רועי ארגמן

ממ״ן 21 כריית מידע

שאלה 1 :

א. **מטרת כריית המידע** בפרויקט זה היא לחזות קיימות של מחלת כליות כרונית בהסתמך על מגוון נתוני דם שנמדדו במעבדה עבור פרטים שאובחנו כבעלי מחלת כליות כרונית ופרטים ללא מחלה. לצורך כך יש ברשותנו בסיס נתונים בו 400 רשומות כאשר כל רשומה מייצגת נבדק אחד ייחודי ובנוסף 25 תכונות כאשר התכונה ה25 היא משתנה המטרה (notckd/ckd) שדה בוליאני שמציין האם לנבדק ישנה מחלת כליות כרונית או לא.

ב. נציג טבלה בה **נתאר את כלל התכונות** סיווגם נדגיש שאלו הנתונים לפי weka לכן יש שוני בין סוג התכונות במאמר ובטבלה שנציג ובנוסף שטווח הערכים טרם נורמל. לאחר הנרמול ערכים מספריים (numeric) יהיו בטווח אחיד של 0-1.

Attribute name	*Description*	*Type*	*Domain*
age	test subject’s age in years	numeric	2-90
bp	blood pressure	numeric	50-180
sg	specific gravity	nominal	1.005-1.025
al	albumin level	nominal	0-5
su	sugar levels	nominal	0-5
rbc	red blood cells	nominal	normal | abnormal
pc	pus cell	nominal	normal | abnormal
pcc	pus cell clumps	nominal	normal | abnormal
ba	bacteria	nominal	normal | abnormal
bgr	blood glucose random	numeric	22-490
bu	blood urea	numeric	1.5-391
sc	serum creatinine	numeric	0.4-76
sod	sodium levels	numeric	4.5-163
pot	potassium levels	numeric	2.5-47
hemo	**haemoglobin** levels	numeric	3.1-17.8
pcv	packed cell volume	numeric	9-54
wbcc	white blood cell count	numeric	2200-26400
rbcc	red blood cell count	numeric	2.1-8
htn	hypertension indication	nominal	yes | no
dm	diabetes mellitus	nominal	yes | no
cad	coronary artery disease	nominal	yes | no
appet	appetite present	nominal	yes | no
pe	pedal edema	nominal	yes | no
ane	does subject has **anaemia**	nominal	yes | no
class	subject having chronic kidney disease (goal)	nominal	yes | no

ג. ככל תהליך KDD ישנם כמה שלבי מפתח, נתאר את עיקרם בסעיף זה ובסעיפים הבאים נתייחס לאופן בו ביצענו את השלב בפרויקט שלנו:

1) **איסוף ושמירת הנתונים:** הנתונים הם נתונים פיזיולוגים שנאספו ע"י צוות המחקר של המחקר במאמר המצורף לקובץ המטלה. הנתונים מציגים מדדים ואינדקציות שונות בדבר מטופלים לרבות בדיקות דם מחלות רקע ריכוז חומרים בדם ובשתן ואת תכונת המטרה (קיום מחלקת כליות כורנית). הנתונים נשמרו בפורמט arff. וכפי שציינו חלק מהתכונות סווגו בצורה שונה מן המאמר בשל השימוש בweka.

2) **ניקוי נתונים:** בחלק זה ננסה "לנקות רעשים" משמע לאתר ערכים חסרים, ערכים שחשודים כלא נכונים וערכים לא חוקיים.

3) **טרנספורמציות/שינוי נתונים:** במידה וניתן להשלים/לתקן נתונים מהסעיף הקודם נעשה זאת למשל ע"י השלמת ערך ממוצע או ערך נפוץ במידה ולא ניתן לקבוע מה הערך הנכון נשקול למחוק את הרשומה ובמידה וישנו מספר גדול של נתונים חסרים עבור שורה/עמודה מסוימת נשקול למחוק לגמרי עמודה זאת. בנוסף נשתמש במגוון שיטות על מנת לסנן תכונות מיותרות. למשל תכונות תלויות לינארית בתכונות אחרות.

4) **בחירת שיטות וכלים לצורך תהליך כריית המידע:** בפרויקט זה נתבקשנו לעבוד עם עצי החלטה נבחר 2 מתוך האלגוריתמים להפקת עצי החלטה מהנתונים.

5) **ביצוע כריית מידע תוך שימוש בכלים שבחרנו:** בפרויקט נשתמש בתוכנת WEKA על מנת להפעיל את האלגוריתמים שציינו ולהכין את המידע לתהליך הכרייה.

6) **ניתוח תוצאות:** נשתמש בכלים ומדדים שונים על מנת לנתח את איכות התוצאות ביניהם, precision , f-score, recall, roc ,accuracy נבחר בחלק מאלו על מנת לאשש את אמינות התוצאות.

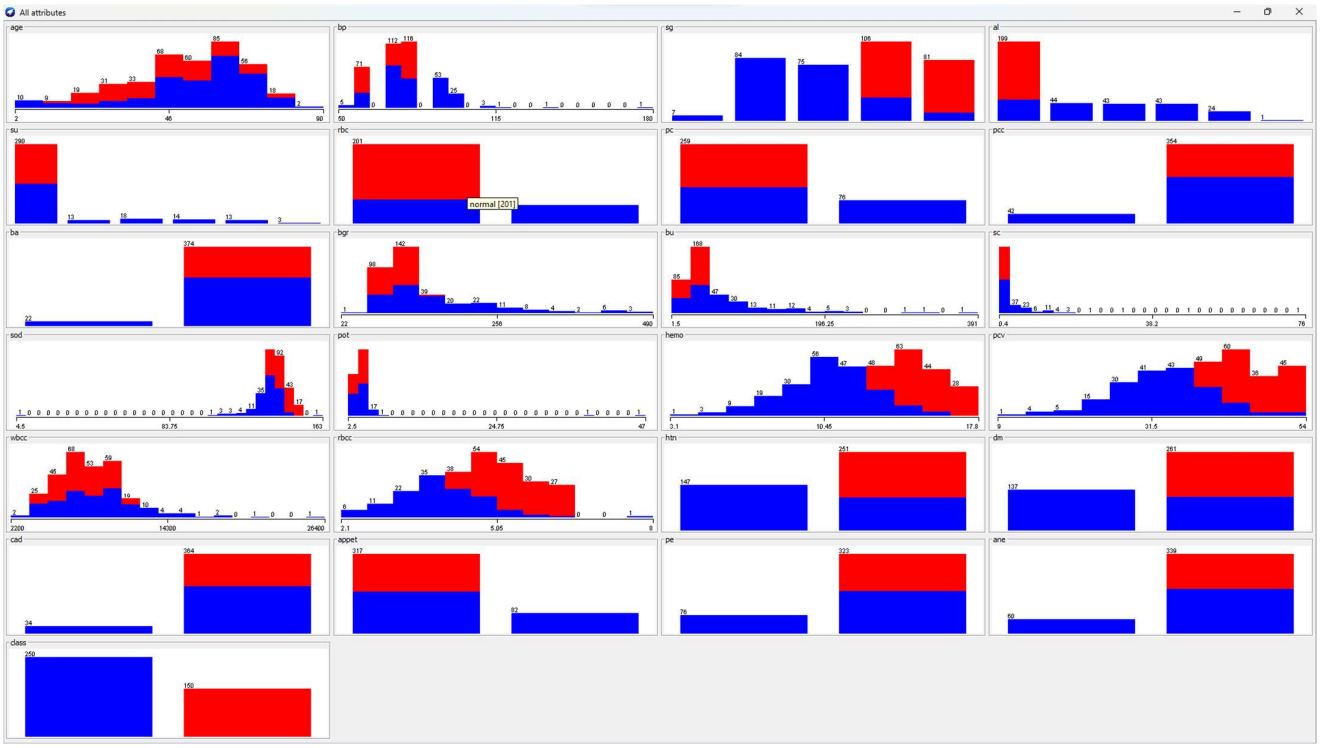
7) **הסקת מסקנות:** נציג את תוצרי תהליך כריית המידע באופן גרפי ונתאר את המסקנות שניתן להסיק מתוצרים אלו.

ד. סקירת השיטות לחיזוי מידע:

שם השיטה	*עץ החלטה ID3*	*מדד החלטה C4.5*	*עץ החלטה cart*	***רגרסיה לינארית***
תיאור השיטה	שיטה לחיזוי ערכים בדידים	שיטה לחיזוי ערכים בדידים	שיטת חיזויי לערכים בדידים ורציפים	שיטתה לחיזוי ערכים רציפים כאשר ישנה התאמה לינארית בין וקטור אחד או יותר בתכונות לבין וקטור המטרה
אילוצים	לא עובד עם תכונות רציפות ולכן יש לבצע דיסקרטיזציה (הפיכה של משתנים רציפים לבדידים) על מנת לעשות בו שימוש	לא עובד עם תכונות רציפות ולכן יש לבצע דיסקרטיזציה (הפיכה של משתנים רציפים לבדידים) על מנת לעשות בו שימוש	לא עובד עם תכונות רציפות ולכן יש לבצע דיסקרטיזציה (הפיכה של משתנים רציפים לבדידים) על מנת לעשות בו שימוש	פועל עם מאפיינים רציפים בלבד
מדד הפיצול	information Gain מרווח אינפורמטיבי חישוב איטרטיבי של כל התכונות שקיימות בכל רמה בעץ לצורך בחירת תוכנת הפיצול הבאה שתהייה בעלת האנטרופיה הנמוכה ביותר	Ratio Gain הינו שיפור **של information** Gain ע"י חישוב חלקם היחסי של התכונות הנבחרות ברווח **אינפורמטיבי** ביחס לאוכלוסיית האימון וכפועל **יוצא הפחתת** ההטיה של הרווח האינפורמטיבי	Gini Index פיצול בינארי בכל צומת בעץ כאשר בחירת תוכנת הפיצול היא ע"י בחירת התכונה בעל פוטנציאל הרעש הנמוך ביותר	שיטה זאת לא יוצרת עץ החלטה אלה פונקציה לינארית המתארת את הקשר הלינארי בין תכונה **מסוימת** לבין משתנה המטרה
יתרונות	פשטות ויעילות זמן ריצה	פועל היטב עם נתונים רציפים. בנוסף גוזם ערכים כאלה וערכים חסרים (כאשר הם מחולקים למקטעים)	נוטה לספק מודלים **קומפקטיים** ומתאים ביותר לחיזוי מידע בינארי (כן או לא) הקומפקטיות של האלגוריתם נובעת גם מתכונת הגיזום של האלגוריתם	לאחר חישוב ניתן להציג את גרף הפונקציה בתחום הגדרתה בצורה גרפית ומאפשרת זיהוי מגמות והערכת ציפיות על סמך **פונקציית** הרגרסיה
חסרונות	עלול ליצור overfitting לאחר דיסקרטיזציה של נתונים שייבחרו בתכונות הפיצול על אף שתרומתם בפועל שולית.	**נטייה** ליצירת עצים לא מאוזנים בשל מדד הפיצול.	עלול ליצור overfitting לאחר דיסקרטיזציה של נתונים שייבחרו בתכונות הפיצול על אף שתרומתם בפועל שולית. בשל העדפתו לבחור נתונים בעלי רעש נמוך וכפועל יוצא קושי להתמודד עם DATA בעלת הרבה תכונות	מאחר ואין ערובה לכך שקיים קשר לינארי שכזה שימוש בה במקרים של פיזור לא אחיד תביא לרמת התאמה נמוכה מאוד ומודל פחות דיוק שלא ניתן לשימוש אמין ויאלץ שימוש במודלים מורכבים כמו קירוב פולינו**מיאלי**

ה. **נציג את שלבי הכנת הנתונים:**

ראשית נציג את הנתונים לפני הכנתם :

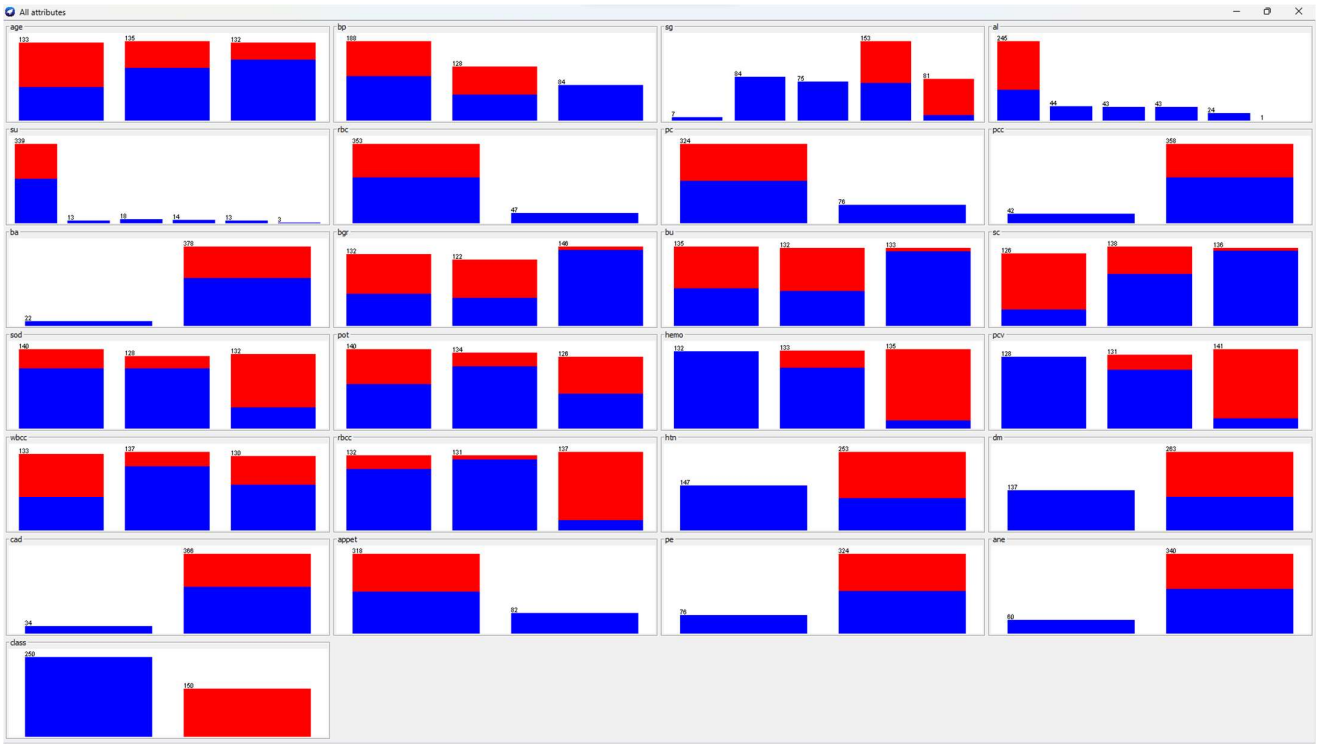


כעת נתחיל לנקות את הנתונים :

1. **טיפול בערכים חסרים** :
   a. השתמשתי בכלים אוטומטים של WEKA כך שישלימו את הערכים המספריים בממוצע התכונה ואת הערכים הנומינליים בערך הנפוץ ביותר. מאחר וזו שיטה מקובלת להשלמת נתונים. בנוסף במאמר השתמשו בשיטה זהה והגיעו לתוצאות טובות מבחינת חיזוי.
   b. שמתי לב כי קיימות 3 תכונות שלכולן יותר מ25% ערכים חסרים. בתחילה השארתי ערכים אלו על מנת לא לאבד מידע אחר ברשומות. זאת, בידיעה שיהיה עלי למחוק את הרשומות בעלות המידע החסר אם בעת תהליך הכרייה יתגלה כי השלמה של נתונים אלו באופן מלאכותי פוגעת באיכות המודל.
      i. red blood cells- **Rbc**
      ii. white blood cell count-**wbcc**
      iii. red blood cell count- **rbcc**
   c. התצוגה הגרפית מראה לנו גם כי קיימים בכמה מדדים ערכי קיצון לאחר בדיקה באינטרנט על טווח המדדים לא מצאתי נתונים חריגים שחשודים כטעות ולכן לא מחקתי רשומות במקרים אלו למשל ערך נמוך יחיד בסך 4.5 במדד sod שהתגלה כערך נמוך בטווח הנורמה.

2. **נרמול נתונים:** בחרתי להשתמש בכלים אוטומטים של weka גם על מנת לנרמל את הנתונים. כפי שציינתי בעת מעבר על הנתונים זיהיתי כי קיימים מספר ערכי קיצוניים (outliers) ונרמול נתונים ידוע כשיטה טובה להקטנת ההשפעה של ערכים כאלו על התוצאות.

3. בנוסף לאחר בחינת אפשרויות שונות החלטתי לבצע דיסקרטיזציה על מנת לקבל חלוקה מאוזנת ככל הניתן של ערכים. בכלים האוטומטים של weka השתמשתי בחלוקת שוות שכיחות על פני חלוקת שוות רוחב. ואכן מבין החלוקות זו החלוקה שוויזואלית נראתה כמאוזנת ביותר במרבית המדדים.



שאלה 2 :

א. בשאלה זו בחרתי מבין 4 השיטות שציינתי בשאלה הקודמת, בחרתי להשתמש ב C4.5 וCART לא בחרתי ברגרסיה לינארית מאחר והיא עובדת רק על ערכים רציפים מאחר ובחרתי לבצע דיסקרטיזציה לא היה זה נכון או אפשרי להשתמש בה. כמו כן, מאחר וC4.5 וCART שניהם שיפורים של ID3 בחרתי בהם על פניו.

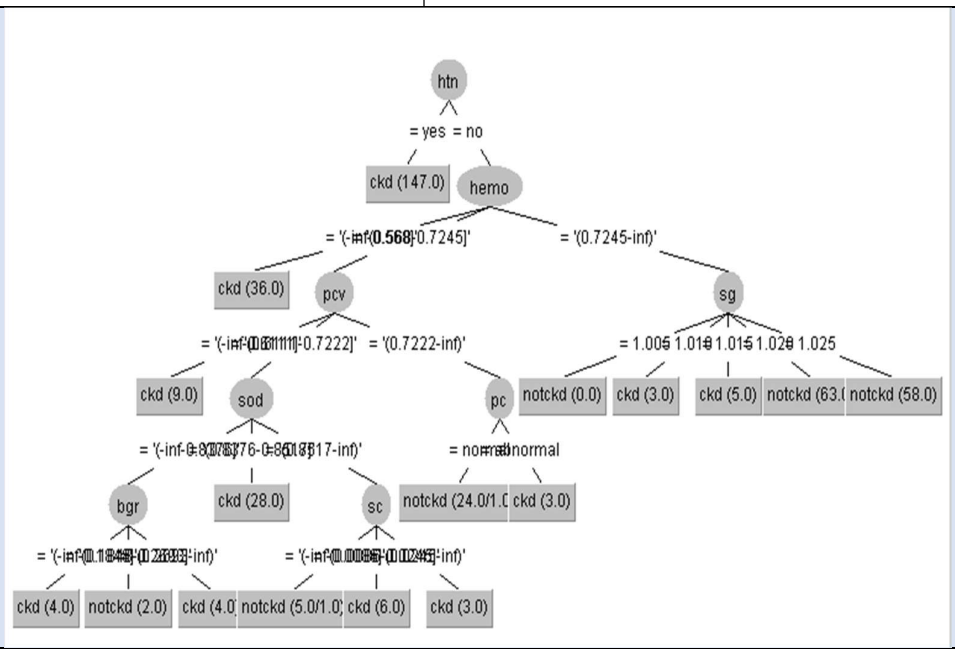
אם כן, בחרתי בשתי שיטות שמשתמשות בעצי החלטה. מאחר ותיארתי כבר את אופן הפעולה של שתיהן בסעיף הקודם רק אציין כי השימוש בהם יפחית את בעיית הoverfitting שמתקיימת בשכיחות גובהה יותר בID3.

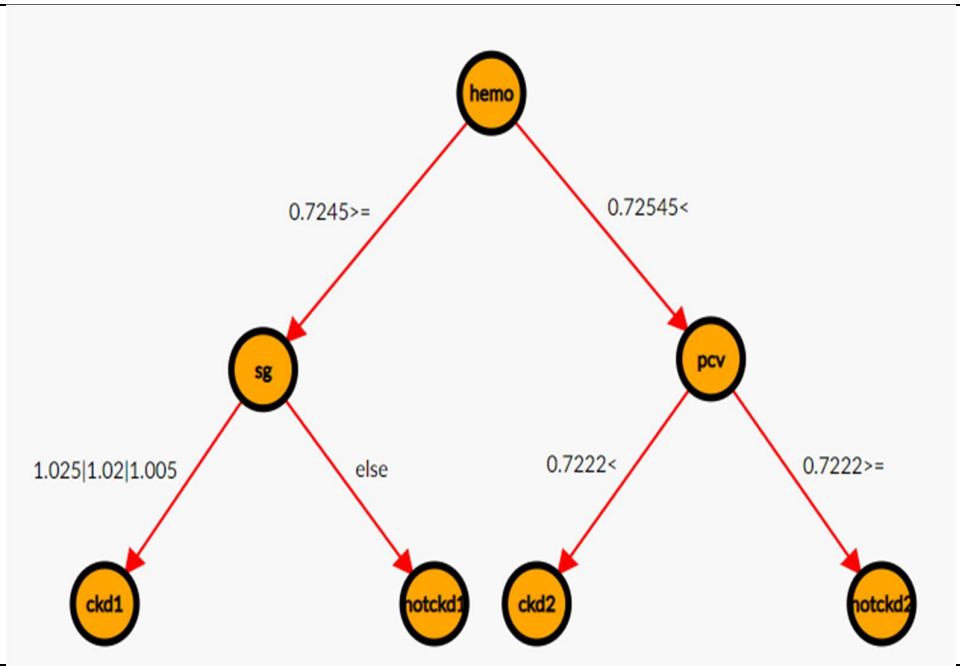
ב. נתאר את שלבי החיזוי של שני האלגוריתמים נציין שבשני האלגוריתמים אנו זוכרים כי ישנן 3 תכונות שאותן בחרנו להשאיר על אף אחוז גבוהה של ערכים חסרים במידה ותכונות אלו יתגלו כרעש נוריד נחזור לתהליך הכנת הנתונים ונוריד אותן. (אבל במקרה שלי הן לא) בשתי השיטות נבצע שימוש ב10-fold כפי שבוצע במאמר ומאחר ושינוי פרמטר ה-k הראה השפעה נמוכה עד לא קיימת ביחס למדדים אחרים :

a. **C4.5 tree** – בWEKA נבחר את אלגוריתם 48J ליצירת עץ החלטה מבוסס אלגוריתם C4.5 באמצעות שינוי פרמטרים כמו כמות Size Batch ורמת ביטחון על מנת לצמצם את אחוז השגיאה.

b. **Cart-Tree** – בWEKA נבחר simpleCartTree ונפעיל את האלגוריתם שתיאורו מצוי בשאלה הקודמת. גם כן נרצה למנוע overfitting ובאותו הזמן להגיע למידת שגיאה נמוכה ככל הניתן. העץ ביסודו יצא קטן יחסית ולכן משחק עם גודל העץ לא הועיל.

ג. נתאר את תוצאות הניתוחים

פרמטר	C4.5 tree	Cart-Tree
precision	**97.75**	**96.5**
מספר שגיאות נומינאלי	**9 מתוך 400**	**14/400**
False positive	**4**	**8**
False Negative	**5**	**6**
כמות צמתים	**25**	**7**
כמות עלים	**17**	**4**
גובה העץ	**5 כולל השורש**	**2 כולל השורש**
Roc	**0.985**	**0.959**
תיאור גרפי C4.5		

	תיאור גרפי Cart

ד. שתי השיטות הראו רמת דיוק גבוהה מאוד ואף קרובה לרמת הדיוק במאמר מבלי להפעיל אלגוריתמים מתקדמים יותר לכריית מידע. הדבר מצביע על כך שרמת האמינות של הנתונים הייתה גבוהה מלכתחילה בעיקר בתכונות הסיווג.

ה. ניתן לראות שאלגוריתם C4.5 הראה תוצאות מדויקות יותר (5 טעויות פחות) וגם שמדד הROC שלו גבוה יותר. אולם יחס זה הוא שולי ולכן כאשר נבחן את המידע בעיניו העסקיות של בית החולים כנראה שנעדיף לבצע בדיקה ע"פ התוצאות שקיבלנו מהעץ של Cart. מאחר ורמת הדיוק של Cart גבוהה מאוד ובנוסף העץ קטן מאוד, המשמעות היא שיידרשו פחות בדיקות לכל מטופל על מנת לאשש או לשלול מחלת כליות כרונית.
עוד מעניין לראות כי בעץ של C4.5 הערך hypertension indication מכריעה כבר בהתחלה האם קיימת מחלה, אולם, בעץ של Cart ניתן לראות שמדד זה לא נבחר כלל כמדד פיצול.
הגדלת מאגר הנתונים וכמובן גיוונם האתני, יכולה לקבוע האם בחירותו של C4.5 באמת מוצדקת, או שמא קיים חשש לoverfitting למאגר זה. הגדלת מאגר הנתונים יכולה להועיל בפרמטר נוסף שלא ניתן לקבוע באופן חד משמעי אבל ניתן להבחין בו בתוצאותינו, בעץ C4.5 יש יותר מקרים של False negative מאשר False Positive ובעץ Cart המצב הפוך. לצערנו מדד זה נמוך מידי על מנת להסיק מסקנות חד משמעיות.
במידה ובכל זאת ירצה חוקר לפתח בדיקה בהסתמך על עצי ההחלטה הללו, קיום כלל זה יכול להיות חשוב להחלטה הסופית באיזו בדיקה עדיף להשתמש בהיבטים של אמינות הבדיקה, עלות (כספית) מול תועלת ועוד.