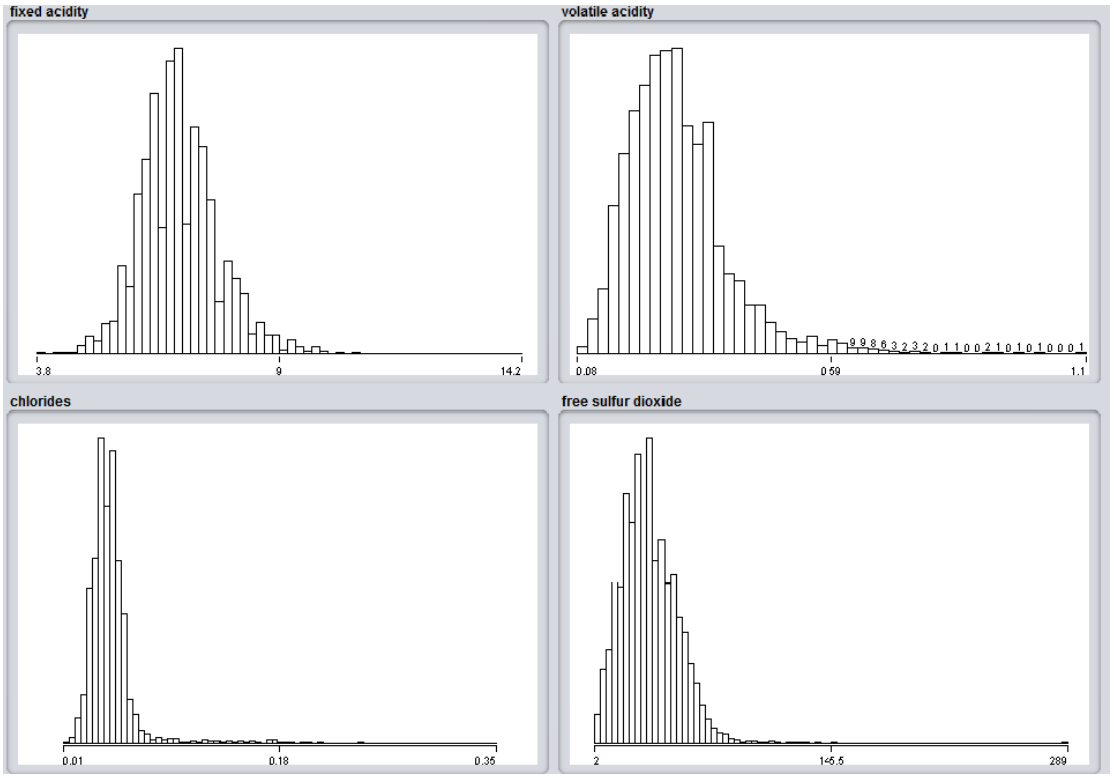
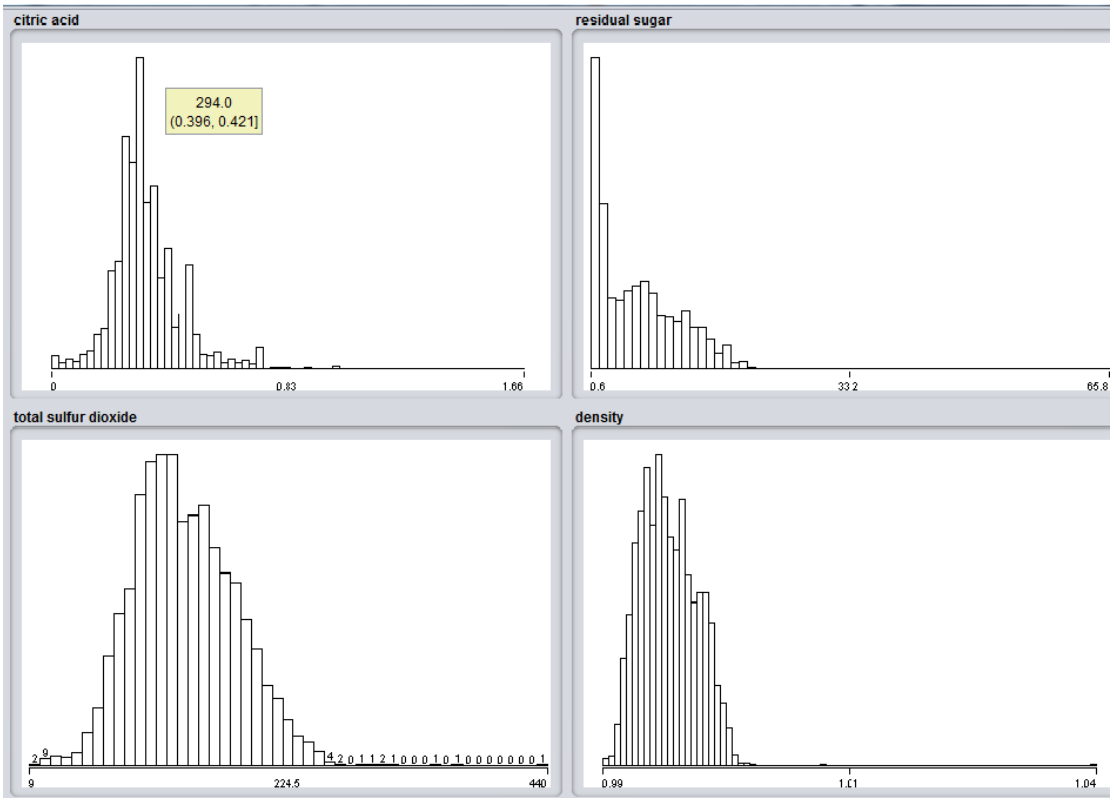
1.א מטרה – ניבוי איכות היין ע"פ נתונים שנמדדו ביקב הסוקרים ערכים ביין – חומציות, כלורידים, אלכוהול ,סוכרים ועוד.

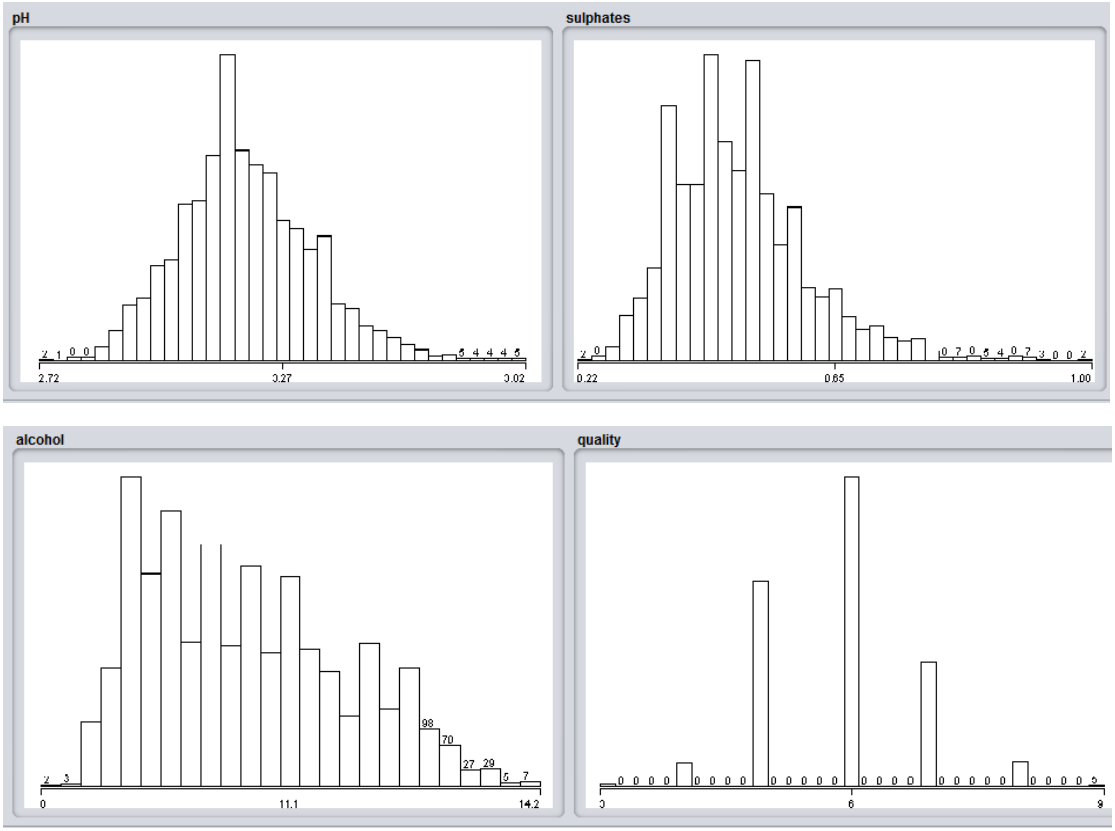
ב. הגדרת הנתונים.

שם	תיאור	סוג	יחידות מדידה	תחום ערכים	ממוצע	סטיית תקן
fixed acidity	חומציות קבועה	נומרי רציף	יחידות מדידה	4.45-9.65	6.855	0.844
volatile acidity	חומציות נדיפה	נומרי רציף	יחידות מדידה	0.137-0.593	0.278	0.101
citric acid	חומצה ציטרית	נומרי רציף	יחידות מדידה	0.074-0.666	0.334	0.121
residual sugar		נומרי רציף	יחידות מדידה	- 3.145 23.505	6.391	5.072
chlorides	כלורידים	נומרי רציף	יחידות מדידה	0.0173-0.0837	0.046	0.022
free sulfur dioxide	דת"ג חופשית	נומרי רציף	יחידות מדידה	13-101	35.308	17.007
total sulfur dioxide	סה"כ דת"ג	נומרי רציף	יחידות מדידה	43.4-310.6	138.361	42.498
density	צמיגות	נומרי רציף	יחידות מדידה	0.988-1.001	0.994	0.003
pH	ערך הגבה	נומרי רציף	יחידות מדידה	2.83-3.71	3.188	0.151
sulphates	סולפטים	נומרי רציף	יחידות מדידה	0.295-0.895	0.49	0.114
alcohol	אלכוהול	נומרי רציף	אחוזים	8.62-13.58	10.514	1.231
Quality	איכות	נומרי שלם	יחידות איכות	3-9	5.878	0.886

מבט על הערכים:







לגבי איכות הנתונים:

בהרצת פילטר IRQ נמצאו:

83 ערכי קיצון.

בנוסף 135 ערכים חריגים.

החלטתי להוריד את כל הערכים האלה מכיוון שנראה שהם ערכים קיצוניים וחריגים מידי ויכולים להטות את התשובה ואת יכולת החיזוי הקיימת.

סה"כ קיבלנו 4690 ערכים תקינים.

ג. שלבי הKDD:

1.**איסוף ושמירת נתונים ומידע.**

בשלב זה, נבחר מה הוא סט הנתונים עליו נעבוד, לרוב, משיקולי יעילות וביצועים נעדיף סטים קטנים של מידע ומשיקולי איכות נעדיף סטים גדולים של מידע.

2.ניקוי הנתונים.

נחפש רשומות בעלות ערכים לא הגיוניים, לדוגמא ערכים מספריים חריגים לפי מינימום ומקסימום של השדה עבור שורות חריגות נחליט אם להשמיט אותן לגמרי (במידה ומדובר בכמות קטנה) , לתקן את הערכים או לשים שם ערך ריק.
אני ביצעתי זאת ע"י הסרת מידע וערכים לא חוקיים (שימוש ב-IQR filter) ואח"כ RemoveWithValue פילטר.

3. ביצוע טרנספורמציה על המידע.

4.בחירת שיטות לכריית המידע.

בשלב זה נחליט על סוג כריית המידע הנדרש לפרוייקט הספציפי הזה בין האפשרויות הקיימות של classification, prediction, clustering ועוד. בנוסף נבחר אלגוריתם אחד או כמה מתוך המשפחה שבחרנו לעיל ונבנה מודל בעזרתו. אני בחרתי להתמקד באלגוריתמים בעלי עצי החלטה.

5. ביצוע דיסקרטיזציה וסיווג הנתונים.

ביצעתי דיסקרטיזציה על הנתונים וסיווגם לפי איכות.

6**.הרצת שיטות לכריית המידע – שבחרנו.**

7**.ניתוח התוצאות.**

שאלה 2 סעיף ה'

8. הסקת מסקנות.

שאלה 2 סעיף ה'

ד. ישנן כמה אפשרויות לבחירת שיטות לכריית מידע:

- **רגרסיה לינארית/לא לינארית**
לפי מודל זה, נבנית נוסחא לניבוי של משתנה המטרה. שיטה זו מתאימה יותר למשתנים נומריים והיא מתאימה כאשר הקשר למשתנה המטרה הינו קשר לינארי.

- **רגרסיה לא לינארית**.

ברגרסיה לא לינארית, הקשר למשתנה המטרה הוא ברך כלל פולינומי.

במקרה שלנו לא ידוע על קשר לינארי וכן אין התפלגות נורמלית ולכן נעדיף להשתמש באלגוריתמים מבוססי עץ החלטה.

- **עץ החלטה Cart(Classification And Regression Tree**)
Cart הינו קבוצה של עצי החלטה המאפשר לבחור את המדד על פיו נבנה את העץ. האלגוריתם מבצע פיצולים בינאריים (בניגוד לJ48) ומשתמש במדד ג'יני לקביעת הפיצול ובנוסף מבצע גיזום למניעת התאמת יתר.

- **עץ החלטה C4.5**
מבוסס על אלגוריתם ID3:

  - בכל צומת נבחר מאפיין אשר מגדיל את הרווח האינפורמטיבי. המדד של רווח אינפורמטיבי מוטה כלפי ערכים רציפים ורבים.
  - שימוש במדד Gain Ratio, מדד שמאפשר להתגבר על בעיית ריבוי הערכים של משתנים, לעומת מדדי האנטרופיה או Information Gain ב-ID3, הנוטים לתת משקל יתר לתכונות מרובות ערכים.
  - אלגוריתם זה מכיל שלב גיזום כך שהוא נוטה לייצר עצים קטנים יותר מ-ID3 ומונע את בעיית ה overfitting ולכן הוא עדיף ונשתמש בו.

ה. **הכנת הנתונים:**

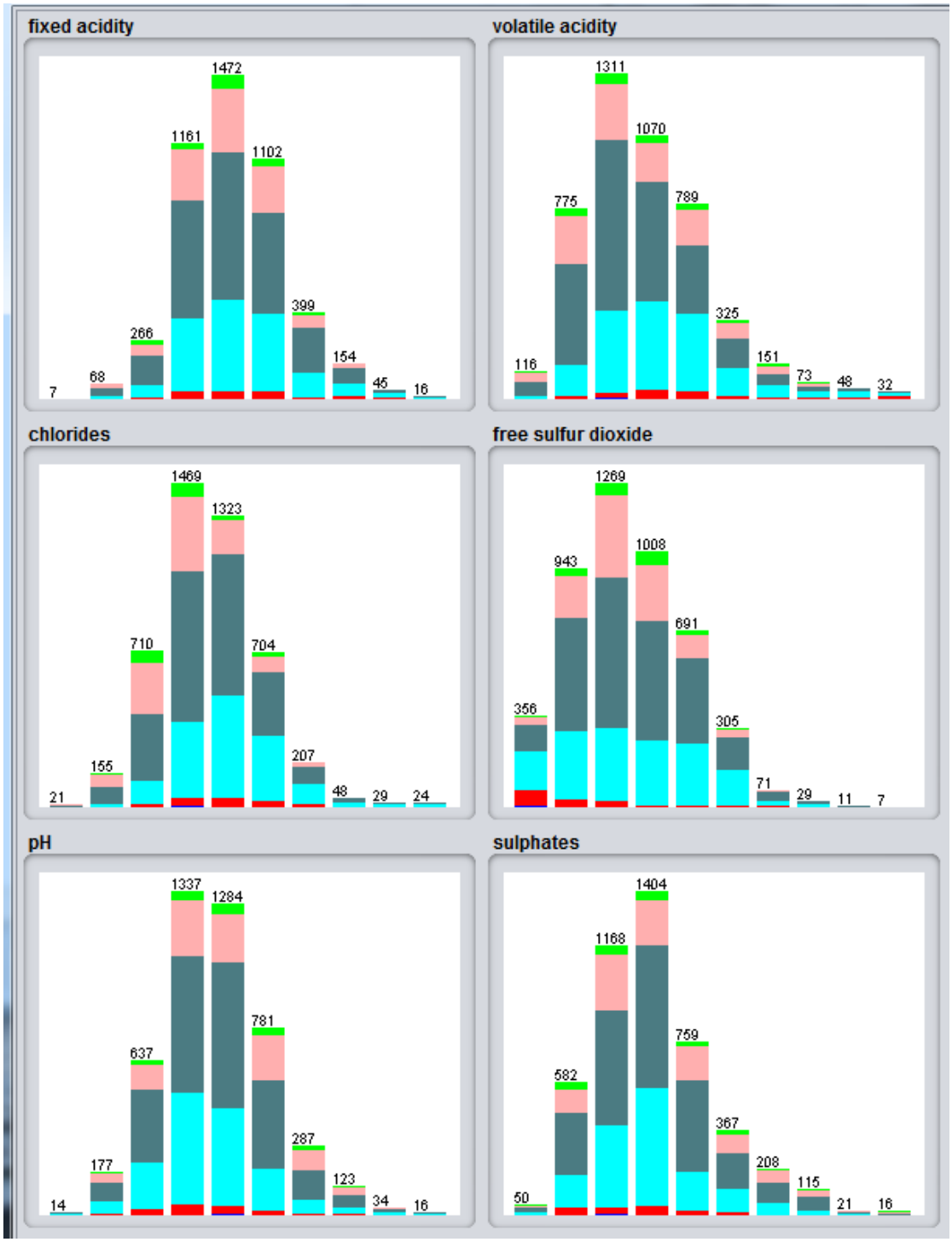
נמחקו 208 רשומות סה"כ המהווים 4.25% מכלל הרשומות.

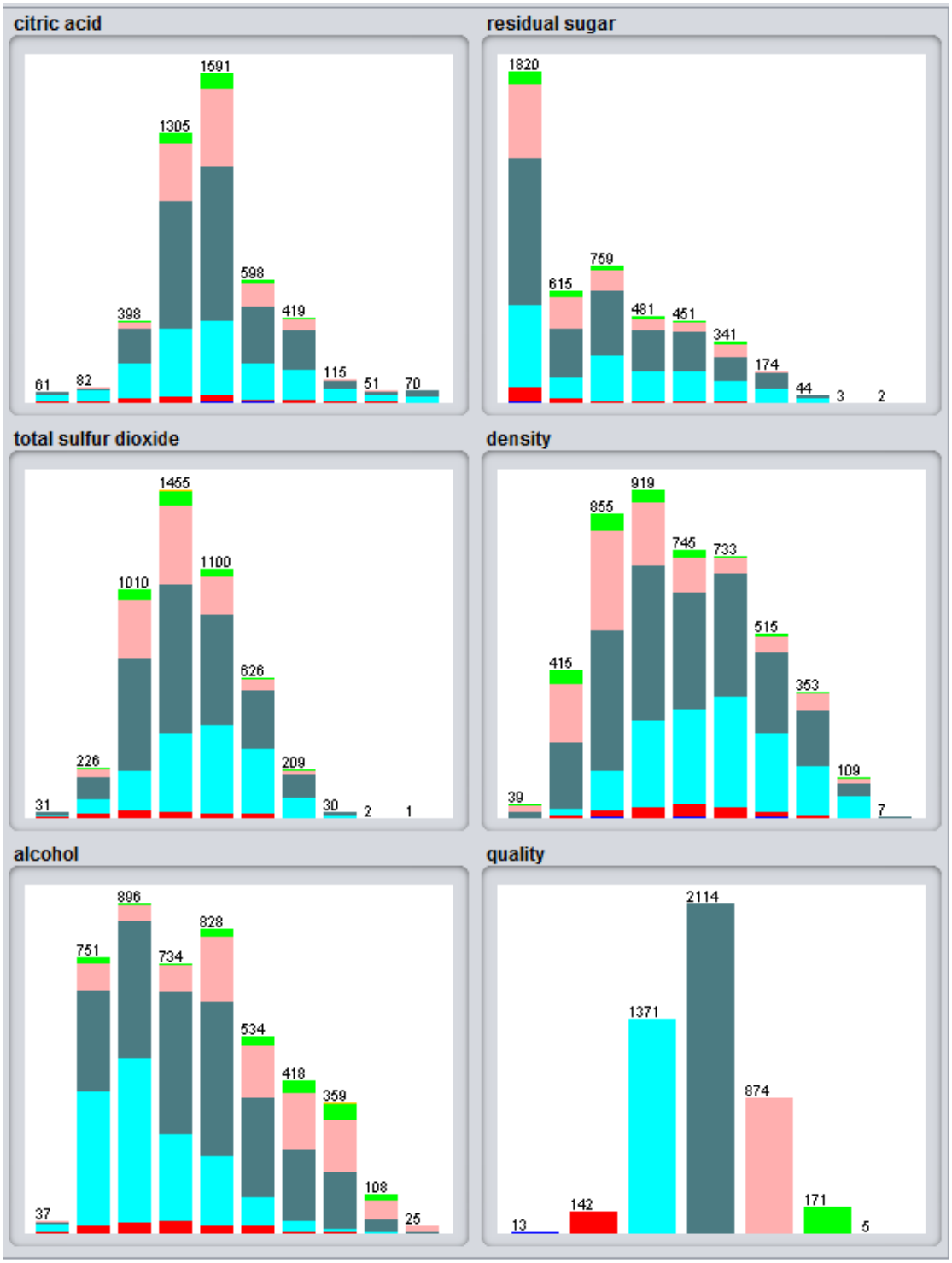
סיבות מחיקה:

ערכים חריגים מאוד או ערכים בלתי חוקיים.

אחרי המחיקה ביצעתי דיסקרטיזציה על כל התכונות למעט איכות, ועל האיכות ביצעתי המרה מנומרי לנומינלי לאחר מכן השתמשתי בפילטר resample כדי לקבל תוצאות טובות יותר.

כך שהנתונים נראים כך:


fixed acidity
7
68
266
1161
1472
1102
399
154
45
16
volatile acidity
116
775
1311
1070
789
325
151
73
48
32
chlorides
21
155
710
1469
1323
704
207
48
29
24
free sulfur dioxide
356
943
1269
1008
691
305
71
29
11
7
pH
14
177
637
1337
1284
781
287
123
34
16
sulphates
50
582
1168
1404
759
367
208
115
21
16



אח"כ נבצע resample בכדי שיעזור לאלגוריתמים מבוססי עץ החלטה שנממש בהמשך.

2.סיווג וחיזוי.

נבחר בשיטת CART ו-C4.5 כי לדעתי הם העדיפות במצב הנוכחי הדורש סיווג יין ע"פ הנתונים וחיזוי לאיכותו.

הסיבות להעדפת C 4.5 ו CART מפורטות בסעיף ד' של שאלה 1 ובקצרה – לא קיים קשר לינארי ובנוסף הפילוג אינו פילוג נורמלי.

שלבי C4.5 :

ריצה	מס'	Confidence Factor	מינימום עלים	Number of Leaves	Size of the tree	ROC Area	F-Measure	Recall	Precision	FP Rate	TP Rate

ראשית נגדיר את תנאי העצירה :

- כל הערכים בתת העץ שייכים לאותה קטגוריה
- אין מאפיין אשר נותן information gain

שלבי האלגוריתם :

1. התחל מהDATA SET המלא
2. אם הגענו לתנאי עצירה, עצור
3. לכל מאפיין a
   a. מצא את הinformation gain המנורמל בחלוקת הצומת הנוכחי ב a
4. קבע את a_best כמאפיין עם הinformation gain הגבוה ביותר
5. פצל את הצומת הנוכחי לפי a
6. הרץ שלבים 2-6 על הבנים של הצומת הנוכחי

:CART

שלבי CART דומים מאוד ל C4.5 עם כמה הבדלים מהותיים :

- הפיצולים הנעשים בCART הינם בינאריים לעומת C4.5 היודע לפצל ליותר מקבוצה אחת.
- CART משתמש במדד ג'יני במקום information gain בסעיפים 3,4 לעיל.

ג. ביצעתי את הריצות על הקובץ הנקי. בכל שיטה הרצתי מס' פעמים את האלגוריתם כדי לקבל תחושה של השפעת הפרמטרים על התוצאות.
בשתי השיטות השתמשתי בcross validation-Fold 10 לבדיקת המודל לפי גישה זו מחלקים את כל סט הנתונים ל-10 קבוצות ומריצים את אחד האלגוריתמים 10 פעמים, כאשר בכל פעם קבוצה אחרת מהווה TestSet ושאר הקבוצות – TrainingSet.להרצת C4.5 השתמשתי באלג' J48 בWeka ולהרצת CART השתמשתי בSimpleCart.

אלו התוצאות:

J48 (Unpruned)	1	0.25	200	4636	5151	0.878	0.767	0.768	0.119	0.768
J48 (Pruned)	2	0.25	200	3160	3511	0.86	0.723	0.726	0.147	0.726
SimpleCart (Pruned)	3		200	650	1299	0.852	0.734	0.737	0.134	0.737
SimpleCart (Unpruned)	4		200	702	1403	0.855	0.738	0.741	0.132	0.741

ד. על מנת להעריך את הדיוק ננתח את התוצאות שקיבלנו.

הפרמטרים החשובים לניתוח הם אלה שהובאו בטבלה, כך שהם מייצגים לנו את הדיוק בסיווג, את שלמותו, אחוז תצפיות שלילות שסווגו כשליליות ושקלול תמורות בין ה True Positive לבין הFalse Positive

לדעתי הנתון הכי חשוב לנו שמבטיח את נכונות הסיווג הוא ה Precision מכיוון שהוא מבטיח לנו את חיזוי איכות היין, כלומר ככל שהדיוק עולה ככה אחוז התצפיות השליליות – רוצה לומר יינות שאמורים להיות מסווגים באיכות מסוימת אך בסוף סווגו באיכות אחרת מהצפוי – קטן, מה שבעצם מאפשר לנו חיזוי טוב יותר וגידול ויישון היין בצורה טובה יותר להבטחת איכותו.

לאחר הגדרה זו ניתן להבחין כי:

J48 – בשיטה זו נראה כי הדיוק המרבי הושג ע"י ריצה 1, הריצה מוגדרת למינימום 200 תצפיות בעלה וללא גיזום.

CART – בשיטה זו הדיוק המרבי הושג ע"י ריצה מס 4 , הריצה מוגדרת למינימום 200 תצפיות ובעלה ללא גיזום.

ה. בשילוב עם הסעיף הקודם שבו ניתחנו חלקית נבצע ניתוח והסקת מסקנות:

א. ככל שיש יותר גיזום של העץ J48 מדויק פחות והעץ בCART כמעט ללא שינוי. הסבר אפשרי הוא כי העצים מבצעים גיזום בצורה שונה :

CART מבצע גיזום ע"י מודל סיבוכיות-עלות עם פרמטרים הנגזרים מהcross validation לעומת J48 המבצע מעבר יחיד על העץ וגוזם לפי איבוד אינפורמציה פוטנציאלי על פי הconfidence level המוגדר (יש לציין שביצעתי מספר ניסויים אמפירים במטרה לבדוק האם יש שינויים לטובה כתלות בCL או במינימום העלים – התוצאות העדיפות נכתבו). בנוסף נראה כי בJ48 לא מתקיים overfitting בגלל הדרישה לכמות סבירה על תצפיות בעלים ולכן איבוד אינפורמציה בגיזום רק פוגע במודל.

ב. חשוב להריץ את האלגוריתמים מס' פעמים עם פרמטרים שונים על מנת להבין יותר טוב את המידע איתו אנו עובדים שכן לפעמים האלגוריתמים יעבדו בצורה שונה ממה שנצפה מהם ומשחק עם הפרמטרים יכול לעזור לקבל הבנה טובה יותר לגבי ה data set שלנו.

ג. בנוסף יש להבחין כי ישנם פרמטרים נוספים שחשובים כמו TP אבסולוטי – במקרה שלנו זה גם הולך יחד – אבל לעיתים זה שיש לי יכולת טובה ומדויקת לחיזוי לא תעזור לנו אם לא יהיו לנו מספיק נתונים ככה שלא נוכל להרכיב "מתכון" ליין באיכות גבוהה.

הצעות לשיפור:

לדעתי ניתן לצמצם את כמות קבוצות האיכות של היין בכדי לקבל חיזוי טוב יותר, כך שאם נצמצם ל4 קבוצות – גרוע, בינוני, טוב, מצוין – דבר כזה יאפשר צורת חיזוי מדויקת יותר מכיוון שאותם הנתונים ישמשו לפחות קבוצות לסיווג האיכות.