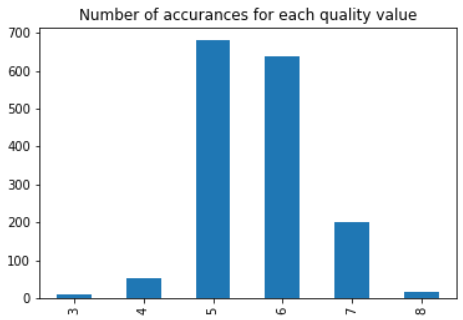
הגשה פרוייקט למידת מכונה:

שם המשתתפים: דביר סעדון

תיאור של המאגר: המאגר מתאר מידע לגבי יין אדום מסוג "Vinho Verde"

המאגר מכיל 12 פיצ'רים: הם מכילים בינהם כאלה שמתארים כמות של סוגי חומצה שונים שנמצאים ביין כמו גם כמו גם סוכר ומלח. בנוסף, כמות הגופרית הדו חמצנית ביין, ריכוז האלכוהול, צפיפות היין ו-PH והמשתנה שארצה לחזות – האיכות של היין מ0-10.

במידע שניתן קיימת התפלגות לא אחידה של הנתונים של האיכות של היין. ניתן לראות זאת בגרף הבא:



השאלות שעליהם עניתי: עניתי על השאלה של חיזוי (classification) של הדירוג איכות של היין (0-10). ניסיתי גם למצוא את המשוואה לregression של האיכות. ניסיתי לענות על אילו הם הפיצרים הכי טובים כדי לענות על השאלה הראשונה.

בנוסף, ניסיתי לבדוק אם אוכל לייצר תוצאות טובות יותר אם אבצע קלסיפיקציה בין קבוצות של תוצאות (1-3) – יין רע, טוב וטוב מאוד.

הטכניקות בהם השתמשתי: Random Forest, KNN, Adaboost וגם Decision tree.

בכדי לענות על השאלה השנייה, השתמשתי בLogistic regression. בנוסף, ניסיתי להריץ PCA ולראום אם אני אוכל לשפר את המודל כך.

אתגרים: האתגר הכי גדול שנתקלתי בו היה לשפר את המודל מ accuracy של כ55% לכ-88%.

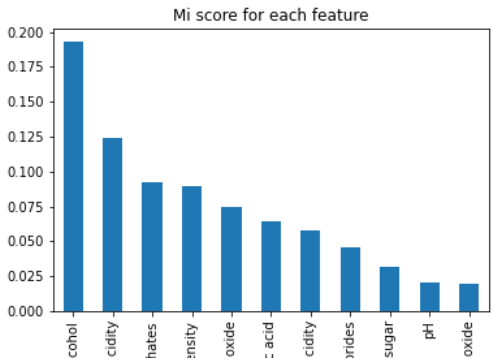
בנוסף, היה לי אתגר לענות על אחת השאלות שתכננתי ("אנסה גם למצוא את המשוואה לregression של האיכות") ולקח לי זמן רב להבין איך אני מוצא את המשוואה.

טכניקות שלא עבדו/ עבדו פחות טוב: טכניקה אחת שעבדה פחות טוב היא Adaboost . המודלים האחרים שהרצתי נתנו תוצאה טובה הרבה יותר ממנה. אני חושב שהסיבה הגדולה ביותר לזה היא שהמודל עושה overfitting ולכן נותן תוצאות פחות טובות.

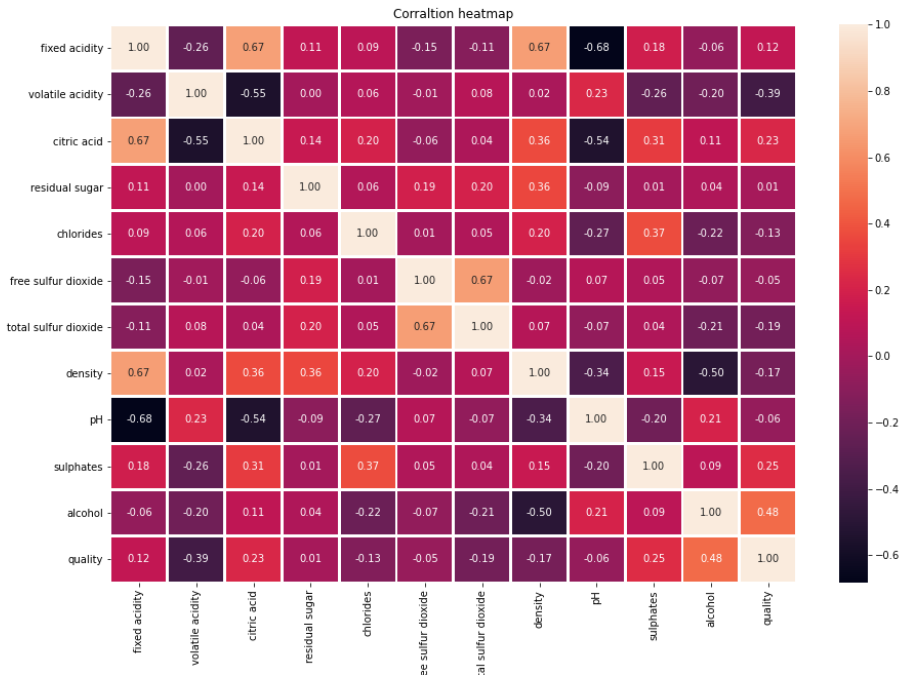
בנוסף, PCA לא הצליח לשפר את המודל. אני חושב שהסיבה לזה היא שהוצאתי את הפיצ'רים הפחות חשובים וכל האלה שנשארו הם חשובים מספיק למודל כך שניתן לראות שכשהורדנו עוד פיצ'רים המודל נהיה פחות טוב.

<u>טכניקות שהצליחו</u>: המודל הכי טוב שהורץ הוא Random Forest Classifier (~88%) אך גם KNN וגם Decision tree נתנו תוצאות יחסית טובות בקלסיפיקציה. המודל שנתן את התוצאה ברגרסיה הוא Logistic regression (Score = 0.63)

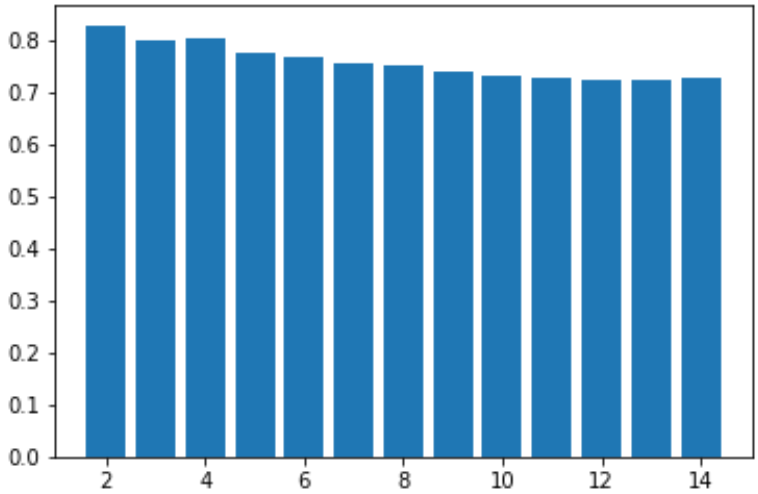
כדי לנסות לענות על השאלה של איזה פיצ'רים הכי טובים כדי לחזות מה הערך איכות של היין ניסיתי את השיטות הבאות: MI :



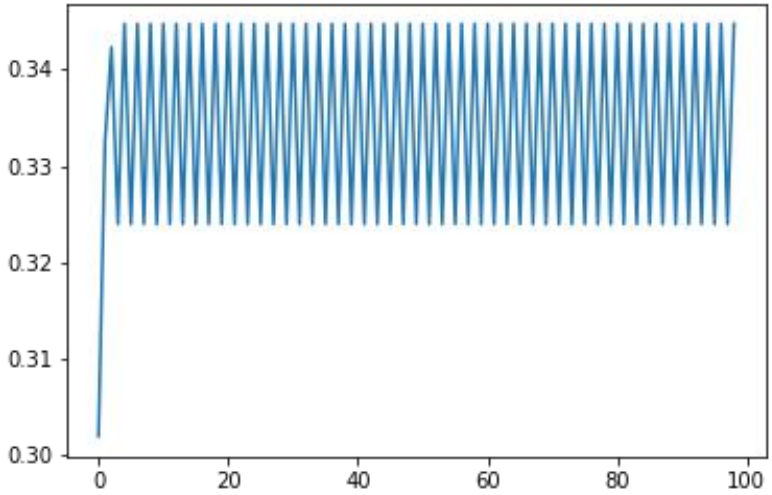
וגם correlation:



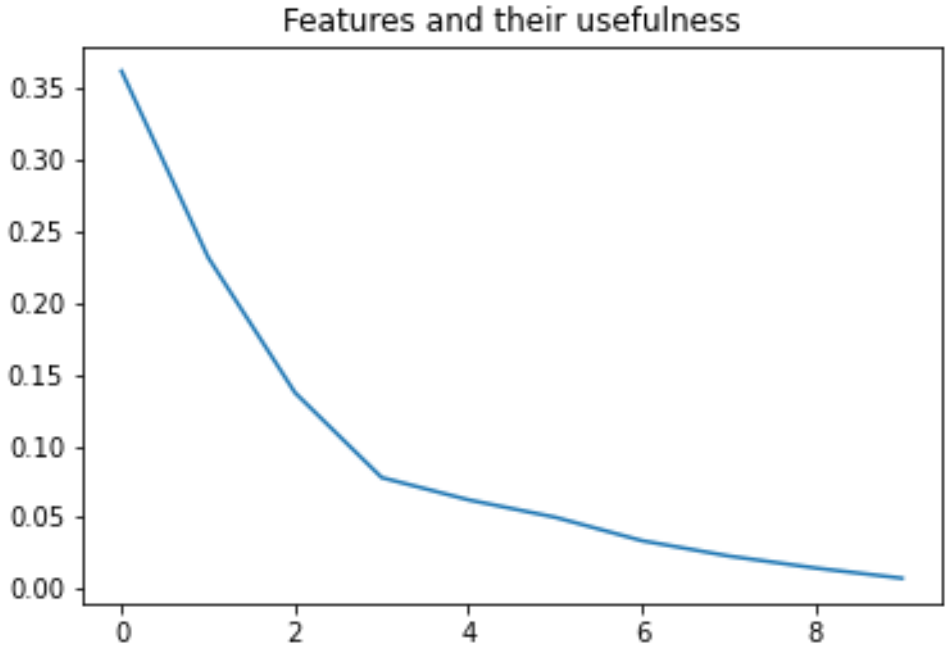
תוצאות של KNN:



ו- ADABOOST :



ניסיתי גם להריץ pca אך נראה שהוא רק הפגע במודל:



המודל שנתן את התוצאות הכי טובות היה Random forest:

```
Accuracy: 0.871638141809291
              precision    recall  f1-score   support

           3       0.98      1.00      0.99       132
           4       0.91      0.98      0.95       123
           5       0.81      0.77      0.79       154
           6       0.70      0.62      0.66       136
           7       0.83      0.89      0.86       135
           8       0.98      1.00      0.99       138

    accuracy                           0.87       818
   macro avg       0.87      0.88      0.87       818
weighted avg       0.87      0.87      0.87       818
```

לבקשת ליעד, הרצתי גם SVM ותוצאות היו:

```
Accuracy: 0.7408312958435208
              precision    recall  f1-score   support

           3       0.90      1.00      0.95       132
           4       0.71      0.89      0.79       123
           5       0.69      0.56      0.62       154
           6       0.48      0.40      0.44       136
           7       0.70      0.63      0.66       135
           8       0.87      1.00      0.93       138

    accuracy                           0.74       818
   macro avg       0.73      0.75      0.73       818
weighted avg       0.73      0.74      0.73       818
```

נראה שהתוצאות של SVM היו פחחות טובות מ Random Forest גם בשימוש עם kernels שונים (מוצג למעלה התוצאה הכי גבוהה).

הרצתי גם Decision tree והתוצאות שלו היו פחות טובות מהRandom Forest אך עדיין לא רעות:

```
Accuracy: 0.7958435207823961
              precision    recall  f1-score   support

           3       0.97      0.97      0.97       132
           4       0.78      0.85      0.81       123
           5       0.70      0.61      0.65       154
           6       0.59      0.56      0.57       136
           7       0.80      0.88      0.84       135
           8       0.92      0.94      0.93       138

    accuracy                           0.80       818
   macro avg       0.79      0.80      0.80       818
weighted avg       0.79      0.80      0.79       818
```

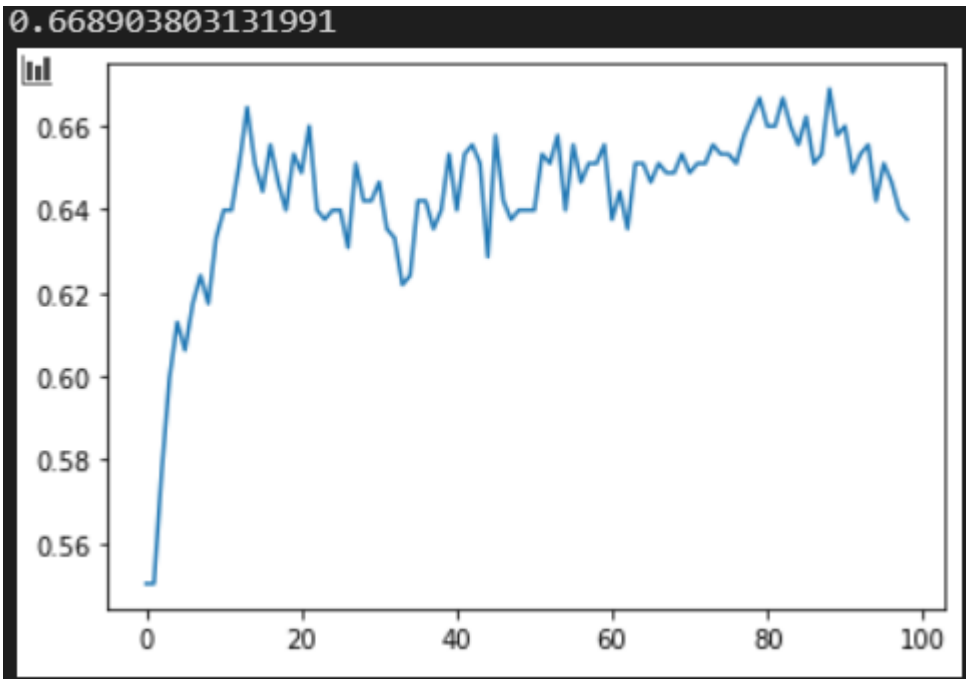
הרצתי את כל המודלים הנ"ל גם על מידע כך שהאיכות מחולקת ל -3 חלקים(איכות לא טובה, איכות טובה ואיכות מצויינת). המודל הכי טוב שיצא הוא Random forest:

```
Accuracy: 0.814317673378076
              precision    recall  f1-score   support

           1       0.81      0.79      0.80       146
           2       0.78      0.68      0.73       149
           3       0.85      0.96      0.90       152

    accuracy                           0.81       447
   macro avg       0.81      0.81      0.81       447
weighted avg       0.81      0.81      0.81       447
```

המודל ADABOOST הגיע לתוצאות יותר טובות מאשר הקלסיפיקציה הקודמת אך עדיין לא טובות:



אחת מהמטרות שלי היו להדפיס את הפונקציה שLogistic regression מחשב, והתוצאות היו:

```
Accuracy: 0.6

1/(1 + exp(-(2.0766995555367567 + 0.1767964626474786*fixed acidity + 2.394775100648505*volatile acidity + -0.2173374498011205*citric acid + 0.35259528619028
19*chlorides + 0.05319018036726719*free sulfur dioxide + -0.04618118538871855*total sulfur dioxide + 0.02918577918769722*density + 0.691437844289132*pH + -
0.4961210991936624*sulphates + -0.7675541567113382*alcohol))

1/(1 + exp(-(2.0766995555367567 + -0.2529448596350562*fixed acidity + 1.9321958218001831*volatile acidity + 0.2660287186851794*citric acid + 0.1428641879581
6852*chlorides + -0.020730635779108657*free sulfur dioxide + -0.00466089584195134*total sulfur dioxide + -0.000673214408010338 3*density + 0.6876419117961449
*pH + -0.9472140266499739*sulphates + -0.4480518991219013*alcohol))

1/(1 + exp(-(2.0766995555367567 + -0.10186595272941004*fixed acidity + 0.6409637963042983*volatile acidity + 0.18879716714860023*citric acid + 0.79560241648
68657*chlorides + -0.020207220603104196*free sulfur dioxide + 0.026822913791096036*total sulfur dioxide + -0.01223459872290774 9*density + -0.179774684761526
47*pH + -1.5308169243694867*sulphates + -0.7656510735501436*alcohol))

1/(1 + exp(-(2.0766995555367567 + 0.031479103665247934*fixed acidity + -1.4319216256986398*volatile acidity + -0.6614518165250921*citric acid + -0.105538111
09115474*chlorides + -0.001140693181859950 4*free sulfur dioxide + 0.011395328541430754*total sulfur dioxide + 0.12388408849822359*density + 0.18006770052267
126*pH + 0.22913012916206035*sulphates + 0.040756309184661894*alcohol))

1/(1 + exp(-(2.0766995555367567 + 0.08355504889339951*fixed acidity + -3.0013312720704564*volatile acidity + 0.02811424936671981*citric acid + -1.0231361107
60197*chlorides + -0.002632710900639388*free sulfur dioxide + 0.005122769708924411 5*total sulfur dioxide + -0.0962348831153170 5*density + -0.476318230666997
74*pH + 1.8514470794945281*sulphates + 0.7794968875982069*alcohol))

1/(1 + exp(-(2.0766995555367567 + 0.06298019715833933*fixed acidity + -0.5346818209805501*volatile acidity + 0.395849131126533*citric acid + -0.162387668782
08635*chlorides + -0.008478919902568997*free sulfur dioxide + 0.007501069189142956*total sulfur dioxide + -0.04392717143989566*density + -0.903054541177912*
pH + 0.8935748415563215*sulphates + 1.1610039325997183*alcohol))
```

ה Accuracy נמוך אך המטרה המרכזית הייתה להצליח להבין איך להדפיס את הנוסחה (ולהבין את הנוסחה עצמה יותר טוב) וזה הצליח.