מדע נתונים ליזמים - עבודה קבוצתית מסכמת 2022

מגישים: יותם גבי (ת.ז 203074224) | איתי גולדמן (ת.ז | מיכאל ידידיה (ת.ז 314988122)
ברק אמזלג (ת.ז 206218513) | דור שלום (ת.ז 316383181)

שלב 1 - הגדרת המטרה

אי ספיקת לב יכולה לפגוע בכל זמן ובכל מקום. גילוי מוקדם של מחלות שכיחות ומסכנות חיים כמו מחלות לב, כלי דם וסרטן הוא הכרחי להשגת רפואה מונעת. חברת PreSee מפעילה פלטפורמה שנבנתה על ידי צוות של רופאים מומחים ומדעני נתונים. המערכת בוחנת בקפידה כל פרמטר מהשאלון שממולא ע"י לקוחות החברה, ומביאה בחשבון את כל גורמי הסיכון הפוטנציאליים לאחר ביצוע מספר בדיקות בעלות שאינה זניחה. לכל לקוח, המערכת מייצרת פרופיל סיכון אישי מפורט הן לטווח הקצר, הבינוני והארוך בהתאם לפרמטרים שהוזנו על ידו ולמדדים הקטגוריים שהוגדרו לאחר ביצוע ארבעת הבדיקות להלן (Corso, 2021; *How Much Does an Ultrasound Cost*, 2021; Slobin, 2022):

1. בדיקות דם - עלות ממוצעת של 50-100 דולר.
2. בדיקת ECG - עלות ממוצעת של 175-299 דולר.
3. בדיקת אולטרסאונד - עלות ממוצעת של 380 דולר.

בעבודה זאת בחרנו להתמקד במזעור הבדיקות שחברת PreSee תעשה לכל לקוח, ללא פגיעה בידיעת רמת הסיכון של הלקוחות, במטרה להוזיל את עלות התהליך של קבלת פרטים מכל לקוח.

שאלת המחקר:
לאחר חלוקה של המטופלים לתתי קבוצות על פי השאלון הרפואי, האם ניתן לנבא את ה- Overall Score באמצעות ביצוע חלק (2 או פחות) מהבדיקות הרפואיות היקרות (דם, ECG, US).

כיום, ה-**Overall Score** מחושב על ידי משקולות קבועות שהוחלטו על ידי הדרג המקצועי בחברה. אנחנו למעשה מציעים מודל שיאפשר לשערך אותו עם פחות בדיקות, שיהיה מבוסס למידה של המשקולות הנוכחיות. במילים אחרות, אם שאלת המחקר תתברר כנכונה, המודל שלנו יאפשר לבצע **שערוך** לציון הכללי על ידי למידה של השקלול שהוחלט על ידי הדרג המקצועי. אם הדרג המקצועי יבחר לשנות את המשקולות, עלינו לבצע שערוך מחדש של רלוונטיות הבדיקות בהתאם למשקולות החדשים (לחילופין, הדרג המקצועי בחברה יכול לקחת את המודל שנציע). ברמה העסקית, הוצאות החברה פר מטופל יקטנו ותהליך ה-Onboarding של מטופל חדש יתקצר.

השערת המחקר:
תחת מידע אפוסטריורי שמבוסס על חלוקה לתתי קבוצות ייחודיות*, ניתן יהיה לקבוע בדיוק גבוה את ציון הoverall score ע"ב 2 בדיקות לכל היותר.

*למשל - קבוצה 1: מעשנים + בעלי היסטוריה של מחלות לב, קבוצה 2: אנשים שאינם מעשנים וללא רקע משפחתי למחלות, קבוצה 3: אנשים שנוטים לעשות פעילות גופנית שאינם מעשנים. אפשר להגיד שההיפותזה שלנו גורסת שבהינתן מידע אפוסטריורי כזה אפשר לשערך מה יהיה הציון הסופי גם עם ביצוע של 2 מתוך 3 הבדיקות.

S.M.A.R.T Goal; **Specific**- שאלת המחקר שלנו ספציפית. הגדרנו תת קבוצה ברורה של בדיקות יקרות שאנחנו מעוניינים לצמצם לכל לקוח, כאשר יש תג מחיר לכל בדיקה. **Measurable** - לכל מטופל נוכל למדוד את השגיאה ב-overall score לאחר הורדת אחת הבדיקות. ה-Data הינו Labeled במובן הזה, לכל קלאסטר נוכל להגדיר Test Set. נקבע את הצלחת שאלת המחקר ע"י אחוז השגיאה- ככל שהוא יקטן כך שאלת המחקר תהיה עם פוטנציאל עסקי גדול יותר. כפי שצוין, עלות הבדיקות ידועה (ההוצאות הגולמיות אינן נמסרו לחברי הקבוצה מהחברה לכן מדובר בשערוך בלבד) ולכן גודל החיסכון בהשוואה לדיוק של המודל הינו נתון מספרי מדיד שיכול לשמש גם להשוואה בין מספר מודלים שונים תחת אותה שיטה. **Achievable**- היעד הינו בר השגה כי קיימים אלגוריתמים מוגדרים היטב למודל. בעזרת אלגוריתם לחלוקה ולאחר מכן שימוש בDecision Tree / Linear Regression - לטובת השערוך, נוכל לומר לכל אחת מתתי הקבוצות על איזו בדיקה ניתן לוותר. **Relevant**- מטרתה של חברת Presee היא לנבא בצורה המדויקת ביותר את הסיכוי של לקוח ספציפי לחלות במחלות מסכנות חיים כיום, בעתיד הקרוב ובעתיד הרחוק תוך מיקסום המשאבים שברשותם (שכן מדובר בחברה למטרות רווח). שאלת המחקר שלנו תאפשר לחברה לחסוך בעלויות השוטפות של הבדיקות ועדיין לעמוד בהצעת הערך שלה ללקוח. **Time Based**- פרק הזמן שבו אנו בודקים את שאלת המחקת שלנו מוגבל - 31.07.2022.

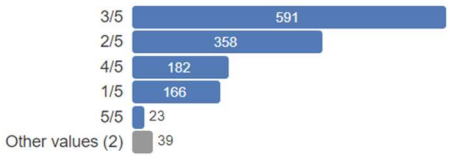
קריטריון הצלחה: דיוק של מעל 75% עבור קבוצת הטסט.

חלוקת התפקידים בצוות ולוחות הזמנים שהגדרנו:

איטרציה 0: ניקוי ה-Data וביצוע Exploration	**איטרציה 1:** בניית מודל להתפלגות תת קבוצות - Clustering	**איטרציה 2:** בחירת מודל וסנכרון	**איטרציה 3:** מידול ה- Prediction לכל תת קבוצה שנמצאה	**איטרציה 4:** בחירת מודל לפרדיקציה, וביצוע ניתוח על בסיס קריטריון ההצלחה ושאלת המחקר
כלל חברי הקבוצה יחד	יותם ואיתי- Community Detection (Gephi)	כלל חברי הקבוצה יחד	יותם איתי ומיכאל- Decision Trees	כלל חברי הקבוצה
	ברק, דור ומיכאל - K Means		ברק ודור - Linear Regression (scikit / Big ML)	

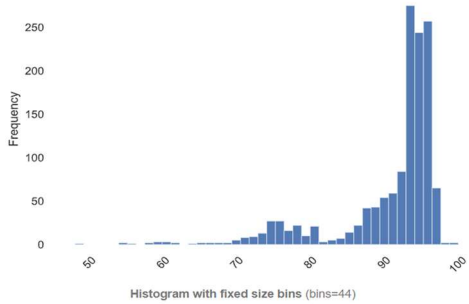
שלב 2 - Data Preparation & Exploration[1]

לפני שהחלטנו על אופן הכנת ה-Data סרקנו וניתחנו אותו כפי שהוא מופיע בצורתו הגולמית. נעזרנו בהרבה כלים כדי להבין איך כל תכונה מתפלגת במרחב המדגם שלנו, את מקדמי המתאם ביניהן (בין אם ליניאריים או יותר) וניסינו לזהות תבניות בסיסיות. מגדר - למרות שבמציאות ההתפלגות למגדר היא אחידה (אקראית לחלוטין) ב-Data היא לא כך. אנחנו לא מבצעים נרמול פירסון כדי למנוע הטיות מגדריות בלקוחות של החברה, שכן רוב הלקוחות נכון להיום הם גברים. גיל, גובה משקל ו-BMI מתפלגים עם גאוסיין אחד. למרות שה-BMI מתפלג נורמלית והינו פונקציה של קודמיו החלטנו לשמור את ה-Feature כי הרגרסיה הליניארית (באם תבוצע בהמשך בתור מודל שערוך) לא תדע לתמחר את הנוסחה הלא ליניארית ל-BMI. בדיקות דם ובדיקת לחץ דם - גאוסיין אחד. אק"ג - 5.81% עם בדיקה שאינה תקינה. אולטרסאונד - 3.6% עם בדיקה שאינה תקינה. שאר התכונות מהשאלון מתפלגות בצורה שאינה ידועה (בוליאנית), או באופן נורמלי עם גאוסיין אחד.



איור 1 – Work Stress Level (בדוגמה) - ניתן לראות שקיים גאוסיין שהממוצע שלו בין 0.6 ל-0.4, באתו אופן ניתן לראות שיש 39 ערכים שנדרש לתקן (שגיאות / פורמט שונה של מחזורת/ ערכים חסרים).

הדאטה מכיל פיצ'רים מהימנים ומדידים כמו גובה, משקל ולחץ דם מחד, ומאידך, מכיל פרמטרים סובייקטיביים שמבוססים על שאלון אישי כמו "רמת הלחץ בעבודה" ו-"מידת רמת הכושר הגופני" שמדורגים ע"י העובדים בטווח של 1-5 אינם מהימנים באופן מוחלט. כמו כן, קיימים ערכים קבועים - המשקולות שאיתן החברה מחשבת את הציון הסופי בממוצע המשוקלל. את המשקלות הנ"ל (עמודות) הסרנו מה-Data, הן נקבעות מראש באופן בלתי תלוי בדגימה לכן אינן תורמות לנו מידע.



איור 2 - Overall Score. תובנה מעניינת שהיוותה תומך החלטה לחלוקה לתתי קבוצות. ניתן לראות שיש 2 גאוסיינים (סביב 75 ו-95). בהקשר ההיפותזה שלנו - אם חלוקה לקלאסטרים על בסיס השאלון תוביל לסינון בסיסי שמחלק את הDATA ל-2 הגאוסיינים, המודל השערוכי יהיה ככל הנראה מדוייק יותר ללא הקלאסטרינג.

בשלב ראשון התכונות הרלוונטיות למציאת חלוקה (Partition) הן התשובות לשאלון והבדיקות הרפואיות הזולות. האופן בו אנחנו בחנו את חלוקת המידע נגזר מתוך המטרה העסקית - רק **לאחר השיוך לתת קבוצה** ספציפית אנחנו מעוניינים להגדיר איזה בדיקות יקרות על המטופל לבצע כדי לשערך את ה-Overall Score. בשלב השני השערוך מתבצע לפי כל ה-Features למעט אחת או שתיים מהבדיקות היקרות. לאור זאת, ובהתאם להיפותזה ולשאלת המחקר שהוגדרה החלטנו שכל ה-Features ב-Data רלוונטיים והואיל ולא ניתן לדעת מראש על איזה בדיקה ניתן לוותר, החלטנו לנרמל את כלל התכונות ל-Scale אחיד.

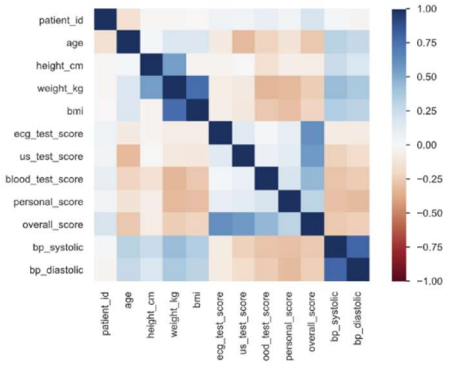
```
normalized_df.head()
```

	gender	age	height_cm	weight_kg	bmi	smoking	heart_disease_hist	heart_disease_family_hist	bp_medication	diabetes	work_stress_level	exe
0	0	0.474359	0.745763	0.293785	0.144	0	0	0	0	0	0.4	
1	1	0.282051	0.771186	0.338983	0.160	1	0	1	0	0	0.4	
2	1	0.435897	0.762712	0.367232	0.176	1	0	0	0	0	0.4	
3	0	0.448718	0.661017	0.378531	0.208	0	0	0	0	0	0.8	
4	1	0.371795	0.796610	0.446328	0.208	0	0	0	0	0	0.4	

איור 3 - Normalized Data

תחילה העברנו את כל הערכים שהיו בפורמט מחזורת לפורמט נומרי. קיומם של מדדים בוליאניים / בינריים (כמו עישון או היסטוריה משפחתית), אילצו אותנו לנרמל את שאר המידע הרציף (משקל, גובה) ואת המידע שנגזר מהשאלון לאותו הטווח, בין 0 ל-1. ביצענו זאת בעזרת Min-Max Normalization לאחר שטיפלנו ידנית ב-Outliers. בגלל שהנרמול מתבצע ביחס לערכי הקיצון כל תכונה שהייתה במרחק של יותר מ-3 **סטיות תקן** מהממוצע נורמלה לערך המתקבל ב-3 סטיות תקן.

כמו כן, חשוב לציין שהייתה לנו דילמה אם למחוק Instances בהם יש ערכים ריקים בתכונות מסיימות או למלא את אותם תאים ריקים בערך חציוני. בגלל שהערכנו שהחלוקה לקלאסטרים תוביל לכך שבכל תת קבוצה יהיו משמעותית פחות דגימות (~25%) החלטנו לא לוותר על אף שורה ב-Data ולבצע השלמה למידע.

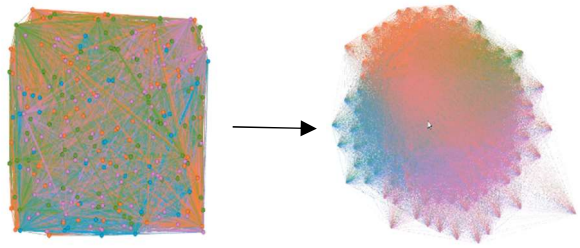


איור 4 - Pearson's r. כשהתבוננו בקשר שבין הפיצ'רים ראינו שיש מקדם מתאם חזק בין תוצאת האק"ג, האולטרסאונד, והציון האישי ל-Overall Score. תוצאה שלא הפתיעה אותנו שכן האחרון פונקציה ליניארית של קודמיו

[1] ניתוח ספציפי של כל feature נמצא **במצגת** שמצורפת בנספחים, כמו כן, צירפנו את הדו"ח שהפקנו - Data_Exploration_Before_Clustering.html.

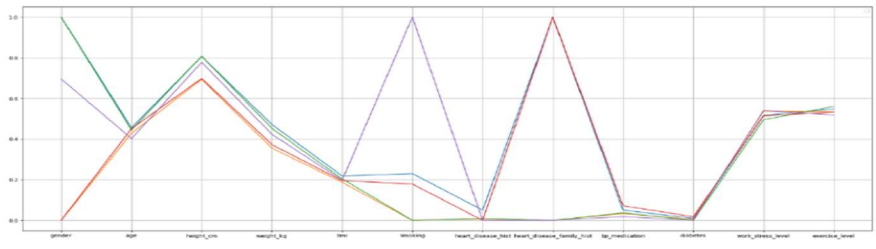
שלב 3 חלק א - Cluster Modeling

בדקנו 2 אופציות לחלוקה של ה-DATA לקבוצות זרות; הראשונה - Community Detection בעזרת Gephi (מגוון אלגוריתמים), והשנייה - K-Means. הניסיון להתאים את ה-Data כדי שיתמוך ב-Community Detection (Louvain method, Girvan Newman algorithm) היווה משימה לא פשוטה עבורנו. בגלל שלא קיימת הגדרה לקשת בין 2 צמתים (צומת = instance) במקרים שקיבלנו, נאלצנו להגדיר את הגרף כגרף מלא. במילים אחרות, בין כל 2 Instances הגדרנו קשת שמשקלה היה כמשקל המרחק האוקלידי ממעלה שנייה. הואיל וכל ה-Data היה מנורמל לאותה אמת מידה (בין 0 ל-1), כל Feature קיבל משקל אחיד בשקלול המרחק בין שני Instances. בגלל סיבוכיות קשתות גבוהה מידי (1350 צמתים - מובילים ל- 910,575=1380C2 קשתות), ובגלל שלא הצלחנו להגדיר גרף בצורה חכמה למעט הפתרון הנאיבי (גרף מלא) נאלצנו להתעלם מרוב ה-Data בכל איטרציה, מה שמנע מאיתנו לדעת את ציון המודולריות האמיתי מחד ומלתת ניתוח בצורה רוחבית מאידך. דוגמה משמאל[2].

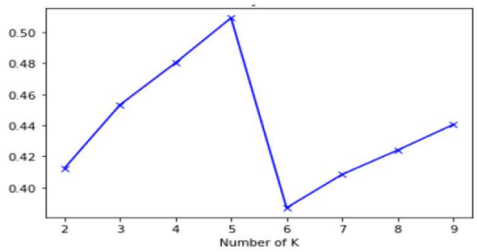


איור 5 - במקרה הנ"ל המודל בחרנו 250 Instances באקראי, הגדרנו גרף מלא (Edges 31,125) כאשר המשקל על כל קשת (Undirected) הוא המרחק האוקלידי ב-Data המנורמל

המודל השני, K-Means הניב תוצאות מספקות יותר. בחרנו K=5 על בסיס ה-Silhouette Score. וחילקנו את כל המקרים ל-5 תתי קבוצות בעלי תכונות משותפות. באופן לא מפתיע, החלוקה שהוגדרה לקבוצות הסתדרה עם האינטואיציה המקדימה שהייתה לנו.



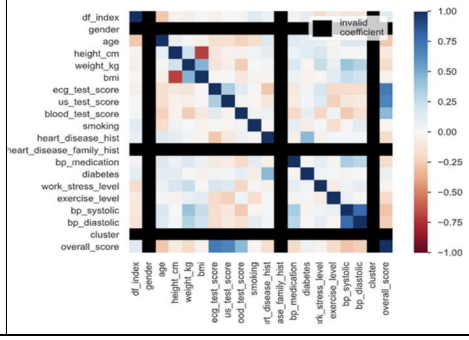
איור 7 - הקורדינטות של כל סנטרואיד במרחב מציגות לנו את למעשה את המאפיינים של כל קלאסטר, תמונה גדולה קיימת תחת תקיית "נספחים". ציר ה-x הוא ה-Features. כל צבע מייצג סנטרואיד של קלאסטר.



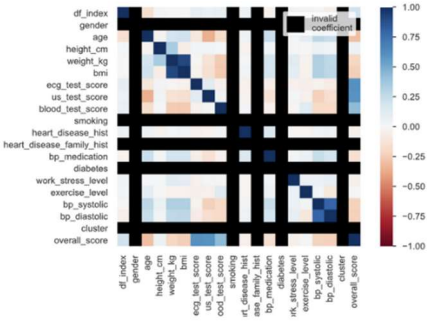
איור 8 - גרף של ציון הסילוהט כפונקציה של כמות הקלאסטרים, ניתן לראות שיש "ברך" ב-K=5.

לאחר השימוש ב-K Means הוספנו עמודה חדשה לכל Instance בה הגדרנו את השיוך שלו לצביר ספציפי. לאחר מכן פיצלנו את ה-Data לחמש תתי קבוצות זרות. לכל תת קבוצה התייחסנו כ-Data Set בלתי תלוי. ביצענו Data Exploration מחדש[3] והסתכלנו בנוסף על מקדמי המתאם החדשים בין ה-Features. להלן פירוק של כלל הצבירים שהתקבלו:

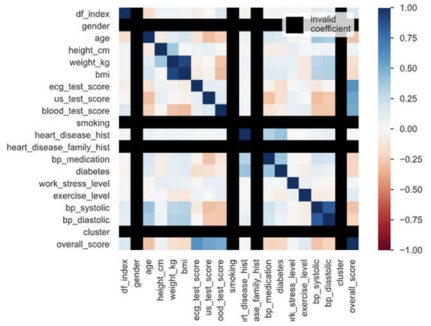
Cluster 1 - מגדר מעורב, מעשנים, ללא היסטוריה של מחלות לב (גם לא במשפחה).



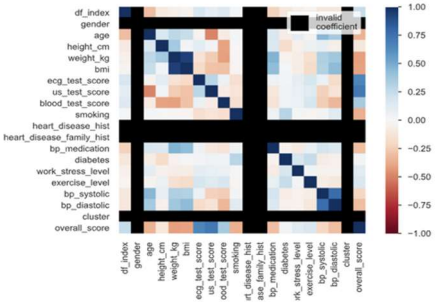
Cluster 2 - נשים, BMI נמוך יחסית, ללא היסטוריה של מחלות לב (גם במשפחה), אינם מעשנים.



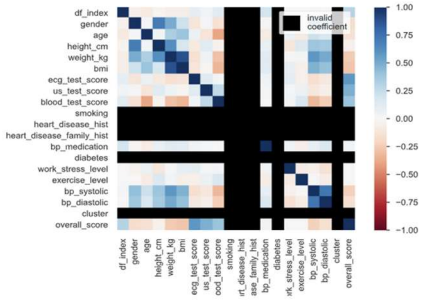
Cluster 3 - נשים, BMI נמוך יחסית, ללא היסטוריה של מחלות לב, אך לאחוז גבוה קיימת היסטוריה במשפחה, אחוז נמוך של מעשנים.



Cluster 4 - גברים, BMI נמוך יחסית, ללא היסטוריה של מחלות לב, אך לרובם היסטוריה משפחתית, אחוז נמוך של מעשנים.



Cluster 5 - גברים, BMI נמוך יחסית, ללא היסטוריה של מחלות לב (גם במשפחה), אינם מעשנים.



---

[2] קישור לסרטון שהוצג בפרזנטציה, הקוד שכתבנו נמצא ב-DS_Python והפלט של הקוד (CSV) נמצא תחת "/נספחים/Exports for Gephi".

[3] לכל קלאסטר בנינו דו"ח מלא, הפלט של הקוד (HTML) נמצא תחת "/נספחים/Data Exploration After Clusterin".

שלב 3 חלק ב – Overall Score Prediction

אחר שחילקנו את הלקוחות לצבירים, לכל צביר החלטנו להריץ מודל שמטרתו לנבא את ה- Overall Score של לקוחות חדשים שנכנסים אליו. ביצענו רגרסיה ליניארית כך שבכל פעם התעלמנו באחת מתוצאות הבדיקות (בדיקת דם, א.ק.ג. ואולטרה-סאונד). מצאנו שרמת הדיוק אינה גבוהה ביחס לכמות המידע הנתון לאימון לעומת רמת הדיוק המקורית. כדי לקבל נוחות גרפית גבוהה ופרמטרי דיוק נוספים ל-R2 פתחנו משתמש באתר Big ML וניסינו לבצע רגרסיות שונות גם בו. כדי לא לצאת ממסגרת הקורס השתמשנו אך ורק באלגוריתמים שלמדנו (לא ביצענו רגרסיה פולינומיאלית). רמות הדיוק שהתקבלו, כפי שניתן לראות מטה[4], אינן מספקות.

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

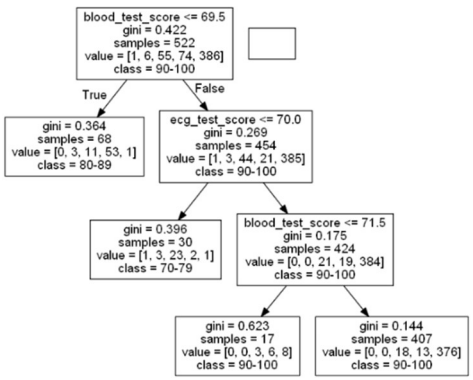
איור 9 - Linear Regression w Pythin, רמות הדיוק אינן טובות עוד לפני הפחתת הבדיקה

איור 10 - Linear Regression w Big ML - ניתן לראות שבאשר דורשים רמת דיוק סבירה מתקבל שערוך שאינו רלוונטי מבחינה עסקית

לאחר קבלת תוצאות שאינן מדויקות ברגרסיה הליניארית בדקנו את התוצאות של אלגוריתם עצי החלטה. תחילה ניסינו לחשב את הציון המדויק שקיבל כל לקוח ע"י חלוקה ל- 100 מחלקות שונות, כך שכל מחלקה מייצגת את ה-Overall Score בטווח בין 0 ל- 100. גילינו שבשיטה זו הדיוק שקיבלנו על קבוצת ה-Test אינו מספק ורחוק מרמת הדיוק שמתקבלת עם ביצוע כל הבדיקות. בחנו את האופציות לחלק את הטווח ל-5, 10, ו-20. התוצאות שהניבו רמת דיוק רלוונטית היו כשחילקנו את הטווח ל-5 ול-10. בגלל שרמת דיוק של 10 ב-Overall Score הינה רלוונטית יותר מבחינה עסקית בחרנו בחלוקה לטווח זה. צמצום הדרישה לדיוק גבוה בפרדיקציה הקטינה משמעותית את אחוז הטעויות. חילקנו את טווח הציונים ל- 10 טווחים המייצגים את מחלקות ציונים 1 עד 10, כאשר 10 היא הגבוהה ביותר. בשיטה זו הצלחנו מחד להגיע לציונים המייצגים בצורה ברורה את רמת הסיכון של לקוח לחלות במחלה מסכנת חיים ומאידך הגענו לרמת דיוק קרובה מספיק לבדיקה המקורית. בחלק מהעצים[5] עץ ההחלטה משתמש בהרבה פיצ'רים כדי להגיע לרמת הדיוק שמצורפת מטה, ובחלק מהעצים העץ שנבנה שטוח מאוד ומשתמש במעט פיצ'רים כדי להגיע למדד אי-ודאות (אנטרופיה/ ג'יני) נמוך.

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

איור 11- התוצאות של המודל (איזה בדיקה להפחית, מה הדיוק לאחר הפחתת הבדיקה שנבחרה) לאחר חלוקה של ה-Overall_Score ל-10 טווחים. לכל קלאסטר יש עץ החלטה שונה, חלק מהעצים משתמשים במעט מאוד תכונות לשערוך ה-Overall_Score



איור 12- עץ ההחלטה שמתקבל עבור Cluster 0. התוצאה לפני הורדת הבדיקה הייתה 0.9, הדיוק לאחר הפחתת בדיקת האולטרסאונד נשאר קרוב - 0.87. במילים אחרות, עבור מעשנים, ללא היסטוריה של מחלות לב (גם לא במשפחה), העץ הנ"ל מנבא את התוצאה הסופית בעזרת שימוש בתוצאת בדיקת הדם והאקג בלבד

[4] הקוד המלא מופיע ב-DS_Python, בכל איטרציה עשינו Report מלא שהגדרנו ובחנו את הדיוק באופן ידני בהתאם לפלט שיצא.
[5] עצי ההחלטה שיצאו במודל מוצגים בפירוט בפרזנטציה – קישור. הפרזנטציה מצורפת גם בקובץ PDF תחת "/נספחים".

שלב 4 – Model Evaluation

כדי לקבל הערכה איכותית על המודל מעבר למדד ה-Accuracy, בנינו Confusion Matrix לכל עץ החלטה. לאחר שחילקנו את ה-Data לתתי קבוצות זרות, נוצר מצב בו בכל תת קבוצה יש רק 4-5 מחלקות (Classes) רלוונטיות לתכונה Overall Score (במקום 10 מחלקות כנדרש), במילים אחרות, ה-Test Set לכל קלאסטר אינו גנרי מספיק. למשל, עבור Cluster 0:

```
Confusion Matrix Tree :
[[  0   1   0   0]
 [  0   1   4   7]
 [  0   0  10   6]
 [  0   0   0 102]]
```

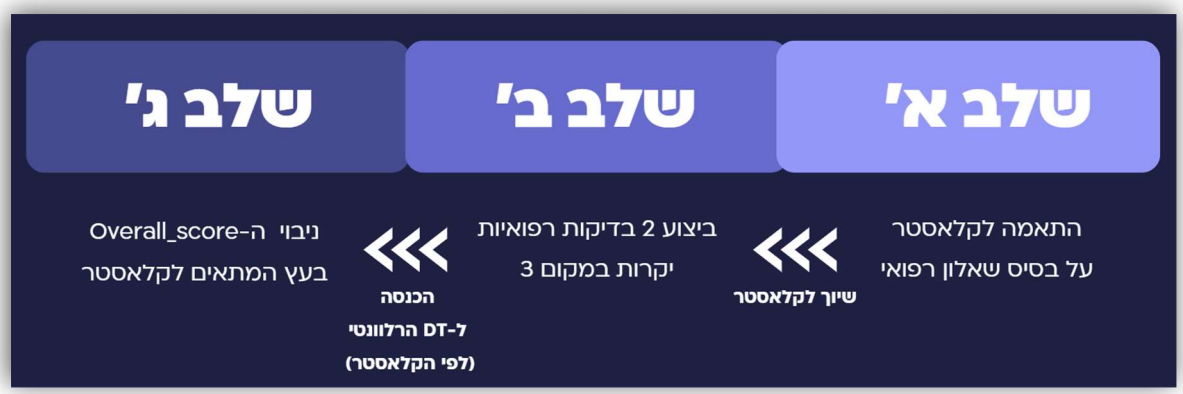
איור 13 - אפשר לראות שרוב הדיוק נמצא במחלקה "90-100" (מיקום 3,3), כלומר, בהינתן שיוך לקלאסטר 0 , ובהינתן דגימה שהקונספט שלה הוא 90-100, הסיכוי להיות צודקים בהיפותזה הוא 0.88. מאידך, אי אפשר להסיק מספיק מידע על סיווג של דגימות שאינן מ-4 המחלקות שהיו ב-Test Set. קיימות 6 מחלקות שהעץ לא סיווג.

נתון זה (כמות ה-TP) מוכיח כי שאלת המחקר שלנו **ככל הנראה** ישימה ואף תוכל לחסוך לחברת Presee משאבים רבים בטווח הארוך. עבור לקוח חדש, כל מה שעלינו לעשות הוא להתאים אותו לאחד מן הצבירים על בסיס השאלון הרפואי אותו הוא ממלא בעצמו, לאחר מכן, לבצע את הבדיקות היקרות הדרושות ללא הבדיקה שהשמטנו עבור אותו צביר ולנבא את ה- Overall Score של הלקוח. נשים לב שמספר הבדיקות שאנחנו חוסכים הוא כמספר הלקוחות החדשים של החברה (כל לקוח מבצע בדיקה אחת פחות). נשים לס\ב שהחיסכון הוא יחסי פר לקוח - ככל שיותר לקוחות חדשים מגיעים כך סכום המשאבים שאנו מסוגלים לחסוך גדל.

סיכום

בכל תתי הקבוצות הצלחנו להשמיט בדיקה שגרמה להרעה של לכל היותר 5% ב-Accuracy ביחס לניבוי המקורי, בדיוק ממוצע של 80%, לכן, עמדנו בקריטריון ההצלחה. עם זאת, חשוב לציין שבבחינה רטרואקטיבית אנחנו מזהים קלאסטרים בהם כמות הדגימות שיש ל-Test Set אינה מספיקה. הדיוק (Accuracy) מוגדר להיות כמות הפעמים שהעץ סיווג נכון. הואיל והסיווג אינו בינארי (מדובר ב-Multiclass Tree) , לא הצלחנו לבצע הערכה מדוייקת ל-Misclassification. במילים אחרות, אנחנו לא יכולים לנתח בצורה איכותית את ה-FP בהינתן כל Class. כדי להעריך את המודל בצורה יותר מדוייקת *(Kautz et al., 2017)*, באופן בו אנחנו יודעים אם קיים Bias ל-Misclassification מסוים, עלינו לקבל כמות דגימות גדולה באופן משמעותי.

תחת ההנחה שקיים סט דגימות שכזה, ניתן לסכם תהליך Onboarding למטופל חדש:



מקורות

1. Corso, A. (2021, September 27). *How Much Does an EKG Cost Without Insurance in 2021?* Mira. https://www.talktomira.com/post/how-much-does-an-ekg-cost-at-urgent-care
2. *How Much Does an Ultrasound Cost?* (2021, August 31). Tripment Health. https://tripment.com/blog/how-much-does-an-ultrasound-cost
3. Slobin, J. (2022, April 17). *How Much Does Bloodwork Cost Without Insurance in 2022?* Mira. https://www.talktomira.com/post/the-cost-of-bloodwork-without-insurance-2021
4. Schneider, A., Hommel, G., & Blettner, M. (2010). Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt international, 107(44), 776–782. https://doi.org/10.3238/arztebl.2010.0776
5. Kautz, T. K., Eskofier, B. M. E., & Pasluosta, C. F. P. (2017). *Pattern Recognition* (Generic performance measure for multiclass-classifiers ed., Vol. 68) [E-book]. Retrieved July 20, 2022, from https://www.sciencedirect.com/science/article/abs/pii/S0031320317301073