למידת מכונה לעיבוד סיגנלים ותמונות רפואי

תרגיל בית 3

❖ מודלים שהשתמשנו

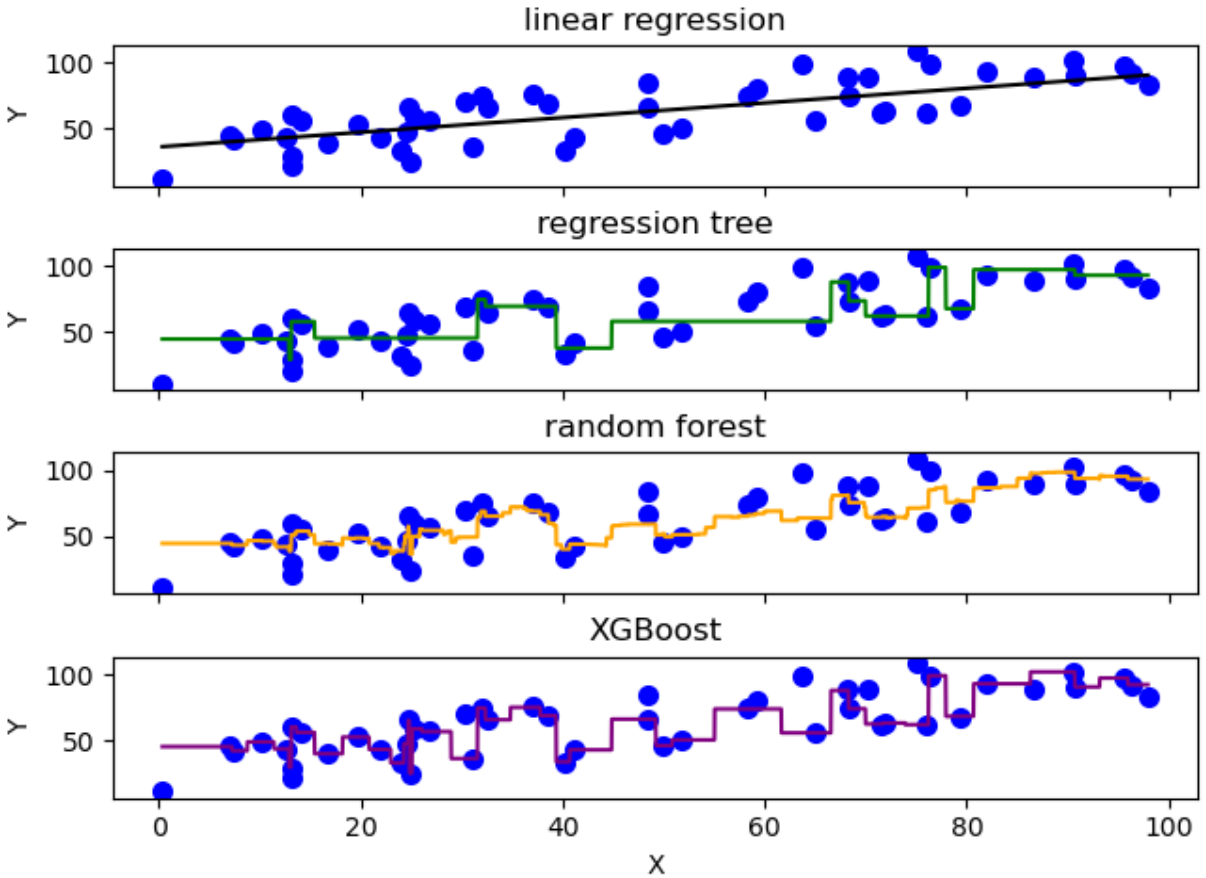
במטרה לשפר את תוצאת מודל הרגרסיה הלינארית (תרגיל מס 1)
,השתמשנו ב-3 מודלים כדי לנסות:

1) Regression tree
2) random forest
3) XGBoost

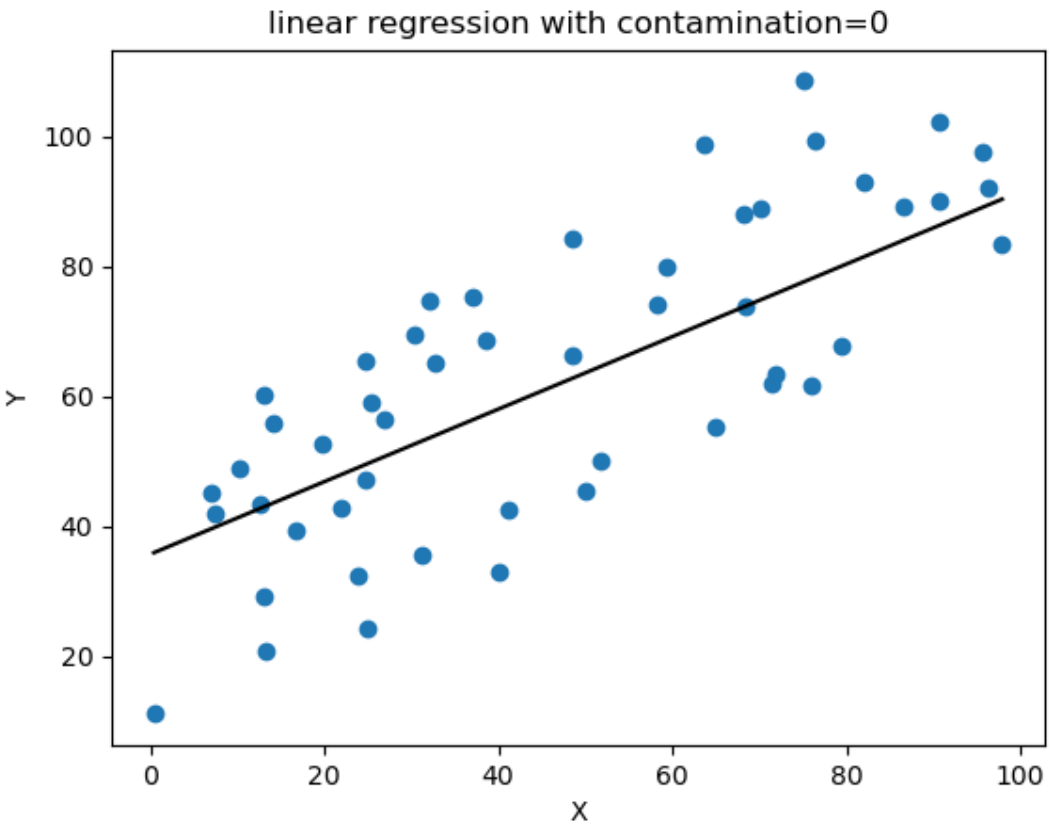
❖ הורדת נקודות מ-data + תוצאות הקוד

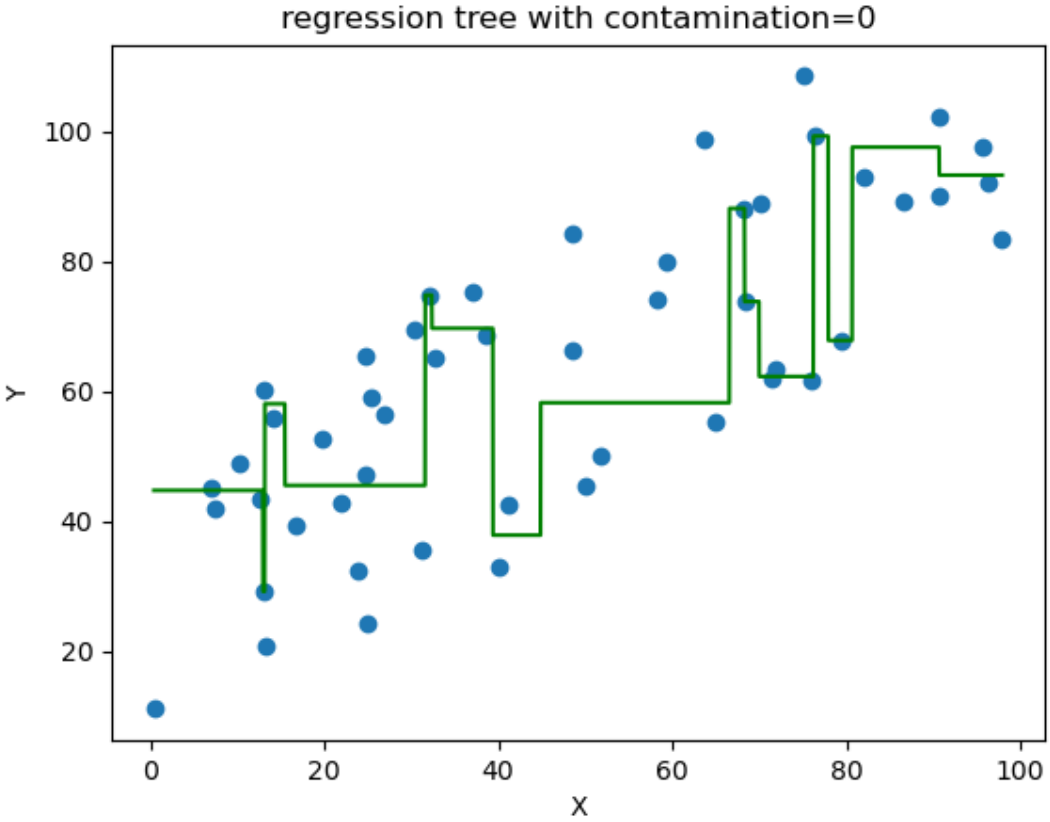
במהלך הניסוי, התכוננו לפעולה של הורדת נקודות מן הנתונים על מנת לשפר את תוצאת המודלים. כל פעם היה נדרש להתחיל מחדש בגלל הורדת נקודה מהנתונים. חלטנו בסוף להוריד רק שלושה נקודות וסיבת הורדת כל נקודה מהשלושה כי היא רחוקה משאר ה-data ומביאים סוג של רעש למודלים , ניסנו ולהוריד נקודות מכל מני סיבות אבל הכי טוב שקיבלנו זה להוריד את שלושת הנקודות הרחוקות.

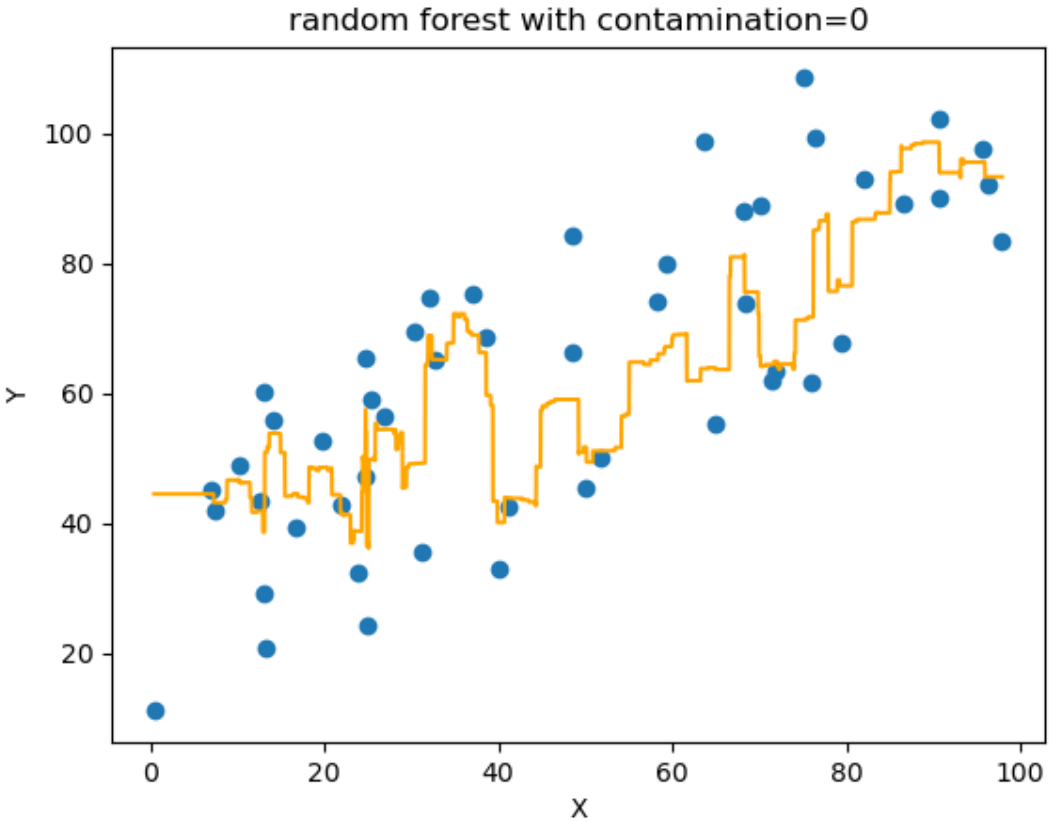
❖ בלי להוריד נקודות :

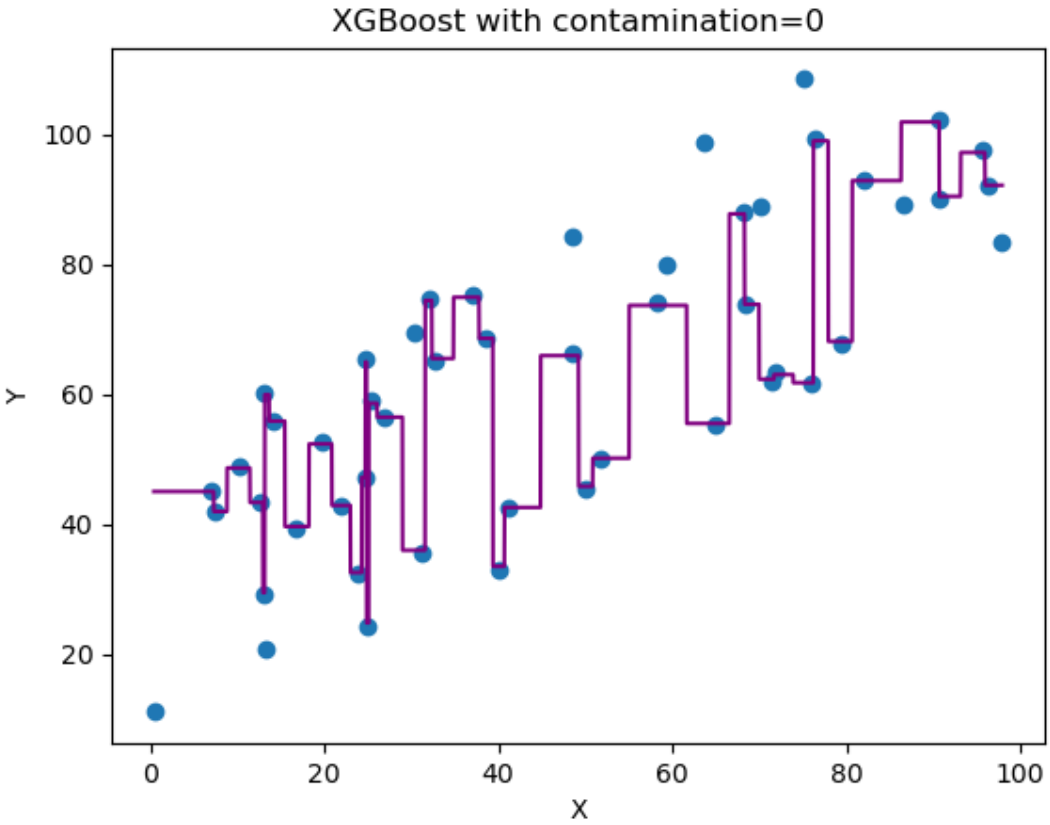


תרשים עם יותר פירוט:




regression tree with contamination=0
Y
X
0
20
40
60
80
100
20
40
60
80
100


random forest with contamination=0
Y
X
0
20
40
60
80
100
20
40
60
80
100



תוצאות הקוד:

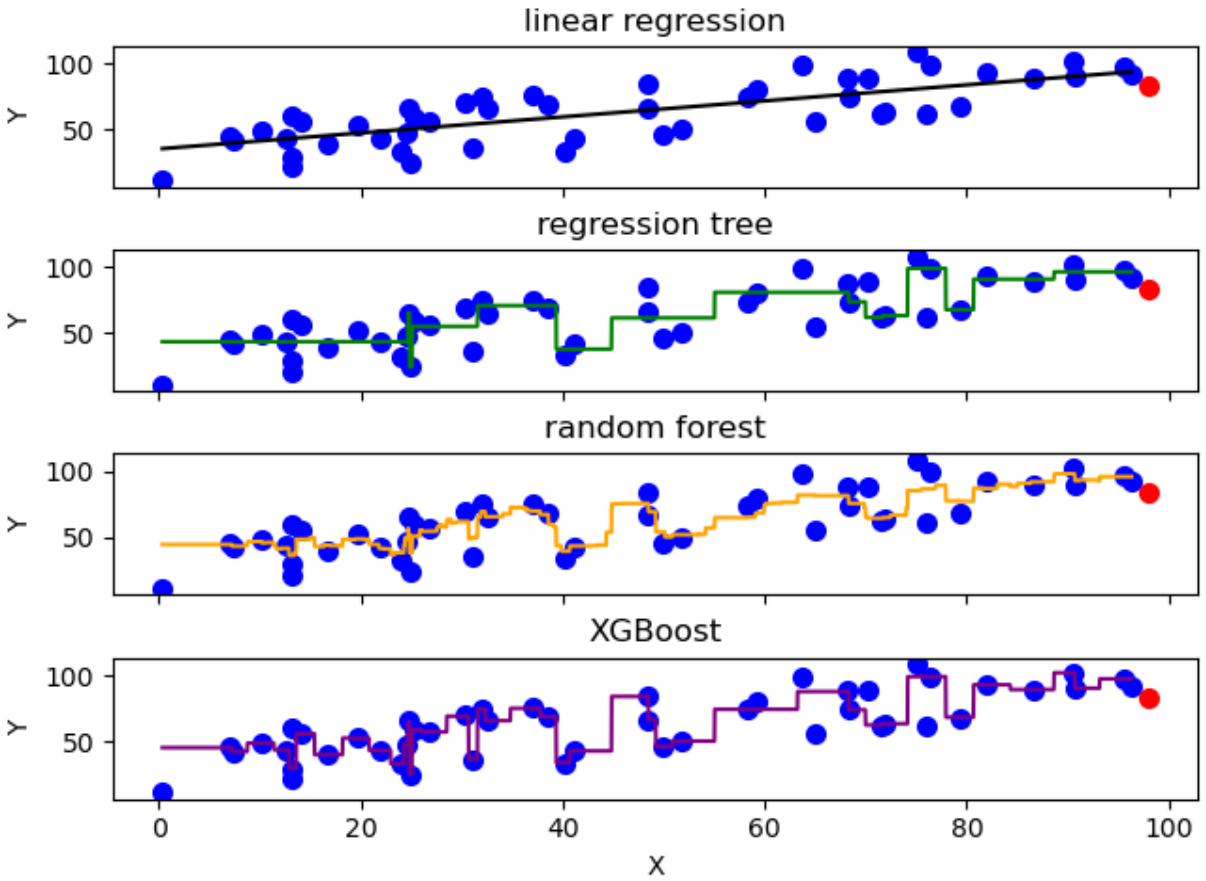
	Train score	Test score
linear regression	0.5685435731131872	0.5718165188063986
regression tree	0.8622874504257273	0.04139730538214548
random forest	0.8957957024803009	0.28603749904340037
XGBoost	0.9998282666474112	0.01117006537470322

	MSE Train	MSE Test
linear regression	177.51678307754625	397.1972636056487
regression tree	56.65992500384094	889.2317987742656
random forest	42.873417860583245	662.2954040788542
XGBoost	0.07065731415489905	917.2715937328169

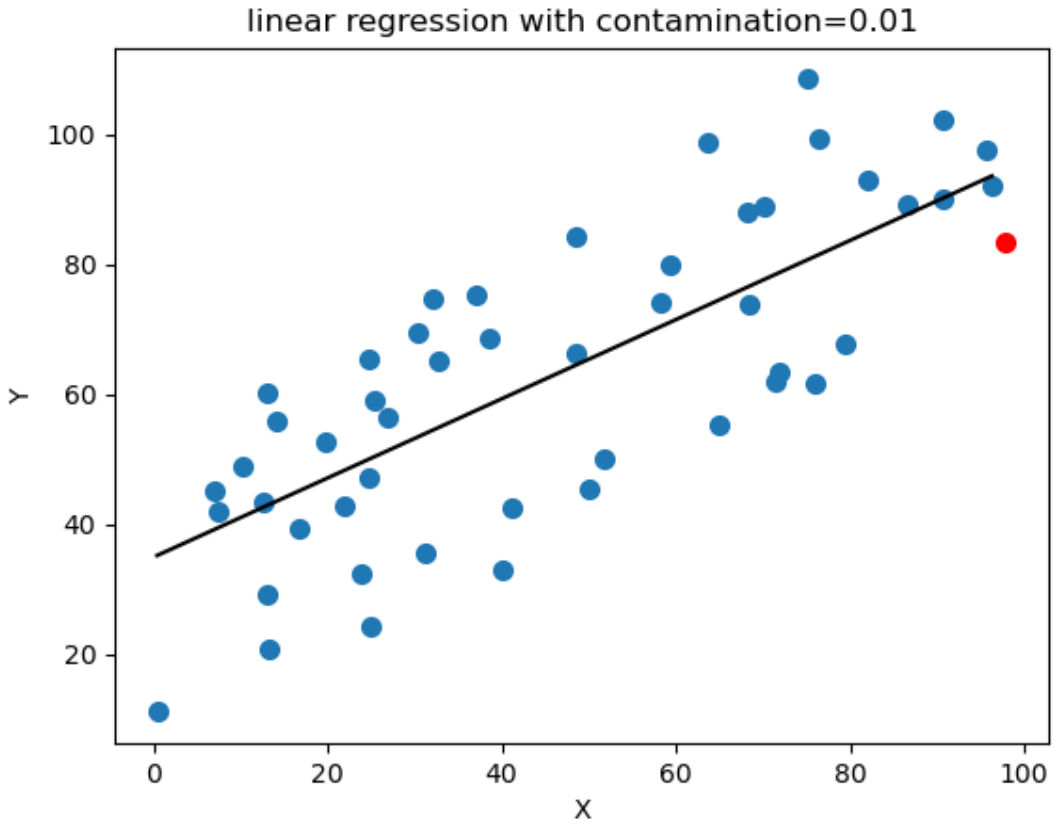
	MAE Train	MAE Test

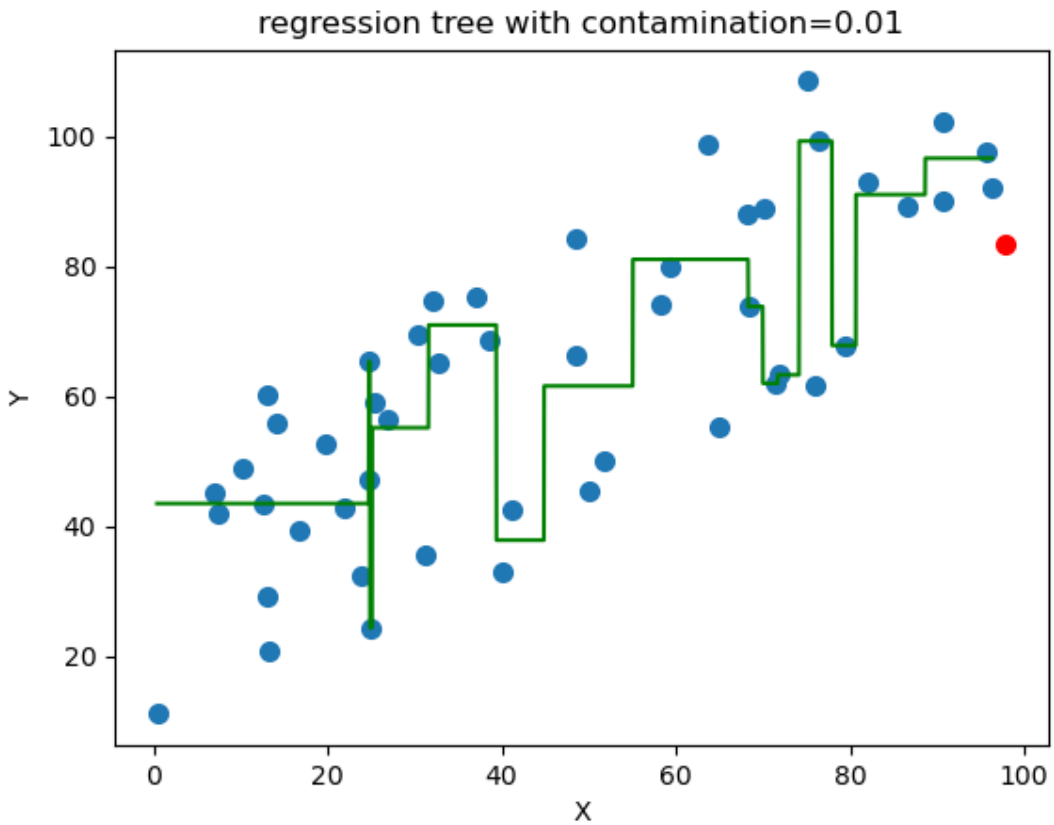
linear regression	11.562717713448263	18.108898475874792
regression tree	5.183106297919835	27.404019058805915
random forest	5.289942920937312	23.703247251593417
XGBoost	0.23001440238946652	26.898697644771993

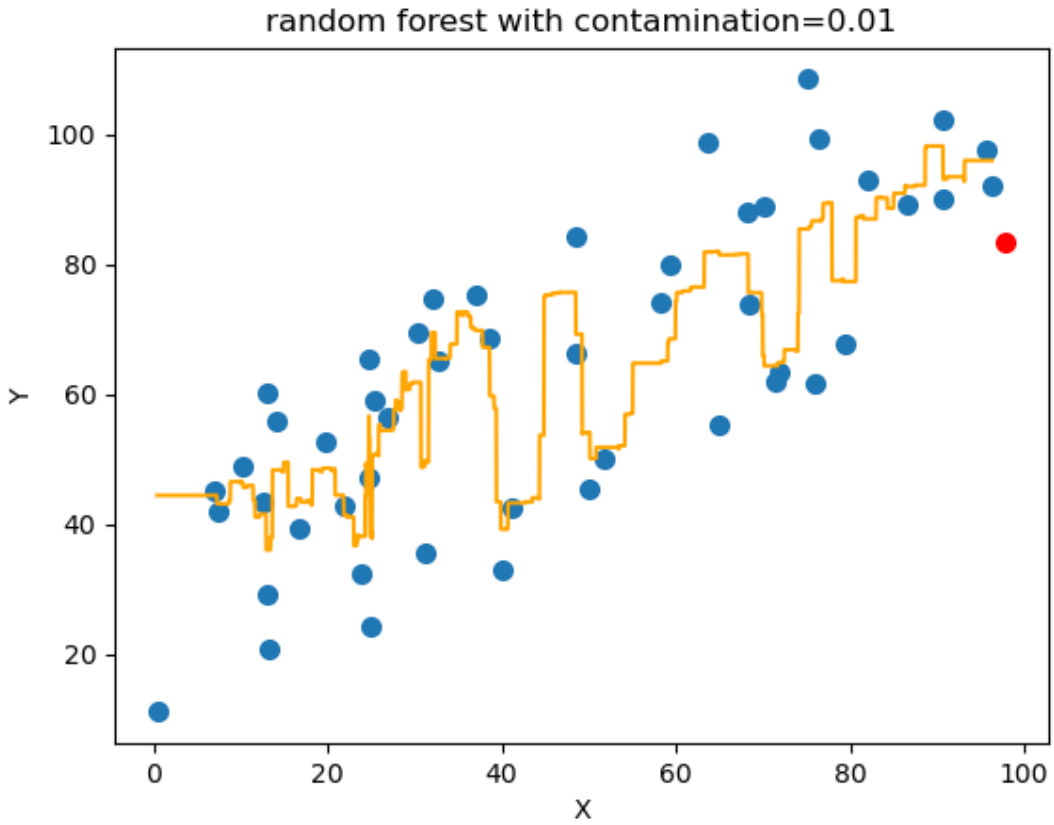
❖ **הורדת הנקודה הרחוקה ביותר :**

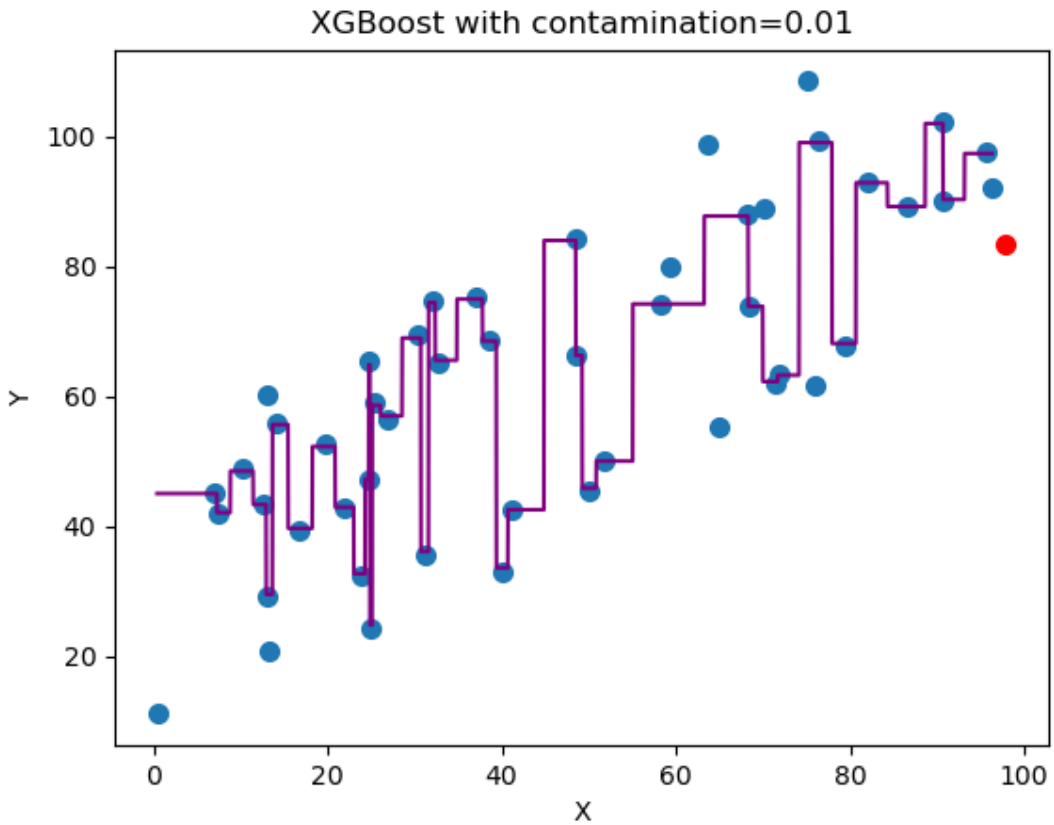


תרשים עם יותר פירוט:


linear regression with contamination=0.01
Y
X
0
20
40
60
80
100
20
40
60
80
100


regression tree with contamination=0.01
Y
X
0
20
40
60
80
100
20
40
60
80
100


random forest with contamination=0.01
100
80
60
40
20
Y
0
20
40
60
80
100
X


XGBoost with contamination=0.01
100
80
60
40
20
Y
0
20
40
60
80
100
X

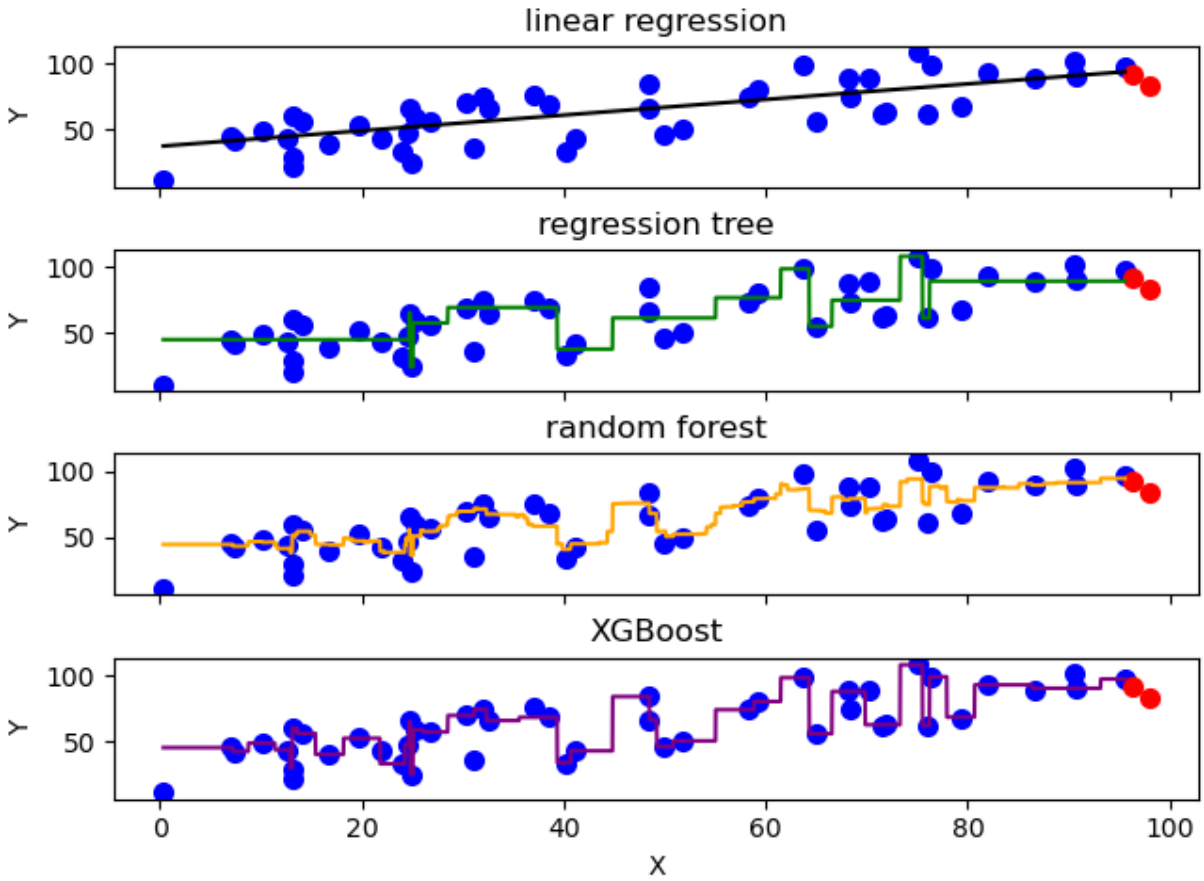
תוצאות הקוד:

	Train score	Test score
linear regression	0.5927677114816616	0.6049987169433757
regression tree	0.8540912168303701	0.463871556554065
random forest	0.9081571019891105	0.4916011084334442
XGBoost	0.9997939424263754	0.40766458095368807

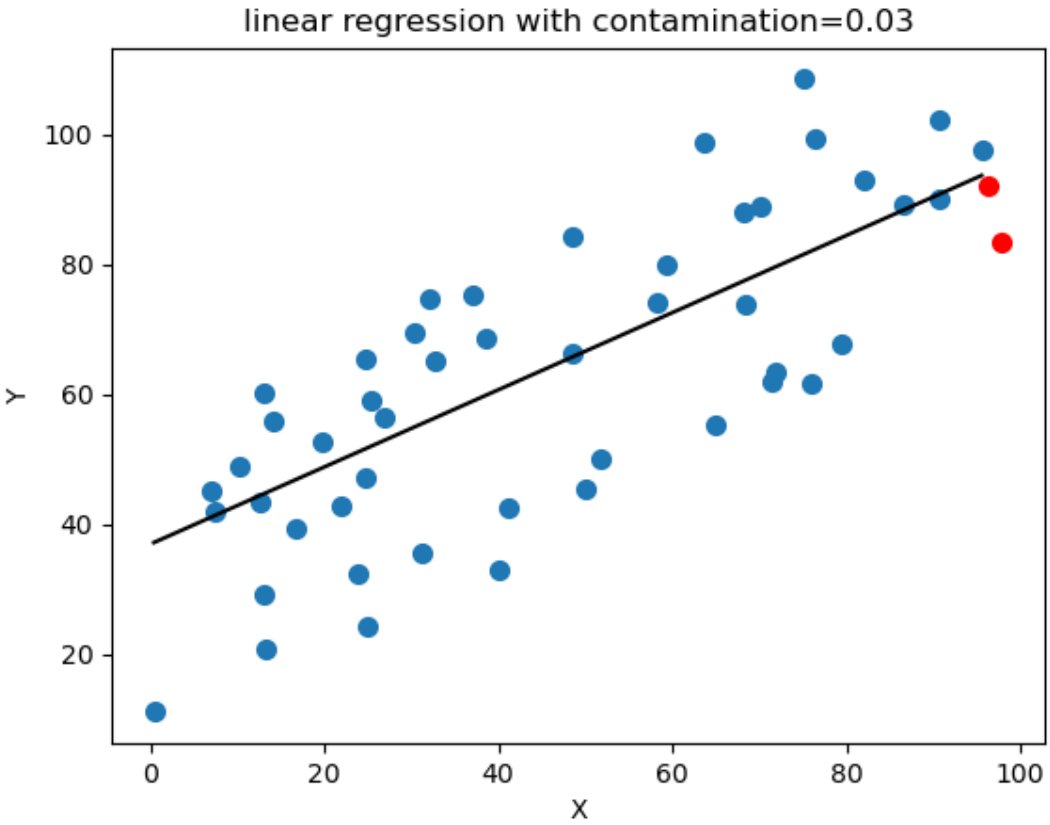
	MSE Train	MSE Test
linear regression	176.12107713321973	373.8027988186832
regression tree	63.1030809185834	507.35610561970236
random forest	39.72050002117133	481.11471211766167
XGBoost	0.08911641547446492	560.5466285212253

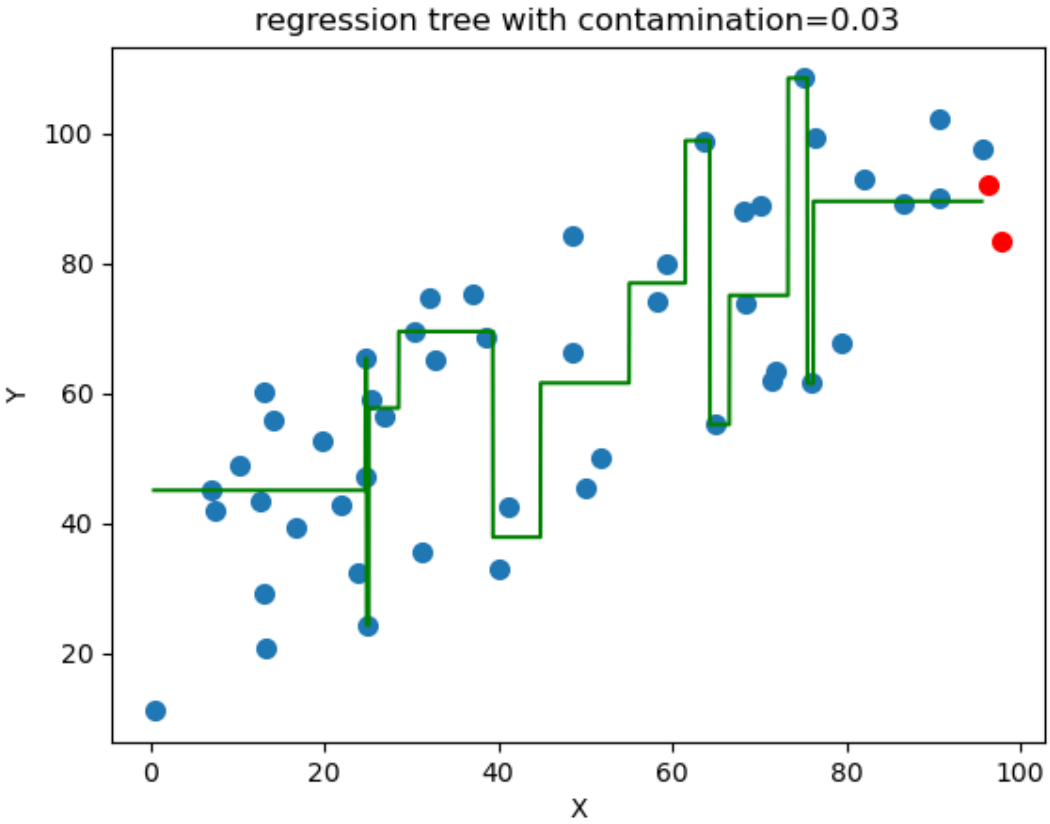
	MAE Train	MAE Test
linear regression	11.222676625477039	17.690964174790402
regression tree	5.512675436969965	19.482530582182704
random forest	5.053561967514597	20.388899182123755
XGBoost	0.2516553690781427	20.136640634056267

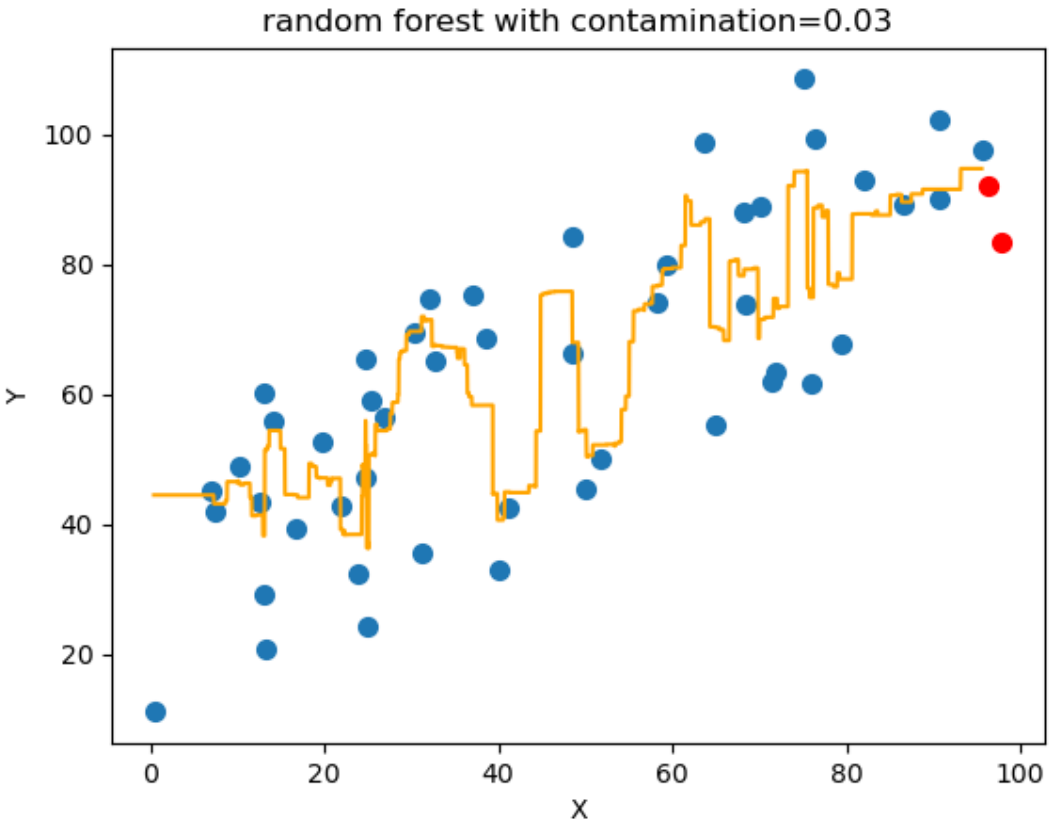
❖ הורדת שתי הנקודה הרחוקה ביותר :

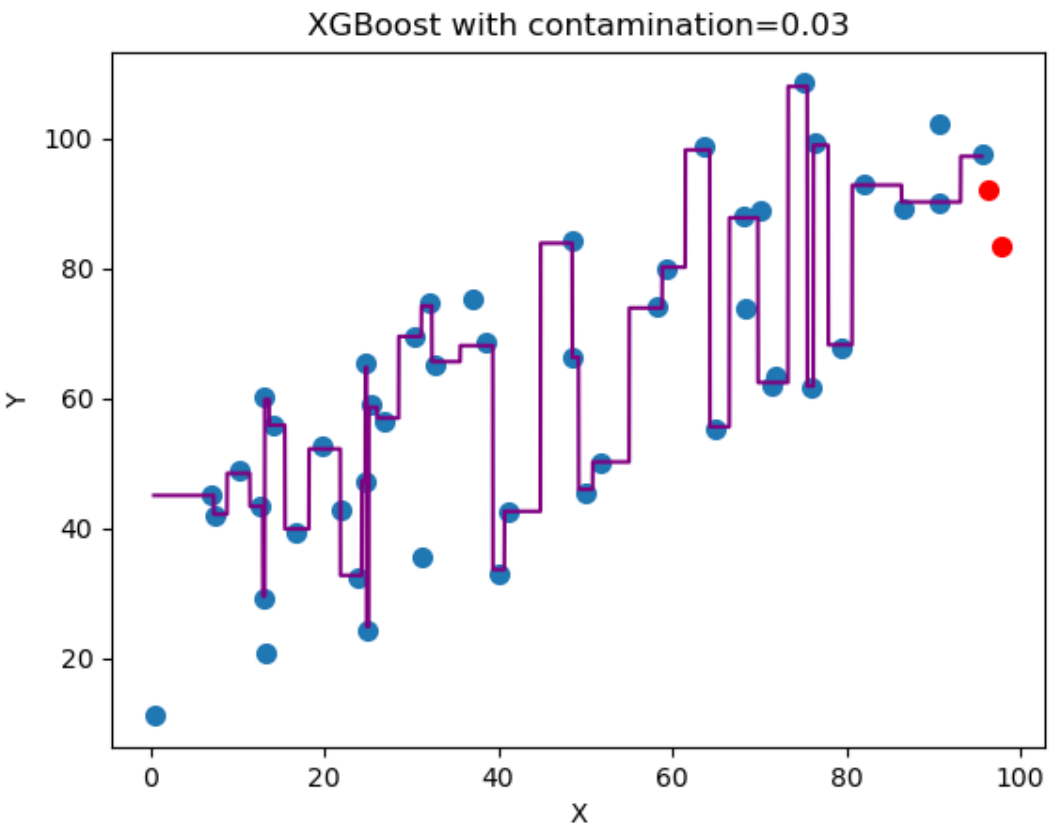


תרשים עם יותר פירוט:


linear regression with contamination=0.03
Y
X
0
20
40
60
80
100
20
40
60
80
100


regression tree with contamination=0.03
Y
X
0
20
40
60
80
100
20
40
60
80
100


random forest with contamination=0.03
100
80
60
40
20
Y
0
20
40
60
80
100
X


XGBoost with contamination=0.03
100
80
60
40
20
Y
0
20
40
60
80
100
X

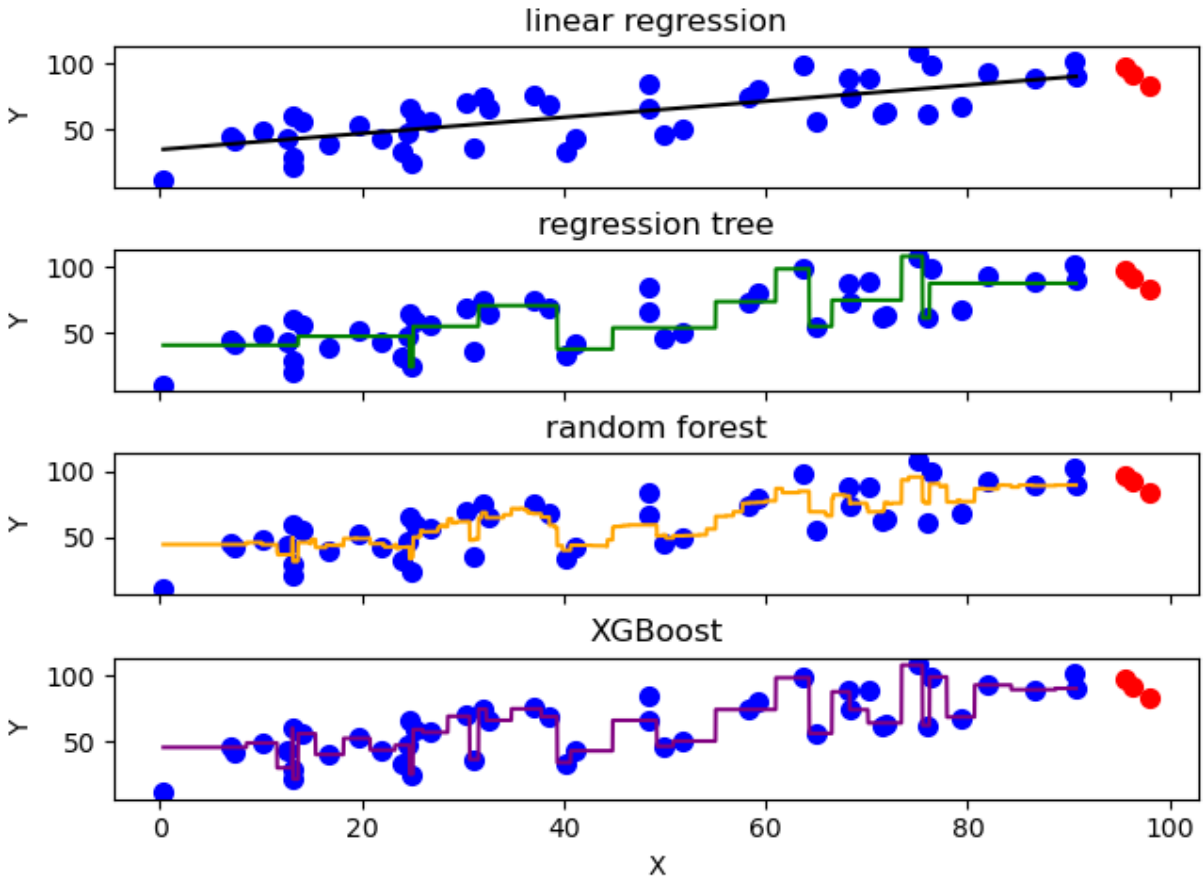
תוצאות הקוד:

	Train score	Test score
linear regression	0.5148103308832741	0.7147239771012123
regression tree	0.8293731273145212	0.6074803216826838
random forest	0.8702022341295622	0.5306097489022341
XGBoost	0.9996873348859151	0.42444292385389737

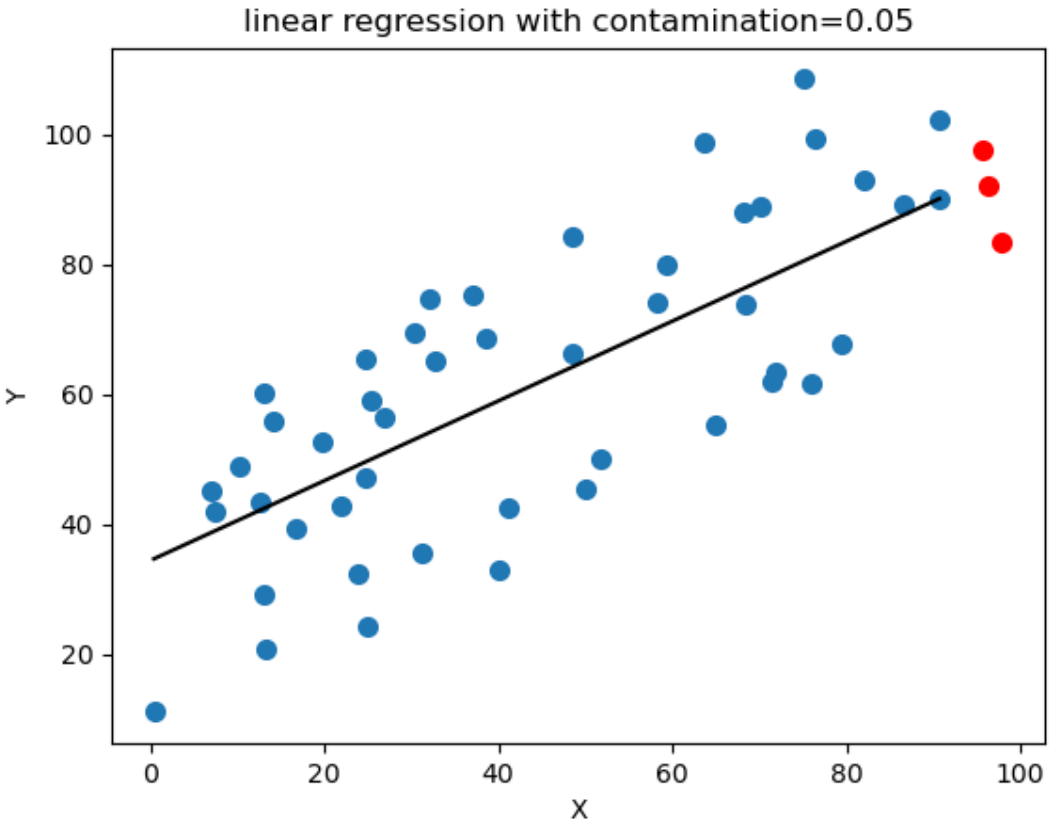
	MSE Train	MSE Test
linear regression	217.40573454048817	248.13634973648837
regression tree	76.45517402722889	341.41810863633356
random forest	58.16030395319077	408.28101263391625
XGBoost	0.14010023939003194	500.6261321533128

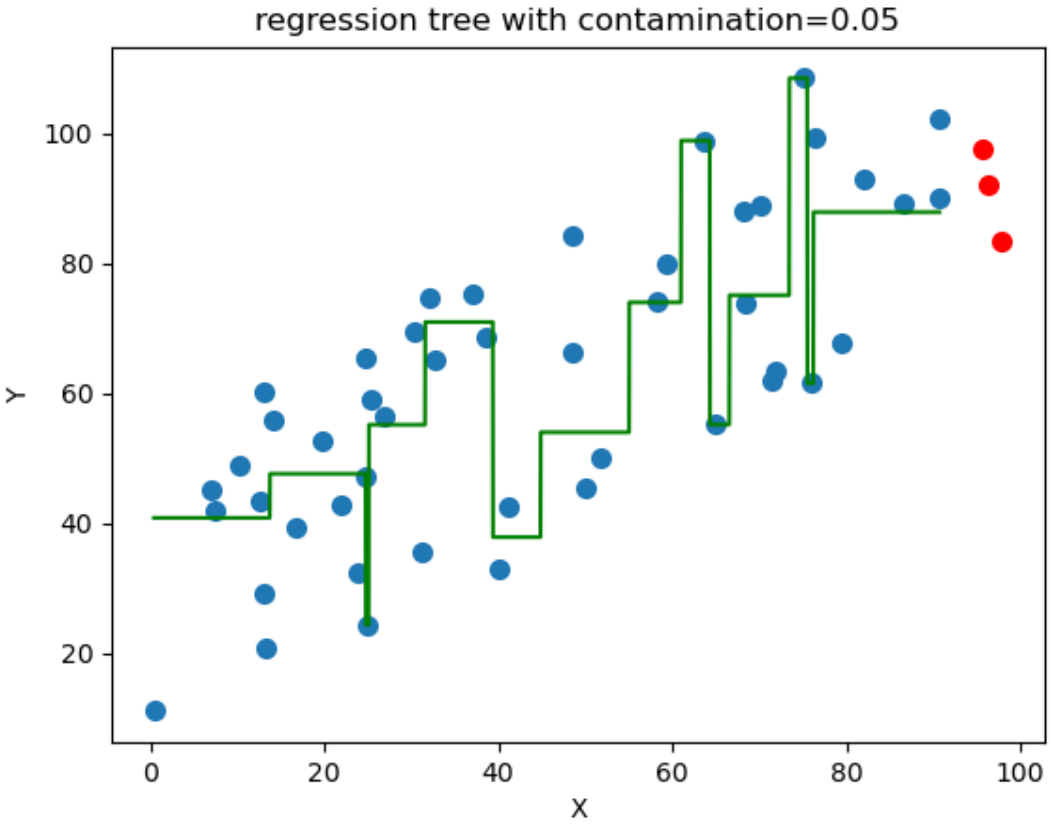
	MAE Train	MAE Test
linear regression	12.354651589813294	13.586911308980778
regression tree	6.099361768444298	13.990750369755759
random forest	6.160195523524941	16.394919168562296
XGBoost	0.322531568798229	17.8353117401157

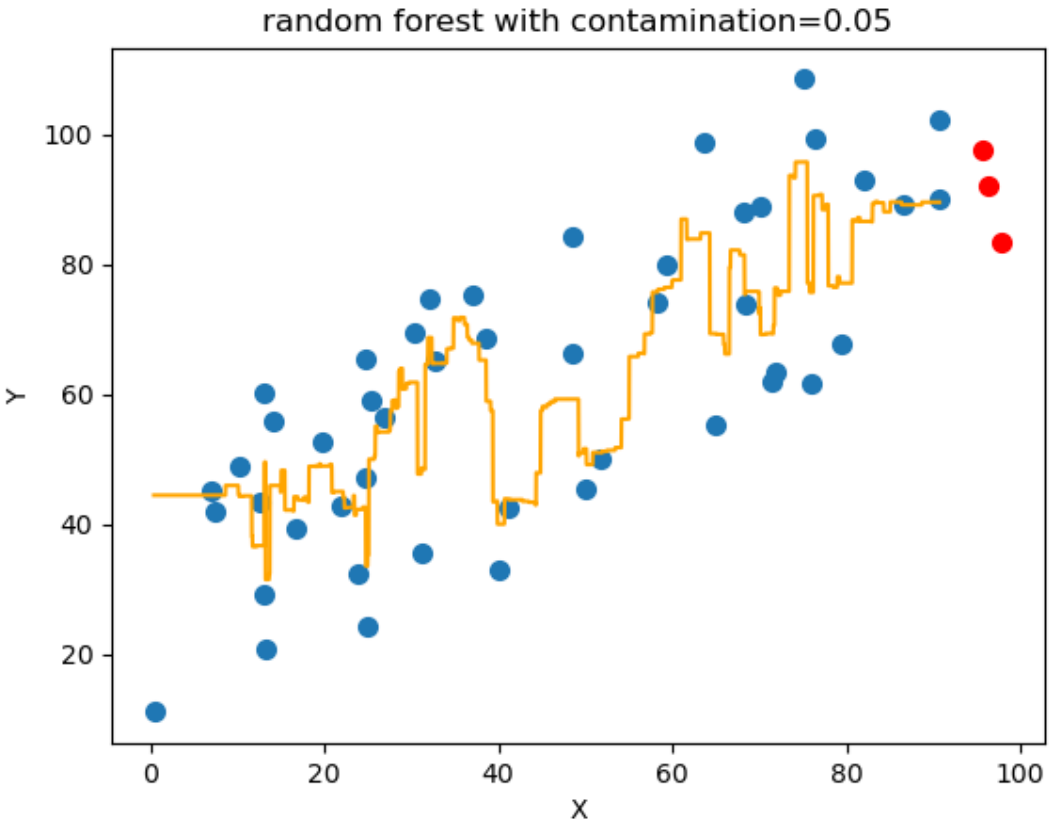
❖ הורדת שלושת הנקודות הרחוקה ביותר :

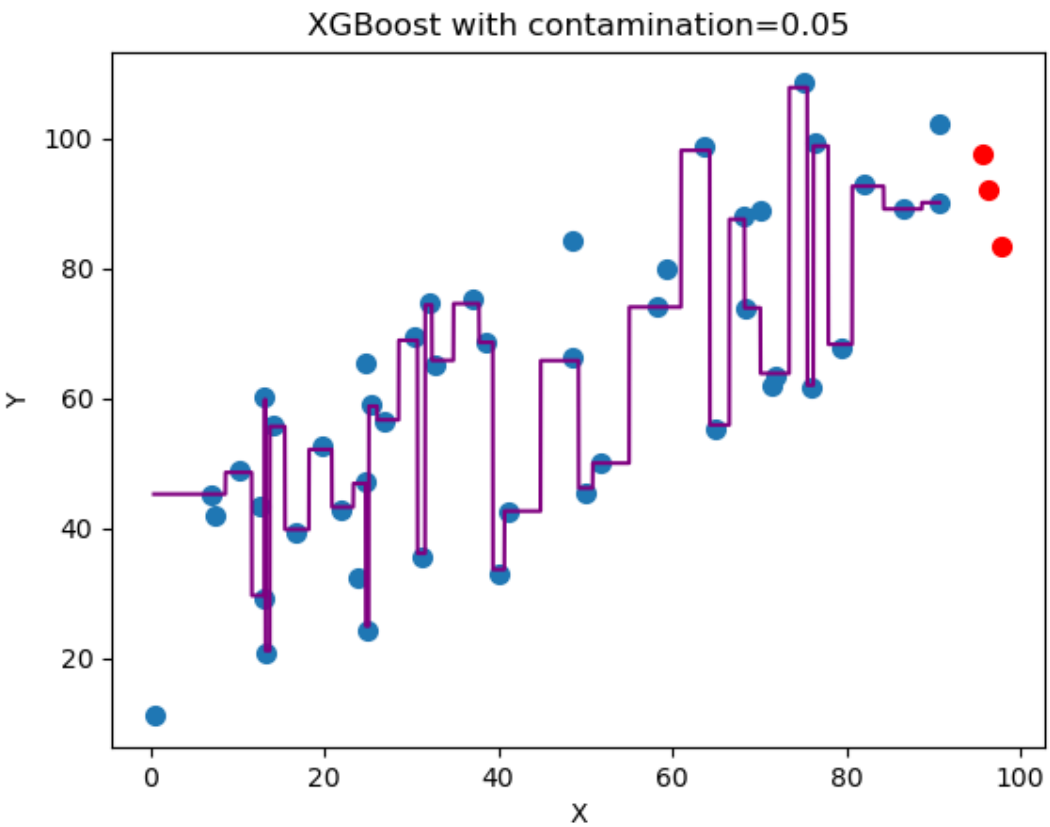


תרשים עם יותר פירוט:


linear regression with contamination=0.05
Y
X
0
20
40
60
80
100
20
40
60
80
100


regression tree with contamination=0.05
Y
X
0
20
40
60
80
100
20
40
60
80
100


random forest with contamination=0.05
Y
X
0
20
40
60
80
100
20
40
60
80
100


XGBoost with contamination=0.05
Y
X
0
20
40
60
80
100
20
40
60
80
100

תוצאות הקוד:

	Train score	Test score
linear regression	0.5059018433114806	0.7122999328980189
regression tree	0.8240602167492828	0.4153935920687314
random forest	0.8677493632300005	0.5247215933645388
XGBoost	0.9995624691093571	0.41016076965935133

	MSE Train	MSE Test
linear regression	228.44653743803838	212.30826203233724
regression tree	81.34584947776474	431.41029368220296
random forest	61.146150081931886	350.7316960704955
XGBoost	0.20229263282310378	435.2718549339963

	MAE Train	MAE Test
linear regression	13.035027154200195	12.895235417581603
regression tree	6.638627043743311	16.67308155620554
random forest	6.54306956321398	15.239600535386094
XGBoost	0.39223755429066026	16.94195279055724

❖ מסקנות :

1) מודל XGBoost מראה שיש לו מצב OverFitting, משאפשר לזהות זאת בקלות על ידי השוואת תוצאות הקוד והסיבה לזאת היא כי אין מספיק נתונים.
2) עבור מודלי random forest ו-regression tree, ניתן לראות שהם הצליחו לקבל ניקוד יותר טוב עבור הנתונים של ה-train מאשר המודל הלינארי של הרגרסיה, וקיבלו ערכי MAE ו-MSE נמוכים יותר מאשר המודל הלינארי של הרגרסיה. עם זאת, עבור נתוני ה-test, הם קיבלו ניקוד נמוך יותר מאשר המודל הלינארי של הרגרסיה וקיבלו ערכי MAE ו-MSE גבוהים יותר מאשר המודל הלינארי של הרגרסיה.

להלן טבלת השוואות בין שני המודלים (regression tree ו- random forest) ביחס למודל linear regression

	Train score	Test score	MSE Train	MSE Test	MAE Train	MAE Test
regression tree	יותר טוב	פחות טוב	יותר טוב	פחות טוב	יותר טוב	פחות טוב
random forest	יותר טוב	פחות טוב	יותר טוב	פחות טוב	יותר טוב	פחות טוב