**שלבים בהכנת ה-Dataset**

1. הסרת שורות ריקות (null בעמודות של ביטוי הגנים).
2. "איחוד" של פרובים שונים השייכים לאותו גן כך שתהיה שורה אחת לכל גן בדאטאסט, רמות הביטוי יחושבו על-פי ממוצע.
3. הסרת גנים שלא מקושרים לאף go term.
4. הורדנו רשימה של go terms המקושרים ל-apoptosis: <http://amigo.geneontology.org/amigo/search/ontology?q=apoptosis>  
   ותייגנו כל גן כ-TRUE אם ברשימת ה-terms המקושרים אליו יש terms המקושרים ל-apoptosis, ו-FALSE אחרת.

סה"כ: 1728 גנים שמקושרים ל-terms של apoptosis, מתוך 16080 גנים מתויגים.

מס' ה-terms בדאטאסט שמקושרים ל-apoptosis הוא 150.

\*\* הרבה מה-terms שמקושרים ל-apoptosis הם דווקא "negative regulation of apoptosis" וניסוחים דומים, האם הגיוני להתייחס אליהם באותה צורה?

אחרי ניסיון הרצה ראשון של decision tree (ללא כוונון פרמטרים):

accuracy score: 0.7949834162520729

recall score: 0.11560693641618497

precision score: 0.1016949152542373

אחרי ניסיון הרצה ראשון של SVM (ללא כוונון פרמטרים):

accuracy score: 0.8853648424543947

recall score: 0.03468208092485549

precision score: 0.2571428571428571

* כלומר יש דיוק יחסית גבוה אבל זה לא אומר שהביצועים באמת טובים (פשוט כי יש הרבה יותר FALSE בלייבלים)

שימוש בדאטאסט המקורי (מס' פרובים לכל גן), לעומת בניית דאטאסט ע"י שימוש בערכים ממוצעים של רמות ביטוי על פני הפרובים של כל גן, לעומת בחירה של פרוב יחיד מכל גן (ללא ממוצע):

|  |  |  |
| --- | --- | --- |
| כל הפרובים | ללא ממוצע | ממוצע |
| Decision Tree  accuracy score: 0.8038971807628524  recall score: 0.12138728323699421  precision score: 0.11392405063291139  SVM  accuracy score: 0.8374792703150912  recall score: 0.10982658959537572  precision score: 0.1503957783641161  KNN with N=3  accuracy score: 0.8638059701492538  recall score: 0.019267822736030827  precision score: 0.06329113924050633 | **ID3**  score: 0.8130182421227198  recall score: 0.10982658959537572  accuracy score: 0.11468812877263582  **SVM**  score: 0.8874378109452736  recall score: 0.019267822736030827  precision score: 0.22727272727272727  **KNN with N=3**  score: 0.87106135986733  recall score: 0.025048169556840076  accuracy score: 0.10077519379844961 | Decision Tree  accuracy score: 0.7941542288557214  recall score: 0.1001926782273603  precision score: 0.08996539792387544  SVM  accuracy score: 0.8146766169154229  recall score: 0.13102119460500963  precision score: 0.13307240704500978  KNN with N=3  accuracy score: 0.8646351575456053  recall score: 0.036608863198458574  precision score: 0.11046511627906977 |

נראה שההבדל לא משמעותי כל כך (כל התוצאות באותו סדר גודל). לכן נמשיך עם ה**ממוצע**.

כיוונים להמשך:

* לעשות צעדי data preparation למשל טרנספורמציית לוג, נרמול...
* כוונון פרמטרים
* להוסיף heat map בסגנון של שיעורי הבית / מה שראינו במצגות

Preprocessing:

* Standard scaler: (מחסירים ממוצע ומחלקים בסטיית התקן)

התוצאות –

Decision Tree

accuracy score: 0.7999585406301825

recall score: 0.10211946050096339

precision score: 0.09601449275362318

SVM

accuracy score: 0.8018242122719734

recall score: 0.22350674373795762

precision score: 0.17339312406576982

KNN with N=3

accuracy score: 0.8644278606965174

recall score: 0.02697495183044316

precision score: 0.08588957055214724

* Log-transform: (=> גרוע, הסבר אפשרי הוא שמלכחילה ההפרשים בין רמות הביטוי אינם דרסטיים כמו שראינו בשיעורי הבית למשל)

Decision Tree

accuracy score: 0.8030679933665008

recall score: 0.10211946050096339

precision score: 0.09869646182495345

SVM

accuracy score: 0.8924129353233831

recall score: 0.0

precision score: 0.0

KNN with N=3

accuracy score: 0.8640132669983416

recall score: 0.028901734104046242

precision score: 0.08982035928143713

הקשרים ביולוגיים:

1. מתוך המאמר: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470986/

failure to activate apoptosis are major contributors leading to malignant cellular transformation [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470986/#pone.0130300.ref002), [3](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4470986/#pone.0130300.ref003)]. Identification of signal transduction targets for apoptosis induction is therefore of importance to provide novel opportunities for therapeutic approaches.

1. מתוך: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2785190/

Protein Kinase Cδ Supports Survival of MDA-MB-231 Breast Cancer Cells by Suppressing the ERK1/2 Pathway

בהתאמה, down regulation של PKCD גורמת ל-apoptosis של תאים סרטניים.

* protein kinase C הוא signaling pathway ולכן מהווה 'מטרה' שניתן לתקוף על מנת למנוע השרדות של תאים סרטניים.

הוספות שהוצעו על ידי לאון

* ~~Check which go terms are more associated with apoptosis and which less by checking which go terms appear mostly in genes labeled true.~~
* Do semi supervised clustering by clustering 70% of the genes and then predicting the rest.
* Check whether the 82% result of genes being in the same cluster is due to the cluster being too big (having most of the genes) or if the result is actually good
* Low priority: Classify using the diff between features. Requires finding out information about the features (samples) by searching them in geo database.
* Make sure to mention that our matrix is transposed!
* ~~Undersampling~~
* ~~Filter genes~~
* After classifying: we’d like to see which genes we classified correctly and which we didn’t, see if they have something in common etc. Also we can pick a few such genes and try to explain why possibly we succeeded or failed in classifying them.

**GSM1338298 -** BT-549 breast cancer cell line  
treatment: siPRKCD

**GSM1338302 -** BT-549 breast cancer cell line  
treatment: siPRKCD

**GSM1338306 –** BT-549 breast cancer cell line  
treatment: siPRKCD

**GSM1338297 -** BT-549 breast cancer cell line  
treatment: siControl

**GSM1338301 -** BT-549 breast cancer cell line  
treatment: siControl

**GSM1338305 -** BT-549 breast cancer cell line  
treatment: siControl

**GSM1338296 -** MDA-mB-468 breast cancer cell line  
treatment: siPRKCD

**GSM1338300 -** MDA-mB-468 breast cancer cell line  
treatment: siPRKCD

**GSM1338304 -** MDA-mB-468 breast cancer cell line  
treatment: siPRKCD

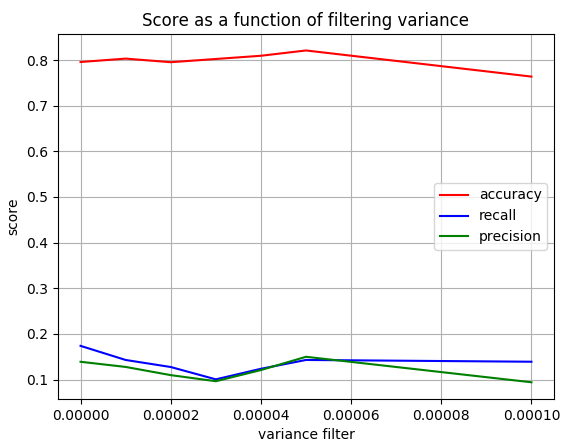
**GSM1338295 -** MDA-MB-468 breast cancer cell line  
treatment: siControl

**GSM1338299 -** MDA-MB-468 breast cancer cell line  
treatment: siControl

**GSM1338303 -** MDA-MB-468 breast cancer cell line  
treatment: siControl

קבלת הדאטאסט:

1. סינון גנים:



הערך שנבחר: 0.00005.

Variance threshold: 5e-05

after filtering, train size is: 8557 test size is: 2412

TRUE in train size: 1005 TRUE in test size: 259

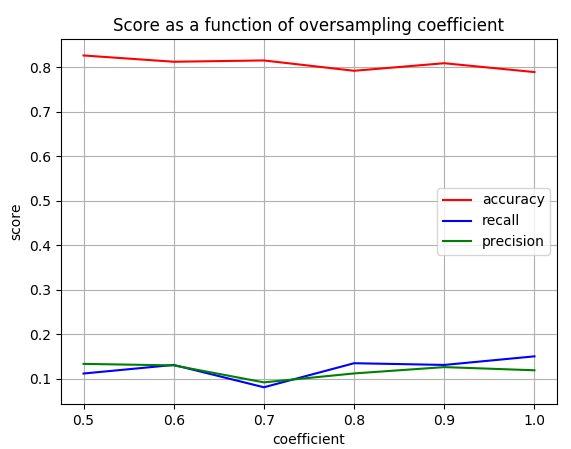
Decision Tree

accuracy score: 0.8208955223880597

recall score: 0.14285714285714285

precision score: 0.14979757085020243

1. Oversampling:



coef is: 1

after oversampling, train size is: 15104

TRUE in train size: 7552

Decision Tree

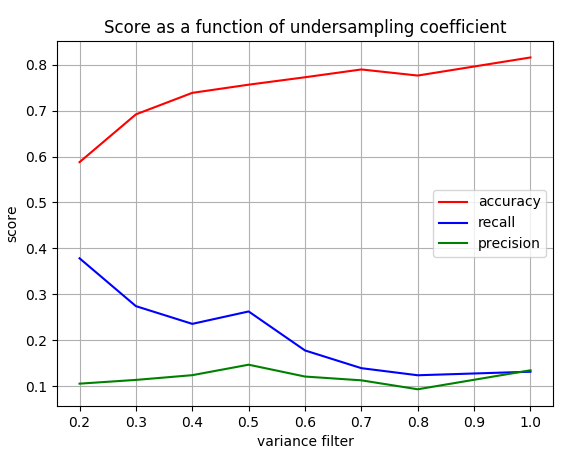
accuracy score: 0.7893864013266998

recall score: 0.15057915057915058

precision score: 0.11926605504587157

ניתן לראות כי התוצאות לא השתפרו, אך ניקח את הגרסה הזו מכיוון שהדאטאסט מאוזן יותר ואנו מעריכים כי זה יעזור בהמשך.

1. Undersampling:



coef is: 0.5

after undersampling, train size is: 4781

TRUE in train size: 1005

Decision Tree

accuracy score: 0.7566334991708126

recall score: 0.2625482625482625

precision score: 0.14655172413793102

אמנם התוצאות יותר טובות אך מבדיקה נוספת במסווגים אחרים, וכן השיקול של הקטנת הדאטאסט, החלטנו לא להשתמש ב-undersampling.

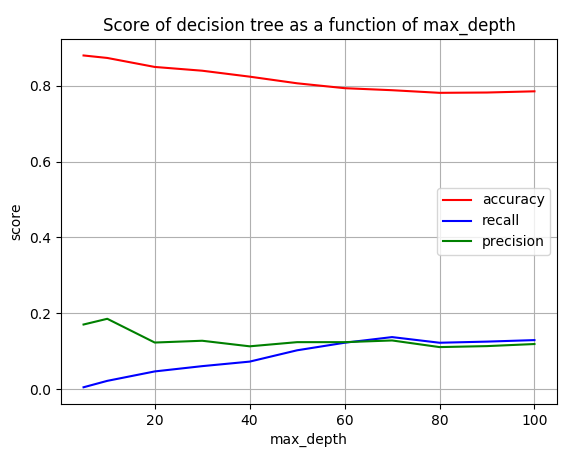
**נבחר: filter+oversample with coef 1**

**train\_oversampled\_1.csv**

\*\* ניסינו גם הרצות עם רק חלק מהדוגמאות, וכן הרצות עם ההפרש בין הדוגמאות, אך לא נראה שיפור ולכן לא נפרט (ניתן לציין בפוסטר ללא פירוט).

כוונון פרמטרים על עץ החלטה

Max\_depth:



VAL is: 70

accuracy score: 0.7883585868241232

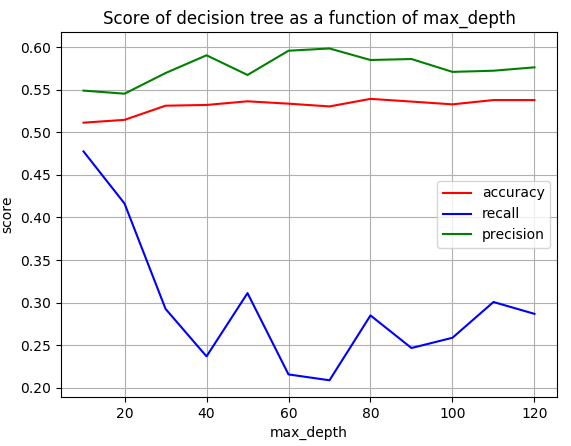
recall score: 0.13731343283582093

precision score: 0.1284234259938783

SVM:

Max\_iter:

ניקח את הערך של 20 (לשים לב שהמדדים כאן קצת "אופטימיים" כי השתמשתי ב-KFOLD ובגלל ה-oversampling יש דוגמאות שהופיעו גם באימון וגם בטסט).



VAL is: 20

accuracy score: 0.5145659424704485

recall score: 0.41630778266224294

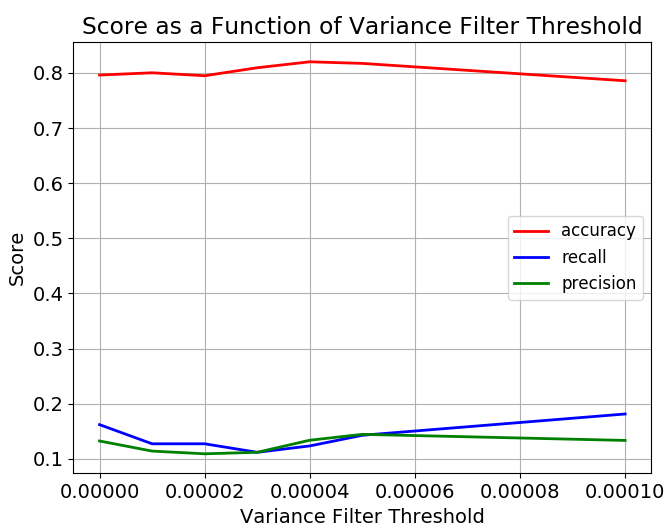
precision score: 0.5452790369493237

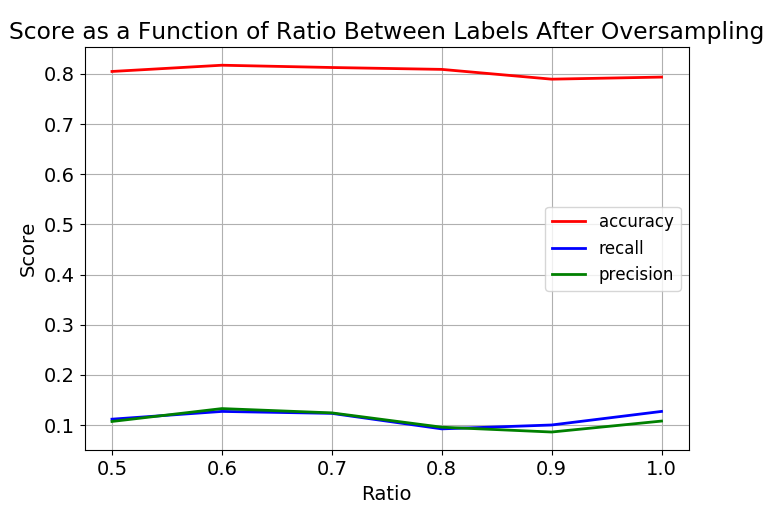
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Conclusions |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

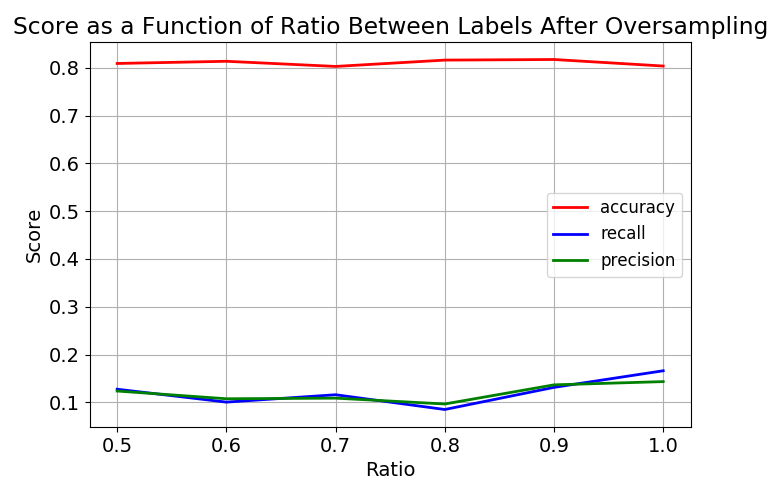
Let’s consider the gene EIF1 which was classified wrong and compare it to TRMT5 which was classified right. As we can see in the part which was marked in orange (which represents controlled MDA-MB cells) the genes TRMT5 and EIF1 have 0.2-0.5 difference for each sample. The blue part which represents the PKC treated MDA-MB cells shows 0.7-1.2 difference for each sample, meaning that EIF1 gene was more affected by the treatment than TRMT5. We can learn from that that either the original labeling FALSE for EIF1 was wrong (because we selected the labeling based on wrong go terms?) or that our classifier considered this change drastic enough to classify as TRUE even though the change is not big enough to be considered true.

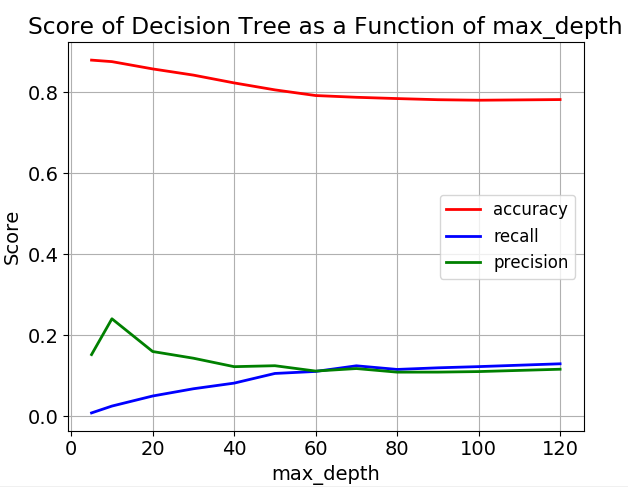
Another thing to note is that the classifier is not aware that some samples are controlled and some are treated so his classification only considers the raw data and not the absolute or relative difference in classification.

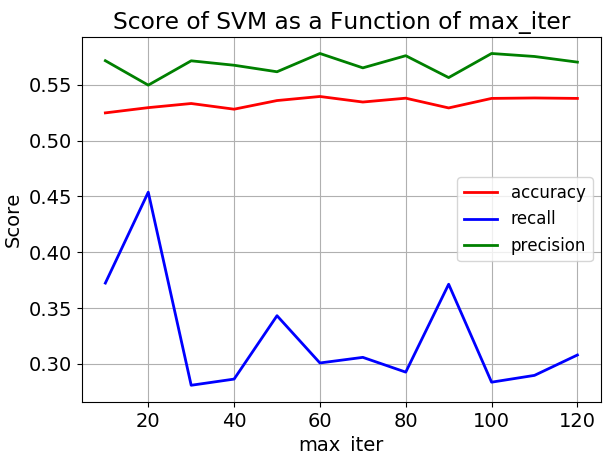
הדפסה מחודשת של הגרפים

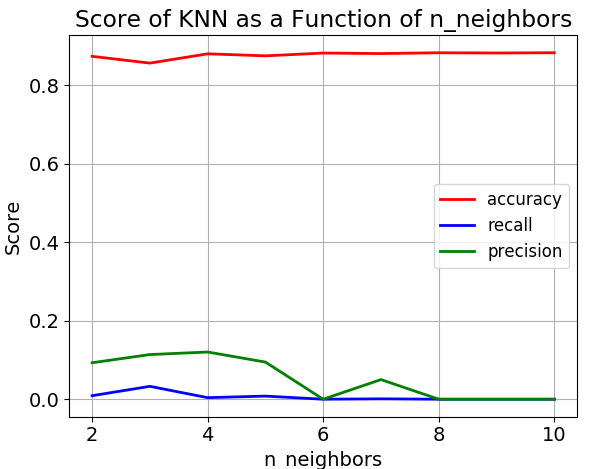


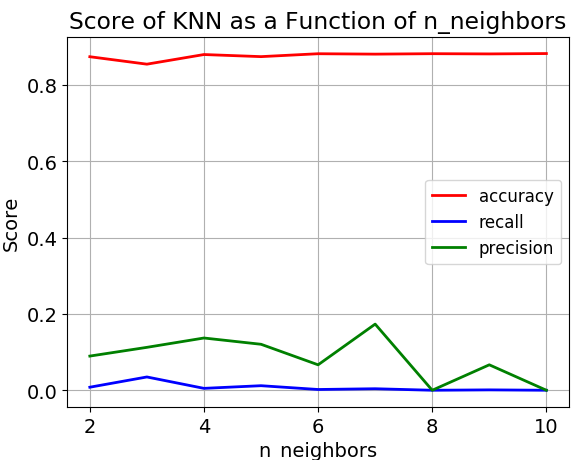


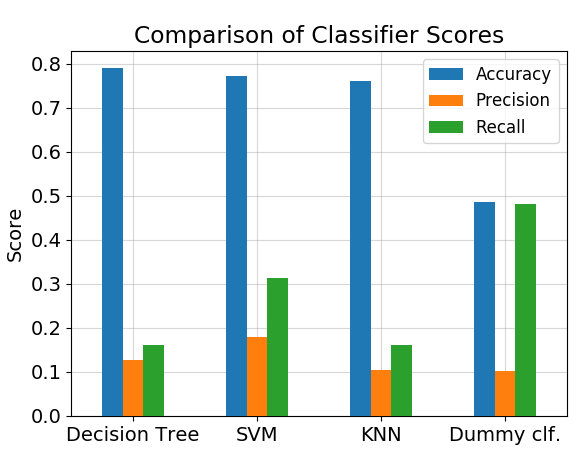


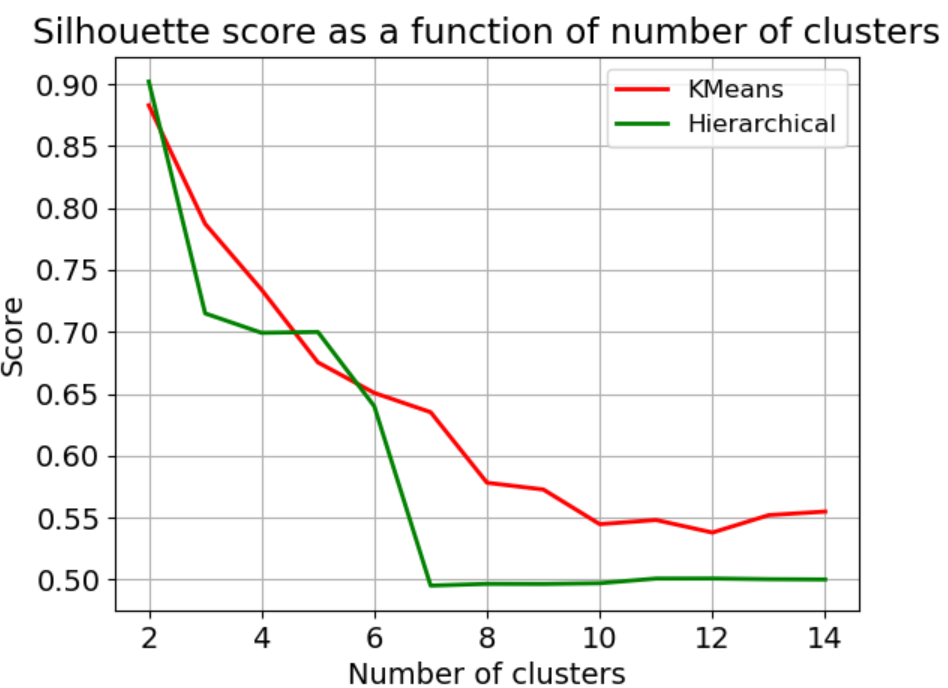












[(0, 2361), (1, 51)]

[(0, 963), (1, 42)]

Cluster: 0 true labels: 0.4078780177890724

Cluster: 1 true labels: 0.8235294117647058

KMeans accuracy: 0.9582089552238806

**After fix to make calculations on test true labels (instead of train):**

**Cluster: 0 true labels: 0.1058873358746294**

**Cluster: 1 true labels: 0.17647058823529413**