**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.

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  accuracy 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