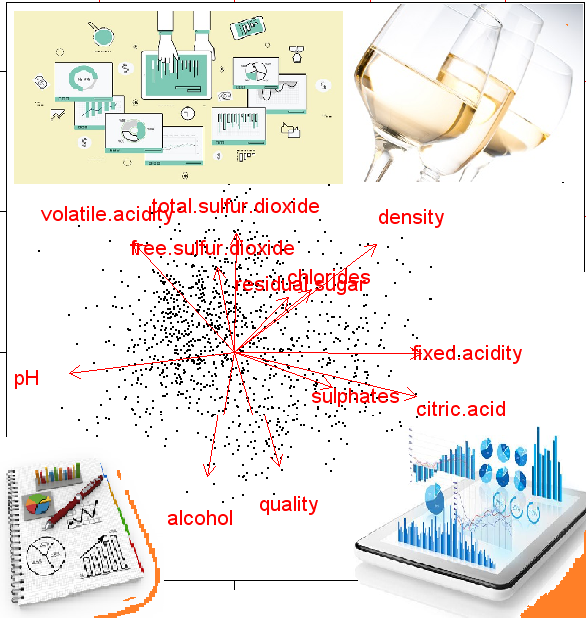
17/04/2019

|  |
| --- |
| מטלת מנחה 21 – פרוייקט גמר | חנן |



|  |  |
| --- | --- |
| חנן וולט | מטלת מנחה (ממ"ן) 21 – פרוייקט גמר |

תוכן עניינים

[1. **הגדרת הבעיה והכנת הנתונים** 2](#_Toc6437335)

[א. הגדרת מטרת כריית המידע 2](#_Toc6437336)

[ב. הגדרת הנתונים 2](#_Toc6437337)

[ג. הגדרת שלבי ה-**KDD** (**K**nowledge **D**iscovery in **D**atabases) 3](#_Toc6437338)

[ד. סקירת אלטרנטיבות שונות לביצוע כריית המידע 4](#_Toc6437339)

[ה. שלבי הכנת הנתונים 5](#_Toc6437340)

[2. **סיווג וחיזוי** 7](#_Toc6437341)

[א. בחירת שיטות לחיזוי הנתונים 7](#_Toc6437342)

[ב. תיאור שלבי החיזוי 7](#_Toc6437343)

[ג. דיווח תוצאות 8](#_Toc6437344)

[ד. הערכת מידת הדיוק 8](#_Toc6437345)

[ה. ניתוח והסקת מסקנות 8](#_Toc6437346)

# הגדרת הבעיה והכנת הנתונים

## הגדרת מטרת כריית המידע

ניתן לדרג את רמת האיכות של יין לפי מדדים שונים דוגמת גיל/שנת הבציר, טעם/ארומה, צבע, מידת חומציות ועוד. דרך נוספת, המסתמכת על מדדים אובייקטיביים יותר הינה שקלול תוצאות של בחינות כימיות-פיסיקליות שונות כגון מידת צפיפות וריכוז האלכוהול, סוכר שאריתי, ותרכובות וחומצות שונות דוגמת גופרית דו-חמצנית וחומצה ציטרית.

מטרת כריית המידע בפרוייקט זה היא לנבא את רמת האיכות של יין לבן על-סמך תכונות כימיות-פיסיקליות שכאלו. בכדי לממש מטרה זו, נבצע כריית מידע מתוך מאגר המכיל מדגם נתוני תצפיות של יינות לבנים עם המאפיינים הכימיים-פיסיקליים שלהם ורמת האיכות שלהם, ונבנה מודל שילמד מתוך מדגם זה כיצד לחזות את רמת האיכות של יינות שדירוגם אינו ידוע.

## הגדרת הנתונים

להלן הגדרות הנתונים המשמשים לחיזוי איכות היין בפרוייקט[[1]](#footnote-1) -

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| # | שם תכונה | תאור התכונה/מאפיין | סוג | תחום ערכים | ממוצע | סטיית  תקן | ערכים לא חוקיים |
| 1 | **Fixed acidity** | חומציות קבועה/יציבה | רציף | מ – 3.8 עד – 14.2 | 6.855 | 0.844 | יש ערכי קיצון/חריגים |
| 2 | **Volatile acidity** | חומציות נדיפה | רציף | מ – 0.08 עד – 1.1 | 0.278 | 0.101 | יש ערכים גבוהים מ-1. |
| 3 | **Citric acid** | חומצה לימונית | רציף | מ – 0 עד – 1.66 | 0.334 | 0.121 | יש ערכים גבוהים מ-1. |
| 4 | **Residual sugar** | סוכר שארי | רציף | מ– 0.6 עד– 65.8 | 6.391 | 5.072 | יש ערכי קיצון/חריגים |
| 5 | **Chlorides** | כלוריד | רציף | מ–0.009 עד–0.346 | 0.046 | 0.022 | יש ערכי קיצון/חריגים |
| 6 | **Free sulfur dioxide** | גופרית דו-חמצנית חופשית | רציף | מ–2 עד– 289 | 35.308 | 17.007 | יש ערכי קיצון/חריגים |
| 7 | **Total sulfur dioxide** | סה"כ גופרית דו-חמצנית | רציף | מ–9 עד– 440 | 138.361 | 42.498 | יש ערכי קיצון/חריגים |
| 8 | **Density** | צפיפות | רציף | מ–0.987 עד– 1.039 | 0.994 | 0.003 | יש ערכים גבוהים מ-1. |
| 9 | **pH** | רמת חומציות | רציף | מ–2.72 עד– 3.82 | 3.188 | 0.151 | אין |
| 10 | **Sulfates** | גופרה | רציף | מ–0.22 עד– 1.08 | 0.49 | 0.114 | יש ערכים גבוהים מ-1. |
| 11 | **Alcohol** | אלכוהול | רציף | מ–8 עד– 14.2 | 10.514 | 1.231 | אין |

משתנה המטרה (פלט) אותו יש לנבא הינו אמת איכות היין שנעה בסקלה שבין 10 (איכות גבוהה ביותר) לבין 1 (איכות נמוכה ביותר).

המשתנים שצויינו עם ערכי קיצון/ערכים חריגים הם משתנים המכילים רשומות בהן הערכים בתכונות אלו סוטים באופן חריג/מובהק מממוצע הערכים לתכונה זו, קישור לקובץ המכיל פירוט מלא של רשומות אלו נמצא בהמשך.

ההנחה היא שכל התכונות המכילות אחוז אמורות להתפלג בין 0 ל-1, לכן ערכים שליליים או ערכים הגבוהים מ-1 אינם חוקיים.

טיפול בנתונים אלו מפורט בהמשך. מתוך סט נתונים של 4,898 רשומות סה"כ, לאחר ניקוי מותירים 4,690 רשומות "נקיות".

## הגדרת שלבי ה-**KDD** (**K**nowledge **D**iscovery in **D**atabases)

להלן שלבי גילוי המידע במאגרי הנתונים עבור הבעיה הנתונה –

1. **הגדרת מטרות כריית מידע**: חיזוי רמת איכות של יין לבן בהסתמך על מאפיינים כימיים/פיסיקליים.
2. **איסוף ושמירת הנתונים**: הנתונים הרלוונטיים לתהליך הכרייה כבר רוכזו בסט נתונים אחיד מיועד, כך שאין צורך באינטגרציה ממקורות שונים ו/או בחירת אוכלוסייה חלקית רלוונטית.
3. **ניקוי הנתונים**: גילוי וטיפול בנתונים רועשים (אין ערכים חסרים/לא חוקיים עבור הסט הנידון), לכן די לטפל ברשומות בעלי ערכי קיצון חריגים שפוגמות באיכות הנתונים ומשבשות את החיזוי, וזאת ניתן לעשות ע"י טכניקות שונות כגון החלקה/ הסרה.
4. **רדוקציה וטרנספורמציה של הנתונים**: בדיקת אפשרות לרדוקציה/כיווץ של היקף הנתונים ע"י סינון רשומות ותכונות שלא תורמות לתהליך (במידה ויש כאלו), וטרנספורמציה ע"פ הצורך ע"י טכניקות כמו נירמול ודיסקרטיזציה והגדרת תכונות חדשות (מחושבות) שיסייעו בתהליך.
5. **בחירת שיטות וכלים לביצוע תהליך כריית המידע**: שימוש בכלי התוכנה Weka עם שיטות הרלוונטיות לחיזוי נתונים שהכלי תומך בהם כגון עצי-החלטה ורגרסיה לינארית.
6. **התאמת ייצוג הנתונים אד-הוק לשיטת כריית מידע הנבחרת**: שיטות מסויימות של כריית מידע עשויות להכיל אילוץ על אופן הנתונים שבהן היא תומכת/מטפלת. למשל שיטת רגרסיה לינארית מתאימה לנתונים נומריים רציפים, ואילו עצי-החלטה מסויימים מתאימים דווקא יותר לנתונים דיסקרטיים וחלקם אף דורשים נתונים בינריים/בוליאניים. לפיכך בשלב זה, נבצע טרנספורמציה ודיסקרטיזציה נוספת ע"פ הצורך כדי להתאים את ייצוג הנתונים בהתאם לאילוצי השיטות שנבחרו.
7. **ביצוע תהליך כריית המידע בפועל**: הרצת השיטות/אלגוריתמים שנבחרו על הנתונים המטוייבים בעזרת כלי התוכנה Weka.
8. **ניתוח התוצאות**: בדיקת ממצאי השיטות/אלגוריתמים שהורצו והערכתן ע"י בדיקת מידת הדיוק, רלוונטיות, פשטות וכד', ע"י שימוש בדו"חות נתונים סטטיסטיים ביצועיים וויזואליזציה גראפית של המודלים/התפלגות נתונים.

עם ניתוח התוצאות, בהתאם למידת שביעות רצון מהמודל שהופק, ייתכנו 2 אפשרויות מרכזיות–

1. אם המודלים לא יוכיחו את עצמם כמשביעי רצון מספק ניתן יהיה לחזור שוב ושוב על התהליך הר"מ ולנסות לבצע את הדברים קצת אחרת במטרה להביא למודל טוב יותר, למשל ע"י בחירות שיטות אלטרנטיביות לכריית המידע / טרנספורמציה לייצוגים אחרים, ו/או לקחת בחשבון רשומות/מאפיינים שקודם סוננו ועוד.
2. אם המודלים כן יוכיחו את עצמם כמשביעי רצון מספק, ניתן יהיה להתקדם לשלב הבא של הסקת מסקנות לתהליך קבלת ההחלטות.
3. **הסקת מסקנות**: ייצוא המודל שהתקבל מתהליך כריית המידע ככלי לחיזוי רמת האיכות ברשומות חדשות עתידיות של יין לבן. בהתאם לשיטה הנבחרת, את המודל ניתן לייצא ולייצג בדמות ויזואלית, נוסחה מתמטית, כללי היסק, עץ החלטה ואף קוד מקור של האלגוריתם.

## סקירת אלטרנטיבות שונות לביצוע כריית המידע

להלן סקירה השוואתית בין אלטרנטיבות שונות של שיטות/אלגוריתמים לביצוע כריית המידע –

1. **עץ החלטה** **ID3 מבוסס Information Gain:** שיטה לסיווג/חיזוי ערכים דיסקרטיים.

* **אילוץ**: רלוונטי למאפיינים דסיקרטיים בלבד (יש לבצע דיסקרטיזציה לערכים רציפים לטווחים/קטגוריות...).
* **מדד פיצול**: Information Gain – רווח אינפורמטיבי. תחילה מחושבת האנטרופיה של מחלקת היעד המבוקשת לסיווג/וחיזוי, ובכל רמה בעץ נבחרת התכונה בעלת האנטרופיה המותנית הנמוכה ביותר, כלומר זו שתביא לרווח האינופרמטיבי המירבי לסיווג/חיזוי משתנה המטרה בהתאם לידיעת הערך של תכונת הפיצול.
* **יתרונות**: אלגוריתם פשוט ויעיל, מבחינת ביצועי זמן ריצה.
* **חסרונות**: המדד המשמש לפיצול מוטה לבחירת תכונות בעלות מספר רב של ערכים אפשריים תכונות עם טווחים רציפים,שאולי באופן מקומי תורמות רווח אינפורמטיבי מירבי, אך במבט כולל יוצרות התאמת יתר Overfitting ואינן תורמות לדיוק הסיווג/חיזוי.

1. **עץ החלטה** **C4.5 מבוסס Gain Ratio:** שיטה זו מתאימה לסיווג/חיזוי ערכים דיסקרטיים.

* **מדד פיצול**: Gain Ratio – *יחס* הרווח האינפורמטיבי. שיטה זו מנסה לשפר את שיטת עץ ID3 ע"י הרחבת מדד הרווח האינפורמטיבי כך שייקח בחשבון את החלק היחסי של הרווח האינפורמטיבי המושג מפיצול לפי תכונה מסויימת ביחס לכלל אוכלוסיית נתוני האימון, ולא רק לתת-האוכלוסייה שתיוותר לאחר הפיצול, וע"י כך שואפת להתגבר על חיסרון ההטייה שמדד רווח אינפורמטיבי סובל ממנו.
* **יתרונות**: צפוי להביא לתוצאות טובות יותר עבור אוכלוסייה המכילה תכונות מרובות ערכים/תכונות המחולקות למקטעי טווח רציפים, וכן בעל יכולות גיזום והתמודדות עם ערכים חסרים ורציפים (במידה והם מחולקים למקטעים).
* **חסרונות**: המדד המשמש לפיצול מוטה לבחירת תכונות המביאות לחלוקה לא מאוזנת שיוצרות ענפים הקטנים משמעותית מכל היתר.

1. **עץ החלטה** **CART מבוסס Gini Index:** שיטה זו מתאימה לסיווג/חיזוי הן ערכים דיסקרטיים והן רציפים (ע"י מציאת נקודות פיצול בין הערכים).

* **אילוץ**: רלוונטי למאפיינים דיסקרטיים בלבד (יש לבצע דיסקרטיזציה לערכים רציפים כל שניתן לקבוע נקודות חלוקה ביניהן).
* **מדד פיצול**: Gini Index – מדד ג'יני. יחס רווח אינפורמטיבי. שיטה זו מבצעת פיצולים בינריים, בדהיינו כל חלוקה היא בדיוק ל-2 קבוצות. בכל שלב, מתוך כלל הצירופים האפשריים של חלוקת האוכלוסייה ל-2 תתי-קבוצות לא ריקות לפי חלוקת הערכים באחת מהתכונות, האלגוריתם בוחר את הפיצול שיביא לרמת רעש/אי-דיוק מינימלית.
* **יתרונות**: מנתאים במיוחד למקרים בהם נדרש פלט בוליאני: כן/לא. שיטה זו כוללת אלגוריתם לגיזום העץ. בהשוואה אמפירית לעצי החלטה אחרים עץ זה נוטה לספק את מודל מדוייק וקומפקטי יותר.
* **חסרונות**: בדומה למדד רווח אינפורמטיבי שבשימוש ID3**,** גם מדד ג'יני מוטה לבחירת תכונות בעלות מספר רב של ערכים אפשריים. כמו כן, הוא נוטה להעדיף פיצולים שיביאו לחלוקות מאוזנות שיביאו לידי רמת רעש שווה פחות או יותר בכל תת-קבוצה, ומתקשה להתמודד עם סט נתונים המכיל מספר רב של מאפיינים.

כמו כן, האלגוריתם של CART משלם על הדיוק וקומפקטיות בביצועי זמן ריצה.

1. **רגרסיה לינארית:** שיטה זו מתאימה לחיזוי ערכים רציפים, במקרה שבו קיימת מידת התאמה לינארית מסויימת בין משתנה המטרה לבין תכונה אחת או יותר בוקטור הרשומה.

* **אילוץ**: רלוונטי רק למאפיינים רציפים.
* **יתרונות**: השיטה ניתנת להמחשה גרפית ישירה, כך שבמידה ואכן קיים קשר לינארי חזק דיו הצגת הרגרסיה מאפשרת לדרג מקבלי ההחלטות לזהות מגמות ולהעריך ציפיות בקלות על סמך תצוגה ויזואלית פשוטה.
* **חסרונות**: השיטה מתאימה רק בהנחה שאכן קיים קשר לינארי במידה די מספקת. במקרים בהם מגמת הפיזור לא אחידה המודל ייתכן והמודל ימצא רק רמות מתאם (קורולציה) נמוכות מאוד, מה שיביא למודל חלש מאוד מבחינת ביצועי דיוק התצפית/חיזוי שאינו מעשי, ויחייב בחינת מודלי רגרסיה מורכבים יותר (כגון חיפוש קירובים לפולינומים).

## שלבי הכנת הנתונים

נתאר את שלבי הכנת נתונים (Data Preprocessing) על סמך סט הנתונים הגולמיים (Raw Dara) –

1. **ניקוי הנתונים**

ראשית נטפל בניקוי הנתונים, דהיינו טיפול בערכים חסרים וערכים חריגים/קיצוניים. כאמור בשלב 1ב, בסט הנתונים הקיים אין ערכים חסרים. נבדוק אם קיימים ערכים חריגים/קיצוניים, נעשה זאת ע"י שימוש בסולם מדידה של [טווח בין-רבעוני (IQR – Interquartile Range)](https://he.wikipedia.org/wiki/%D7%98%D7%95%D7%95%D7%97_%D7%91%D7%99%D7%9F-%D7%A8%D7%91%D7%A2%D7%95%D7%A0%D7%99) . ניישם זאת ב-Weka עם פילטר למידה לא-מודרכת ברמת בתכונה, המסמן עבור כל רשומה/תכונה האם היא נחשבת ע"פ מדד זה כרשומה בעלת ערך קיצון/חריג. להלן קובץ שבו מסומנות באדום בהיר הרשומות בעלות ערך קיצוני/חריג, וסימון אדום כהה מסמן את המאפיינים בעלי ערכים אלו וכן האם השורה סומנה באדום בהיר בשל הקיצוניות/חריגות או גם וגם –



מתוך כלל 4,898 הרשומות שבסט הנתונים, ישנן רק 208 רשומות שלפחות אחד המאפיינים שבה מכיל ערך חריג/קיצוני, המהווים רק כ-4.2% מסט הנתונים כולו. בכדי שמיעוט זה של ערכים חריגים/קיצוניים לא יגרום לפגם בדיוק החיזוי, נסיר רשומות אלו, המהוות יחסית חלק די זניח, וניוותר עם 4,690 רשומות "נקיות":



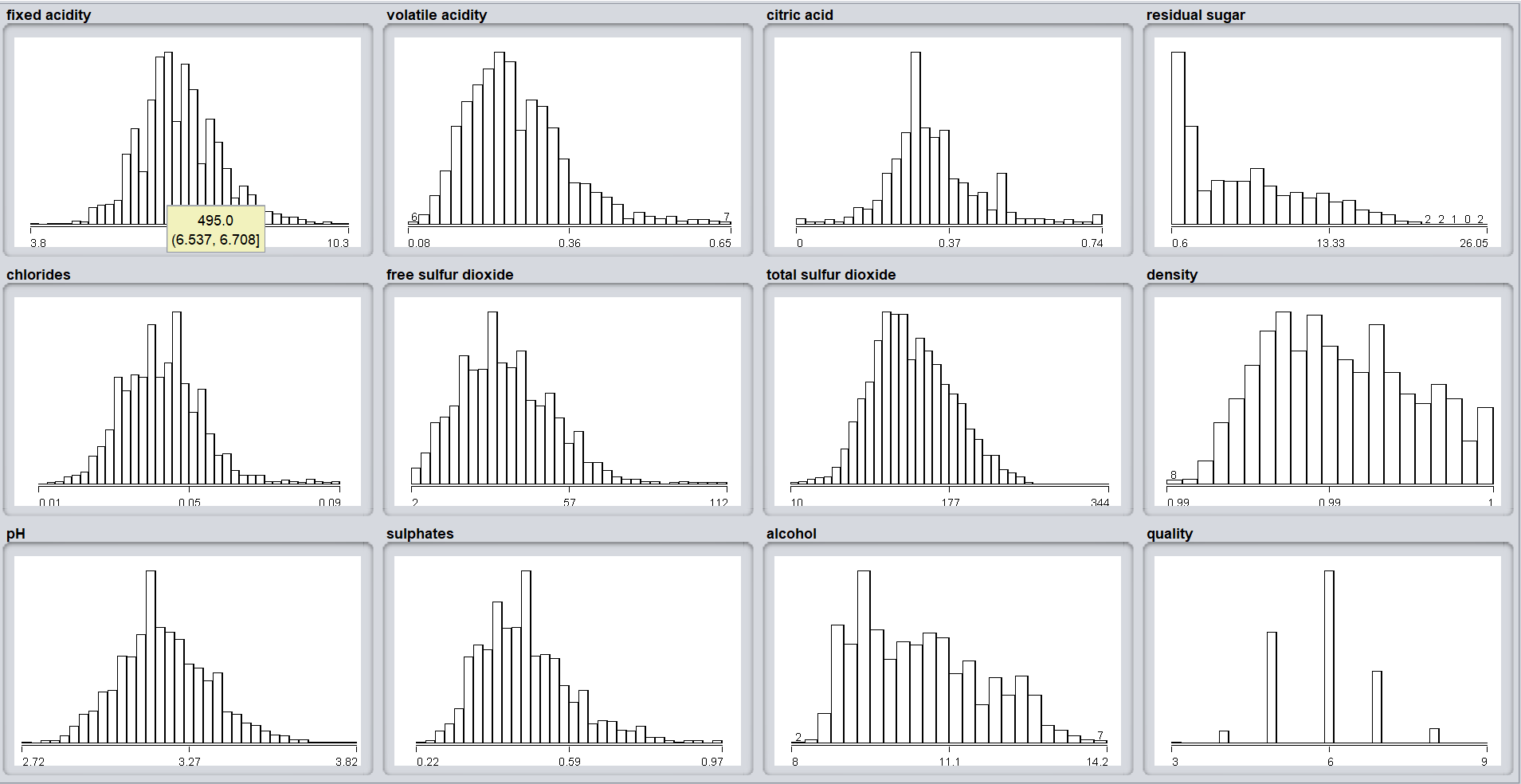
בבדיקה לאחר הניקוי האוטומטי של Weka, נית לראות שמלאכת "ביעור החמץ" עדיין לא תמה – ישנן 71 רשומות שלא הוסרו, שמכילות בתכונה Density (צפיפיות) ערך הגבוהה מ-1.000:



סביר להניח שמידת הצפיפות נורמלה לטווח ערכים בין 0.00 ל-1.00, לפיכך כל ערך שחורג מטווח זה, ובפרט 71 רשומות אלו מכילות ערך שאינו תקף. מבדיקה של רשומות אלו ניכר שהסטיות מינוריות (ברמת האלפית או פחות), לפיכך סביר גם להניח שמדובר בשגיאות מדידה קלות, על כן נעגל את הערך ברשומות אלו לערך 1.00. להלן סט הנתונים המעודכן לעת עתה (עדיין 4,690 רשומות) –



זהו המצב הנוכחי של התפלגות רמת האיכות ביחס לכל האחד מהתכונות האחרון באופן אינדיווידואלי –



1. **רדוקציה וטרנספורמציה של הנתונים**

במצב הנוכחי הנתונים עוד קשים לניתוח, שכן הם נמדדים בסקלות (סולמות מדידה) שונות. למשל, תכונת החומציות הקבועה (fixed acidity) נעה בין 3.8 לבין 10.3, בעוד שתכונת החומציות הנדיפה (volatile acidity) נעה בין 0 לבין 1, לפיכך בשלב ראשון, כדי להביא לקנה מידה אחיד, נבצע המרה וננרמל את כל תכונות כל הנתונים (להוציא את משתנה המטרה) כך שיתפלגו נורמלית סטנדרטית בין 0 ל-1. פעולה זו מצמצמת את היקף ייצוג הנתונים (רדוקציה) ואינה משפיעה על ההתפלגות. תוכן עדכני -

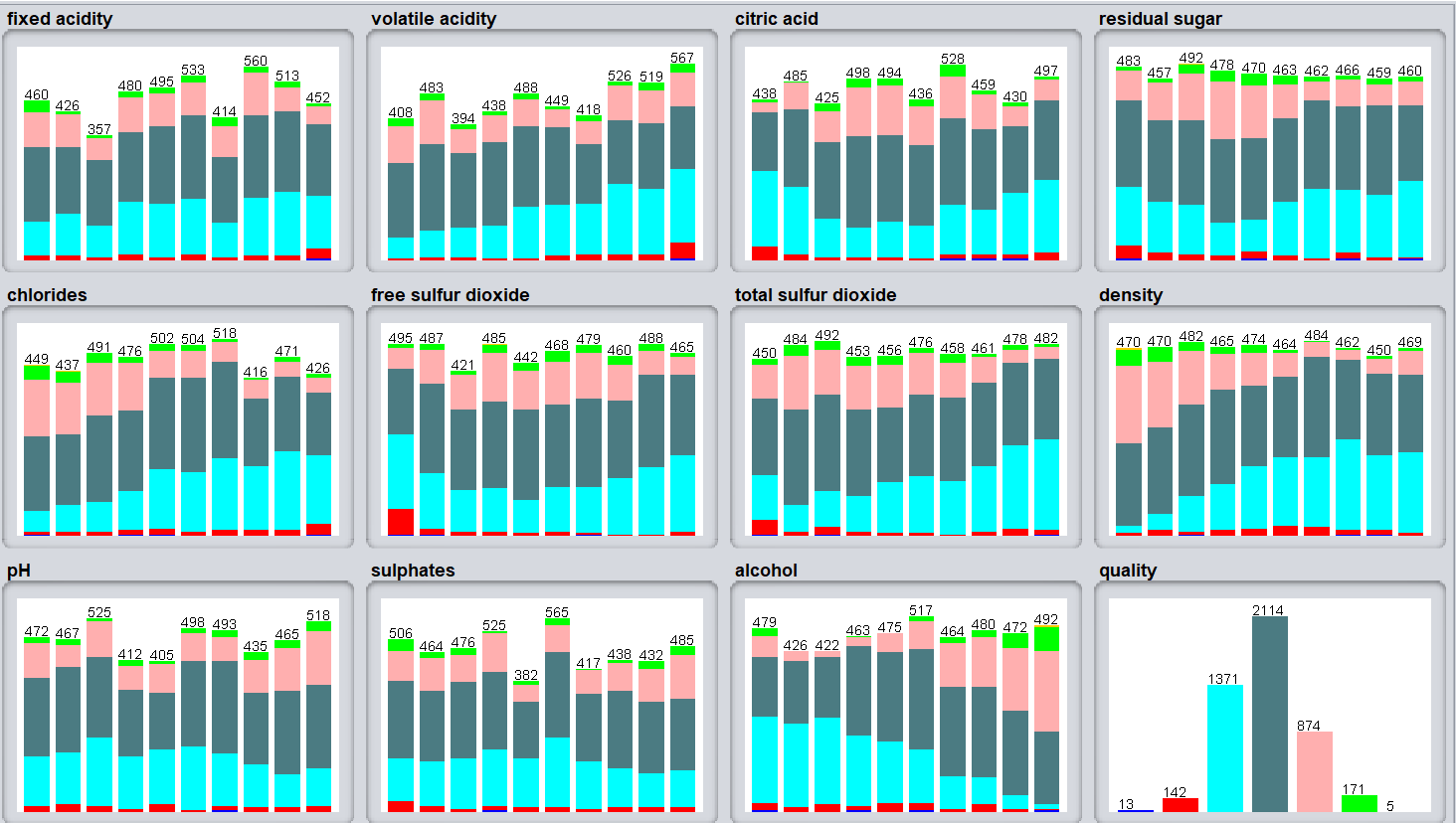


כעת הסקלה אחידה, אך יש בכל תכונה יש עשרות עד מאות ערכים שונים המתפלגים לאורך כל הסקלה הזו. מאחר שהערך הקונקרטי המסויים של כל מאפיין בכל רשומה פחות מעניין ופחות שימושי לחיזוי, ומה שמעניין באמת זה כיצד התפלגויות הערכים עשויה להשפיע על חיזוי משתנה המטרה, נבצע דיסקרטיזציה של הנתונים ע"י חלוקתם למחלקות סדרי גודל שונות, כאשר כל מחלקה תכיל שילוב של ריבוי רשומות הנמצאות בטווח ערכים מסויים, כלומר דיסקרטיזציה בשיטת Binning.

ניתן לבצע חלוקה *שוות רוחב (Equal-width)*, שבה הטווח מחולק לתתי-מקטעים שווי אורך, או חלוקת *שוות עומק/שכיחות (Equal-depth \ frequency)*, שבה הטווח מחולק אולי לתתי-מקטעים באורכים משתנים, אך בכל מקטע/תת-מחלקה יש פחות או יותר מספר מאוזן של רשומות.

החלוקה בשיטה שוות הרוחב פחות מתאימה לסט הנתונים הנתון, מכיוון שעבור מרבית התכונות ניתן להתרשם שההתפלגות לא אחידה, אלא מוטה, מה שעלול לגרום לכך שבחלוקה שוות רוחב עלול להינתן דגש יתר לרשומות חריגות. לפיכך נבצע חלוקה שוות עומק, לאחריה נקבל התפלגות פחות או יותר אחיד במחלקות מאוזנות פחות או יותר בכל תכונה.

כמו כן, כדי לקבל תמונת מצב בהירה יותר שתקל על הניתוח, נבצע דיסקרטיזציה של משתנה המטרה מערך נומרי לנומינלי (קטגוריאלי). תמונת מצב עדכנית של התפלגות רמת איכות היין הלבן ביחס לכל אחת מהתכונות בנפרד (הצבע בראש העמודות של כל היסטוגרמה מייצג את רמת האיכות הגבוהה יותר והצבע בתחתית העמודה את אמת האיכות הנמוכה ביותר) -



הנתונים המעובדים כעת נקיים ומסודרים וניתן להתקדם הלאה לשלב הבא. להלן סט הנתונים המטוייב –

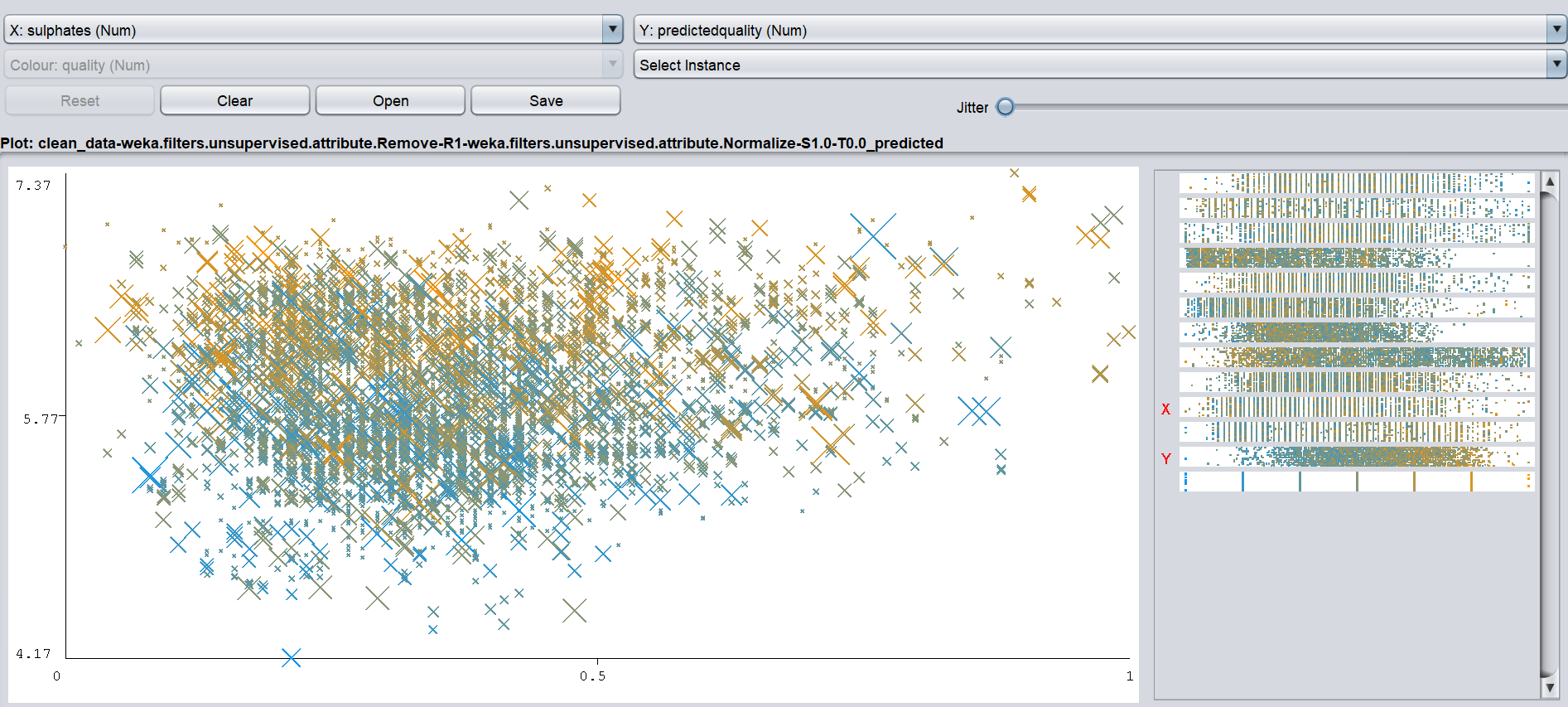


# סיווג וחיזוי

## בחירת שיטות לחיזוי הנתונים

נבחן את החלופות שהוזכרו לעיל לביצוע כריית מידע - עצי החלטה של ID3, C4.5 ו-CART, ורגרסיה לינארית. נתחיל דווקא באחרונה, רגרסיה לינארית:

כאמור לעיל, רגרסיה רלוונטית לערכים לנתונים נומריים רציפים, כפי שהיו בסט הנתונים טרם ביצוע הדיסקרטיזציה. אם נבחן את הנתונים הנקיים טרם ביצוע הדיסקרטיזציה וננסה לדלות איזשהו קשר לינארי בין ערכי המשתנים, נגלה כי אין כזה, הנתונים פשוט מפוזרים במעין "תוהו-ובוהו" ולא מלמדים על מגמה ברורה שתוכל להצביע על קשר לחיזוי רמת איכות היין על סמך חיוביות או שליליות יתר התכונות. למשל אם נסתכל על מידת חיזוי רמת האיכות הציר האנכי (ציר ה-Y) לעומת רמת הגופרה בציר האופקי (ציר ה- X) ניכר בבירור שאין מגמה ברורה ויש המון חריגות שיגרמו לסיווג רשומות באופן שגוי -



תמונת מצב דומה מתקבלת גם כאשר מנסים להציג את משתנה המטרה כפונקציה של יתר התכונות. במילים אחרות, מקדם המתאם הוא בעל מידת קשר עם ערך מוחלט חלש במיוחד (בערך 0.5), ולכן שיטת רגרסיה לינארית לא תוכל לתרום הרבה לבעיה הנידונה. לפיכך נבחר בשיטות של עצי החלטה: מבין שני עצי ההחלטה הראשונים שהוזכרו - ID3 עם מדד Information Gain ו-C4.5 עם מדד Gain Ratio נעדיף את C4.5, שכן כאמור לעיל הוא עוצב במטרה לשפר ולהתעלות על ביצועי הדיוק של ID3, לכן צפוי לתת ביצועים טובים יותר.

בנוסף ננסה גם את שיטת העץ השלישי של CART עם מדד ג'יני, בתקווה שגיזום ימנע התאמת-יתר ויביא לשיפור דיוק.

## תיאור שלבי החיזוי

1. **עץ C4.5** – בהחל צומת השורש הכולל את סט הנתונים כולו, בכל שלב מחושבים הן הרווח האינפורמטיבי של כל תכונה והן את מדד הפיצול שלה, כדי לבדוק איזו תכונה תביא ליחס רווח אינפורמטיבי מרבי (מקסימאלי), ולפיה מבצעים את הפיצול שיכול להיות פיצול בינרי או מרובה. ממשיכים בפיצול עד שכל הענפים הגיעו סווג/נחזו לאחד מערכי משתני החיד ע"י הגעה לעלה (צומת טרמינלי בודד) או עד אשר מגיעים לתנאי עצירה מוגדר מראש (למשל מספר עלים מינימלי), ומקבלים החלטה למשל לפי רוב הצמתים תחת ענף הפיצול האחרון.
2. **עץ CART**– החל מצומת השורש שכולל את סט הנתונים כולו, בכל שלב עוברים על התכונות ועל כל הערכים שלהם ומחשבים איזו תכונה עם איזו נקודת פיצול תביא לחלוקה הבינארית שתניב תתי-קבוצות בעלות חוסר דיוק מזערי (מינימאלי) ביחס להסתברות לחיזוי משתנה היעד. החלוקה היא תמיד בינארית ל-2 קבוצות לא-ריקות. הפיצול ממשיך באופן רקורסיבי על תתי-קבוצות הנתונים שנותרו לאחר הפיצול עד לחיזוי טרמינלי או תנאי עצירה כמתואר לעיל, ואז נבנית סדרת עצים מגוזמים ומבינהם נבחר עץ המהווה פשרה סבירה בין רמת דיוק רצויה לבין זמן ריצה יעיל תוך כדי הימנעות במידת האפשר מהתאמת יתר.

## דיווח תוצאות

בשל היקף הנתונים הרחב (4.7K רשומות) ב-2 השיטות ההרצות בוצעו עם גיזום, ותנאי סף של 10 רשומות לפחות בכל עלה טרמינלי, במטרה להביא לעץ שאינו גדול מידי מחד, אך עדיין מדוייק מספיק מאידך. להלן דיווח תוצאות הריצה המלא -



* **תוצאות הרצה של שיטת עץ C4.5[[2]](#footnote-2):** מתקבל עץ בעל 201 צמתים מתוכם 181 עלים.
* **תוצאות הרצה של שיטת עץ CART:** תקבל עץ בעל 107 צמתים מתוכם 54 עלים.

## הערכת מידת הדיוק

בשני האלגוריתמים ההערכה בוצעה עם שיטת k-fold-cross-validation עם פרמטר k=10:

10 פעמים מחלקים את סט הנתונים ל-10 חלקים שווים, כאשר בכל פעם 9/10 חלקים משמשים כנתוני אימון והחלק העשירי כנתוני מבחן להערכת המודל שנבנה. בסיום 10 ההרצות משקללים את התוצאות ומשווים מול המודל המקורי שנבנה על סמך האוכלוסייה במלואה.

שיטת Bootstrap פחות התאימה בשל היקף הרחב, ושיטת Holdout של חלוקת הנתונים לנתוני אימון (2/3( ונתוני מבחן (1/3) פחות התאימה בשל אופי התפלגות הנתונים שאינו שאינו אחיד, מה שעלול היה לגרום להתאמת יתר בהתאם לאוכלוסייה הנדגמת בנתוני אימון.

להלן מידת הדיוק (precision) המוגדרת כמספר היחסי של רשומות שסווגו נכונה מתוך כלל הרשומות (אלו שסווגו נכונה ואלו שסווגו באופן שגוי)  -

* **מידת דיוק בעץ C4.5**: 
* **מידת דיוק בעץ CART**: 

## ניתוח והסקת מסקנות

ראשית נבדוק האם המודלים בכלל תרמו מעבר לחוק הרוב (ה-Base line"").

רמת האיכות השכיחה ביותר בסט הנתונים הינו 6, המופיעה ב-2,114 רשומות מתוך 4,690, כך שהיא מהווה 45% מהכלל, לפיכך שני המודלים אכן עוברים את סף המינימום בכדי להועיל בשיפור חיזוי במעבר לברירת המחדל הטריוויאלית.

כמו כן, אם נבחן את שיטת OneR שעובדת לחיזוי/סיווג לפי תכונה בודדת שדיוקו מרבי מבין כל יתר התכונות, במקרה זה התכונה אלכוהול, נגלה שרמת הדיוק של המודלים שלנו עולה גם על שיטה בסיסת זו, שהינה בדיוק של 49.59%.

כעת להשוואה מבין 2 השיטות, שניהם הניבו עצים די גדולים עם מספר לא מבוטל של עלים, וניכר שניתן לצמצם עוד משמעותית את העץ ומספר העלים (למשל ע"י קביעת 20 רשומות במינימום לכל עלה) אולם יש כאן Tradeoff : ככל שנגדיל את מספר הרשומות בכל עלה, תוצאת החיזוי תהיה פחות מדוייקת שכן פיזור הרשומות בכל רמה יגדל, מה שיגרום לירידה באיכות הדיוק. מאחר שממילא הדיוק בשני המודלים לא גבוה מידי, המצב הנוכחי מהווה פשרה מספקת ונעדיף את המודל של CART שכן הוא גם מניב עץ חסכוני יותר וגם מניב רמת דיוק גבוהה יותר.

ניתן לנקוט באחת מהשיטות לנסות לשפר את איכות בדיוק במעט, אך מאחר שהנתונים רועשים ואינם מאוזנים (המון יינות בינוניים ורק טיפה טובים או גרועים מאוד) השיפור יהיה די זניח.

1. Paulo Cortez, University of Minho, Guimarães, Portugal, <http://www3.dsi.uminho.pt/pcortez>   
   A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009 [↑](#footnote-ref-1)
2. מסווג כ- J48 ב-Weka, על שם [המימוש בשפת Java](http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/J48.html) [↑](#footnote-ref-2)