**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.

פילטור גנים בעלי variance נמוך:

צריך לקבוע סף ל-variance שמתחתיו הגנים יסוננו החוצה. התוצאות עם מס' ספים שונים (**אחרי נרמול)**: (לקחתי את סט האימון והמבחן שכבר חולקו לפני כן)

1. סף = 0.005:

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 10251 test size is: 4372

TRUE in train size: 1210 TRUE in test size: 519

Decision Tree

accuracy score: 0.7861390667886551

recall score: 0.17341040462427745

precision score: 0.15100671140939598

1. סף = 0.001:

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 11246 test size is: 4818

TRUE in train size: 1210 TRUE in test size: 519

Decision Tree

accuracy score: 0.7841427978414279

recall score: 0.1252408477842004

precision score: 0.09984639016897082

1. סף = 0.01

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 3362 test size is: 1418

TRUE in train size: 1209 TRUE in test size: 518

Decision Tree

accuracy score: 0.7708039492242595

recall score: 0.722007722007722

precision score: 0.6738738738738739

1. סף = 0.05

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 2368 test size is: 1008

TRUE in train size: 1173 TRUE in test size: 505

Decision Tree

accuracy score: 0.8541666666666666

recall score: 0.8574257425742574

precision score: 0.8523622047244095

1. סף = 0.07

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 2175 test size is: 919

TRUE in train size: 1127 TRUE in test size: 487

Decision Tree

accuracy score: 0.8672470076169749

recall score: 0.8788501026694046

precision score: 0.8716904276985743

1. סף = 0.08

before filtering, train size is: 11256 test size is: 4824

after filtering, train size is: 2081 test size is: 869

TRUE in train size: 1089 TRUE in test size: 471

Decision Tree

accuracy score: 0.8411967779056386

recall score: 0.8598726114649682

precision score: 0.8490566037735849

* אנחנו רואים כי ניתן לצמצם משמעותית את הדאטאסט ולהביא לתוצאות הרבה יותר טובות. עם זאת ראינו שהיחס בין סט האימון למבחן קטן ולכן נחזור על הפעולה על הדאטאסט כולו, לאחר הצמצום שלו נחלק לאימון ומבחן ואז נבדוק ביצועים שוב:

after filtering, train size is: 2177 test size is: 933

Decision Tree

accuracy score: 0.8563772775991425

recall score: 0.8677685950413223

precision score: 0.8571428571428571

הבעיה במקרה הזה היא שכדי לקבל תוצאות טובות אני מנרמלת אבל נרמול על כל הדאטאסט הוא לא "נכון" כי צריך לנרמל את סט המבחן לפי סט האימון. **לשאול את ליאון**