זיהוי סרטן השד באמצעות למידת מכונה

לינוי כהן
21/08/22
ת.ז: 207797531

שאלת המחקר

האם ניתן, באמצעות מודלי למידת מכונה, לחזות

האם בגוף מסוים יתפתח סרטן השד?

מהם הפקטורים התומכים ביותר בשאלה של

סרטן השד.

קצת עובדות על סרטן השד

- סרטן השד הוא סוג הסרטן הכי נפוץ כיום בעולם.
- בארה"ב יש כ- 264 אלף מקרים של סרטן השד כל שנה. מתוכם רק כ- 2400 מקרים מאובחנים אצל גברים.
- כ- 42 אלף נשים וכ- 500 גברים מתים כל שנה כתוצאה מסרטן השד.
- גילוי מוקדם של סרטן השד יכול להבטיח החלמה בכמעט 100% לעומת רק 15% אחוזי החלמה בגילוי מאוחר.
- על פי המחקר העולמי לחקר הסרטן, במהלך שני העשורים האחרונים, מספר האנשים שחלו בסרטן השד הכפיל את עצמו והגיע מ- 10 מיליון מאובחנים בשנת 2000 ל- 19.3 מיליון בשנת 2020. חוקרים משערים שכמות המקרים יכפיל את עצמו עד שנת 2040.
- נכון להיום בממוצע אחת מכל חמישה נשים עלולה לסבול מסרטן השד במהלך חייה.

תהליך העבודה

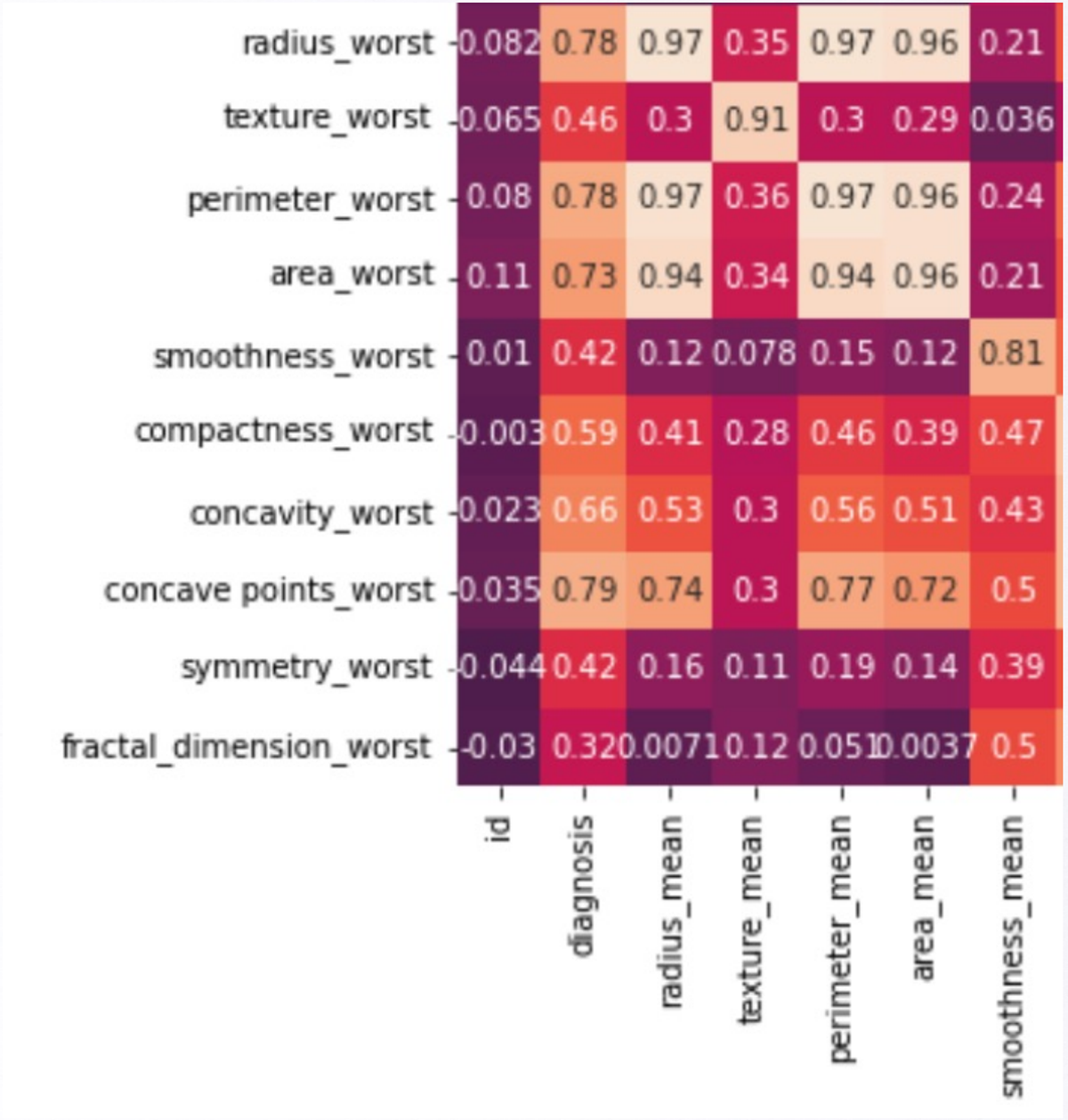
1. איסוף דאטה
2. ניקוי וניתוח הדאטה
3. ויזוליזציות וניתוח הנתונים
4. הפעלת מודלי למידת מכונה
5. ניתוח התוצאות
6. בחירת המודל הטוב ביותר ומימושו

הדאטה

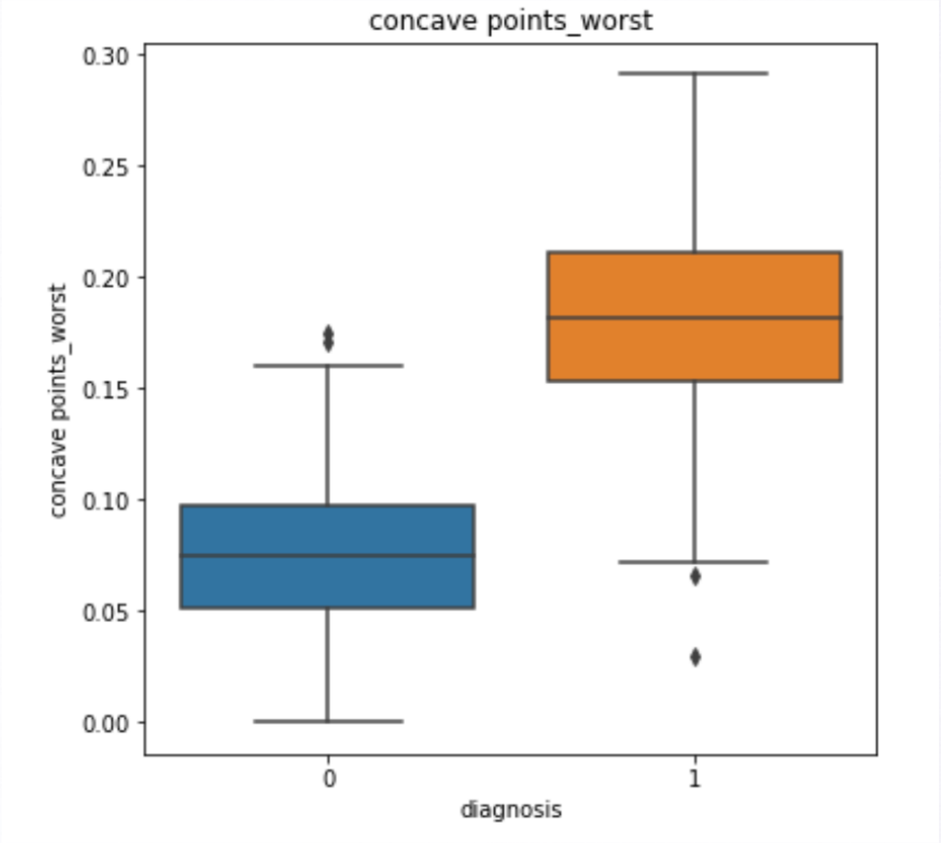
- Wisconsin Diagnostic dataset from University of Wisconsin
- הדאטה לקוח מKaggle אך קיים גם ב - UCI Machine Learning Repository
- 32 מאפיינים שונים של שד
- תמונות של סרטן השד
- 569 תצפיות, מתוכם 357 שפירים ו- 212 ממאירים (62.74% שפיר ו- 37.26% ממאיר)

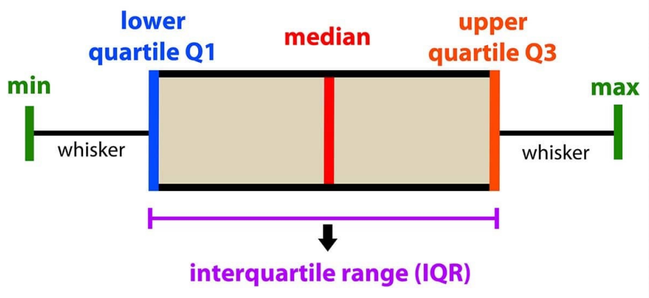
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980

HEATMAP


radius_worst -0.082 0.78 0.97 0.35 0.97 0.96 0.21
texture_worst -0.065 0.46 0.3 0.91 0.3 0.29 0.036
perimeter_worst -0.08 0.78 0.97 0.36 0.97 0.96 0.24
area_worst -0.11 0.73 0.94 0.34 0.94 0.96 0.21
smoothness_worst -0.01 0.42 0.12 0.078 0.15 0.12 0.81
compactness_worst -0.003 0.59 0.41 0.28 0.46 0.39 0.47
concavity_worst -0.023 0.66 0.53 0.3 0.56 0.51 0.43
concave points_worst -0.035 0.79 0.74 0.3 0.77 0.72 0.5
symmetry_worst -0.044 0.42 0.16 0.11 0.19 0.14 0.39
fractal_dimension_worst -0.03 0.32 0.0071 0.12 0.051 0.0037 0.5
id
diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean

Boxplot


concave points_worst
0.30
0.25
0.20
0.15
0.10
0.05
0.00
concave points_worst
0
1
diagnosis


lower
quartile Q1
median
upper
quartile Q3
min
max
whisker
whisker
interquartile range (IQR)

Fit transform

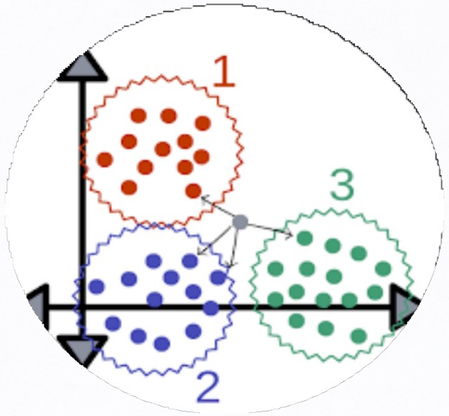
Min Max Scaler

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization

$$x_{\text{scaled}} = \frac{x - mean}{sd}$$

Machine Learning Model's



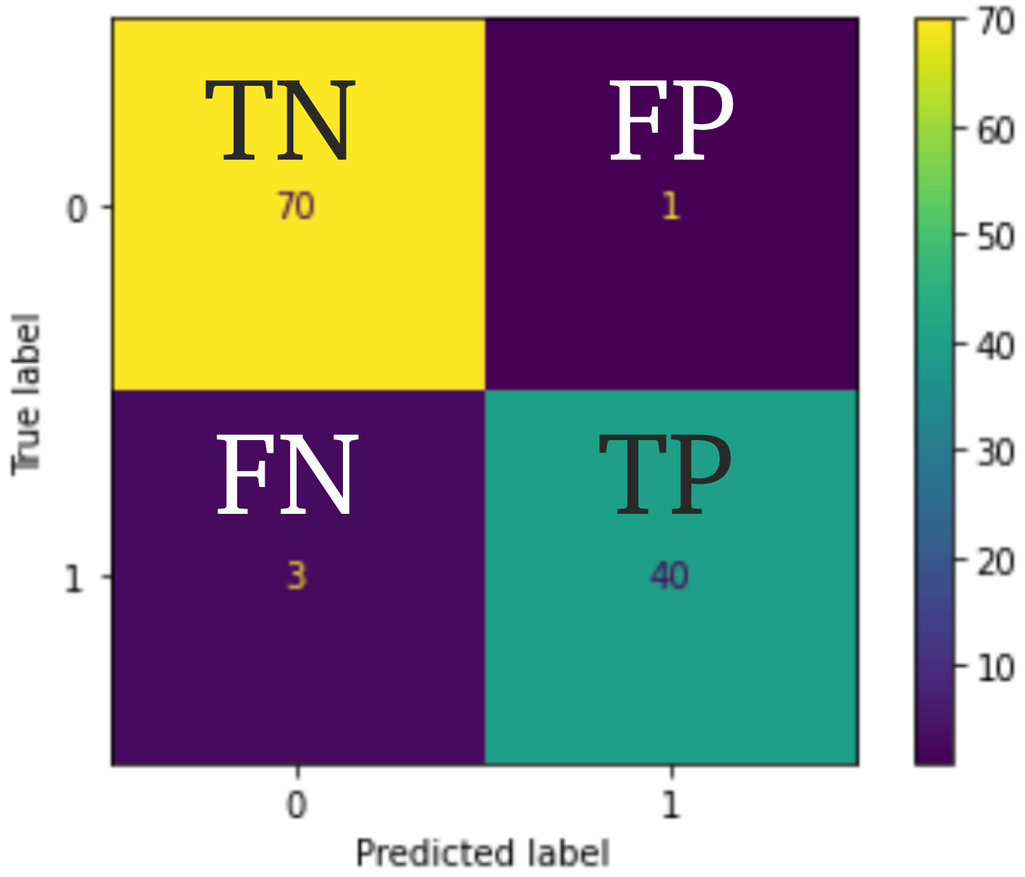
k-nearest neighbors (k-NN)

Random Forest Classification

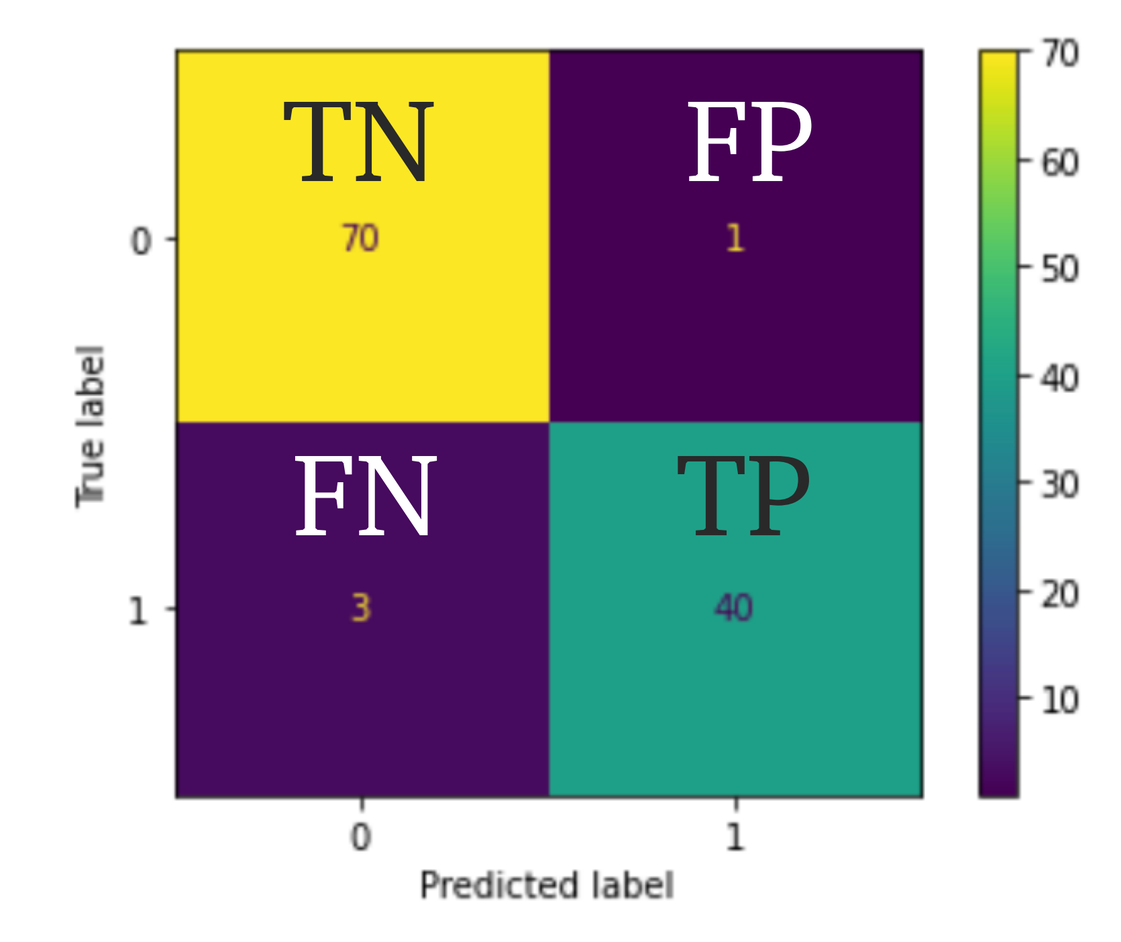
Logistic Regression

k-nearest neighbors (k-NN)

	K	Test Score	Train Score
0	1	0.813298	1.000000
1	2	0.850639	0.959992
2	3	0.875532	0.946656
3	4	0.883311	0.929398
4	5	0.871549	0.926950
5	6	0.874495	0.918938
6	7	0.865880	0.914701
7	8	0.858223	0.910129
8	9	0.863085	0.903655
9	10	0.848772	0.899556
10	11	0.849713	0.898239
11	12	0.854528	0.890632
12	13	0.858151	0.885643
13	14	0.867975	0.886796
14	15	0.873375	0.886869
15	16	0.874997	0.883065



Random Forest



Grid Search

Hyperparameter models

```python
grid_cv = GridSearchCV(estimator=rf, param_grid={'n_estimators': [10, 50, 100], 'max_depth': [3, 5, 7]}, scoring='accuracy', cv=2, n_jobs=3)
# n_jobs is the number of jobs to run in parallel, -1 means all available CPUs
# cv is the number of folds that we want to use for the cross validation

# Since it takes a long time to run, we will be using few options to speed up the process (n_jobs, cv)
grid_cv.fit(X_train, y_train)
```

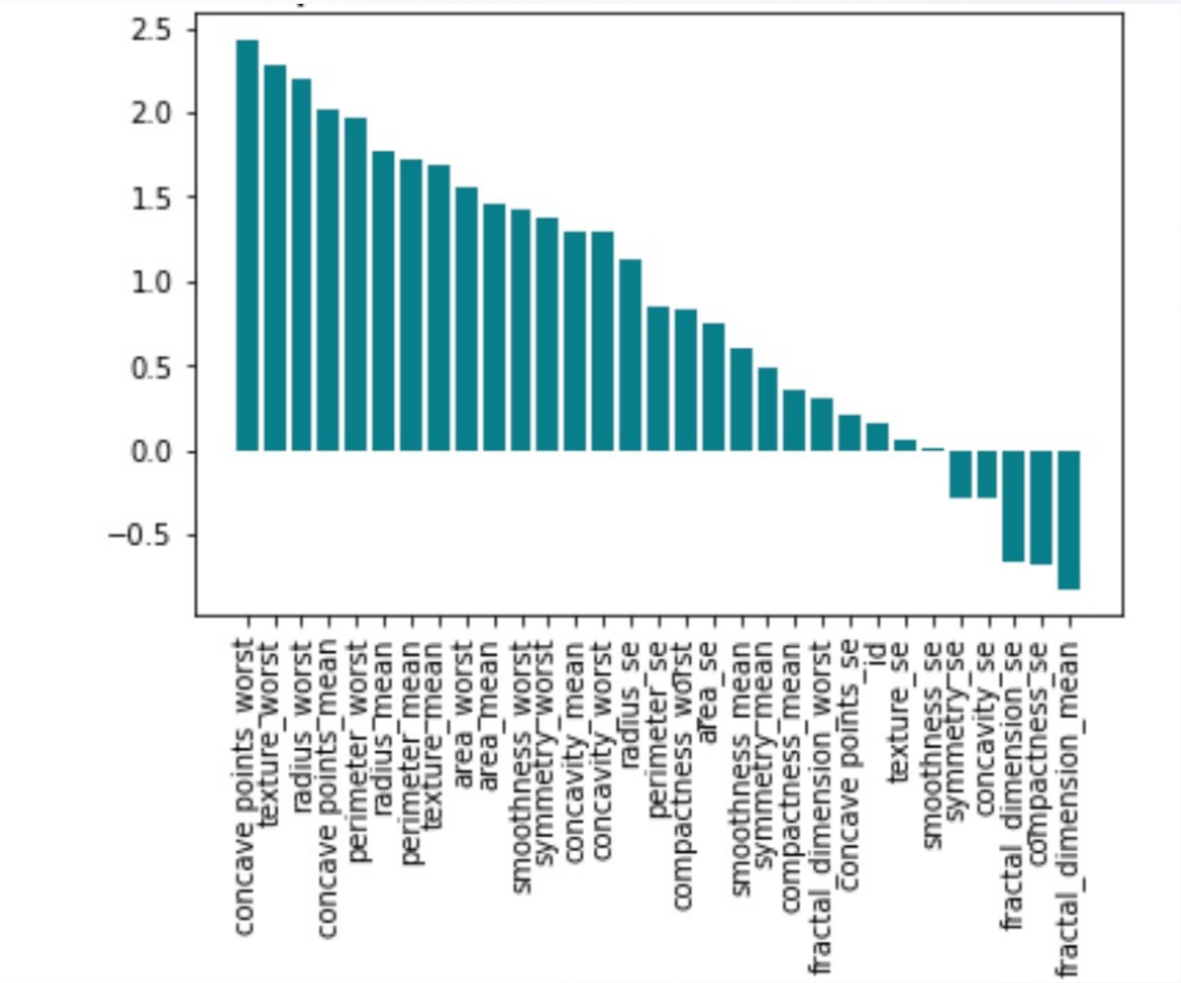
```
GridSearchCV(cv=2,
             estimator=RandomForestClassifier(max_depth=5, n_estimators=50),
             n_jobs=3,
             param_grid={'max_depth': [3, 5, 7], 'n_estimators': [10, 50, 100]},
             scoring='accuracy')
```

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Feature Importance's

using permutation

importance



Model's Scores

איך ממשיכים מפה?

1. הרחבה ודיוק המודלים הקיימים
2. הוספת נתונים שונים (כמו תמונות נוספות)
3. מימוש מודלים נוספים