

רגרסיות לחג הפסח

1. פתור באמצעות **polynomial-regression**

פיצ'ר 1

חוקר חקלאי בודק את הקשר בין כמות הדשן (בקילוגרם) לבין יכול תפוחי אדמה (בטונות לדונם). הוא מגלה שיש יחס לא-ליניארי - יותר מדי דשן עלול לפגוע ביבול. הנה הנתונים:

fertilizer Kg (x_i)	Potatoes t(y_i)	x_i^2	$x_i \cdot y_i$	$x_i^2 \cdot y_i$
50	7.5	—	—	—
100	10.2	—	—	—
150	12.8	—	—	—
200	14.5	—	—	—
250	15.6	—	—	—
300	16.0	—	—	—
350	15.8	—	—	—
400	15.0	—	—	—
450	13.5	—	—	—
500	11.2	—	—	—
TOTAL:	TOTAL:	TOTAL:	TOTAL:	TOTAL:

1. חשב את הערכים החסרים בטבלה (באמצעות קוד פייטון)
2. בנה מודל רגרסיה פולינומיאלית (מדרגה 2) שמתאר את הקשר בין כמות הדשן לבין היבול
3. מהי כמות הדשן האופטימלית שתביא ליבול מקסימלי?
4. מה היבול המקסימלי הצפוי לפי המודל?

המשך בעמוד הבא ...

2. פתור באמצעות **polynomial-regression** (מספר פיצ'רים)

חוקר חקלאי בודק את הקשר בין שלושה גורמים - כמות הדשן (בק"ג), השקיה (ליטרים ליום), וטמפרטורה (מעלות צלזיוס) - לבין יכול תירס (בטונות לדונם). הוא חושד שיש יחסים לא-ליניאריים מורכבים בין הגורמים. הנה חלק מהנתונים:

יכול תירס (טון/דונם)	טמפרטורה (°C)	השקיה (ליטר/יום)	דשן (ק"ג)
2.5	20	5	100
3.2	20	5	150
3.8	20	5	200
4.1	20	5	250
3.9	20	5	300
3.0	20	3	200
3.5	20	4	200
3.9	20	6	200
3.7	20	7	200
3.2	15	5	200
3.6	18	5	200
3.7	22	5	200
3.5	25	5	200
3.1	28	5	200

1. בונה מודל רגרסיה פולינומיאלית עם שלושה משתנים מסבירים (מדרגה 3)
2. מצא את הערכים האופטימליים של דשן, השקיה וטמפרטורה שיובילו ליכול תירס מקסימלי

המשך בעמוד הבא ...

3. תרגיל: נרמול וסטנדרטיזציה

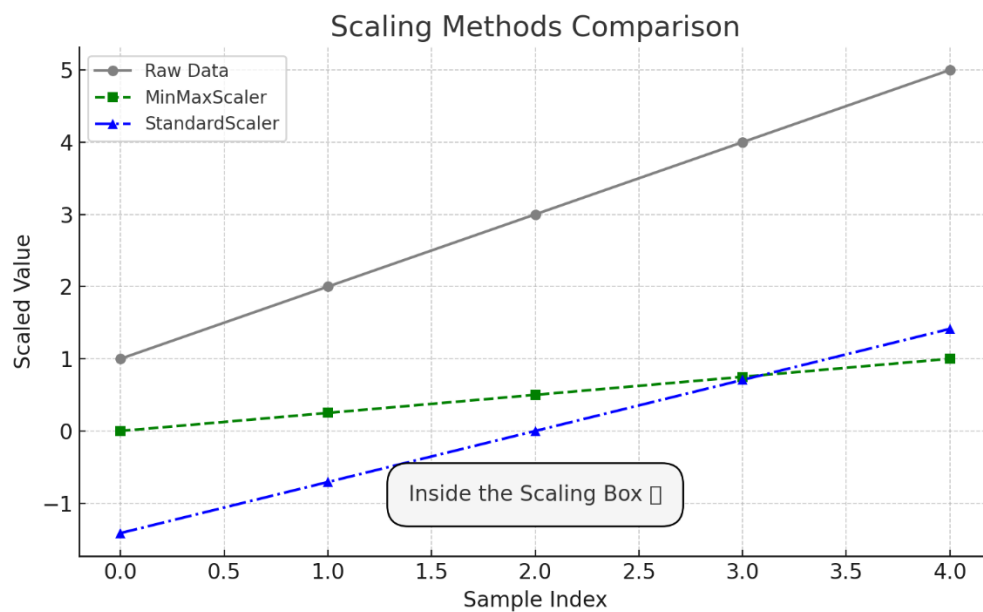
לפניך 2 פיצ'רים:

$X1 = [5, 10, 15, 20, 25, 30, 35, 40, 45, 50]$

$X2 = [100, 90, 80, 70, 60, 50, 40, 30, 20, 10]$

- הסבר בקצרה מדוע יש צורך לעשות scale לפיצ'רים במודלים
- אחד את הרשימות לתוך מטריצת פיצ'רים (כל רשימה היא עמודה)
- בצע Min-Max Scaling (כך שכל ערך יהיה בין 0 ל-1), MinMaxScaler
- בצע Standardization, StandardScaler
- הסבר בקצרה מה ההבדל בין שתי השיטות

****ציור להמחשה של פעולת scale:**



המשך בעמוד הבא

4. זיהוי ומימוש המודל הנכון

מקרה 1:

חוקר תזונה בדק את השפעת כמות שעות פעילות גופנית בשבוע על כמות הקלוריות שסטודנטים שורפים-

Exercise = [0.0, 0.53, 1.05, 1.58, 2.11, 2.63, 3.16, 3.68, 4.21, 4.74,
5.26, 5.79, 6.32, 6.84, 7.37, 7.89, 8.42, 8.95, 9.47, 10.0]

Calories = [7.99, 9.17, 11.91, 12.89, 9.70, 11.32, 15.49, 17.53, 20.27, 20.92,
24.76, 24.80, 26.36, 26.21, 29.41, 33.71, 31.72, 36.96, 32.63, 35.60]

מקרה 2:

פסיכולוג חוקר את הקשר בין רמות חרדה לבין רמת ביטחון

anxiety level = [-4.0, -3.58, -3.16, -2.74, -2.32, -1.89, -1.47, -1.05, -0.63, -0.21,
0.21, 0.63, 1.05, 1.47, 1.89, 2.32, 2.74, 3.16, 3.58, 4.0]

Self esteem = [41.02, 43.27, 39.14, 25.96, 23.08, 16.41, 14.31, 12.26, 1.91, 5.65,
3.80, -5.23, -10.89, -2.85, -3.41, -3.34, -0.75, 13.14, 11.03, 17.77]

עליך לחקור את 2 המקרים ולהחליט מה המודל המתאים ביותר:

- הצג scatter plot לשני המקרים. האם ניתן לראות מגמה?
- בנה את המודל Linear Regression למקרה הראשון:
 - חשב מה ה-R בריבוע?
 - חשב מה ה-MSE?
 - הצג גרף של נקודות התחזית + קו המתאר קו ישר
 - האם המודל מתאים?
- בנה את המודלים Linear Regression וגם Polynomial Regression למקרה השני:
 - חשב מה ה-R בריבוע לכל אחד?
 - חשב מה ה-MSE לכל אחד?
 - הצג גרף של נקודות התחזית + קו המתאר קו ישר
 - כעת כתוב מה המסקנה אליה הגעת – מי המודלים יותר מתאים

