**Mini-Project Number 1.** **What Breast cancer will recur?**

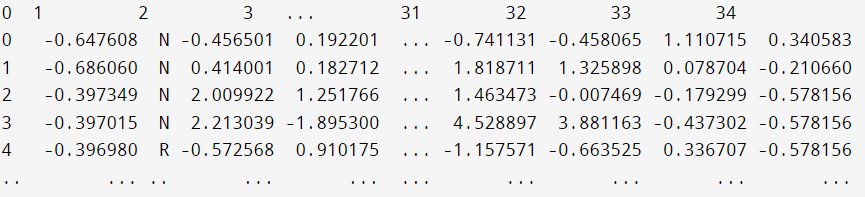
דוח מעבדה נכתב ע"י מיתר פוקס ת.ז 315325605 , רבקה קקון ת.ז 204901417 , שרה חננייב 039197488

בפרויקט מנתחים את הדאטה של אוניברסיטת ויסקונסין על נתוני החלמה סרטן השד .

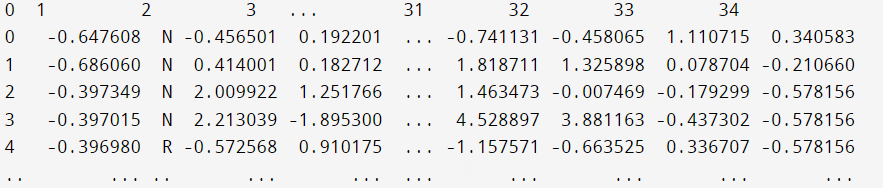
קישור למידע: https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data

הסבר על המידע : <https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.names>

צילום חלקי של המידע הנ"ל (סהכ המידע מכיל 198 שורות על 35 עמודות)



R = recurrent, N = nonrecurrent בעמודה השניה מוחלף כך N= -1, R= 1



בפרויקט זה מיישמים 3 אלגוריתמים שנלמדו :

1. **perceptron algorithm** שמיועד לנתונים שמופרדים לינארית, מה שלא היה נכון לגבי הדאטה שהובא במטלה ולכן יש התמודדות אם עובדה הזו. נעזרנו במידע שיש בקישור זה

https://machinelearningmastery.com/implement-perceptron-algorithm-scratch-python/

1. **Adaline algorithm** נעזרנו במידע שיש בקישור זה <https://dzone.com/articles/adaline-explained-with-python-example-data-analyti>
2. **Backpropagation algorithm** – בחלק זה של מימוש האלגוריתמים השתמשנו באלגוריתם קיים

from sklearn.neural\_network import MLPClassifier

**נעזרנו במידע שיש בקישור הבא**

[**https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/**](https://analyticsindiamag.com/a-beginners-guide-to-scikit-learns-mlpclassifier/)

במידע שסופק ניתן לראות הפרש משמעותי בכמות השורות שמייצגות חיובים ושלילים (ראה דיאגרמה) . אי אפשר למחוק שורות מידע כדי לאזן את הכמות כי נזייף מידע שהוא ככל הנראה לא נכון ולעובדה שיש פער בכמות השלילים והחיובים יש משמעות .

כדי לחלק לקבוצות הוגנות של טסט ואימון חילקנו אותם בנפרד ורק אז איחדנו כלומר חילקנו את 47 החיובים לקבוצות של 66% ו33% ואז את 151 השלילים לקבוצות של 66% ו33% ואז חיברנו בין אלה שהוקצו ל66% ואת אלו שהוקצו ל33%. כדי להרגיש בטוחים בקבוצות שלנו נבצע את הפעולות 3 פעמים והאחוז הצלחה שלנו יהיה ממוצע של 3 אחוזי ההצלחה שנקבל ב3 הסבבים.

חלק ראשון:

התבקשנו לרשום כמה זמן ריצה לקח לכל אלגוריתם :

1. **perceptron algorithm** - 

2.  **Adaline algorithm**- 

1. **Backpropagation algorithm**-



חלק שני:

התבקשנו לרשום את הפרמטרים שבחרנו לכל אלגוריתם, לכל אלגוריתם עשינו ניסוי וטעייה, ניסינו כמה וכמה פרמטרים שונים כדי לקבל את האחוז הצלחה המקסימלי.

1. **perceptron algorithm** - סה"כ 50 הרצות , עם קצב למידה של 0.01.

2.  **Adaline algorithm**- עשינו סה"כ 65 הרצות .

קצב הלמידה הוא 0.05.

**Backpropagation algorithm3.**-

במודל זה יש שכבות פנימיות שיש לתאר , אנחנו בחרנו שכבה פנימית ראשונה בגדול 200 , שכבה שניה בגודל 100 ושכבה שלישית בגודל 65.

מס' הרצות מקסימלי שבחרנו היה 650.

האלפה שבחרנו היא 0.01.

נעזרנו ב

https://www.kaggle.com/ahmethamzaemra/mlpclassifier-example

חלק שלישי:

התבקשנו לרשום מהם אחוזי ההצלחה של כל אלגוריתם והסבר של התוצאות שקיבלנו .

1. **perceptron algorithm** -

אחוזי הצלחה : 74%

סטיית תקן: 

טבלאות נדרשות:

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 46 | 0 |
| Actual: no | 13 | 2 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 36 | 10 |
| Actual: no | 10 | 5 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 45 | 1 |
| Actual: no | 12 | 3 |

2.  **Adaline algorithm**-

אחוזי הצלחה : 78%

סטיית תקן: 

טבלאות נדרשות:

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 34 | 12 |
| Actual: no | 13 | 2 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 40 | 6 |
| Actual: no | 9 | 6 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 36 | 10 |
| Actual: no | 12 | 3 |

**Backpropagation algorithm 3.**-

אחוזי הצלחה : 78%

סטיית תקן: 

טבלאות נדרשות:

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 44 | 2 |
| Actual: no | 10 | 5 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 46 | 0 |
| Actual: no | 13 | 2 |

|  |  |  |
| --- | --- | --- |
|  | Predicted: yes | Predicted: no |
| Actual: yes | 46 | 0 |
| Actual: no | 13 | 2 |

חלק רביעי:

ללא ספק המטלה היתה מאוד מאתגרת , מכמה בחינות: א. לא התנסינו בעבר במשהו דומה שמצריך ניתוח נתונים. או עבודה עם דאטה סט כפי שניתן לנו. לקח לנו לא מעט זמן להבין איך לעשות שינויים בדאטה סט כך שלא יפגמו בתוצאות של הניסוי (למשל '?' ) ב. לא היה לנו ניסיון רב בפייתון. ג. כתיבת האלגוריתמים עצמם היוותה אתגר.

עשינו חיפוש רב מאוד באינטרנט וראינו סרטונים ביוטיוב כדי להבין איך הכי נכון לכתוב את הקודים כך שננתח אותם לכל הנתונים שדרשת. הבנו שיש צורך בפונקציות fit, predict, train, score אבל בכל אלגוריתם הפונקציות האלו עושות חישוב שונה ופעולה שונה ולכן האלגוריתמים צריכים להכיל את פונקציות אבל בהתאמה החישובית הנכונה לכל אלגוריתם.

כדי להבין מה הפרמטרים שנותנים תוצאות מספקות הכנסנו נותנים שונים בכל פעם ועשינו רישום של התוצאות שמתקבלות כדי להבין מה הכי נכון . אנחנו עדין לא יודעת אם הצלחנו באמת להגיע לתוצאות שאתה ציפית להן אבל אנחנו מקוות שכן.

**perceptron algorithm** היה מאתגר כי המידע שניתן לנו לא היה באמת ניתן להפרדה לינארית , לאחר חיפוש באינטרנט הבנו שההפרדה לא תתבצע ב100 אחוז אלא באחוז חלקי , וזה בעצם הפתרון שלנו.