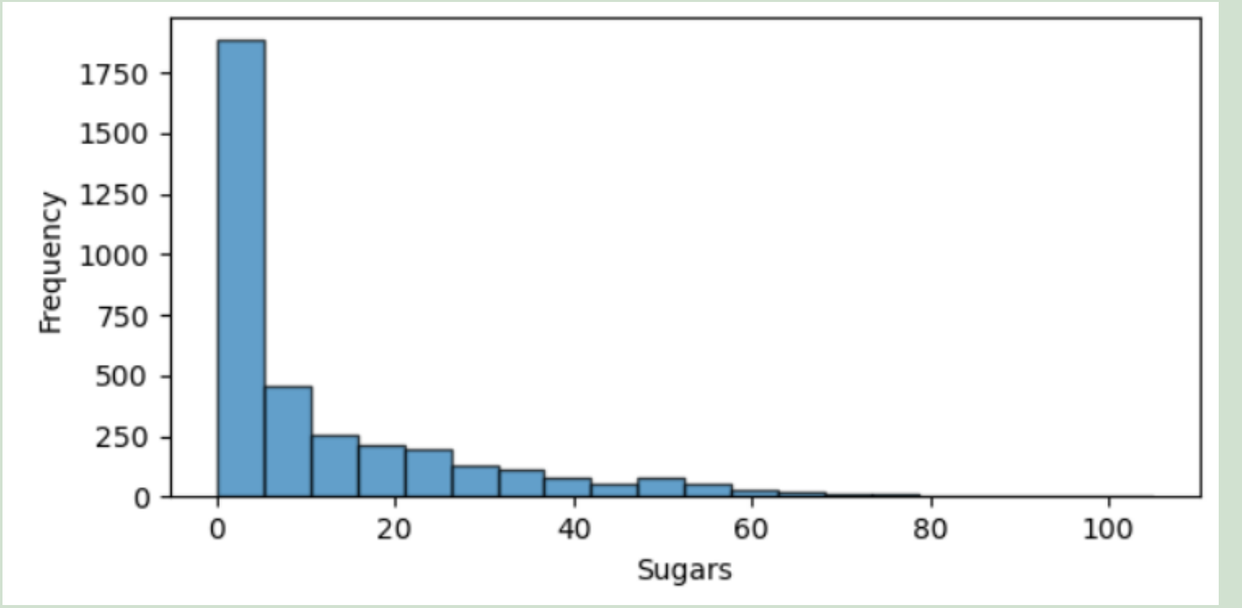

Frequency
Energy
0
100
200
300
400
0
500
1000
1500
2000
2500
3000
3500

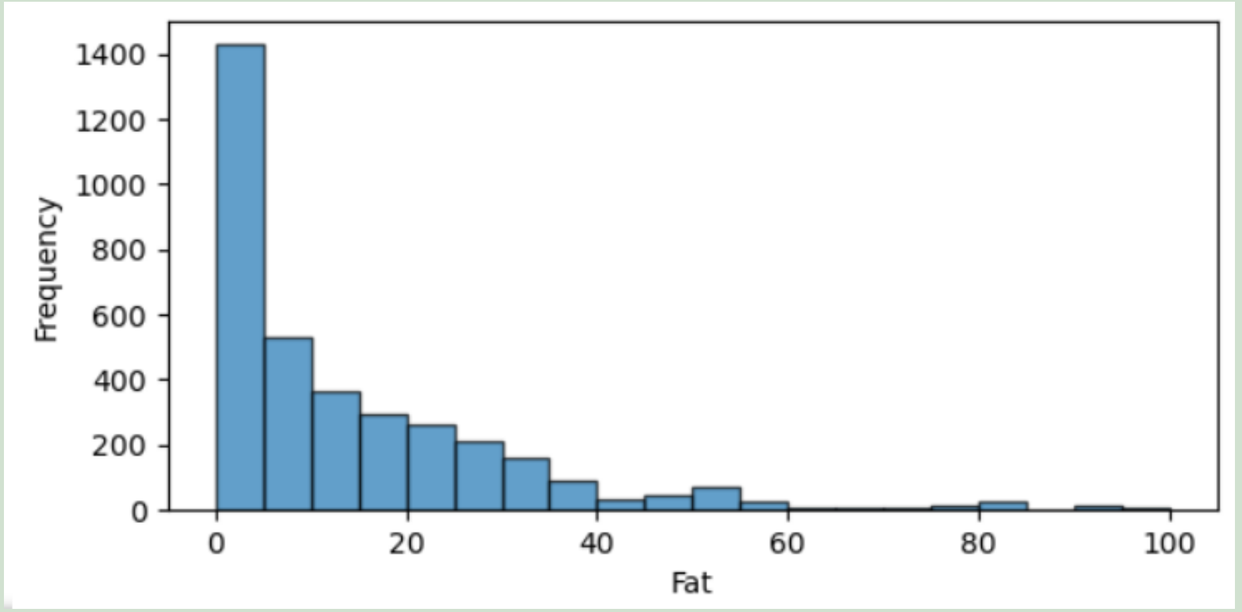

Frequency
Proteins
0
200
400
600
800
1000
1200
0
10
20
30
40
50
60
70
80

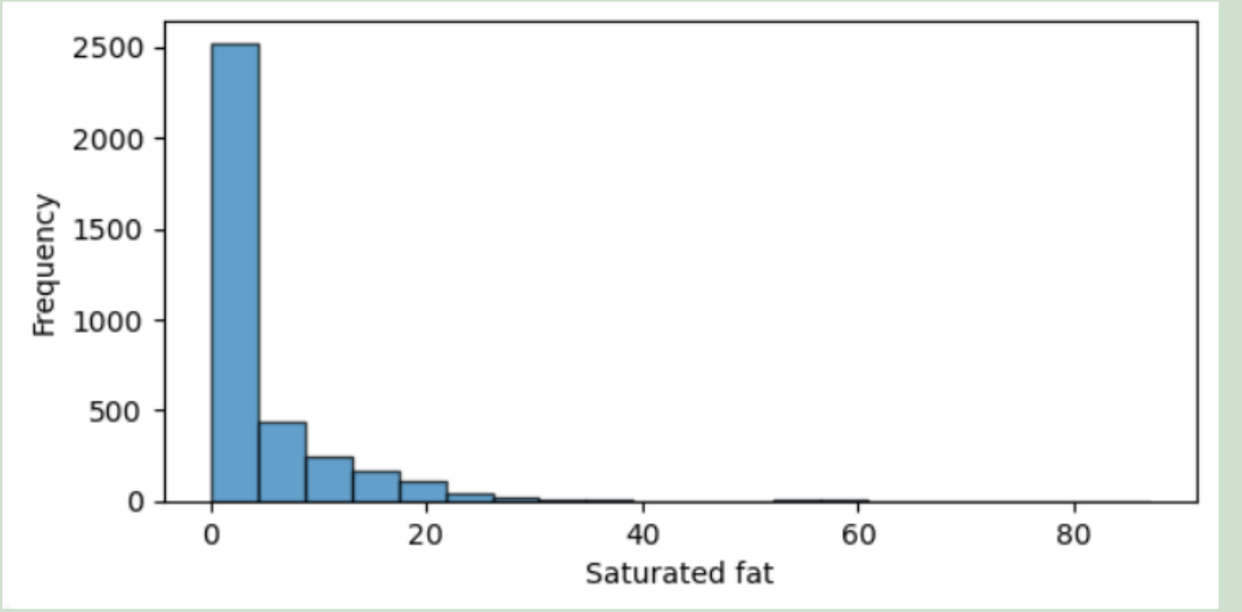

Frequency
Fiber
0
500
1000
1500
2000
2500
0
20
40
60
80

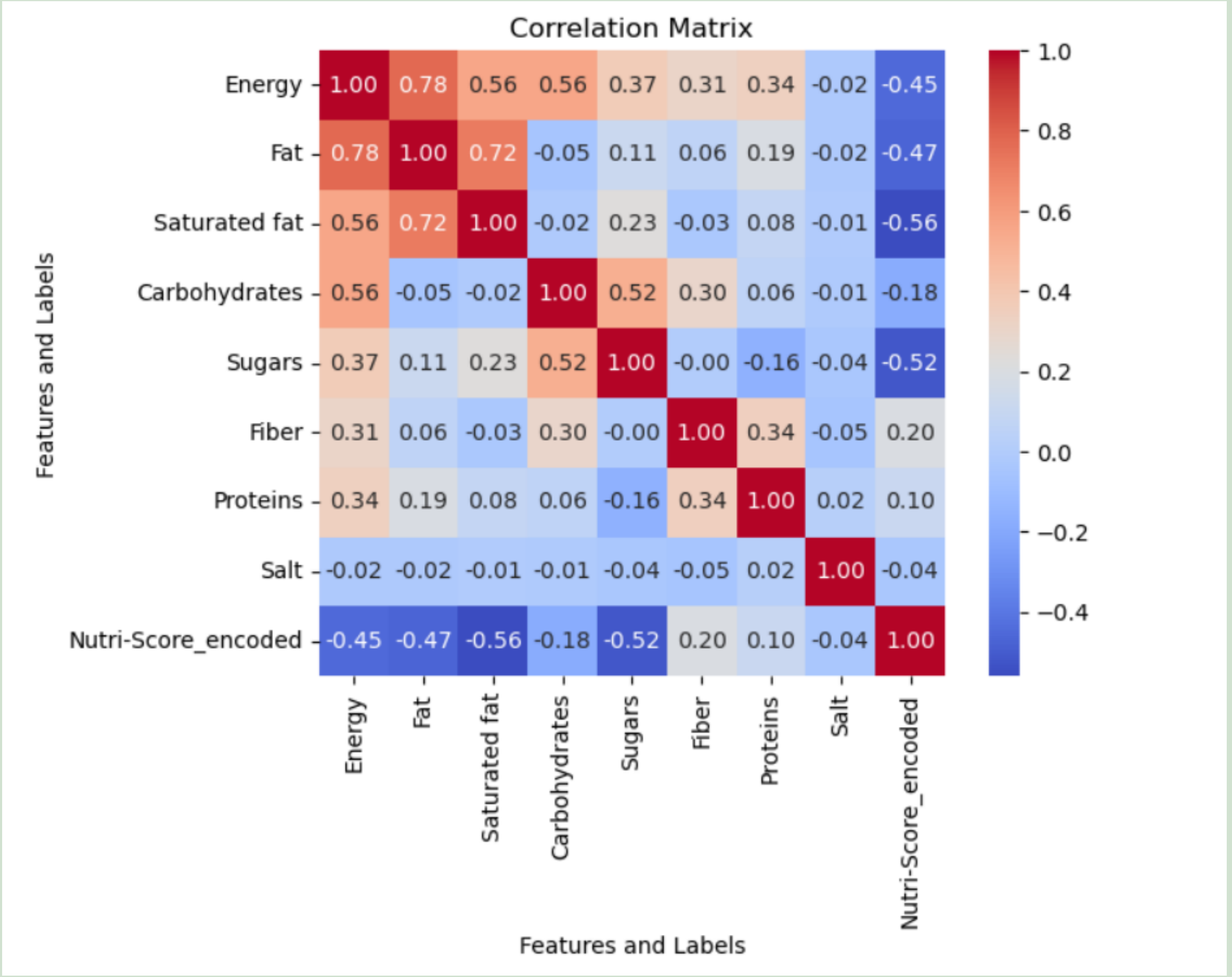

Frequency
Salt
0
500
1000
1500
2000
2500
3000
3500
0
10
20
30
40
50
60

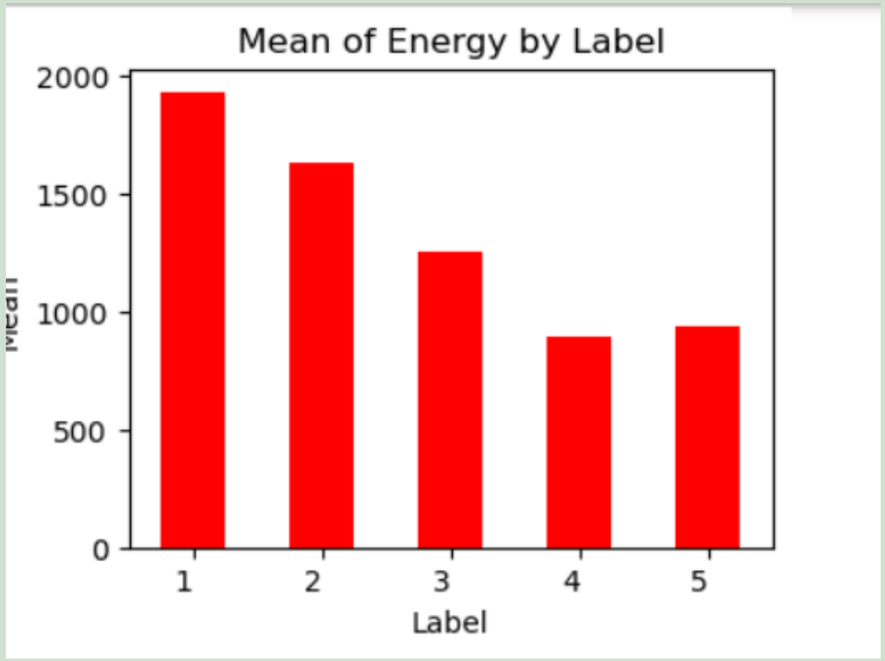

Frequency
Carbohydrates
0
100
200
300
400
500
600
700
0
20
40
60
80
100

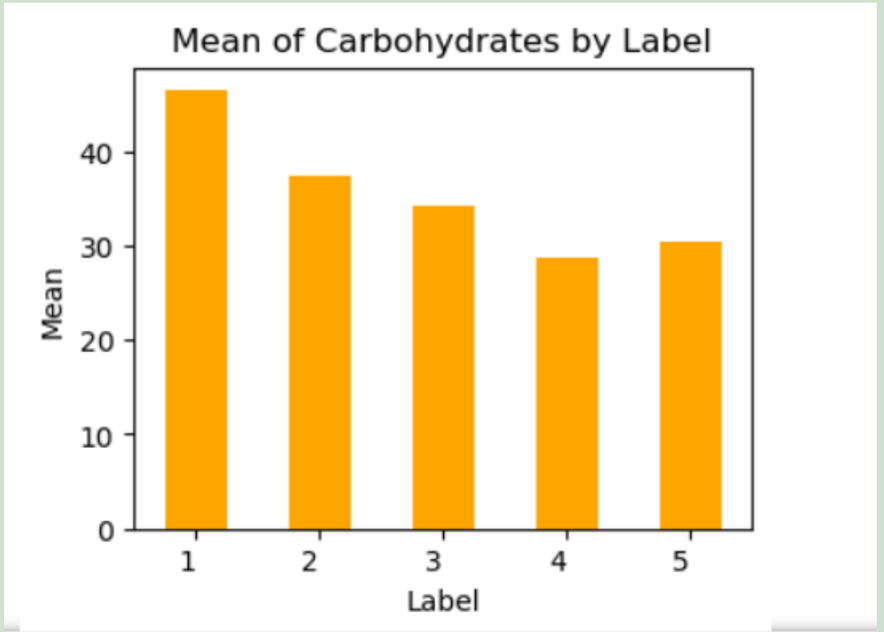

Frequency
Sugars
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100

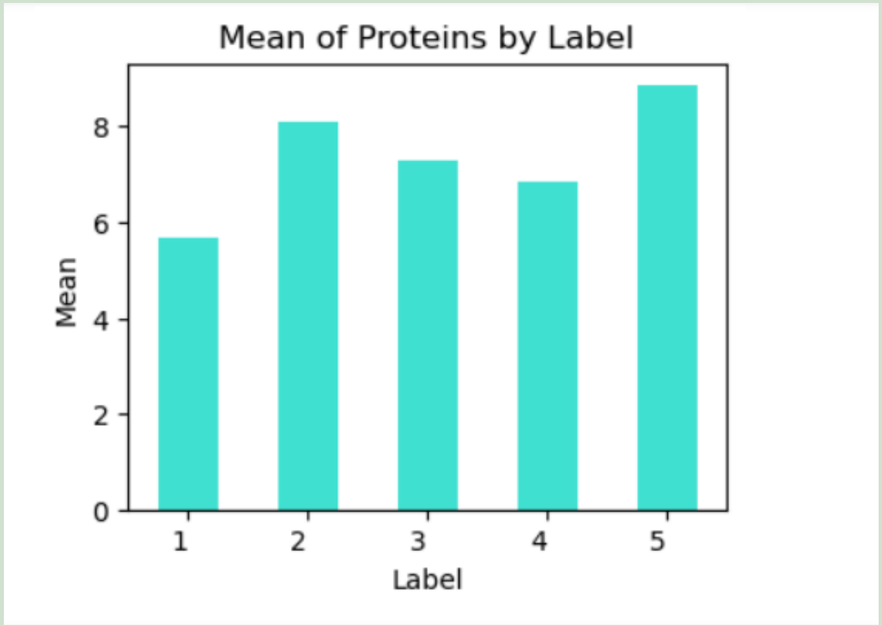

Frequency
Fat
0
200
400
600
800
1000
1200
1400
0
20
40
60
80
100

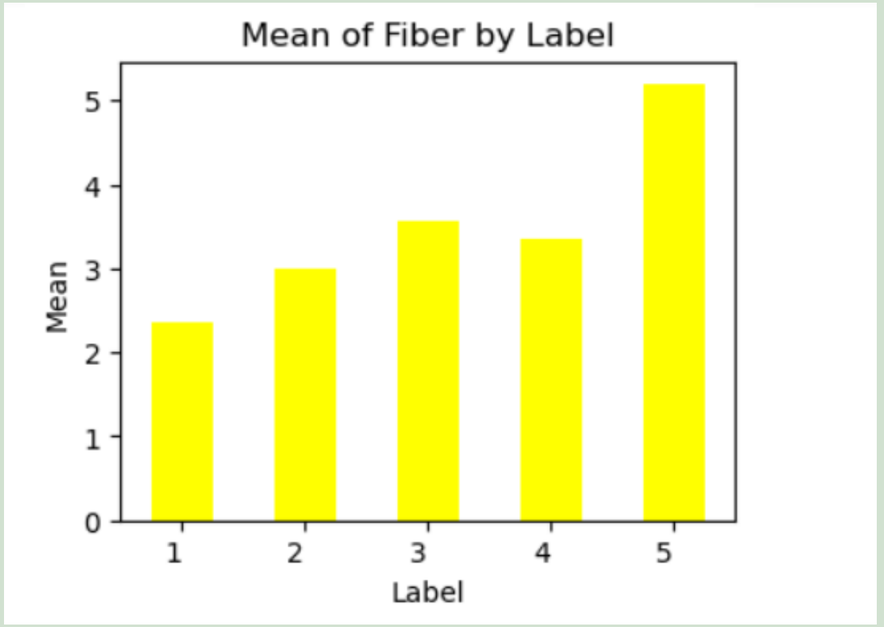

Frequency
Saturated fat
0
500
1000
1500
2000
2500
0
20
40
60
80

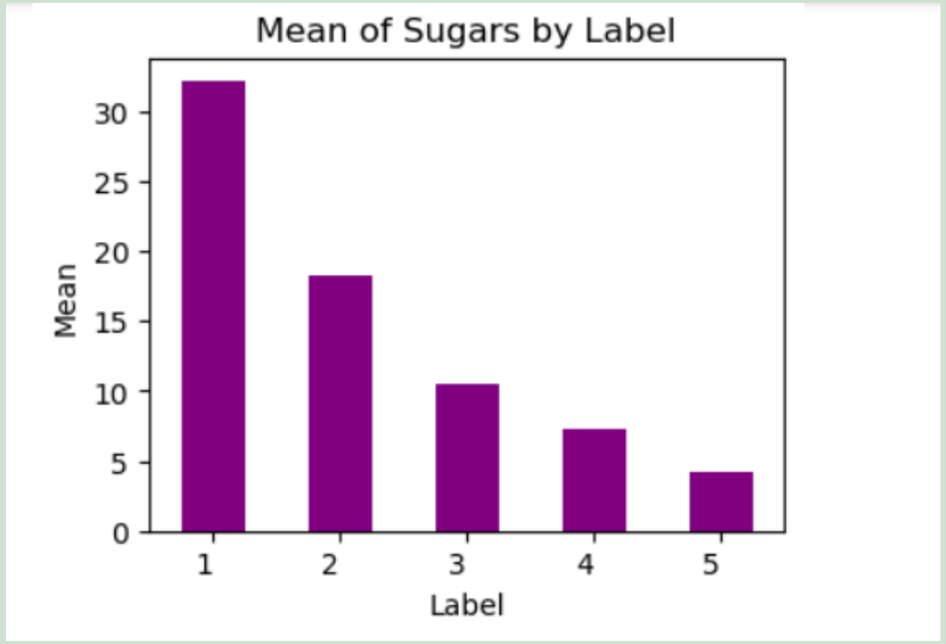

Correlation Matrix
Energy 1.00 0.78 0.56 0.56 0.37 0.31 0.34 -0.02 -0.45
Fat 0.78 1.00 0.72 -0.05 0.11 0.06 0.19 -0.02 -0.47
Saturated fat 0.56 0.72 1.00 -0.02 0.23 -0.03 0.08 -0.01 -0.56
Carbohydrates 0.56 -0.05 -0.02 1.00 0.52 0.30 0.06 -0.01 -0.18
Sugars 0.37 0.11 0.23 0.52 1.00 -0.00 -0.16 -0.04 -0.52
Fiber 0.31 0.06 -0.03 0.30 -0.00 1.00 0.34 -0.05 0.20
Proteins 0.34 0.19 0.08 0.06 -0.16 0.34 1.00 0.02 0.10
Salt -0.02 -0.02 -0.01 -0.01 -0.04 -0.05 0.02 1.00 -0.04
Nutri-Score_encoded -0.45 -0.47 -0.56 -0.18 -0.52 0.20 0.10 -0.04 1.00
Features and Labels
Energy
Fat
Saturated fat
Carbohydrates
Sugars
Fiber
Proteins
Salt
Nutri-Score_encoded
Features and Labels
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

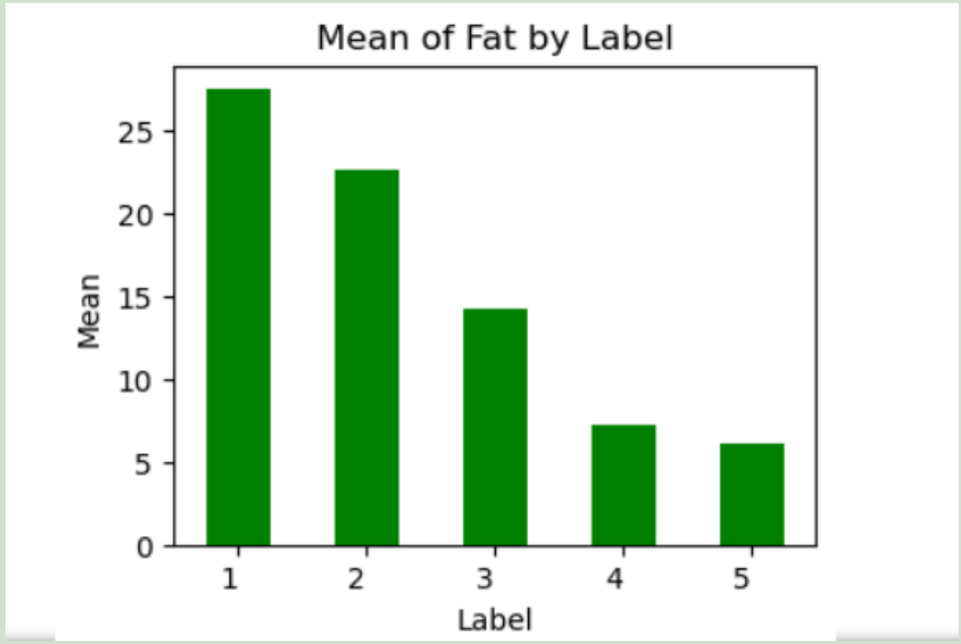

Mean of Energy by Label
2000
1500
1000
500
0
1
2
3
4
5
Label

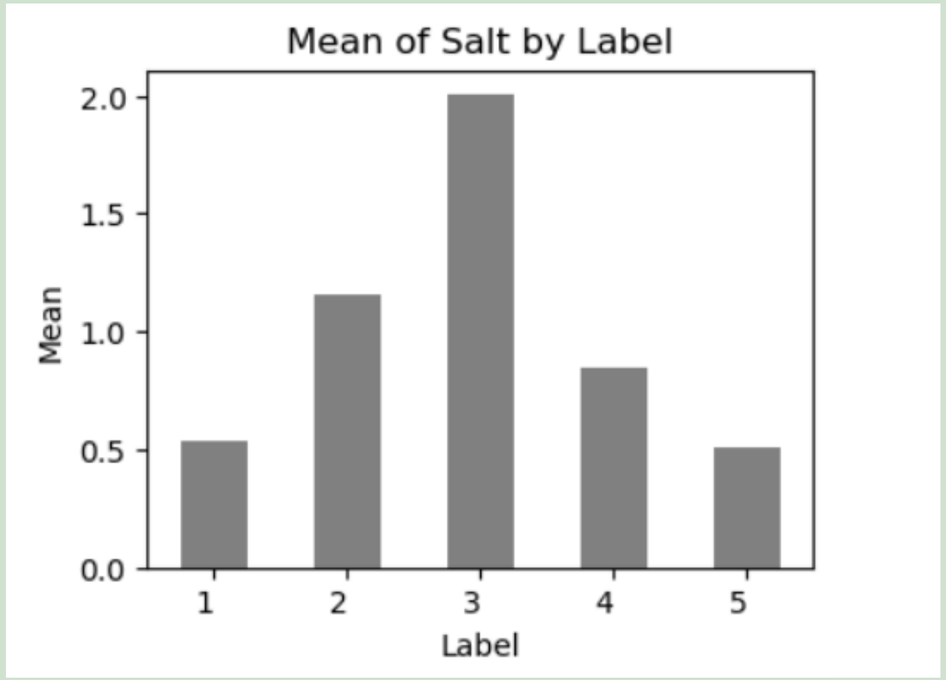

Mean of Carbohydrates by Label
40
30
20
10
0
Mean
1
2
3
4
5
Label

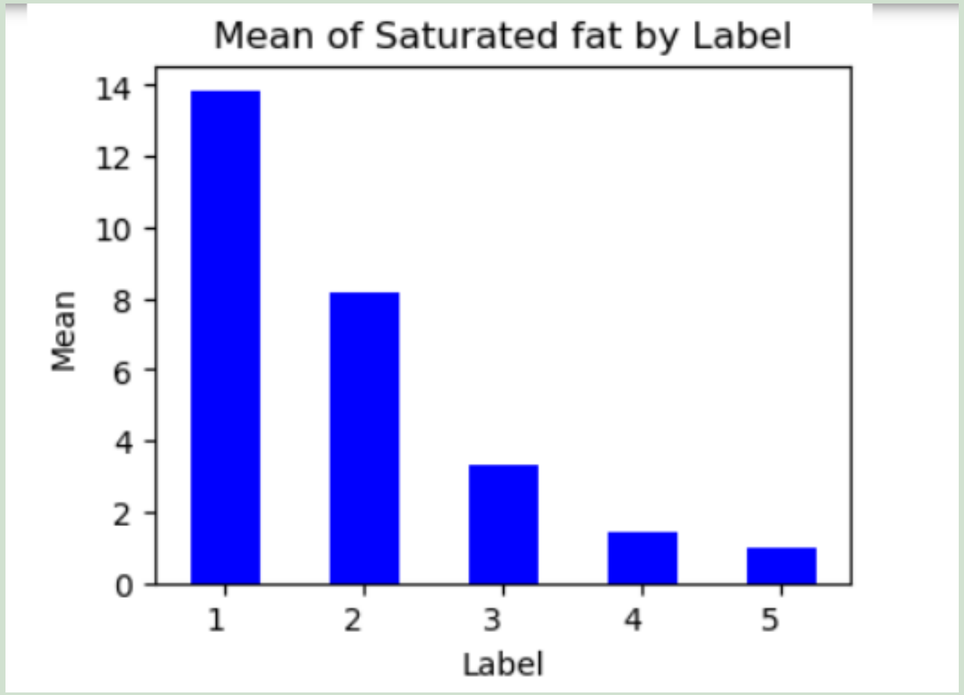

Mean of Proteins by Label
8
6
4
2
0
Mean
1
2
3
4
5
Label


Mean of Fiber by Label
5
4
3
2
1
0
Mean
1
2
3
4
5
Label

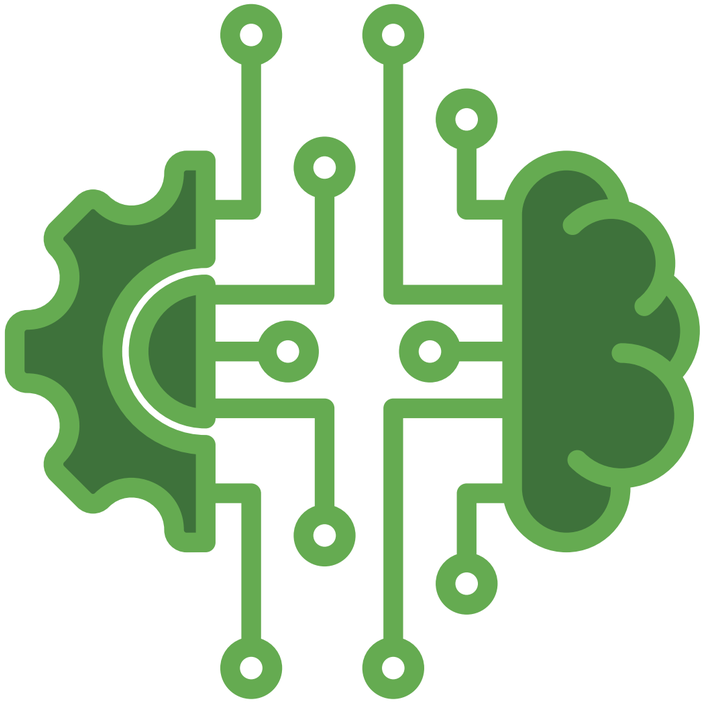

Mean of Sugars by Label
Mean
0
5
10
15
20
25
30
1
2
3
4
5
Label


Mean of Fat by Label
Mean
0
5
10
15
20
25
1
2
3
4
5
Label


Mean of Salt by Label
Mean
0.0
0.5
1.0
1.5
2.0
1
2
3
4
5
Label


Mean of Saturated fat by Label
Mean
0
2
4
6
8
10
12
14
1
2
3
4
5
Label

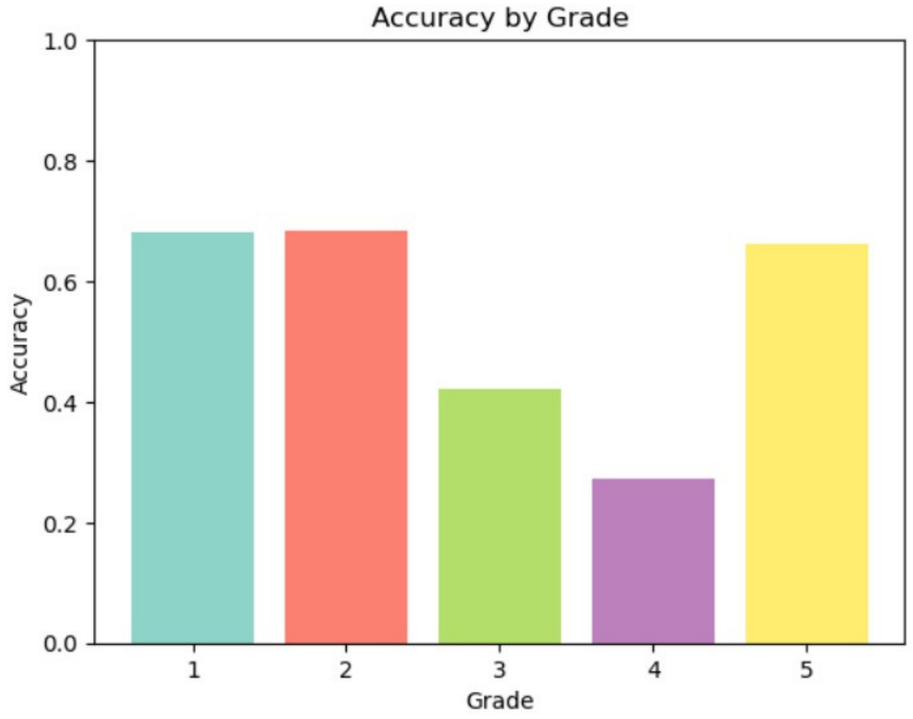
Machine Learning



- KNN
- Naive Bayes
- Decision Tree
- Random Forest

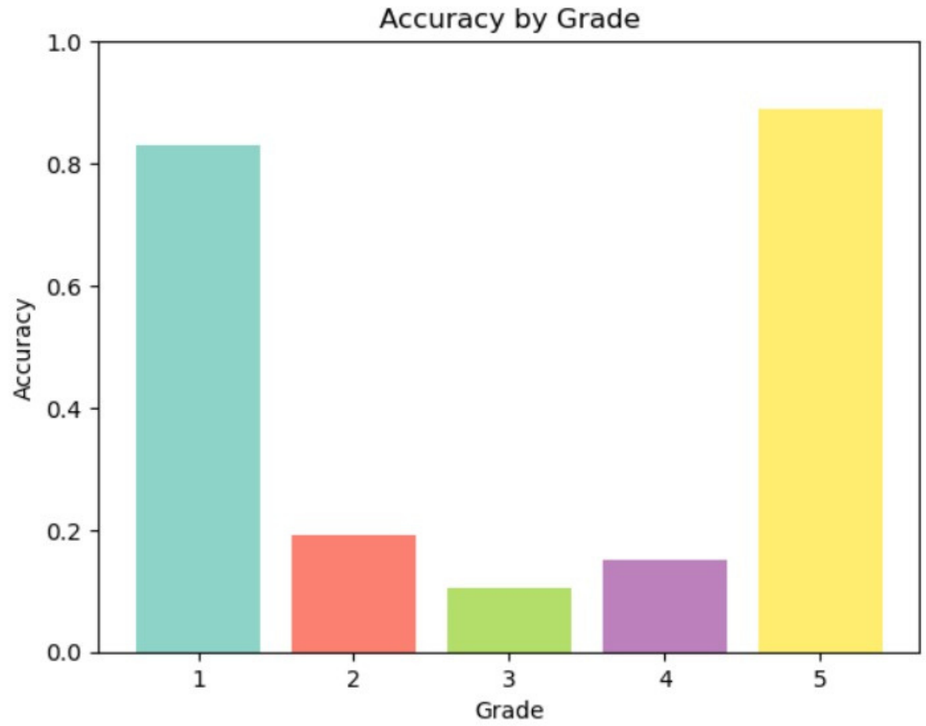
KNN

```
R2 score: 0.37049427034166293
MAE: 0.6532593619972261
Precision: 0.5392938582251237
Recall: 0.5444668250649637
F1 score: 0.5375789668419108
accuracy: 0.536754507628294
```



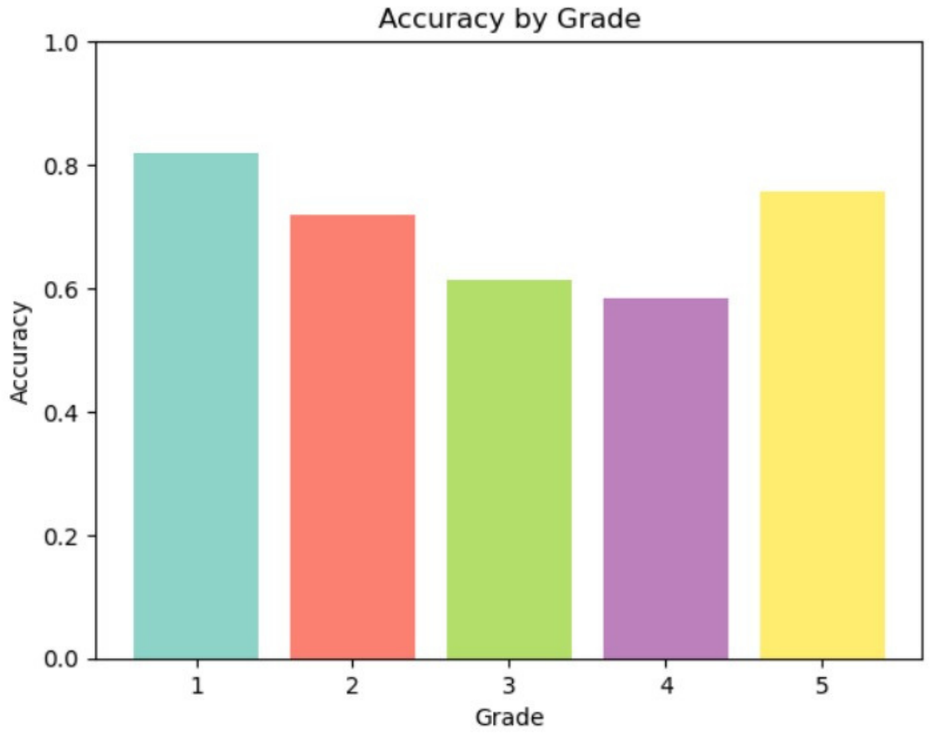
Naive Bayes

```
R2 score: 0.11029856874955024
MAE: 0.8765603328710125
Precision: 0.4445115771488696
Recall: 0.4331062713264401
F1 score: 0.35239351044950473
accuracy: 0.4160887656033287
```



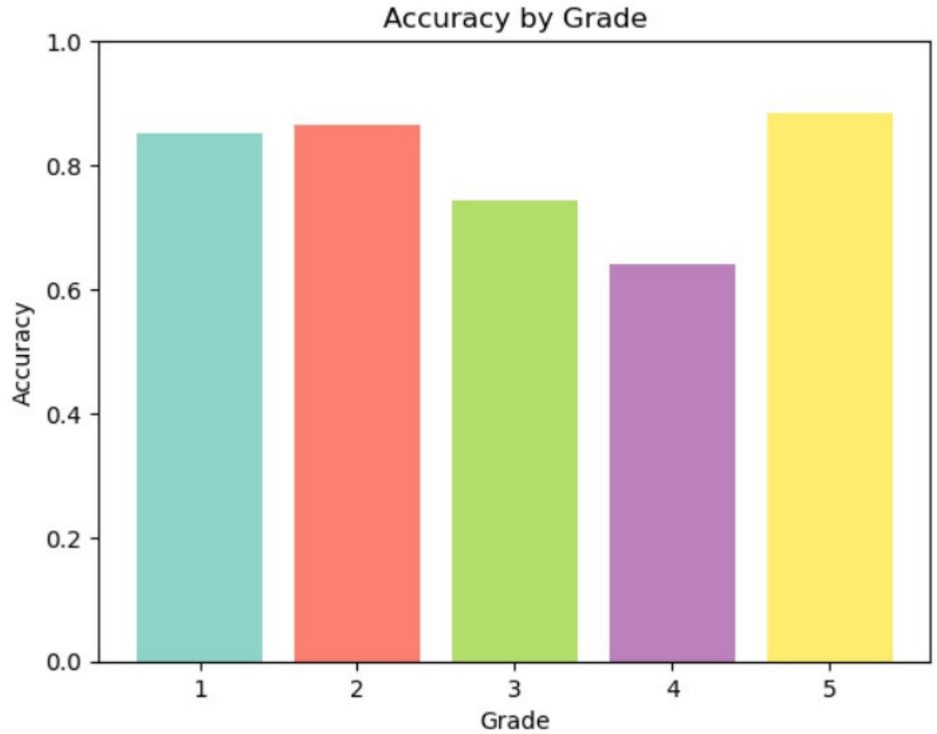
Decision Tree

R2 score: 0.7130979947254124
MAE: 0.37447988904299584
Precision: 0.7015875652646616
Recall: 0.6980333310707506
F1 score: 0.699554880406946
accuracy: 0.6893203883495146



Random Forest

R2 score: 0.8130558742226757
MAE: 0.24271844660194175
Precision: 0.804068781786253
Recall: 0.797085127837502
F1 score: 0.7992743296896199
accuracy: 0.7961165048543689



סיכום ומסקנות –

הציון התזונתי ניתן לחיזוי בדיוק של **80%**
באמצעות מודל **Random Forest**

הציונים **'D'** , **'A'** ו-**'E'** הם הציונים הניתנים
לחיזוי בדיוק הגבוהה ביותר

ערכים תזונתיים כמו סוכר, שומן ושומן רווי בעלי השפעה שלילית
על הציון התזונתי

ערכים תזונתיים כמו חלבונים וסיבים הם בעלי השפעה חיובית
על הציון התזונתי