Breast Cancer Classification

Group 10 Project
Jancey Liu
Weibing Wang
Gege Zhang
Junyi Liu
Zephyr Xiao

Catalog:

- Question
- Methods & Model
- Visualization of Features
- Proportion
- ROC
- Coefficients of Features

Question:

Classifying whether a suspicious lump is benign or malignant based on 5 variables?

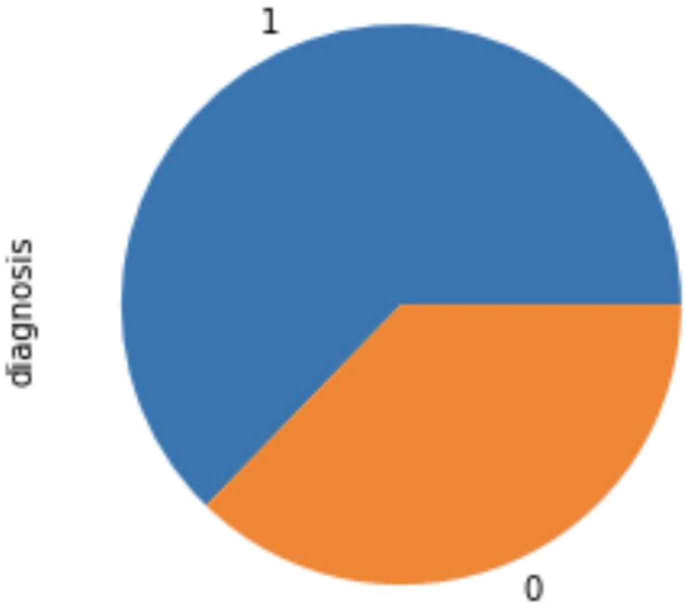
Mean Radius(mm): Mean of distances from center to points on the perimeter
Mean Texture: Standard deviation of gray-scale values
Mean Perimeter(mm): Mean size of the core tumor
Mean Area(mm^2): Mean slice area of the tumor
Mean Smoothness: Mean of local variation in radius lengths

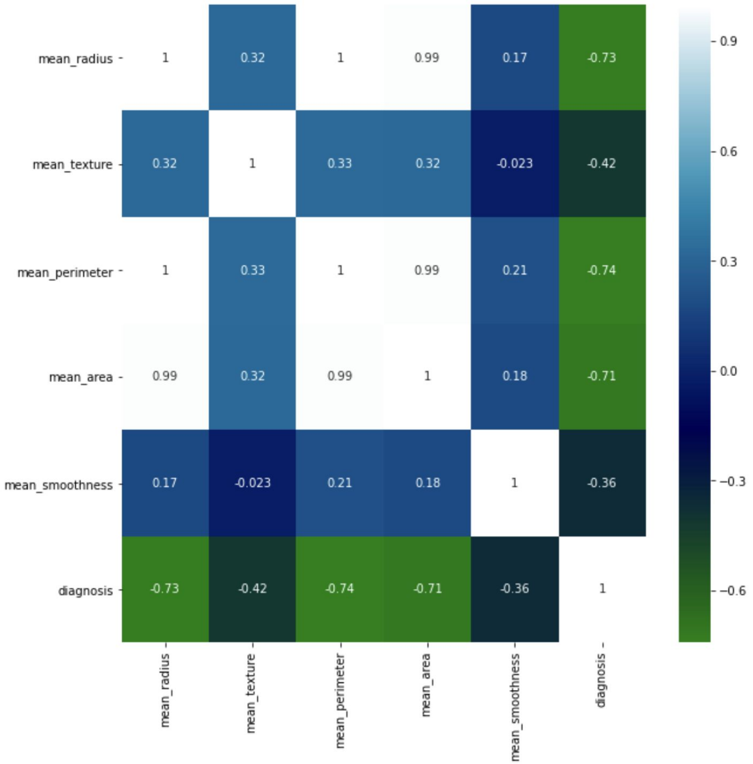
Categorical data:
Diagnosis : The diagnosis of breast tissues (0 = malignant, 1 = benign)

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

Question:

Classifying whether a suspicious lump is benign or malignant based on 6 variables?



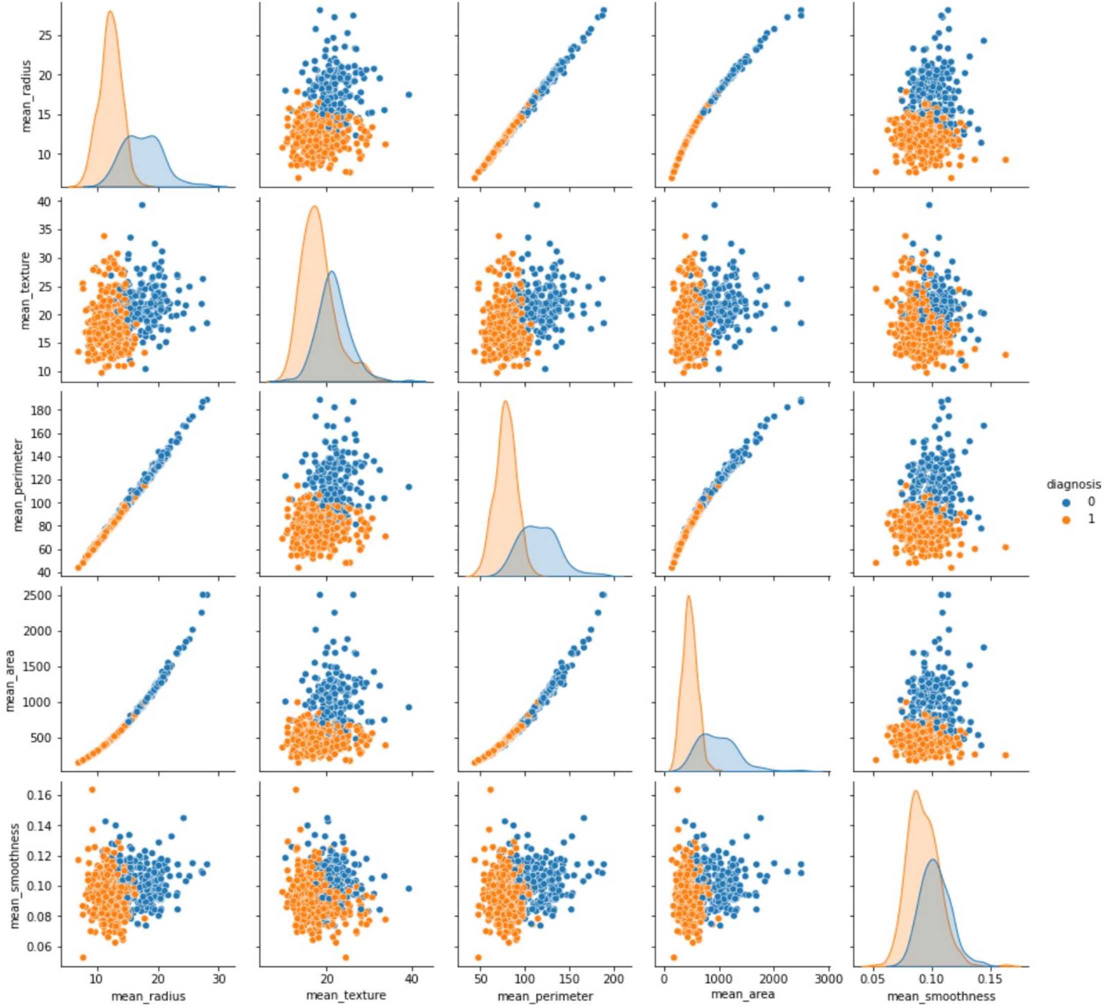


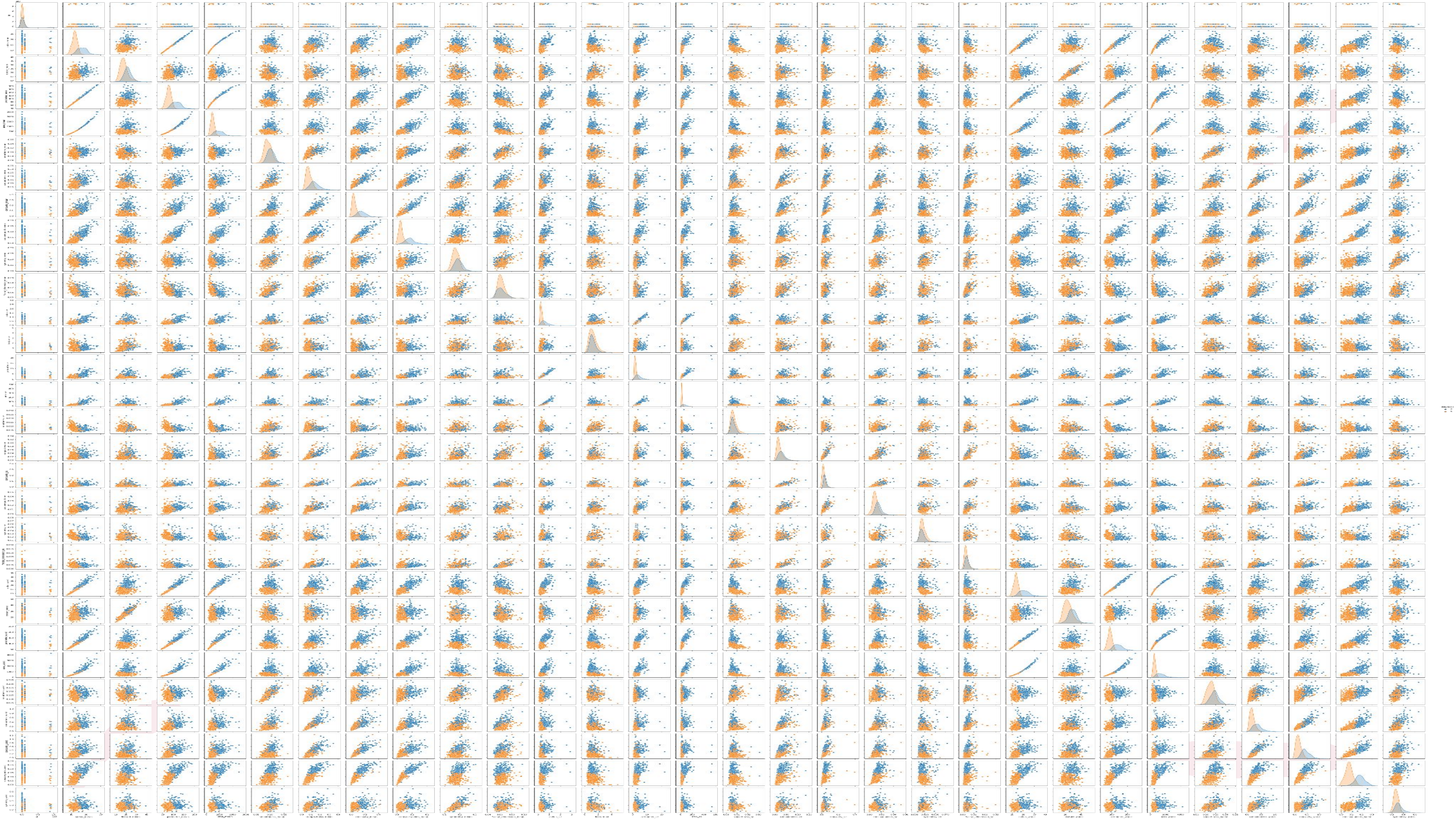
	mean_radius	mean_texture	mean_perimeter	mean_area	mean_smoothness	diagnosis
mean_radius	1	0.32	1	0.99	0.17	-0.73
mean_texture	0.32	1	0.33	0.32	-0.023	-0.42
mean_perimeter	1	0.33	1	0.99	0.21	-0.74
mean_area	0.99	0.32	0.99	1	0.18	-0.71
mean_smoothness	0.17	-0.023	0.21	0.18	1	-0.36
diagnosis	-0.73	-0.42	-0.74	-0.71	-0.36	1

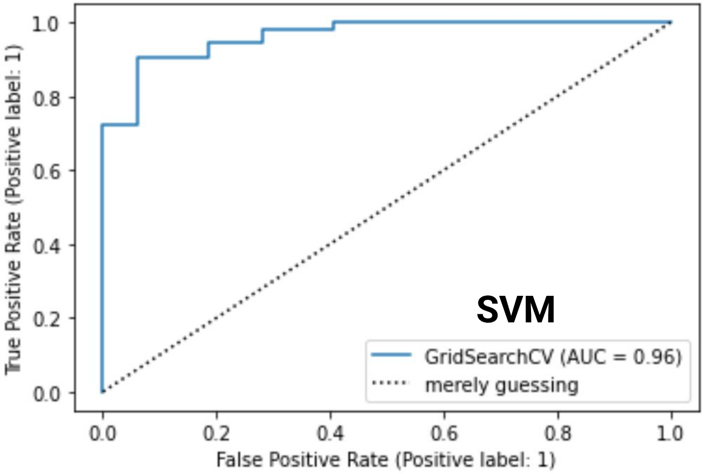
Methods & Model

LogisticRegression	SVM	KNN	RandomForest
91.8%	91.8%	88.2%	94.1%

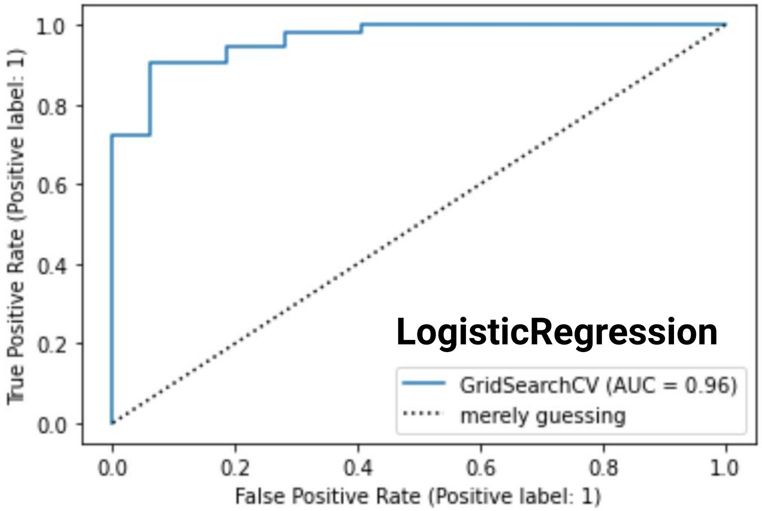
Visualization of Features



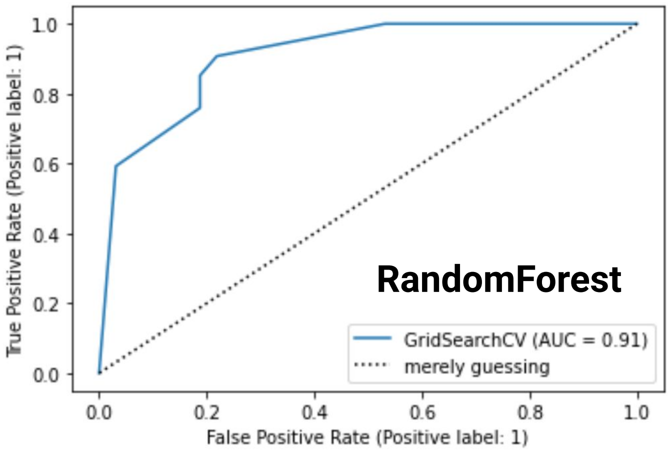




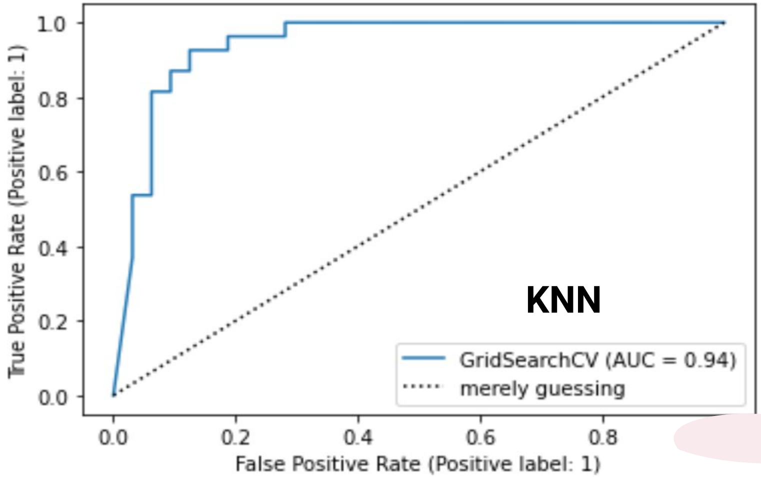
SVM



LogisticRegression



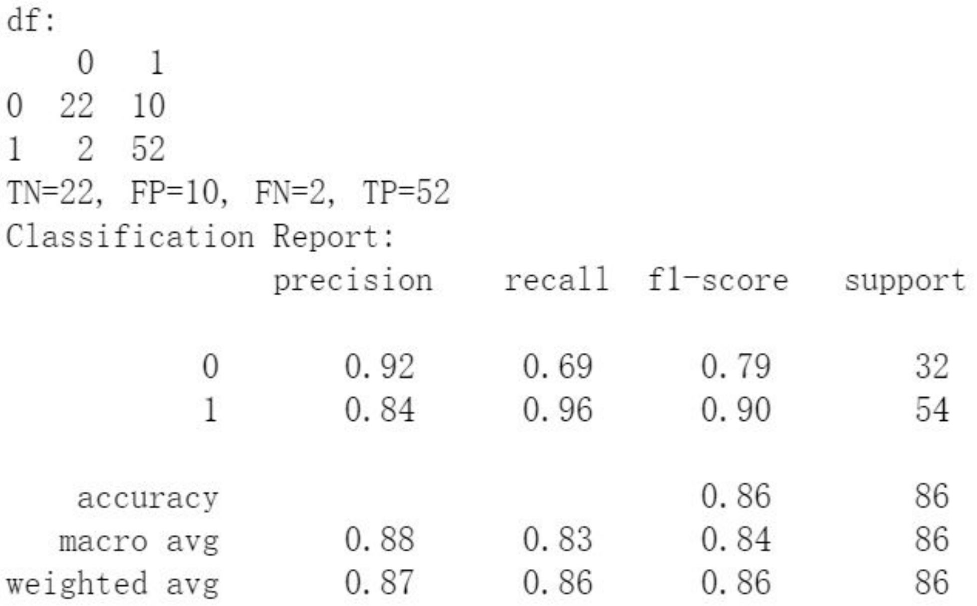
RandomForest



KNN

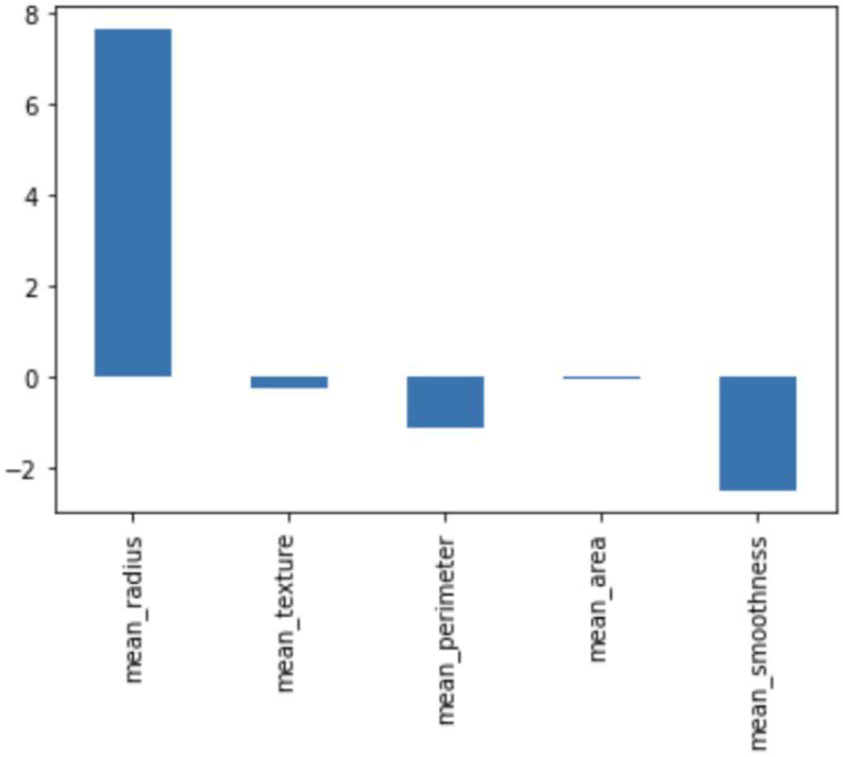
Proportion

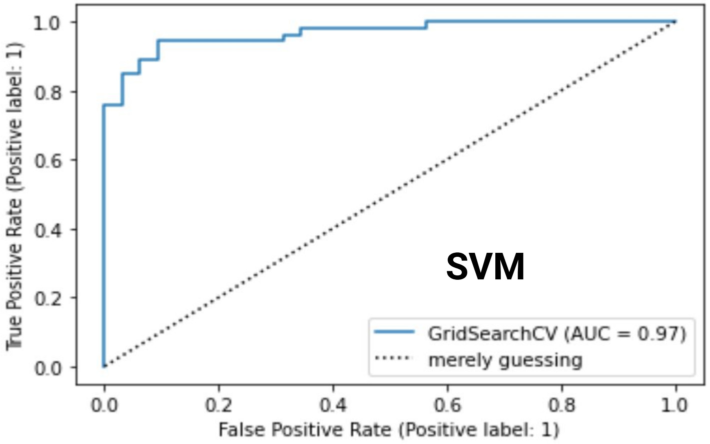
Confusion Matrix

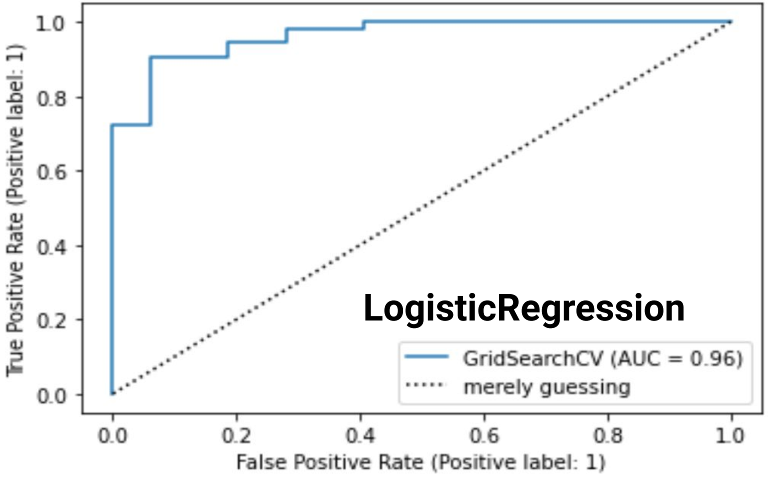


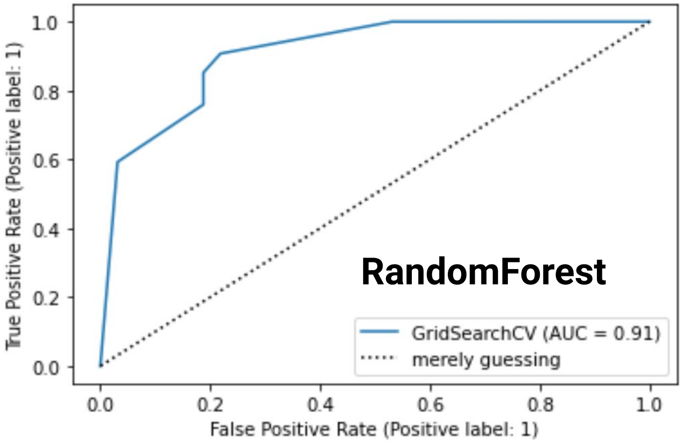
```
df:
    0   1
0  22  10
1   2  52
TN=22, FP=10, FN=2, TP=52
Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.69      0.79        32
           1       0.84      0.96      0.90        54

    accuracy                           0.86        86
   macro avg       0.88      0.83      0.84        86
weighted avg       0.87      0.86      0.86        86
```

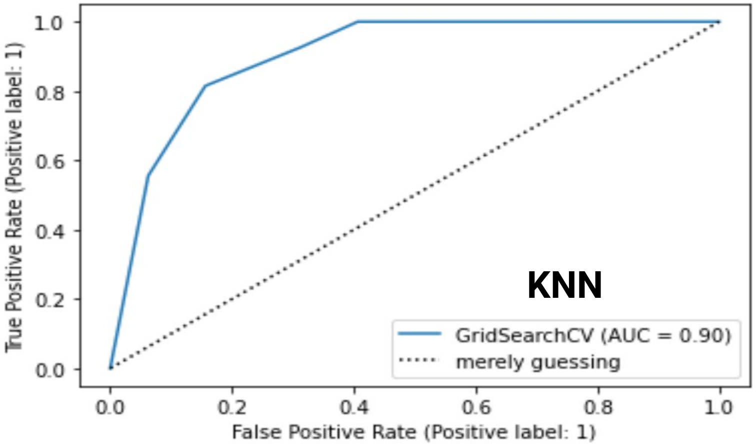
Coefficients of Features (Importance of features) Before




SVM
GridSearchCV (AUC = 0.97)
merely guessing
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

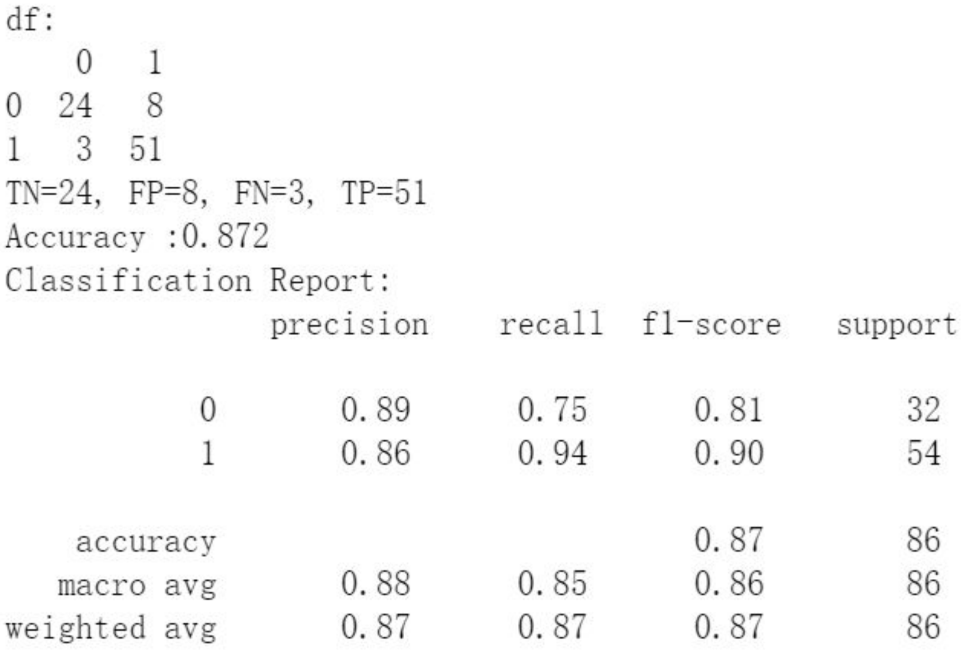

LogisticRegression
GridSearchCV (AUC = 0.96)
merely guessing
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0


RandomForest
GridSearchCV (AUC = 0.91)
merely guessing
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0


KNN
GridSearchCV (AUC = 0.90)
merely guessing
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0

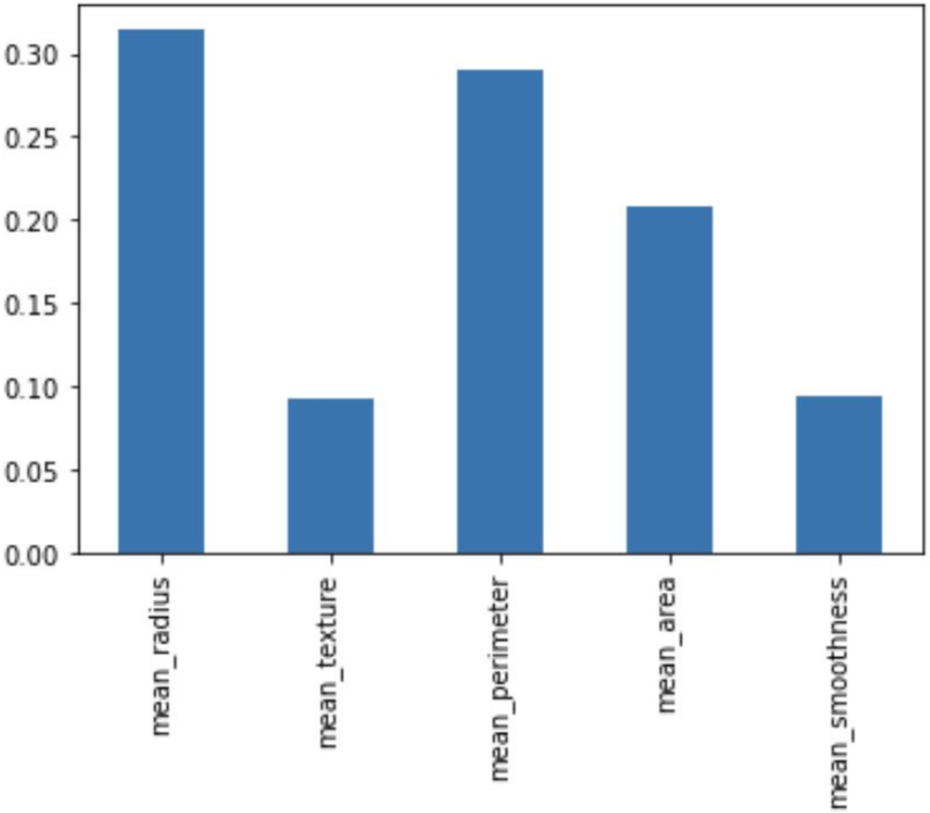
Proportion

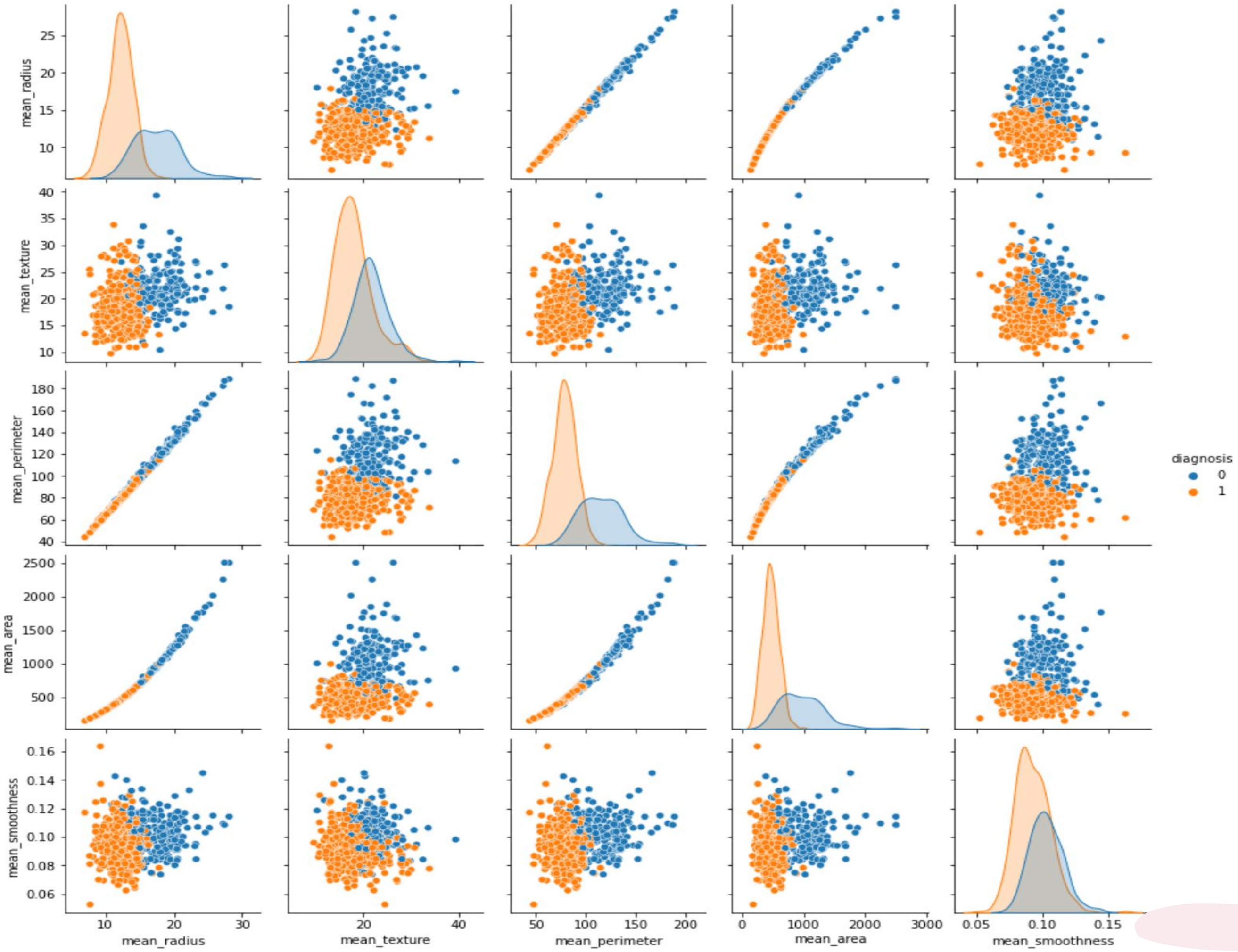
Confusion Matrix

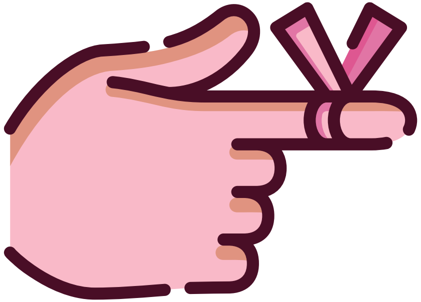


```
df:
    0   1
0  24   8
1   3  51
TN=24, FP=8, FN=3, TP=51
Accuracy :0.872
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.75      0.81        32
           1       0.86      0.94      0.90        54

    accuracy                           0.87        86
   macro avg       0.88      0.85      0.86        86
weighted avg       0.87      0.87      0.87        86
```

Coefficients of Features (Importance of features) After




mean_radius
mean_texture
mean_perimeter
mean_area
mean_smoothness
diagnosis
0
1
25
20
15
10
40
35
30
180
160
140
120
100
80
60
2500
2000
1500
1000
500
0.16
0.14
0.12
0.10
0.08
0.06
10 20 30
20 40
50 100 150 200
0 1000 2000 3000
0.05 0.10 0.15



Methods & Model

LogisticRegression	SVM	KNN	RandomForest
98.4%	98.9%	95.9%	99.1%

Thank you for your listening ！