כריית נתונים

שימוש בתוכנת Orange

לינוי עזר

תיאור קובץ הנתונים

הקובץ מתוך המודל מכיל נתונים הלקוחים מדוחות מאזן ורו"ה של עשרות חברות בארץ.

הנתונים הינם נתונים מחושבים ומציגים יחסים פיננסים (כגון: "יחס שוטף", "יחס מהיר", "רוח גולמי חלקי מכירות").

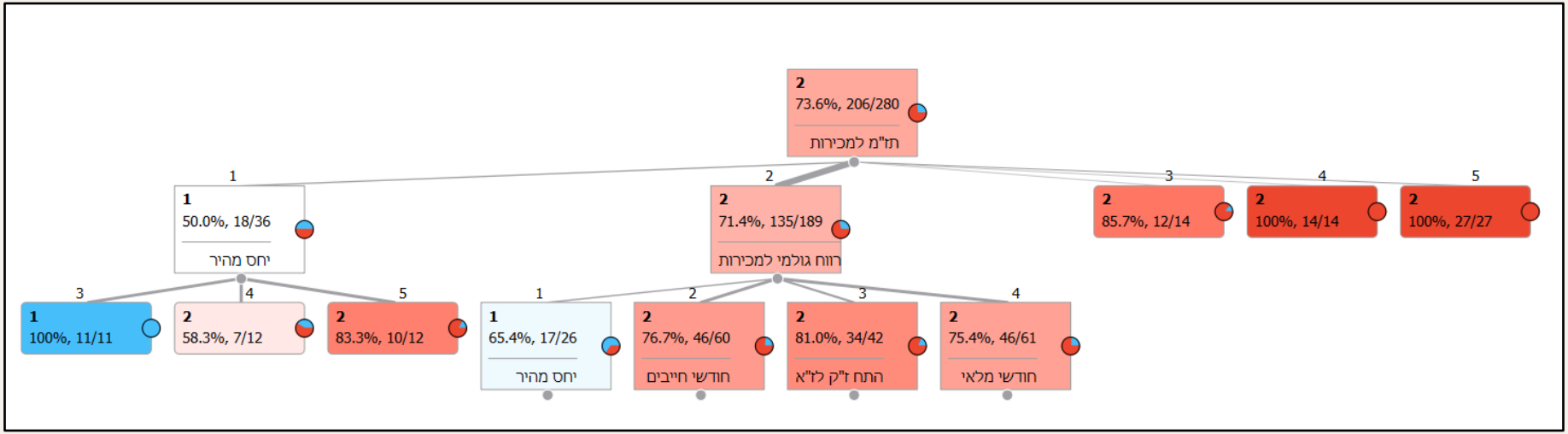
כלל הפרמטרים הינם קטגוריאליים כאשר פרמטר המטרה הינו - "ציון", עם ערכים 1, 2 (1-חוסן פיננסי חלש, 2- חוסן חזק).

	ציון	תז"מ למכירות	הון למאזן	יחס שוטף	יחס מהיר	חודשי מלאי	חודשי חייבים	התח ז"ק לז"א	התפתחות המחזור	רווח גולמי למכירות
1	2	2	2	5	5	3	2	1	4	4
2	1	2	1	3	2	4	1	3	3	1
3	1	2	5	5	3	5	4	1	3	2
4	1	2	2	4	4	1	5	2	3	1
5	2	1	5	4	?	2	4	5	3	2
6	2	2	2	5	?	3	2	5	3	4
7	2	2	2	5	5	1	3	5	3	4
8	2	3	2	5	4	1	2	5	4	4
9	2	2	2	5	5	1	2	5	3	4
10	2	2	5	4	3	3	2	5	4	4
11	2	2	3	3	3	1	3	5	2	4
12	2	2	2	5	4	2	5	5	3	3

חלק ראשון

ברירת מחדל

ציור עץ החלטות

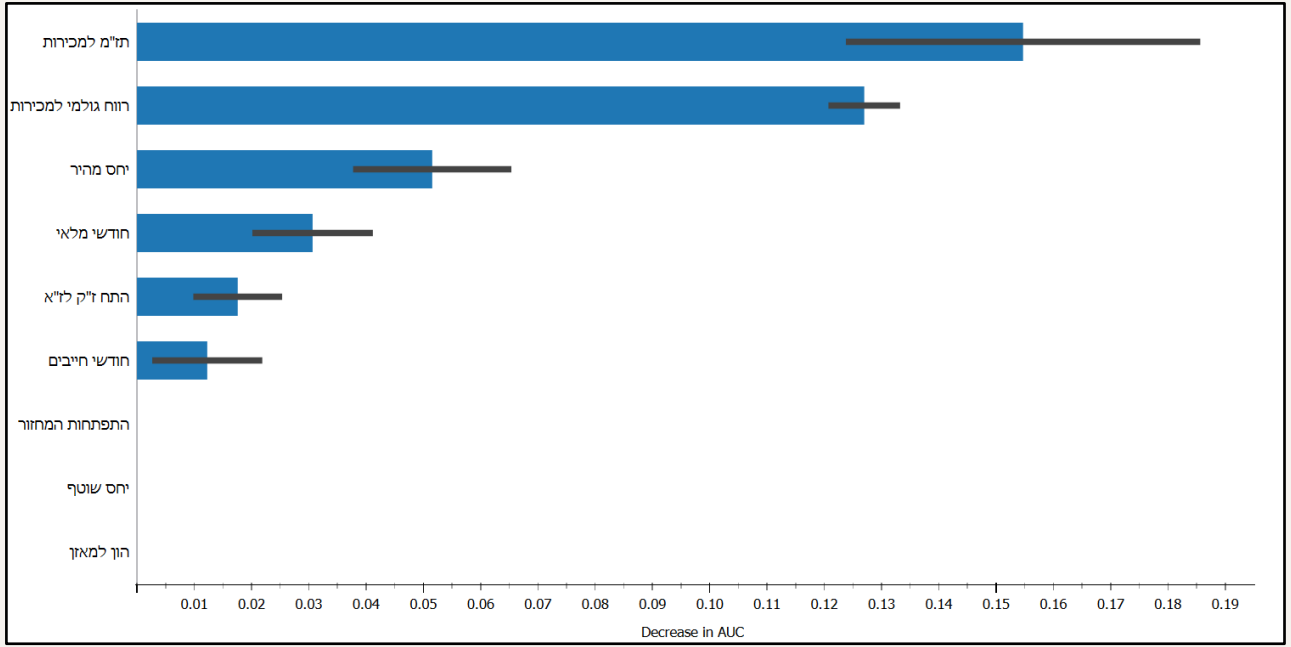


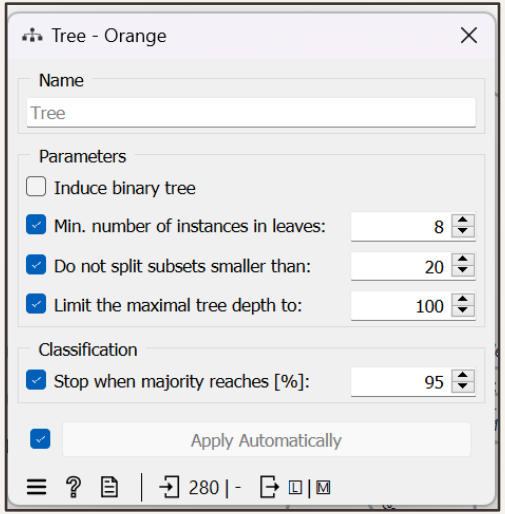
רשימת פרמטרים מסבירים

עץ החלטות

ניתן לראות כי הפרמטר המסביר ביותר הוא 'תז"מ למכירות'.

מיד אחריו נמצא הפרמטר 'רווח גולמי למכירות'.



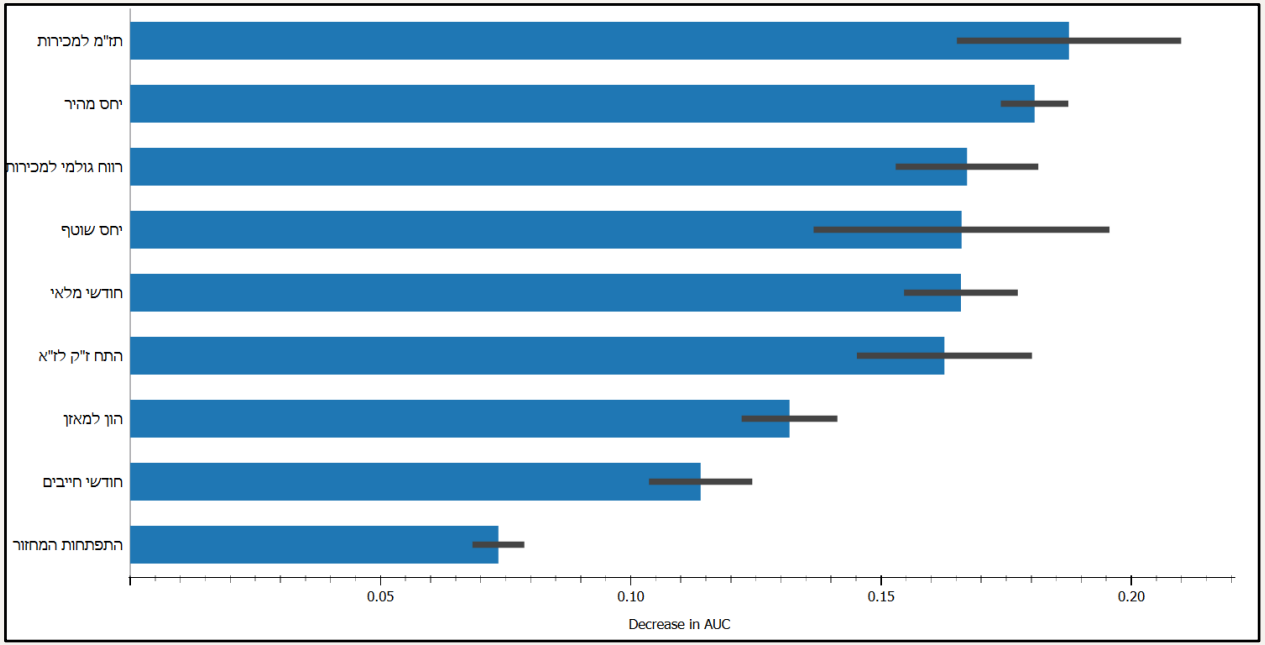


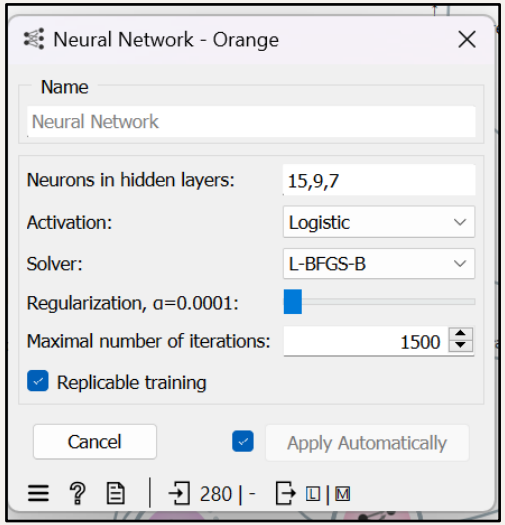
רשימת פרמטרים מסבירים

רשת נוירונים

ניתן לראות כי הפרמטר המסביר ביותר הוא 'תז"מ למכירות'.

מיד אחריו נמצא הפרמטר 'יחס מהיר'.



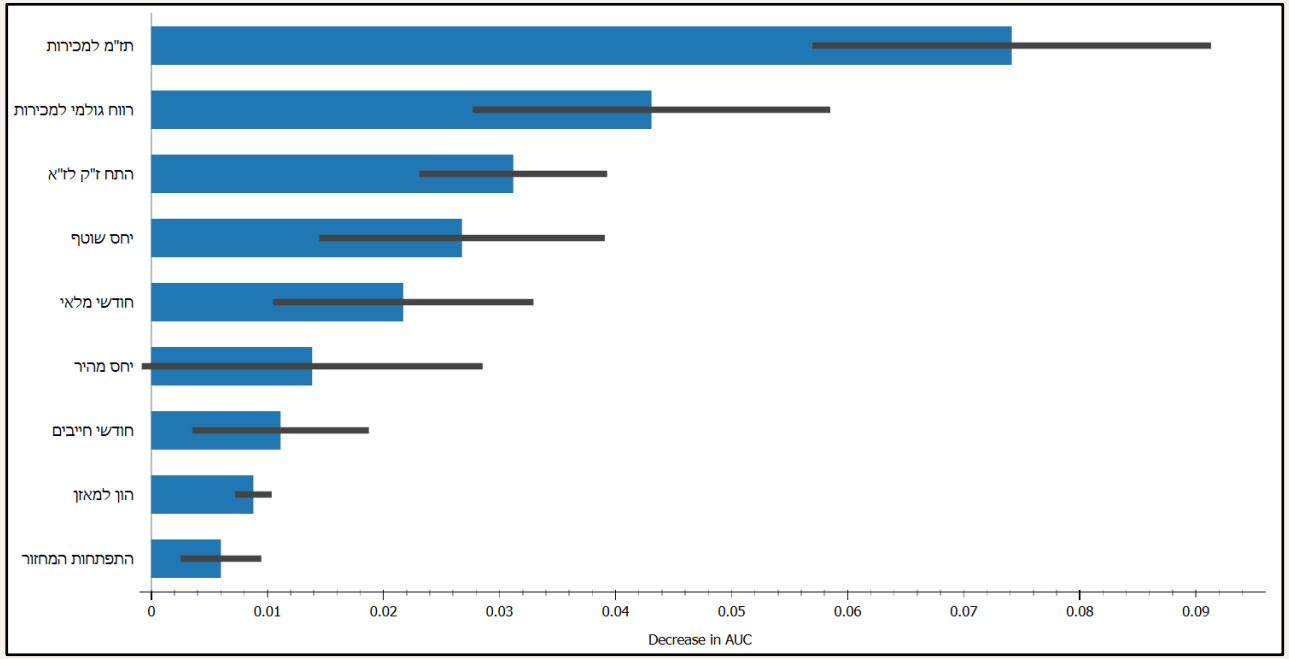


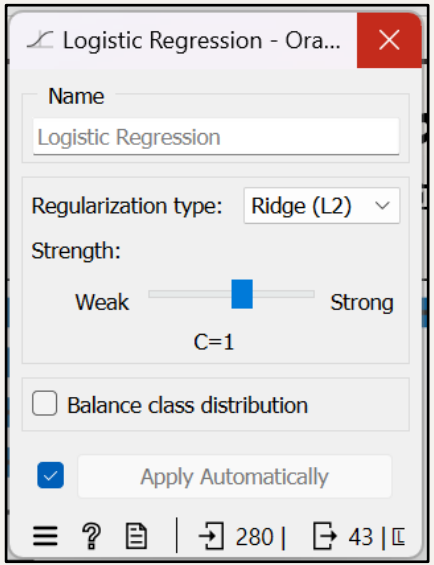
רשימת פרמטרים מסבירים

רגרסיה לוגיסטית

ניתן לראות כי הפרמטר המסביר ביותר הוא 'תז"מ למכירות'.

מיד אחריו נמצא הפרמטר 'רווח גולמי למכירות'.





מדדי דיוק המודל - כללי

Test on train data

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.996	0.968	0.968	0.968	0.968	0.945	0.056
Logistic Regression	0.840	0.846	0.831	0.848	0.846	0.624	0.408
Tree	0.829	0.800	0.776	0.791	0.800	0.530	0.408

רשת נוירונים: בעל רמת הדיוק הגבוהה ביותר (0.968).

מצביע על כך שהאלגוריתם הצליח לחזות נכונה את התוצאה

עבור 96.8% מהדוגמאות באוסף האימון.

רגרסיה לוגיסטית: בעל רמת דיוק בינונית (0.846).

עץ החלטה: בעל רמת הדיוק הנמוכה ביותר (0.8).

Test on test data

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.923	0.914	0.915	0.915	0.914	0.870	1.149
Logistic Regression	0.816	0.829	0.814	0.822	0.829	0.605	0.429
Tree	0.771	0.794	0.770	0.781	0.794	0.514	0.638

רשת נוירונים: רשת הנוירונים שמרה על דיוק גבוה גם

באוסף הנתונים החדש (0.914).

רגרסיה לוגיסטית: בעל רמת דיוק בינונית (0.829).

עץ החלטה: בעל רמת הדיוק הנמוכה ביותר (0.794).

מדדי דיוק המודל - כללי

Test on train data

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.996	0.968	0.968	0.968	0.968	0.945	0.056
Logistic Regression	0.840	0.846	0.831	0.848	0.846	0.624	0.408
Tree	0.829	0.800	0.776	0.791	0.800	0.530	0.408

Test on test data

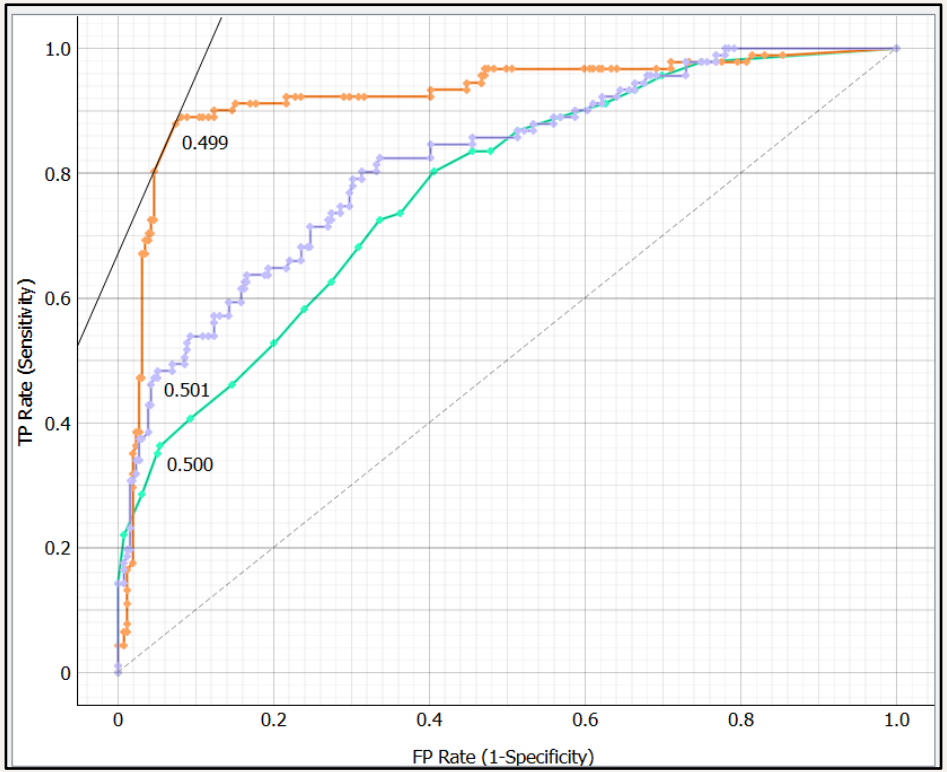
Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.923	0.914	0.915	0.915	0.914	0.870	1.149
Logistic Regression	0.816	0.829	0.814	0.822	0.829	0.605	0.429
Tree	0.771	0.794	0.770	0.781	0.794	0.514	0.638

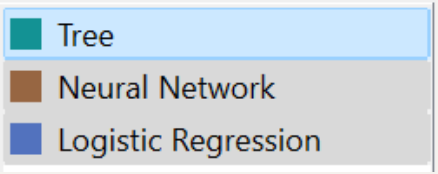
נרצה לראות דמיון ברמות הדיוק בכמה שיותר מדדים בין קבוצת האימון לקבוצת המבחן.

ניתן לראות שארבעת המדדים המרכזיים שנלמדו, AUC,CA,Prec,Spec, אכן יצאו דומים.

יחד עם זאת, ניתן לראות סטיות משמעותיות יותר במדד הדיוק AUC של אלוגריתם עץ ההחלטה ובמדד הדיוק Spec של אלגוריתם רשת נוירונים. עפ"י סטיות אלו, קבוצת המבחן קיבלה אחוזי הצלחה נמוכים יותר.

עקומת ROC





העקומה מתייחסת הן לקבוצת האימון והן לקבוצות המבחן ומציגה מודל חיזוי אופטימלי (גרף שחור).

המודל הקרוב ביותר לגרף זה הוא המודל בעל החיזוי הטוב ביותר – רשת נוירונים.

בנוסף מודל זה רחוק באופן מובהק מהגרפים של שני המודלים האחרים דבר אשר מחזק את המסקנה שהוא בעל החיזוי הטוב ביותר.

מדדי דיוק המודל - מפורט

עץ החלטות

Actual \ Predicted	1	2	Σ
1	33	58	91
2	14	245	259
Σ	47	303	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 79.4% מהתצפיות שחזה המודל היו נכונות.

Precision : 80.6% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 303 תצפיות בעלות ציון "2", רק 80.6% מתוכן היו בעלות ציון זה באמת (245).)

Sensitivity : המודל חזה נכונה 94.6% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 94.6% מתוכן (245).)

Specifity : המודל חזה נכונה 36.3% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 36.3% מתוכן (33).)

מדדי דיוק המודל - מפורט

רשת נוירונים

Actual \ Predicted	1	2	∑
1	77	14	91
2	16	243	259
∑	93	257	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 91.4% מהתצפיות שחזה המודל היו נכונות.

Precision : 94.5% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 257 תצפיות בעלות ציון "2", 94.5% מתוכן היו בעלות ציון זה באמת (243).)

Sensitivity : המודל חזה נכונה 98.8% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 93.8% מתוכן (243).)

Specifity : המודל חזה נכונה 84.6% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 84.6% מתוכן (77).)

מדדי דיוק המודל - מפורט

רגרסיה לוגיסטית

Actual \ Predicted	1	2	∑
1	44	47	91
2	13	246	259
∑	57	293	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 82.8% מהתצפיות שחזה המודל היו נכונות.

Precision : 83.9% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 293 תצפיות בעלות ציון "2", 83.9% מתוכן היו בעלות ציון זה באמת (246).)

Sensitivity : המודל חזה נכונה 95% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 95% מתוכן (246).)

Specifity : המודל חזה נכונה 48.3% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 48.3% מתוכן (44).)

חיזוי

בטבלה הבאה נוכל לראות השוואה בין ערך האמת לבין הערכים החזויים שהתקבלו ע״י האלגוריתמים.

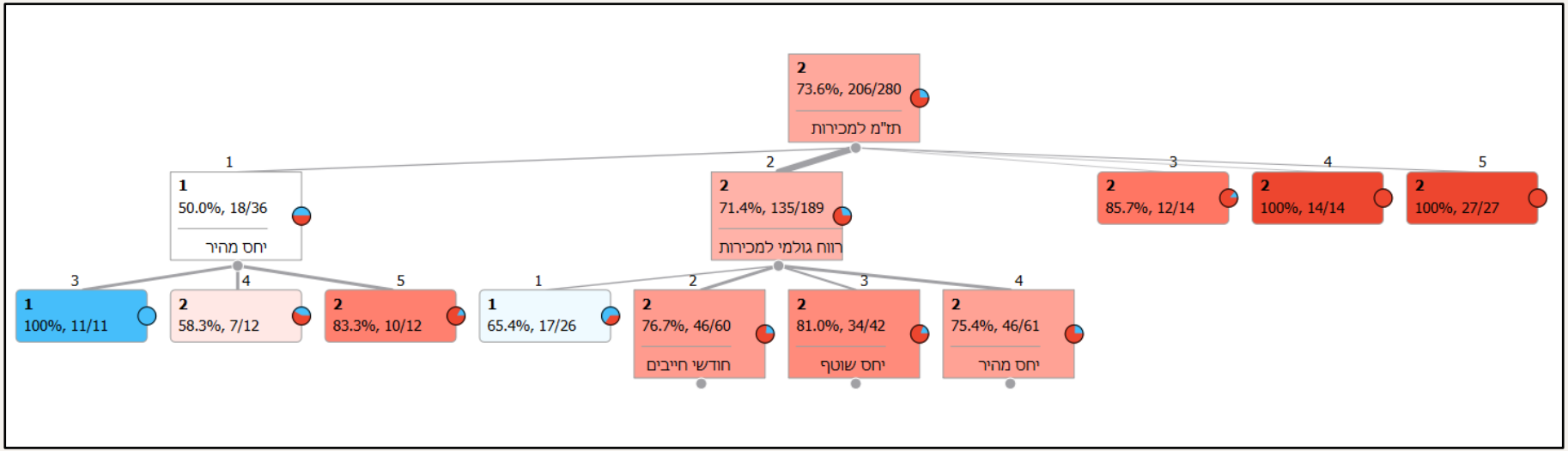
ניתן לראות שעבור התצפית המודגשת, ערך האמת ״1״ התקבל ע״י האלגוריתם רשת נוירונים בלבד. המודלים האחרים, עץ החלטה ורגרסיה לוגיסטית, חזו את לעומת זאת את הערך ״2״ עבור התצפית.

ערך אמת	ערכים חזויים ע״י האלגוריתמים		
ציון	Tree	Neural Network	ogistic Regressio
2	2	2	2
1	1	1	1
1	2	1	2
1	1	1	1
2	1	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
1	2	2	1
2	2	2	2
2	2	2	2
1	2	2	2
2	2	2	2
1	2	1	1
1	2	1	1

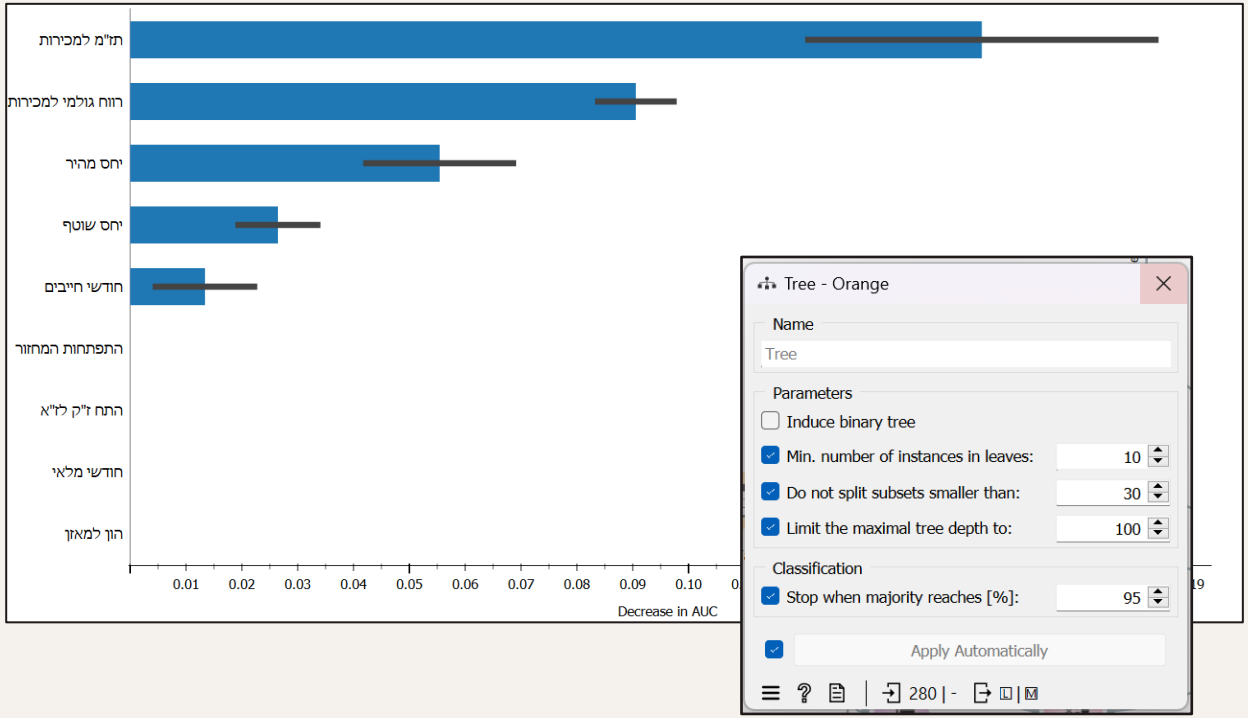
חלק שני

שינוי המודלים

ציור עץ החלטות



רשימת פרמטרים מסבירים



עץ החלטות

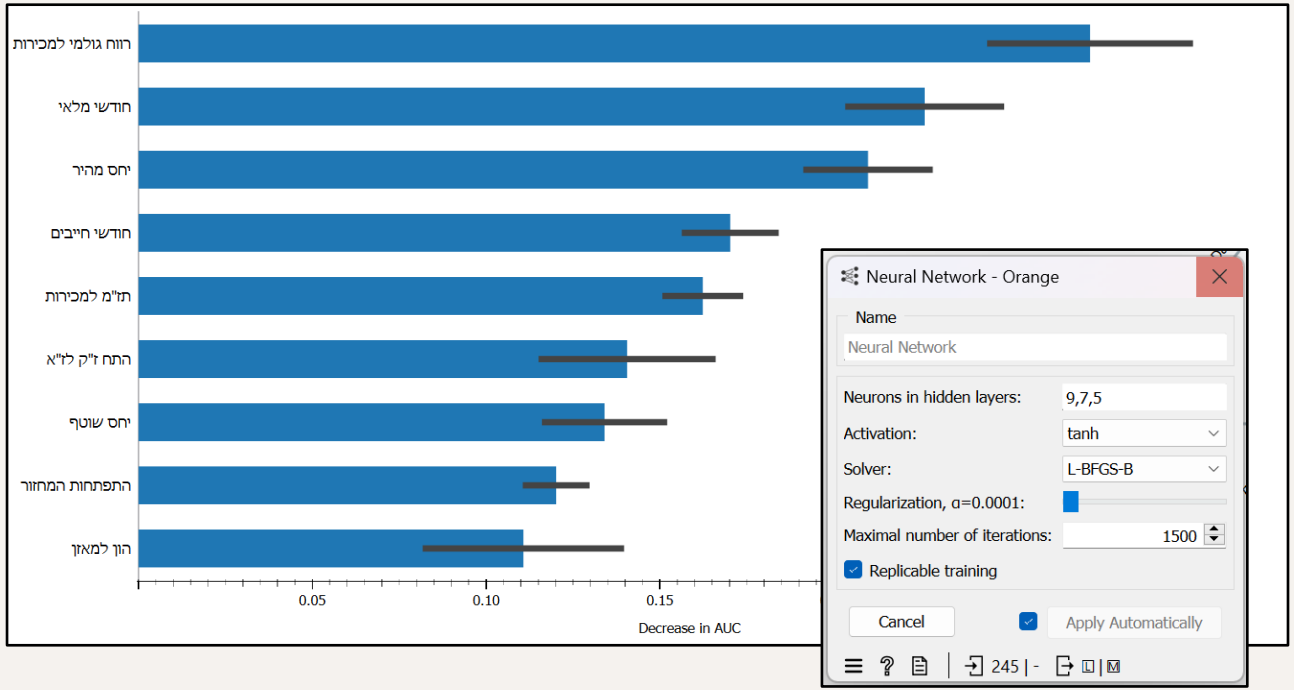
ניתן לראות כי הפרמטר המסביר ביותר הוא 'תז"מ למכירות'.

מיד אחריו נמצא הפרמטר 'רווח גולמי למכירות'.

בניגוד לחלק הראשון בתרגיל בו התקבלו 6 פרמטרים מסבירים, בניתוח זה התקבלו 5 פרמטרים בלבד.

בדומה לחלק הראשון, 3 הפרמטרים המסבירים ביותר נותרו זהים גם בסדר חשיבותם.

רשימת פרמטרים מסבירים



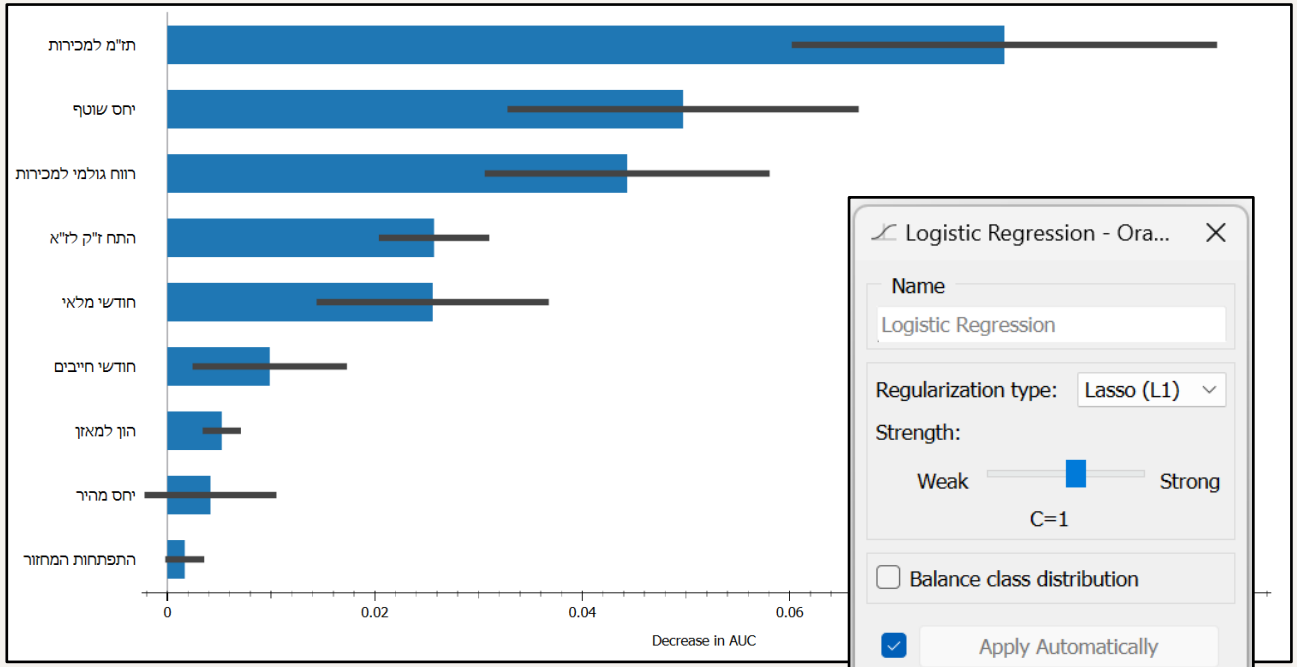
רשת נוירונים

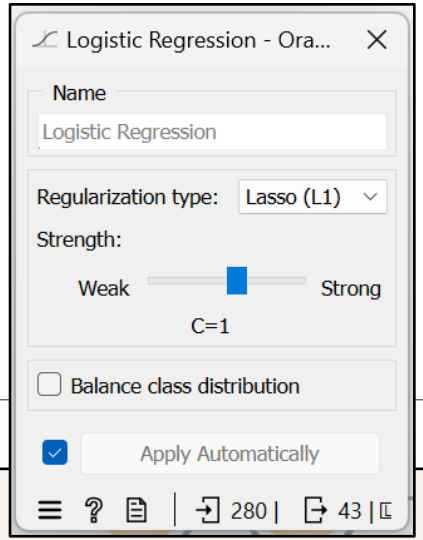
ניתן לראות כי הפרמטר המסביר ביותר הוא 'רווח גולמי למכירות'.

מיד אחריו נמצא הפרמטר 'חודשי מלאי'.

ביחס לחלק הראשון, כל הפרמטרים השתנו ברמת חשיבותם בניתוח זה (מלבד הפרמטר 'התח ז"ק לז"א' ששמר על מיקומו).

רשימת פרמטרים מסבירים





רגרסיה לוגיסטית

ניתן לראות כי הפרמטר המסביר ביותר הוא 'תז"מ למכירות'.

מיד אחריו נמצא הפרמטר 'יחס שוטף'.

בדומה לחלק הראשון, הפרמטר המסביר ביותר נשאר זהה ברמת חשיבותו.

הפרמטרים חודשי מלאי והתפתחות המחזור שמרו על מיקומם גם כן, שאר הפרמטרים השתנו מבחינת רמת חשיבות.

מדדי דיוק המודל - כללי

Test on train data

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.997	0.968	0.967	0.968	0.968	0.928	0.046
Logistic Regression	0.832	0.832	0.819	0.827	0.832	0.619	0.418
Tree	0.815	0.800	0.776	0.791	0.800	0.530	0.422

Test on test data

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.886	0.894	0.894	0.895	0.894	0.835	0.996
Logistic Regression	0.810	0.814	0.802	0.804	0.814	0.600	0.436
Tree	0.782	0.794	0.770	0.781	0.794	0.514	0.541

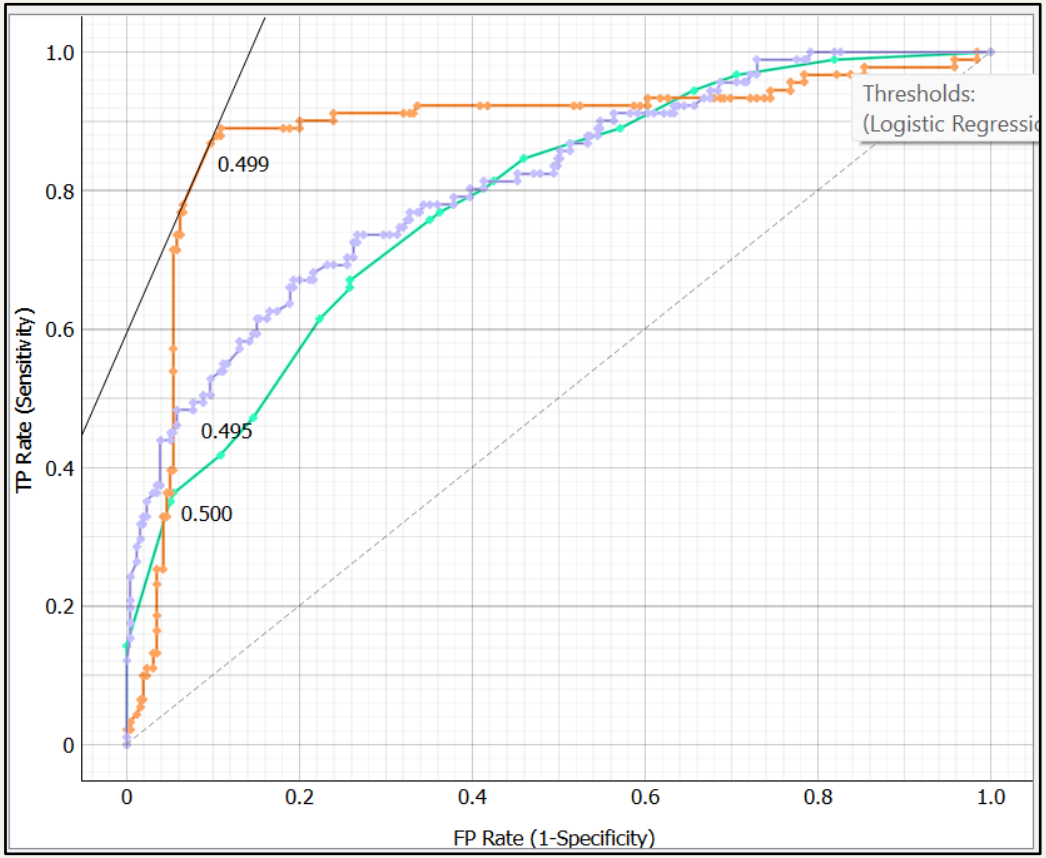
ניתוח זה בולט בהבדלים בין תוצאות כלל המדדים של קבוצות האלגוריתם רשת נוירונים. לפיהם, קבוצת המבחן קיבלה אחוזי הצלחה נמוכים יותר.

יחד עם זאת, בניתוח הזה התקבלו אותן המסקנות:

רשת נוירונים: בעל רמת הדיוק הגבוהה ביותר.

רגרסיה לוגיסטית: בעל רמת דיוק בינונית.

עץ החלטה: בעל רמת הדיוק הנמוכה ביותר.



עקומת ROC

גם בניתוח זה התקבלו אותן המסקנות –

המודל הקרוב ביותר לגרף האופטימלי הוא המודל בעל החיזוי הטוב ביותר – רשת נוירונים.

בנוסף מודל זה רחוק באופן מובהק מהגרפים של שני המודלים האחרים דבר אשר מחזק את המסקנה שהוא בעל החיזוי הטוב ביותר.

Classifiers

- Tree
- Neural Network
- Logistic Regression

מדדי דיוק המודל - מפורט

עץ החלטות

Actual \ Predicted	1	2	Σ
1	33	58	91
2	14	245	259
Σ	47	303	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 79.4% מהתצפיות שחזה המודל היו נכונות.

Precision : 80.8% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 303 תצפיות בעלות ציון "2", רק 80.8% מתוכן היו בעלות ציון זה באמת (245).)

Sensitivity : המודל חזה נכונה 94.6% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 94.6% מתוכן (245).)

Specifity : המודל חזה נכונה 36.3% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 36.3% מתוכן (33).)

מדדי דיוק המודל - מפורט

רשת נוירונים

Actual \ Predicted	1	2	Σ
1	73	18	91
2	19	240	259
Σ	92	258	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 89.4% מהתצפיות שחזה המודל היו נכונות.

Precision : 93% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 258 תצפיות בעלות ציון "2", רק 93% מתוכן היו בעלות ציון זה באמת (240).)

Sensitivity : המודל חזה נכונה 92.7% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 92.7% מתוכן (240).)

Specifity : המודל חזה נכונה 80.2% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 36.3% מתוכן (73).)

מדדי דיוק המודל - מפורט

רגרסיה לוגיסטית

		Predicted		
		1	2	∑
Actual	1	44	47	91
	2	18	241	259
	∑	62	288	350

Accuracy : מדד לרמת הדיוק הכוללנית – לפיו 81.4% מהתצפיות שחזה המודל היו נכונות.

Precision : 83.7% מהתצפיות שחזה המודל כבעלות ציון"2" אכן היו כאלה. (המודל חזה שיש 288 תצפיות בעלות ציון "2", רק 83.7% מתוכן היו בעלות ציון זה באמת (241).)

Sensitivity : המודל חזה נכונה 93% מתוך כלל התצפיות בעלות ציון "2". (במודל הנוכחי 259 תצפיות בעלות ציון "2". המודל גילה 93% מתוכן (241).)

Specifity : המודל חזה נכונה 48.3% מתוך כלל התצפיות בעלות ציון "1". (במודל הנוכחי 91 תצפיות בעלות ציון "1". המודל חזה 36.3% מתוכן (44).)

חיזוי

בטבלה הבאה נוכל לראות השוואה בין ערך האמת לבין הערכים החזויים שהתקבלו ע״י האלגוריתמים.

ניתן לראות שעבור התצפית המודגשת, ערך האמת ״1״ התקבל ע״י רשת נוירונים ורגרסיה לוגיסטית. המודלים האחרון, עץ החלטה, חזה לעומתם את הערך ״1״ עבור התצפית.

ערכים חזויים ע״י האלגוריתמים			ערך אמת
.ogistic Regressio	Neural Network	Tree	ציון
2	2	2	2
1	1	1	1
2	1	2	1
1	1	1	1
2	2	1	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2
1	1	2	1
2	2	2	2
2	2	2	2
1	2	2	1
2	2	2	2
1	1	2	1
1	1	2	1

סיכום מדדי דיוק כלליים

במבחן Test on train data התקבלו מידות דיוק דומות בין ניתוח א' ל-ב'. (בין החלק הראשון של התרגיל לבין החלק השני)

A

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.996	0.968	0.968	0.968	0.968	0.945	0.056
Logistic Regression	0.840	0.846	0.831	0.848	0.846	0.624	0.408
Tree	0.829	0.800	0.776	0.791	0.800	0.530	0.408

B

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.997	0.968	0.967	0.968	0.968	0.928	0.046
Logistic Regression	0.832	0.832	0.819	0.827	0.832	0.619	0.418
Tree	0.815	0.800	0.776	0.791	0.800	0.530	0.422

לעומת זאת, במבחן Test on test data הוצגו פערים בין מידות הדיוק של הניתוחים השונים ושל מודל רשת נוירונים בפרט. ניתוח א' שמר על מידת הדיוק באופן יחסי בעוד שניתוח ב' הציג סטיות גדולות יותר.

A

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.923	0.914	0.915	0.915	0.914	0.870	1.149
Logistic Regression	0.816	0.829	0.814	0.822	0.829	0.605	0.429
Tree	0.771	0.794	0.770	0.781	0.794	0.514	0.638

B

Model	AUC	CA	F1	Prec	Recall	Spec	LogLoss
Neural Network	0.886	0.894	0.894	0.895	0.894	0.835	0.996
Logistic Regression	0.810	0.814	0.802	0.804	0.814	0.600	0.436
Tree	0.782	0.794	0.770	0.781	0.794	0.514	0.541

משני הניתוחים עלו אותן המסקנות: מודל רשת נוירונים בעל מידת הדיוק הגבוה ביותר, מודל עץ החלטה בעל מידת הדיוק הנמוכה ביותר.

סיכום מדדי דיוק מפורטים

	רשת נוירונים		רגרסיה לוגיסטית		עץ החלטה	
שם המדד	ניתוח א	ניתוח ב	ניתוח א	ניתוח ב	ניתוח א	ניתוח ב
Accuracy	91.4%	89.4%	82.8%	81.4%	79.4%	
Precision	94.5%	93%	83.9%	83.7%	80.6%	80.8%
Sensitivity	98.8%	92.7%	95%	93%	94.6%	
Specifity	84.6%	80.2%	48.3%		36.3%	
	ניתן לראות שמודל הניתוח <u>הראשון</u> סיפק אחוזי דיוק גבוהים יותר.		ניתן לראות ששני המודלים סיפקו אחוזי דיוק יחסית דומים.		ניתן לראות ששני המודלים סיפקו אחוזי דיוק דומים מאוד.	