**ממן 21: כריית מידע**

**הגדרת הבעיה והכנת הנתונים:**

1. מטרות כריית המידע: לנבא את איכותו של היין הלבן ע"פ הנתונים הפיזיים שנמדדו (12 תכונות)
2. הגדרת הנתונים בהם השתמשתי:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **שם תכונה** | **תאור התכונה** | **סוג** | **תחום ערכים** | **ממוצע** | **סטיית**  **תקן** | **ערכים לא חוקיים** |
| **fixed acidity** | חומציות קבוע | נומרי - רציף | מ-(3.8)  עד-(14.2) | 6.855 | 0.844 | אין |
| **volatile acidity** | חומציות נדיפה | נומרי - רציף | מ-(0.08)  עד-(1.1) | 0.278 | 0.101 | אין |
| **citric acid** | חומצת לימון | נומרי - רציף | מ-(0)  עד-(1.66) | 0.334 | 0.121 | אין |
| **residual sugar** | סוכר שיורי | נומרי - רציף | מ-(0.6)  עד-(65.8) | 6.391 | 5.072 | אין |
| **chlorides** | כלורידים | נומרי - רציף | מ-(0.009)  עד-(0.346) | 0.046 | 0.022 | אין |
| **free sulfur dioxide** | דו תחמוצת הגופרית | נומרי - שלם | מ-(2)  עד-(289) | 35.308 | 17.007 | אין |
| **total sulfur dioxide** | סה"כ דו תחמוצת הגופרית | נומרי - שלם | מ-(9)  עד-(440) | 138.361 | 42.498 | אין |
| **density** | צפיפות | נומרי - רציף | מ-(0.987)  עד-(1.039) | 0.994 | 0.003 | אין |
| **pH** | חומציות | נומרי - רציף | מ-(2.72)  עד-(3.82) | 3.188 | 0.151 | אין |
| **sulphates** | סולפטים | נומרי - רציף | מ-(0.22)  עד-(1.08) | 0.49 | 0.114 | אין |
| **alcohol** | אחוז אלכוהול | נומרי - רציף | מ-(8)  עד-(14.2) | 10.514 | 1.231 | אין |
| **quality** | איכות היין | נומרי – שלם | מ-(1)  עד-(10) ערכים תקינים, בסט הנתונים שלנו הטווח הוא מ-(3) עד (9) | 5.878 | 0.886 | אין |

1. **הגדרת ותיאור שלבי ה-KDD עבור הבעיה:**
2. איסוף ושמירת הנתונים-
   1. בשלב הראשון, נצטרך לבחור מה הוא סט הנתונים עליו אנו עובדים.
   2. הנתונים התקבלו מ- <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/>
   3. ביצעתי המרה לקובץ כדי שיתמוך בפורמט CSV לשם הזנה ושמירה נוחה בכלי העבודה EXCEL ו-WEKA.
3. ניקוי הנתונים –
   1. בשלב זה נעבור על כלל הנתונים, ונחפש רשומות בעלות ערכים חסרים או 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.
   2. בסט הנתונים שהתקבל לא היו ערכים חסרים או ערכים לא חוקיים. הסרתי נתונים חריגים וקיצוניים
4. ביצוע טרנספורמציות על הנתונים –
   1. בשלב זה נבחן האם ניתן לבצע שינוים על המידע אשר יתרמו למודל בין אם בשינויי המאפיינים הקיימים או ביצירת מאפיינים חדשים.
5. בחירת שיטות לכריית מידע (כגון רגרסיה, עצי החלטה וכו)
   1. בשלב זה, נבחן את האפשרויות השונות, מכיוון שישנן מספר משפחות אלגוריתמים שונות המבצעות דברים שונים, ולכל אלגוריתם יכולת אחרת כך שהוא יודע למקסם דברים אחרים.
6. ביצוע דיסקרטיזציות וסיווג הנתונים
   1. נבחן האם ישנן עמודות עליהן ניתן לבצע דיסקרטיזציה על מנת מקסום המידע ובניית מודל טוב יותר, בהתאם למטרת כריית המידע.
7. הרצת שיטות לכריית מידע שנבחרו
   1. על פי כמות הנתונים שברשותינו, נחלק את המידע ל-trining ו-test. נריץ את האלגוריתמים שנבחר על קבוצת ה-training ונבחן את הדיוק על קבוצת ה-test. את האלגוריתם נריץ מספר פעמים עם קונפיגורציה שונה כך שנוכל למצוא את ההתנהגות האופטימלית שלו והמדוייקת ביותר על המידע שלנו.
8. ניתוח התוצאות והסקת מסקנות:
   1. נבחן את התוצאות של האלגוריתמים השונים כדי להבין מאיזה אלגוריתם (ומאיזו קונפיגורציה) התקבל המודל המדוייק ביותר. בעזרת מודל זה נוכל לבחון כל שורת נתונים חדשה שתתקבל ונוכל להבין, במקרה שלנו, האם היין הלבן (רשומת המידע שהתקבלה) יהיה בעל איכות נמוכה, בינונית או גבוהה (בהתאם לדיסקרטיזציה שנבצע על עמודת ה-quality - ספויילר)
9. סקירה השוואתית **לכלל החלופות האפשריות לביצוע כריית מידע**:

נציג ארבע אפשרויות:

1. **רגרסיה לינארית מרובת משתנים**

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

1. **רגרסיה לא לינארית**

כמו רגרסיה לינארית, אך משתמשים בשיטה זו כאשר הנתונים לא מתנהגים לפי מודל לינארי, אלא לפי מודל אחר, למשל פולינומי.

1. **עץ החלטה Cart(Classification And Regression Tree**)

Cart הינו קבוצה של עצי החלטה המאפשר לבחור את המדד על פיו נבנה את העץ. במקרה זה ננתח עץ החלטה מבוסס מדד ג'יני. האלגוריתם מבצע פיצולים בינאריים (בניגוד לJ48) וגיזום למניעת התאמת יתר.

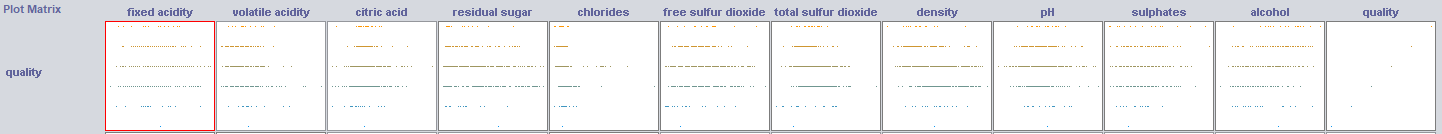
1. **עץ החלטה J48 – (C4.5)**

C4.5 הינו אלגוריתם לבניית עץ החלטה המשפר את אלגוריתם ID3. עץ זה, בשונה מID3 מבצע Pruning ע"י כך שמנסה להחליף צמתים בעלים ולראות האם ישנה פגיעה משמעותית. כמו כן, עץ זה תומך בערכים חסרים. לאור 2 שיפורים אלו, נעדיף את C4.5 על ID3.

1. **תיאור שלבי הכנת הנתונים:**

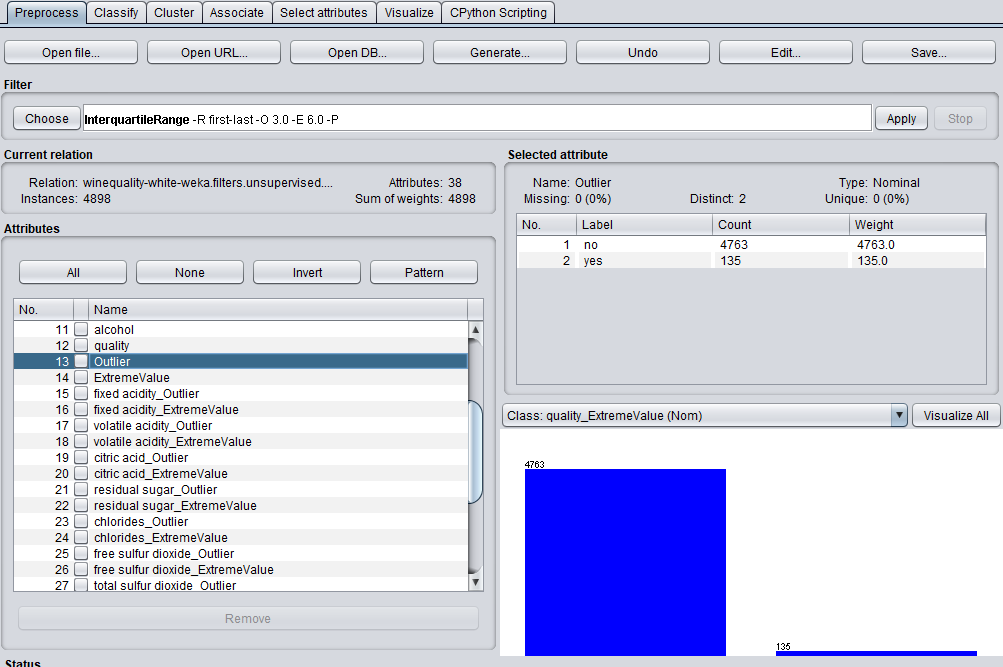
הצגה גרפית של הנתונים ההתחלתיים:

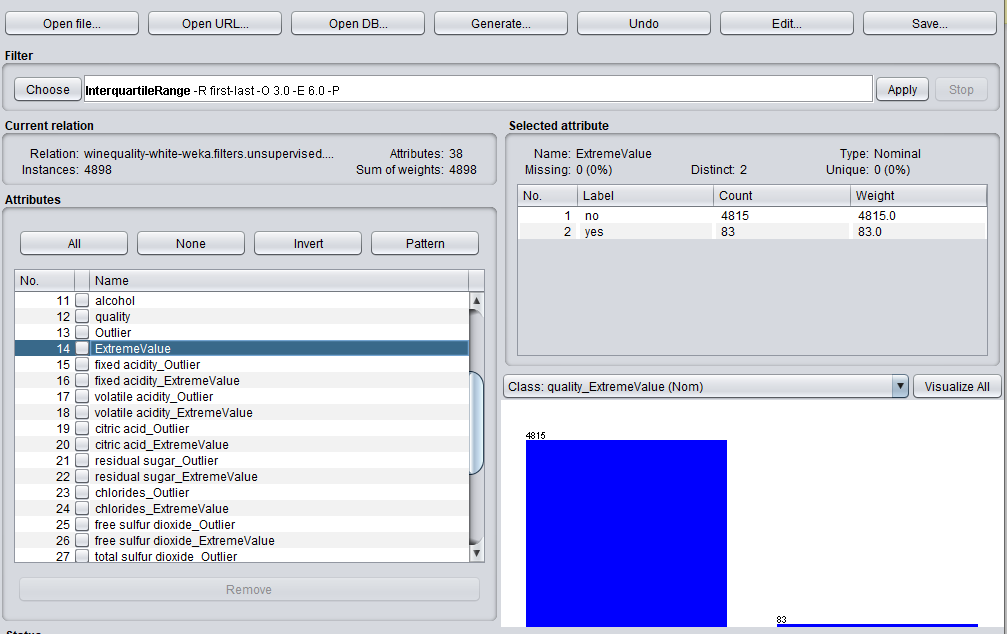
|  |  |
| --- | --- |
| שם תכונה | התפלגות הנתונים |
| fixed acidity |  |
| volatile acidity |  |
| citric acid |  |
| residual sugar |  |
| Chlorides |  |
| free sulfur dioxide |  |
| total sulfur dioxide |  |
| Density |  |
| pH |  |
| Sulphates |  |
| alcohol |  |
| Quality |  |

הצגה גרפית של "איכות היין" כפונקציה של כל אחת מהתכונות:

איכות הנתונים והכנתם:

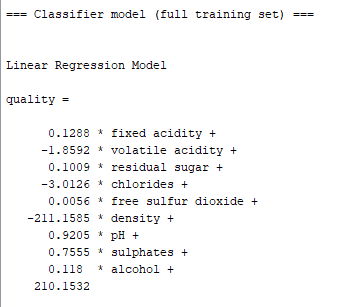
* כל הנתונים שהגיעו תקינים. לא קיימים ערכים חסרים וכמו כן אין הגבלות על נתונים מסויימים.
* נשתמש בפילטר בתוכנת WEKA על מנת למצוא נתונים חריגים. נסיר אותם בהתאם:

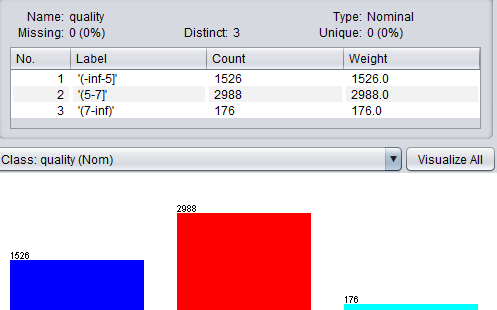


כמו כן נסיר גם את ערכי הקיצון: 

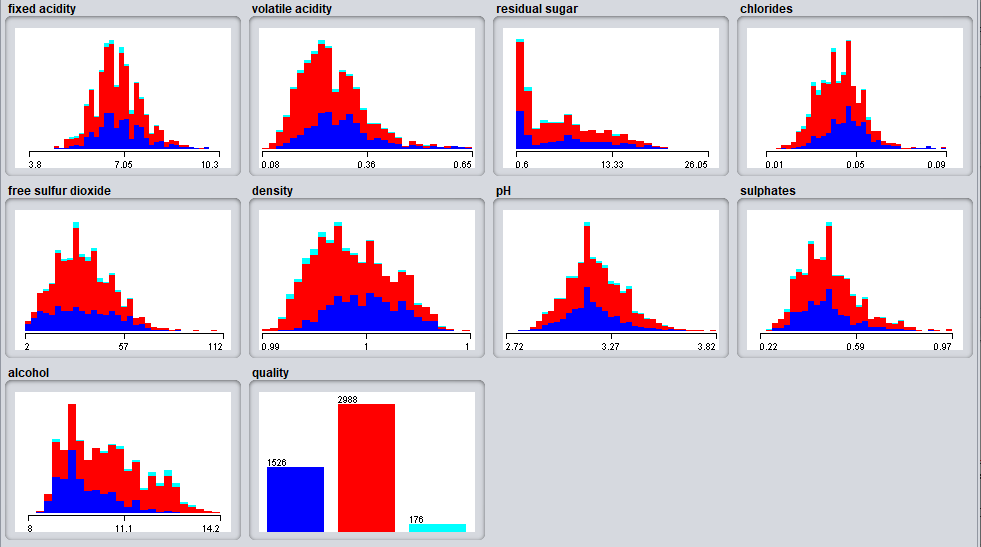
**לפני ניקוי הנתונים: סכ"ה 4898 רשומות.**

**לאחר הניקוי קיימות סכ"ה 4690 רשומות.**

* סינון עמודות:
  + הרצתי אלגוריתמים שונים למציאת עמודות זניחות (ע"י FETURE SELECTION בתוכנת WEKA). לא מצאתי עמודות חד משמעיות שניתן להסיר.
  + הרצתי רגרסיה לינארית מרובת משתנים, והתוצר הסופי השמיט 2 עמודות אותן מצא כזניחות:
  + 
  + לכן על פי זה השמטתי את העמודות total sulfur dioxide, citric acid .
* ביצוע דיסקרטיזציות וסיווג הנתונים:
  + נחלק את מאפיין היעד quality ל-3 תחומים, באמצעות binning של עומק שווה, כך שנוכל לסווג בצורה יותר טובה את איכות היין – איכות נמוכה (1), איכות בינונית (2) ואיכות גבוהה (3).



וכעת נציג את המידע לאחר הכנת הנתונים:



2**. סיווג וחיזוי:**

א. נבחר בשיטות: J48 (C4.5) -simpleCart (השיטות הוסברו לעיל). לא נבחר ברגרסיה לינארית מכיוון שלא קיים קשר לינארי (המשתנים לא מתפלגים נורמאלית).

**ב. J48:**

* בתוכנת WEKA נבחר באלגוריתם לסיווג J48. לאחר מכן נריץ אותו עם פרמטרים שונים, כמו מספר רשומות מינימלי בעלה ומקדם הבטחון ועוד. נבדוק את עץ ההחלטה שנבנה ואת הפרמטרים שלו, כמו גודל העץ, מספר עלים, דיוק, שגיאה.

עבור הבדיקה של המודל נשתמש בשיטה של 10 Fold CrossValidation, שמחלקת את המידע ל-FOLDS כך שמחלק מכינה את ה-trainning set ומחלק מהמידע בודקת אותו.

נעצור את הבדיקה לאחר שגודל העץ נראה לנו סביר ,לא גדול (overfiting) ולא קטן(הדיוק יורד).

**CART:**

* בתוכנת WEKA נבחר באלגוריתם לסיווג simpleCart. עבור הבדיקה של המודל נשתמש בשיטה של 10 Fold CrossValidation, שמחלקת את המידע ל-FOLDS כך שמחלק מכינה את ה-trainning set ומחלק מהמידע בודקת אותו. תהליך האלגוריתם הוא פיצול כל צומת בעזרת מדד ג'יני, קבלת החלטה שהעץ הושלם (כאשר אין פיצול נוסף שישפר אותו), והקצאת כל צומת סופי למחלקה.

**ג. תוצאות הדיווחים:**

Precision – "דייקנות" – כמה תצפיות סווגו באופן חיובי

ROC Area : השטח תחת עקומת הROC אשר מייצגת את שקלול התמורות שבין שיעור ה True positive לבין שיעור הFalse positive.

Acurracy – אחוז התצפיות שסווגו נכונה.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Num of leaves | Tree size | Accuracy | Specificity | ROC Area | F-Measure | Recall | Precision | FP Rate | TP Rate | Pruning | Confidence factor | Leaf min | RUN |
| 301 | 601 | **73.71%** | **65.79%** | 0.751 | 0.733 | 0.737 | 0.732 | 0.295 | 0.737 | TRUE | 0.25 | 2 | 1 J48 |
| 395 | 789 | 73.475% | 66.84% | 0.758 | 0.734 | 0.735 | 0.733 | 0.284 | 0.735 | FALSE | 0.25 | 2 | 2 J48 |
| 357 | 713 | 73.838% | 66.44% | 0.755 | 0.737 | 0.738 | 0.736 | 0.286 | 0.738 | TRUE | 0.5 | 2 | 3 J48 |
| 249 | 497 | 73.219% | 64.87% | 0.758 | 0.728 | 0.732 | 0.726 | 0.300 | 0.732 | TRUE | 0.5 | 4 | 4 J48 |
| 55 | 109 | **74.904%** | **61.92%** | **0.759** | 0.737 | 0.749 | 0.741 | 0.319 | 0.749 | TRUE |  | 2 | 1 CART |
| 438 | 875 | 75.756% | 66.71% | 0.766 | 0.735 | 0.758 | 0.751 | 0.283 | 0.758 | FALSE |  | 2 | 2 CART |
| 58 | 115 | 74.328% | 58.65% | 0.760 | 0.727 | 0.743 | 0.737 | 0.339 | 0.743 | TRUE |  | 4 | 3 CART |

**ד. מידת הדיוק של כל שיטה:**

|  |  |
| --- | --- |
| תוצאות הדיוק | RUN |
|  | 1 J48 |
|  | 2 J48 |
|  | 3 J48 |
|  | 4 J48 |
|  | 1 CART |
|  | 2 CART |
|  | 3 CART |

**מסקנות:**

* ראינו שהתוצאות של אלגוריתם CART ללא גיזום גורר תוצאות יותר מדוייקות – ולעומת זאת אם מריצים את אלגוריתם J48 ללא גיזום מקבלים תוצאות קצת פחות טובות
* בהרצת J48, העלאת ה-confidence factor הניבה תוצאות פחות מדויקות בהרצה ללא גיזום.
* הגדלת מספר המינימלי לעלה הניב תוצאות נמוכות יותר בשני האלגוריתמים.

נסתכל על התוצאות שקיבלנו. מצד אחד ההרצה השניה של אלגוריתם CART הביא לנו תוצאות טובות מאוד, גם סיווג נכונה הרבה מתוך הרשומות שבעלות איכות ברמה גבוהה, אך העץ שקיבלנו ללא הגיזום הינו גדול מאוד. כך גם ההרצות ה-2 וה-3 של אלגוריתם J48 – סיווגנו נכונה הרבה רשומות בעלות איכות ברמה גבוהה אך קיבלנו בתמונה עץ גדול מאוד (overfitting). לכן נבחר את ההרצה הראשונה של של אלגוריתם CART, (או אחריו לחילופין בהרצה הראשונה של אלגוריתם J48),מכיוון שבעבור עץ קטן קיבלנו ערכים גבוהים לעומת הרצות אחרות, ומכיוון שמעניין אותנו לנבא את איכותו של היין – בין אם האיכות היא נמוכה, בינונית או גבוהה – ובהנחה הזאת נרצה לבחור את ההרצה שנותנת לנו כמה שיותר יתרונות לעומת חסרונות.