בס"ד

למידת מכונה – תרגיל מספר 2

הפיכת הפיצ'ר המציין את סוג היין מפיצ'ר קטגוריאלי לפיצ'ר מספרי:

בכל דוגמה שבה סוג היין הוא W הצבתי במקום W את הספרה 1 ובכל דוגמה שבה סוג היין הוא R הצבתי במקום R את הספרה 0 .

יצירת סט אימון וסט ולידציה:

על מנת שאוכל לקבוע hyper-parameters ולאמן את המודל, השתמשתי ב k-folds באמצעות חבילה שקיימת בפיתון, עם k=7. כיוון שאת הקוד בו בחנו את תוצאות הדיוק לא צריך להגיש, השתמשתי בחבילה החיצונית הזו.

נרמול לסט האימון והמבחן:

בחנתי 2 שיטות נירמול שונות: Min-Max ו Z-score.

להלן תוצאות הדיוק עבור כל אחת משיטות הנרמול:

שיטת נירמול	אלגוריתם	דיוק על סט הולידציה
Min-Max	KNN	80%
	Perceptron	71.3%
	Passive_aggressive	72.1%
z-score	KNN	66.5%
	Perceptron	67.6%
	Passive_aggressive	44.8%

ניתן לראות בבירור כי **Min-Max** מניב תוצאות גבוהות יותר לכן בחרתי להשתמש בנרמול זה.

feature selection – ברירת פיצ'רים:

השתמשתי בכלי weka על מנת לבחון אילו פיצ'רים ניתן להוריד על מנת לקבל דיוק גבוהה יותר.
לפי אלגוריתם classifier attribute eval עם ranker נתן תוצאות כך שעלי להחסיר את פיצ'ר מספר 0 בלבד.

תוצאות weka לפי ranker :

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 12 class):
        Classifier feature evaluator

        Using   Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.rules.ZeroR
        Scheme options:
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5


Ranked attributes:
  0  11 alcohol
  0   3 citric
  0   2 volatile
  0   5 chlorides
  0   4 residual
  0   6 free
  0  10 sulphates
  0   9 pH
  0   8 density
  0   7 total
  0   1 fixed_acidity

Selected attributes: 11,3,2,5,4,6,10,9,8,7,1 : 11
```

הרצת ה DATA עם feature selection :

feature selection	אלגוריתם	דיוק על סט הולידציה
סט האימון והמבחן עם כל הפיצ'רים	KNN	80%
	Perceptron	71.3%
	Passive_aggressive	72.1%
ללא פיצ'ר מספר 0	KNN	79.4%
	Perceptron	68.2%
	Passive_aggressive	70.7%

תוצאות weka לפי best first :

```
Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 71
        Merit of best subset found:    0.494

Attribute Subset Evaluator (supervised, Class (nominal): 12 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,4,5,6,7,8,9,10 : 9
                     volatile
                     citric
                     residual
                     chlorides
                     free
                     total
                     density
                     pH
                     sulphates
```

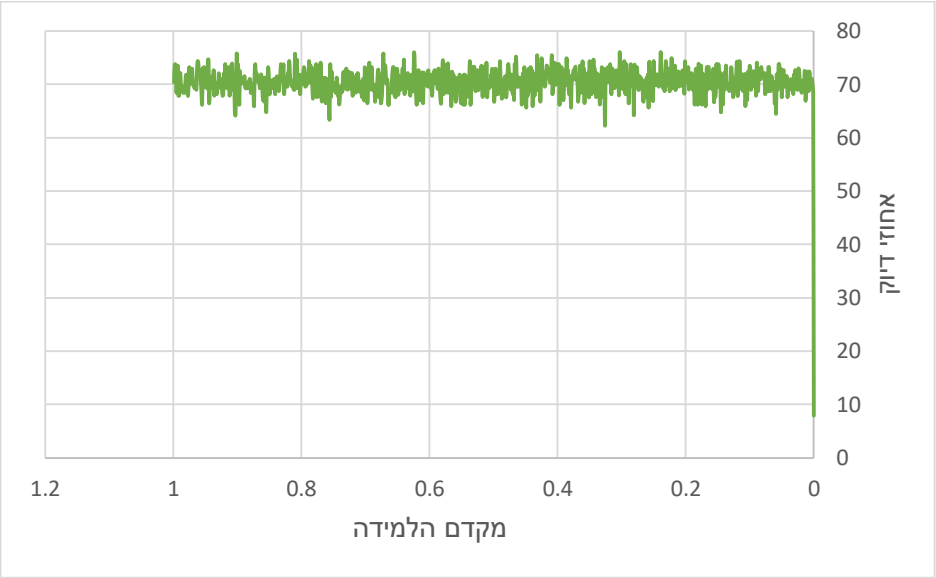
הרצת ה DATA עם feature selection :

	אלגוריתם	דיוק על סט הולידציה
סט האימון והמבחן עם כל הפיצ'רים	KNN	80%
	Perceptron	71.3%
	Passive_aggressive	72.1%
ללא פיצ'רים מספר 0, 10, 11	KNN	76.6%
	Perceptron	67.3%
	Passive_aggressive	71%

ביצוע feature selection על סט האימון והמבחן **לא שיפר את תוצאות הדיוק, לכן בחרתי שלא לממש feature selection.**

בחירת מקדם הלמידה:

בדקתי דיוק על מקדמי למידה שונים בין 0 לבין 0.999 בסט הולידציה ולהלן התוצאות:

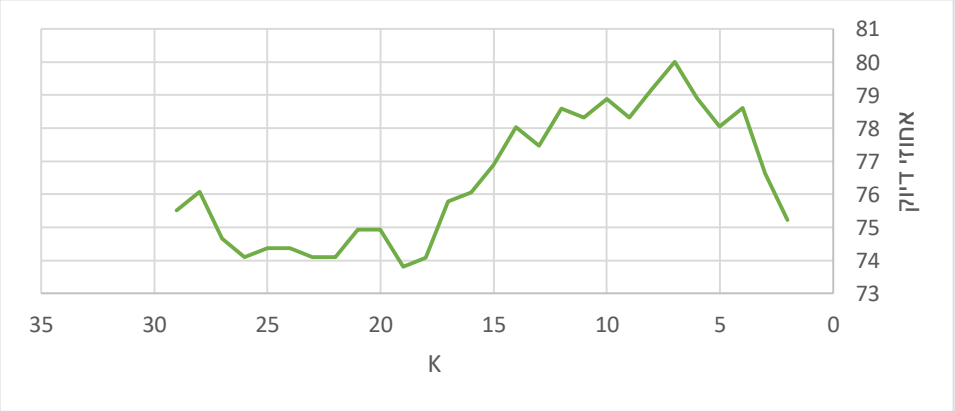


דיוק מקסימלי שהתקבל :

מקדם הלמידה	דיוק
0.239	76.03921569

בחירת K עבור מימוש KNN:

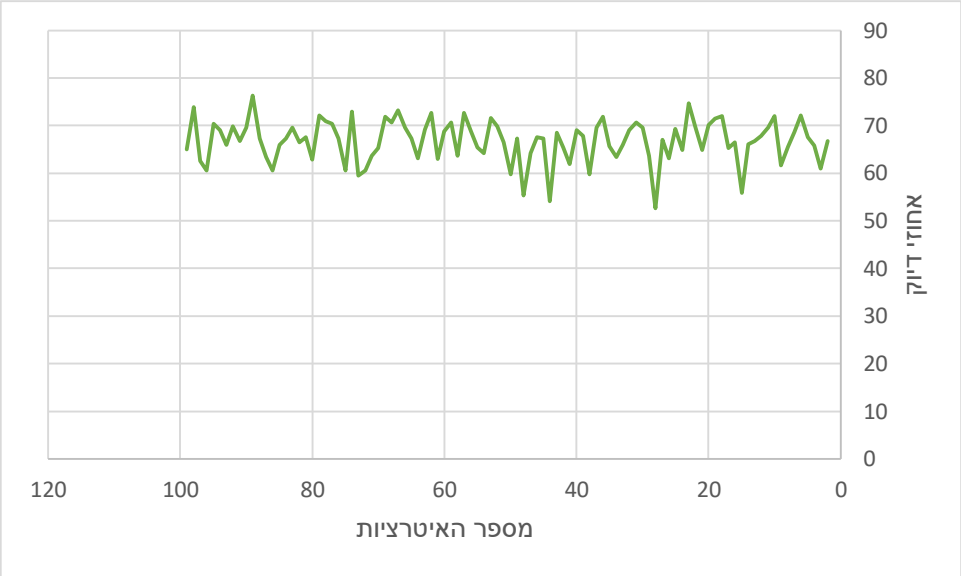
בדקתי דיוק עבור ערכי K שונים באלגוריתם KNN ולהלן התוצאות:



דיוק מקסימלי שהתקבל :

K	דיוק
7	80.0056

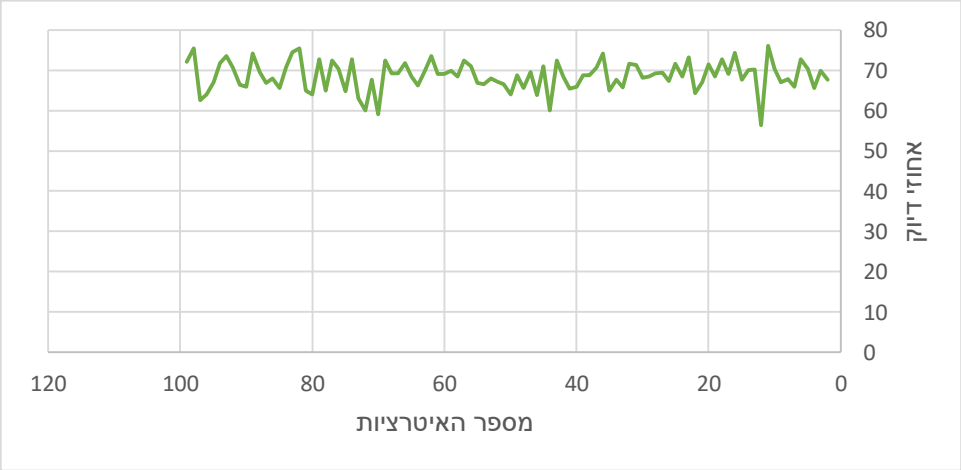
בחירת מספר האיטרציות עבור מימוש האלגוריתם Passive Aggressive:



דיוק מקסימלי שהתקבל :

מספר האיטרציות	דיוק
89	76.28011

בחירת מספר האיטרציות עבור מימוש האלגוריתם Perceptron:



דיוק מקסימלי שהתקבל :

מספר האיטרציות	דיוק
11	76.03922

סה"כ תוצאות הדיוק עבור כל אחד מהאלגוריתמים עם hyper-parameters שנקבעו לפי ההסברים הקודמים :

אלגוריתם	דיוק
KNN	80.0056
Perceptron	76.03922
Passive Aggressive	76.28011