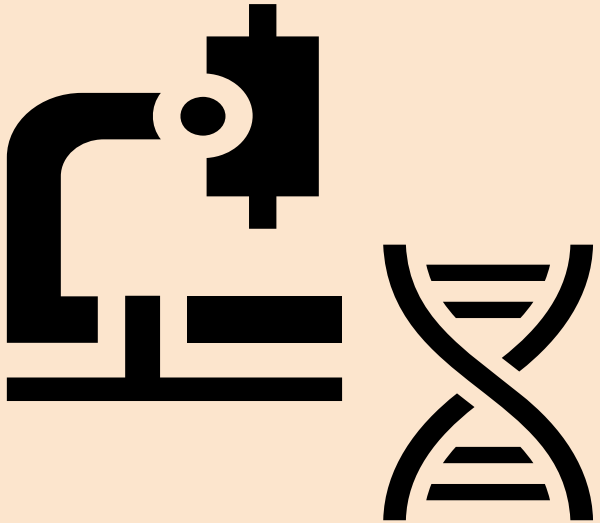
Breast Cancer prediction



Alícia Pinho Santos and Candela Palacios

Introduction and prediction task

Data set: Breast cancer
Variables: 32
Target: **diagnosis**

Variable description: Different **measures** (mean, SE and worst) of cell observation **parameters**

Diagnosis:

- **M**, Malignant
- **B,** Benign

<u>Prediction task:</u>
Predict the probability of an observed tumor being malignant or benign, based on different measures of cell observation parameters.

Variable	Description
ID	Unique ID
diagnosis	**Target: M - Malignant B - Benign**
radius_mean	Radius of Lobes
texture_mean	Mean of Surface Texture
perimeter_mean	Outer Perimeter of Lobes
area_mean	Mean Area of Lobes
smoothness_mean	Mean of Smoothness Levels
compactness_mean	Mean of Compactness
concavity_mean	Mean of Concavity
concave points_mean	Mean of Concave Points
symmetry_mean	Mean of Symmetry
fractal_dimension_mean	Mean of Fractal Dimension
radius_se	SE of Radius
texture_se	SE of Texture
perimeter_se	Perimeter of SE
area_se	Area of SE
smoothness_se	SE of Smoothness
compactness_se	SE of compactness
concavity_se	SE of concavity
concave points_se	SE of concave points
symmetry_se	SE of symmetry
fractal_dimension_se	SE of Fractal Dimension
radius_worst	Worst Radius
texture_worst	Worst Texture
perimeter_worst	Worst Permimeter
area_worst	Worst Area
smoothness_worst	Worst Smoothness
compactness_worst	Worse Compactness
concavity_worst	Worst Concavity
concave points_worst	Worst Concave Points
symmetry_worst	Worst Symmetry
fractal_dimension_worst	Worst Fractal Dimension

Exploratory Data Analysis

Missing Values

13 zeros on concavity related variables.
Median imputation

Distributions and outliers

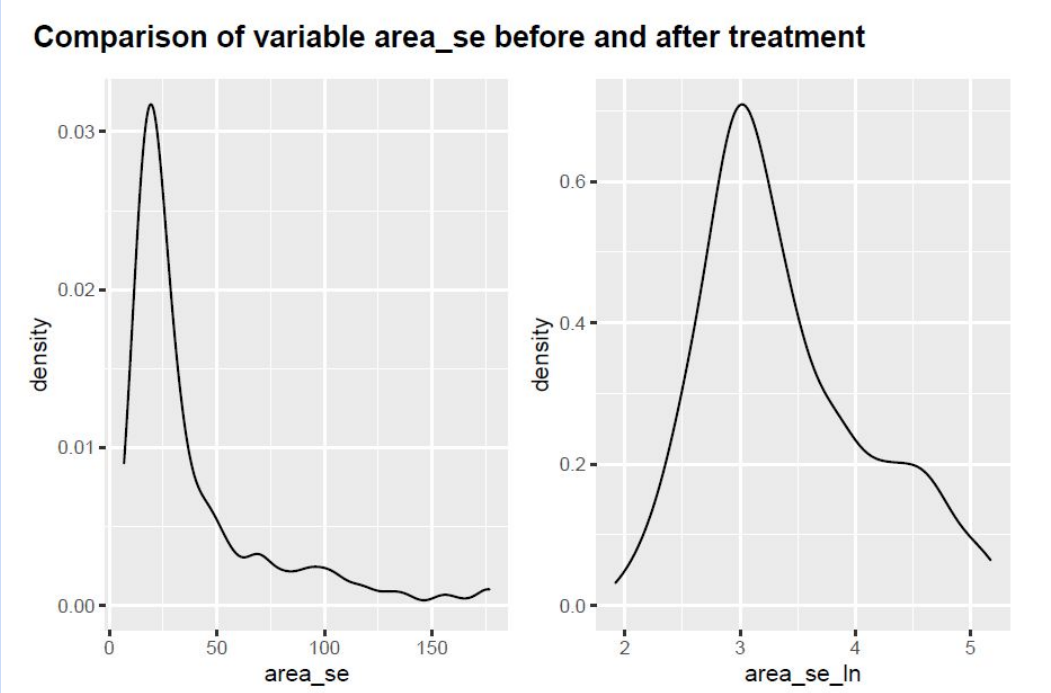
1° Loop to visualize all distributions
2° Manual transformation of each variable

Main problem

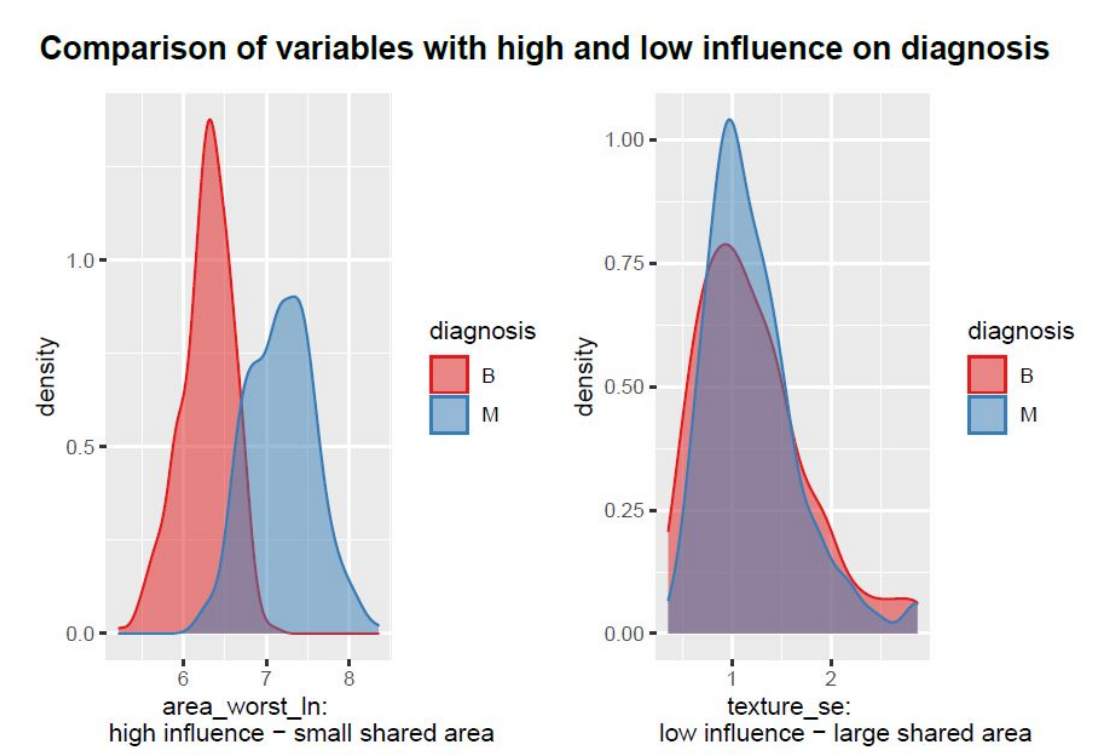
Non-normal distributions

New Features

- Log-transformed variables: radius_mean_ln, perimeter_mean_ln, area_mean_ln, perimeter_se_ln, area_se_ln,radius_worst_ln, perimeter_worst_ln and area_worst_ln
- Numeric Binary variable: **diagnosis_n**



Exploratory Data Analysis



Visualizing influence

1° Creating a loop to plot density lines of every variable for each class of diagnosis.

2° Manually classifying variables by their influence on the target variable.

Influence tables

TABLE 1 - Influence by variable

High Influence	Mid influence	Low influence
radius_mean	smoothness_mean	texture_se
perimeter_mean	compactness_se	smoothness_se
area_mean	concavity_se	symmetry_se
compactness_mean	concave points_se	fractal_dimension_se
concavity_mean	texture_worst	symmetry_mean
concave points_mean	smoothness_worst	fractal_dimension_mean
radius_se	symmetry_worst	
perimeter_se	fractal_dimension_worst	
area_se texture_mean		
radius_worst		
perimeter_worst		
area_worst		
compactness_worst		
concavity_worst		
concave points_worst		

Variables classified by **three levels** of influence.

TABLE 2 - Influence by parameter

Parameter	Type of Influence
radius	**high**
texture	one variable in each cat.
perimeter	**high**
area	**high**
smoothness	mid-low
compactness	*mid*
concavity	*mid*
concave.points	*mid*
fractal_dimension	mid-low

Combination of the three **measures of each parameter** to see the parameter's general influence

Modeling

1. Logistic regression, 3 parameters
2. Logistic regression, 9 parameters
3. Model 2 + ridge
4. All high influence variables + Lasso
5. All high influence variables on a Decision Tree
6. Model 5 with pruning

Model 1

Only the **three most influential** variables

3 Predictors: radius_worst_ln, perimeter_worst_ln and area_worst_ln

Model 2 and 3

All **the variables** of the **parameters** that scored "**high**" on our **table 2**

9 Predictors: radius_worst_ln, radius_mean_ln, radius_se_ln, perimeter_worst_ln, perimeter_mean_ln, perimeter_se_ln, area_worst_ln, area_mean_ln and area_se_ln

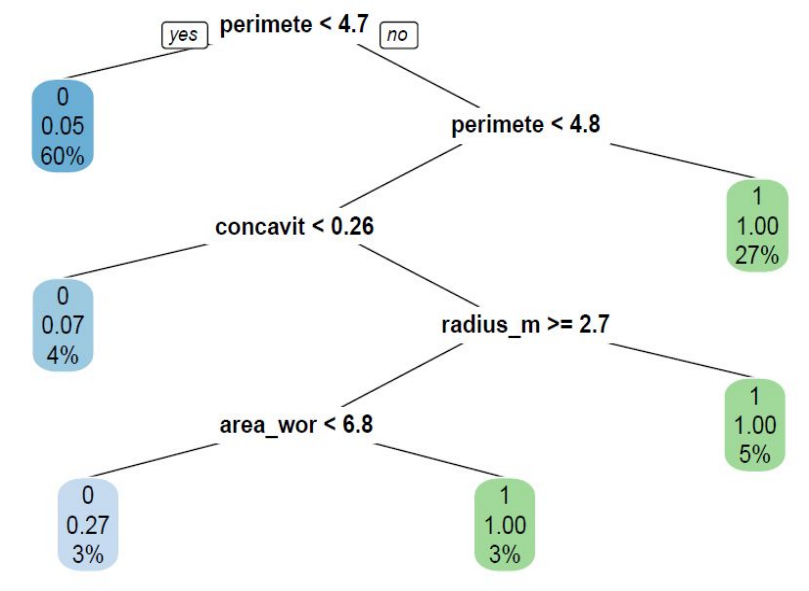
Model 4

All the **variables** that scored "**high**" on our **table 1**

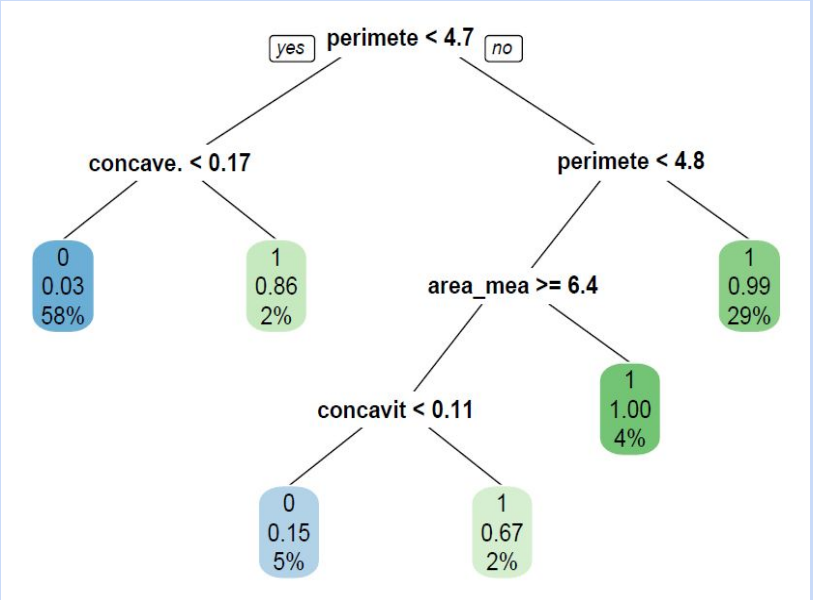
15 Predictors: radius_worst_ln, radius_mean_ln, radius_se_ln, perimeter_worst_ln, perimeter_mean_ln, perimeter_se_ln, area_worst_ln, area_mean_ln, area_se_ln, compactness_mean, concavity_mean, concave points_mean, compactness_worst, concavity_worst, concave points_worst

Modeling - Decision Trees

Model 5: All high influence variables, **without pruning**.



Model 6: All high influence variables, **with pruning**.
Parameters: cp = 0 --- minsplit = 20



Feature Selection Evaluation

	[,1]
perimeter_worst_ln	140.427364
area_worst_ln	126.485221
radius_worst_ln	126.485221
perimeter_mean_ln	123.018151
area_mean_ln	117.564021
radius_mean_ln	116.950995
concave.points_worst	15.764330
concavity_worst	7.464850
concavity_mean	7.336930
concave.points_mean	6.232332
compactness_mean	4.897750
compactness_worst	3.786689

Model 6, the **pruned** decision tree, not only works as a **predictive** model, but also allows us to evaluate if our manual **feature selection** was accurate.

All of these variables were **classified as high** influence on our **manual selection**.

The **first three** variables were the ones we used on model 1.

We conclude our selection was **pertinent**.

Evaluation and conclusion

Naive classifier

Accuracy: 0.6432749
RMSE: 0.4795689
AUC: 0.5

Root Mean Square Error (RMSE)
Model 1: 0.2196695
Model 2: 0.2139374
Model 3: 0.2328456
Model 4: 0.1954697
Model 5: 0.2440556
Model 6 0.2282298

Accuracy
Model 1: 0.9356725
Model 2: 0.9239766
Model 3: 0.9356725
Model 4: 0.9356725
Model 5: 0.9356725
Model 6: 0.9415205

Area Under Curve (AUC)
Model 1: 0.9796
Model 2: 0.9773
Model 3: 0.9745
Model 4: 0.9854
Model 5: 0.9417
Model 6: 0.9499

There is **no unconditional leader**.

Best performing model: **Model 4** - All high influence parameters with LASSO regularization.

Thank you for your attention

-Alícia and Cande-