

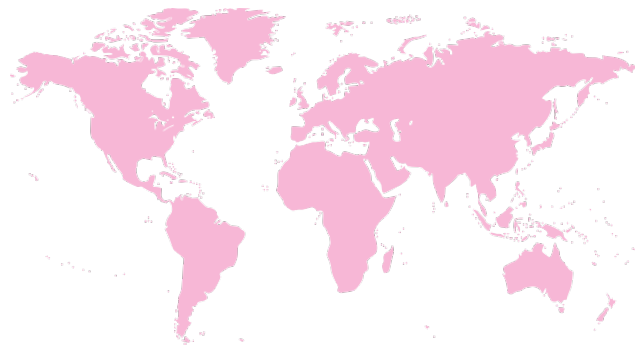
# BREAST CANCER WISCONSIN (DIAGNOSTIC) DATASET

**Supervised Analysis for Malignancy Detection**

*Authors: Pablo Laso Mielgo, Helena Sofía Yaben*

# 1. PRESENTATION OF THE PROBLEM

- Breast Cancer is the **most common type of cancer** among women across the world.
- **Leading cause of death from cancer in women.**
- **Biopsy is essential to distinguish between malignant and benignant tissue** but it requires **expensive** and **bulky equipment**, and **highly trained professionals**.
- **Digitalization** of pathology slides and **application of AI** can make diagnosis faster, cheaper, and provides a useful tool for phathologists.
- **Inferential analysis** can help statistically determine **differences between malignant and benignant populations**. It can also help determining **important features** for further AI modelling.



# 1. PRESENTATION OF THE PROBLEM

## 1.1 Source of the Database

UCI Machine Learning Repository

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.  
University of Wisconsin, Clinical Sciences Center  
Madison, WI 53792  
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.  
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.  
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706  
olvi '@' cs.wisc.edu

Donor:

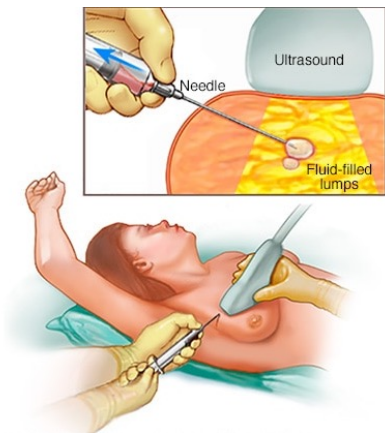
Nick Street

**Is there any difference between  
benignant and malignant  
populations for each feature?**

**In other words, which features  
could significantly help  
determining diagnosis of each  
sample?**

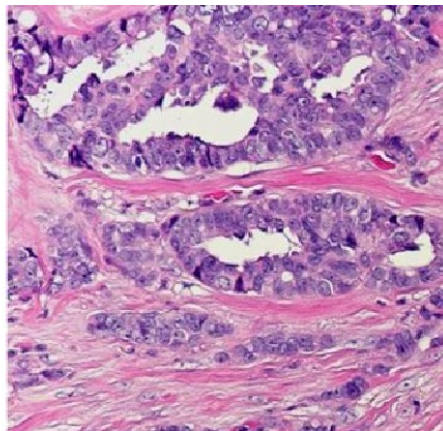
## 2 MATERIALS

### 2.2 Dataset Information: acquisition of the data



#### 1. Tissue sample from Breast Tumor by Fine-Needle Aspiration (FNA)

569 tissue samples



#### 2. Histopathological analysis of 10 characteristics for each nucleus:

- **Radius** (mean of distances from the center to points on the perimeter)
- **Texture** (std of gray-scale values)
- **Perimeter**
- **Area**
- **Smoothness** (local variation in radius lengths)
- **Compactness** ( $\text{perimeter}^2 / \text{area} - 1$ )
- **Concavity** (severity of concave portions of the contour)
- **Concave points** (number of concave portions of the contour)
- **Simmetry**
- **Fractal Dimension** ("coastline approximation"-1)

*\* No units were provided*

#### 3. Mean, Ste and Worst Values of all nuclei characteristics in each sample

30 features

## 2 MATERIALS

### 2.2 Data set Information: predictors and target variable

For each nucleus in each sample:

1. **Radius** (mean of distances from the center to points on the perimeter)
2. **Texture** (std of gry-scale values)
3. **Perimeter**
4. **Area**
5. **Smoothness** (local variation in radius lentghs)
6. **Compactness** ( $\text{perimeter}^2/\text{area} - 1$ )
7. **Concavity** (severity of concave portions of the contour)
8. **Concave points** (number of concave portions of the contour)
9. **Simmetry**
10. **Fractal Dimension** ("coastline approximation"-1)

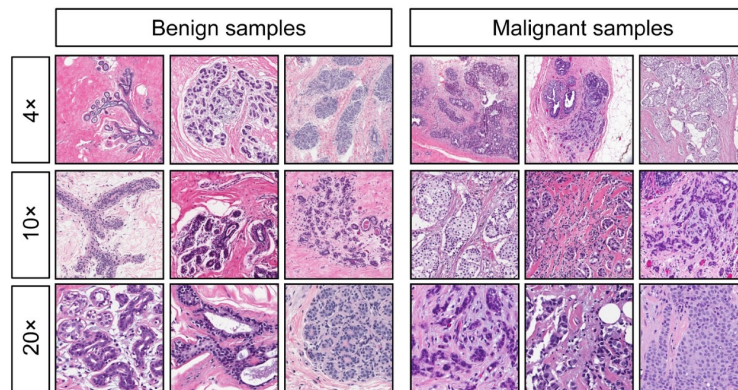
Mean, Ste and Worst Values of nuclei characteristics in each sample

30 features

+

Diagnosis:  
1. Benignant (B)  
2. Malignant (M)

1Target Variable



# 3. EXPLORATORY DATA ANALYSIS

## 3.1 Data Types

### Identifier



Drop Column

### Target Variable



Transform to numerical discrete

Numerical and Continuous

Predictors

Mean,  
Std and  
Worst

Radius	Area	Concavity	Fractal dimension
Texture	Smoothness	Concave points	
Perimeter	Compactness	Simmetry	



Drop Column

# 3. EXPLORATORY DATA ANALYSIS

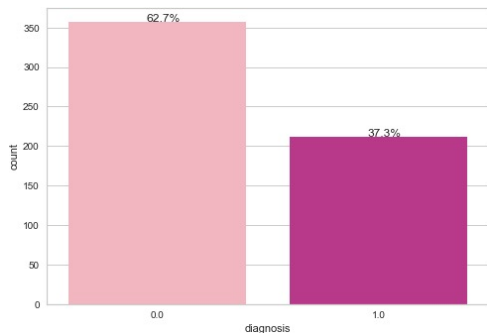
## 3.2 Descriptive statistics of the dataset

In order to summarize the main and most basic statistical characteristics of the dataset, we will use the method **describe**:

**No abnormal values for max/min values were initially identified(e.g 0 values or max/min values highly above/below the mean).**

By plotting count of samples for each class we find that:

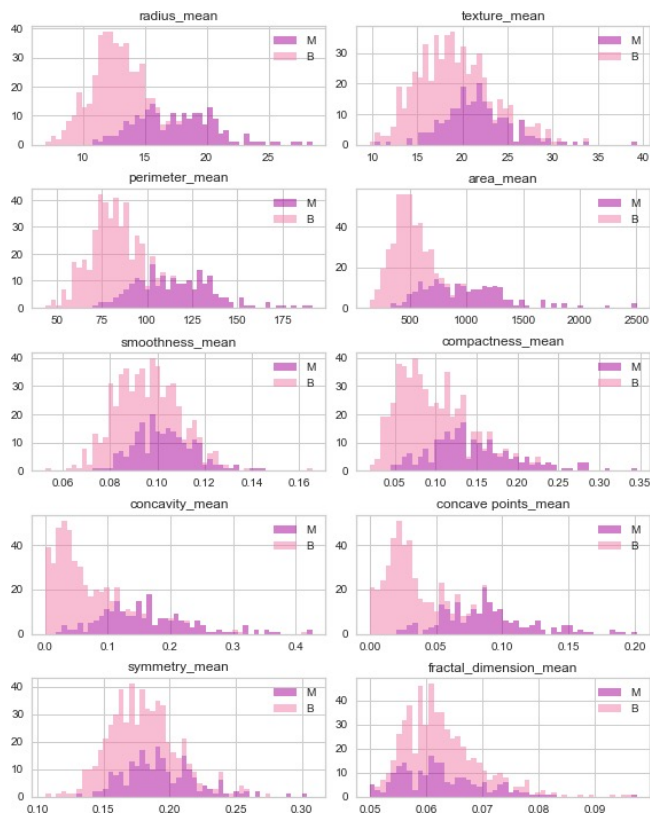
**Class imbalance (62.7 %(B, majority class)/ 37.3% (M, minority class): moderate**



# 3. EXPLORATORY DATA ANALYSIS

## 3.3 Univariate Analysis: Graphical

1. **Mean values** for radius, texture, perimeter, area, compactness, concavity and concave points **seem to be larger in malignant tissue.**
2. Features (distinguishing between malignant/benign) follow, approximately, a **normal distribution.**



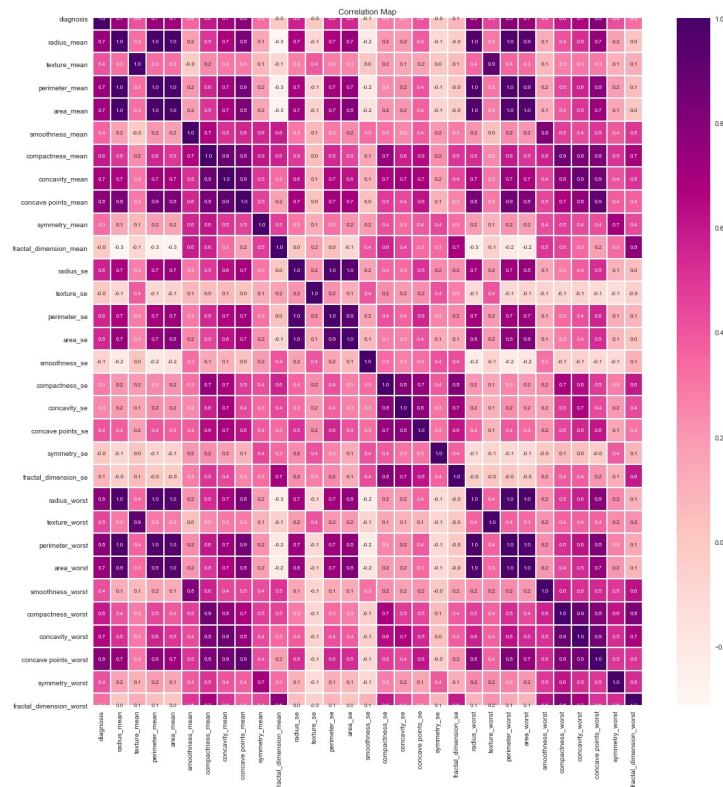


# 3. EXPLORATORY DATA ANALYSIS

## 3.4 Multivariate Analysis: Correlation

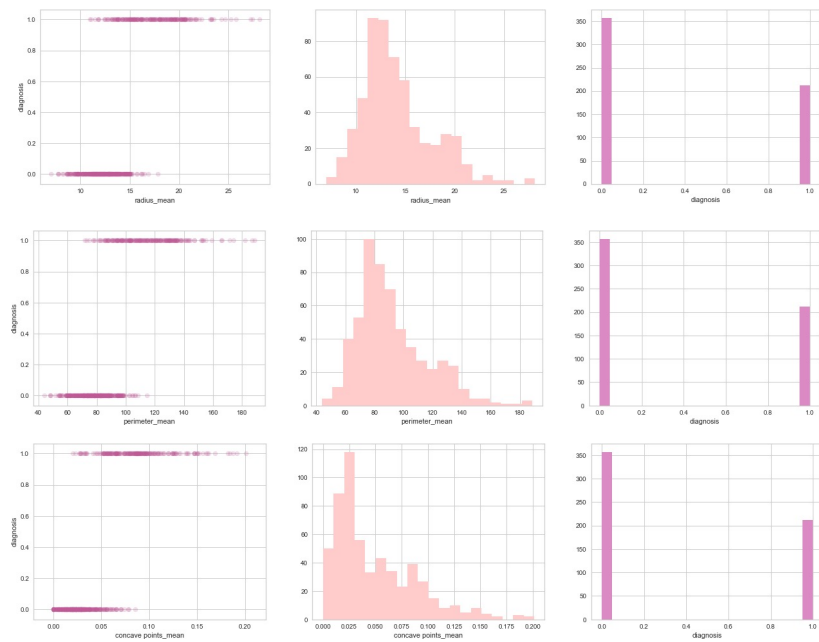
1. **Strong** positive relationship between target variable and mean and worst values for **radius**, **area**, **perimeter**, **concavity** and **concave points** (P. Correlation coefficient  $>0.7$ ).
2. **Strongest** relationship with **worst value for concave points**.
3. **Strong** correlation between radius, perimeter and area.
4. **Strong** correlation between concave points, concavity and compactness.

Correlation between features may imply redundant information during diagnosing.

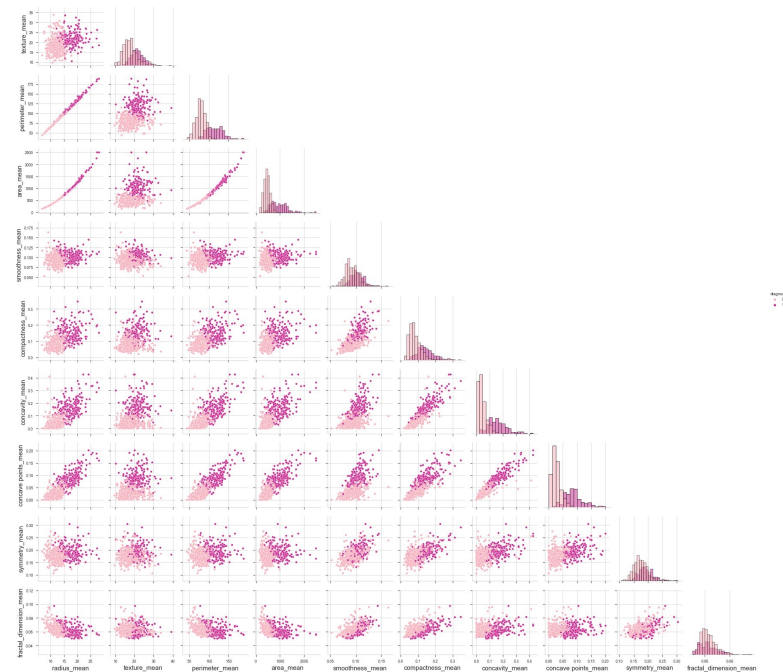


# 3. EXPLORATORY DATA ANALYSIS

## 3.5 Multivariate Analysis: Scatter Plots (Diagnosis vs each feature/ Feature vs Feature)



**Only 1 feature suitable for Logistic Regression**



**Only 2 features suitable for models such as KNN**

# 3. EXPLORATORY DATA ANALYSIS

## 3.6 Test for Normality

**Kolmogorov-Smirnov (KS) test** for normality ( $n > 50$ ) for each population (malignant/benignant), for each feature:

*H<sub>0</sub>: The sample data distribution is not significantly different than a normal population.*

*H<sub>1</sub>: The sample data distribution is significantly different than a normal population.*

In general, normally/non-normally distribution between malignant and benignant populations for one feature.

**Non-parametric methods should be used for hypothesis testing.**

	Normally distributed B	Normally distributed M
radius_mean	Yes	Yes
texture_mean	No	Yes
perimeter_mean	Yes	Yes
area_mean	Yes	Yes
smoothness_mean	Yes	Yes
compactness_mean	No	No
concavity_mean	No	No
concave points_mean	No	No
symmetry_mean	Yes	Yes
fractal_dimension_mean	No	No
radius_se	No	No
texture_se	No	No
perimeter_se	No	No
area_se	No	No

	Normally distributed B	Normally distributed M
smoothness_se	No	No
compactness_se	No	No
concavity_se	No	No
concave points_se	No	Yes
symmetry_se	No	No
fractal_dimension_se	No	No
radius_worst	Yes	Yes
texture_worst	Yes	Yes
perimeter_worst	Yes	Yes
area_worst	Yes	Yes
smoothness_worst	Yes	Yes
compactness_worst	No	No
concavity_worst	No	No
concave points_worst	Yes	Yes
symmetry_worst	Yes	No
fractal_dimension_worst	No	Yes

## 4. PRE- PROCESSING OF THE DATASET

### 4.1 Missing Values

1. **No missing values**, neither explicit (NaN values) nor implicit (e.g. repeated 0 values for an instance for different features).
2. Samples with 0 values show the same behavior, all associated with diagnosis=B and same zero features, which may imply that these values are indeed correct.

	diagnosis	concavity_mean	concave points_mean
101	0.0	0.0	0.0
140	0.0	0.0	0.0
174	0.0	0.0	0.0
175	0.0	0.0	0.0
192	0.0	0.0	0.0

Data columns (total 31 columns):			
Data #	Column	Non-Null Count	Dtype
0	diagnosis	569 non-null	float64
1	radius_mean	569 non-null	float64
2	texture_mean	569 non-null	float64
3	perimeter_mean	569 non-null	float64
4	area_mean	569 non-null	float64
5	smoothness_mean	569 non-null	float64
6	compactness_mean	569 non-null	float64
7	concavity_mean	569 non-null	float64
8	concave points_mean	569 non-null	float64
9	symmetry_mean	569 non-null	float64
10	fractal_dimension_mean	569 non-null	float64
11	radius_se	569 non-null	float64
12	texture_se	569 non-null	float64
13	perimeter_se	569 non-null	float64
14	area_se	569 non-null	float64
15	smoothness_se	569 non-null	float64
16	compactness_se	569 non-null	float64
17	concavity_se	569 non-null	float64
18	concave points_se	569 non-null	float64
19	symmetry_se	569 non-null	float64
20	fractal_dimension_se	569 non-null	float64
21	radius_worst	569 non-null	float64
22	texture_worst	569 non-null	float64
23	perimeter_worst	569 non-null	float64
24	area_worst	569 non-null	float64
25	smoothness_worst	569 non-null	float64
26	compactness_worst	569 non-null	float64
27	concavity_worst	569 non-null	float64
28	concave points_worst	569 non-null	float64
29	symmetry_worst	569 non-null	float64
30	fractal_dimension_worst	569 non-null	float64

## 4. PRE- PROCESSING OF THE DATASET

### 4.2 Outliers

1. Outlier detection **independently for Benignant/Malignant samples**, as they showed different distributions.
2. For detection, **considered as outliers those values with  $\text{abs}(z\text{-score}) \geq 2.5$ .**
3. **Very low percentage of outliers per feature (0.3-2%).**
4. 20.5% (117 instances) of rows with, at least, one outlier. **We shouldn't consider dropping this quantity of data.**
5. Since variables are highly correlated, random/mean/median imputation methods can introduce bias in the analysis  
→ **Tailored Imputation Method** → Computationally expensive
6. ***From previous work with AI methods: Performance metrics after dropping outliers are not improved. They are similar → Outliers may not be deleted incorrect values, just values far from the population.***
7. We decide to maintain these values.

Original dataset

	Train Accuracy	Validation Accuracy	Test Accuracy
<b>Logistic Regression</b>	0.985915	0.971847	0.979021
<b>KNN Classifier</b>	0.974178	0.967141	0.965035
<b>Decision Tree Classifier</b>	1.000000	0.925007	0.923077
<b>Neural Network Classifier</b>	0.995305	0.978906	0.986014
<b>Random Forest Classifier</b>	1.000000	0.962435	0.944056

Dataset without outliers (deletion)

	Train Accuracy	Validation Accuracy	Test Accuracy
<b>Logistic Regression</b>	0.988201	0.973442	0.982301
<b>KNN Classifier</b>	0.979351	0.967559	0.973451
<b>Decision Tree Classifier</b>	1.000000	0.967603	0.893805
<b>Neural Network Classifier</b>	0.994100	0.976383	0.973451
<b>Random Forest Classifier</b>	1.000000	0.967515	0.973451

## 5. HYPOTHESIS TESTING

**Two populations for each feature: malignant and benignant.**

Parametric:

*H<sub>0</sub>: Mean values for malignant and benignant populations are the same.*

*H<sub>1</sub>: Mean values for malignant are different than those for benignant.*

Non-parametric:

*H<sub>0</sub>: Median for malignant and benignant populations are the same.*

*H<sub>1</sub>: Median for malignant are different than those for benignant.*

**Two independent samples, two-sided hypothesis testing problem.**

**Non-parametric methods preferred.**

Both parametric and non-parametric ( Mann-Whitney-Wilcoxon) methods will be compared for learning purposes.

## 5. HYPOTHESIS TESTING

*E.g. : Non-parametric vs Parametric (just some features)*

	P-value	H_0	H_1
radius_mean	1.346471e-68	reject	accept
texture_mean	1.714313e-28	reject	accept
perimeter_mean	1.776935e-71	reject	accept
area_mean	7.698902e-69	reject	accept
smoothness_mean	3.896503e-19	reject	accept
compactness_mean	4.475996e-48	reject	accept
concavity_mean	1.082274e-68	reject	accept
concave points_mean	5.031619e-77	reject	accept
symmetry_mean	1.134025e-15	reject	accept
fractal_dimension_mean	2.685928e-01	accept	reject
radius_se	3.108570e-49	reject	accept
texture_se	3.218464e-01	accept	reject
perimeter_se	2.549719e-51	reject	accept
area_se	2.883912e-65	reject	accept
smoothness_se	1.068158e-01	accept	reject
compactness_se	5.840307e-20	reject	accept

	Statistic	P-value	H_0	H_1
radius_mean	22.208798	1.684459e-64	reject	accept
texture_mean	11.022087	3.019055e-25	reject	accept
perimeter_mean	22.935314	1.023141e-66	reject	accept
area_mean	19.640990	3.284366e-52	reject	accept
smoothness_mean	9.297355	5.573331e-19	reject	accept
compactness_mean	15.818246	9.607863e-42	reject	accept
concavity_mean	20.332425	3.742121e-58	reject	accept
concave points_mean	24.844810	3.127316e-71	reject	accept
symmetry_mean	8.112198	5.957651e-15	reject	accept
fractal_dimension_mean	-0.296866	7.667216e-01	accept	reject
radius_se	13.300706	1.491133e-30	reject	accept
texture_se	-0.207865	8.354171e-01	accept	reject
perimeter_se	12.832763	6.868553e-29	reject	accept
area_se	12.155556	2.983568e-26	reject	accept
smoothness_se	-1.622869	1.052970e-01	accept	reject
compactness_se	7.082641	6.341807e-12	reject	accept

Taking into account results from non-parametric method, mean and median values are different between benignant and benignant samples for every feature but for "fractal\_dimension\_mean", "texture\_se", and "smoothness\_se".

## 6. CONCLUSION

Dataset **with excellent quality** for its purpose:

1. Simple data visualization allows to get a great insight into data distribution.

We saw behavior of malignant nuclei just by scatter plotting B vs M for each feature.

**Malignant samples tend to have greater values than benignant. It was statistically determined that malignant mean and median values were different from those of benignant samples.**

**Statistically concluded that nuclei characteristics vary depending on malignancy/benignancy of the sample.** Along with AI models, inferential analysis can help determining decision boundaries and translate them to clinical practice.

2. Exhaustive and complicated pre-processing is not needed to perform inferential analysis and draw relevant conclusions.
3. Results show that some features are more relevant than others when determining the diagnosis, which is relevant for feature selection → discard "fractal\_dimension\_mean", "texture\_se", and "smoothness\_se".
4. Feature selection can improve the efficiency of the histopathological analysis by discarding non-relevant features.