Non-invasive **Breast Tumor** Diagnosis in Machine Learning

Olivia Fan & Alicia Gong

1.

Introduction

Background

Breast cancer is the most common cancer worldwide. Each year in the US, about 264,000 cases of breast cancer are diagnosed in women and about 2,400 in men.

In clinical ontology, to diagnose breast cancer, doctors usually perform morphimetric analysis of mammographic images of breast masses.

However, diagnosing cancer is a challenging task even for the most experienced ontologist…

Why important? To increase survival rate and relieve patients' suffering!

Aim and Data

We aimed to build machine learning models to predict the breast cancer diagnosis, understanding the importance of different predictors, to assist doctors to make decisions.

We decided to build: 1. LASSO Penalized Logistic model; 2. SVM model; 3. Random Forest model.

We used the dataset from Breast Cancer Wisconsin (Diagnostic) Data Set. The dataset contains diagnosis results and features of the cell nuclei computed from a digitized image of a fine needle aspirate (FNA) of a breast mass for 568 patients.

We partitioned 70% of the original data into training set, and 30% into test set.

Models

01

LASSO Penalized Logistic

02

SVM

03

Random Forest

2. Exploratory Data Analysis

- ❑ Correlation between predictors
- ❑ Separating hyperplane
- ❑ Distribution by diagnosis type

EDA (Cont.)

Correlation

- Based on the correlation matrix, we have several major observations:
- Radius_mean, perimeter_mean, and area_mean are highly correlated.
- Compactness_mean, concavity_mean and concave_points_mean are highly correlated.

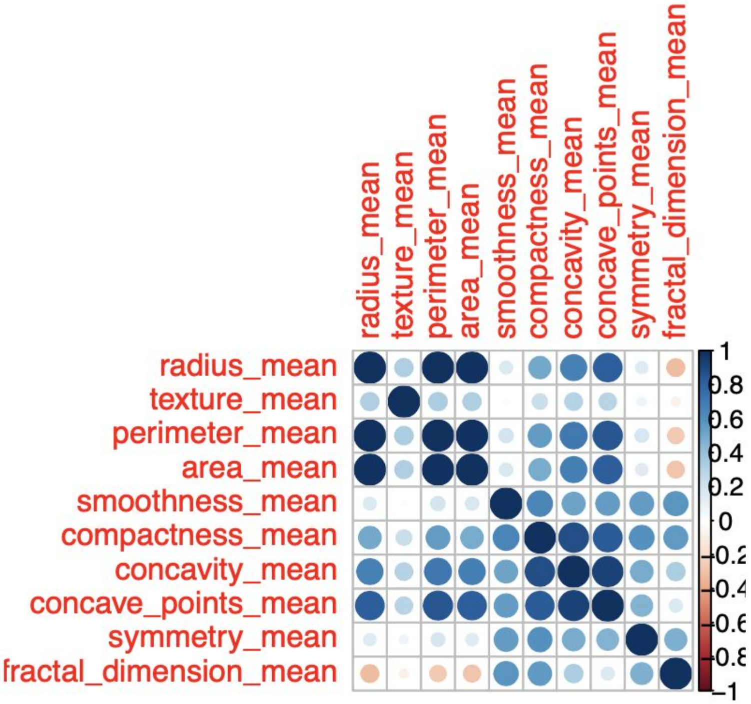


Figure 2: Correlation between mean predictors in the dataset

EDA (Cont.)

Scatterplot

- Separating Hyperplane
  - Overall linear relationship
  - M associates with higher values in predictors, B with lower values
  - ✔ Clear separation
    - *radius_mean* and *perimeter_mean*
  - ✖ Mixed
    - *texture_mean* and *symmetry_mean*
    - *smoothness_mean* and *fractal_dimension_mean*

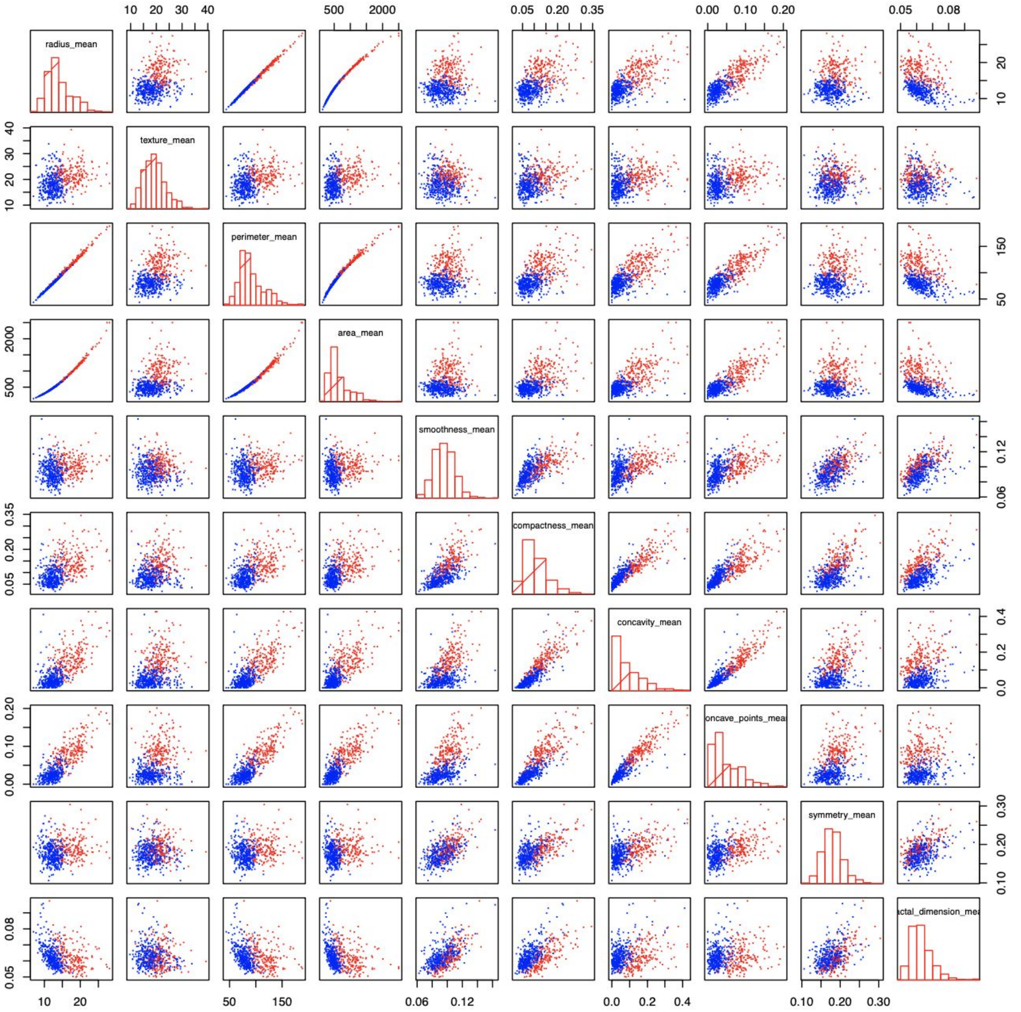


Figure 3: Scatterplot of mean predictors by malignant (red) and benign (blue) tumors

EDA (Cont.)

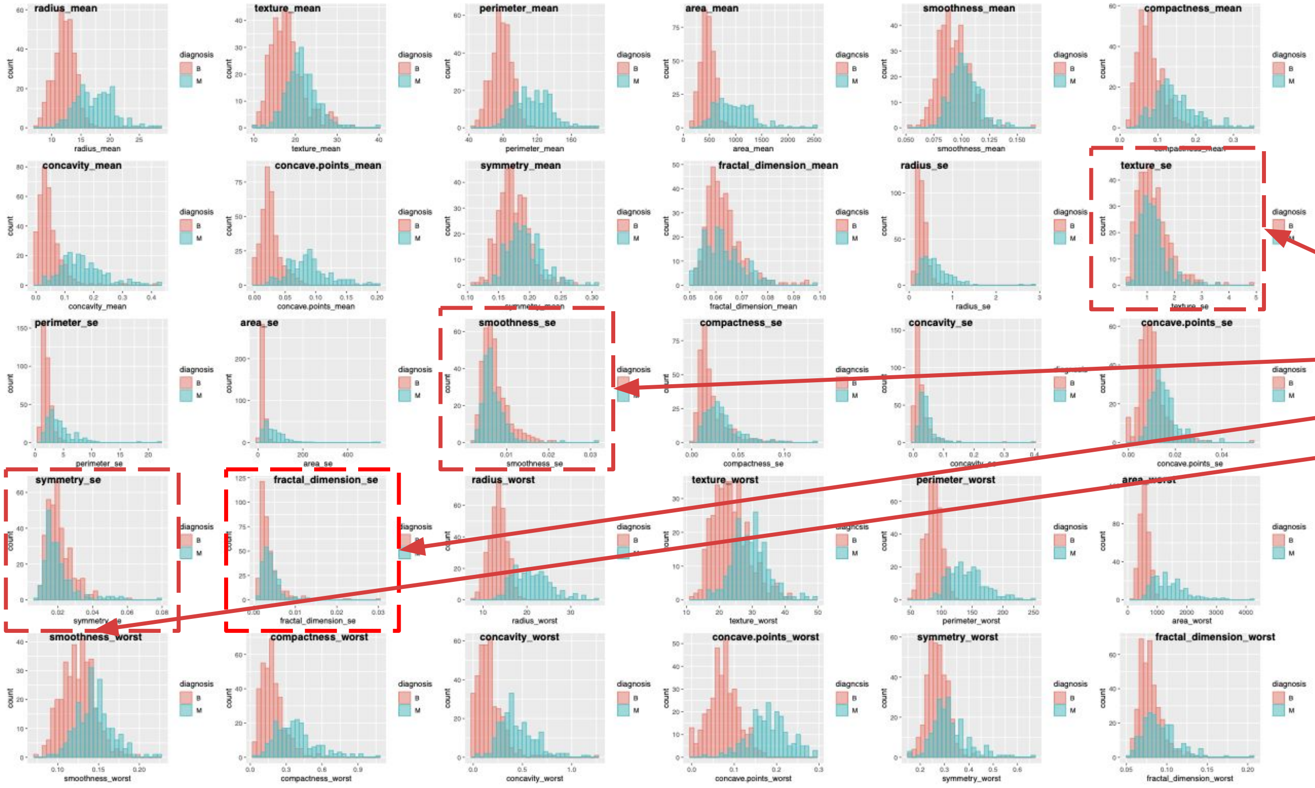


Figure 5: Distribution of predictors by diagnosis

Distribution

By Diagnosis

- ❑ Wish to find features with little overlap between M and B
- ❑ ✖ Overlapping (Eliminate)
  - ❑ *texture_se*
  - ❑ *smoothness_se*
  - ❑ *fractal_dimension_se*
  - ❑ *symmetry_se*

Eliminate predictors associated with texture, smoothness, symmetry and fractal dimension

3. Methodology

LASSO-Penalized Logistic Regression

- ❑ Model selection:
  - ❑ ✖ *texture, smoothness, fractal dimension and symmetry* (lack of separation)
  - ❑ ✖ Correlated predictors
- ❑ Hyperparameter tuning
  - ❑ Cross validation
- ❑ Interaction term
  - ❑ *compactness_mean*concave_points_mean*

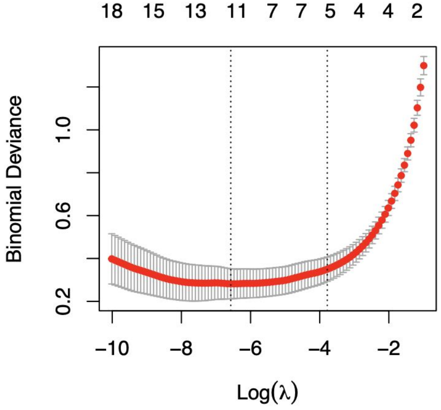
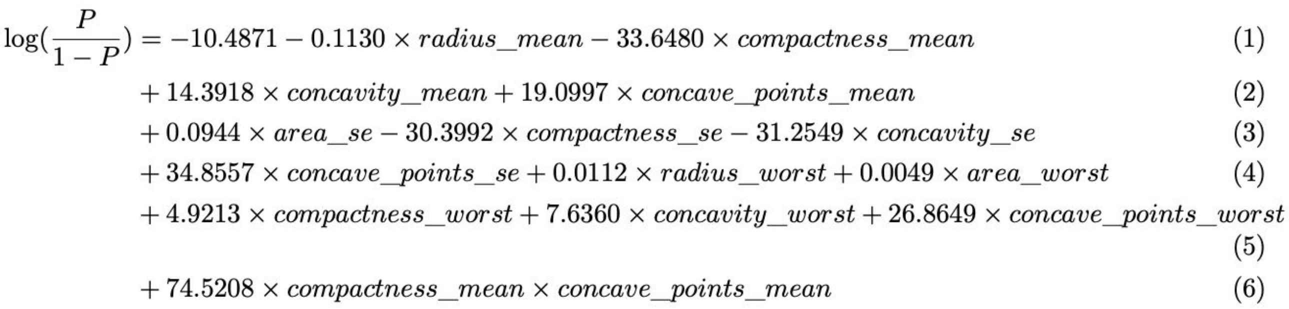


Figure 6: LASSO parameter tuning

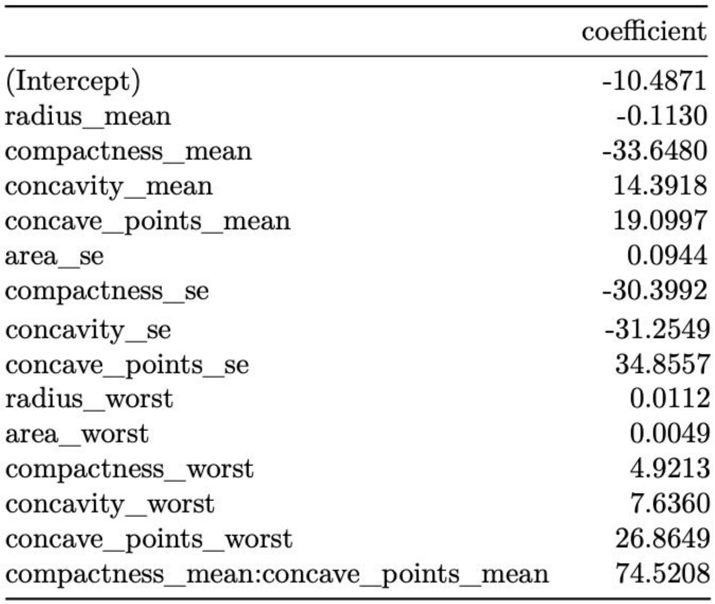
$\lambda = 0.001399$

Logistic Model



$$
\begin{aligned}
\log(\frac{P}{1-P}) = &-10.4871 - 0.1130 \times radius\_mean - 33.6480 \times compactness\_mean && (1)\\
&+ 14.3918 \times concavity\_mean + 19.0997 \times concave\_points\_mean && (2)\\
&+ 0.0944 \times area\_se - 30.3992 \times compactness\_se - 31.2549 \times concavity\_se && (3)\\
&+ 34.8557 \times concave\_points\_se + 0.0112 \times radius\_worst + 0.0049 \times area\_worst && (4)\\
&+ 4.9213 \times compactness\_worst + 7.6360 \times concavity\_worst + 26.8649 \times concave\_points\_worst && (5)\\
&+ 74.5208 \times compactness\_mean \times concave\_points\_mean && (6)
\end{aligned}
$$

- ❑ *radius_mean*, *compactness_mean*, *compactness_se* negatively associated with response log-odds while others positive
- ❑ Most significant:
  - ❑ *concave_points_se* (+)
  - ❑ *compactness_mean* (-)



	coefficient
(Intercept)	-10.4871
radius_mean	-0.1130
compactness_mean	-33.6480
concavity_mean	14.3918
concave_points_mean	19.0997
area_se	0.0944
compactness_se	-30.3992
concavity_se	-31.2549
concave_points_se	34.8557
radius_worst	0.0112
area_worst	0.0049
compactness_worst	4.9213
concavity_worst	7.6360
concave_points_worst	26.8649
compactness_mean:concave_points_mean	74.5208

Interaction Term **Visualization**

- ❑ Three fixed levels of *compactness_mean*
  - ❑ Low: 0.01938
  - ❑ Medium: 0.26
  - ❑ High: 0.28670
- ❑ Observations:
  - ❑ Effect of *concave_points_mean* on P changes on different levels of *compactness_mean*
  - ❑ High → *concave_points_mean* has strongest effect on P
  - ❑ Medium → *concave_points_mean* has strong effect on P
  - ❑ Low → *concave_points_mean* has weak effect on P

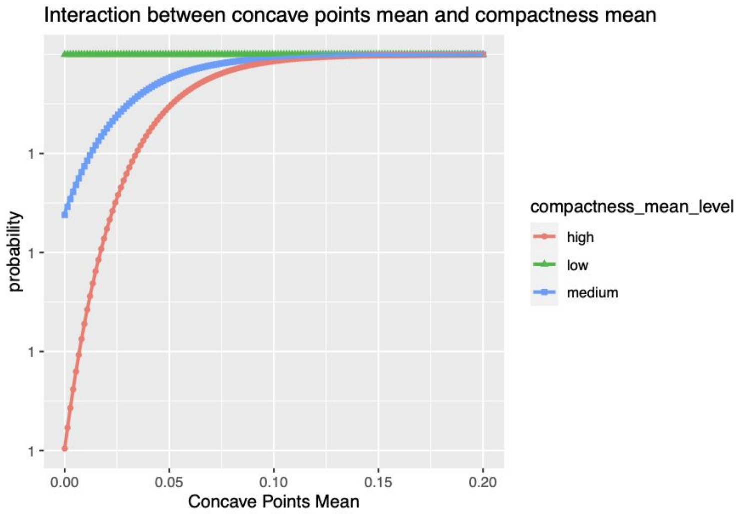
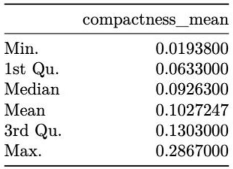
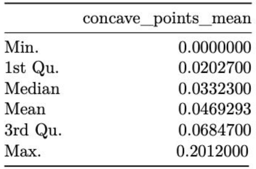


Figure 7: Interaction Term Visualization

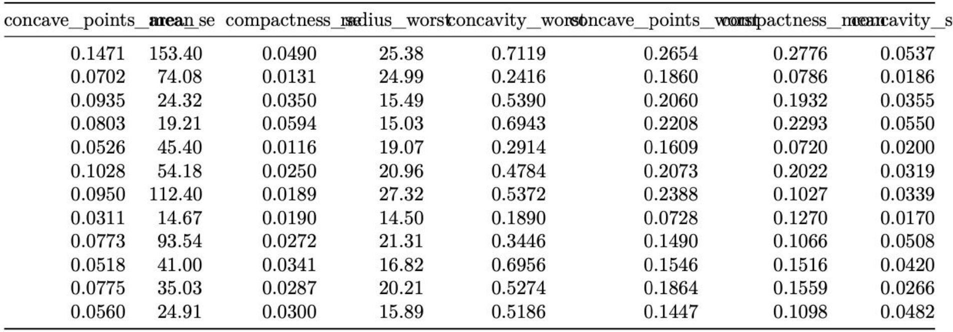


	compactness_mean
Min.	0.0193800
1st Qu.	0.0633000
Median	0.0926300
Mean	0.1027247
3rd Qu.	0.1303000
Max.	0.2867000

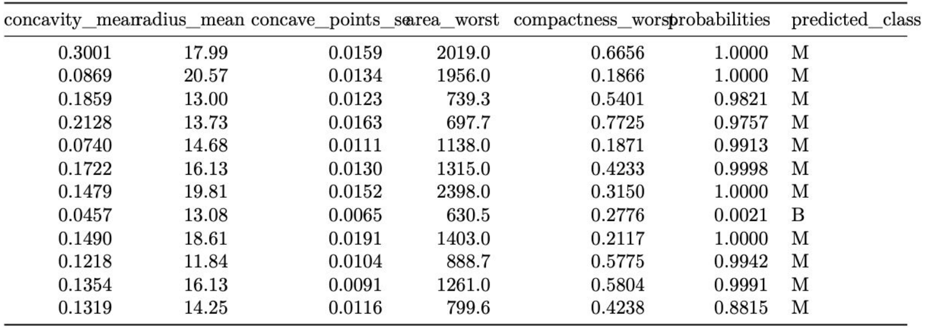


	concave_points_mean
Min.	0.0000000
1st Qu.	0.0202700
Median	0.0332300
Mean	0.0469293
3rd Qu.	0.0684700
Max.	0.2012000

Prediction



concave_points_mean	area_se	compactness_se	radius_worst	concavity_worst	concave_points_worst	compactness_mean	concavity_se
0.1471	153.40	0.0490	25.38	0.7119	0.2654	0.2776	0.0537
0.0702	74.08	0.0131	24.99	0.2416	0.1860	0.0786	0.0186
0.0935	24.32	0.0350	15.49	0.5390	0.2060	0.1932	0.0355
0.0803	19.21	0.0594	15.03	0.6943	0.2208	0.2293	0.0550
0.0526	45.40	0.0116	19.07	0.2914	0.1609	0.0720	0.0200
0.1028	54.18	0.0250	20.96	0.4784	0.2073	0.2022	0.0319
0.0950	112.40	0.0189	27.32	0.5372	0.2388	0.1027	0.0339
0.0311	14.67	0.0190	14.50	0.1890	0.0728	0.1270	0.0170
0.0773	93.54	0.0272	21.31	0.3446	0.1490	0.1066	0.0508
0.0518	41.00	0.0341	16.82	0.6956	0.1546	0.1516	0.0420
0.0775	35.03	0.0287	20.21	0.5274	0.1864	0.1559	0.0266
0.0560	24.91	0.0300	15.89	0.5186	0.1447	0.1098	0.0482



concavity_mean	radius_mean	concave_points_se	area_worst	compactness_worst	probabilities	predicted_class
0.3001	17.99	0.0159	2019.0	0.6656	1.0000	M
0.0869	20.57	0.0134	1956.0	0.1866	1.0000	M
0.1859	13.00	0.0123	739.3	0.5401	0.9821	M
0.2128	13.73	0.0163	697.7	0.7725	0.9757	M
0.0740	14.68	0.0111	1138.0	0.1871	0.9913	M
0.1722	16.13	0.0130	1315.0	0.4233	0.9998	M
0.1479	19.81	0.0152	2398.0	0.3150	1.0000	M
0.0457	13.08	0.0065	630.5	0.2776	0.0021	B
0.1490	18.61	0.0191	1403.0	0.2117	1.0000	M
0.1218	11.84	0.0104	888.7	0.5775	0.9942	M
0.1354	16.13	0.0091	1261.0	0.5804	0.9991	M
0.1319	14.25	0.0116	799.6	0.4238	0.8815	M

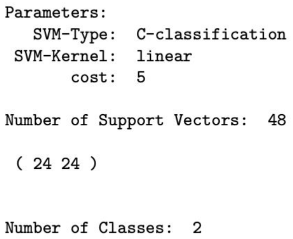
Prediction Accuracy = 0.9883

❑ Patient in row 1: 100% probability of being diagnosed as malignant cancer

❑ Patient in row 5 to last: 0.2089% probability of being diagnosed as malignant

❑ Patient in last row: 88.1% probability of being diagnosed as malignant

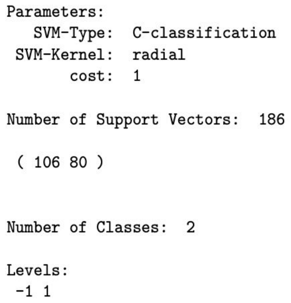
Model 2: SVM



```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  5

Number of Support Vectors:  48

 ( 24 24 )


Number of Classes:  2
```

Linear Kernel

Radial Kernel



```
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  186

 ( 106 80 )


Number of Classes:  2

Levels:
 -1 1
```

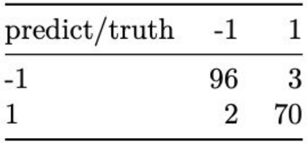
- ❑ Predictors selected by LASSO-penalized logistic regression
- ❑ Hyperparameter tuning via 10-fold cross validation

Linear Kernel:

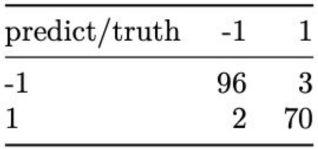
- ❑ Cost from 0.001 to 100
  - ❑ Optimal cost = 5
- ❑ Misclassification% = 0.02924

Radial Kernel:

- ❑ Cost from 0.001 to 1000, gamma from 0.5 to 4
  - ❑ Optimal cost = 1, gamma=0.5
- ❑ Misclassification% = 0.02924



predict/truth	-1	1
-1	96	3
1	2	70

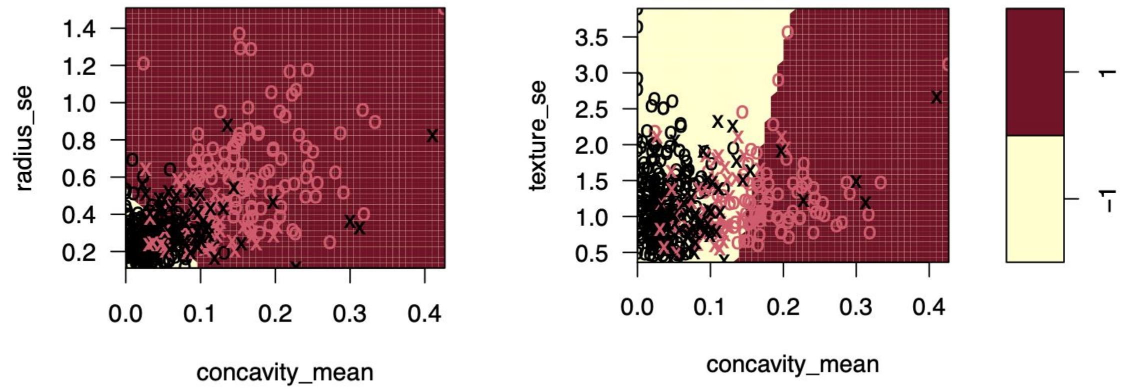


predict/truth	-1	1
-1	96	3
1	2	70

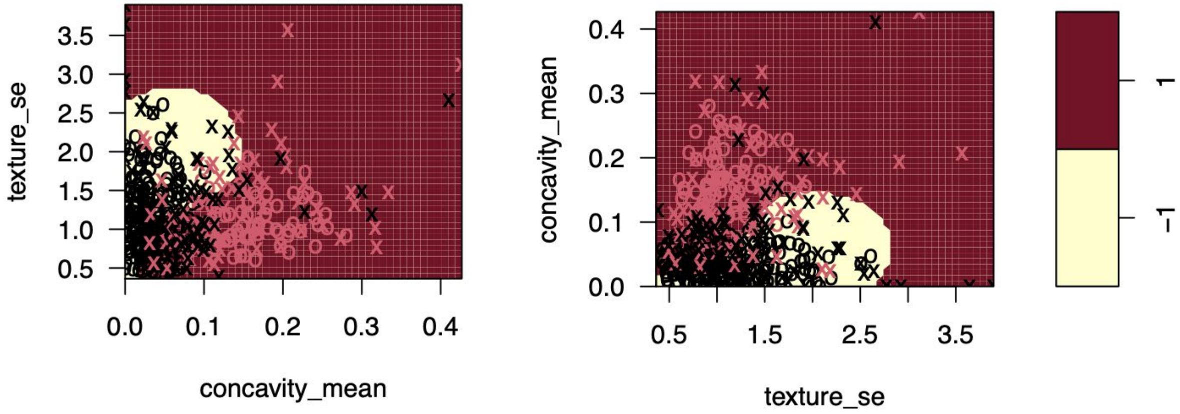
1: malignant
-1: benign

SVM **Decision Boundary**

Linear

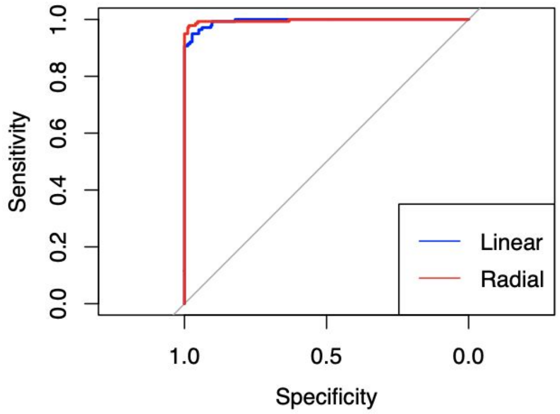


Radial



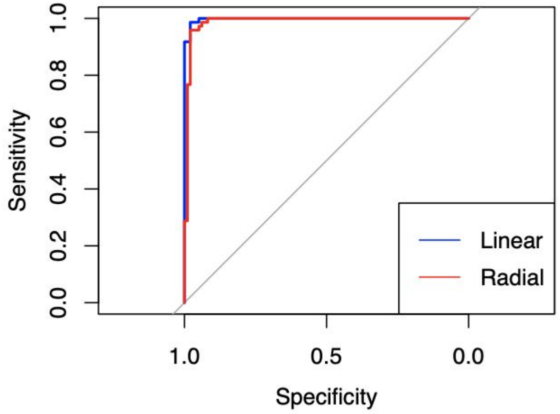
ROC Diagnosis

Train ROC



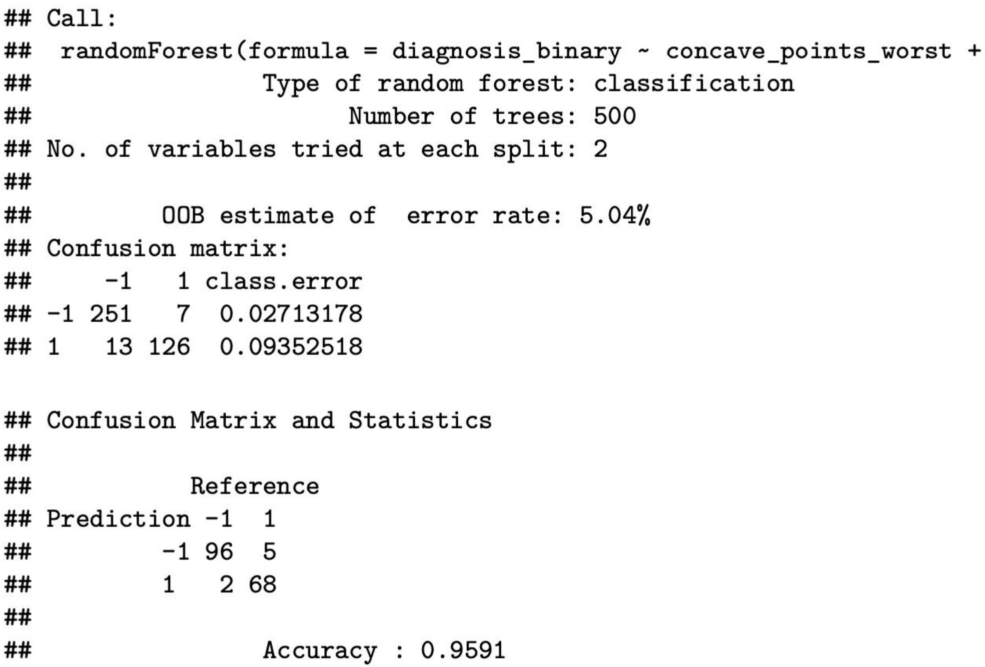
Test ROC

On test data:



- Radial kernel fits more closely on train data, linear performs better on test data
- Select the SVM model linear kernel
  - Lower variance
  - Better performance on future data

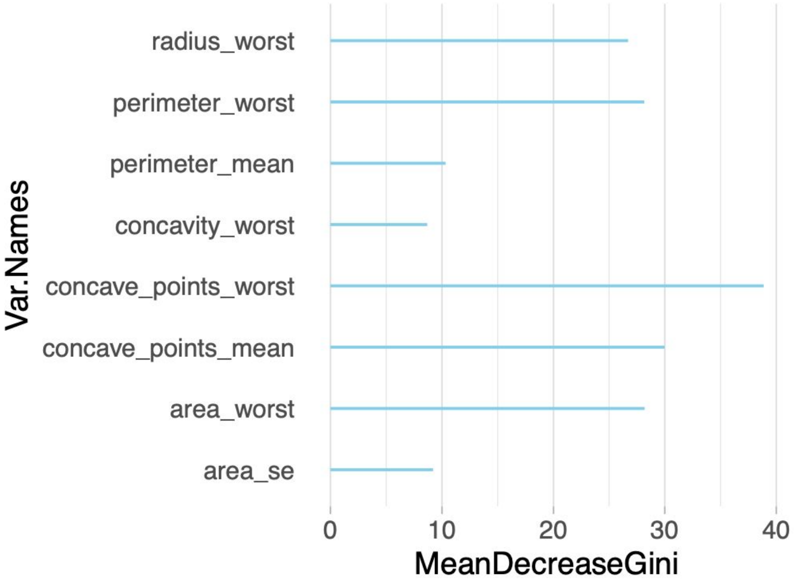
Model 3: Random Forest



```
## Call:
##  randomForest(formula = diagnosis_binary ~ concave_points_worst +
##                 Type of random forest: classification
##                       Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 5.04%
## Confusion matrix:
##     -1   1 class.error
## -1 251   7  0.02713178
## 1    13 126  0.09352518
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction -1  1
##         -1 96  5
##         1   2 68
##
##                 Accuracy : 0.9591
```

Model 3 : Random Forest



🟦 **Concave Points**

🟦 **Area**

🟦 **Perimeter**

🟦 **Radius**

5.

Conclusion

Conclusion

Important factors in differentiating benign hyperplasia from malignant carcinoma:

- Concave point
- Radius, area, perimeter
- Compactness

The ductal carcinoma cells showed higher values for nuclear area, perimeter, diameter, compactness, and concave points when compared to benign hyperplasia.

Conclusion

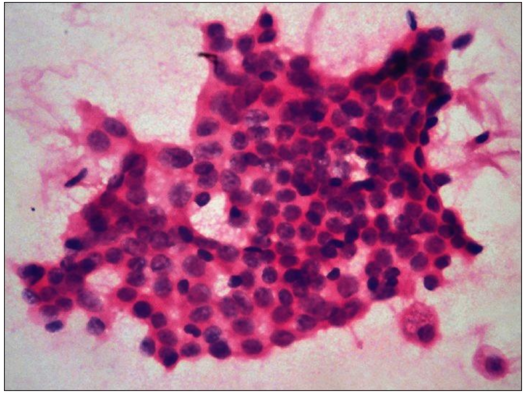
Concavity: the severity of concave portions of the contour. A high concavity means that the boundary of the cell nucleus has indentations, and thus is rather rough than smooth.

Concave points: the number of concave portions of the contour of the cell nucleus. If the contour contains one real cell, the added concave point separates one cell into two parts.

The odds of malignant diagnosis increases as perimeter, area, radius increases.

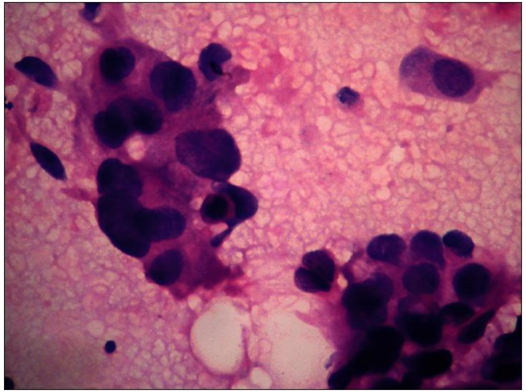
Interaction: `compactness_mean:concave_points_mean`

Clinical Evidence



Mild variation in size and shape.

Benign



Deformed and increased indentation.

Malignant

Conclusion

- LASSO-Penalized Logistic: 98.8% accuracy.
- Linear and Radial SVM: 97% accuracy,
- Random Forest : 96% accuracy.

Be careful about the False Negative rate! -For cancer diagnosis, it is much more dangerous to have a high type 2 error than type 1.

The false negative rate in the both the linear and radial SVM is: 0.041, and in the RF is: 0.068. Therefore, we decided that SVM is a better classification model than RF in this case.

Limitation and Future Work

- Few predictors and high multicollinearity.
- Only Wisconsin.
- Potential omitted variables that might cause endogeneity.

References

Centers for Disease Control and Prevention. (2022, September 26). Breast cancer. Centers for Disease Control and Prevention. https://www.cdc.gov/cancer/breast/index.htm

Mayo Foundation for Medical Education and Research. (2022, December 14). Breast cancer. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

Narasimha A, Vasavi B, Kumar HM. Significance of nuclear morphometry in benign and malignant breast aspirates. Int J Appl Basic Med Res. 2013 Jan;3(1):22-6. doi: 10.4103/2229-516X.112237. PMID: 23776836; PMCID: PMC3678677.

Wolberg WH, Street WN, Mangasarian OL. Importance of nuclear morphology in breast cancer prognosis. Clin Cancer Res. 1999;5:3542–8.