

## 1 הגדרת הבעיה והכנת הנתונים

## 1.1 – מטרת כריית המידע

מטרת כריית המידע היא לחזות את סיווג תפקוד בלוטת התריס של נבדק. ישנם שלושה סיווגים, תפקוד תקין, פעילות יתר (*hyperthyroid*), ותת פעילות (*hypothyroid*).

## 1.2 – הגדרת הנתונים

נגדיר כעת את הנתונים הגולמיים בהם נשתמש. להלן התכונות שנאספו על הנבדקים.

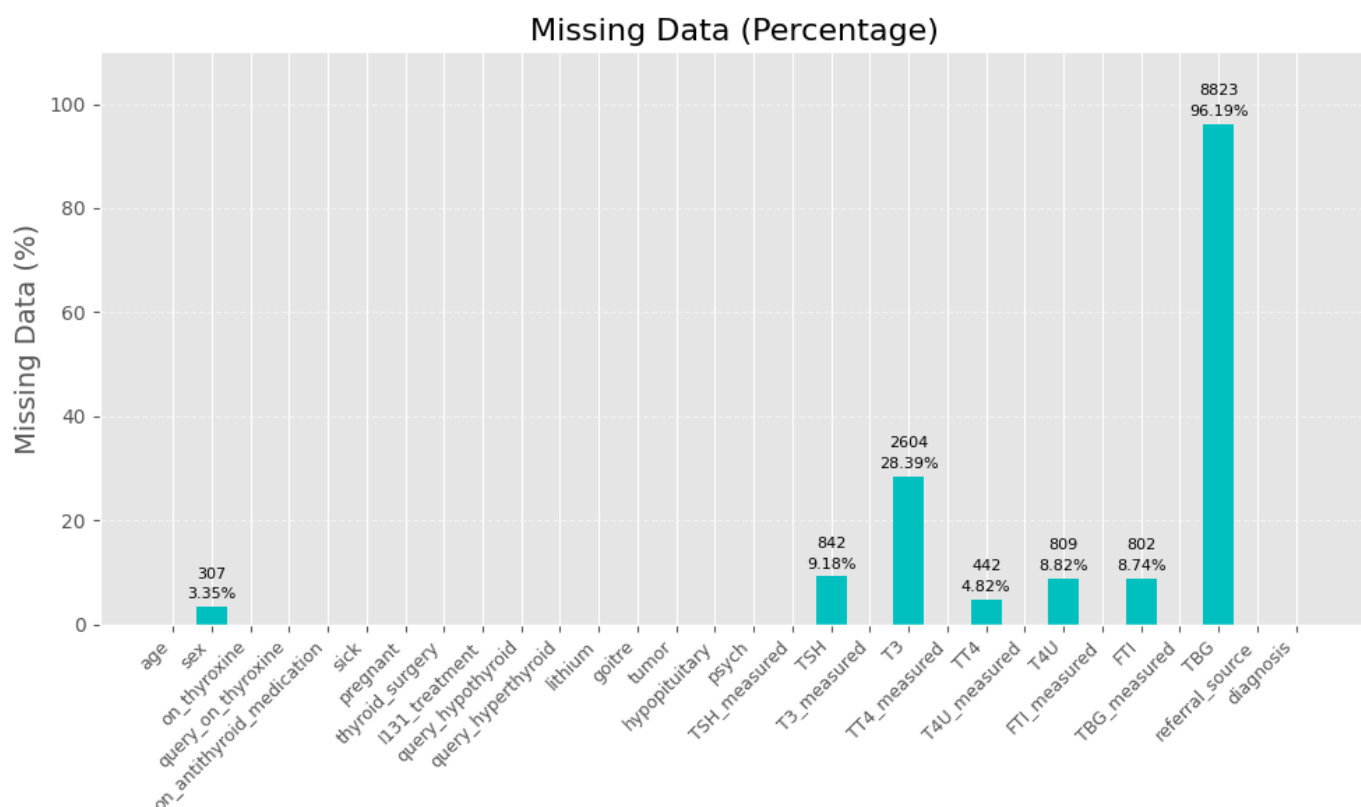
שם תכונה	תיאור תכונה	סוג הנתונים	תחומי ערכים	ממוצע	סטיית תקן	נתונים חסרים
Age	גיל הנבדק	נומרי	2-65526	73.56	1183.9	0%
Sex	מין הנבדק	קטגורי	Male – זכר Female – נקבה			3%
On Thyroxine	האם הנבדק מקבל תרופות תירוקסין	קטגורי	True, False			0%
Query on Thyroxine	האם יש צורך לברר על שימוש ב-Thyroxine	קטגורי	True, False			0%
On Antithyroid Medication	האם האדם מקבל תרופות נגד בלוטת התריס (שמורידות <i>thyroid</i> )	קטגורי	True, False			0%
Sick	האם הנבדק חולה	קטגורי	True, False			0%
Pregnant	האם הנבדקת בהריון	קטגורי	True, False			0%
Thyroid Surgery	האם הנבדק עבר ניתוח בלוטת התריס	קטגורי	True, False			0%
I131 Treatment	האם לנבדק היה טיפול I131 (נגד פעילות יתר)	קטגורי	True, False			0%
Query Hypothyroid	שאלתה לגבי מצב היפותירואיד	קטגורי	True, False			0%
Query Hyperthyroid	האם יש צורך לברר על <i>Hyperthyroid</i> (מצב יתר של בלוטת התריס)	קטגורי	True, False			0%
Lithium	האם האדם נוטל תרופות ליתיום	קטגורי	True, False			0%
Goitre	נוכחות זפק	קטגורי	True, False			0%
Tumor	נוכחות של גידול	קטגורי	True, False			0%
Hypopituitary	נוכחות של מצב <i>hypopituitary</i> (חוסר בייצור הורמונים רלוונטיים)	קטגורי				0%
Psych	מצב פסיכולוגי	קטגורי				0%
TSH	רמות הורמונים ממריצות בלוטת התריס	רציף	0.005-530	5.218	24.184	9%
T3	רמות טרייודותרונין	רציף	0.05-18	1.971	0.888	28%
TT4	שיעור ניצול תירוקסין	רציף	2-600	108.7	37.523	5%
T4U	מדד תירוקסין חינם	רציף	0.17-2.33	0.976	0.2	9%
FTI	רמות גלובולין קושר בלוטת התריס	רציף	1.4-839	113.641	41.552	9%
Referral Source	מידע על האופן שבו מטופלים הופנו לספק שירותי הבריאות	קטגורי	WEST, STMW, SVHC, SVI, SVHD, Other			

שם התכונה	תיאור הבעיה	צורת התמודדות	שיטה
Age	גילאים גבוהים ברמה מופרזת	<ul style="list-style-type: none"> <li>ערכים מעל 120 שונו לגיל החציוני</li> <li>הבחירה ב-120 היא שרירותית</li> <li>הבחירה בשינוי הערכים ולא במחיקת כל הרשומה היא מחמת שסביר שהערכים הוכנסו בצורה שגויה, ולא כל הרשומה</li> <li>בחירת החציון עדיפה לטיפול בחריגים מכיוון שהיא פחות מושפעת מערכי קיצון בהשוואה לממוצע</li> </ul>	Python
TBG	מעל 96% מהערכים חסרים	מחיקת העמודה	Python (מצ"ב גרף התפלגות הנתונים החסרים)
Sex	3% מערכים חסרים	<ul style="list-style-type: none"> <li>אם מצוין שהנבדק בהריון – נשנה את העמודה ל-f</li> <li>עבור שאר הערכים נכניס בצורה רנדומלית ערכים (F או M) לפי יחס הגברים והנשים הכללי</li> </ul>	Python (מצ"ב תמונת ההדפסות לפני ואחרי השינויים)
עמודות מסוג 'measured'	דאטה מיותר	מחיקת העמודה	Python (מצ"ב תמונת ההדפסות לפני ואחרי השינויים)
Referral Source	מידע לא רלוונטי לגבי חיזוי רשומות חדשות	מחיקת העמודה	Python
נבדקים שאובחנו כ-R	נבדקים שהבדיקות שלהם יצאו סותרות	מחיקה	Python (הפונקציה הרלוונטית היא <code>remove_specific_diagnoses</code> בקובץ <code>data_preparation.py</code> )
נבדקים שאובחנו כ-S	נבדקים שנמדדו להם מדדי ה-TBG בלבד	כיוון שמחקנו את העמודה ה-TBG נשמיט נבדקים אלו (מחיקה)	Python (הפונקציה הרלוונטית היא <code>remove_specific_diagnoses</code> בקובץ <code>data_preparation.py</code> )

תכונה	Bins	תוויות	נימוק
Age	טווחי גילאים בהפרשים של 10, עד גיל 80	1 – 9, ..., 70 – 79, 80 +	הגיל משפיע באופן משמעותי על תפקוד בלוטת התריס; דיסקרטיזציה כזו מאפשרת ניתוח לפי קבוצת גיל, ומקלה על זיהוי מגמות הקשורות לגיל. קטגוריית '+80' מתייחסת במיוחד לאנשים מבוגרים עם מאפיינים בריאותיים מובהקים.
TSH	<ul style="list-style-type: none"> <li>Low: &lt;0.4 (hyperthyroidism)</li> <li>Normal: 0.4-4.0</li> <li>High: &gt;4.0 (hypothyroidism)</li> </ul>	Low, Normal, High	<p>משקף סף קליני:</p> <ul style="list-style-type: none"> <li>נמוך (&lt; 0.4) מצביע על תת פעילות של בלוטת התריס</li> <li>נורמלי (0.4 – 4.0) מצביע על תפקוד בריא של בלוטת התריס</li> <li>גבוה (&gt; 4.0) מצביע על תת פעילות של בלוטת התריס.</li> </ul> <p>דיסקרטיזציה כזו מפשטת הערכה קלינית ומתואמת עם סטנדרטים אבחוניים נפוצים.</p>
T3	<ul style="list-style-type: none"> <li>Low: &lt;0.8 (hypothyroidism)</li> <li>Normal: 0.8-2.0</li> <li>High: &gt;2.0 (hyperthyroidism)</li> </ul>	Low, Normal, High	T3 חיוני לוויסות חילוף החומרים. רמות מחוץ לטווח של 0.8 עד 2.0 ננוגרם/מ"ל יכולות להצביע על תפקוד לקוי של בלוטת התריס, כאשר רמות גבוהות קשורות לעיתים קרובות להיפר-תירואידיזם.
TT4	<ul style="list-style-type: none"> <li>Low: &lt;50 (hypothyroidism)</li> <li>Normal: 50-120</li> <li>High: &gt;120 (hyperthyroidism)</li> </ul>	Low, Normal, High	T4 נותן מדד מקיף של תירוקסין בדם. הטווח התקין הוא בדרך כלל בין 50 ל-120 מק"ג/ד"ל, עם סטיות המעידות על בעיות אפשריות בבלוטת התריס.
T4U	<ul style="list-style-type: none"> <li>Low: &lt;0.7 (ספיגה מופחתת)</li> <li>Normal: 0.7-1.3</li> <li>High: &gt;1.3 (ספיגה גבוהה)</li> </ul>	Low, Normal, High	בדיקות ספיגת T4 (T4U) עוזרות להעריך חלבונים קושרים לבלוטת התריס. הטווח שבין 0.7 ל-1.3 נחשב נורמלי; ערכים מחוץ לטווח זה עשויים לרמז על חריגות בקשירת חלבון או ברמות ההורמונים.
FTI	<ul style="list-style-type: none"> <li>Low: &lt;70 (hypothyroidism)</li> <li>Normal: 70-130</li> <li>High: &gt;130 (hyperthyroidism)</li> </ul>	Low, Normal, High	FTI מעריכה את פעילות בלוטת התריס על ידי תיקון TT4 עבור T4U. הטווח התקין הסטנדרטי הוא בין 70 ל-130, המספק מדד מותאם יותר של זמינות התירוקסין.

סיווג מצב בלוטת התריס (Thyroid Condition Classification)

סימון	מצב רפואי	סיווג	הסבר
-	No condition	בריא	מציין שאין בעיה בבלוטת התריס או מצב המחייב הערה.
A	Hyperthyroid	פעילות יתר של בלוטת התריס (hyperthyroid)	אינדיקציה ישירה למצב יתר של בלוטת התריס.
B	T3 toxic	פעילות יתר של בלוטת התריס (hyperthyroid)	סוג ספציפי של פעילות יתר של בלוטת התריס שבו רמות T3 מוגברות.
C	Toxic goitre	פעילות יתר של בלוטת התריס (hyperthyroid)	זפק הקשור להיפרתירואידיזם.
D	Secondary toxic	פעילות יתר של בלוטת התריס (hyperthyroid)	פעילות יתר של בלוטת התריס עקב סיבות משניות.
E	Hypothyroid	תת פעילות של בלוטת התריס (hypothyroid)	אינדיקציה ישירה למצב של תת פעילות בלוטת התריס.
F	Primary hypothyroid	תת פעילות של בלוטת התריס (hypothyroid)	מחסור ראשוני בהורמון בלוטת התריס, המעיד על תת פעילות של בלוטת התריס.
G	Compensated hypothyroid	תת פעילות של בלוטת התריס (hypothyroid)	תת פעילות של בלוטת התריס בשלב מוקדם שבו הגוף פיצה.
H	Secondary hypothyroid	תת פעילות של בלוטת התריס (hypothyroid)	תת פעילות של בלוטת התריס עקב סיבה משנית (לא בלוטת התריס).
I	Increased binding protein	בריא	המאפיין קשור לרמות חלבון, לא ישירות למצב של בלוטת התריס.
J	Decreased binding protein	בריא	באופן דומה, המאפיין מתייחס לרמות החלבון, לא לבריאות בלוטת התריס ישירות.
K	Concurrent non-thyroidal illness	בריא	מציין שקיימת מחלה שאינה בלוטת התריס.
L	Consistent with replacement therapy	בריא	מציין שהאדם נמצא בטיפול חלופי, לא מטבעו היפו/היפר-תירואיד.
M	Underreplaced	תת פעילות של בלוטת התריס (hypothyroid)	מצביע על תחלופה לא מספקת של הורמונים, ככל הנראה מוביל לתסמינים של תת פעילות בלוטת התריס.
N	Overreplaced	פעילות יתר של בלוטת התריס (hyperthyroid)	מצביע על החלפת הורמונים מוגזמת, שכנראה מובילה לתסמיני יתר של בלוטת התריס.
O	Antithyroid drugs	פעילות יתר של בלוטת התריס (hyperthyroid)	טיפול בפעילות יתר של בלוטת התריס, מרמז על מצב של יתר בלוטת התריס המטופל.
P	I131 treatment	פעילות יתר של בלוטת התריס (hyperthyroid)	טיפול ביוז רדיואקטיבי משמש בדרך כלל ליפרתירואידיזם.
Q	Surgery	פעילות יתר של בלוטת התריס (hyperthyroid)	ניתוח בלוטת התריס מתייחס לרוב לטיפול בהיפרת התריס.
T	Elevated thyroid hormones	פעילות יתר של בלוטת התריס (hyperthyroid)	מצביע ישירות על רמות הורמונים גבוהות, בדרך כלל היפרתירואידיזם.



תוצאות הפעלת הפונקציות 'standardize\_sex\_column', 'preprocessed\_data' על מסד הנתונים:

Data loaded successfully. Number of rows removed: 6 Rows containing 'measured' data have been removed.	Total values: 9172 Missing data before: 0 (0.00%) Data changed: 0 (0.00%) Missing data after: 0 (0.00%)	Total values: 9172 Missing data before: 307 (3.35%) Data changed: 4 (0.04%) Missing data after: 303 (3.30%)
--	--	--

Figure 1 - תוצאות שינוי עמודות 'sex' ביחס לעמודות 'pregnant'

Figures 2-3 - תוצאות ניקוי והכנת הדאטה ע"פ המפורט בטבלה לעיל

## 1.3 – הגדרת ותיאור שלבי ה-KDD

### שלבי ה-KDD

שלב ראשון – הגדרת יעדי כריית נתונים

אנו שואפים לפתח מודל שיכול לחזות אם לבדק יש הפרעה בבלוטת התריס, כגון תת פעילות של בלוטת התריס, בהתבסס על נתוני בדיקות רפואיות.

שלב שני – בחירת הנתונים (Data Selection)

בשלב זה, המטרה היא לבחור את תת-קבוצת הנתונים הרלוונטיים שיהיו שימושיים לחזות הפרעות בבלוטת התריס בהתבסס על תוצאות בדיקות רפואיות. קיימים 6 בסיסי נתונים שנתרמו על ידי המוסד Garavan, סידיני אוסטרליה. במקרה זה נפעל ע"פ המלצתה של ד"ר מיה הרמן ונבחר בבסיס הנתונים 'thyroid0387' המכיל 9172 רשומות וכ-20 תכונות.

נתחיל בבחינת הקבצים 'thyroid0387-data.csv' ו-'thyroid0387-names.csv' כדי להבין את התכונות, סוגי הנתונים, והכנת הנתונים הנדרשים כגון טיפול בערכים חסרים או חוסר איזון נתונים. את הטיפול בנתונים נבצע באמצעות Python (קוד מצורף לפרויקט).

נתאר את שלבי עיבוד והכנת הנתונים באמצעות Python.

- נרצה לקבל תפיסה ויזואלית של טיב וצורת הנתונים (ובשלב זה בהתעלם ממשמעותם), לכן נשרטט את גרף התפלגות הערכים החסרים, וסטטיסטיקות כלליות כדי לראות את הנתונים בצורה ויזואלית.
- נרצה להבין את התפלגות מין וגיל הנבדקים בכמה אופנים, לכן ננתח את הדאטה בהתאם ונשרטט את הגרפים הרלוונטיים.
- בשלב זה נשים לב כי יש כמה נתונים בעייתיים. פירוט הנתונים הבעייתיים וצורת ההתמודדות עמם מפורטת בטבלה לעיל.

### שלב רביעי – עיבוד מקדים (Preprocessing)

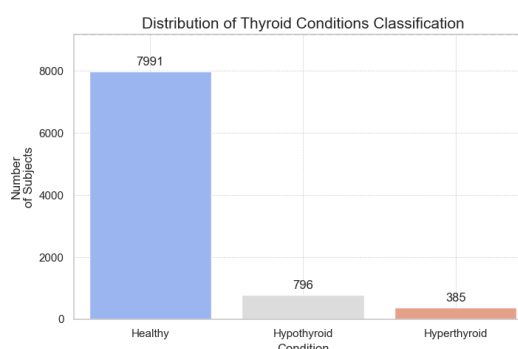
שלב זה כולל ניקוי הנתונים והכנתם לתהליך הכרייה. נצטרך לוודא שהנתונים נקיים מחוסר עקביות, ערכים חסרים וחריגים שעלולים להטות את התוצאות. נשים לב כי יש נתונים בעייתיים, לא סבירים, לא רלוונטיים ומיותרים.

פעולות שבוצעו:

- טיפול בערכים חסרים: השלמה ומחיקה.
- המרת נתונים: המרנו סוגי נתונים כראוי, למשל, המרת רמות הורמונים ממחרוזות לערכים מספריים.
- הסרת/תיקון ערכים חריגים: הערכה וניהול ערכים חריגים ולא סבירים.
- הסרת ערכים מיותרים.

פירוט הנתונים הבעייתיים וצורת ההתמודדות עמם מפורטת בטבלה לעיל.

התוצאות לאחר הסיווג הן:



### שלב חמישי – טרנספורמציה של הנתונים (Transformation)

שלב שינוי נתונים כרוך בשינוי או יצירת תכונות חדשות שהופכות את תהליך הכרייה ליעיל יותר. נרצה לשפר את כוח הניבוי של האלגוריתם על ידי תכונות הלוכדות היבטים חשובים של הנתונים. המרנו משתנים רציפים כמו גיל,  $TSH$ ,  $T3$ ,  $FTI$ ,  $T4U$  לקטגוריות כמפורט בטבלה. חלוקת ערכי ה- $bins$  הספציפיים לכל ערך נקבעה על פי מחקר אישי שחלק משיקוליו מפורטים בטבלה.

את סיווג התוצאות ביצענו על פי טבלת סיווג הנתונים דלעיל (Thyroid Condition Classification).

## שלב שישי – כריית המידע (Data Mining)

- נשתמש באקסל וב-Python וב-Weka כדי לעבד את המידע ולהריץ אלגוריתמים שונים.
- חלק משיטות כריית המידע מצריכות אילוצים ומניפולציות על אופן הצגת ואכסון הנתונים – נבצע טרנספורמציות ומניפולציות מתאימות כדי להתאים את הנתונים לאלגוריתמים השונים.
- נתבונן בכמה משיטות כריית המידע באמצעות Weka כדי למקסם את תהליך ותוצאות כריית המידע.
- נבצע את שיטות כריית המידע שנבחרו.

## שלב שביעי – ניתוח התוצאות (Result Analysis)

- ניתוח נתונים סטטיסטיים של תוצאות הפעלת האלגוריתמים על הנתונים.
- ביצוע הערכה לפי מדדים כמו מידת דיוק, רלוונטיות, פשטות, וכו'.
- לאחר ניתוח התוצאות קיימות שני אפשרויות:
  - התוצאות אינן משביעות רצון – נחזור על התהליך לעיל עם שינויים בחלקים אחרים בתהליך (עיבוד נתונים, שינוי הפרמטרים בדיסקרטיזציה, בחירת אלגוריתם אחר), ובכך ננסה להגיע למודל חיזוי טוב יותר.
  - התוצאות משביעות רצון – ניתן להתקדם לשלב של הסקת המסקנות.

## שלב שמיני – הסקת מסקנות (Transformation)

בסוף התהליך נקבל מודל על פי האלגוריתם שהרצנו. באמצעות מודל זה נוכל לחזות האם הנבדק סובל מבעיה בבלוטת התריס כדוגמת פעילות יתר של בלוטת התריס. ניתן להציג את המודל בצורה ויזואלית.

## 1.4 – סקירה השוואתית של 4 חלופות אפשרויות לביצוע כריית מידע

נשווה ארבעה אלגוריתמים שונים של כריית נתונים: *ID3 Decision Tree*, רגרסיה לינארית, *CART Decision Tree* ו-*C4.5 Decision Tree*. ההשוואה תתמקד בהתאמתם למערך הנתונים ולבעיה שעל הפרק, בהתחשב ביתרונות ובחסרונות שלהם.

### עץ החלטות *ID3* (Information Gain)

*ID3* הוא אלגוריתם לבניית עץ החלטה. הוא בונה עץ החלטות על ידי בחירת התכונה שמניבה את ה-*Information Gain* הגבוה ביותר, המודד עד כמה תכונה יכולה להפריד בין המחלקות. נקודות ההחלטה מתקבלות באמצעות חיפוש מלמעלה למטה, בצורה חמדנית דרך מערכי הנתונים המסופקים. תכונת הפיצול נקבעת על פי מדד ה-*Information Gain* הגבוה ביותר. ערך זה הוא השיפור באנטרופיה, כלומר, כמה אי הוודאות ירדה אם התפצלנו על פי תכונה זו.

#### יתרונות:

פשטות ושקוף: *ID3* הוא פשוט וקל להבנה, מה שהופך את החלטות המודל לשקופות. מטפל בנתונים קטגוריים היטב: מכיוון ש-*ID3* משתמש ב-*Information Gain* כדי לבצע פיצולים, הוא מתאים היטב לנתונים קטגוריים.

טוב לניתוח חקרני: מספק תובנה ברורה לגבי התכונות שהכי אינפורמטיביות לסיווג.

#### חסרונות:

נוטה להתאמת יתר: במיוחד עם נתונים רועשים, האלגוריתם יכול להתאים יתר על המידה לנתוני האימון אם לא מבצעים אותו בצורה מתאימה.

אינו מטפל בערכים חסרים: מצריך עיבוד מקדים לניהול כל הנתונים החסרים.

לא מטפל בנתונים מספריים באופן ישיר: נתונים מספריים צריכים לעבור דיסקרטיזציה, מה שעלול לגרום לאובדן מידע.

## רגרסיה לינארית (Linear Regression)

רגרסיה לינארית היא גישה סטטיסטית למידול הקשר בין משתנה תלוי למשתנה בלתי תלוי אחד או יותר על ידי התאמת משוואה לינארית לנתונים שנצפו. המקדמים של המשוואה הלינארית נגזרים על סמך מזעור סכום ההפרשים בריבוע בין הערכים הנצפים והחזויים.

### יתרונות:

יעילות: יעילה מבחינה חישובית, מה שהופך אותו לכדאי עבור מערכי נתונים גדולים.

פרשנות: קל לפרש את ההשפעה של כל משתנה.

טיפול רציף בנתונים: מטפל ישירות במשתנים רציפים ללא צורך בקטגוריאציה.

### חסרונות:

הנחה לינארית: מניחה קשר ליניארי, שאולי לא תמיד מתקיים, במיוחד במערכים רפואיים מורכבים.

מגבלות לא ליניאריות: לא יעיל עבור מודל קשרים לא ליניאריים.

בעיות תוצאה קטגוריות: לא מתאים למשתני יעד בינאריים או קטגוריים.

## עץ החלטות CART - Gini Index

עץ סיווג ורגרסיה CART הוא עץ החלטות רב-תכליתי שניתן להשתמש בו הן לסיווג והן לרגרסיה. הוא משתמש בפיצולים בינאריים כדי לפצל את הנתונים. ההחלטה היכן לפצל מבוססת על מדד ג'יני, שמטרתו למקסם את ההומוגניות של הצמתים. מדד זה בוחן את טוהר המידע, כלומר, נקבל ציון יותר "טוב" (נמוך) ככל שבחלוקה שבחרנו יש אחוז גבוה יותר מקטגוריה אחת, ומגיע ל-0 כאשר כל הערכים הם בדיוק מקטגוריה אחת.

### יתרונות:

פיצולים בינאריים: יעיל בטיפול בסוגים מגוונים של פיצולים.

גמישות נתונים: מעבד ביעילות הן נתונים מספריים והן נתונים קטגוריים.

חוסן חריג: פחות רגיש לחריגים בהשוואה לשיטות הנשענות על אנטרופיה.

### חסרונות:

מודלים מורכבים: יכולים להיווצר מודלים מורכבים ועמוקים, מה שמוביל להתאמת יתר.

סכנת התאמת יתר: דורש כוונן וגזוזם קפדניים.

אי יציבות מודל: רגיש לשינויים קטנים בנתונים, המובילים לתוצאות שונות.

## עץ החלטות C4.5 (Information Gain)

C4.5 הוא הרחבה של האלגוריתם ID3 הבונה עצי החלטה תוך שימוש במושג  $Gain Ratio$ , המנרמל את רווח המידע באמצעות המידע הפנימי של פיצול. שיטה זו נועדה לתת מענה להטיה כלפי תכונות מרובות רמות שקיימות ב-ID3.

### יתרונות:

הפחתת הטיה: מתגבר על הטיה כלפי תכונות בעלות רמות רבות.

טיפול ישיר בנתונים: מסוגל לעבד נתונים רציפים ודיסקרטיים כאחד.

מנגנון גזוזם: משלב מנגנונים לגזוזם העץ, משפר את ההכללה.

### חסרונות:

אינטנסיביות חישובית: מורכבת יותר לחישוב, במיוחד עבור מערכי נתונים גדולים.

מורכבות גזוזם: דורש גזוזם קפדני כדי לאזן את דיוק ומורכבות המודל.

רגישות לרעש: למרות שיפור, עדיין רגיש לנתונים רועשים.

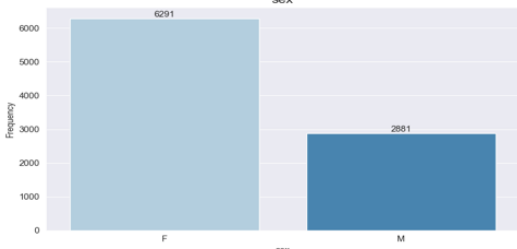
חוסר איזון: חלוקה פחות מאוזנת עם הרבה ענפים קטנים ביחס לעץ.

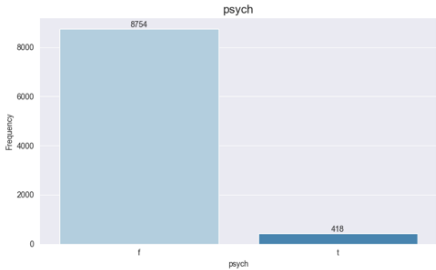
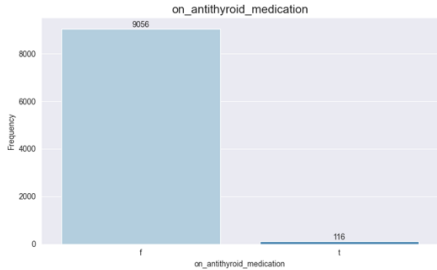
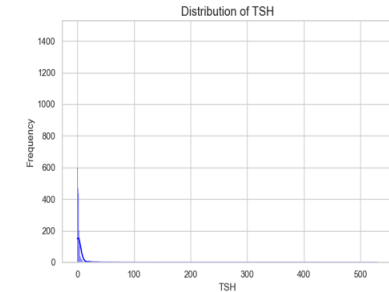
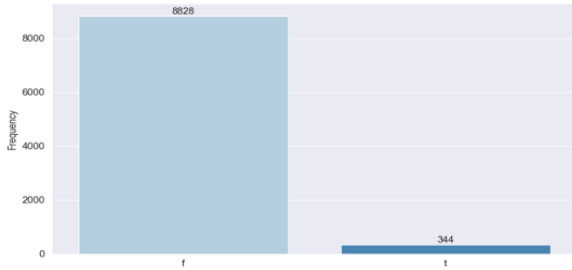
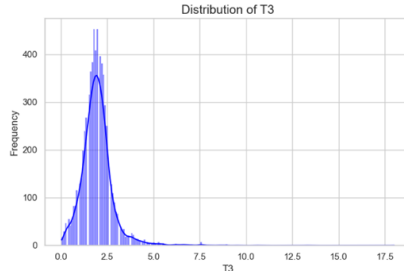
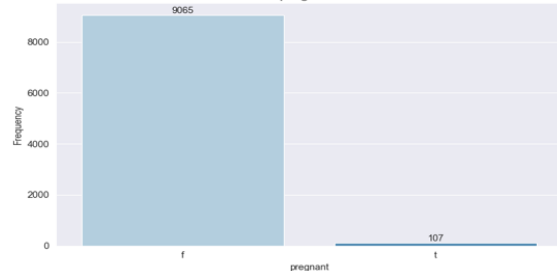
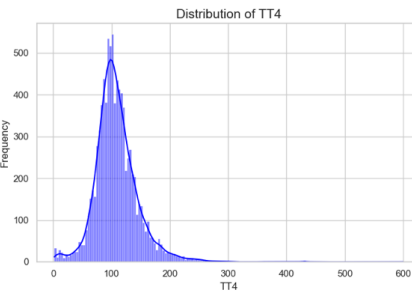
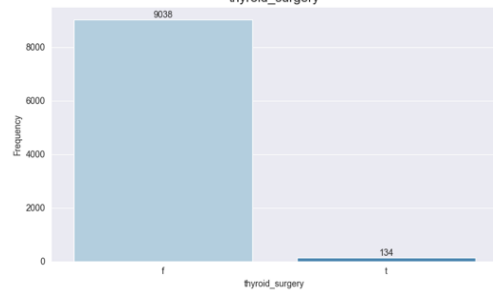
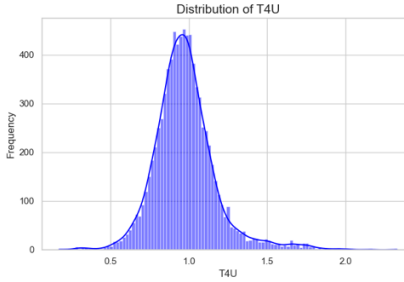
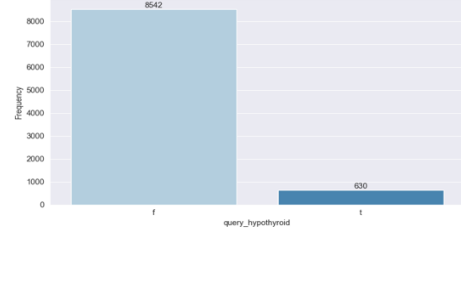


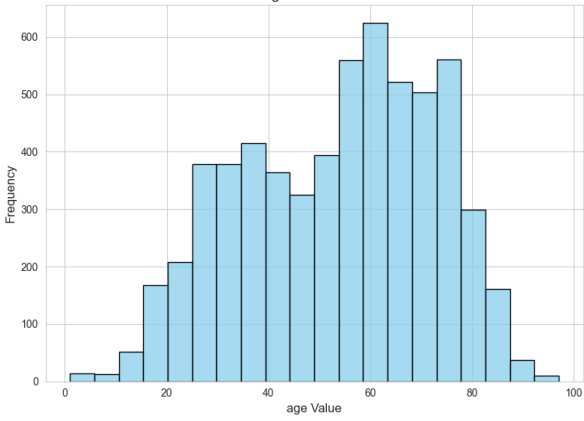
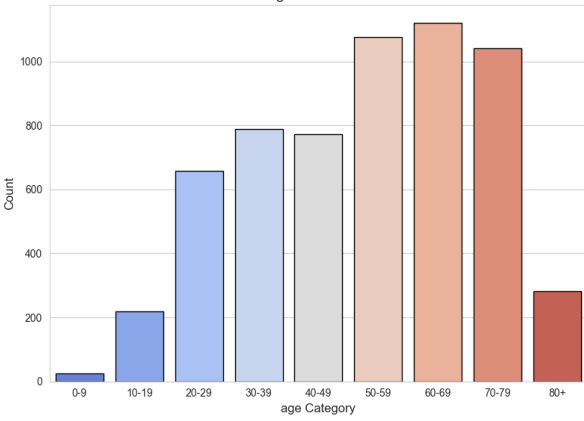
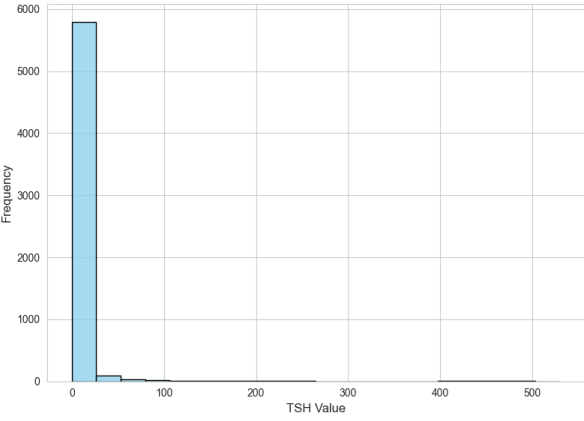
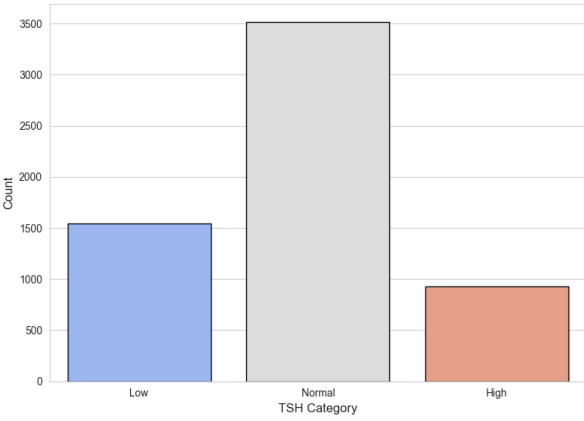
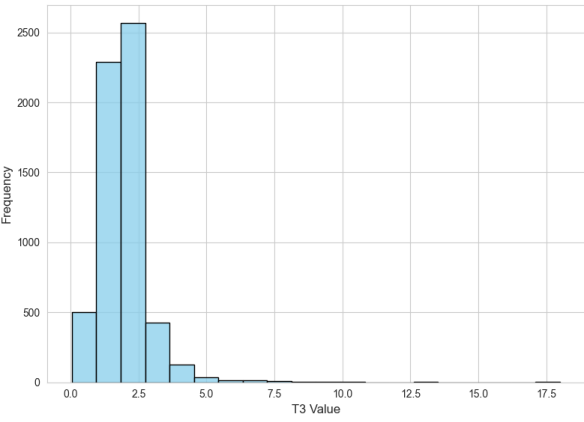
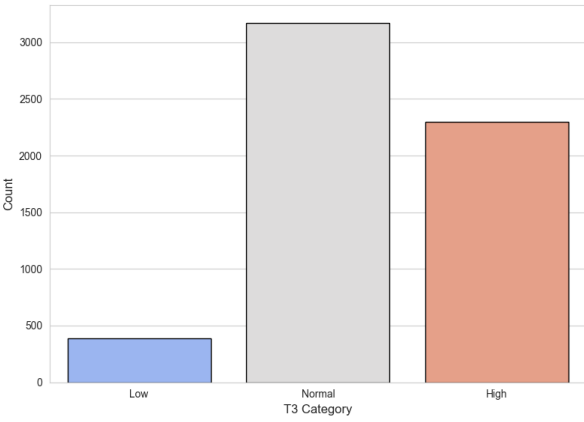
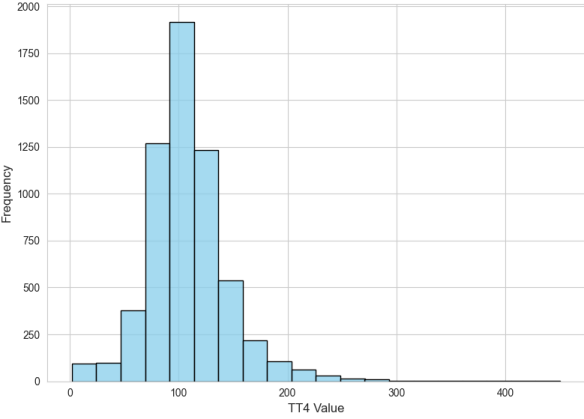
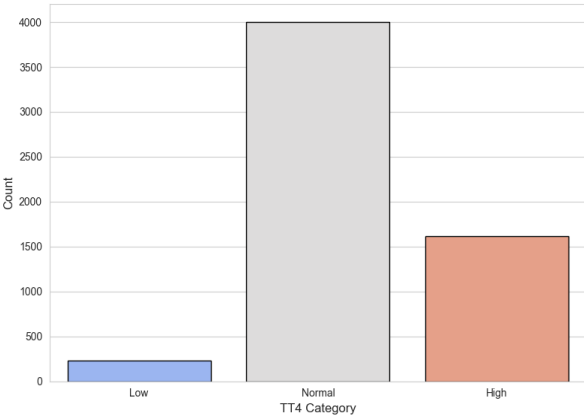
## 1.5 – תיאור הכנת הנתונים והצגתם הגרפית

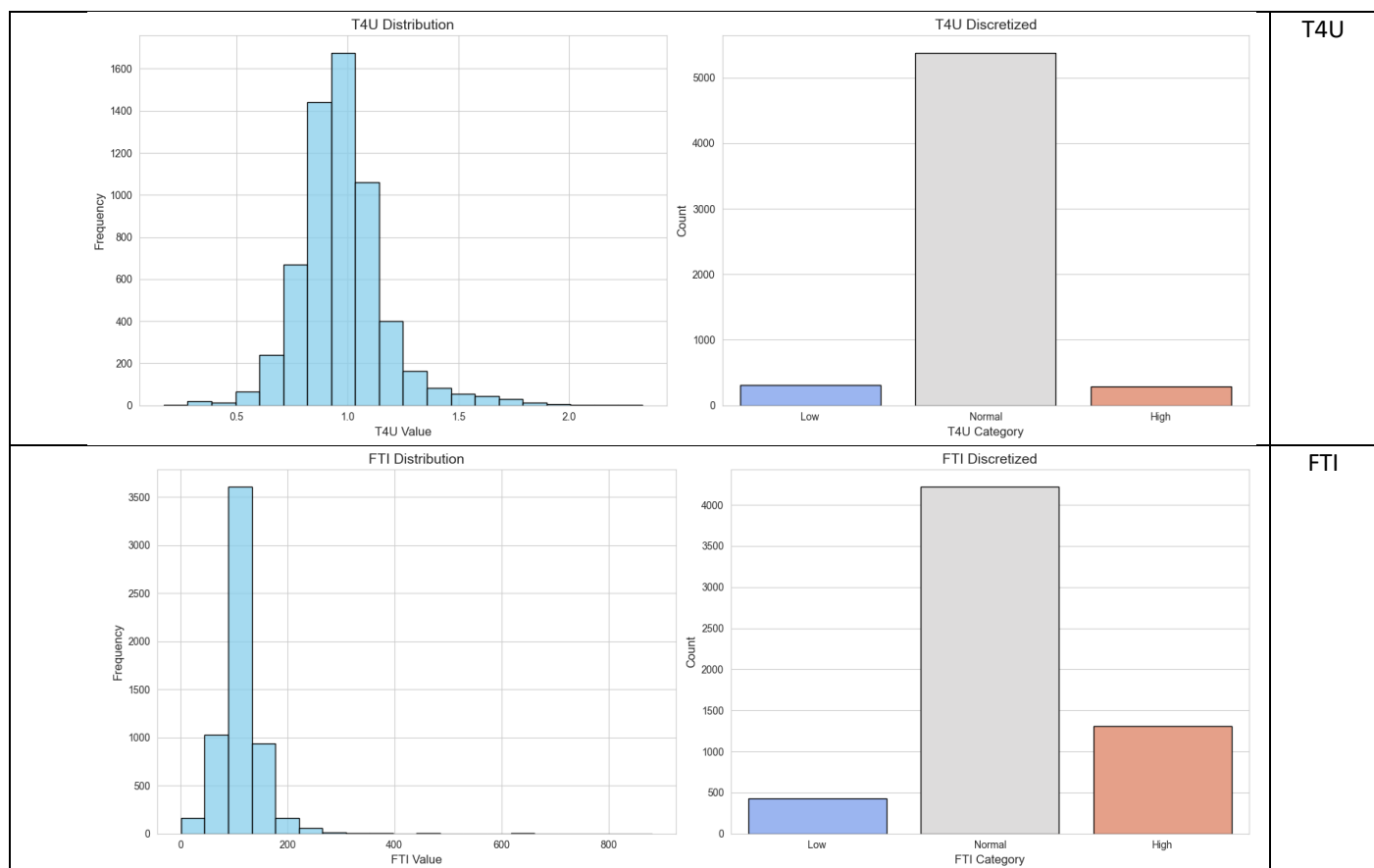
אופן הכנת הנתונים מפורט לעיל. בנוסף, כדי להשתמש בספריית pandas נאלצנו להמיר ערכים חסרים שמסומנים באמצעות '?' ל-*NaN*. הקוד שנועד להציג את הערכים בצורה ויזואלית נמצא בקובץ *'data\_visualization.py'*.

### הצגה גרפית של הנתונים

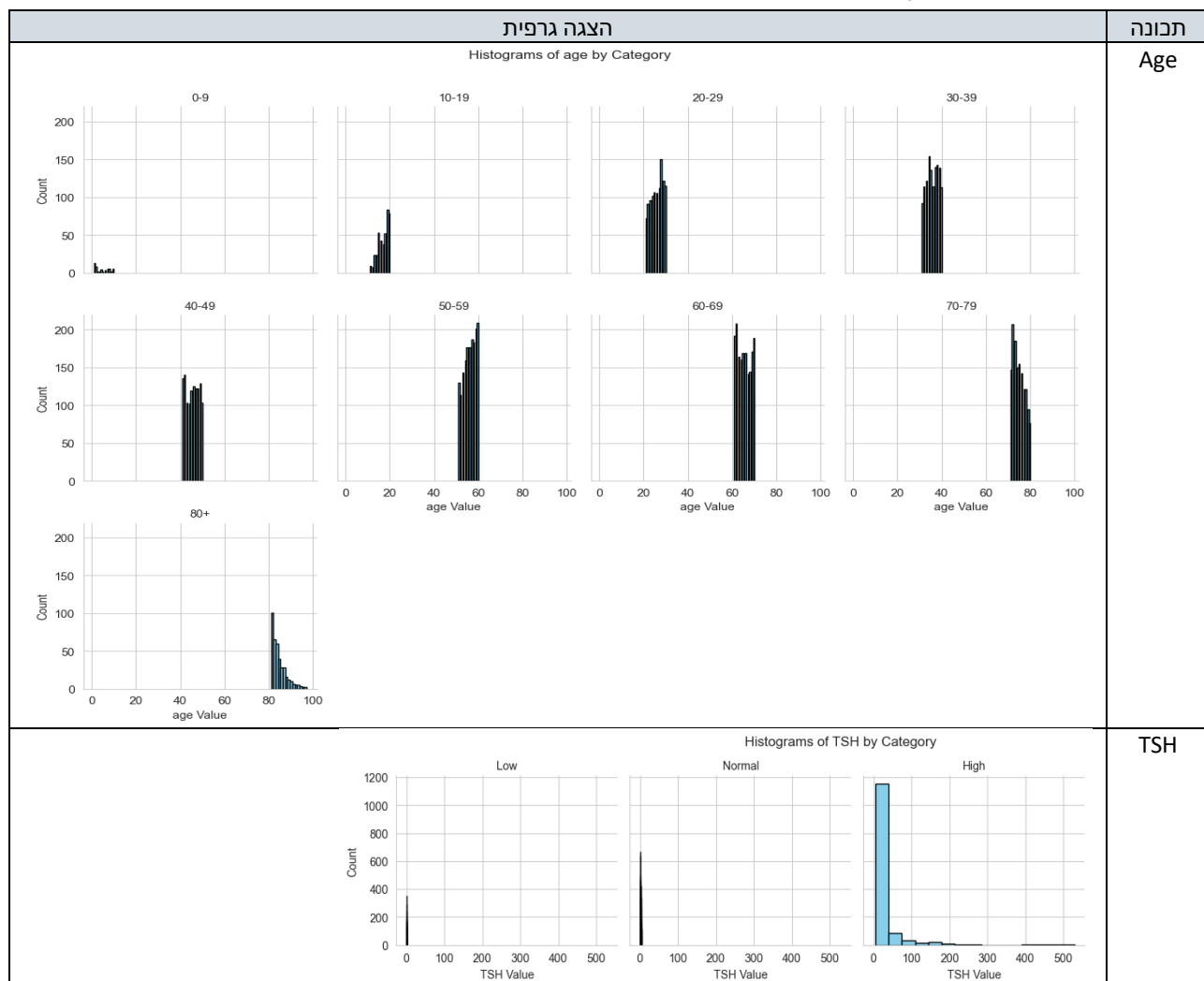
שם תכונה	הצגה גרפית	שם תכונה	הצגה גרפית
Age		Lithium	
Age by Gender			
Sex		Goitre	
On Thyroxine		Tumor	
Query on Thyroxine		Hypopituitary	

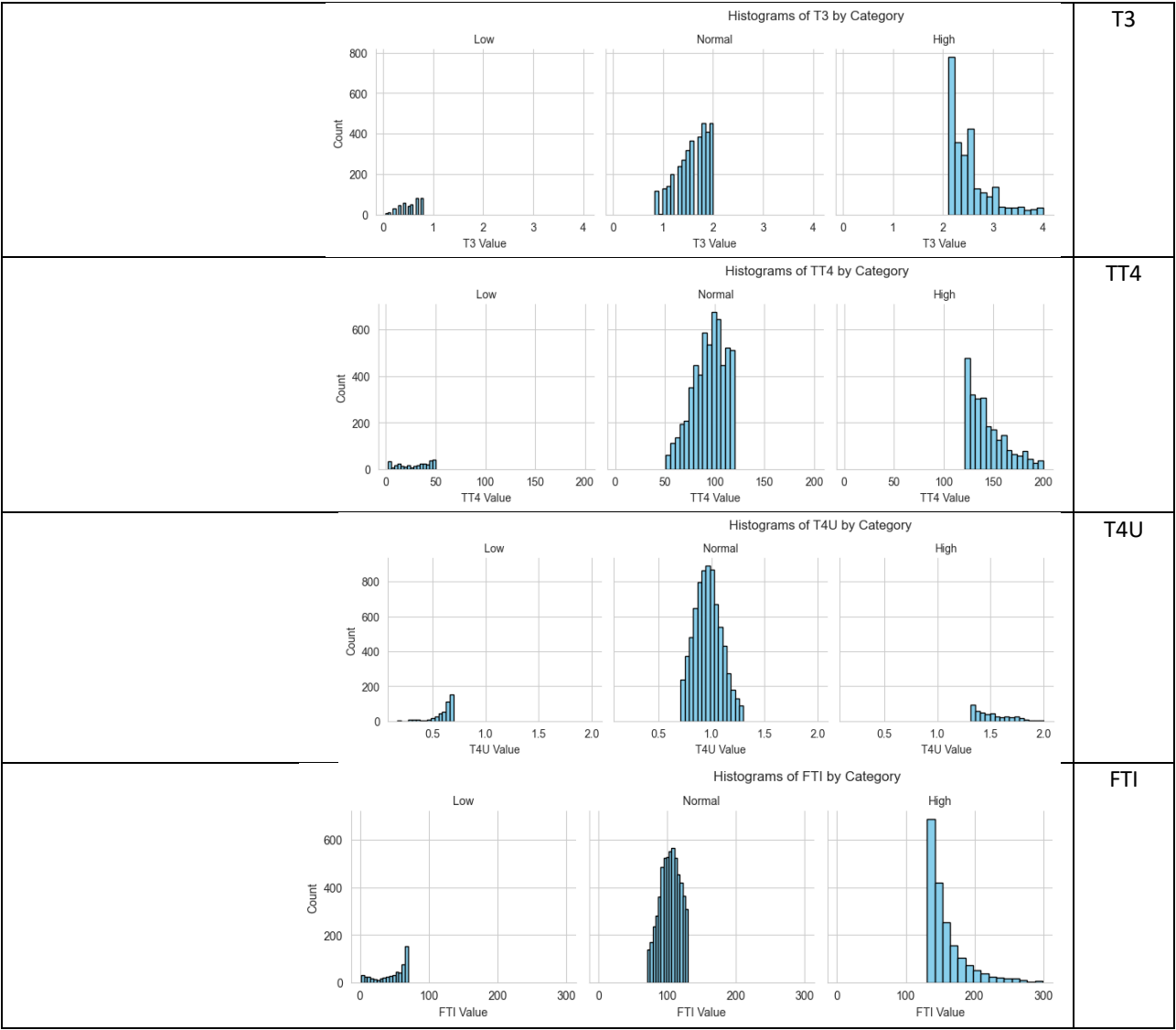
 <p>psych</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>8754</td></tr><tr><td>t</td><td>418</td></tr></table>	Category	Frequency	f	8754	t	418	<i>Psych</i>	 <p>on_antithyroid_medication</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>9056</td></tr><tr><td>t</td><td>116</td></tr></table>	Category	Frequency	f	9056	t	116	<i>On Antithyroid Medication</i>
Category	Frequency														
f	8754														
t	418														
Category	Frequency														
f	9056														
t	116														
 <p>Distribution of TSH</p>	<i>TSH</i>	 <p>sick</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>8828</td></tr><tr><td>t</td><td>344</td></tr></table>	Category	Frequency	f	8828	t	344	<i>Sick</i>						
Category	Frequency														
f	8828														
t	344														
 <p>Distribution of T3</p>	<i>T3</i>	 <p>pregnant</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>9065</td></tr><tr><td>t</td><td>107</td></tr></table>	Category	Frequency	f	9065	t	107	<i>Pregnant</i>						
Category	Frequency														
f	9065														
t	107														
 <p>Distribution of TT4</p>	<i>TT4</i>	 <p>thyroid_surgery</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>9038</td></tr><tr><td>t</td><td>134</td></tr></table>	Category	Frequency	f	9038	t	134	<i>Thyroid Surgery</i>						
Category	Frequency														
f	9038														
t	134														
 <p>Distribution of T4U</p>	<i>T4U</i>	 <p>query_hypothyroid</p> <table><tr><th>Category</th><th>Frequency</th></tr><tr><td>f</td><td>8542</td></tr><tr><td>t</td><td>630</td></tr></table>	Category	Frequency	f	8542	t	630	<i>Query Hypothyroid</i>						
Category	Frequency														
f	8542														
t	630														

הצגה גרפית		תכונה
<p>age Distribution</p> 	<p>age Discretized</p> 	Age
<p>TSH Distribution</p> 	<p>TSH Discretized</p> 	TSH
<p>T3 Distribution</p> 	<p>T3 Discretized</p> 	T3
<p>TT4 Distribution</p> 	<p>TT4 Discretized</p> 	TT4



## דיסקרטיזציה – פילוח פנימי לפי קטגוריה





## 2.1 – שיטות סיווג הנתונים

### השיטות הנבחרות

שיטות הסיווג שנבחנו הן:

- CART
- C4

כריית המידע הזו דורשת טיפול בסיווגים קטגוריים לכן רגרסיה לינארית לא מתאימה. נעדיף את C4.5 על פני ID3 כיוון שC4.5 הוא הרחבה של ID3.

### CART (Classification and Regression Trees)

CART בונה עץ החלטות בינארי המפצל את מערך הנתונים לתת-קבוצות קטנות יותר על סמך התכונה המספקת את הגידול המשמעותי ביותר בהומוגניות לגבי משתנה היעד. הוא מעריך פיצולים פוטנציאליים באמצעות אינדקס ג'יני, מדד שבודק את טוהר הנתונים (*impurity or purity*); המטרה היא למקסם את ההומוגניות (או למזער את ה-*impurity*) בתוך כל ענף לאחר הפיצול.

הצדקת הבחירה: עץ בינארי של CART נבחר בשל הפשטות והיעילות שלו, בנוסף הוא קל לפירוש, מה שהופך אותו לאידיאלי עבור בעיות שבהן הבנת היגיון ההחלטה של המודל חיונית. בנוסף, היכולת שלו להתמודד עם מערכי נתונים גדולים במהירות והחוסן שלו לחריגים הופכים אותו לכלי יעיל במקרים אלו.

### C4.5

C4.5 מתבסס על העקרונות של אלגוריתם ID3 על ידי שימוש באנטרופיית *Information Gain* כקריטריון לבחירת התכונה המפרידה בצורה הטובה ביותר בין המחלקות במערך נתונים נתון. זה יוצר עץ רב-כיווני, מוצא את התכונה הקטגורית הטובה ביותר לפצל את הנתונים בכל צומת, ומשתמש בגיזום לאחר כדי להפחית התאמה יתר ולשפר את יכולת ההכללה של המודל.

הצדקה: שיטה זו מתאימה למערכי נתונים מורכבים יותר, שכן היא יכולה להתמודד עם נתונים רציפים ודיסקרטיים, ומתמודדת ביעילות עם ערכים חסרים. השימוש שלו באנטרופיה מאפשר הערכה מגוונת לגבי אילו פיצולים יארגן את הנתונים בצורה היעילה ביותר, ומספק גישה מעודנת יותר מ-CART בתרחישים רבים.

## 2.2 – פסאודו-קוד עבור כל אחת מהשיטות

### אלגוריתם CART

1. התחל עם כל מערך הנתונים כצומת הבסיס.
2. בחר את התכונה הטובה ביותר באמצעות מדד *Gini* שמחלק את מערך הנתונים לשתי קבוצות. חפש את התכונה שיוצרת את הענפים ההומוגניים ביותר.
3. פצל את מערך הנתונים לשתי קבוצות משנה באמצעות התכונה שנבחרה, יצירת שני צמתים צאצאים מתחת לצומת הנוכחי.
4. חזור על התהליך באופן רקורסיבי עבור כל צומת צאצא עם תת-קבוצת הנתונים הקשורים לצומת זה.
5. הפסק את הפיצול כאשר אחד מהם:
  - כל אלמנט בתת-הקבוצה שייך לאותה מחלקה.
  - אין עוד תכונות שצריך לקחת בחשבון.
6. במקרים בהם לא מושגת חלוקה ברורה (למשל, שאר התכונות אינן מתאימות), הקצה את המחלקה הנפוצה ביותר של נקודות הנתונים בתת-הקבוצה לצומת.

### C4.5 אלגוריתם

1. התחל עם כל מערך הנתונים כשורש העץ.
2. בחר את התכונה עם רווח המידע הגבוה ביותר לפיצול הנתונים. רווח מידע נמדד על ידי הפחתת האנטרופיה או *impurity* בהתפלגות המעמדות של מערך הנתונים.
3. פצל את הנתונים על סמך התכונה שנבחרה למספר קבוצות משנה המתאימות לכל ערך של התכונה, יצירת צומת צאצא עבור כל תת קבוצה.
4. החל באופן רקורסיבי את האלגוריתם על כל צומת צאצא, תוך שימוש רק בתת-קבוצת הנתונים הרלוונטיים לאותו צומת.
5. Pruning - גזום את העץ לאחר שצמח במלואו כדי להסיר ענפים שאינם תורמים לדיוק בנתונים שלא נראים.
6. הפסק את הרקורסיה אם:
  - כל המופעים בצומת שייכים לאותה מחלקה.
  - אין אפשרות להשיג מידע נוסף.
  - אין עוד תכונות לבחירה, אבל המופעים עדיין לא שייכים לאותה מחלקה. הקצה את הכיתה הנפוצה ביותר.

## 2.3 – תוצאות הניתוח

הערות:

- בשני המקרים השתמשנו ב- 10 fold cross-validation
- ויזואליזציה של העצים ושאר התוצאות בנספח בסוף הקובץ.

### CART תוצאות

```
Number of Leaf Nodes: 37
Size of the Tree: 73
Time taken to build model: 1.72 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      5377      89.8563 %
Incorrectly Classified Instances    607      10.1437 %
Kappa statistic                    0.5455
Mean absolute error                0.0945
Root mean squared error            0.2215
Relative absolute error            57.3143 %
Root relative squared error        77.1699 %
Total Number of Instances          5984

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.960    0.458    0.929      0.960    0.944      0.555    0.884     0.972    Healthy
               0.085    0.004    0.511      0.085    0.145      0.194    0.749     0.202    Hyperthyroid
               0.743    0.038    0.663      0.743    0.701      0.670    0.965     0.727    Hypothyroid
Weighted Avg.   0.899    0.398    0.885      0.899    0.884      0.548    0.885     0.913

=== Confusion Matrix ===
      a    b    c  <-- classified as
4949  22  185 |  a = Healthy
 240  24   20 |  b = Hyperthyroid
 139   1  404 |  c = Hypothyroid
```

מורכבות העץ: לעץ CART יש 37 צמתים עלים וגודל כולל

של 73.

- *Healthy*: ערכי true positive rate (TPR) ו-

*Precision* גבוהים, המצביעים על זיהוי יעיל.

- *Hyperthyroid*: *Recall* נמוך אך דיוק סביר, מה

שמצביע על קושי בזיהוי כל מקרי פעילות היתר של

בלוטת התריס, אך בצורה אמינה כאשר זה קורה.

- *Hypothyroid*: ערכי *Recall* טובים, מראה יעילות

בזיהוי רוב מקרי תת פעילות של בלוטת התריס, אם כי

עם כמה שגיאות.

מדדי ביצועים:

דיוק: 89.8563% מהמקרים סווגו נכון.

סטטיסטיקת *Kappa*: 0.5455, מצביע על הסכמה מתונה.

*ROC Area*: ערכי אזור ROC בסך הכל הגונים, מה

שמצביע על יכולת סיווג טובה בין המחלקות.



```

Number of Leaves :      67
Size of the tree :      106

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      5402      90.2741 %
Incorrectly Classified Instances    582      9.7259 %
Kappa statistic                     0.5801
Mean absolute error                 0.0844
Root mean squared error             0.2116
Relative absolute error             51.1659 %
Root relative squared error         73.7285 %
Total Number of Instances          5984

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.955	0.412	0.935	0.955	0.945	0.577	0.927	0.982	Healthy
	0.229	0.007	0.613	0.229	0.333	0.357	0.918	0.393	Hyperthyroid
	0.757	0.037	0.673	0.757	0.713	0.684	0.961	0.737	Hypothyroid
Weighted Avg.	0.903	0.359	0.896	0.903	0.895	0.577	0.930	0.932	

```

=== Confusion Matrix ===
      a    b    c  <-- classified as
4925   39   192 |    a = Healthy
 211    65     8 |    b = Hyperthyroid
 130     2   412 |    c = Hypothyroid

```

מורכבות העץ: לעץ C4.5 יש 67 עלים וגודל כולל של 106.

ביצועים ספציפיים לסיווג:  
 - *Healthy*: ביצועים דומים מאוד ל-CART, חזקים בזיהוי נבדקים בריאים.

מדדי ביצועים:

- *Hyperthyroid*: ערכי *Recall* טובים יותר מ-CART, דבר המצביע על שיפורים בזיהוי מקרי פעילות יתר של בלוטת התריס.

דיוק: 90.2741% מהמקרים סווגו נכון.

סטטיסטיקת *Kappa*: 0.5801, שהוא מעט טוב יותר מ-

- *Hypothyroid*: ערכי *Recall* דומים לאלו של CART, דבר שמעיד על שיעורי זיהוי טובים.

CART, מה שמצביע על הסכמה בינונית עד טובה.

*ROC Area*: מעט יותר טוב מ-CART, במיוחד בהבחנה בין סיווגים, מה שיכול להצביע על תהליך קבלת החלטות ניואנסים יותר.

## 2.4 – אומדן מידת הדיוק

הדיוק של כל שיטה כפי שחושבה במהלך הניתוח:

**CART: 89.8563%**

**C4.5: 90.2741%**

היתרון הקל בדיוק של C4.5 ניתן לייחס לטיפול המתוחכם שלו בבחירת תכונות וגיוזום, אשר נוטה להימנע מהתאמת יתר טוב יותר מ-CART.

## 2.5 – ניתוח השוואתי ומסקנות

### ניתוח השוואתי

- דיוק: C4.5 עולה במעט על ה-CART ברמת הדיוק הכללית ובסטטיסטיקת קאפה, מה שמצביע על איזון טוב יותר בין רגישות וספציפיות.
- מורכבות: ל-C4.5 יש עץ גדול יותר, מה שעשוי להצביע על דגם מורכב יותר מ-CART. מורכבות זו עשויה להוות גורם לביצועים מעט טובים יותר שלה, אך עשויה גם להצביע על סיכון גבוה יותר להתאמה יתר למרות תהליך הגיוזום.
- אזורי ROC ו-Precision-Recall (PRC): C4.5 מציג אזורי ROC ו-PRC טובים יותר, מה שמרמז על כך שהוא מסוגל יותר להבחין בין המחלקות מאשר CART.

## מסקנות

שני האלגוריתמים מתפקדים היטב, אבל C4.5 עוקף את ה-CART ברוב המדדים, מה שהופך אותו לבחירה קצת יותר טובה עבור מערך הנתונים הזה. המורכבות הגבוהה יותר של C4.5 עשויה לדרוש משאבי חישוב רבים יותר, ויש לשקול את יתרונם בביצועים מול גורם זה.

## הצעות ייעול

הפרויקט מאשר את ישימותם של מודלים של עצי החלטה לסיווג פעילות בלוטת התריס. בהתחשב בביצועים המעט מעולים של C4.5, הוא מומלץ למשימות דומות במסגרות קליניות שבהן יש חשיבות מכרעת לפרשנות ודיוק. עם זאת, עבודה עתידית יכולה לחקור את הדברים הבאים כדי לשפר עוד יותר את ביצועי המודל והיישום:

- כוונן פרמטרים: שני הדגמים עשויים להפיק תועלת מכוונן מעמיק יותר של הפרמטרים שלהם. עבור C4.5, התאמת גורם הביטחון לגיזום ומספר המקרים המינימלי לכל עלה עשויה להניב שיפורים. עבור CART, ניסוי עם פיצולים מינימליים שונים ושיפורי טוהר יכול ליעל את הביצועים.
- הנדסת תכונות: ניתוח מעמיק יותר ליצירה ובחירת תכונות שעשויות ללכוד את הניואנסים של תפקוד לקוי של בלוטת התריס בצורה יעילה יותר.
- אימות צולב: שימוש בטכניקות אימות צולב חזקות יותר כדי להבטיח שהמודלים יתכללו היטב לנתונים בלתי נראים. תובנות והמלצות אלו נגזרים מהתוצאות שסופקו, במטרה להנחות חידוד נוסף ואופטימיזציה של מודלים לסיווג.

## ניתוח השוואתי עם הספרות המחקרית

הסיווג של הפרעות בבלוטת התריס באמצעות למידת מכונה מתועד היטב בספרות, כאשר מחקרים שונים מדגישים את הפוטנציאל של אלגוריתמים כמו עצי החלטה בשל יכולת הפירוש והיעילות שלהם. מחקרים מדגישים לעתים קרובות את החשיבות של בחירת תכונות ואיכות נתונים, שהיו גם היבטים קריטיים בפרויקט שלנו.

המחקר הנוכחי על סיווג תפקודי בלוטת התריס הסתמך יותר ויותר על טכניקות למידת מכונה בשל יכולתן להתמודד עם מערכי נתונים גדולים ולזהות דפוסים מורכבים שאולי אינם ניכרים באמצעות שיטות סטטיסטיות מסורתיות. מחקרים הראו שאלגוריתמים של עצי החלטות, כמו CART ו-C4.5, שימושיים במיוחד מכיוון שהם מספקים מודל ברור של קבלת החלטות שקל לרופאים לפרש.

עם זאת, נושא שכיח אחד המודגש במחקר סיווג בלוטת התריס הוא האתגר של התמודדות עם נתונים לא מאוזנים, כאשר השכיחות של מחלקה אחת מאפילה באופן משמעותי על האחרים. זה רלוונטי במיוחד מכיוון שהפרעות בתפקוד בלוטת התריס כמו פעילות יתר של בלוטת התריס והיפותירואידיזם שכיחות פחות מתפקוד תקין של בלוטת התריס. גם CART וגם C4.5 יכולים להיות רגישים לחוסר איזון זה, מה שמוביל לרוב לדיוק גבוה יותר בחיזוי מעמד הרוב, אך לביצועים גרועים יותר בקבוצות הסיווג הקטנות יותר.

התוצאות שלנו עולות בקנה אחד עם הממצאים הללו, ומוכיחות שעצי החלטה יכולים להבחין ביעילות בין מצבים שונים של פעילות בלוטת התריס, עם מדדי ביצועים תחרותיים עם הסטנדרטים הנוכחיים. יתר על כן, ההצלחה היחסית של C4.5 בפרויקט זה מאששת מחקרים המצביעים על כך ששיטות המסבירות הן את איכות הפיצולים והן את המורכבות של המודל (באמצעות מנגנונים כמו גיזום) נוטות לבצע ביצועים טובים יותר, במיוחד במערך נתונים עם שילוב של תכונות סוגים ומספר לא מבוטל של מקרים.

## סיכום

לסיכום, פרויקט זה הוכיח את ישימותם של אלגוריתמי עצי ההחלטה *CART* ו-*C4.5* בסיווג מצבי בלוטת התריס בהצלחה ניכרת. עם זאת, ניתן לבצע שיפורים, במיוחד בסיווג מדויק של הפרעות בתפקוד בלוטת התריס, על ידי שילוב טכניקות נתונים מתקדמות וחקירת אלגוריתמים מתוחכמים יותר של למידת מכונה. מאמץ זה לא רק מקדם את ההבנה של ניתוח תפקוד בלוטת התריס אלא גם תורם לתחום הרחב יותר של אבחון רפואי באמצעות בינה מלאכותית.

תודה לבודק/ת ☺

