**ממן 22: כריית מידע**

1. **חוקי הקשר**

נבחר 2 אלגוריתמים של חוקי הקשר:

1. **אלגוריתם אפריורי:**

אלגוריתם אפריורי מסתמך על התכונה האפריורית, כלומר העיקרון של האלגוריתם הוא שבמידה וקיים סט נתונים שכיח בבסיס הנתונים אז כל קבוצה חלקית של אותה קבוצת פריטים שכיחה חייבת גם היא להיות שכיחה.

זהו אלגוריתם איטרטיבי, כאשר במחזור k מוצאים את כל קבוצות הפריטים השכיחות בגודל k, כלומר אלה שמכילות k פריטים. בכל מחזור באלגוריתם האפריורי מתבצעים שלושה מעברים עיקריים:

**1**. יצירת מועמדים. מציאת מועמדים בגודל k על בסיס קבוצות הפריטים השכיחות שנמצאו בשלב k – 1. בשלב זה מתבצע צירוף של שתי קבוצות שנמצאו בשלב קודם לצורך מציאת קבוצה חדשה.

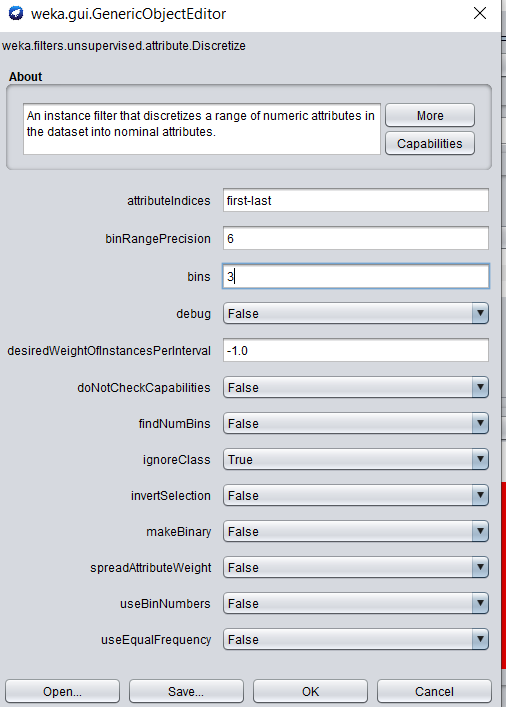
**2**. גיזום מועמדים שלא מקיימים את התכונה האפריורי, כלומר מתעלמים מהצירופים שאינם שכיחים

**3**. בדיקת התמיכה של קבוצות המועמדים שלא נגזמו בשלב 2 ע"י מניית מספר התנועות שבהן הם מופיעים בתוך בסיס הנתונים עצמו, כלומר בדיקה האם הקבוצות המועמדות באמת שכיחות.

2. **אלגוריתם FPGrowth:**

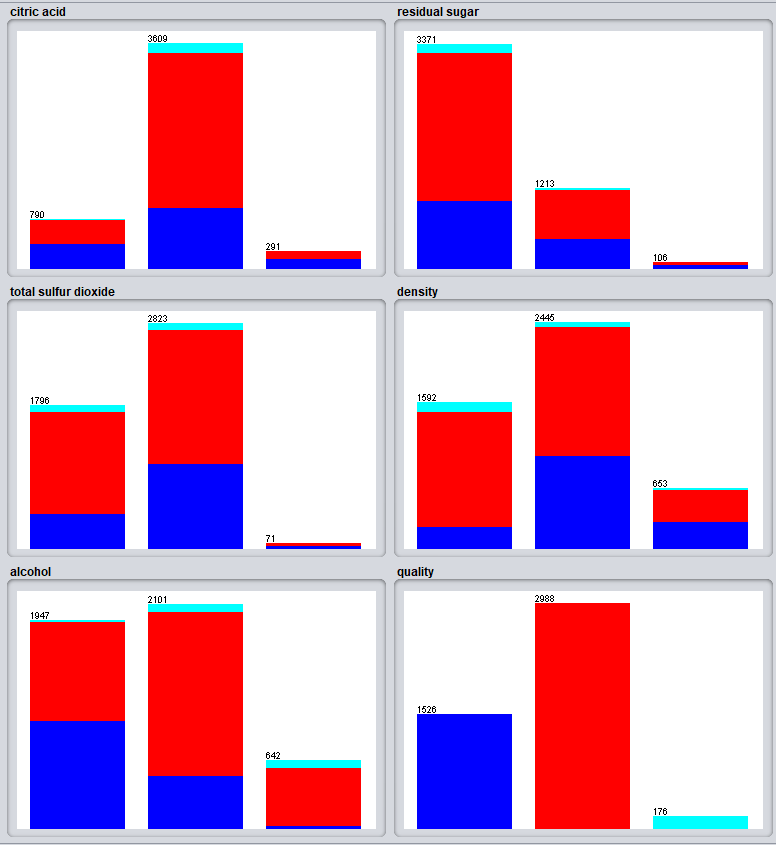
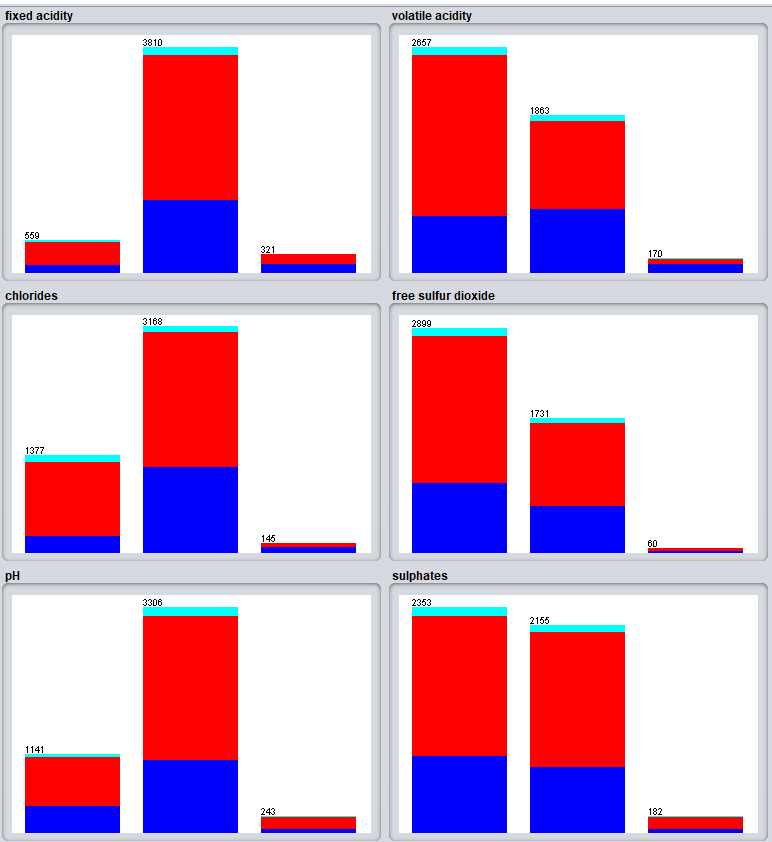
אלגוריתם זה משתמש בחיפוש לעומק, בניגוד לאלגוריתם אפריורי המבצע חיפוש לרוחב. הוא מוצא קבוצות שכיחות בלי לייצר מועמדים, תוך שימוש במבנה נתונים של עץ השומר מידע סטטיסטי על הפריטים בבסיס הנתונים. האלגוריתם יעיל במיוחד כאשר העץ קומפקטי – כאשר יש קבוצות שכיחות עם מספר פריטים גדול יחסית. האלגוריתם משתמש בשיטת הפרד ומשול:

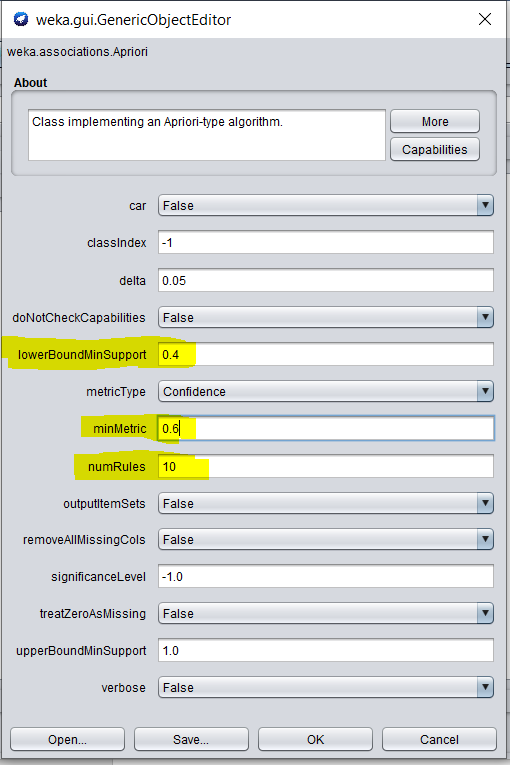
1. תחילה הוא דוחס את המידע שבבסיס הנתונים על ידי יצירת עץ תבניות שכיחות
2. לאחר מכם הוא מפריד את המידע בבסיס הנתונים לבסיסי נתונים מותנים, כל אחד מקושר לתבנית שנמצאה בשלב הראשון.
3. בשלב השלישי הוא מבצע כרייה נפרדת, לעומק, עבור כל בסיס נתונים שנוצר.

נשתמש בקובץ הנתונים המטוייב המכיל 4690 רשומות (הקובץ לאחר ניקוי). תחילה נבצע דיסקרטיזציה כדי שנוכל להריץ עליו את האלגוריתמים שנבחרו.

את כל התכונות חילקתי ל-3 bins, כך שהתקבל:

|  |  |  |
| --- | --- | --- |
| **שם תכונה** | **תאור התכונה** | **ערכים לאחר דיסקרטיזציה ל-3 בינים** |
| **fixed acidity** | חומציות קבוע |  |
| **volatile acidity** | חומציות נדיפה |  |
| **citric acid** | חומצת לימון |  |
| **residual sugar** | סוכר שיורי |  |
| **chlorides** | כלורידים |  |
| **free sulfur dioxide** | דו תחמוצת הגופרית |  |
| **total sulfur dioxide** | סה"כ דו תחמוצת הגופרית |  |
| **density** | צפיפות |  |
| **pH** | חומציות |  |
| **sulphates** | סולפטים |  |
| **alcohol** | אחוז אלכוהול |  |
| **quality** | איכות היין |  |

ובתצוגה ויזואלית:



**נתון: min\_confidence = 60%,**

**min\_support = 40%**

נריץ את אלגוריתם **אפריורי** על הנתונים.

בהתחלה הרצתי את האלגוריתם עם הגבלה של 10 חוקים. לאחר שנמצאו 10 חוקים, הרצתי גם עם 50, 200, 100 חוקים. לבסוף נמצאו כ-87 כללי הקשר.

כל החוקים שנמצאו הם כללי הקשר חזקים – מכיוון שהם מקיימים את ערכי הסף של שני המדדים (מכיוון שבהרצת האלגוריתם אלו ערכי הסף שהגדרנו).

נציג את הפלט של הרצת האלגוריתם עם מספר חוקים 100:

=== Run information ===

Scheme: weka.associations.Apriori -N 100 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.4 -S -1.0 -c -1

Relation: winequality-white-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-O3.0-E6.0-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-Llast-weka.filters.unsupervised.attribute.Remove-R14-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-Llast-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6-unset-class-temporarily

Instances: 4690

Attributes: 12

fixed acidity

volatile acidity

citric acid

residual sugar

chlorides

free sulfur dioxide

total sulfur dioxide

density

pH

sulphates

alcohol

quality

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.4 (1876 instances)

Minimum metric <confidence>: 0.6

Number of cycles performed: 12

Generated sets of large itemsets:

Size of set of large itemsets L(1): 14

Size of set of large itemsets L(2): 31

Size of set of large itemsets L(3): 8

Best rules found:

1. chlorides='(0.036667-0.064333]' total sulfur dioxide='(121.333333-232.666667]' 2162 ==> fixed acidity='(5.966667-8.133333]' 1892 <conf:(0.88)> lift:(1.08) lev:(0.03) [135] conv:(1.5)

2. citric acid='(0.246667-0.493333]' chlorides='(0.036667-0.064333]' 2366 ==> fixed acidity='(5.966667-8.133333]' 2041 <conf:(0.86)> lift:(1.06) lev:(0.03) [118] conv:(1.36)

3. total sulfur dioxide='(121.333333-232.666667]' 2823 ==> fixed acidity='(5.966667-8.133333]' 2425 <conf:(0.86)> lift:(1.06) lev:(0.03) [131] conv:(1.33)

4. fixed acidity='(5.966667-8.133333]' quality='(5-7]' 2438 ==> citric acid='(0.246667-0.493333]' 2072 <conf:(0.85)> lift:(1.1) lev:(0.04) [195] conv:(1.53)

5. chlorides='(0.036667-0.064333]' pH='(3.086667-3.453333]' 2281 ==> fixed acidity='(5.966667-8.133333]' 1937 <conf:(0.85)> lift:(1.05) lev:(0.02) [83] conv:(1.24)

6. volatile acidity='(-inf-0.27]' 2657 ==> citric acid='(0.246667-0.493333]' 2248 <conf:(0.85)> lift:(1.1) lev:(0.04) [203] conv:(1.49)

7. chlorides='(0.036667-0.064333]' 3168 ==> fixed acidity='(5.966667-8.133333]' 2675 <conf:(0.84)> lift:(1.04) lev:(0.02) [101] conv:(1.2)

8. citric acid='(0.246667-0.493333]' pH='(3.086667-3.453333]' 2576 ==> fixed acidity='(5.966667-8.133333]' 2169 <conf:(0.84)> lift:(1.04) lev:(0.02) [76] conv:(1.18)

9. citric acid='(0.246667-0.493333]' quality='(5-7]' 2479 ==> fixed acidity='(5.966667-8.133333]' 2072 <conf:(0.84)> lift:(1.03) lev:(0.01) [58] conv:(1.14)

10. density='(0.99239-0.99767]' 2445 ==> fixed acidity='(5.966667-8.133333]' 2034 <conf:(0.83)> lift:(1.02) lev:(0.01) [47] conv:(1.11)

11. quality='(5-7]' 2988 ==> citric acid='(0.246667-0.493333]' 2479 <conf:(0.83)> lift:(1.08) lev:(0.04) [179] conv:(1.35)

12. citric acid='(0.246667-0.493333]' 3609 ==> fixed acidity='(5.966667-8.133333]' 2986 <conf:(0.83)> lift:(1.02) lev:(0.01) [54] conv:(1.09)

13. pH='(3.086667-3.453333]' 3306 ==> fixed acidity='(5.966667-8.133333]' 2723 <conf:(0.82)> lift:(1.01) lev:(0.01) [37] conv:(1.06)

14. volatile acidity='(-inf-0.27]' 2657 ==> fixed acidity='(5.966667-8.133333]' 2172 <conf:(0.82)> lift:(1.01) lev:(0) [13] conv:(1.03)

15. quality='(5-7]' 2988 ==> fixed acidity='(5.966667-8.133333]' 2438 <conf:(0.82)> lift:(1) lev:(0) [10] conv:(1.02)

16. free sulfur dioxide='(-inf-38.666667]' 2899 ==> residual sugar='(-inf-9.083333]' 2342 <conf:(0.81)> lift:(1.12) lev:(0.06) [258] conv:(1.46)

17. residual sugar='(-inf-9.083333]' pH='(3.086667-3.453333]' 2507 ==> fixed acidity='(5.966667-8.133333]' 2024 <conf:(0.81)> lift:(0.99) lev:(-0) [-12] conv:(0.97)

18. citric acid='(0.246667-0.493333]' residual sugar='(-inf-9.083333]' 2653 ==> fixed acidity='(5.966667-8.133333]' 2126 <conf:(0.8)> lift:(0.99) lev:(-0.01) [-29] conv:(0.94)

19. fixed acidity='(5.966667-8.133333]' residual sugar='(-inf-9.083333]' 2656 ==> citric acid='(0.246667-0.493333]' 2126 <conf:(0.8)> lift:(1.04) lev:(0.02) [82] conv:(1.15)

20. fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 2723 ==> citric acid='(0.246667-0.493333]' 2169 <conf:(0.8)> lift:(1.04) lev:(0.02) [73] conv:(1.13)

21. residual sugar='(-inf-9.083333]' 3371 ==> fixed acidity='(5.966667-8.133333]' 2656 <conf:(0.79)> lift:(0.97) lev:(-0.02) [-82] conv:(0.88)

22. residual sugar='(-inf-9.083333]' 3371 ==> citric acid='(0.246667-0.493333]' 2653 <conf:(0.79)> lift:(1.02) lev:(0.01) [58] conv:(1.08)

23. residual sugar='(-inf-9.083333]' pH='(3.086667-3.453333]' 2507 ==> citric acid='(0.246667-0.493333]' 1971 <conf:(0.79)> lift:(1.02) lev:(0.01) [41] conv:(1.08)

24. fixed acidity='(5.966667-8.133333]' 3810 ==> citric acid='(0.246667-0.493333]' 2986 <conf:(0.78)> lift:(1.02) lev:(0.01) [54] conv:(1.06)

25. free sulfur dioxide='(-inf-38.666667]' 2899 ==> fixed acidity='(5.966667-8.133333]' 2264 <conf:(0.78)> lift:(0.96) lev:(-0.02) [-91] conv:(0.86)

26. fixed acidity='(5.966667-8.133333]' total sulfur dioxide='(121.333333-232.666667]' 2425 ==> chlorides='(0.036667-0.064333]' 1892 <conf:(0.78)> lift:(1.16) lev:(0.05) [253] conv:(1.47)

27. free sulfur dioxide='(-inf-38.666667]' 2899 ==> citric acid='(0.246667-0.493333]' 2259 <conf:(0.78)> lift:(1.01) lev:(0.01) [28] conv:(1.04)

28. pH='(3.086667-3.453333]' 3306 ==> citric acid='(0.246667-0.493333]' 2576 <conf:(0.78)> lift:(1.01) lev:(0.01) [32] conv:(1.04)

29. density='(0.99239-0.99767]' 2445 ==> chlorides='(0.036667-0.064333]' 1882 <conf:(0.77)> lift:(1.14) lev:(0.05) [230] conv:(1.41)

30. total sulfur dioxide='(121.333333-232.666667]' 2823 ==> chlorides='(0.036667-0.064333]' 2162 <conf:(0.77)> lift:(1.13) lev:(0.05) [255] conv:(1.38)

31. citric acid='(0.246667-0.493333]' pH='(3.086667-3.453333]' 2576 ==> residual sugar='(-inf-9.083333]' 1971 <conf:(0.77)> lift:(1.06) lev:(0.03) [119] conv:(1.2)

32. fixed acidity='(5.966667-8.133333]' chlorides='(0.036667-0.064333]' 2675 ==> citric acid='(0.246667-0.493333]' 2041 <conf:(0.76)> lift:(0.99) lev:(-0) [-17] conv:(0.97)

33. fixed acidity='(5.966667-8.133333]' residual sugar='(-inf-9.083333]' 2656 ==> pH='(3.086667-3.453333]' 2024 <conf:(0.76)> lift:(1.08) lev:(0.03) [151] conv:(1.24)

34. pH='(3.086667-3.453333]' 3306 ==> residual sugar='(-inf-9.083333]' 2507 <conf:(0.76)> lift:(1.06) lev:(0.03) [130] conv:(1.16)

35. total sulfur dioxide='(121.333333-232.666667]' 2823 ==> citric acid='(0.246667-0.493333]' 2114 <conf:(0.75)> lift:(0.97) lev:(-0.01) [-58] conv:(0.92)

36. chlorides='(0.036667-0.064333]' 3168 ==> citric acid='(0.246667-0.493333]' 2366 <conf:(0.75)> lift:(0.97) lev:(-0.02) [-71] conv:(0.91)

37. residual sugar='(-inf-9.083333]' 3371 ==> pH='(3.086667-3.453333]' 2507 <conf:(0.74)> lift:(1.06) lev:(0.03) [130] conv:(1.15)

38. fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 2723 ==> residual sugar='(-inf-9.083333]' 2024 <conf:(0.74)> lift:(1.03) lev:(0.01) [66] conv:(1.09)

39. citric acid='(0.246667-0.493333]' residual sugar='(-inf-9.083333]' 2653 ==> pH='(3.086667-3.453333]' 1971 <conf:(0.74)> lift:(1.05) lev:(0.02) [100] conv:(1.15)

40. quality='(5-7]' 2988 ==> residual sugar='(-inf-9.083333]' 2215 <conf:(0.74)> lift:(1.03) lev:(0.01) [67] conv:(1.09)

41. citric acid='(0.246667-0.493333]' 3609 ==> residual sugar='(-inf-9.083333]' 2653 <conf:(0.74)> lift:(1.02) lev:(0.01) [58] conv:(1.06)

42. fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2986 ==> pH='(3.086667-3.453333]' 2169 <conf:(0.73)> lift:(1.03) lev:(0.01) [64] conv:(1.08)

43. fixed acidity='(5.966667-8.133333]' chlorides='(0.036667-0.064333]' 2675 ==> pH='(3.086667-3.453333]' 1937 <conf:(0.72)> lift:(1.03) lev:(0.01) [51] conv:(1.07)

44. total sulfur dioxide='(121.333333-232.666667]' 2823 ==> pH='(3.086667-3.453333]' 2038 <conf:(0.72)> lift:(1.02) lev:(0.01) [48] conv:(1.06)

45. chlorides='(0.036667-0.064333]' 3168 ==> pH='(3.086667-3.453333]' 2281 <conf:(0.72)> lift:(1.02) lev:(0.01) [47] conv:(1.05)

46. volatile acidity='(-inf-0.27]' 2657 ==> residual sugar='(-inf-9.083333]' 1902 <conf:(0.72)> lift:(1) lev:(-0) [-7] conv:(0.99)

47. fixed acidity='(5.966667-8.133333]' 3810 ==> pH='(3.086667-3.453333]' 2723 <conf:(0.71)> lift:(1.01) lev:(0.01) [37] conv:(1.03)

48. citric acid='(0.246667-0.493333]' 3609 ==> pH='(3.086667-3.453333]' 2576 <conf:(0.71)> lift:(1.01) lev:(0.01) [32] conv:(1.03)

49. volatile acidity='(-inf-0.27]' 2657 ==> quality='(5-7]' 1892 <conf:(0.71)> lift:(1.12) lev:(0.04) [199] conv:(1.26)

50. fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2986 ==> residual sugar='(-inf-9.083333]' 2126 <conf:(0.71)> lift:(0.99) lev:(-0) [-20] conv:(0.98)

51. fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 2723 ==> chlorides='(0.036667-0.064333]' 1937 <conf:(0.71)> lift:(1.05) lev:(0.02) [97] conv:(1.12)

52. fixed acidity='(5.966667-8.133333]' chlorides='(0.036667-0.064333]' 2675 ==> total sulfur dioxide='(121.333333-232.666667]' 1892 <conf:(0.71)> lift:(1.18) lev:(0.06) [281] conv:(1.36)

53. volatile acidity='(-inf-0.27]' 2657 ==> chlorides='(0.036667-0.064333]' 1879 <conf:(0.71)> lift:(1.05) lev:(0.02) [84] conv:(1.11)

54. quality='(5-7]' 2988 ==> pH='(3.086667-3.453333]' 2102 <conf:(0.7)> lift:(1) lev:(-0) [-4] conv:(0.99)

55. fixed acidity='(5.966667-8.133333]' 3810 ==> chlorides='(0.036667-0.064333]' 2675 <conf:(0.7)> lift:(1.04) lev:(0.02) [101] conv:(1.09)

56. fixed acidity='(5.966667-8.133333]' 3810 ==> residual sugar='(-inf-9.083333]' 2656 <conf:(0.7)> lift:(0.97) lev:(-0.02) [-82] conv:(0.93)

57. free sulfur dioxide='(-inf-38.666667]' 2899 ==> pH='(3.086667-3.453333]' 2017 <conf:(0.7)> lift:(0.99) lev:(-0.01) [-26] conv:(0.97)

58. residual sugar='(-inf-9.083333]' 3371 ==> free sulfur dioxide='(-inf-38.666667]' 2342 <conf:(0.69)> lift:(1.12) lev:(0.06) [258] conv:(1.25)

59. fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2986 ==> quality='(5-7]' 2072 <conf:(0.69)> lift:(1.09) lev:(0.04) [169] conv:(1.18)

60. quality='(5-7]' 2988 ==> fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2072 <conf:(0.69)> lift:(1.09) lev:(0.04) [169] conv:(1.18)

61. pH='(3.086667-3.453333]' 3306 ==> chlorides='(0.036667-0.064333]' 2281 <conf:(0.69)> lift:(1.02) lev:(0.01) [47] conv:(1.05)

62. citric acid='(0.246667-0.493333]' 3609 ==> quality='(5-7]' 2479 <conf:(0.69)> lift:(1.08) lev:(0.04) [179] conv:(1.16)

63. fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2986 ==> chlorides='(0.036667-0.064333]' 2041 <conf:(0.68)> lift:(1.01) lev:(0.01) [24] conv:(1.02)

64. chlorides='(0.036667-0.064333]' 3168 ==> total sulfur dioxide='(121.333333-232.666667]' 2162 <conf:(0.68)> lift:(1.13) lev:(0.05) [255] conv:(1.25)

65. total sulfur dioxide='(121.333333-232.666667]' 2823 ==> fixed acidity='(5.966667-8.133333]' chlorides='(0.036667-0.064333]' 1892 <conf:(0.67)> lift:(1.18) lev:(0.06) [281] conv:(1.3)

66. chlorides='(0.036667-0.064333]' 3168 ==> residual sugar='(-inf-9.083333]' 2097 <conf:(0.66)> lift:(0.92) lev:(-0.04) [-180] conv:(0.83)

67. residual sugar='(-inf-9.083333]' 3371 ==> quality='(5-7]' 2215 <conf:(0.66)> lift:(1.03) lev:(0.01) [67] conv:(1.06)

68. pH='(3.086667-3.453333]' 3306 ==> fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2169 <conf:(0.66)> lift:(1.03) lev:(0.01) [64] conv:(1.06)

69. citric acid='(0.246667-0.493333]' 3609 ==> chlorides='(0.036667-0.064333]' 2366 <conf:(0.66)> lift:(0.97) lev:(-0.02) [-71] conv:(0.94)

70. free sulfur dioxide='(-inf-38.666667]' 2899 ==> quality='(5-7]' 1889 <conf:(0.65)> lift:(1.02) lev:(0.01) [42] conv:(1.04)

71. chlorides='(0.036667-0.064333]' 3168 ==> fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2041 <conf:(0.64)> lift:(1.01) lev:(0.01) [24] conv:(1.02)

72. fixed acidity='(5.966667-8.133333]' 3810 ==> quality='(5-7]' 2438 <conf:(0.64)> lift:(1) lev:(0) [10] conv:(1.01)

73. fixed acidity='(5.966667-8.133333]' 3810 ==> total sulfur dioxide='(121.333333-232.666667]' 2425 <conf:(0.64)> lift:(1.06) lev:(0.03) [131] conv:(1.09)

74. pH='(3.086667-3.453333]' 3306 ==> quality='(5-7]' 2102 <conf:(0.64)> lift:(1) lev:(-0) [-4] conv:(1)

75. quality='(5-7]' 2988 ==> volatile acidity='(-inf-0.27]' 1892 <conf:(0.63)> lift:(1.12) lev:(0.04) [199] conv:(1.18)

76. quality='(5-7]' 2988 ==> chlorides='(0.036667-0.064333]' 1889 <conf:(0.63)> lift:(0.94) lev:(-0.03) [-129] conv:(0.88)

77. quality='(5-7]' 2988 ==> free sulfur dioxide='(-inf-38.666667]' 1889 <conf:(0.63)> lift:(1.02) lev:(0.01) [42] conv:(1.04)

78. residual sugar='(-inf-9.083333]' 3371 ==> fixed acidity='(5.966667-8.133333]' citric acid='(0.246667-0.493333]' 2126 <conf:(0.63)> lift:(0.99) lev:(-0) [-20] conv:(0.98)

79. citric acid='(0.246667-0.493333]' 3609 ==> free sulfur dioxide='(-inf-38.666667]' 2259 <conf:(0.63)> lift:(1.01) lev:(0.01) [28] conv:(1.02)

80. citric acid='(0.246667-0.493333]' 3609 ==> volatile acidity='(-inf-0.27]' 2248 <conf:(0.62)> lift:(1.1) lev:(0.04) [203] conv:(1.15)

81. residual sugar='(-inf-9.083333]' 3371 ==> chlorides='(0.036667-0.064333]' 2097 <conf:(0.62)> lift:(0.92) lev:(-0.04) [-180] conv:(0.86)

82. pH='(3.086667-3.453333]' 3306 ==> total sulfur dioxide='(121.333333-232.666667]' 2038 <conf:(0.62)> lift:(1.02) lev:(0.01) [48] conv:(1.04)

83. pH='(3.086667-3.453333]' 3306 ==> fixed acidity='(5.966667-8.133333]' residual sugar='(-inf-9.083333]' 2024 <conf:(0.61)> lift:(1.08) lev:(0.03) [151] conv:(1.12)

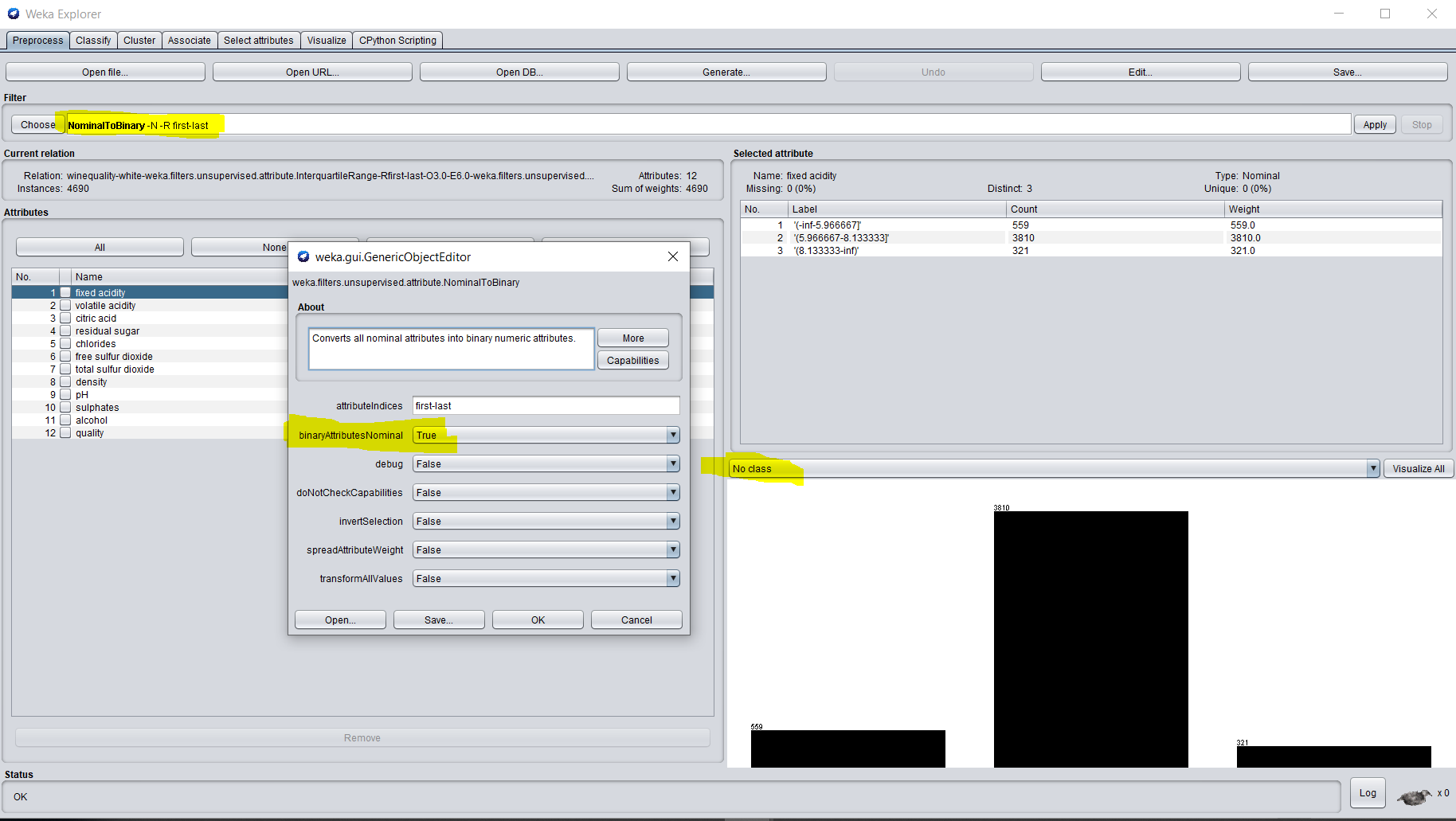
84. chlorides='(0.036667-0.064333]' 3168 ==> fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 1937 <conf:(0.61)> lift:(1.05) lev:(0.02) [97] conv:(1.08)

85. pH='(3.086667-3.453333]' 3306 ==> free sulfur dioxide='(-inf-38.666667]' 2017 <conf:(0.61)> lift:(0.99) lev:(-0.01) [-26] conv:(0.98)

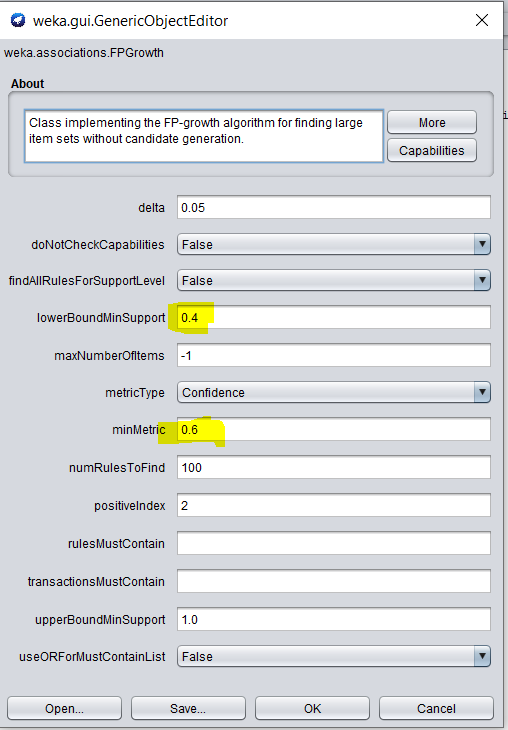
86. citric acid='(0.246667-0.493333]' 3609 ==> fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 2169 <conf:(0.6)> lift:(1.04) lev:(0.02) [73] conv:(1.05)

87. residual sugar='(-inf-9.083333]' 3371 ==> fixed acidity='(5.966667-8.133333]' pH='(3.086667-3.453333]' 2024 <conf:(0.6)> lift:(1.03) lev:(0.01) [66] conv:(1.05)

נריץ את אלגוריתם FP Growth על הנתונים.

על מנת שנוכל להריץ את האלגוריתם, נבצע שינוי על הנתונים:

גם באלגוריתם זה נשנה את הקונפיגורציה המבוקשת לפני ההרצה, ונריץ.



גם פה קיבלנו 87 חוקי הקשר:

=== Run information ===

Scheme: weka.associations.FPGrowth -P 2 -I -1 -N 100 -T 0 -C 0.6 -D 0.05 -U 1.0 -M 0.4

Relation: winequality-white-weka.filters.unsupervised.attribute.InterquartileRange-Rfirst-last-O3.0-E6.0-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-Llast-weka.filters.unsupervised.attribute.Remove-R14-weka.filters.unsupervised.instance.RemoveWithValues-S0.0-Clast-Llast-weka.filters.unsupervised.attribute.Remove-R13-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-Rfirst-last-precision6-unset-class-temporarily-weka.filters.unsupervised.attribute.NominalToBinary-N-Rfirst-last

Instances: 4690

Attributes: 36

fixed acidity='(-inf-5.966667]'

fixed acidity='(5.966667-8.133333]'

fixed acidity='(8.133333-inf)'

volatile acidity='(-inf-0.27]'

volatile acidity='(0.27-0.46]'

volatile acidity='(0.46-inf)'

citric acid='(-inf-0.246667]'

citric acid='(0.246667-0.493333]'

citric acid='(0.493333-inf)'

residual sugar='(-inf-9.083333]'

residual sugar='(9.083333-17.566667]'

residual sugar='(17.566667-inf)'

chlorides='(-inf-0.036667]'

chlorides='(0.036667-0.064333]'

chlorides='(0.064333-inf)'

free sulfur dioxide='(-inf-38.666667]'

free sulfur dioxide='(38.666667-75.333333]'

free sulfur dioxide='(75.333333-inf)'

total sulfur dioxide='(-inf-121.333333]'

total sulfur dioxide='(121.333333-232.666667]'

total sulfur dioxide='(232.666667-inf)'

density='(-inf-0.99239]'

density='(0.99239-0.99767]'

density='(0.99767-inf)'

pH='(-inf-3.086667]'

pH='(3.086667-3.453333]'

pH='(3.453333-inf)'

sulphates='(-inf-0.47]'

sulphates='(0.47-0.72]'

sulphates='(0.72-inf)'

alcohol='(-inf-10.066667]'

alcohol='(10.066667-12.133333]'

alcohol='(12.133333-inf)'

quality='(-inf-5]'

quality='(5-7]'

quality='(7-inf)'

=== Associator model (full training set) ===

FPGrowth found 87 rules (displaying top 87)

1. [chlorides='(0.036667-0.064333]'=t, total sulfur dioxide='(121.333333-232.666667]'=t]: 2162 ==> [fixed acidity='(5.966667-8.133333]'=t]: 1892 <conf:(0.88)> lift:(1.08) lev:(0.03) conv:(1.5)

2. [citric acid='(0.246667-0.493333]'=t, chlorides='(0.036667-0.064333]'=t]: 2366 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2041 <conf:(0.86)> lift:(1.06) lev:(0.03) conv:(1.36)

3. [total sulfur dioxide='(121.333333-232.666667]'=t]: 2823 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2425 <conf:(0.86)> lift:(1.06) lev:(0.03) conv:(1.33)

4. [fixed acidity='(5.966667-8.133333]'=t, quality='(5-7]'=t]: 2438 ==> [citric acid='(0.246667-0.493333]'=t]: 2072 <conf:(0.85)> lift:(1.1) lev:(0.04) conv:(1.53)

5. [pH='(3.086667-3.453333]'=t, chlorides='(0.036667-0.064333]'=t]: 2281 ==> [fixed acidity='(5.966667-8.133333]'=t]: 1937 <conf:(0.85)> lift:(1.05) lev:(0.02) conv:(1.24)

6. [volatile acidity='(-inf-0.27]'=t]: 2657 ==> [citric acid='(0.246667-0.493333]'=t]: 2248 <conf:(0.85)> lift:(1.1) lev:(0.04) conv:(1.49)

7. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2675 <conf:(0.84)> lift:(1.04) lev:(0.02) conv:(1.2)

8. [citric acid='(0.246667-0.493333]'=t, pH='(3.086667-3.453333]'=t]: 2576 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2169 <conf:(0.84)> lift:(1.04) lev:(0.02) conv:(1.18)

9. [citric acid='(0.246667-0.493333]'=t, quality='(5-7]'=t]: 2479 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2072 <conf:(0.84)> lift:(1.03) lev:(0.01) conv:(1.14)

10. [density='(0.99239-0.99767]'=t]: 2445 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2034 <conf:(0.83)> lift:(1.02) lev:(0.01) conv:(1.11)

11. [quality='(5-7]'=t]: 2988 ==> [citric acid='(0.246667-0.493333]'=t]: 2479 <conf:(0.83)> lift:(1.08) lev:(0.04) conv:(1.35)

12. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2986 <conf:(0.83)> lift:(1.02) lev:(0.01) conv:(1.09)

13. [pH='(3.086667-3.453333]'=t]: 3306 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2723 <conf:(0.82)> lift:(1.01) lev:(0.01) conv:(1.06)

14. [volatile acidity='(-inf-0.27]'=t]: 2657 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2172 <conf:(0.82)> lift:(1.01) lev:(0) conv:(1.03)

15. [quality='(5-7]'=t]: 2988 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2438 <conf:(0.82)> lift:(1) lev:(0) conv:(1.02)

16. [free sulfur dioxide='(-inf-38.666667]'=t]: 2899 ==> [residual sugar='(-inf-9.083333]'=t]: 2342 <conf:(0.81)> lift:(1.12) lev:(0.06) conv:(1.46)

17. [residual sugar='(-inf-9.083333]'=t, pH='(3.086667-3.453333]'=t]: 2507 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2024 <conf:(0.81)> lift:(0.99) lev:(-0) conv:(0.97)

18. [citric acid='(0.246667-0.493333]'=t, residual sugar='(-inf-9.083333]'=t]: 2653 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2126 <conf:(0.8)> lift:(0.99) lev:(-0.01) conv:(0.94)

19. [fixed acidity='(5.966667-8.133333]'=t, residual sugar='(-inf-9.083333]'=t]: 2656 ==> [citric acid='(0.246667-0.493333]'=t]: 2126 <conf:(0.8)> lift:(1.04) lev:(0.02) conv:(1.15)

20. [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 2723 ==> [citric acid='(0.246667-0.493333]'=t]: 2169 <conf:(0.8)> lift:(1.04) lev:(0.02) conv:(1.13)

21. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2656 <conf:(0.79)> lift:(0.97) lev:(-0.02) conv:(0.88)

22. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [citric acid='(0.246667-0.493333]'=t]: 2653 <conf:(0.79)> lift:(1.02) lev:(0.01) conv:(1.08)

23. [residual sugar='(-inf-9.083333]'=t, pH='(3.086667-3.453333]'=t]: 2507 ==> [citric acid='(0.246667-0.493333]'=t]: 1971 <conf:(0.79)> lift:(1.02) lev:(0.01) conv:(1.08)

24. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [citric acid='(0.246667-0.493333]'=t]: 2986 <conf:(0.78)> lift:(1.02) lev:(0.01) conv:(1.06)

25. [free sulfur dioxide='(-inf-38.666667]'=t]: 2899 ==> [fixed acidity='(5.966667-8.133333]'=t]: 2264 <conf:(0.78)> lift:(0.96) lev:(-0.02) conv:(0.86)

26. [fixed acidity='(5.966667-8.133333]'=t, total sulfur dioxide='(121.333333-232.666667]'=t]: 2425 ==> [chlorides='(0.036667-0.064333]'=t]: 1892 <conf:(0.78)> lift:(1.16) lev:(0.05) conv:(1.47)

27. [free sulfur dioxide='(-inf-38.666667]'=t]: 2899 ==> [citric acid='(0.246667-0.493333]'=t]: 2259 <conf:(0.78)> lift:(1.01) lev:(0.01) conv:(1.04)

28. [pH='(3.086667-3.453333]'=t]: 3306 ==> [citric acid='(0.246667-0.493333]'=t]: 2576 <conf:(0.78)> lift:(1.01) lev:(0.01) conv:(1.04)

29. [density='(0.99239-0.99767]'=t]: 2445 ==> [chlorides='(0.036667-0.064333]'=t]: 1882 <conf:(0.77)> lift:(1.14) lev:(0.05) conv:(1.41)

30. [total sulfur dioxide='(121.333333-232.666667]'=t]: 2823 ==> [chlorides='(0.036667-0.064333]'=t]: 2162 <conf:(0.77)> lift:(1.13) lev:(0.05) conv:(1.38)

31. [citric acid='(0.246667-0.493333]'=t, pH='(3.086667-3.453333]'=t]: 2576 ==> [residual sugar='(-inf-9.083333]'=t]: 1971 <conf:(0.77)> lift:(1.06) lev:(0.03) conv:(1.2)

32. [fixed acidity='(5.966667-8.133333]'=t, chlorides='(0.036667-0.064333]'=t]: 2675 ==> [citric acid='(0.246667-0.493333]'=t]: 2041 <conf:(0.76)> lift:(0.99) lev:(-0) conv:(0.97)

33. [fixed acidity='(5.966667-8.133333]'=t, residual sugar='(-inf-9.083333]'=t]: 2656 ==> [pH='(3.086667-3.453333]'=t]: 2024 <conf:(0.76)> lift:(1.08) lev:(0.03) conv:(1.24)

34. [pH='(3.086667-3.453333]'=t]: 3306 ==> [residual sugar='(-inf-9.083333]'=t]: 2507 <conf:(0.76)> lift:(1.06) lev:(0.03) conv:(1.16)

35. [total sulfur dioxide='(121.333333-232.666667]'=t]: 2823 ==> [citric acid='(0.246667-0.493333]'=t]: 2114 <conf:(0.75)> lift:(0.97) lev:(-0.01) conv:(0.92)

36. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [citric acid='(0.246667-0.493333]'=t]: 2366 <conf:(0.75)> lift:(0.97) lev:(-0.02) conv:(0.91)

37. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [pH='(3.086667-3.453333]'=t]: 2507 <conf:(0.74)> lift:(1.06) lev:(0.03) conv:(1.15)

38. [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 2723 ==> [residual sugar='(-inf-9.083333]'=t]: 2024 <conf:(0.74)> lift:(1.03) lev:(0.01) conv:(1.09)

39. [citric acid='(0.246667-0.493333]'=t, residual sugar='(-inf-9.083333]'=t]: 2653 ==> [pH='(3.086667-3.453333]'=t]: 1971 <conf:(0.74)> lift:(1.05) lev:(0.02) conv:(1.15)

40. [quality='(5-7]'=t]: 2988 ==> [residual sugar='(-inf-9.083333]'=t]: 2215 <conf:(0.74)> lift:(1.03) lev:(0.01) conv:(1.09)

41. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [residual sugar='(-inf-9.083333]'=t]: 2653 <conf:(0.74)> lift:(1.02) lev:(0.01) conv:(1.06)

42. [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2986 ==> [pH='(3.086667-3.453333]'=t]: 2169 <conf:(0.73)> lift:(1.03) lev:(0.01) conv:(1.08)

43. [fixed acidity='(5.966667-8.133333]'=t, chlorides='(0.036667-0.064333]'=t]: 2675 ==> [pH='(3.086667-3.453333]'=t]: 1937 <conf:(0.72)> lift:(1.03) lev:(0.01) conv:(1.07)

44. [total sulfur dioxide='(121.333333-232.666667]'=t]: 2823 ==> [pH='(3.086667-3.453333]'=t]: 2038 <conf:(0.72)> lift:(1.02) lev:(0.01) conv:(1.06)

45. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [pH='(3.086667-3.453333]'=t]: 2281 <conf:(0.72)> lift:(1.02) lev:(0.01) conv:(1.05)

46. [volatile acidity='(-inf-0.27]'=t]: 2657 ==> [residual sugar='(-inf-9.083333]'=t]: 1902 <conf:(0.72)> lift:(1) lev:(-0) conv:(0.99)

47. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [pH='(3.086667-3.453333]'=t]: 2723 <conf:(0.71)> lift:(1.01) lev:(0.01) conv:(1.03)

48. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [pH='(3.086667-3.453333]'=t]: 2576 <conf:(0.71)> lift:(1.01) lev:(0.01) conv:(1.03)

49. [volatile acidity='(-inf-0.27]'=t]: 2657 ==> [quality='(5-7]'=t]: 1892 <conf:(0.71)> lift:(1.12) lev:(0.04) conv:(1.26)

50. [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2986 ==> [residual sugar='(-inf-9.083333]'=t]: 2126 <conf:(0.71)> lift:(0.99) lev:(-0) conv:(0.98)

51. [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 2723 ==> [chlorides='(0.036667-0.064333]'=t]: 1937 <conf:(0.71)> lift:(1.05) lev:(0.02) conv:(1.12)

52. [fixed acidity='(5.966667-8.133333]'=t, chlorides='(0.036667-0.064333]'=t]: 2675 ==> [total sulfur dioxide='(121.333333-232.666667]'=t]: 1892 <conf:(0.71)> lift:(1.18) lev:(0.06) conv:(1.36)

53. [volatile acidity='(-inf-0.27]'=t]: 2657 ==> [chlorides='(0.036667-0.064333]'=t]: 1879 <conf:(0.71)> lift:(1.05) lev:(0.02) conv:(1.11)

54. [quality='(5-7]'=t]: 2988 ==> [pH='(3.086667-3.453333]'=t]: 2102 <conf:(0.7)> lift:(1) lev:(-0) conv:(0.99)

55. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [chlorides='(0.036667-0.064333]'=t]: 2675 <conf:(0.7)> lift:(1.04) lev:(0.02) conv:(1.09)

56. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [residual sugar='(-inf-9.083333]'=t]: 2656 <conf:(0.7)> lift:(0.97) lev:(-0.02) conv:(0.93)

57. [free sulfur dioxide='(-inf-38.666667]'=t]: 2899 ==> [pH='(3.086667-3.453333]'=t]: 2017 <conf:(0.7)> lift:(0.99) lev:(-0.01) conv:(0.97)

58. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [free sulfur dioxide='(-inf-38.666667]'=t]: 2342 <conf:(0.69)> lift:(1.12) lev:(0.06) conv:(1.25)

59. [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2986 ==> [quality='(5-7]'=t]: 2072 <conf:(0.69)> lift:(1.09) lev:(0.04) conv:(1.18)

60. [quality='(5-7]'=t]: 2988 ==> [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2072 <conf:(0.69)> lift:(1.09) lev:(0.04) conv:(1.18)

61. [pH='(3.086667-3.453333]'=t]: 3306 ==> [chlorides='(0.036667-0.064333]'=t]: 2281 <conf:(0.69)> lift:(1.02) lev:(0.01) conv:(1.05)

62. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [quality='(5-7]'=t]: 2479 <conf:(0.69)> lift:(1.08) lev:(0.04) conv:(1.16)

63. [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2986 ==> [chlorides='(0.036667-0.064333]'=t]: 2041 <conf:(0.68)> lift:(1.01) lev:(0.01) conv:(1.02)

64. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [total sulfur dioxide='(121.333333-232.666667]'=t]: 2162 <conf:(0.68)> lift:(1.13) lev:(0.05) conv:(1.25)

65. [total sulfur dioxide='(121.333333-232.666667]'=t]: 2823 ==> [fixed acidity='(5.966667-8.133333]'=t, chlorides='(0.036667-0.064333]'=t]: 1892 <conf:(0.67)> lift:(1.18) lev:(0.06) conv:(1.3)

66. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [residual sugar='(-inf-9.083333]'=t]: 2097 <conf:(0.66)> lift:(0.92) lev:(-0.04) conv:(0.83)

67. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [quality='(5-7]'=t]: 2215 <conf:(0.66)> lift:(1.03) lev:(0.01) conv:(1.06)

68. [pH='(3.086667-3.453333]'=t]: 3306 ==> [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2169 <conf:(0.66)> lift:(1.03) lev:(0.01) conv:(1.06)

69. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [chlorides='(0.036667-0.064333]'=t]: 2366 <conf:(0.66)> lift:(0.97) lev:(-0.02) conv:(0.94)

70. [free sulfur dioxide='(-inf-38.666667]'=t]: 2899 ==> [quality='(5-7]'=t]: 1889 <conf:(0.65)> lift:(1.02) lev:(0.01) conv:(1.04)

71. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2041 <conf:(0.64)> lift:(1.01) lev:(0.01) conv:(1.02)

72. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [quality='(5-7]'=t]: 2438 <conf:(0.64)> lift:(1) lev:(0) conv:(1.01)

73. [fixed acidity='(5.966667-8.133333]'=t]: 3810 ==> [total sulfur dioxide='(121.333333-232.666667]'=t]: 2425 <conf:(0.64)> lift:(1.06) lev:(0.03) conv:(1.09)

74. [pH='(3.086667-3.453333]'=t]: 3306 ==> [quality='(5-7]'=t]: 2102 <conf:(0.64)> lift:(1) lev:(-0) conv:(1)

75. [quality='(5-7]'=t]: 2988 ==> [volatile acidity='(-inf-0.27]'=t]: 1892 <conf:(0.63)> lift:(1.12) lev:(0.04) conv:(1.18)

76. [quality='(5-7]'=t]: 2988 ==> [chlorides='(0.036667-0.064333]'=t]: 1889 <conf:(0.63)> lift:(0.94) lev:(-0.03) conv:(0.88)

77. [quality='(5-7]'=t]: 2988 ==> [free sulfur dioxide='(-inf-38.666667]'=t]: 1889 <conf:(0.63)> lift:(1.02) lev:(0.01) conv:(1.04)

78. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [fixed acidity='(5.966667-8.133333]'=t, citric acid='(0.246667-0.493333]'=t]: 2126 <conf:(0.63)> lift:(0.99) lev:(-0) conv:(0.98)

79. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [free sulfur dioxide='(-inf-38.666667]'=t]: 2259 <conf:(0.63)> lift:(1.01) lev:(0.01) conv:(1.02)

80. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [volatile acidity='(-inf-0.27]'=t]: 2248 <conf:(0.62)> lift:(1.1) lev:(0.04) conv:(1.15)

81. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [chlorides='(0.036667-0.064333]'=t]: 2097 <conf:(0.62)> lift:(0.92) lev:(-0.04) conv:(0.86)

82. [pH='(3.086667-3.453333]'=t]: 3306 ==> [total sulfur dioxide='(121.333333-232.666667]'=t]: 2038 <conf:(0.62)> lift:(1.02) lev:(0.01) conv:(1.04)

83. [pH='(3.086667-3.453333]'=t]: 3306 ==> [fixed acidity='(5.966667-8.133333]'=t, residual sugar='(-inf-9.083333]'=t]: 2024 <conf:(0.61)> lift:(1.08) lev:(0.03) conv:(1.12)

84. [chlorides='(0.036667-0.064333]'=t]: 3168 ==> [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 1937 <conf:(0.61)> lift:(1.05) lev:(0.02) conv:(1.08)

85. [pH='(3.086667-3.453333]'=t]: 3306 ==> [free sulfur dioxide='(-inf-38.666667]'=t]: 2017 <conf:(0.61)> lift:(0.99) lev:(-0.01) conv:(0.98)

86. [citric acid='(0.246667-0.493333]'=t]: 3609 ==> [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 2169 <conf:(0.6)> lift:(1.04) lev:(0.02) conv:(1.05)

87. [residual sugar='(-inf-9.083333]'=t]: 3371 ==> [fixed acidity='(5.966667-8.133333]'=t, pH='(3.086667-3.453333]'=t]: 2024 <conf:(0.6)> lift:(1.03) lev:(0.01) conv:(1.05)

התוצאות של שני האלגוריתמים זהות – נמצאו אותם חוקי הקשר. רמת ה-confidence הגבוהה ביותר היא 0.88 ולא נמצאו חוקים טריוויאלים.

מדובר בשני אלגוריתמים דומים – ולכן לא מפתיע שהתוצאות הן דומות. ההבדלים העיקריים ביניהם הם דרך החיפוש, כשאפריורי מונחה ע"י יצירת מועמדים ואילו FP Growth מחפש אחרי תבניות שכיחות. אופן השימוש בתכונות דרש המרה שלהן על מנת שנוכל להריץ את FPGrowth.

עבור אותה הקונפיגורציה הנתונה האלגוריתמים מספקים תוצאות זהות.

**חלק 2 – ניתוח אשכולות**

1. ניתוח אשכולות היא טכניקה שעוסקת בפילוח אוכלוסיית אובייקטים גדולה לאוסף כלשהו של אשכולות בעלי הגיון פנימי כלשהו. אשכול היא קבוצת רשומות. המטרה בתהליך אשכול הא ליצור חלוקה שבה איברי האשכולות דומים זה לזה, בעוד האשכולות עצמם שונים זה מזה - אובייקטים שונים ישויכו לאשכולות שונים.

בסיום הניתוח אמורות להתקבל קבוצות ברורות של נתונים, אשר מראות על מגמה מסוימת, שפעמים רבות לא הייתה נודעת בצורת ניתוח אחרת.

קיימים שני סוגי אישכול עיקריים: אישכול היררכי אשר מותאם לפרטים המועמדים לאישכול, ויוצר היררכיה בין פריט אחד לאוסף הפריטים הסופי; אישכול חלוקתי המגדיר מראש את מספר החלוקות הנדרש ומשתמש בדרך כלל בחישובי מרחק לשם החלוקה.

1. בשביל מדידת איכות האשכולות נצטרך לבדוק את מידת הדמיון בין האובייקטים שבתוך האשכול, וכמו כן את מידת חוסר-הדמיון שבין האשכולות. לשם מדידת איכות האשכול אפשר להשתמש במטריקת דמיון(- הדמיון מתבטא במונחים של פונקציית מרחק) ובפונקציית איכות( – המודדת את טיב האישכול). סכום השגיאות המרובעות המתקבל בסיום הרצה של כל אלגוריתם יראה על אשכול טוב ככל שהסכום נמוך יותר.
2. קיימות שתי גישות עיקריות לאישכול: **חלוקה והיררכיה.**

באישכול חלוקה נעשה שימוש במדידות מרחק – בהתחלה בוחרים מספר נקודות, אקראיות או קבועות מראש, ובהמשך מצרפים אליהן את האובייקטים הקרובים ביותר. עבור כל איטרציה מחושב המרכז של כל אשכול מחדש, בהתחשב באובייקטים החדשים שצורפו אליו, ולפיו נמדד המרחק עבור אובייקטים חדשים המועמדים להצטרף לאשכול.

לשיטת אישכול חלוקה קיימות מספר שיטות טיפוסיות ליישום ניתוח האשכולות.

בחרתי בשיטת ה- **K-Means**. היתרון באלגוריתם זה הוא חוסר ההתערבות הסובייקטיבי ביצירת אשכולות.

האלגוריתם בוחר בעצמו את נקודות המרכז הראשוניות, בצורה אקראית - כך שמובטח כי יתקבלו אשכולות, אפילו אם הנתונים ההתחלתיים אינם מפולגים בצורה ברורה.

שלבי האלגוריתם:

* 1. מחלקים את האובייקטים המצויים בבסיס הנתונים ל-K אשכולות בצורה אקראית
  2. מחשבים ערך ממוצע "מוביל" לכל קבוצה K' על פי החלוקה שביצענו
  3. מבצעים בדיקת שיוך של האובייקטים בכל אזור אל מול ערכי הממוצע של כל אשכול. האובייקטים עם המרחקים/ההפרשים הקטנים אל מול אותו ממוצע ישויכו לאשכול K' המתאים
  4. במידה ושיוך האובייקטים בהתאם לערכי הממוצע השתנה חוזרים לסעיף 2
  5. תהליך מחזורי זה יסתיים רק לאחר שלא יתקיימו יותר תזוזות של האובייקטים בין האשכולות בהתאם להגדרת ערכי הממוצע

האלגוריתם הזה הוא נוח לשימוש וגמיש באופן יצירת הקבוצות המתאימות ולכן בחרתי בו.

באישכול היררכי נוצרים אשכולות מקוננים בצורה של עץ היררכי – הרמה העליונה היא סט הנתונים כולו, המהווה אשכול אחד גדול, והרמה התחתונה היא החלוקה הגדולה ביותר של האובייקטים בסט הנתונים, כאשר בכל אשכול מספר אובייקטים קטן יחסית.

בחרתי באלגוריתם ( EM (expectation maximization.

EM הוא אלגוריתם היררכי-הסתברותי (Probabilsic Hierarchical), אגלומרטיבי-מצטבר (bottom-up). האלגוריתם מקצה חלוקה הסתברותית עבור כל משתנה, לגבי סיכוייו להכלל בכל אחד מן האשכולות. הוא יכול להחליט אילו אובייקטים דומים מספיק כדי להכלל באותו אשכול, ומהי ההיררכיה שבין האשכולות.

האלגוריתם משתמש במשתנה LogLikelihood, שמחשב את ההסתברות הממוצעת עבור כל סט הנתונים, בכל שלב בו נבדקים אובייקטים נוספים מהסט. עלייה בערך המשתנה גורמת ליצירת אשכול חדש. תקציר האלגוריתם שנלקח מ-WEKA:

1. the number of clusters is set to 1

2. the training set is split randomly into 10 folds.

3. EM is performed 10 times using the 10 folds the usual CV way.

4. the loglikelihood is averaged over all 10 results.

5. if loglikelihood has increased the number of clusters is increased by 1 and the program continues at step 2.

אלגוריתמי אשכול חלוקה מחלקים את הרשומות לאשכולות כך שכל רשומה שייכת לאשכול אחד, ואלגוריתמי אשכול היררכיים יוצרים אשכולות מקוננים המסודרים כעץ היררכי.

היתרון של אלגוריתם היררכי הסתברותי על-פני אלגוריתמים היררכיים אחרים, הוא בעיקר בכך שאלגוריתמים היררכיים מחפשים אפשרויות אשכול בצורה מקומית-נקודתית, עבור כל שלב בו הם נמצאים, ואילו היררכי-הסתברותי בוחן עצמו תמיד ביחס גלובלי (ההסתברות הכללית).

1. ה. ו.

**k-Means**

נשתמש בקובץ הנתונים לאחר הניקוי (ממן 21), וכמו כן לאחר ביצוע הדיסקרזציה (שאלה 1). הפרמטרים הרלוונטים עבור אלגוריתם זה הם מספר האשכולות – k, פונקציית המרחק ו-seed. נריץ את האלגוריתם עבור

ערכי פרמטרים שונים ונבדוק את התוצאות:

|  |  |
| --- | --- |
|  | K = 10  SEED = 10  Euclidean distance |
|  | K = 5  SEED = 10  Euclidean distance |
|  | K = 3  SEED = 10  Euclidean distance |
|  | K = 3  SEED = 13  Euclidean distance |
|  | K = 3  SEED = 15  Euclidean distance |
|  | K = 3  SEED = 15  Manhattan distance |
|  | K = 3  SEED = 13  Manhattan distance |
|  | **K = 3**  **SEED =20**  **Euclidean distance** |

ניתן לראות שפונקצית המרחק, בהתבסס על הנתונים הקיימים, המדידה ע"י מרחק אוקלידי או מרחק מנהטן לא השפיעה על התוצאות. שלמרות שבחלוקה ל-10 אשכולות מספר השגיאות נמוך יותר – הרבה מהאשכולות הם קטנים ולא בעלי משמעות רבה. לאחר הרצות שונות, עם k = 3, 4, 5 ניתן לראות שכאשר בחרנו ב-3 אשכולות עם seed=20 מתקבלים 3 אשכולות יחסית בגודל זהה, ועם כמות שגיאות נמוכה במקצת ביחס להרצות האחרות של 3 אשכולות. נבדוק על המאפיינים השונים האם ניתן למצוא מאפיין שכאשר נבחר אותו ונתמקד בחלוקה לאשכולות לפיו- נקבל תוצאות טובות.

לאחר הרצת האלגוריתם על המאפיינים השונים, מצאתי מספר מאפיינים שבהם אחוז השגיאות הוא הנמוך ביותר:

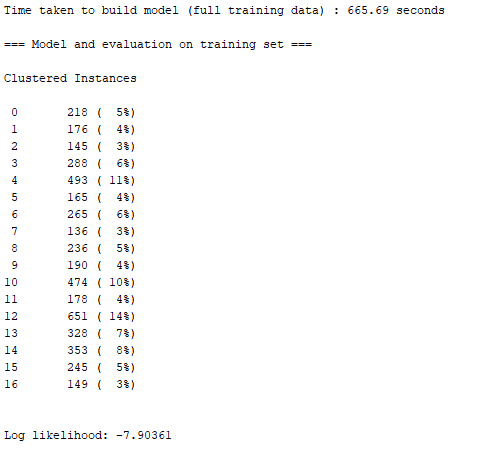
|  |  |
| --- | --- |
|  | מאפיין alcohol |
|  | מאפיין chlorides |
|  | מאפיין residual sugar |

כאשר ניסיתי להגדיל את מספר האשכולות ליותר מ-3 (4 ו-5) , וכמו כן כאשר הקטנתי את ערך ה-seed – ראיתי שאחוז השגיאה עלה משמעותית – יותר מ-60 אחוז שגיאה - דבר שחיזק את הבחירה שלי ב-3 אשכולות ובפרמטרים הנ"ל.

כאשר בחרנו במאפיין alcohol – כמות השגיאות שלו קטנה במקצת – אך ה- incorrectly clustered instances גדול במקצת – לעומת בחירה במאפיין chlorides – שבו כמות השגיאות גבוהה במקצת אבל אחוזי ה- incorrectly clustered instances קטנים במקצת.

**EM:**

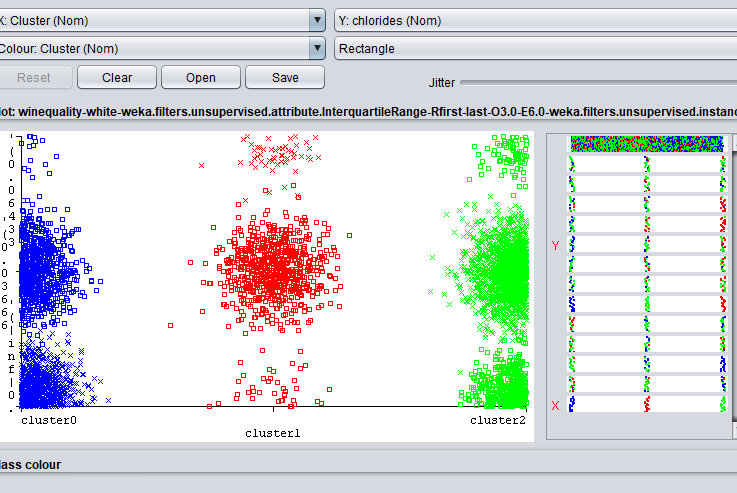
הרצת אלגוריתם זה לוקחת זמן ומשאבים רבים. תחילה, הרצתי אותו עם ערכים דיפולטים – לא הגדרתי לו את כמות האשכולות כך שאראה מה הפלט של האלגוריתם ללא שימוש בתוצאות ההרצה שמצאתי ע"י אלגוריתם k-means.



כעת, בהיעזר בנתונים שקיבלנו לאחר הרצת ה-k-means, נריץ את האלגוריתם ונגדיר לו 3 אשכולות. כמו כן נריץ כל פעם בהדגשת אחד מהמאפיינים שמצאנו, ונבדוק מה אחוז השגיאות.

|  |  |
| --- | --- |
|  | הרצה כללית – חלוקה ל3 אשכולות |
|  | הרצת 3 אשכולות , מאפיין acohol |
|  | הרצת 3 אשכולות , מאפיין chlorides |
|  | הרצת 3 אשכולות , מאפיין residual sugar |

ואכן ניתן לראות שקיבלנו תוצאה ממש טובה – כאשר בחרנו במאפיין chlorides – רק 41% שגיאות !



ובעצם בשאלה זו נעזרנו ב-2 סוגי האלגוריתמים כדי לקבל את התוצאה הטובה ביותר – תחילה , הרצנו את אלגוריתם k-means , קיבלנו תוצאות טובות – ולאחר מכן נעזרנו בתוצאות של אלגוריתם זה כדי להגדיר את הפרמטרים לאלגוריתם EM – ובאמצעות פרמטרים אלה קיבלנו תוצאות אפילו טובות יותר.

**שאלה 3 -**

**מסקנות:**

* תחילה, כדי לדעת לכרות את המידע בצורה הטובה ביותר ולמצות אותו, חייב לדעת מה המטרה, מה אנחנו מחפשים לדעת – מי הלקוח ובמה הוא מעוניין – כך שבאמת נוכל לתת פיתרון ותשובה שיהיו משמעותיים עבורו ועבור צרכיו. בממן 21 והממן הנוכחי קיבלנו קובץ ומעליו הנחתי הנחות – אך במצב רגיל זה לא יעבוד כי חייבת להיות מטרה ברורה לכריית המידע. הרגשתי פער בהבנה שלי של הנתונים ומה אנחנו מעוניינים להוציא מהם.
* כל סט נתונים ניתן לחקור ולשפר כל הזמן – ולכן כדאי להגדיר במידת האפשר מטרות שניתן למדוד אותן – דבר שיעזור לנו להגדיר את תוצאות כריית המידע בצורה טובה.
* קיימים אלגוריתמים רבים ושונים, וכל אחד מהם מייצר פלט שונה. כדי שנוכל לתת את התוצאות הטובות ביותר חשוב שנכיר את האלגוריתמים ושנדע לשחק עם הפרמטרים שלהם – כך שנדע למצות את המידע ונקבל תוצאות טובות. כל אלגוריתם יכול להועיל במצבים שונים – לא הכל עובד כמו שהיינו מצפים שהוא יעבוד – ולכן יש הרבה ניסוי וטעייה בתהליך הזה. במהלך הפרוייקטים השונים הרצתי הרבה פעמים את האלגוריתמים בפרמטרים שונים – ולא תמיד כאשר אני מבינה לעומק מה המשמעות של השינויים האלה – כך שלמדתי עם ההרצות איך זה משפיע. כדי לדעת להריץ בצורה נכונה את האלגוריתמים חשוב רק להבין מה משמעות הפרמטרים השונים, וכמו כן להבין את המגבלות והיתרונות של כל מודל ושיטה ולהבין איך הם משפיעים על המקרה שלנו.