

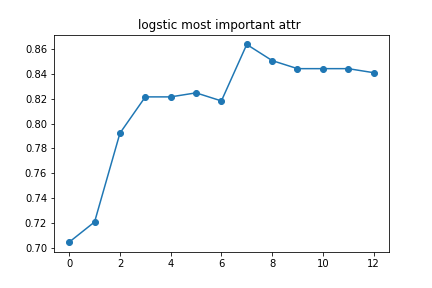
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | | | | |
|  | | **עבודת סיום**  **למידה חישובית** | | | | |  | |
|  |  | | | | | | |  |
|  | | | |  |  | | | |
|  | | | | לואי ג'בר : 315427740מוסא חסונה : 315606843 |  | | | |
|  | | | | מועד הגשה :03.03.2020מרצה קורס :חסין יהודה |  | | | |
|  | | |  | | |  | | |

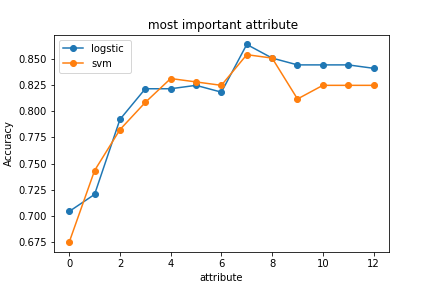
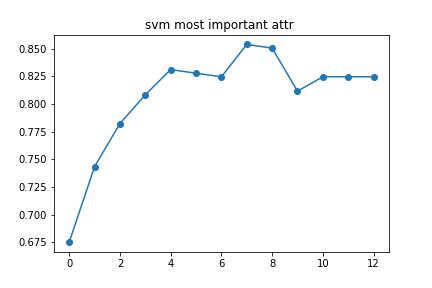
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | |  |  | | |  |
|  | מבוא | | | | | | |  |
|  |  | | |  |  | | |  |
|  |  | |  | | |  | |  |
|  |  |  | מחלות לב גורמות לתחלואה ותמותה נרחבות ומהוות נטל כבד על מערכת הבריאות ועל החולה ומשפחתו. מחלות לב הן הגורם השני למוות בישראל (לאחר סרטן) ובשנת 2013 היוו כ -16% מכלל מקרי המוות בישראל.  זה הופך את מחלות הלב לדאגה העיקרית שיש לטפל בה. אך קשה לזהות מחלות לב בגלל מספר גורמי סיכון תורמים כמו סוכרת, לחץ דם גבוה, כולסטרול גבוה, קצב דופק חריג ועוד גורמים רבים אחרים. בשל אילוצים כאלה, מדענים פנו לגישות מודרניות כמו כריית נתונים ולמידה חישובית לניבוי המחלה.  למידת חישובית (ML) מתגלה כיעילה בסיוע בקבלת החלטות ותחזיות מכמות הנתונים הגדולה שמייצרת ענף הבריאות.  בפרויקט זה, אנו נשתמש בגישות של למידת לצורך סיווג האם אדם סובל ממחלת לב או לא | | |  |  |  |
|  | | | | |
|  |  |  |  |
|  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | |  |  |
|  | | התהליך | | |  | |
|  | |  |  |  |  | |
|  | ניתוח וקבלת ה נתונים מערך הנתונים מורכב מ- 1025 נתוני אנשים. במערך הנתונים יש 14 עמודות, המתוארות להלן  1. גיל  2. מין  3. סוג כאב בחזה (4 ערכים)  4. לחץ דם במנוחה  5. סרום כולסטרול ב-מ"ג  6. סוכר בדם בצום  7. תוצאות אלקטרו קרדיו גרפיות מנוחות (ערכים 0,1,2)  8. קצב לב מרבי שהושג  9. אנגינה הנגרמת על ידי אימונים  10. old peak = דיכאון ST שנגרם על ידי פעילות גופנית יחסית למנוחה  11. slope of the peak exercise ST segment  12. number of major vessels (0-3) colored by fluoroscopy  13. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect | | | | |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ניתוח וקבלת הנתונים אפשר לרואות בתוצאות למעלה שהנתונים שלנו מאוזנים יחסית (המחלקות מאוזנות) , דבר שמאוד רצוי ב שלב הלמידה ו תמיד שואפים אליו מספר האנשים החולים הוא 526 ו הלא חולים 499 לכן בחרנו להסתכל על מדד ה (accuracy).  אז בשלב הראשון נערבב את ה נתונים שיש לנו ואז נחלק את הנתונים ל 30% שהם הולכות ל בדיקה ו 70% ללמידה . שלב הלמידה אנו נריץ כמה אלגורתמים ונבחר לבדוק כמה מדדים חשובים , נשחק עם כמה פרמטרים ונראה את השבעתם על התוצאות ובסוף נראה את ההבדל בין כל האלגורתמים שהרצנו ובאיז אלגרותם התקבלה התואצאה הכי טובה יחסית ל נתונים שיש לנו.  מדד ההערכה המשמש הוא (confusion matrix), מציגה את הערכים החזוי כהלכה כמו גם הערכים שנבנו באופן שגוי על ידי המסווג.  סכום ה- TP ו- TN, מה confusion matrix , הוא מספר הערכים המסווגים כהלכה על ידי המסווג.   |  | | --- | | Logistic Regression | | Accuracy | F1-score | Recall | Precision | | 0.86039 | 0.859303 | 0.914286 | 0.851064 |  Logistic Regression  |  | | --- | | SVM | | Accuracy | F1-score | Recall | Precision | | 0.853896 | 0.85298 | 0.91875 | 0.821229 |   **LOGSTIC REGRESSION** תמונה שמכילה מפה, טקסט  התיאור נוצר באופן אוטומטיתמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטי ***SVM***  *תמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטי*  ***בדיקת פרמטרים שונים כדי להבין כיצד זה משפיע על התוצאות***   |  | | --- | |  | | Accuracy | F1-score | Recall | Precision | | 0.86039 | 0.859303 | 0.914286 | 0.851064 |  1. ***אנו משתמשים בפרמטר C כפרמטר הרגולציה שלנו. פרמטר C = 1 / λ****.*   C הוא פרמטר המאלץ אתנו לחפש מקדמים טיטיות קטנות מ C הוא מחליט מורכבות ה מודל  *עבור ערכים קטנים של C, אנו מגדילים את חוזק ההסדר אשר ייצר מודלים פשוטים אשר יגרום ל (* *underfit) לנתונים. עבור ערכים גדולים של C, אנו מנמיכים את כוחה של הסדירה שמרמז על כך שהמודל מותר להגדיל את המורכבות שלו, ולכן זה גורם ל (overfit) על הנתונים. עשינו השוואה בין train \test accuracy בשביל לראות איך השינוי ב פרמטר משפיע על ה overfit underfit אפשר לראות ש קבלנו דיוק מקסימלי לא בהכרח שמגדלים את הפרמטר אלה באמצע נתן לראות אם אנו לוקחים פרמטרים גדולים לפי הפרש הדיוק נהיה לנו underfit אך הדיוק נהיה פחות*  **LOGSTIC REGRESSION**  תמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, מפה, צילום מסך  התיאור נוצר באופן אוטומטי  תמונה שמכילה מפה, טקסט  התיאור נוצר באופן אוטומטי  ***SVM***  ***תמונה שמכילה טקסט, מפה  התיאור נוצר באופן אוטומטיתמונה שמכילה מפה, טקסט, צילום מסך  התיאור נוצר באופן אוטומטי***  ***תמונה שמכילה מפה, טקסט  התיאור נוצר באופן אוטומטי***  בגרף שלמעלה אפשר לראות ש ב SVM ה גודל של ה פרמטר גרם ל הפרש יותר ומכוון שדיוק ה train הוא גדול יש לנו over   1. ***אנו משתמשים בפרמטר TOL***   *אומרת למודל להפסיק לחפש MAX MIN כשתנאי עצירה מתקיים*  *במקרה שלנו כשאנחנו מחפשים* coefficients  *מקדמי של בודקים את הרכיב המקסימלי ב גראדינט אם הוא נהיה קטן מה ה TOL עוצרים*    **LOGSTIC REGRESSION**  תמונה שמכילה טקסט, צילום מסך, מפה  התיאור נוצר באופן אוטומטיתמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטי**תמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטי**  ***SVM***  תמונה שמכילה טקסט, מפה, צילום מסך  התיאור נוצר באופן אוטומטיתמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטי  תמונה שמכילה צילום מסך, מפה  התיאור נוצר באופן אוטומטי   1. Class weight   פרמטר זה נותן משקל לכל מחלקה במקרה שלנו יש 2 מחלקות מחלקה 0 לא חולה 1 חולה  ואז הוא מכפיל את השגיאה של המחלקה שלנו עם ה פרמטר שנתנו נותנים את זה כ DICT  Class\_weight={0:max\_w,1:1} מכוון שיש לנו 2 מחלקות רק עשינו בדיקה על אחד המחלקות עם שנוי ה משקל הטווח שבדקנו בין 0.5 ל 4 בקפיצות של 0.05  **LOGSTIC REGRESSION**  תמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטיתמונה שמכילה מפה, טקסט  התיאור נוצר באופן אוטומטי  תמונה שמכילה מפה  התיאור נוצר באופן אוטומטי  ***SVM***  תמונה שמכילה צילום מסך  התיאור נוצר באופן אוטומטיתמונה שמכילה מפה, טקסט  התיאור נוצר באופן אוטומטי  תמונה שמכילה מפה  התיאור נוצר באופן אוטומטי   |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | *itteration* | *Fasting*  *Blood Sugar* | *Cholesterol* | *Resting Blood Pressure* | *Resting*  *ECG* | *Exercise Induced Angina* | *Age* | *Max*  *Heart*  *Rate*  *Achieved* | *Thalassemia* | *Gender* | *Old peak* | *Slope* | *Major Vessels* | *Chest Pain* | | *1* |  |  |  |  |  |  |  |  |  | סימן ביקורת |  |  |  | | *2* |  |  |  |  |  |  | סימן ביקורת |  |  | סימן ביקורת |  |  |  | | *3* |  |  |  |  |  |  | סימן ביקורת | סימן ביקורת |  | סימן ביקורת |  |  |  | | *4* |  |  |  |  |  |  | סימן ביקורת | סימן ביקורת |  | סימן ביקורת |  |  | סימן ביקורת | | *5* |  |  |  |  |  |  | סימן ביקורת | סימן ביקורת |  | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | | *6* |  |  | סימן ביקורת |  |  |  | סימן ביקורת | סימן ביקורת |  | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | | *7* |  | סימן ביקורת | סימן ביקורת |  |  |  | סימן ביקורת | סימן ביקורת |  | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | | *8* |  | סימן ביקורת | סימן ביקורת |  |  |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | | *9* |  | סימן ביקורת | סימן ביקורת |  |  |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | | *10* |  | סימן ביקורת | סימן ביקורת | סימן ביקורת |  |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | | *11* |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | | *12* |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | | *13* |  | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | | *14* | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת | סימן ביקורת |   ***בדיקת* פיצ'רים (תכונות) חשובים/לא חשובים**   |  | | --- | |  | | Accuracy | F1-score | Recall | Precision | | 0.86039 | 0.859303 | 0.914286 | 0.851064 |   *נריץ אלגוריתם שיודע לבחור את הפיצ'רים הכי טובים, בחרנו ב RFE שהוא אלגוריתם רקורסיבי נותנים לו כמה תכונות לחפש ונותנים לו מודל במקרה שלנו 'SVM OR LOGSTIC\_REGRESSION ' אחרי שנותנים לו את מה שבקש מחשבים את K התכונות החשובות דרך ה COEEF שלנו כל פעם מורדים את ה תכונה של COEFF הנמוך עד שמגעים ל K התכונות שדרשנו*  *הטבלה למטה מראה לנו איזה תכונות בחרנו בכל איטרציה לפי כמות תכונות חשובות בשני המודלים שחשבנו היה הבדל בהרצה הראשונה מה התוכנה הכי חשובה 1 LOGSTIC נתן OLDPEAK*  *SVM נתן MAX HEART RATE*  *אך בהמשך ההרצות שניהם מתחלים להוסיף וקבלנו מה שנתן למטה* |  |

ציור של רמת דיוק ה מודל בכל פעם שהוספנו תכונה למודל שלנו





***נעבור על שלושת הפרמטרים ונבחר את התוצאות שנותנים תוצאות מקסימליות***

|  |  |  |
| --- | --- | --- |
| SVM | LOGISTIC\_REGRESSION | PARAMETER |
| 10 | 0.6 | C |
| 0.1 | 1.0 | TOL |
| 1.4 | 1.5 | CLASS\_WEIGHT |
| 0.87337662 | 0.840909 | ACCURACY |

***נעבור על שלושת הפרמטרים ונבחר תוצאה מקסימלית ביותר עבור האלגוריתמים***

|  |  |  |
| --- | --- | --- |
| SVM | LOGISTIC\_REGRESSION | PARAMETER |
| 20 | 20 | C |
| 2 | 0.1 | TOL |
| 2.85 | 1.8 | CLASS\_WEIGHT |
| 0.91233766 | 0.86638961 | ACCURACY |

<https://www.kaggle.com/johnsmith88/heart-disease-dataset>

<https://www.kaggle.com/joparga3/2-tuning-parameters-for-logistic-regression>

<https://towardsdatascience.com/heart-disease-prediction-73468d630cfc>

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

ספרות וקישורים