PreSee
Data Detection Diagnosis

PreSee

מדע נתונים יישומים ליזמים

ד"ר גייל גלבוע פרידמן

ד"ר נווה אשכנזי

PreSee
Data Detection Diagnosis

שאלת המחקר

לאחר חלוקה של הלקוחות לתתי קבוצות
על פי השאלון הרפואי,
<u>האם ניתן לנבא את ה- Overall Score</u> עם
ביצוע <u>רק חלק</u> מהבדיקות הרפואיות
היקרות?

PreSee
Data Detection Diagnosis

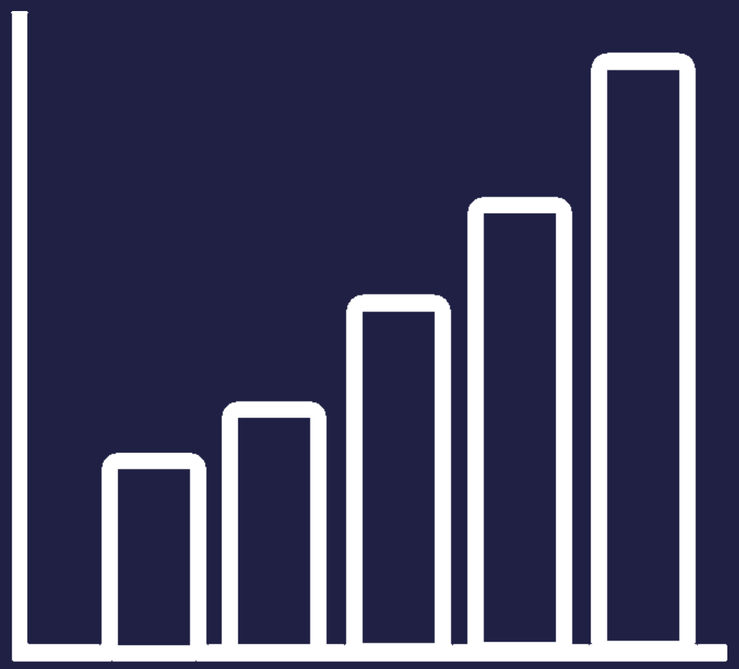
הסתכלות עסקית

הציון הסופי מחושב על ידי משקולות קבועות שהוחלטו על ידי החברה. אנחנו מציעים (על בסיס המשקולות שנקבעו עד היום) מודל שמאפשר לשערך אותו עם פחות בדיקות.

-> קיצור ה-ONBOARDING של לקוח חדש, וחסכון בכסף.

PreSee
Data Detection Diagnosis

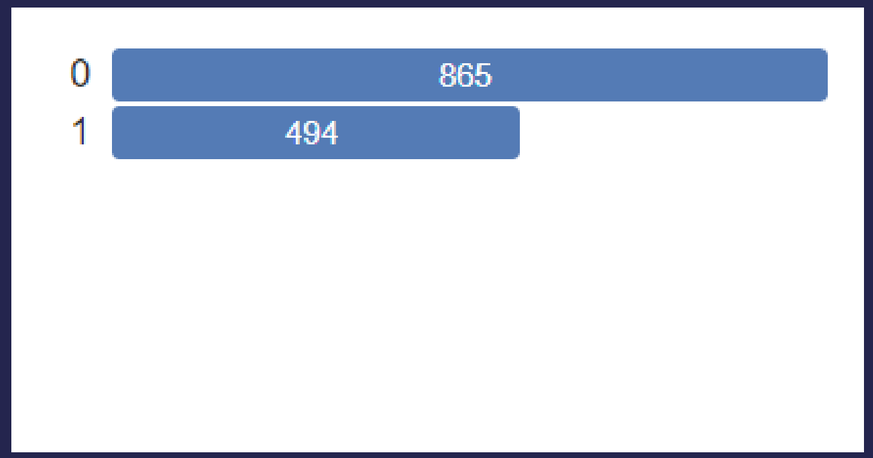
שלב א':



Data Exploration

Gender (Boolean)

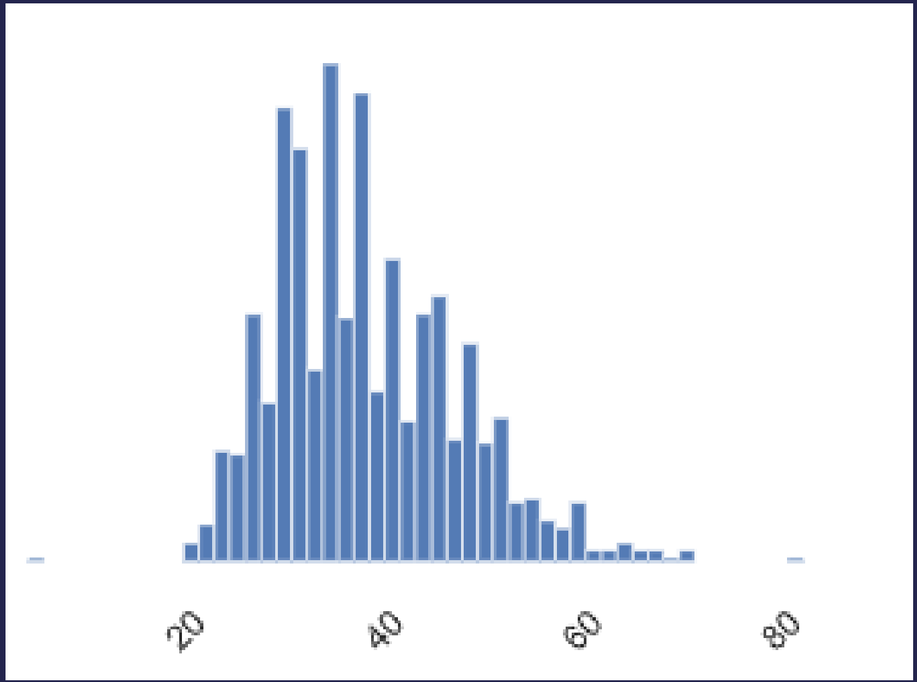
Distinct	2	Mean	
Distinct (%)	0.1	Minimun	-
Missing	0	Maximum	-
			-



**נשים לב שההתפלגות למגדר היא אחידה למרות שב-
DATA היא לא כך , אנחנו לא מבצעים נרמול
(פירסון/גאוסיאני) כדי למנוע הטיות מגדריות בלקוחות של
החברה (רוב הלקוחות נכון להיום הם גברים)

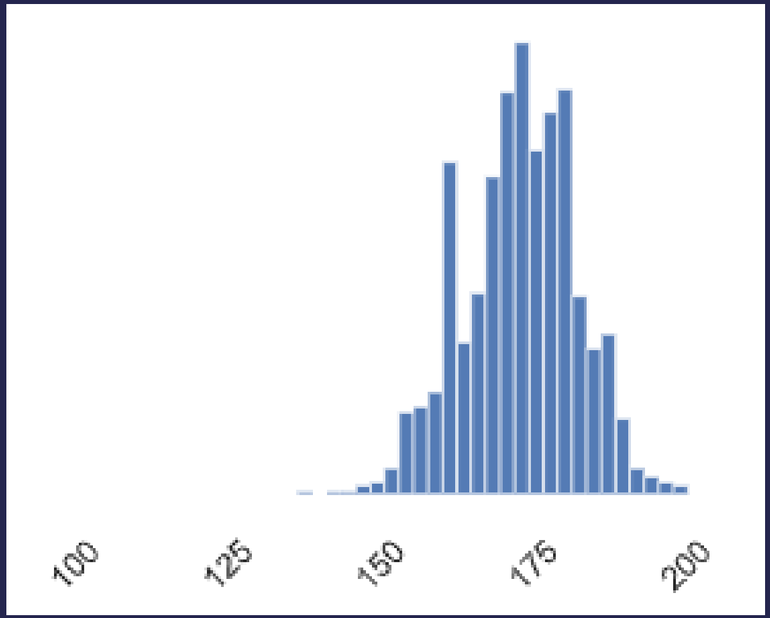
Age

Distinct	51	Mean	38.015
Distinct (%)	3.8	Minimun	4
Missing	0	Maximum	82



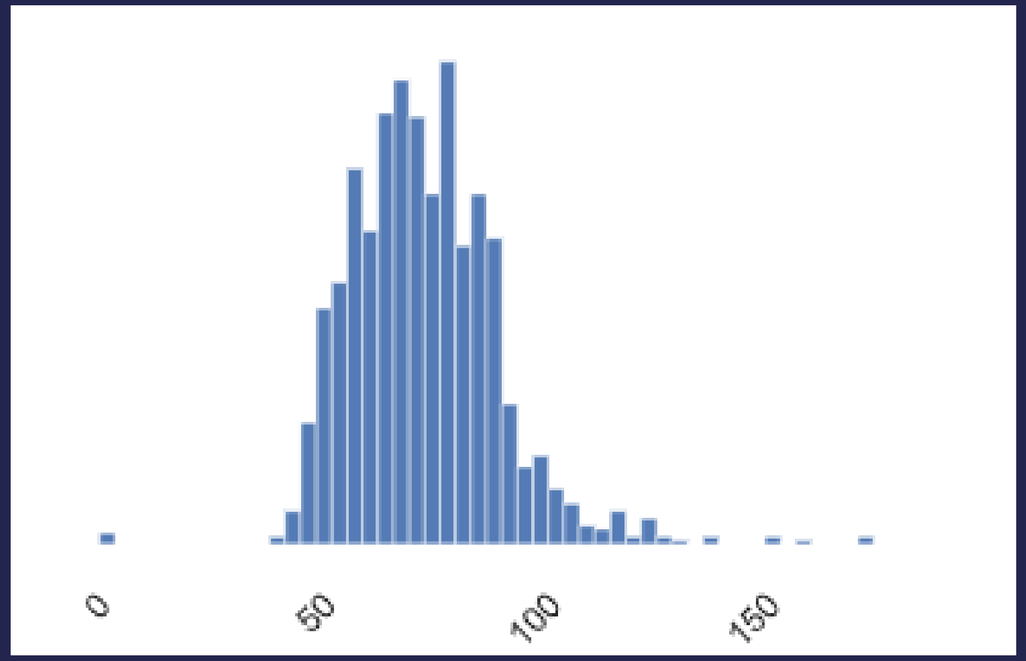
Heigth (cm)

Distinct	51	Mean	172.48
Distinct (%)	3.8	Minimun	80
Missing	0	Maximum	200



Wiegth (Kg)

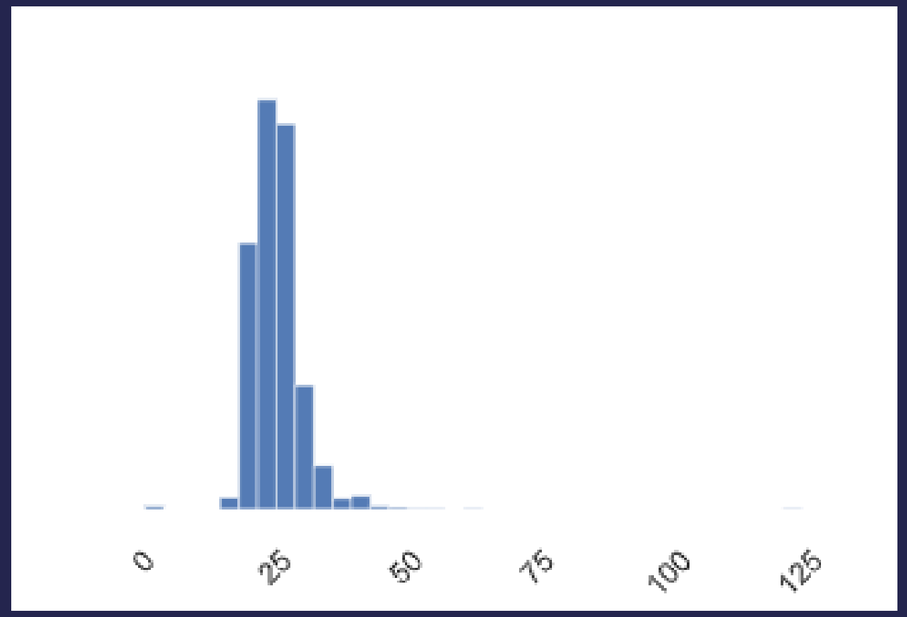
Distinct	86	Mean	74.249
Distinct (%)	6.3	Minimun	0
Missing	0	Maximum	177



BMI

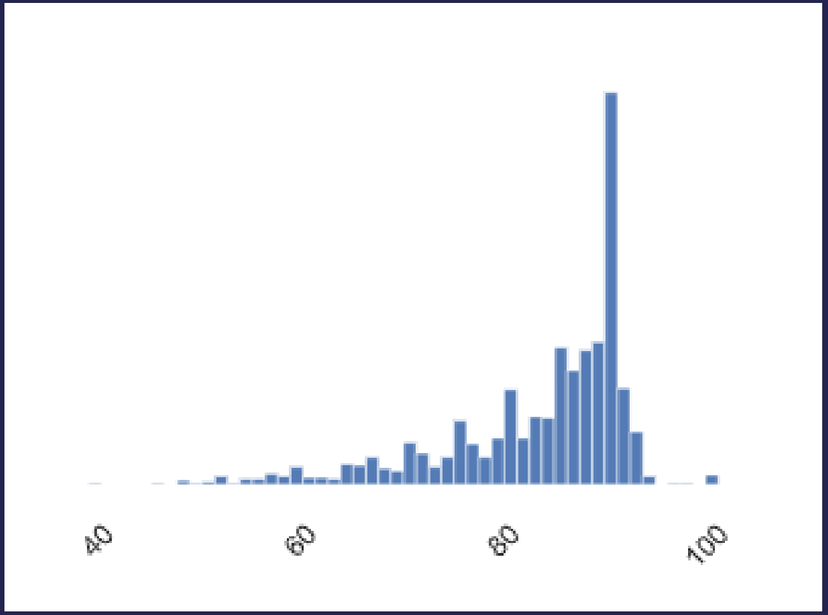
Distinct	35	Mean	24.877
Distinct (%)	2.6	Minimun	0
Missing	0	Maximum	125

נשים לב שה BMI מתפלג נורמלית, אך בנוסף לזה הוא
פונקציה (לא ליניארית) של משקל וגובה. אנחנו מעוניינים
לשמור את הFEATURE הזה כי הרגרסיה הליניארית לא תדע
לתמחר את הנוסחה $\mathrm{BMI} = \frac{\mathrm{weight}}{\mathrm{height}^2}$



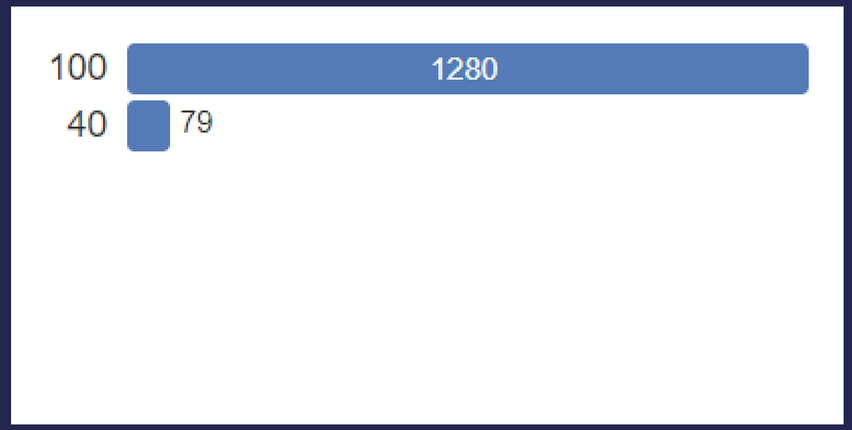
Blood-Test

Distinct	51	Mean	83.320
Distinct (%)	3.8	Minimun	39
Missing	0	Maximum	101



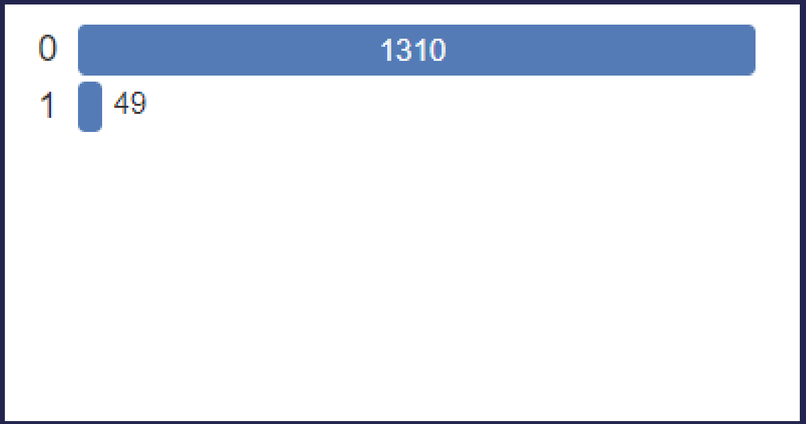
ECG Test Score

Distinct	2
Distinct (%)	5.1
Missing	0



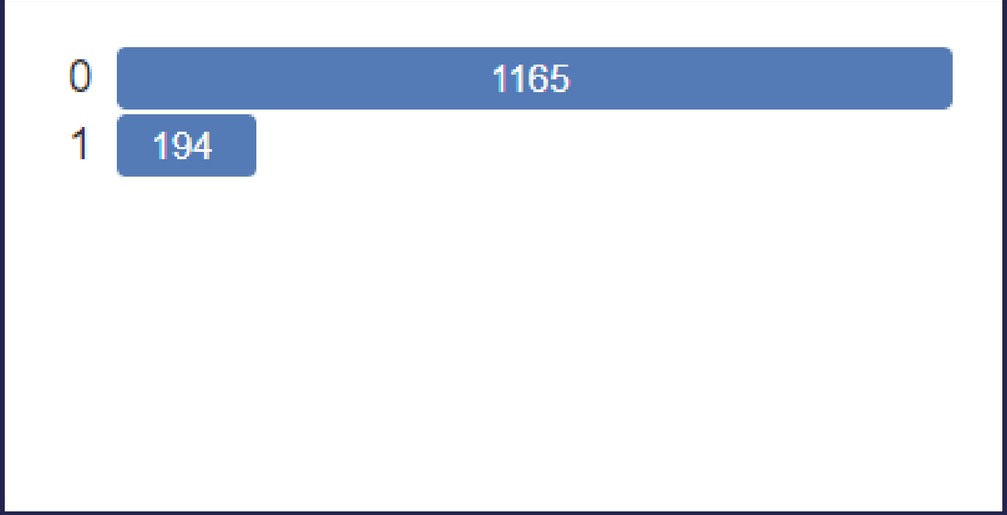
US - Test score

Distinct	7	Mean	97.317
Distinct (%)	3.6	Minimun	36
Missing	0	Maximum	100



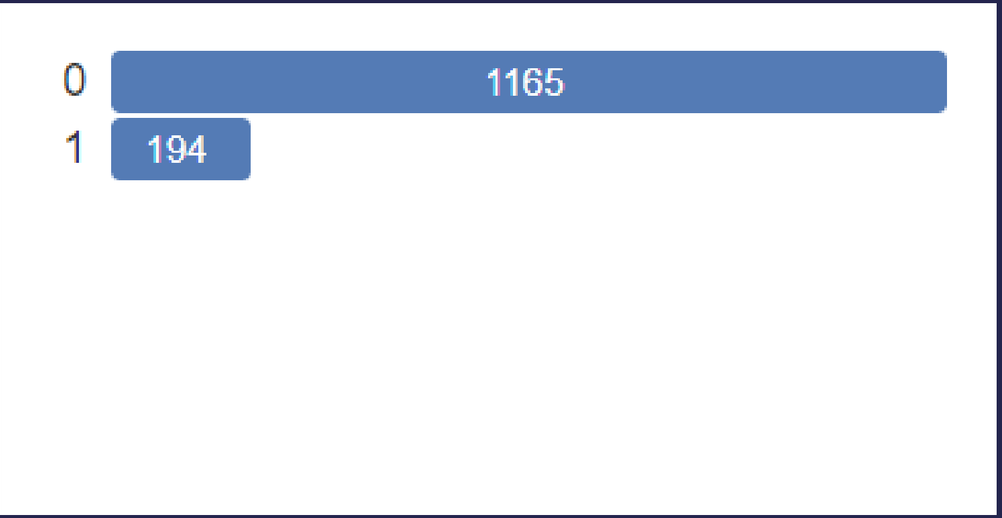
Smoking

Distinct	2
Distinct (%)	14.2
Missing	0



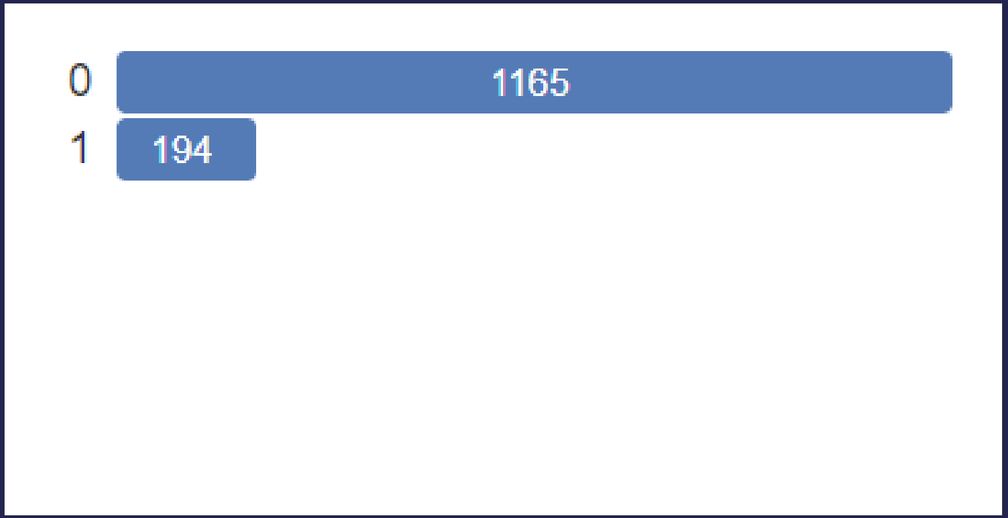
Heart Diseases

Distinct	2
Distinct (%)	0.1
Missing	0



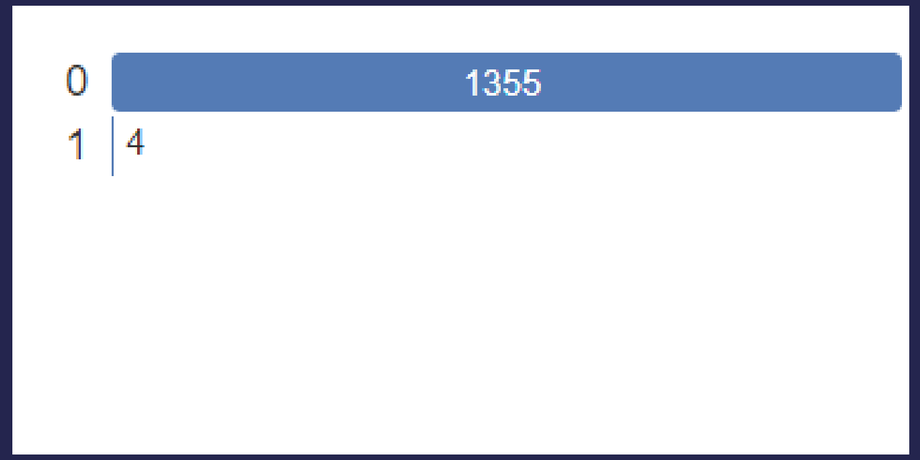
BP Medication

Distinct	2
Distinct (%)	0.1
Missing	0



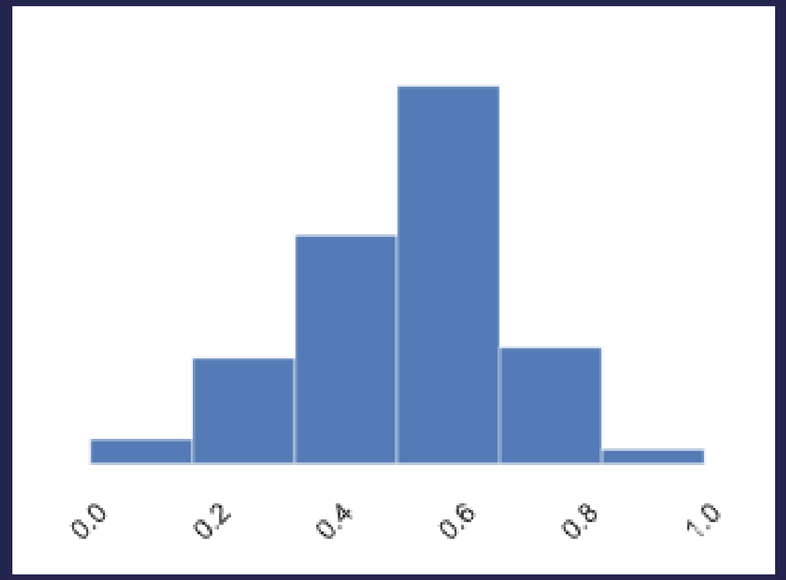
Diabetes

Distinct	2
Distinct (%)	0.1
Missing	0



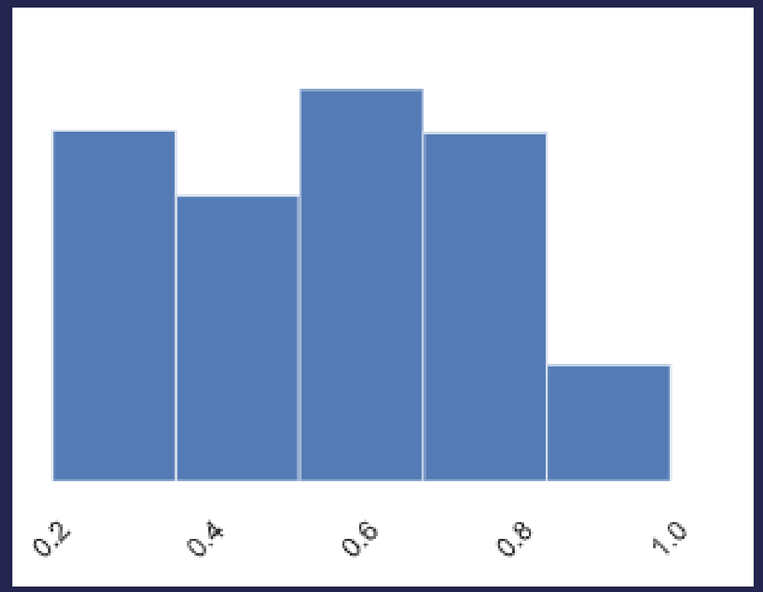
Work Stress Level

Distinct	6	Mean	0.514
Distinct (%)	0.4	Minimun	0
Missing	0	Maximum	1



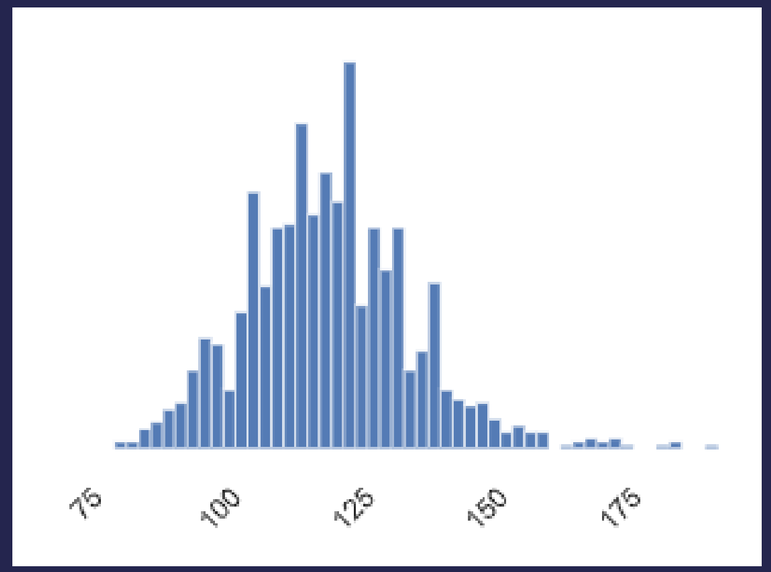
Excersie Level

Distinct	5	Mean	0.545
Distinct (%)	0.4	Minimun	0.2
Missing	0	Maximum	1



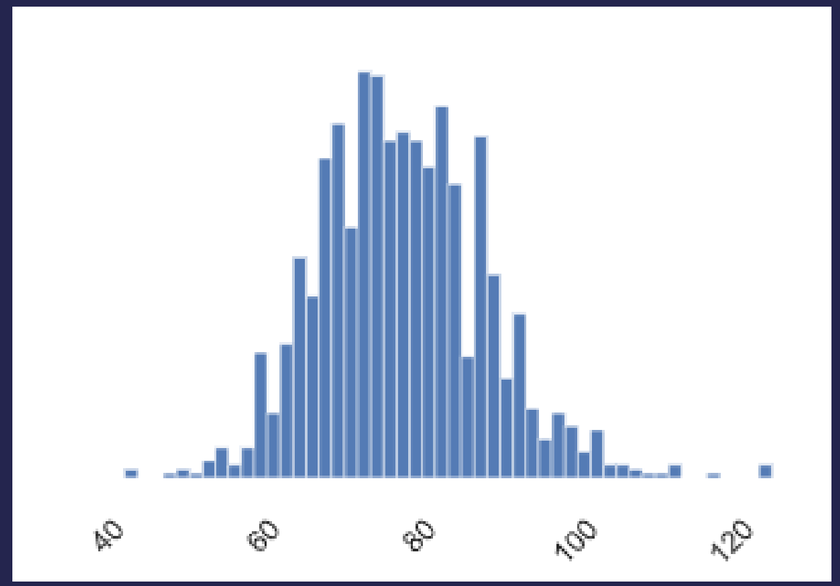
BP - Systolic

Distinct	90	Mean	120.209
Distinct (%)	6.6	Minimun	80
Missing	0	Maximum	193



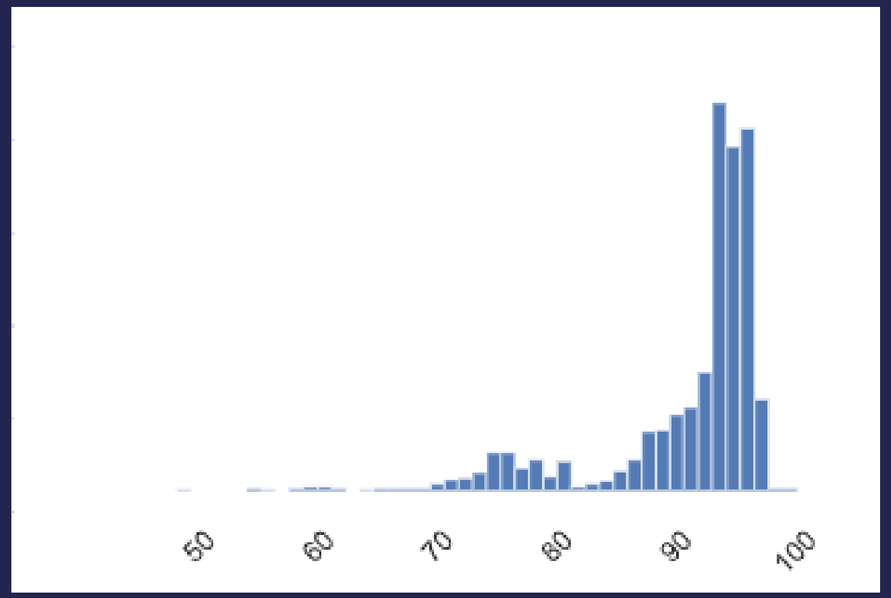
BP - Diastolic

Distinct	65	Mean	77.578
Distinct (%)	4.8	Minimun	42
Missing	0	Maximum	125



Overall-Score

Distinct	44	Mean	90.873
Distinct (%)	3.2	Minimun	48
Missing	0	Maximum	100



ניתן לראות שיש שני גאוסיינים בפונקציית המטרה
הנחת עבודה: אחרי חילוק ל-K קבוצות יהיה לנו גאוסיין אחד
לכל היותר (בציון הסופי) לכל קבוצה

PreSee
Data Detection Diagnosis

01 נרמול ערכים בצורה נכונה עבור קלאסטרינג

קיום של ערכים בוליאניים (בין 0 ל-1) אילץ אותנו לנרמל את שאר הDATA הרציף בשאלון

02 טיפול ב-Outliers שנוצרו בגלל מידע חסר בשאלונים

BMI ממוצע בדאטה הינו 24.8
משקל ממוצע 74.2
גובה ממוצע 172.4

03 MIN-MAX Normalizaion

נרמול כל הערכים הרצפים להיות בין 0 ל-1 לאחר השלמת המידע

	gender	age	height_cm	weight_kg	bmi	smoking	heart_disease_hist	heart_disease_family_hist	bp_medication	diabetes	work_stress_level	exercise_level
0	0	0.474358974	0.745762712	0.293785311	0.144	0	0	0	0	0	0.4	0.8
1	1	0.282051282	0.771186441	0.338983051	0.16	1	0	1	0	0	0.4	0.2
2	1	0.435897436	0.762711864	0.367231638	0.176	1	0	0	0	0	0.4	0.8
3	0	0.448717949	0.661016949	0.378531073	0.208	0	0	0	0	0	0.8	0.6
4	1	0.371794872	0.796610169	0.446327684	0.208	0	0	0	0	0	0.4	0.6

PreSee
Data Detection Diagnosis

שלב ב':

התאמת המודל/ים

PreSee
Data Detection Diagnosis

חלק א

איך נבצע Clustering?

Community Detection - Gephi מתוך השאלון הרפואי בלבד

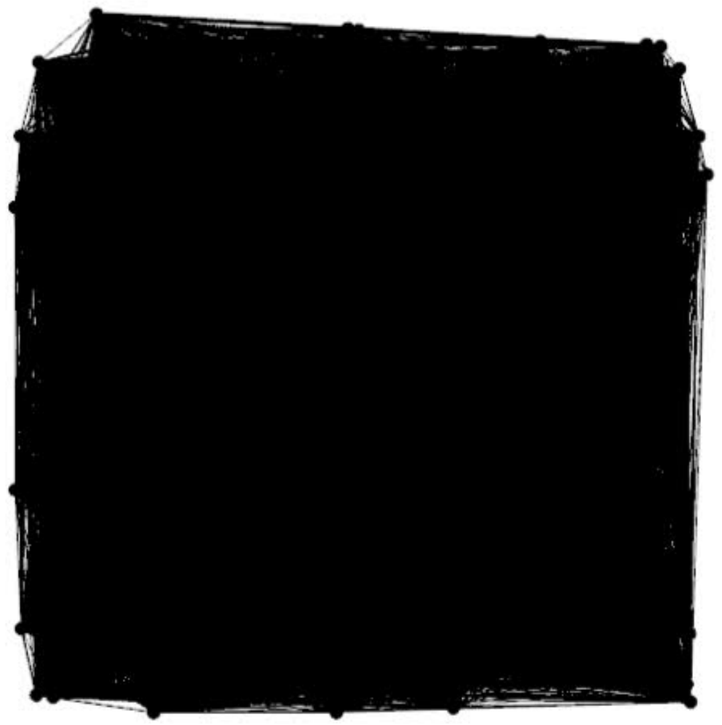
?

K-Means מתוך השאלון הרפואי בלבד

Gephi

רצינו לבדוק את המודלים שיש ל-Gephi להציע עבור Community Detection (בין היתר - Louvain method, Girvan Newman algorithm) בגלל סיבוכיות קשתות גבוהה מידי (NODES 1350 מוביל לסדר גודל של מיליון EDGES) נאלצנו להתעלם מרוב ה-DATA בכל פעם, מה שמנע מאיתנו לדעת את ציון המודולריות האמיתי.

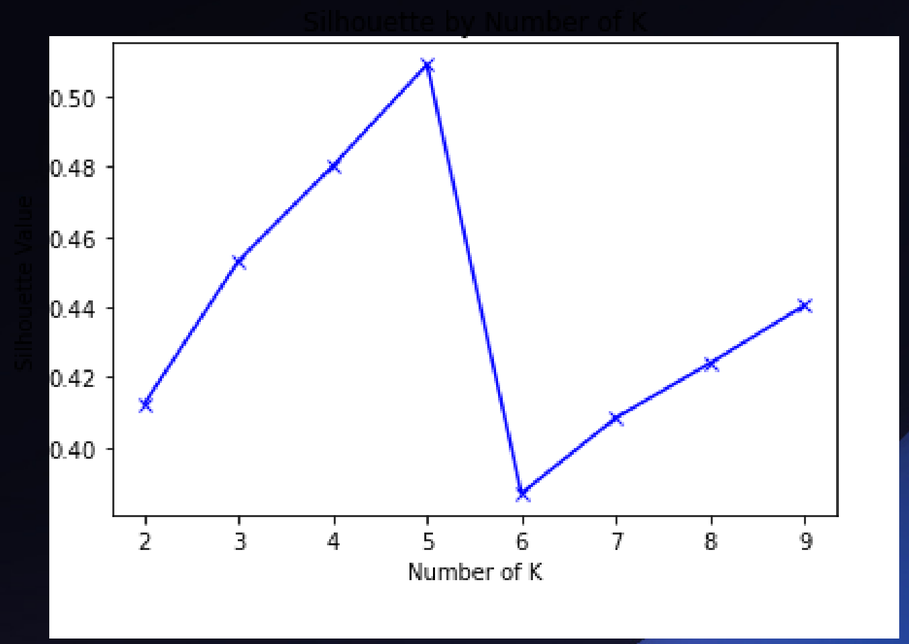
כדי לבדוק את המודל בחרנו 250 instances באקראי, הגדרנו גרף מלא (Edges 31,125) כאשר המשקל על כל קשת (Undirected) הוא המרחק האוקלידי ב-Data המנורמל.



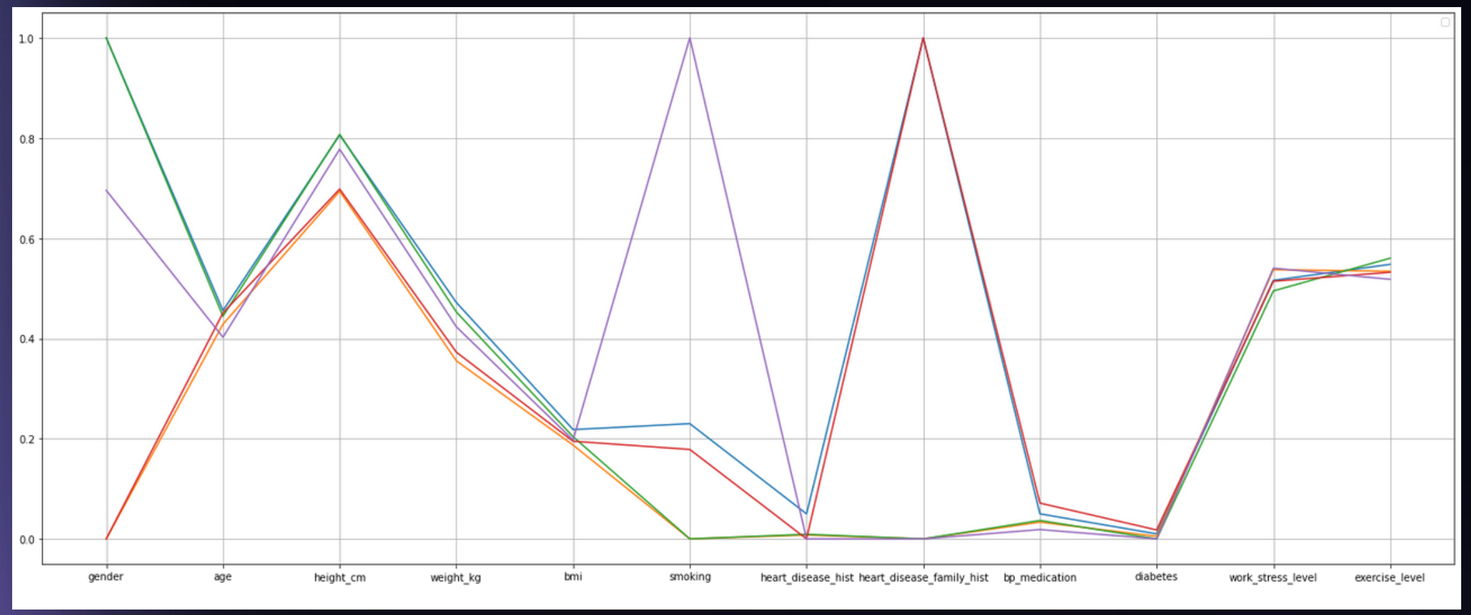
** מידול 250 INSTANCES

מודל K-Means

בחרנו K=5 על בסיס ה-Silhouette Score



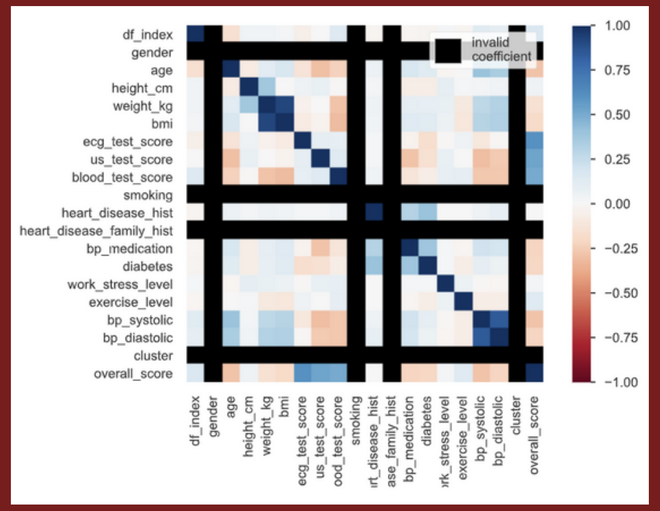
חלוקה ל Clustring לפי השאלון הרפואי:



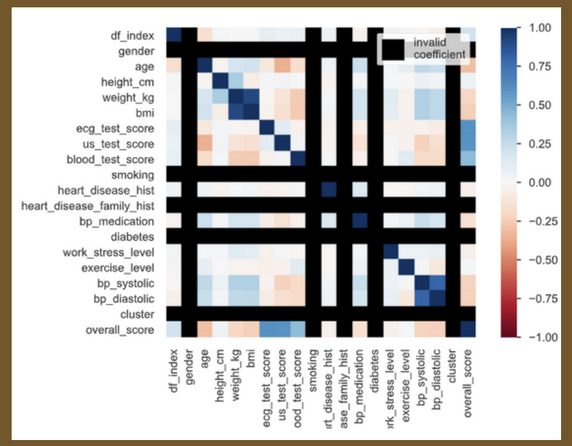
ניתוח Clustering

אדום

נשים
BMI נמוך יחסית
ללא היסטוריה של מחלות לב,
אך לאחוז גבוה קיימת היסטוריה במשפחה
אחוז נמוך של מעשנים



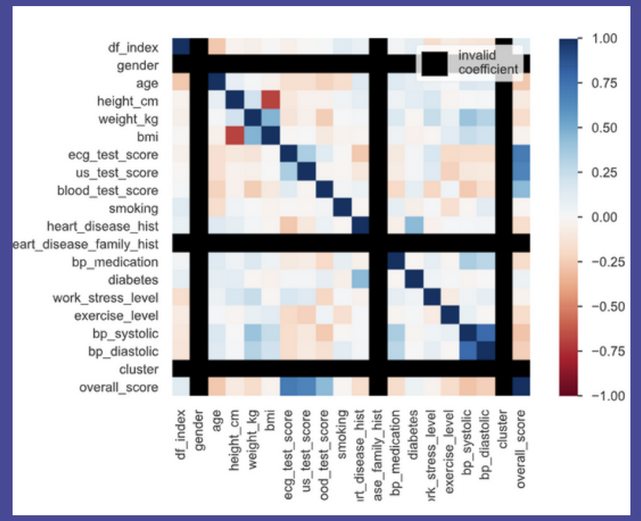
כתום



נשים
BMI נמוך יחסית
ללא היסטוריה של מחלות לב, גם במשפחה
אינם מעשנים

סגול

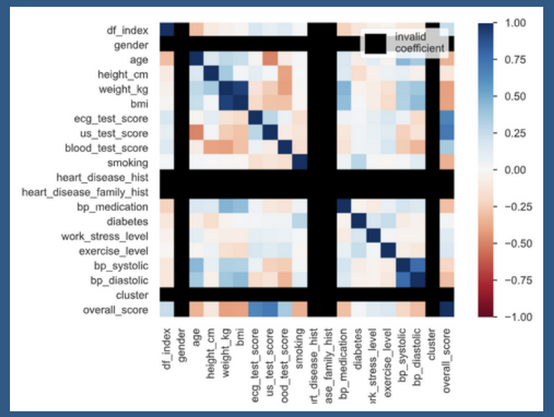
מגדר מעורב
מעשנים
ללא היסטוריה של מחלות לב
(גם לא במשפחה)



ניתוח Clustering

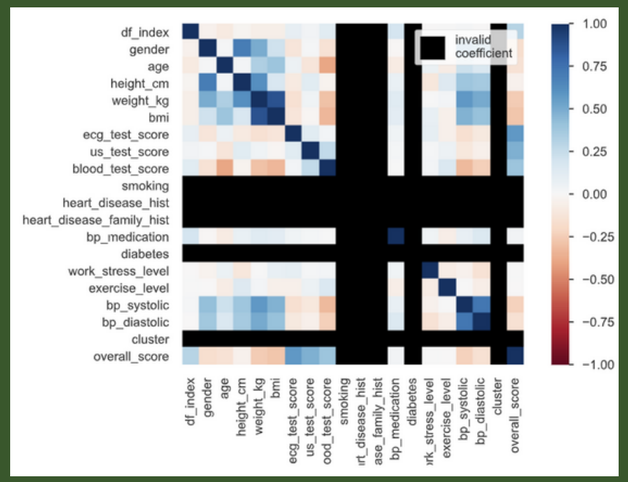
כחול

גברים
BMI נמוך יחסית
ללא היסטוריה של מחלות לב,
אך לרובם היסטוריה משפחתית
אחוז נמוך של מעשנים



ירוק

גברים
BMI נמוך יחסית
ללא היסטוריה של מחלות לב, גם במשפחה
אינם מעשנים



PreSee
Data Detection Diagnosis

חלק ב

ניבוי ה-Overall Score

Linear Regression for each Cluster

?

Decision Tree for each Cluster

רגרסיה לינארית

ביצענו רגרסיה כשבכל פעם התעלמנו באחת מתוצאות הבדיקות (בדיקת דם, א.ק.ג. ואולטרה-סאונד)
מצאנו שרמת הדיוק אינה גבוהה ביחס לכמות המידע הנתון לאימון, בחרנו לעבוד עם עצי החלטה

Linear Regression w python

R2-Score after deducting the medical test	הבדיקה שעבורה התקבל ה-SCORE הגדול ביותר לאחר שהפחיתו אותה	R2-Score before deducting the medical test	Clustering
0.731	blood_test_score	0.895	0
0.727	blood_test_score	0.817	1
0.428	blood_test_score	0.358	2
0.512	blood_test_score	0.632	3
0.639	blood_test_score	0.728	4

Linear Regression w BigML

עץ החלטה

נוכל לראות כי אחוז הדיוק נמוך יותר מאשר במודל הרגרסיה הלינארית, עצי החלטה בנויים לקסלסיפיקציה, לכן החלטנו לבצע התאמות ב-Overall_score.

Clustering	ללא הפחתת הבדיקה	הבדיקה שעבורה התקבל ה-SCORE הגדול ביותר לאחר שהפחיתו אותה	הציון לאחר הפחתת הבדיקה
0	0.572	blood_test_score	0.213
1	0.551	us_test_score	0.551
2	0.347	us_test_score ecg_test_score	0.347
3	0.25	us_test_score ecg_test_score	0.166
4	0.5	us_test_score ecg_test_score	0.5

PreSee
Data Detection Diagnosis

עץ החלטה מעודכן

חילקנו את הציון הכללי ל-10 טווחים אפשריים מתוך הנחה שציון ה overall score בטווח זה יספיק לחברה מבחינה עסקית*

לכל קלסטר, ניתן לראות את הבדיקה שהשפיעה הכי פחות על הציון הכללי, לאחר ההפחתה שלה

Cluster	ללא הפחתת הבדיקה	ecg_test_score	us_test_score	blood_test_score
0	0.908	0.862	**0.870**	0.832
1	0.910	0.884	**0.8974**	0.871
2	0.913	**0.913**	**0.913**	0.782
3	0.916	0.833	**0.916**	0.666
4	0.9	0.833	**0.9**	0.8

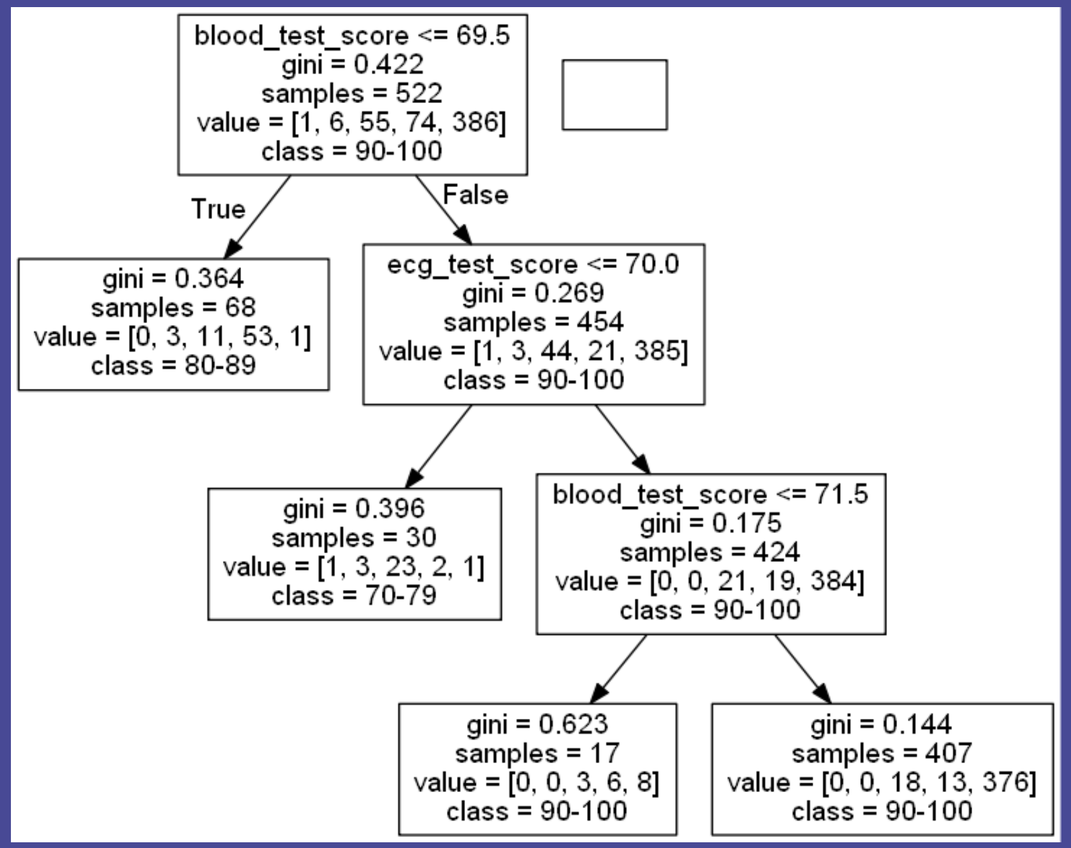
מקרא:

0 0-9
1 10-19
2 20-29
3 30-39
4 40-49
5 50-59
6 60-69
7 70-79
8 80-89
9 90-100

*ניתן לשנות בהתאם לצרכי החברה

תהליך הניבוי

עץ החלטה - Cluster 0



התוצאה לפניי הורדת הבדיקה הייתה : **0.9083969465**

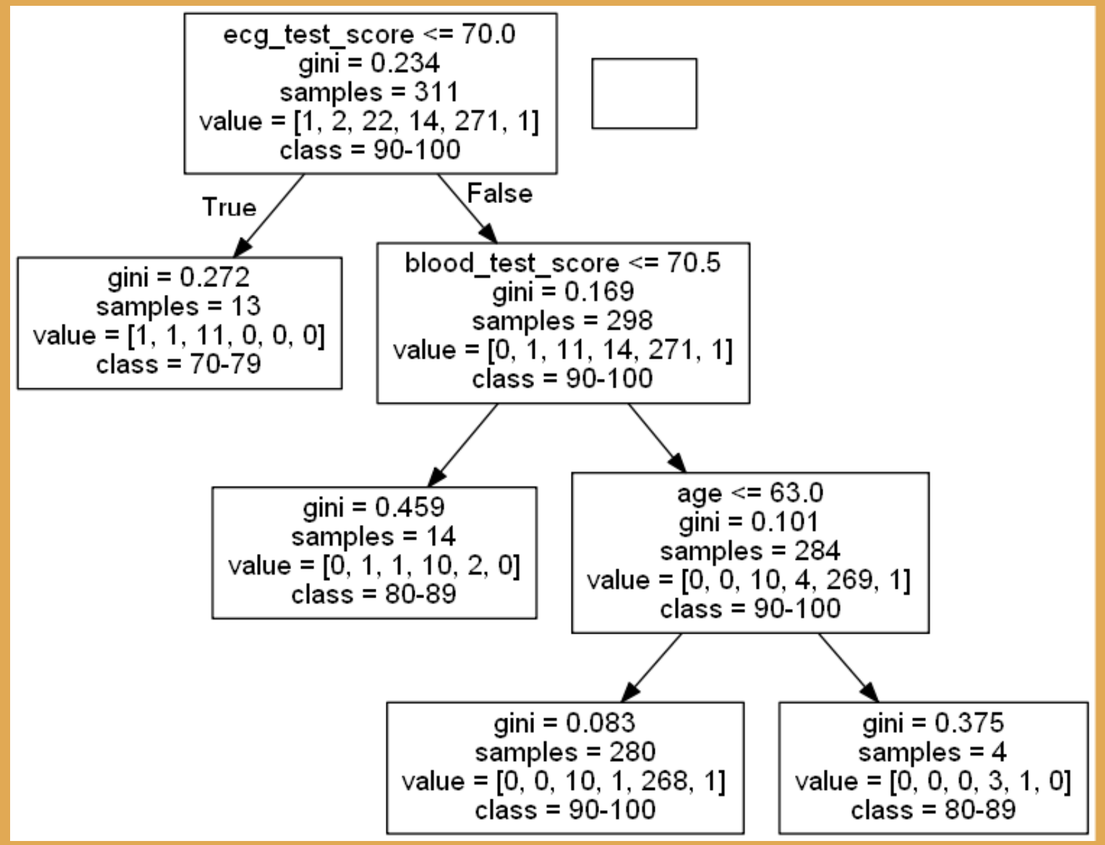
לאחר הורדת **בדיקת האולטרסאונד** התקבל: **0.8702290076**

לכן נמליץ עבור Cluster 0 -

להוריד את בדיקת האולטסאונד

תהליך הניבוי

עץ החלטה - Cluster 1



התוצאה לפניי: **0.91025641025**

לאחר הורדת **בדיקת האולטרסאונד** התקבל:

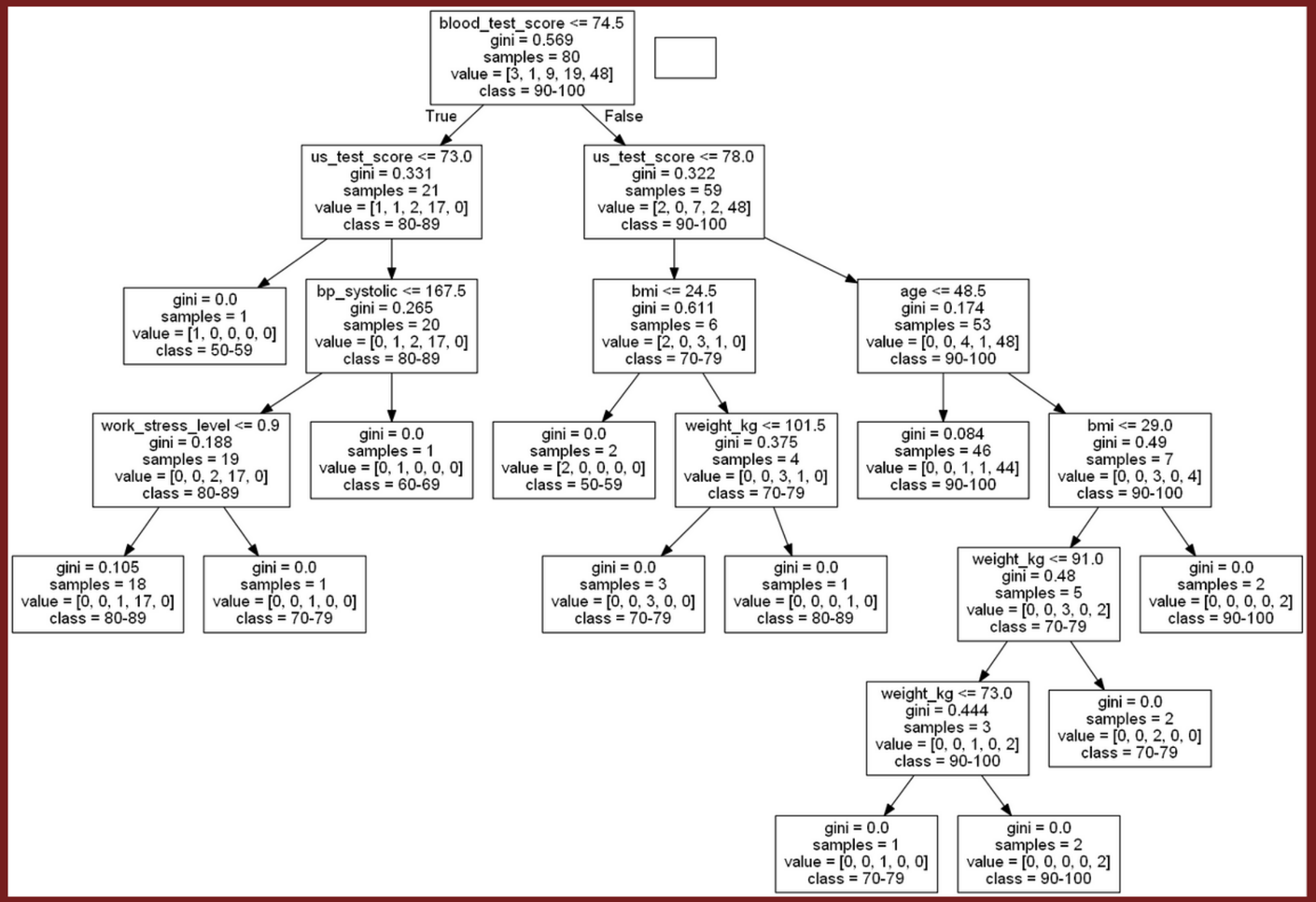
0.89743589743

לכן נמליץ עבור Cluster 1 -

להוריד את בדיקת האולטסאונד

תהליך הניבוי

עץ החלטה - Cluster 2



נשים לב כי התוצאה לפניי : **0.91304347826**

לאחר הורדת **בדיקת האק"ג** התקבל:

0.91304347826

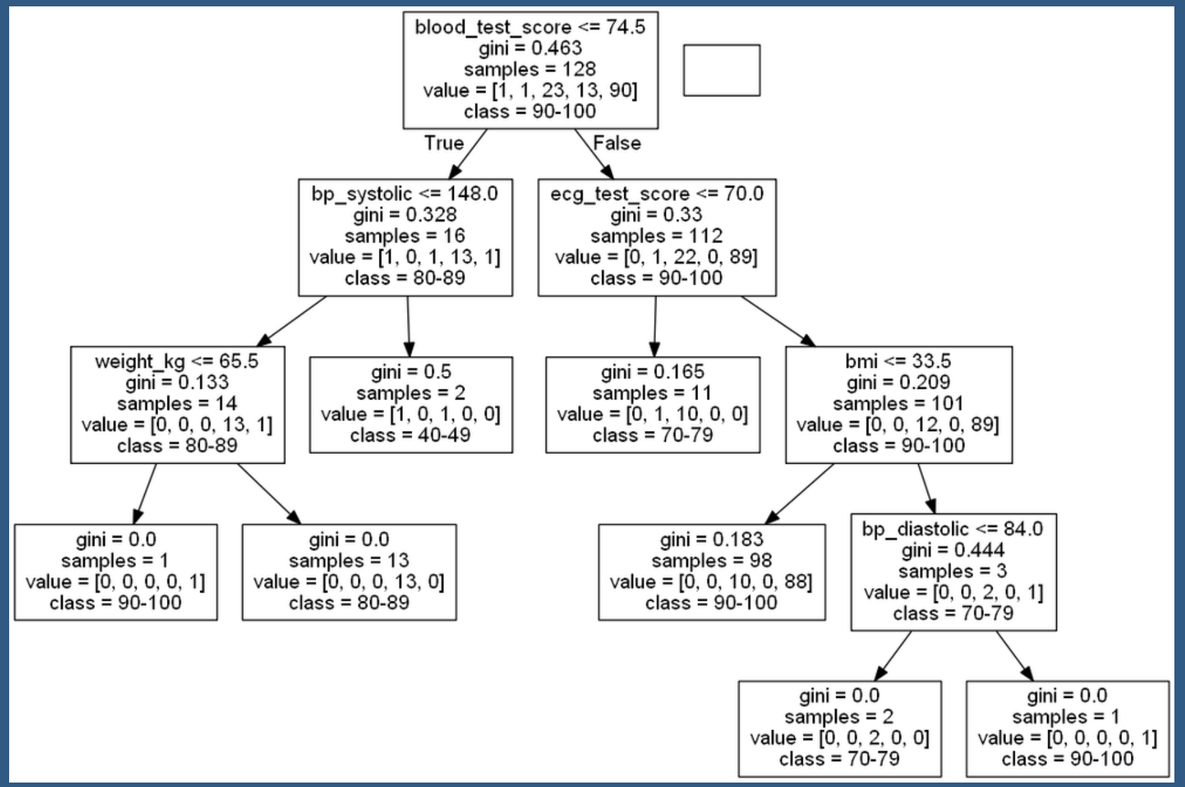
לכן נמליץ עבור Cluster 2 -

להוריד את בדיקת האולטסאונד

או להוריד את בדיקת האק"ג

תהליך הניבוי

עץ החלטה - Cluster 3



התוצאה לפניי : **0.91666666666**

לאחר הורדת **בדיקת האולטרסאונד** התקבל:

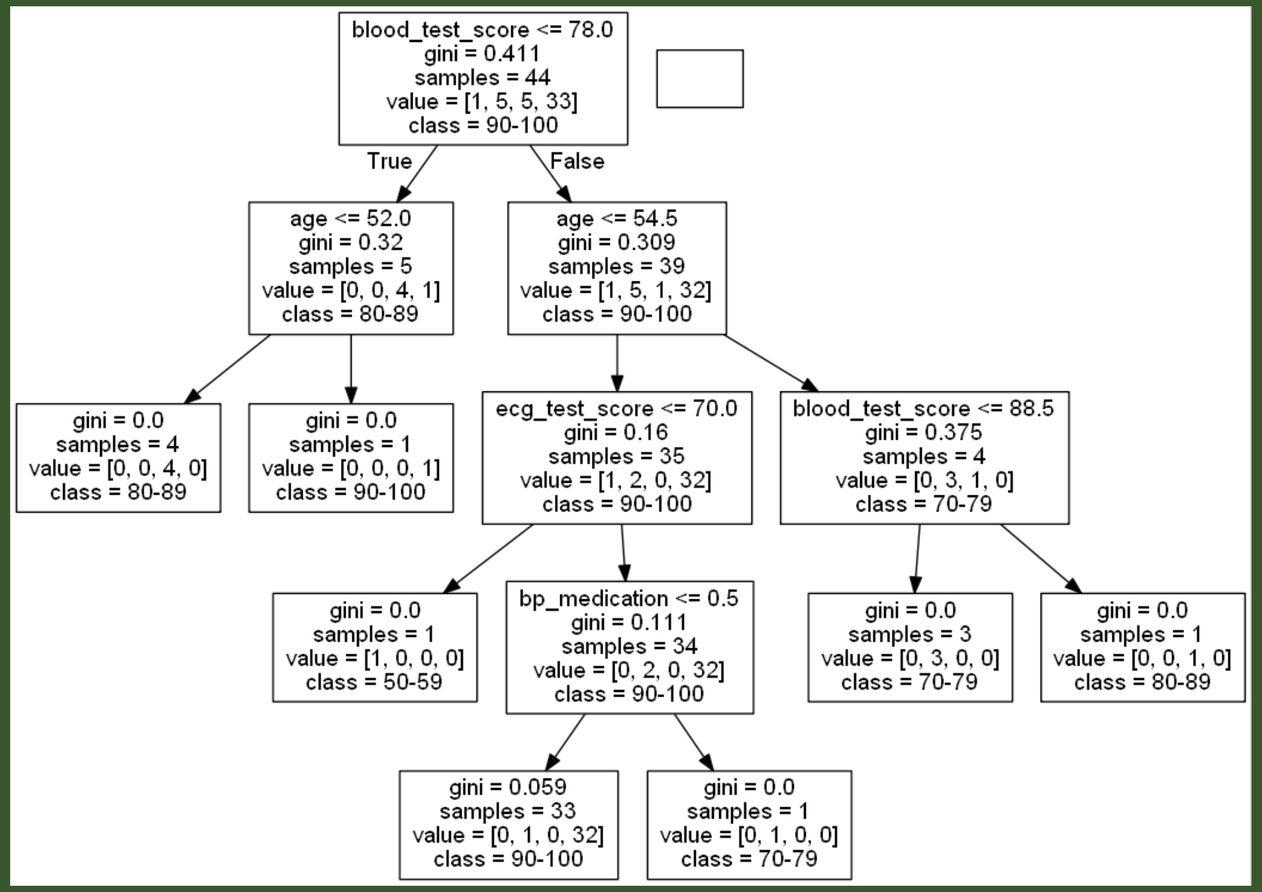
0.91666666666

לכן נמליץ עבור Cluster 3 -

להוריד את בדיקת האולטסאונד

תהליך הניבוי

עץ החלטה - Cluster 4



התוצאה לפניי : **0.9**

לאחר הורדת **בדיקת האולטרסאונד** התקבל: **0.9**

לכן נמליץ עבור Cluster 4 -

להוריד את בדיקת האולטסאונד

PreSee
Data Detection Diagnosis

סיכום

Flow של לקוח חדש בחברה

שלב א'

התאמה לקלאסטר
על בסיס שאלון רפואי

שיוך לקלאסטר

שלב ב'

ביצוע 2 בדיקות רפואיות
יקרות במקום 3

הכנסה
ל-DT הרלוונטי
(לפי הקלאסטר)

שלב ג'

ניבוי ה-Overall_score
בעץ המתאים לקלאסטר

הצוות

יותם גבי

איתי גולדמן

מיכאל ידידיה

ברק אמזלג

דור שלום