# 

# ממ"ן 21 – גיאורגי מירזויאן

# 309096485

# 

# שאלה 1

# א.

# מטרת כריית מידע היא לחזות ,לפי תצפיות שנעשו על אוכלוסיה, האם אדם חולה במחלת: פעילות יתר של בלוטת התריס(Hyperthyroidism), תת פעילות של בלוטת התריס(Hypothyroidism) או בריא.

לרשותנו עומד מאגר נתונים אשר נאסף מכמה מקורות ומעודכן משנת 1987, מכון גראוון סידני אוסטרליה, במאגר הנתונים יש 29 תכונות ועמודת סיווג.

# הנחות: הנחתי כי הנתונים נאספו בצורה מהימנה וכל טעות או ערך חסר היא לא נעשתה בזדון כלומר הנתונים נאספו על ידי אנשי מקצוע (רופאים, אחיות) וכן תהליך הבדיקה נעשה על ידי מכונה מכוילת ועומדת בסטנדרט.

# ב.

# 

| ערכים חסרים | יחידות מידה | תחום ערכים | סוג תכונה | תיאור תכונה | תכונה |
| --- | --- | --- | --- | --- | --- |
| כן | שנים | 1-65526 | נומרי- רציף | גיל | age |
| כן |  | male, female | קטגורי | מגדר | sex |
|  |  | ture, false | קטגורי | שימוש בתרופה תירוקסין | on thyroxine |
|  |  | ture, false | קטגורי | בדיקת שימוש בתירוקסין | query on thyroxine |
|  |  | ture, false | קטגורי | בדיקה שימוש בתרופות לבלוטת התריס | on antithyroid medication |
|  |  | ture, false | קטגורי | מצב בריאותי | sick |
|  |  | ture, false | קטגורי | בהריון | pregnant |
|  |  | ture, false | קטגורי | ניתוח הקשור לבלוטת התריס | thyroid surgery |
|  |  | ture, false | קטגורי | טיפול ביוד רדיואקטיבי | I131 treatment |
|  |  | ture, false | קטגורי | בדיקת תת פעילות בלוטת התריס | query hypothyroid |
|  |  | ture, false | קטגורי | בדיקת פעילות יתר בלוטת התריס | query hyperthyroid |
|  |  | ture, false | קטגורי | שימוש בתרופה לתיום | lithium |
|  |  | ture, false | קטגורי | זפקת | goitre |
|  |  | ture, false | קטגורי | גידול | tumor |
|  |  | ture, false | קטגורי |  | hypopituitary |
|  |  | ture, false | קטגורי | היסטוריה של פסיכיאטריה | psych |
|  |  |  | קטגורי | בדיקה של הורמון מגרה בלוטת התריס | TSH measured |
| כן | מילי־יחידות לליטר | 0.005-530 | נומרי- רציף | מדידה של הורמון מגרה בלוטת התריס | TSH |
|  |  |  | קטגורי | בדיקה של טרי-יודו-תירונין | T3 measured |
| כן | פיקומול לליטר | 0.05-18 | נומרי- רציף | מדידה של טרי-יודו-תירונין | T3 |
|  |  |  | קטגורי | בדיקה של תירוקסין חופשי ותירוקסין קשור | TT4 measured |
| כן | ננוגרם לדציליטר | 2-600 | נומרי- רציף | מדידה של תירוקסין חופשי ותירוקסין קשור | TT4 |
|  |  |  | קטגורי | בדיקה של תירוקסין חופש | T4U measured |
| כן | לא ידוע | 0.17-2.33 | נומרי- רציף | מדידה של תירוקסין חופשי | T4U |
|  |  |  | קטגורי | בדיקה של תירוקסין חופש | FTI measured |
| כן | ננוגרם לדציליטר | 1.4-881 | נומרי- רציף | מדידה של תירוקסין חופשי | FTI |
|  |  |  | קטגורי | בדיקה של גלובולין קושר תירוקסין | TBG measured |
| כן | ננוגרם לדציליטר | 0.1-200 | נומרי- רציף | מדידה של גלובולין קושר תירוקסין | TBG |
|  |  |  | קטגורי | מקור הבדיקה | referral source |
|  |  |  | קטגורי | אבחון | diagnosis |

# 

ג.

שלבי ה-KDD :

1. איסוף ושמירת הנתונים: הנתונים נאספו מכמה מקורות ונאגדו עבורנו תחת קובץ אחד.

2. ניקוי הנתונים: הנתונים מכילים ערכים רקים, שורות כפולות, עמודות המכילות ערכים שליליים בלבד וערכים מחוץ לטווח. טיפול בערכים אלה יתבצע בסעיף ה.

3. ביצוע טרנספורמציות על הנתונים: ישנן עמודות מיותרות אך הנתונים מתפלגים נורמלית לכן אין צורך בנרמול ערכים. במידת הצורך ניתן גם לעשות דסקרטיזציה על הנתונים.

4. בחירת שיטות לכריית מידע: נשתמש בפייטון עם ספיריות הבאות: sklearn, pandas,numpy,matplotlib,seaborn

5. ביצוע דיסקרטיזציות וסיווג הנתונים: נבצע שינויים במידע על מנת להתאים את המידע לשיטת כריית המידע.

6. הרצת שיטות לכריית מידע שנבחרו:נשתמש בפייטון עם הספריות מסעיף 4 על מנת להריץ את האלגוריתמים.

7. ניתוח התוצאות: נשתמש במדדי precision, recall, sensitivity, specificity על מנת לקבוע האם השיטות של כריית המידע אכן מצליחות לנבא אבחנה של תת/יתר פעילות של בלוטת התריס.

8. הסקת מסקנות:לאחר ניתוח התוצאות ניתן להבין אלו תכונות חשובות לאיבחון חולה/בריא, האם המודל מספק ניבוי טוב של תצפיות חדשות.

ד.

חלופות אפשריות לכריית מידע:

* אלגוריתם ID3–מדד Information Gain

אלגוריתם חמדן ורקורסיבי שיודע לבצע חיזוי/סיווג לערכים בדידים, דורש דיסקרטיזציה לכל הערכים הרציפים מחזיר עץ לא בהכרח בינארי.

* + יתרונות:בניית עץ מהירה ושימוש בכל מאגר המידע.
  + חסרונות:אינו מתמודד עם תכונות רציפות, ערכים חסרים.משוחד כלפי תכונות עם מספר רב של ערכים שונים.
* אלגוריתם C4.5–מדד Gain Ratio

אלגוריתם מבוסס על ID3 עם שיפור בהתמודדות עם תכונות מרובות ערכים.

* + יתרונות:מתמודד עם ערכים חסרים ותכונות קטגוריות ונומריות. נמנע משימוש בתכונות עם ערכים מרובים.
  + חסרונות:מתאים למאגר נתונים קטן, מעדיף פיצול תכונות עם מדד פיצול נמוך.
* אלגוריתם CART–מדד Gini

אלגוריתם זה מייצר עצים קטנים ביותר ובעל שגיאה נמוכה משאר האלגוריתמים. מדד הפיצול שואף לפצל לשתי קבוצות שוות גודל.

* + יתרונות:מתמודד עם ערכים חסרים ותכונות קטגוריות ונומריות.
  + חסרונות:מייצר עצים לא יציבים, מעדיף תכונות עם ערכים מרובים, חישוביות גבוהה.
* רגרסייה לינארית

שיטה זו מחפשת קשר לינארי בין משתנים רציפים.

* + יתרונות:קל למימוש, סיבוכיות נמוכה.
  + חסרונות:רמת דיוק גבוהה רק כאשר ישנו קשר לינארי בין המשתנה מטרה לתכונות.

# ה.

# טיפול בנתונים:

1. מחיקת נתונים:
   * תכונה hypopituitary מכילה רק ערכים שליליים, לא תורמת לנו מידע.
   * תכונה TBH מכילה 96% ערכים חסרים.
   * תכונות המכילות measured אלה תכונות המראות האם בוצע בדיקת דם, אם הערך שלילי אז לא בוצעה בדיקה ולא יהיה ערך מספרי של בדיקה.
   * תכונה referral source לא רלוונטי מאיפה הבדיקה הגיע.
   * מחיקת ערכים: ישנם הרבה מאוד ערכים חסרים לכן הסרתי אותם ממאגר הנתונים למרות הצמצם במספר הרשומות נשארנו עם יותר מחצי מהרשומות המקוריות.
   * רשומות כפולות: ישנן שלוש רשימות המכילות את אותן ערכים בדיוק גם מהן נפטר.
   * ערכים חריגים: יש מספר ערכי גיל חריגים שניתן להיתר מהם.
2. טרנספורמציית נתונים:
   * כלל הנתונים הקטגוריים שונו מ-t/f ל-1/0 בגלל אילוצים של הסיפריות בפייטון.
   * תכונת האיבחון מכילה הרבה ערכים שונים ביחד עם מספר האבחון/תיק רפואי, לכן נמחק את מספר האבחון ונשנה את הופעת הערכים לפי קבוצות אבחנה המופיעות בקובץ הסברים על מאגר הנתונים(התעלמתי משילובים והשארות באבחונים):

hyperthyroid conditions - מקבל ערך 2

A hyperthyroid

B T3 toxic

C toxic goitre

D secondary toxic

hypothyroid conditions - מקבל ערך 1

E hypothyroid

F primary hypothyroid

G compensated hypothyroid

H secondary hypothyroid

אדם ללא הבחנה מקבל ערך 0

# 

טבלת התפלגות נתונים נומריים:

|  | **age** | **TSH** | **T3** | **TT4** | **T4U** | **FTI** |
| --- | --- | --- | --- | --- | --- | --- |
| **count** | 4490 | 4490 | 4490 | 4490 | 4490 | 4490 |
| **mean** | 53.1 | 2.12 | 2.03 | 107.96 | 0.97 | 112.14 |
| **std** | 18.39 | 4.13 | 0.76 | 30.46 | 0.16 | 31.58 |
| **min** | 1 | 0.005 | 0.1 | 4 | 0.25 | 3.5 |
| **25%** | 38 | 0.42 | 1.6 | 90 | 0.87 | 95. |
| **50%** | 56 | 1.3 | 2 | 104 | 0.97 | 108 |
| **75%** | 68 | 2.3 | 2.3 | 122 | 1.06 | 125 |
| **max** | 97 | 98 | 18 | 430 | 2.12 | 642 |

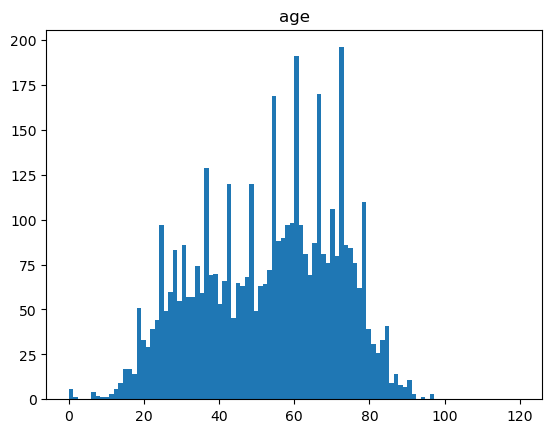
# מטריצת קורלציה:

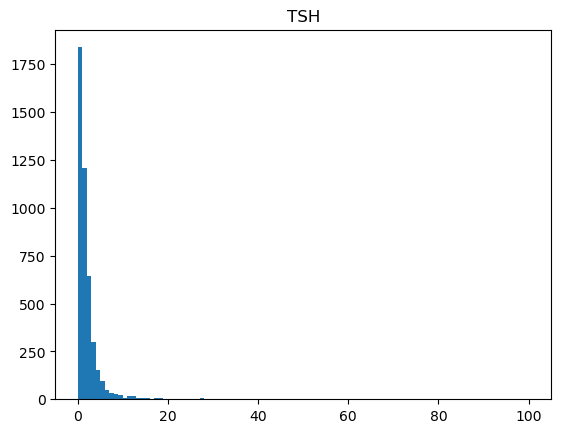
# 

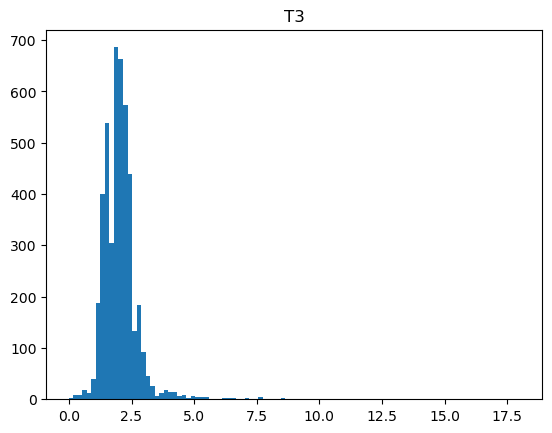
# קורלציה:

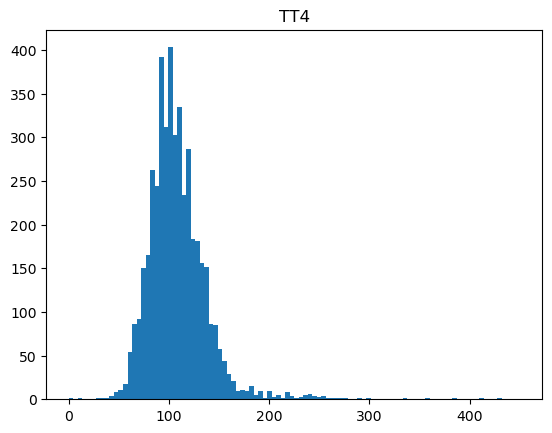
* בין TT4 וFTI ישנה קורלציה חיובית גבוהה
* בין T3 לTT4 FTI T4U, בין TT4 לT4U ישנה קורלציה חיובית חלשה
* בין T4U לFTI, בין TSH לFTI ישנה קורלציה שלילית חלשה

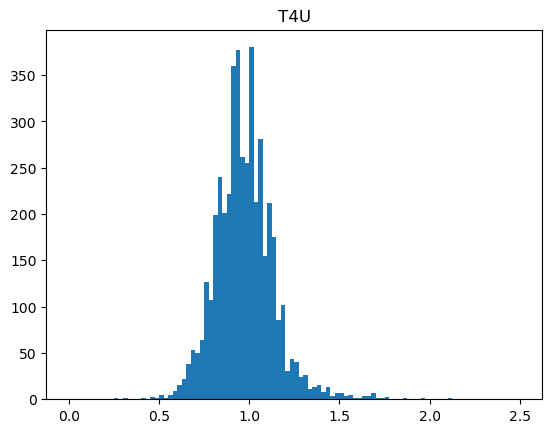
היסטוגרמות:











ספירת ערכים של תכונות קטגוריות:

sex

0.0 2930

1.0 1563

Name: count, dtype: int64

on thyroxine

0 4032

1 461

Name: count, dtype: int64

query on thyroxine

0 4450

1 43

Name: count, dtype: int64

on antithyroid medication

0 4419

1 74

Name: count, dtype: int64

sick

0 4317

1 176

Name: count, dtype: int64

pregnant

0 4470

1 23

Name: count, dtype: int64

thyroid surgery

0 4431

1 62

Name: count, dtype: int64

I131 treatment

0 4404

1 89

Name: count, dtype: int64

query hypothyroid

0 4244

1 249

Name: count, dtype: int64

query hyperthyroid

0 4131

1 362

Name: count, dtype: int64

lithium

0 4433

1 60

Name: count, dtype: int64

goitre

0 4449

1 44

Name: count, dtype: int64

tumor

0 4390

1 103

Name: count, dtype: int64

psych

0 4179

1 314

Name: count, dtype: int64

תובנות מהנתונים:

* כל הנתונים הנומרים מתפלגים נורמלי, תכונה FTI בעלת צידוד שלילי.
* יש פי 2 יותר נשים מגברים בבדיקה.
* כל הנתונים הקטגורים, פרט לגיל ומין, מכילים יותר ערכים שלילים(FALSE=)

שאלה 2

א+ב.

שיטות אנסמבל יותר מדויקות עקב ריבוי של מסווגים "חלשים" מאשר שימוש במסווגים "חלשים" בנפרד, אך שיטות אלה צריכות יותר כוח חישוב וזמן אימון, למרות זאת החלטתי לבחור בשתי שיטות אנסמל. השתמשתי ב10-fold cross validation על מנת להימנע מהתאמת יתר וGridSearchCV אשר מריצה מספר פעמים את האלגוריתם עם מטה פרמטרים שונים בשביל למצוא את הטובים ביותר.

שיטה ראשונה AdaBoost:

שיטה זו היא אחת משיטות אנסמבל והיא יעילה כאשר יש מחלקות סיווג לא מאוזנות, כמו במקרה שלנו כאשר רוב האיבחונים הם של אנשים בריאים ומעט איבחונים של אנשים חולים.

בשיטה זו מאתחלים כל תצפית במאגר המידע עם משקולות שווים.

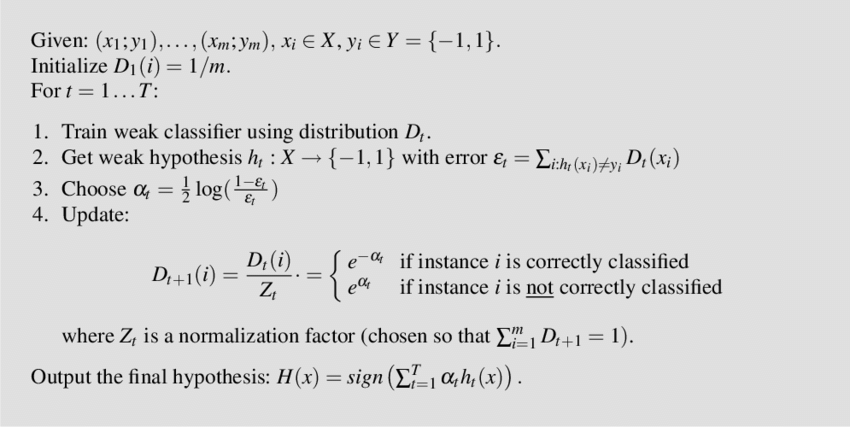
מייצרים k מסווגים ב-k איטרציות, בכל איטרציה נבחר תת-מערך בגודל |di| מתוך מאגר המידע.

ההסתברות של תצפית להיבחר לקבוצת האימון מבוססת על משקלה.

מייצרים מסווג, במקרה שלנו משתמשים במסווג עץ החלטה מבוסס אלגוריתם CART עם מדד GINI).

מחשבים אחוז שגיאה של מודל הסיווג על קבוצת המבחן מתוך מאגר המידע, תצפיות אשר מסווגות לא נכון מגדילות את משקל התצפית ולהיפך.

לשיטה זו יש מטה פרמטרים שניתן לשנות לקבל תוצאות טובות יותר, לדוגמא: כמות המסווגים(=מס' איטרציות), מדד למידה משגיאה, גודל קבוצת הנתונים.

פסודו-קוד:

שיטה שנייה RandomForests:

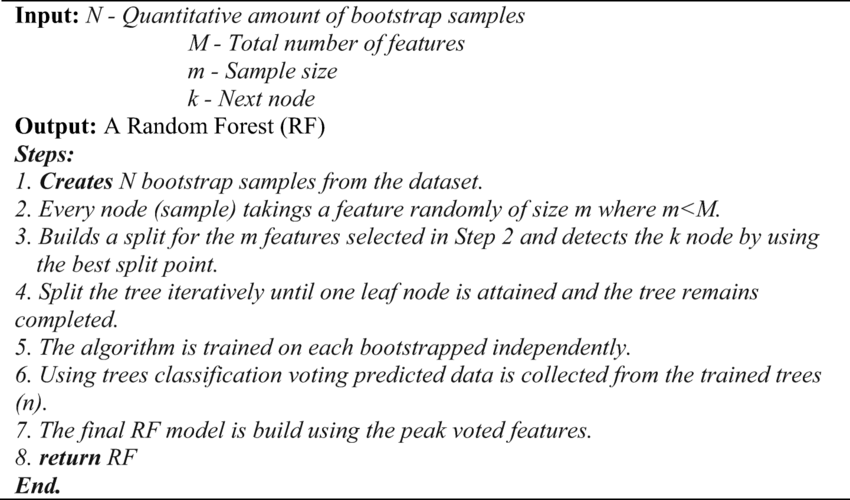
יער אקראי הוא אוסף של עצי החלטה(מסווגים "חלשים"), הנוצרים מתת-קבוצה של קב' האימון.

במקרה שלנו משתמשים במסווג עץ החלטה מבוסס אלגוריתם CART עם מדד GINI.

כל עץ נבנה על תת-קבוצת אימון שונה אשר מכילה תצפיות חדשות וישנות, בכל צומת נבחר מאפיין מפצל הטוב ביותר מקבוצה אקראית של מאפיינים. סיווג של תצפית חדשה נעשה על ידי הצבעה של כלל העצים, הסיווג בעל מספר הגבוה ביותר של הצבעות הוא הסיווג של התצפית.

גם כאן יש מטה פרמטרים שניתן לשחק איתם על מנת לקבל סיווג מוצלח, לדוגמא: מספר העצים לייצור הסיווג, מספר מינימלי של דגימות לפיצול של צומת, גודל קבוצת הנתונים..

פסודו-קוד:



ג+ד.

השתמשתי GridSearchCV, פונקציה זו מריצה עבור כל מטה פרמטר ועבור כל קב' אימון וקב' מבחן מ10-foldCrossValidation. נבחר את מידת ההצלחה של אלגוריתם למי מידת הדיוק(accuracy) שלו.

יער אקראי:

מטה פרמטרים הם:

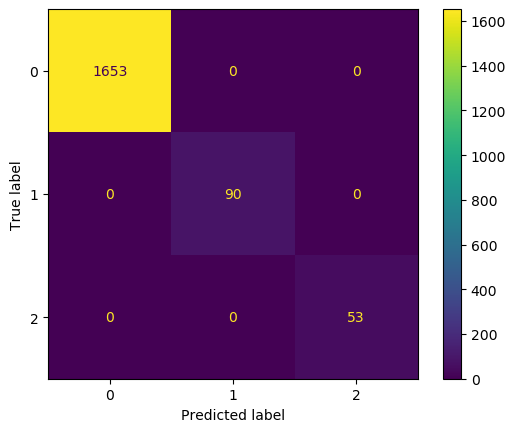
מס' העצים בכל איטרציה של בניית יער, בחרתי ב: 100, 200, 300, 400.

גודל מינימלי לפיצול של צומת: 3, 5, 7.

כל ריצה לקחה בערך שנייה אחת. זמן ריצה כולל 120 שניות = 10 \* 4 \* 3.

|  | **healthy** | **hyper\_cond** | **hypo\_cond** | **accuracy** | **macro avg** | **weighted avg** |
| --- | --- | --- | --- | --- | --- | --- |
| **precision** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| **recall** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| **f1-score** | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

|  | **class** | **sensitivity** | **specificity** |
| --- | --- | --- | --- |
| **0** | 0 | 1.0 | 1.0 |
| **1** | 1 | 1.0 | 1.0 |
| **2** | 2 | 1.0 | 1.0 |



AdaBoost:

מטה פרמטרים הם:

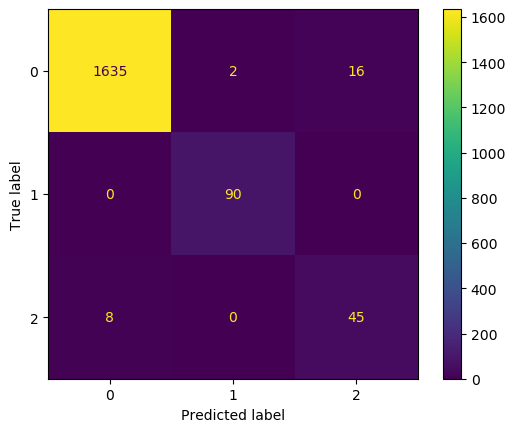
כמות המסווגים(מס' איטרציות) , בחרתי ב: 160, 200 ,240.

מדד למידה משגיאה:0.5, 0.7, 0.8, 0.9.

כל ריצה לקחה בערך שנייה אחת. זמן ריצה כולל 120 שניות = 10 \* 4 \* 3.

|  | **healthy** | **hyper\_cond** | **hypo\_cond** | **accuracy** | **macro avg** | **weighted avg** |
| --- | --- | --- | --- | --- | --- | --- |
| **precision** | 0.994559 | 0.978022 | 0.745098 | 0.986637 | 0.905893 | 0.987211 |
| **recall** | 0.990964 | 1.000000 | 0.808511 | 0.986637 | 0.933158 | 0.986637 |
| **f1-score** | 0.992758 | 0.988889 | 0.775510 | 0.986637 | 0.919052 | 0.986881 |

|  | **class** | **sensitivity** | **specificity** |
| --- | --- | --- | --- |
| **healthy** | 0 | 0.933824 | 0.990964 |
| **hyper\_cond** | 1 | 0.998828 | 1.000000 |
| **hypo\_cond** | 2 | 0.992567 | 0.808511 |



ה.

הדיוק של יער אקראי מאוד גבוהה, הסכנה שיש התאמת יתר למרות שהשתמשנו ב-k-fold cross validation. אחוזי הדיוק נשמרו לאורך כל הבדיקות.

הפרמטרים שלהם התקבל הדיוק הגבוהה ביותר: פיצול מינימלי של 3, 300 עצים בריצה של האלגוריתם.

עבור adaboost רואים אחוזי דיוק קצת נמוכים יותר אך עדיין גבוהים.

הפרמטרים שהניבו את התוצאות הטובות ביותר: 200 איטרציות עם קצב למידה של 0.8.

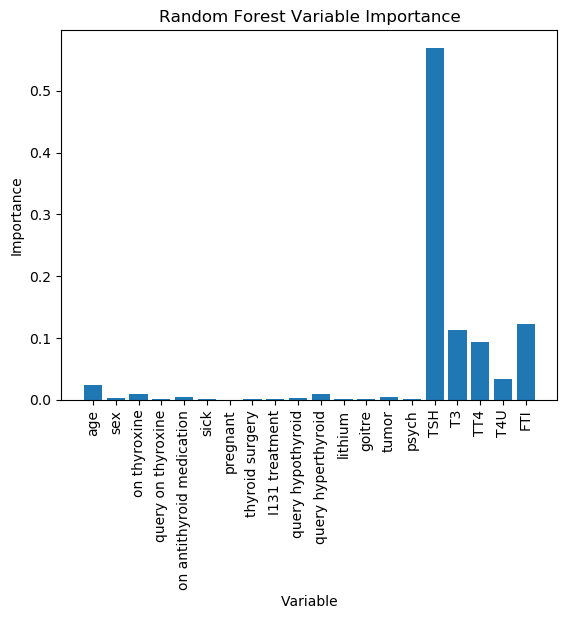
ישנם עוד פרמטרים איתם ניתן לשפר את זמן הריצה, כמו עומק העץ, אלגוריתם של עץ החלטה(מדד הפיצול), התמורה תיהיה גדולה עבור מאגרי נתונים גדולים יותר משלנו.

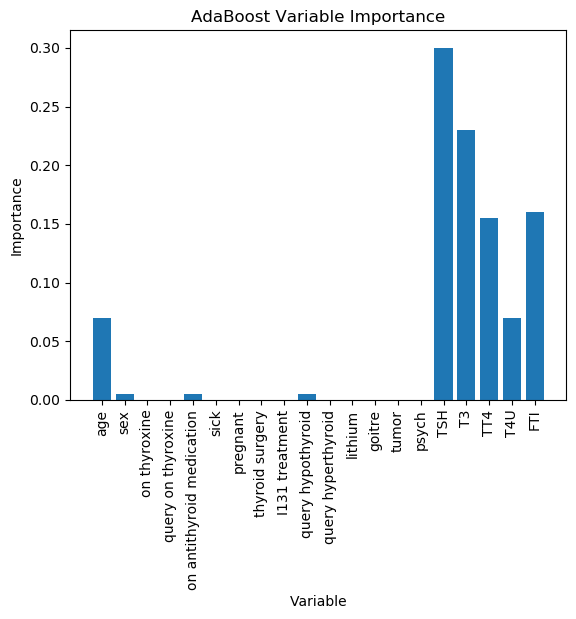
שיטות אלה מניבות תוצאות מאוד מדויקות אך אינן ניתנות לתצוגה גרפית.

מעניין לראות את בחירת המאפיינים החשובים משני האלגוריתמים:

יער אקראי מסתמך בעיקר על התכונה TSH לעומת adaBoost שמסתמך על התכונות הנומריות.

שניי האלגוריתמים אינם מסתמכים על התכונות הבינריות, וזה עומד עם התובנות ממטריצת הקורלציה.





הצעות לשיפור:

קשה למצוא משהו לשיפור עם אחוזי דיוק כה גבוהים, אך אולי היה שווה להחליק את הנתונים הנומרים על מנת לייצר התפלגות אחידה יותר, זה יכול היה לתרום לאלגוריתם של adaBoost. בעיה אחרת היא האיזון בין כמות האבחנות של בריאים לחולים, הייתי מציעה במקום למחוק נתונים עם ערכים חסרים לנסות למלא נתונים על ידי ממוצע ולייצר נתונים סינטטים של חולים.