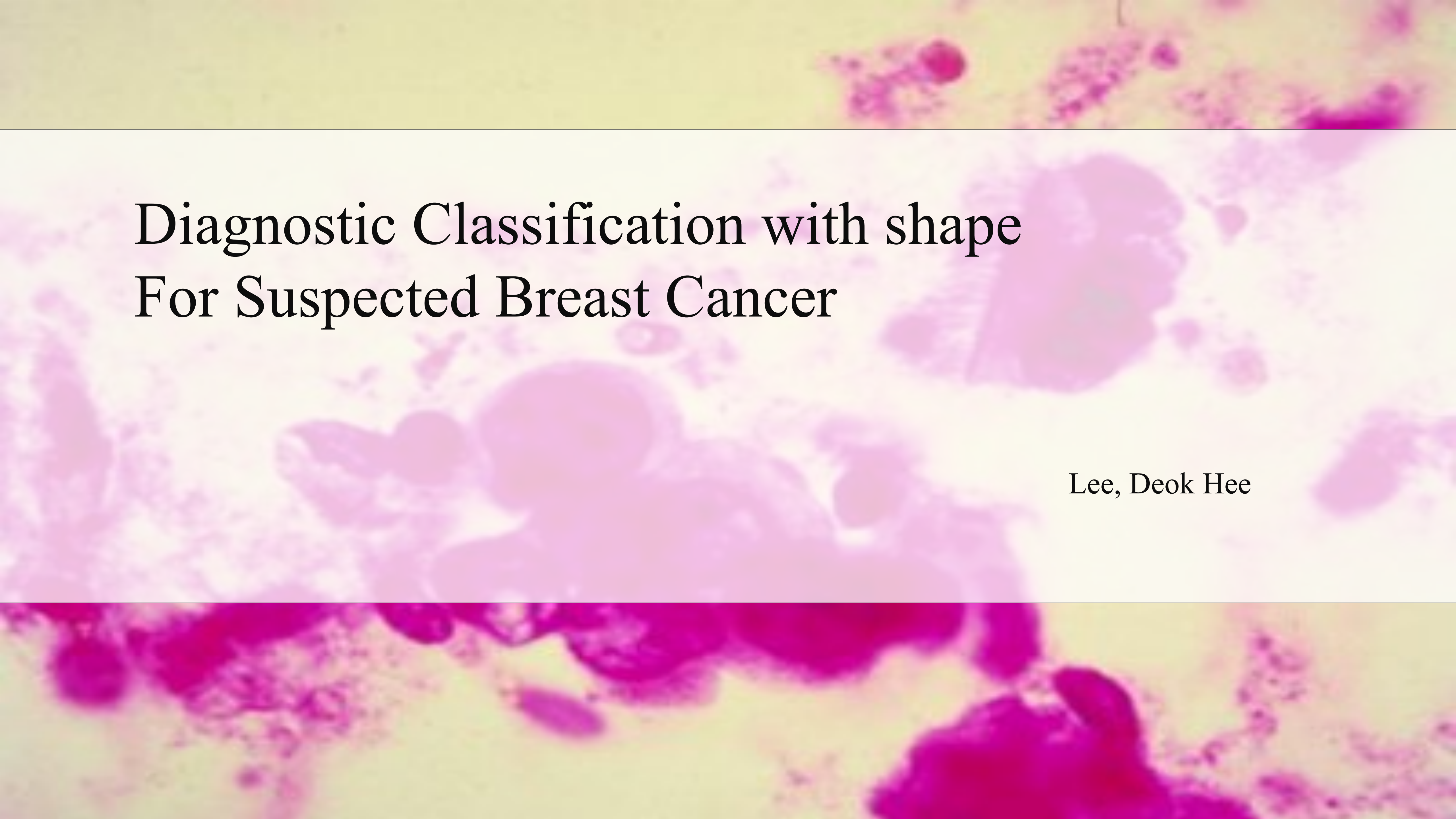


Diagnostic Classification with shape For Suspected Breast Cancer

Lee, Deok Hee

CONTENTS

- Danger of Breast Cancer
- How to Preprocessing
- Analysis of Breast Cancer Dataset
- Result

Breast Cancer

- Breast cancer is the most common disease in women worldwide. (IARC, 2013)
- Excluding thyroid cancer, breast cancer is the most frequently diagnosed cancer in women living in Korea. (NCIC, 2013)
- Approximately 77% of women with breast cancer are over the age of 50 at the time of diagnosis (USDHHS, 2008, Aug).
- If current rates stay the same, a woman born today has about a 1 in 8 chance of developing breast cancer over the course of her lifetime (NCI, 2010, Sep)

Danger of Breast Cancer lurks in women

Breast Cancer

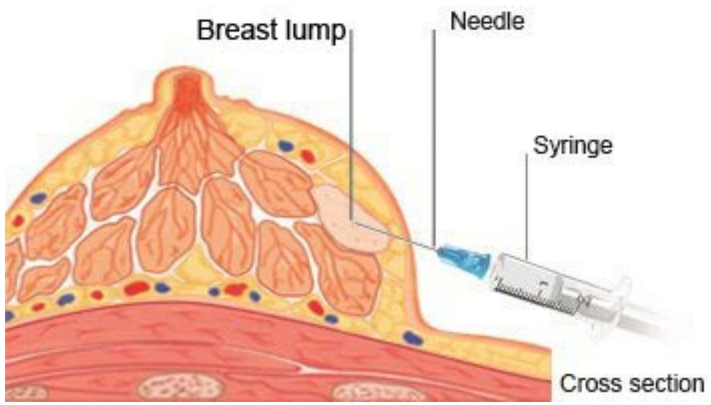
- Among breast cancer patients, 37.9% were in the first stage and 35.7% in the second stage, This show that patients with relatively early breast cancer account for more than 70%. (KBCS, 2008)

- The 5-year survival rate of breast cancer patients was 99% in group 0 and 1, 89% in group 2, and 59% and 28% in group 3 and 4 rapidly. (KBCS, 2008)
- In order to deal with the uncertainty of whether or not you have cancer, it is best to push for an early and proper diagnosis. The earlier cancer is diagnosed and treated, the better the chances of it being cured.

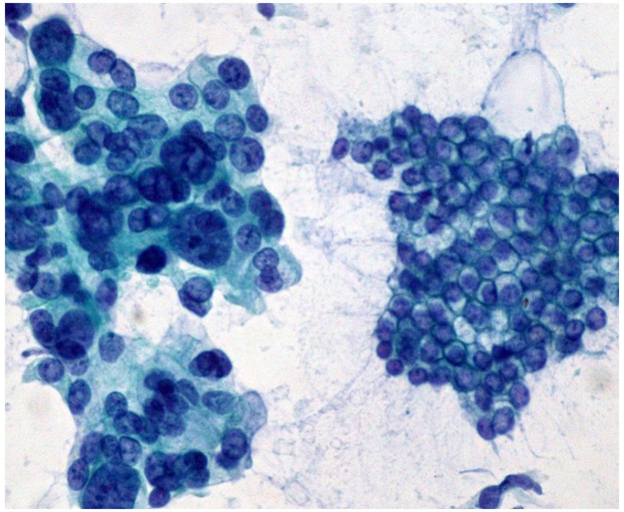
Early detection and accurate diagnosis are very important

Diagnosis Cancer

- Fine-needle Aspiration(FNA) is a diagnostic procedure used to investigate lumps or masses.
- During FNA, a long, thin needle is inserted into the suspicious area. A syringe is used to draw out fluid and cells for analysis.



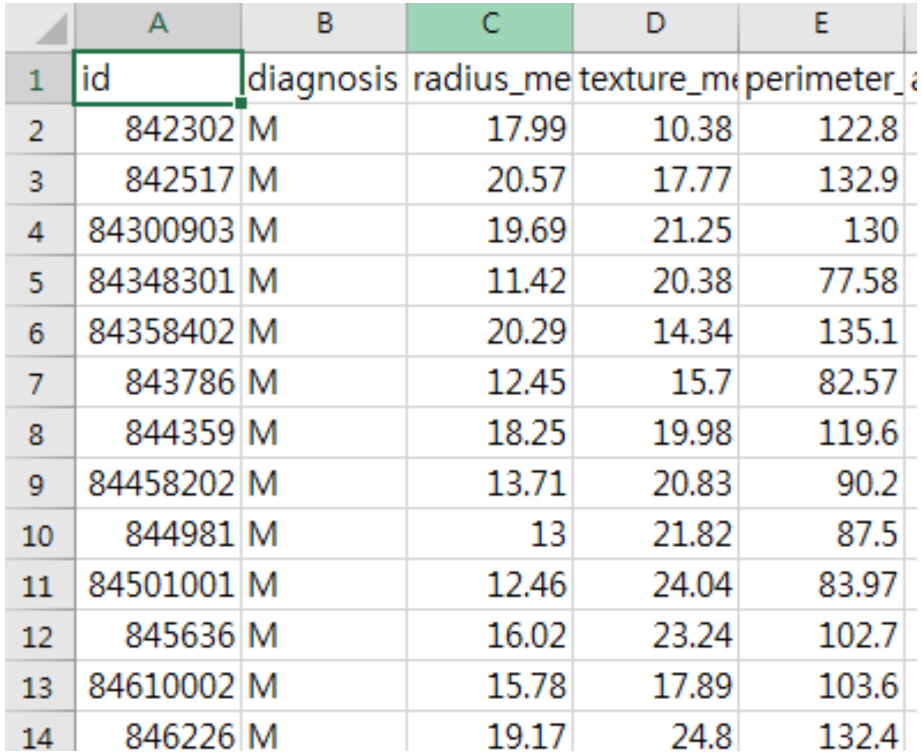
<Fine Needle Aspiration>



< FNA of Tissue >

A Cancer is seen on the left, normal cells on the right.

Data Gathering



	A	B	C	D	E
1	id	diagnosis	radius_me	texture_m	perimeter_
2	842302	M	17.99	10.38	122.8
3	842517	M	20.57	17.77	132.9
4	84300903	M	19.69	21.25	130
5	84348301	M	11.42	20.38	77.58
6	84358402	M	20.29	14.34	135.1
7	843786	M	12.45	15.7	82.57
8	844359	M	18.25	19.98	119.6
9	84458202	M	13.71	20.83	90.2
10	844981	M	13	21.82	87.5
11	84501001	M	12.46	24.04	83.97
12	845636	M	16.02	23.24	102.7
13	84610002	M	15.78	17.89	103.6
14	846226	M	19.17	24.8	132.4

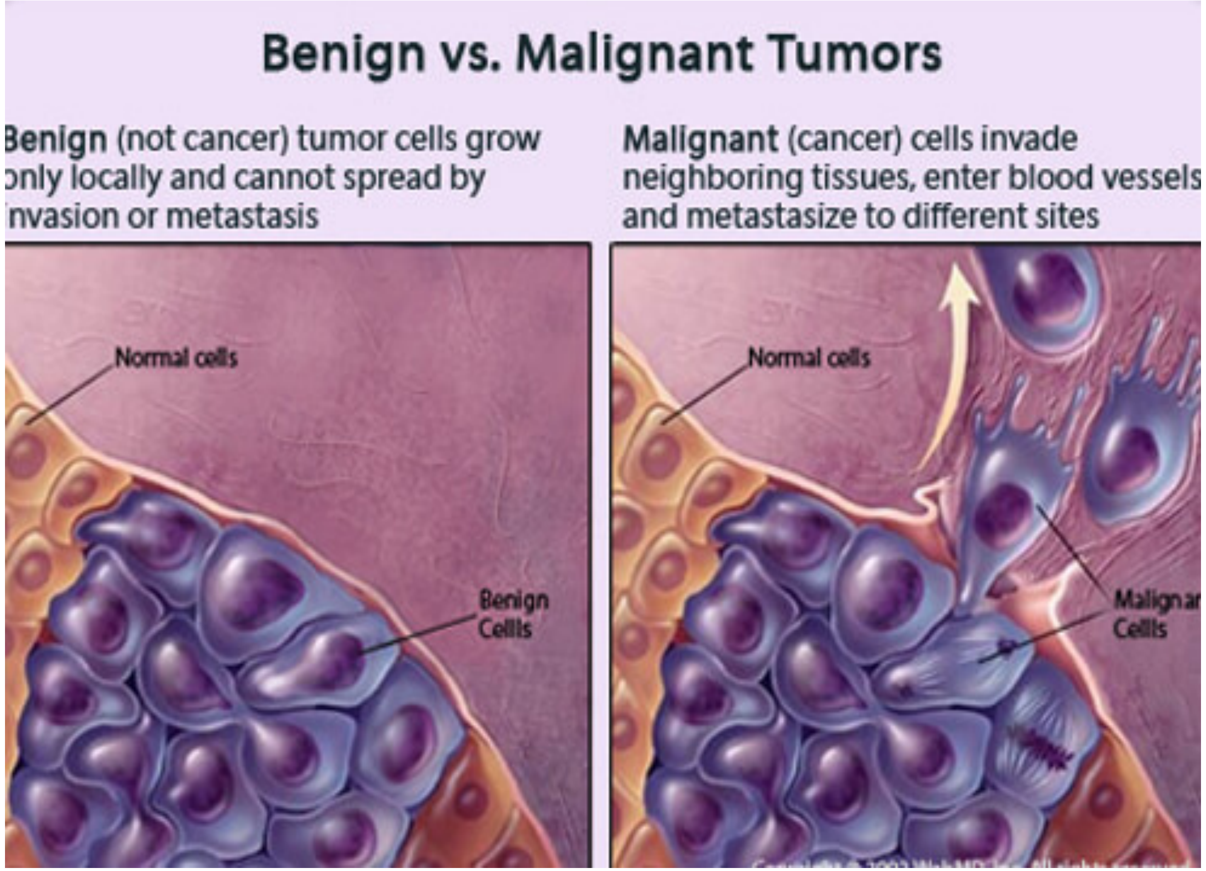
Kaggle : Breast Cancer Wisconsin Data Set

Number of Records :	569
Number of Attributes :	32
Data set Characteristics :	Multivariate
Attribute Characteristics :	Real
Associated Tasks :	Classification
Missing Values ?	None

Variables Information – Dependent variable

-Diagnosis

M = malignant (cancer)
B = benign (not cancer, just tumor)



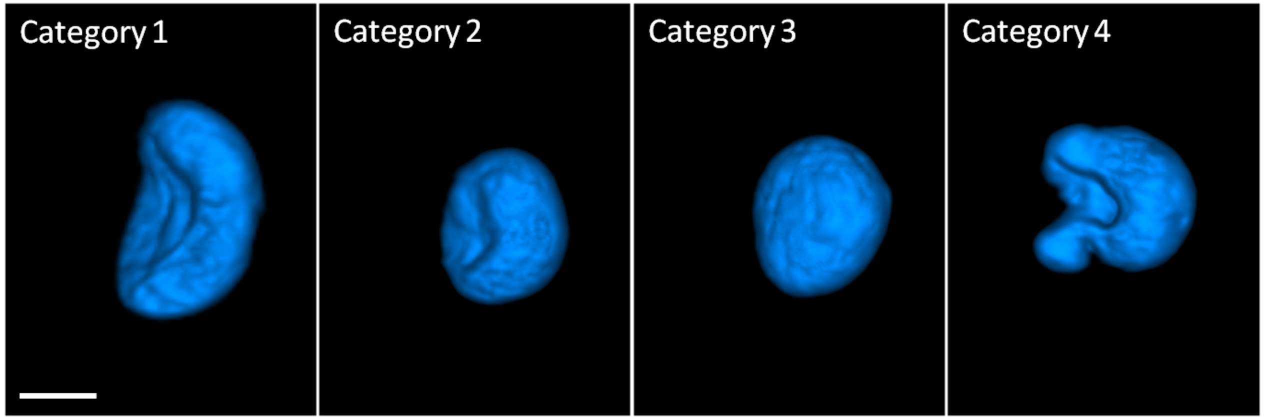
Benign vs. Malignant Tumors

Benign (not cancer) tumor cells grow only locally and cannot spread by invasion or metastasis

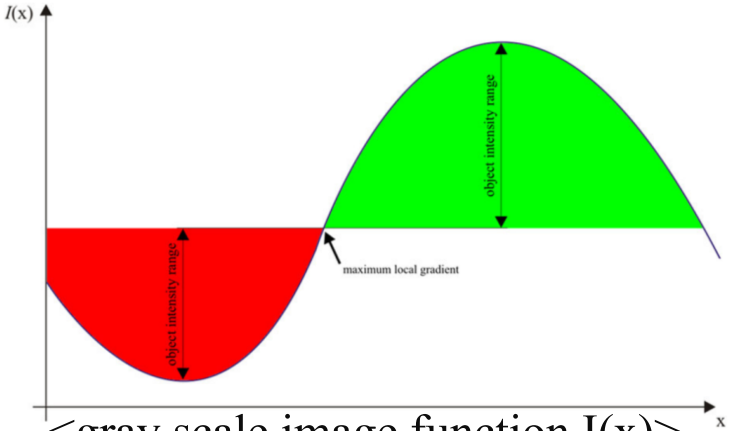
Malignant (cancer) cells invade neighboring tissues, enter blood vessels and metastasize to different sites

Variables Information – independent vaiables

-Concavity
: severity of concave portions of the contour



- Texture : standard deviation of gray-scale values



<gray scale image function I(x)>

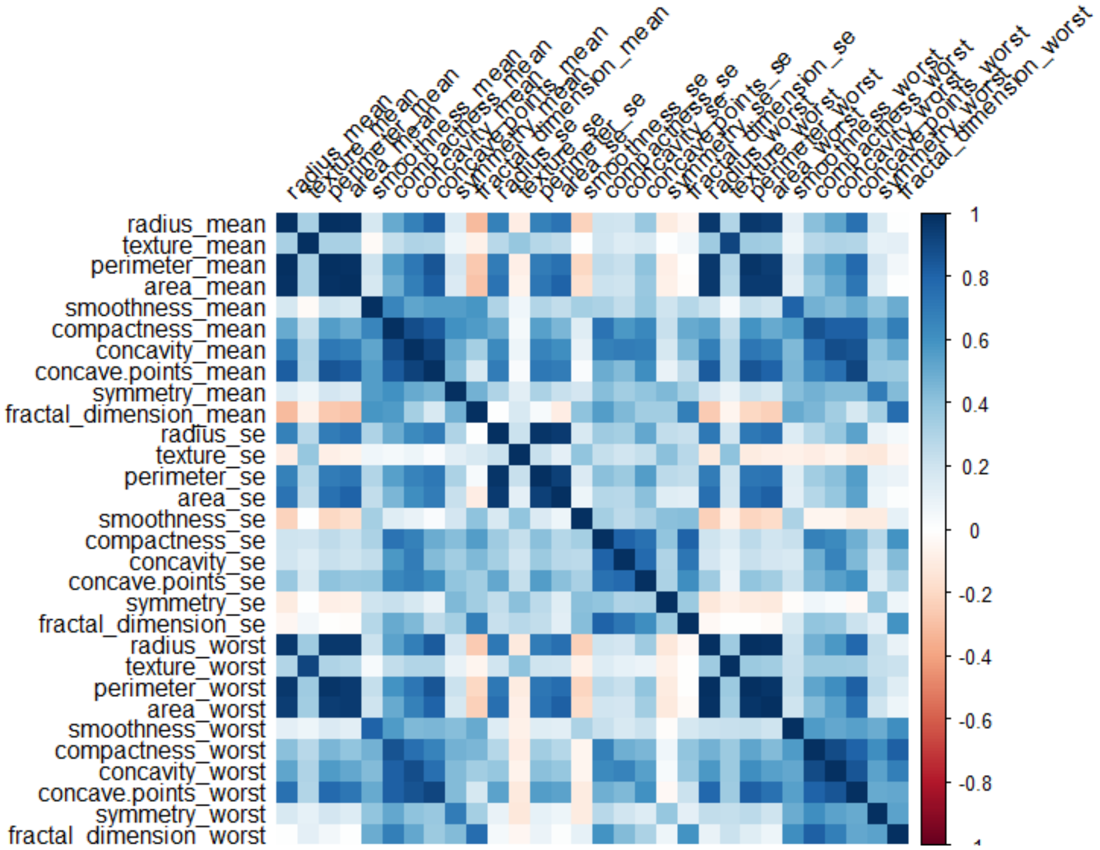
Variables Information

Number of Records :	569
Number of Attributes :	32

(Attributes:
ID, diagnosis,
30 real-valued input features)

1) ID number
2) Diagnosis (M = malignant(cancer), B = benign(just tumor))
3) Ten real-valued features are computed for each cell nucleus:
    A. Radius : mean of distances from center to points on the perimeter
    B. Texture : standard deviation of gray-scale values
    C. Perimeter : The outer limits of an area
    D. Area
    E. Smoothness : local variation in radius lengths
    F. Compactness : perimeter^2 / area - 1.0
    G. Concavity : severity of concave portions of the contour
    H. Concave points : number of concave portions of the contour
    I. Symmetry
    J. Fractal dimension : "coastline approximation" - 1

Correlation analysis



A graph that correlates 30 variables.

The higher the blue color,

the higher the positive correlation.

The higher the red color,

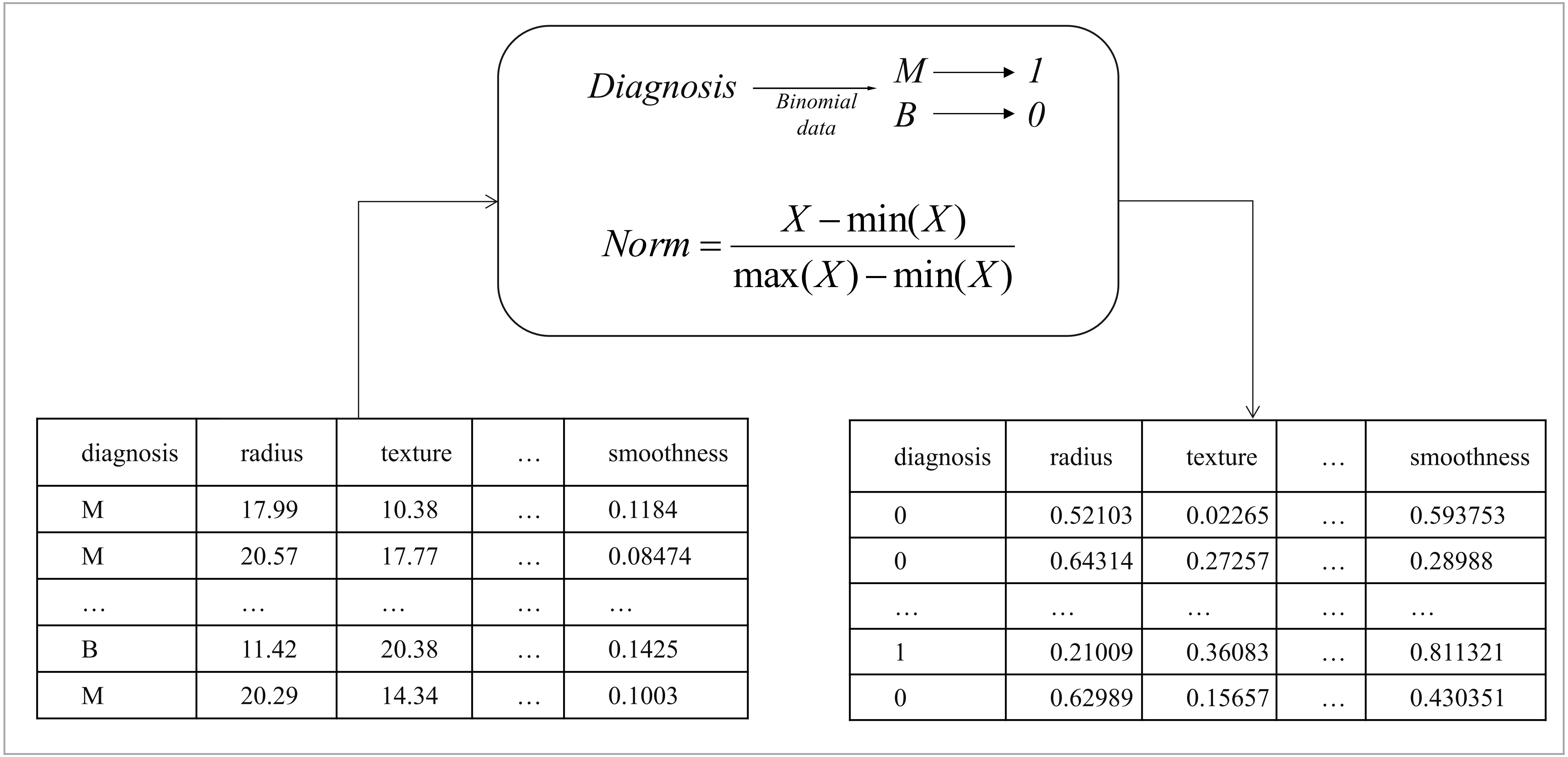
the higher the negative correlation

Variables within classification model

Characteristic	Mean	Standard error	Worst (Farthest)
Radius	radius_mean	radius_se	radius_worst
Texture	texture_mean	texture_se	texture_worst
Perimeter	perimeter_mean	perimeter_se	perimeter_worst
Area	area_mean	area_se	area_worst
Smoothness	smoothness_mean	smoothness_se	smoothness_worst
Compactness	compactness_mean	compactness_se	compactness_worst
Concavity	concavity_mean	concavity_se	concavity_worst
concave points	concave points_mean	concave points_se	concave points_worst
Symmetry	symmetry_mean	symmetry_se	symmetry_worst
fractal_dimension	fractal_dimension_mean	fractal_dimension_se	fractal_dimension_worst

Delete variable which correlation coefficient is bigger than 0.7 → 11 variables left!

Data Transformation



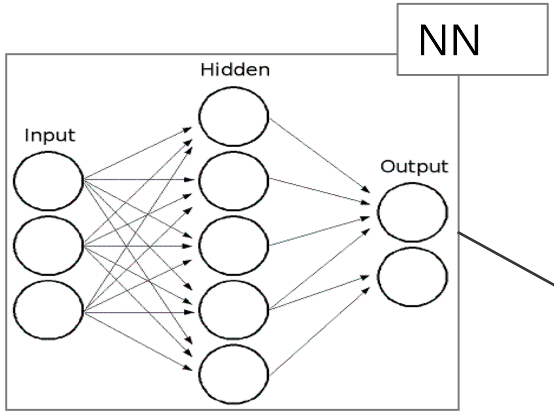
$$Diagnosis \xrightarrow[\text{data}]{\text{Binomial}} \begin{array}{l} M \longrightarrow 1 \\ B \longrightarrow 0 \end{array}$$

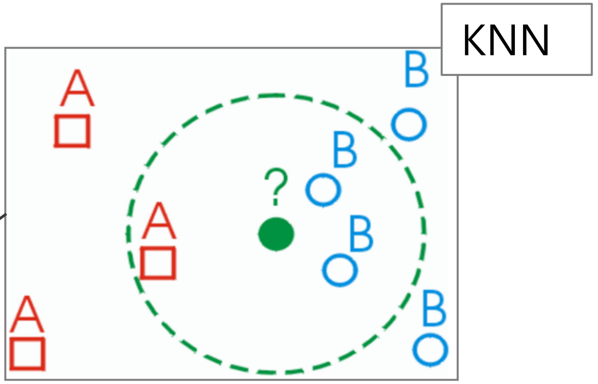
$$Norm = \frac{X - \min(X)}{\max(X) - \min(X)}$$

diagnosis	radius	texture	…	smoothness
M	17.99	10.38	…	0.1184
M	20.57	17.77	…	0.08474
…	…	…	…	…
B	11.42	20.38	…	0.1425
M	20.29	14.34	…	0.1003

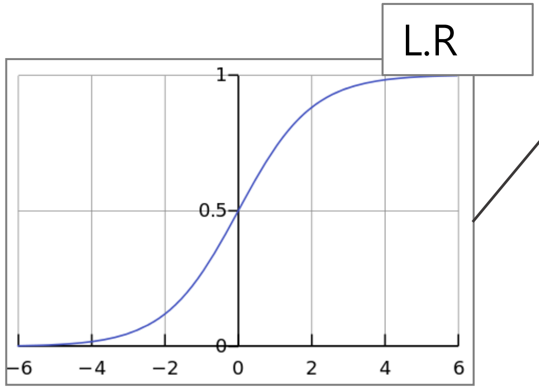
diagnosis	radius	texture	…	smoothness
0	0.52103	0.02265	…	0.593753
0	0.64314	0.27257	…	0.28988
…	…	…	…	…
1	0.21009	0.36083	…	0.811321
0	0.62989	0.15657	…	0.430351

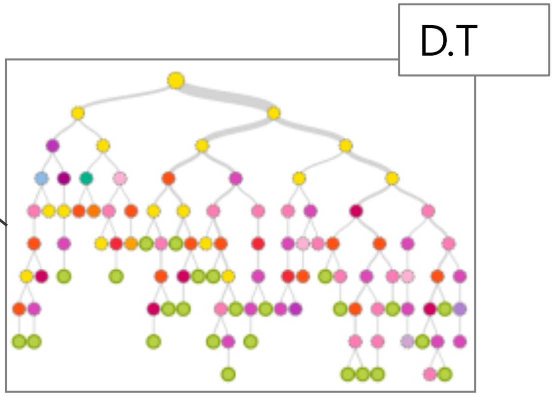
Data - Mining Algorithm(python sklearn)


NN

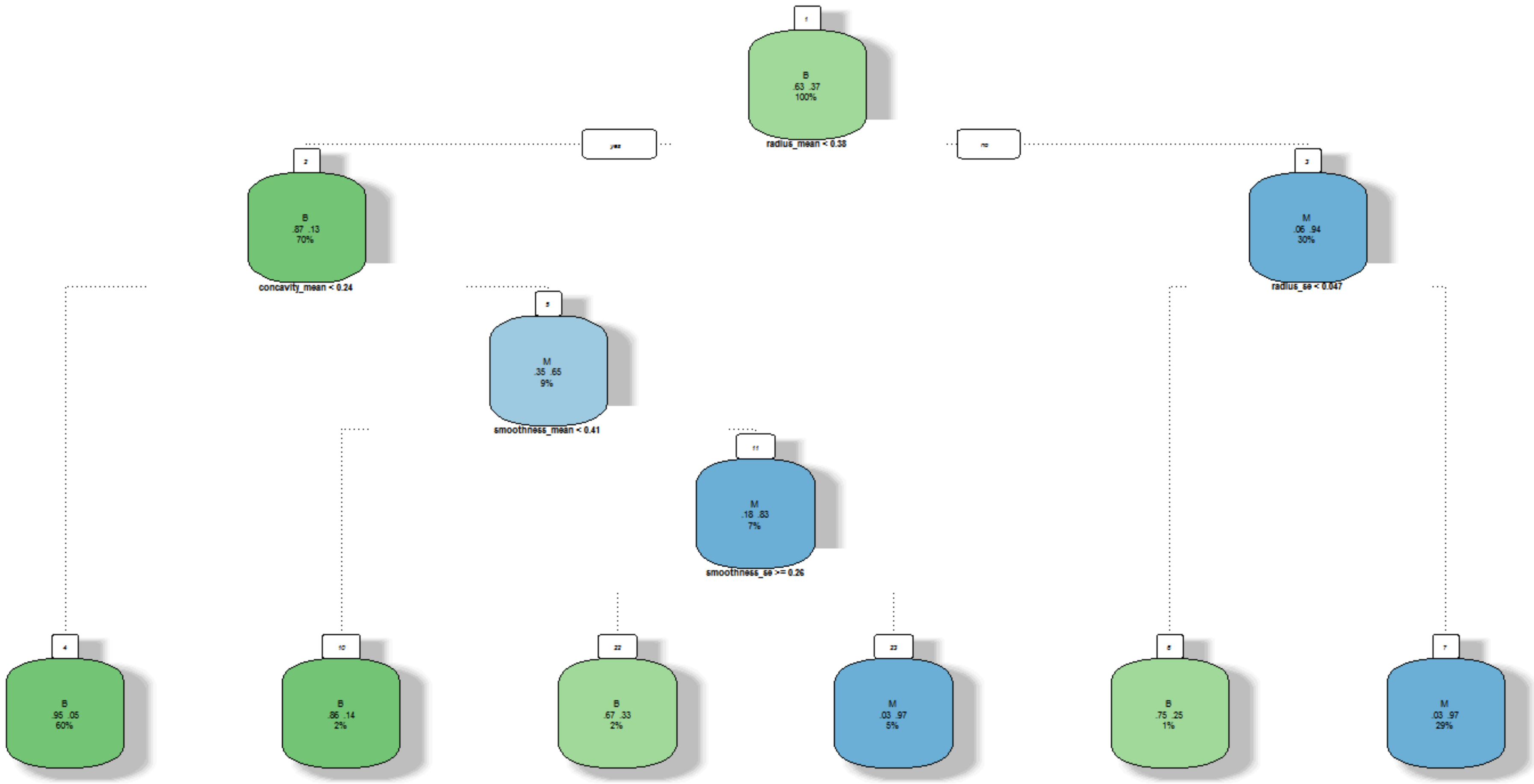

KNN

Y	Diagnosis
X_1	radius_mean
…	…
X_{11}	symmetry_se


L.R

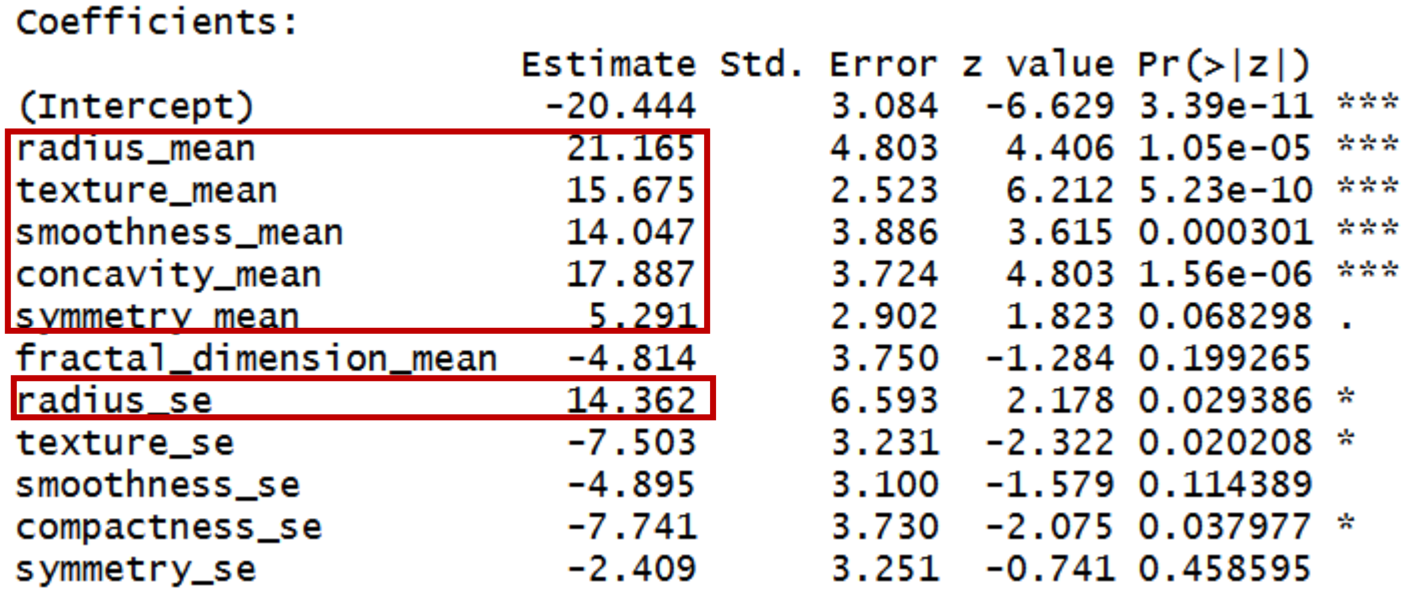

D.T

Algorithm - Decision Tree



Rattle 2016-12-06 00:05:32 user

Algorithm – Logistic Regression

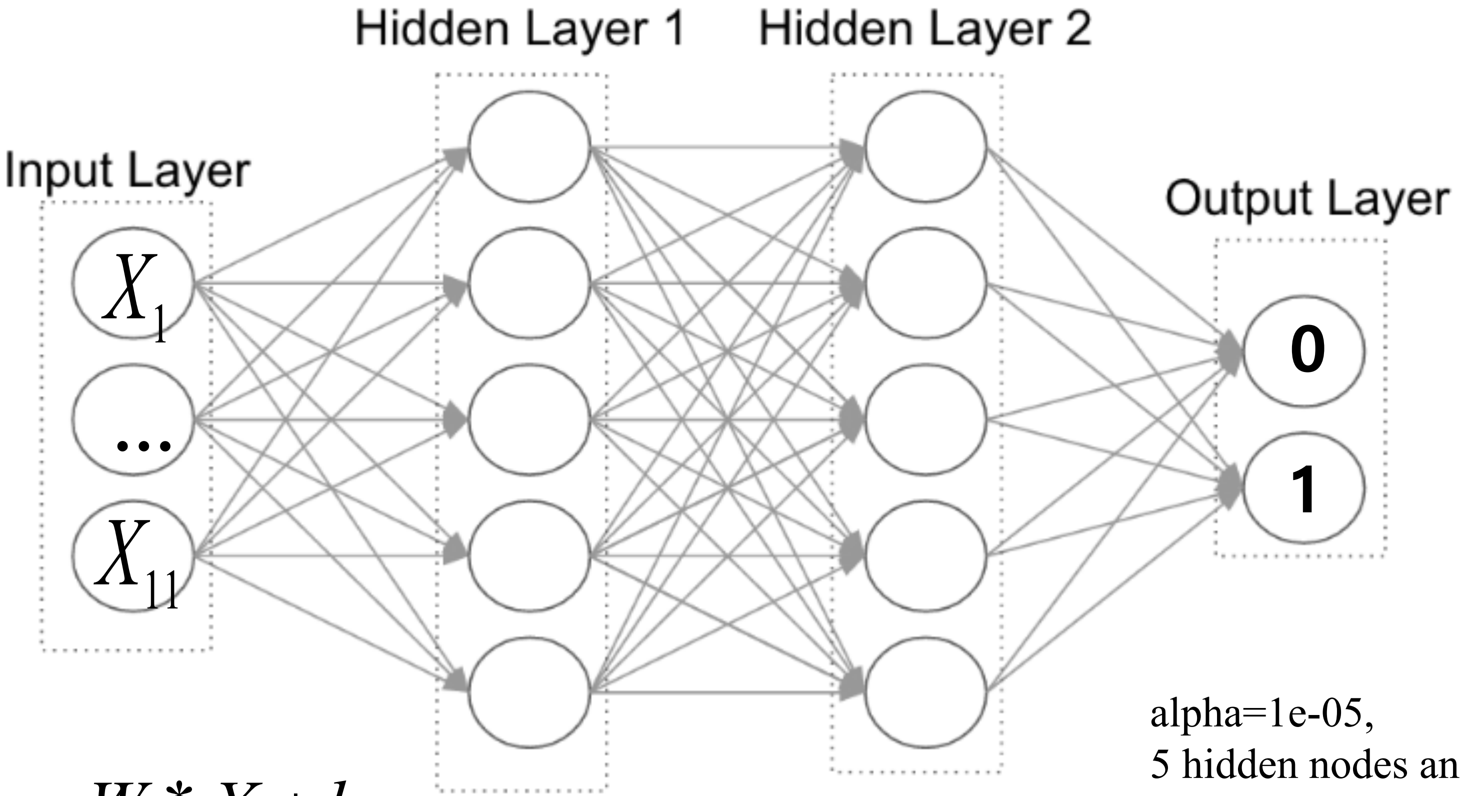


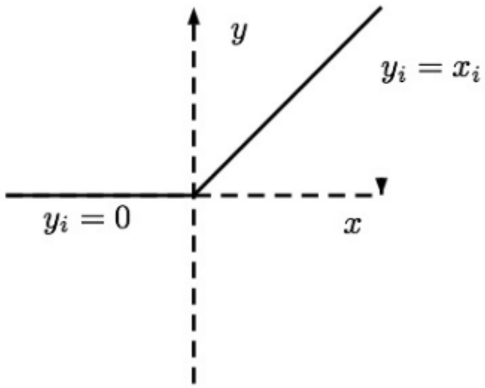
```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -20.444      3.084  -6.629 3.39e-11 ***
radius_mean              21.165      4.803   4.406 1.05e-05 ***
texture_mean             15.675      2.523   6.212 5.23e-10 ***
smoothness_mean          14.047      3.886   3.615 0.000301 ***
concavity_mean           17.887      3.724   4.803 1.56e-06 ***
symmetry_mean             5.291      2.902   1.823 0.068298 .  
fractal_dimension_mean   -4.814      3.750  -1.284 0.199265    
radius_se                14.362      6.593   2.178 0.029386 *  
texture_se               -7.503      3.231  -2.322 0.020208 *  
smoothness_se            -4.895      3.100  -1.579 0.114389    
compactness_se           -7.741      3.730  -2.075 0.037977 *  
symmetry_se              -2.409      3.251  -0.741 0.458595    
```

$$\ln\left(\frac{p}{1-p}\right) = -20.4 + 21.2 * radius\_mean + 15.7 * texture\_mean + \ldots - 2.4 * symmetry\_se$$

$$p = prob.\text{ of malignant}(1)$$

Algorithm – Neural Network





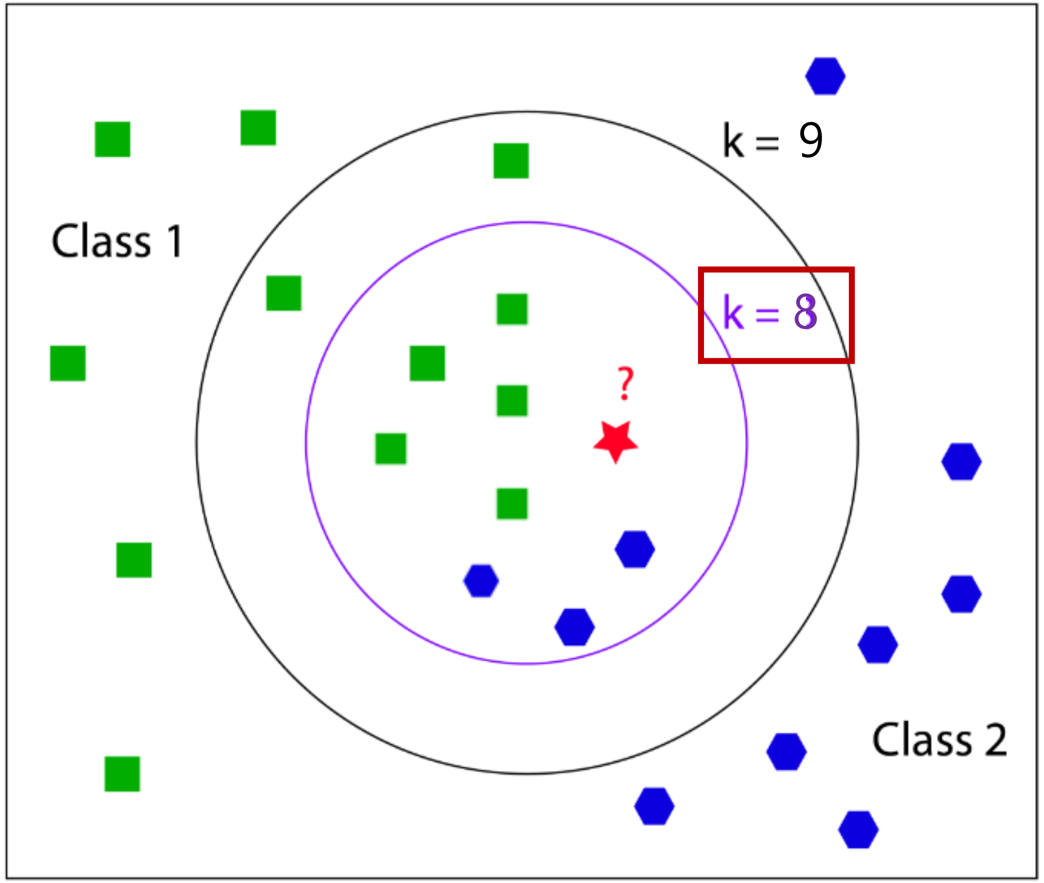
<ReLU>

alpha=1e-05,
5 hidden nodes and 2 layers

$$y = W * X + b$$

$$y = diagnosis,\ \ \mathrm{X} = [\mathrm{radius\_mean} \ldots \mathrm{symmetry\_se}]$$

Algorithm – KNN



Result train : test =(60:40)

Decision Tree

	Y_pred	
Y_true	**0**	**1**
0	70	13
1	12	133

Accuracy Score = 0.89

Logistic Regression

	Y_pred	
Y_true	**0**	**1**
0	66	17
1	3	142

Accuracy Score = 0.91

Result train : test =(60:40)

K-Nearest Neighbor (k=8)

	Y_pred	
Y_true	**0**	**1**
0	73	10
1	5	140

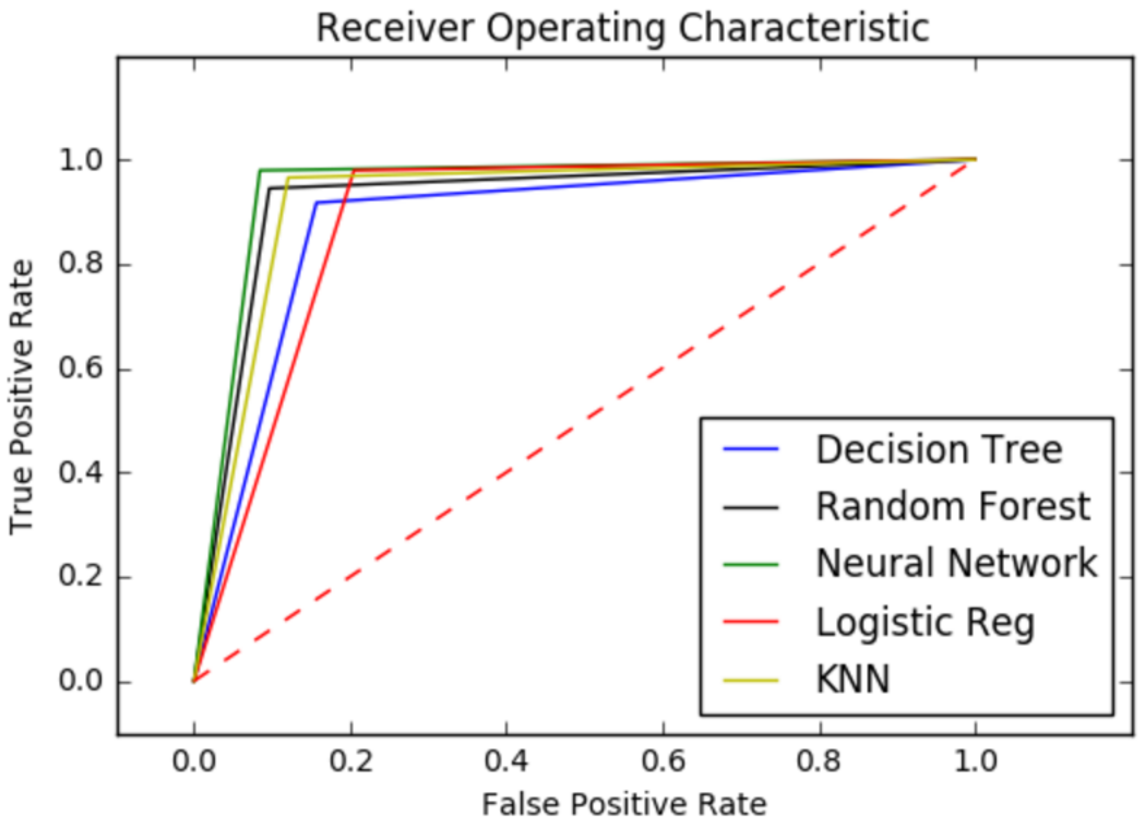
Accuracy Score = 0.93

Neural Network (activation='relu', alpha=1e-05, hidden_layer_sizes=(5, 2))

	Y_pred	
Y_true	**0**	**1**
0	76	7
1	3	142

Accuracy Score = 0.96

Model Comparison



ROC Curve

Best Algorithm is

Neural Network

Thank you