**CIMVHR**
Canadian Institute for Military and Veteran Health Research

**ICRSMV**
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Machine Learning for Health Research

Sunday October 20th, 2019

9:00 am – 12:00 pm

Krieghoff room

# DEMOS SESSION

Presented by :
Dr. Mohamed Sami Rakha

CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Assumption…

At this point we assume that you get your environment setup successfully. This means you have a Jupyter NoteBook instance running.


CIMVHR | ICRSMV

# So First…  What we need?

CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# So First… What we need?

Dataset

CIMVHR | ICRSMV

# We have two Datasets

## Diabetes Dataset

442 Patients

## Breast Cancer Dataset

569 Patients

CIMVHR | ICRSMV

# We have two Datasets

**Diabetes Dataset**

442 Patients

**Breast Cancer Dataset**

569 Patients

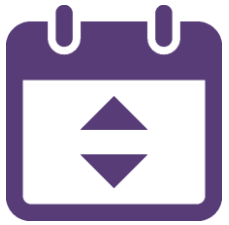
CIMVHR | ICRSMV

# 1# Diabetes Dataset

**Number of Patients**

442 Patients

10 Features



Age

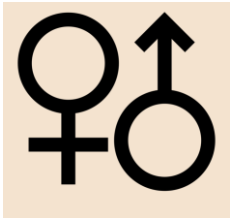
# 1# Diabetes Dataset

**Number of Patients**

442 Patients

10 Features

Age



Gender

CIMVHR | ICRSMV

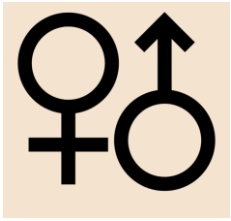
# 1# Diabetes Dataset

**Number of Patients**

442 Patients

10 Features

Age



Gender

BMI

CIMVHR | ICRSMV

# 1# Diabetes Dataset

**Number of Patients**

442 Patients

10 Features

Age

Gender

BMI

Blood Pressure

CIMVHR | ICRSMV

# 1# Diabetes Dataset

**Number of Patients**

442 Patients

10 Features

Age

Gender

BMI

Blood Pressure

S1 S2 S3 S4 S5 S6

Six Blood Serum Measurements

CIMVHR | ICRSMV

# 1# Diabetes Dataset

**Number of Patients**

442 Patients

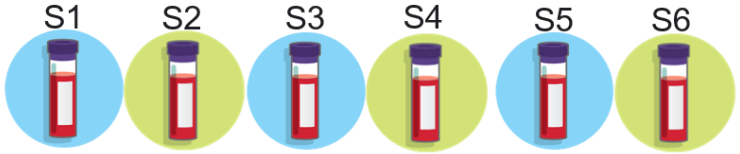
10 Features

Age

Gender

BMI

Blood Pressure




Six Blood Serum Measurements

11 Columns

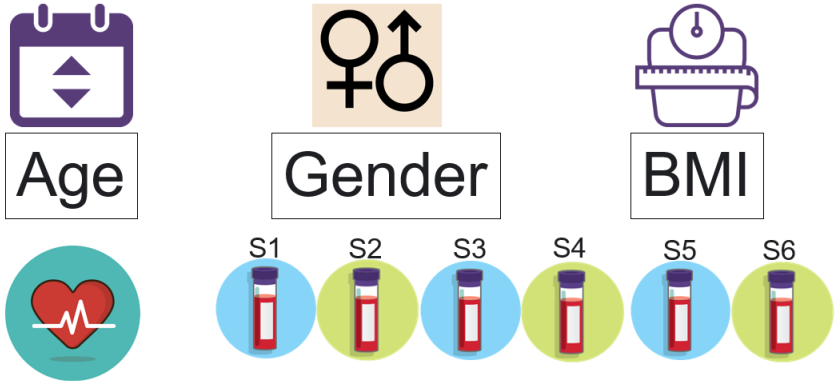
Outcome

Quantitative Measure of the Disease

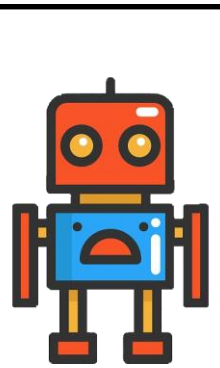
CIMVHR | ICRSMV

# 1# Diabetes Dataset

10 Features




Training



Machine Learning Model

Outcome

Regression Model

CIMVHR | ICRSMV

# We have two Datasets

## Diabetes Dataset

442 Patients

## Breast Cancer Dataset

569 Patients

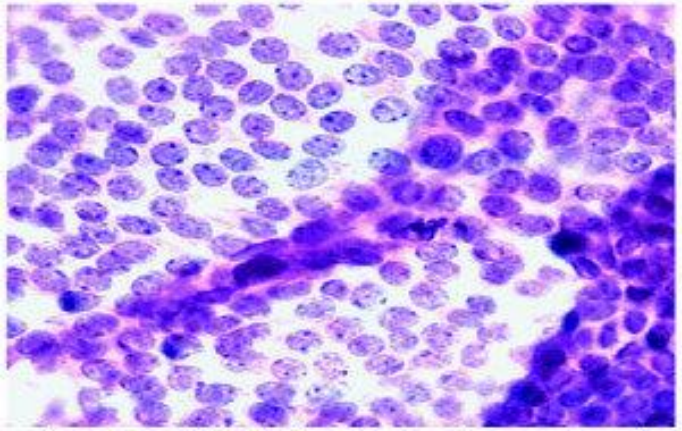
CIMVHR | ICRSMV

# 2# Breast Cancer Dataset

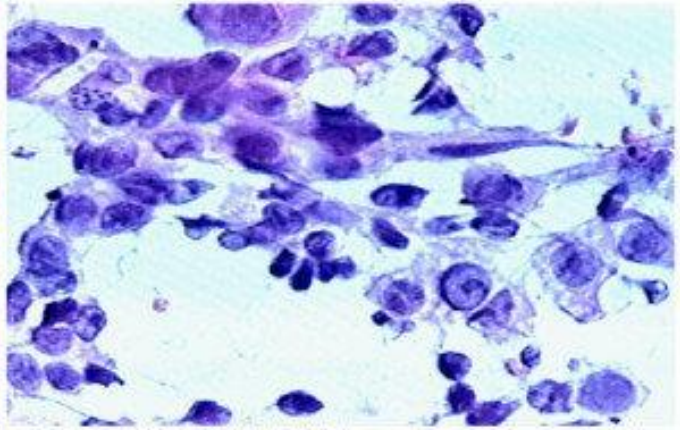
## Number of Patients

569 Patients

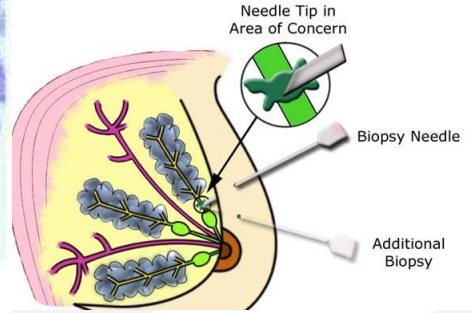
## 30 Features



Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size



Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size



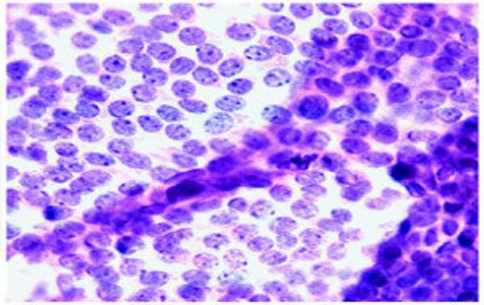

| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

# 2# Breast Cancer Dataset

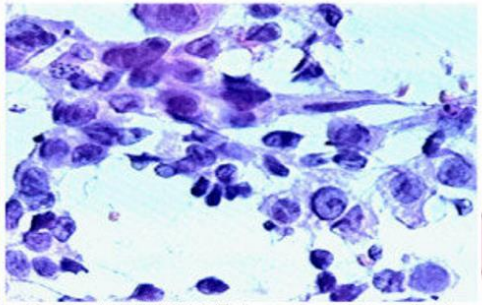
## Number of Patients

569 Patients

## 30 Features



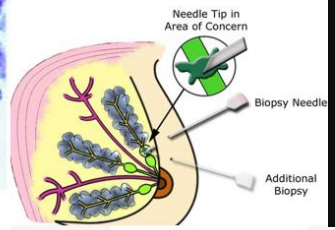
Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size



Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size

| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

## Outcome




**2 Classes:**

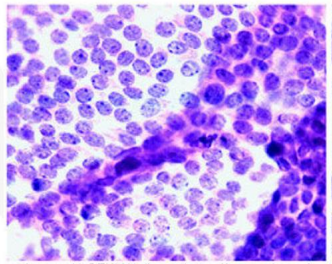
- Malignant
- Benign

## 31 Columns

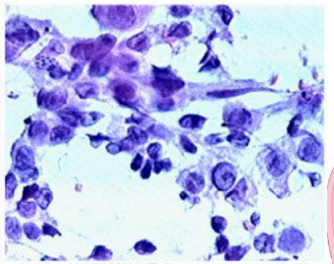
CIMVHR | ICRSMV

# 2# Breast Cancer Dataset

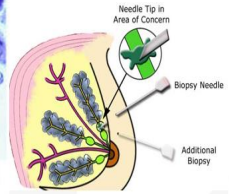
29 Features



Smear with BENIGN diagnosis – uniform nucleus of cells, symmetrical, homogeneous, with areas within normal size

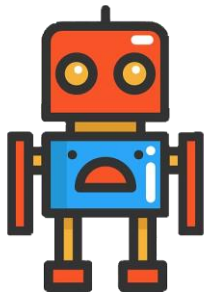


Smear with MALIGNANT diagnosis – nucleus of cells without uniformity, asymmetrical, not homogeneous (multiple sizes) and with areas above normal size




| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

Training



**Machine Learning Model**

Outcome

**Classification Model**

CIMVHR | ICRSMV

# Supervised Learning Demos

**Regression Demo**

**Classification Demo**

# Supervised Learning Demos

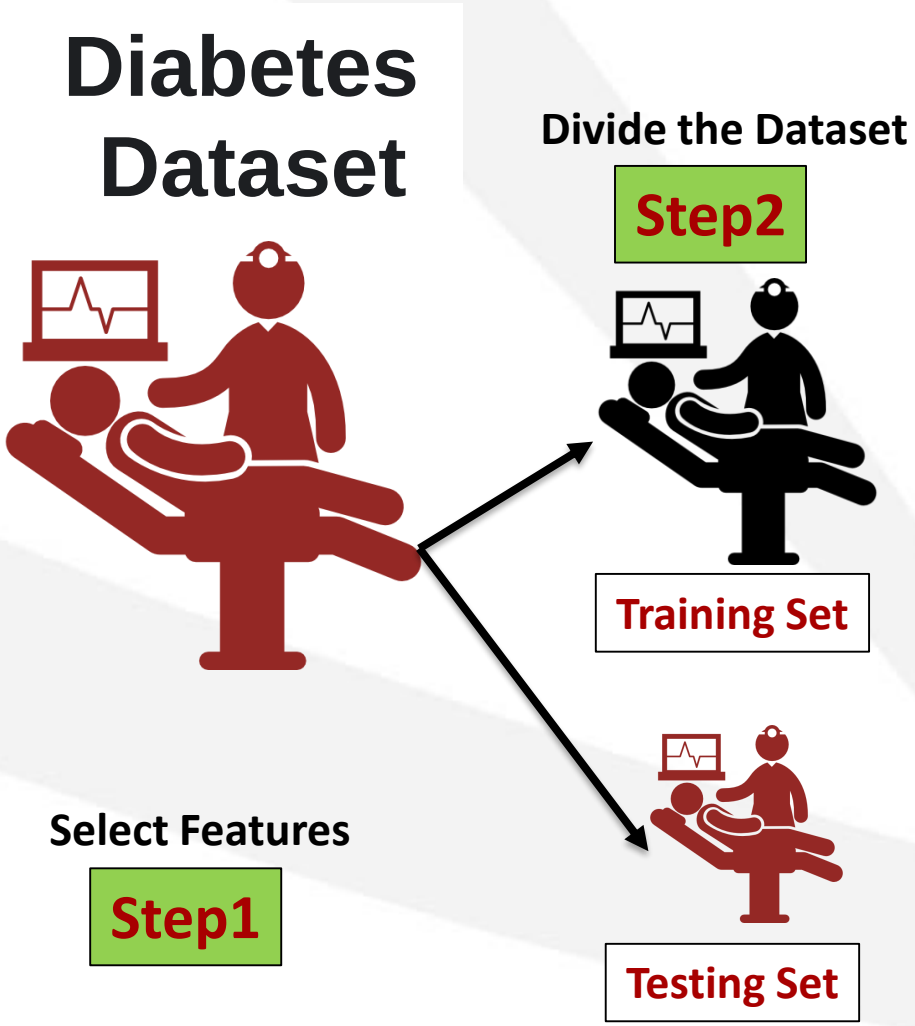
**Regression Demo**

**Classification Demo**

# DEMO#1

## Regression Modeling

# Demo#4: Regression Modeling




CIMVHR | ICRSMV

# Demo#4: Regression Modeling

**Diabetes Dataset**

**Select Features**

**Step1**

**Divide the Dataset**

**Step2**

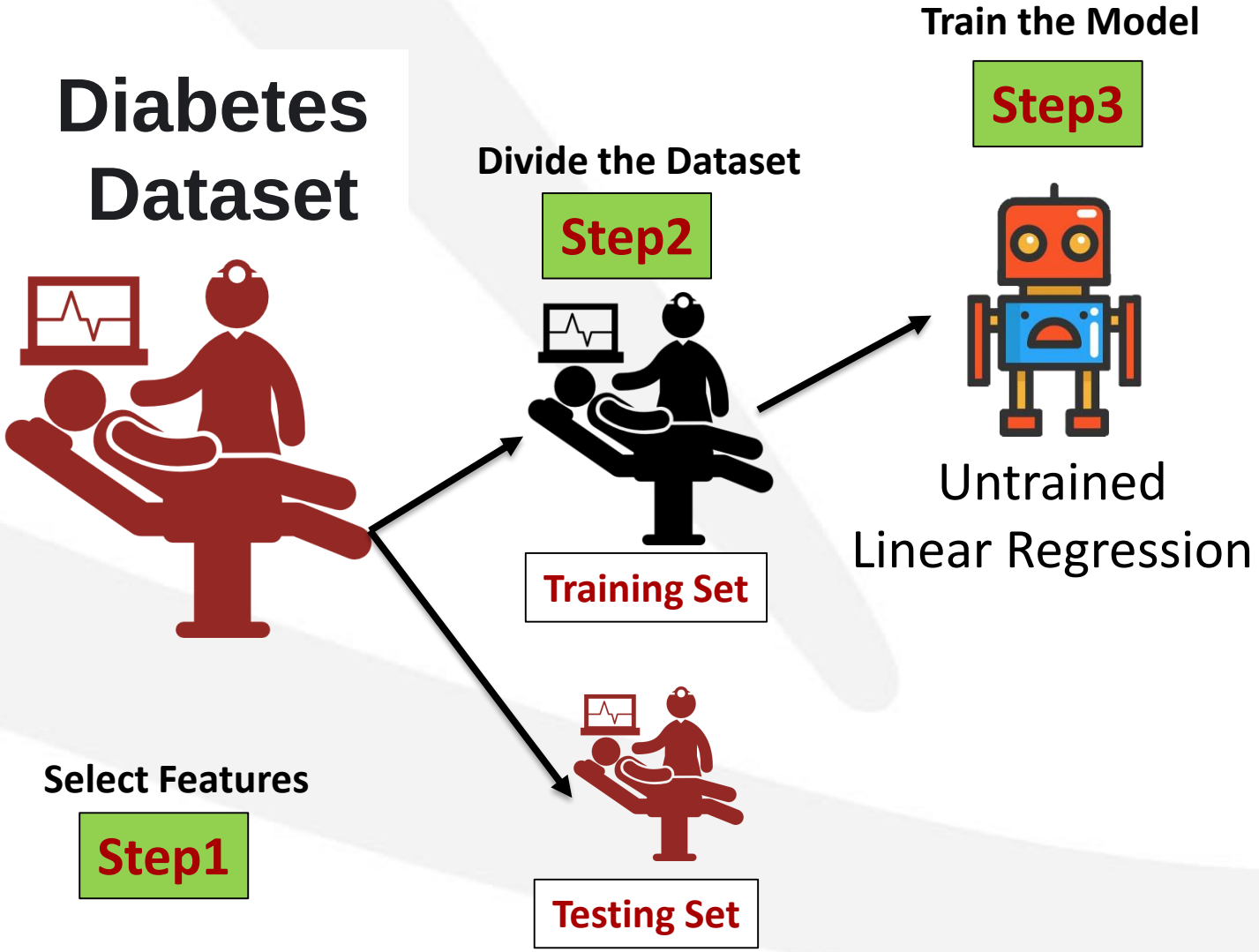
**Training Set**

**Testing Set**

The features and the Target are continuous values. We want to model a linear relation between the features and the target

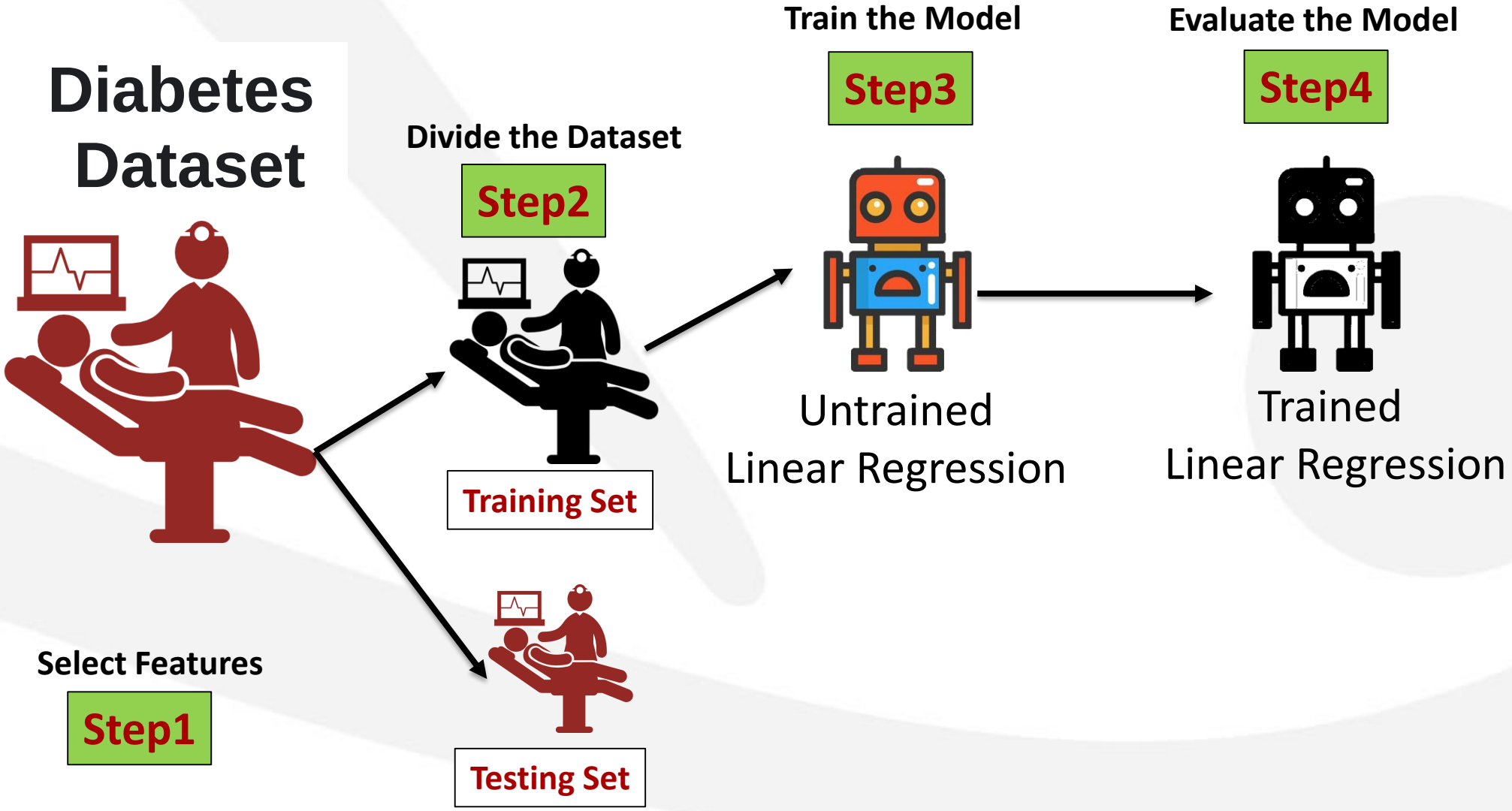
CIMVHR | ICRSMV

# Demo#4: Regression Modeling



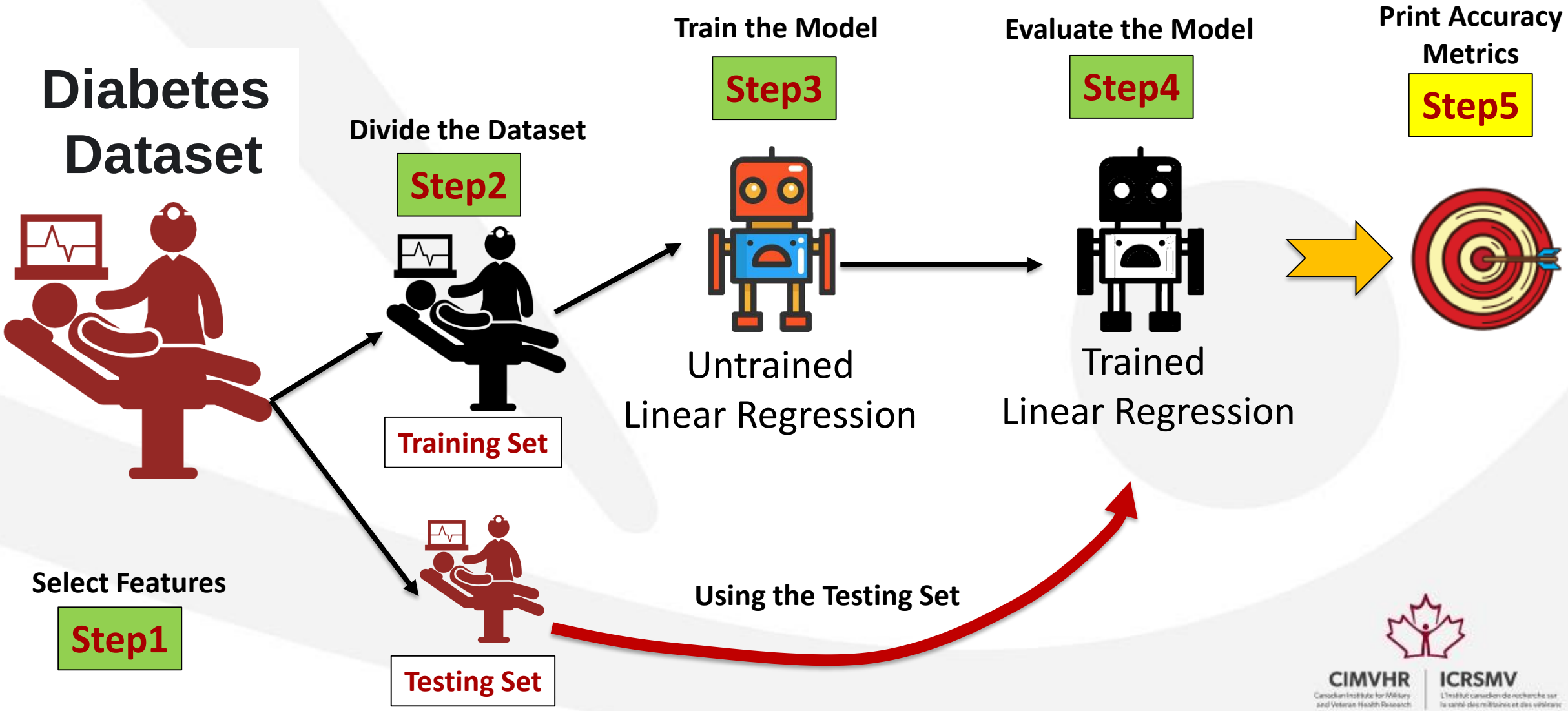

CIMVHR | ICRSMV

# Demo#4: Regression Modeling




CIMVHR | ICRSMV

# Demo#4: Regression Modeling




CIMVHR | ICRSMV

# Demo#1: Linear Regression


**Lets Do Some Coding!**

# Supervised Learning Demos

Regression Demo

Classification Demo

# DEMO#2

**First Classification**

CIMVHR | ICRSMV

# Demo#2: Classification

## Breast Cancer Dataset

**Select Features**

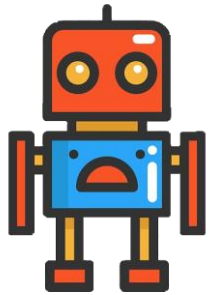
**Step1**

| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

CIMVHR | ICRSMV

# Demo#2: Classification

**Breast Cancer Dataset**

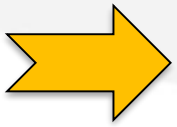


Untrained
Naïve Bayes

Our goal is train this model, so it predict future the class of future data

**Select Features**

**Step1**



| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

CIMVHR | ICRSMV

# Demo#2: Classification

**Breast Cancer Dataset**

**Select Features**

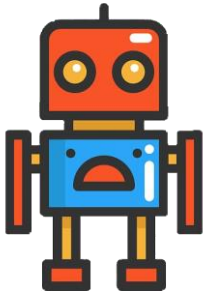
**Step1**

**Divide the Dataset**

**Step2**

**Training Set**

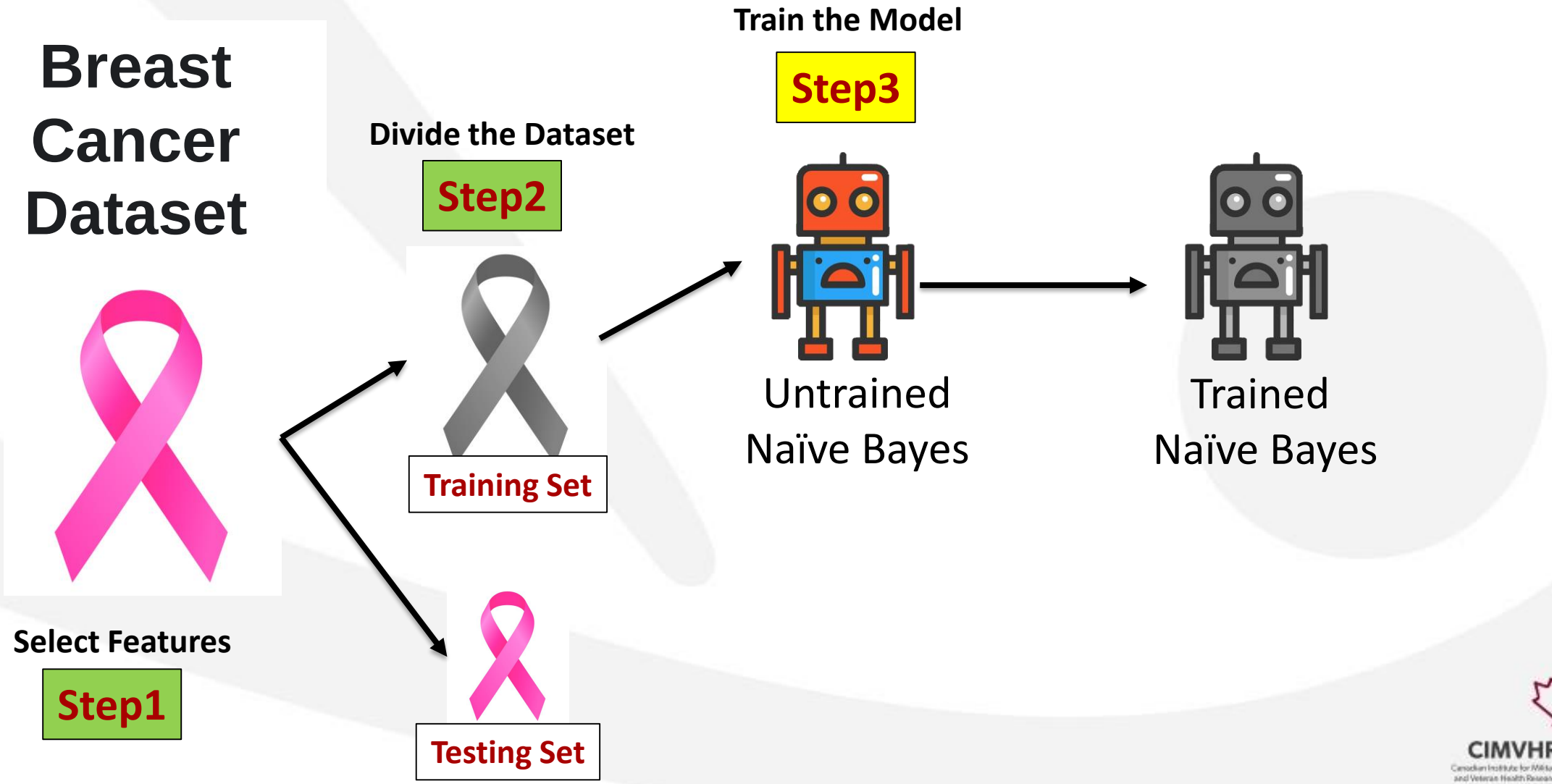
**Testing Set**



Untrained
Naïve Bayes

CIMVHR | ICRSMV

# Demo#2: Classification




CIMVHR | ICRSMV

# Demo#2: Classification

Breast Cancer Dataset

Select Features

Step1

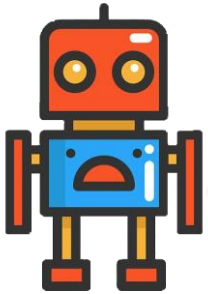
Divide the Dataset

Step2

Training Set

Train the Model

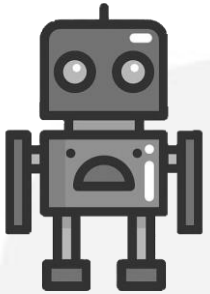
Step3



Untrained Naïve Bayes

Evaluate the Model

Step4



Trained Naïve Bayes

Testing Set

Using the Testing Set

CIMVHR | ICRSMV

# Demo#2: Classification

Breast Cancer Dataset

Select Features

Step1

Divide the Dataset

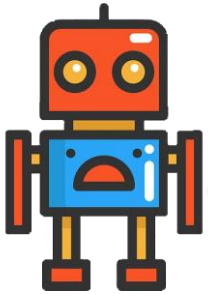
Step2

Training Set

Testing Set

Train the Model

Step3

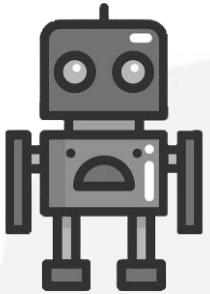


Untrained Naïve Bayes

Using the Testing Set

Evaluate the Model

Step4



Trained Naïve Bayes

Print Accuracy Metrics

Step5

CIMVHR | ICRSMV

# Demo#2: Classification


**Lets Do Some Coding!**

# DEMO#3

**Trying another Classifer**

CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#3: Trying Another Classifier

Breast Cancer Dataset

Select Features

Step1

Divide the Dataset

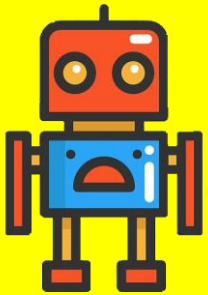
Step2

Training Set

Testing Set

Train the Model

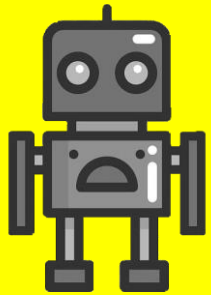
Step3



Untrained Naïve Bayes

Evaluate the Model

Step4

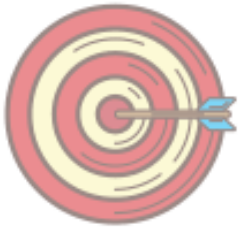


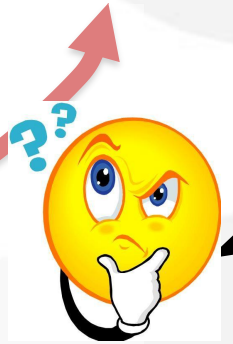
Trained Naïve Bayes

Using the Testing Set

Print Accuracy Metrics

Step5





Lets try another Learning Algorithm…. **Random Forest?**

CIMVHR Canadian Institute for Military and Veteran Health Research

ICRSMV L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#3: Trying Another Classifier

Breast Cancer Dataset

Select Features

Step1

Divide the Dataset

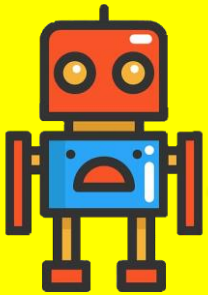
Step2

Training Set

Testing Set

Train the Model

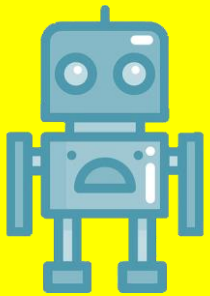
Step3



Untrained **RandomForest**

Evaluate the Model

Step4

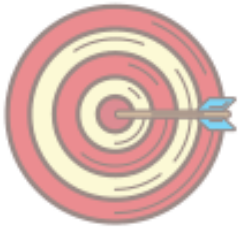


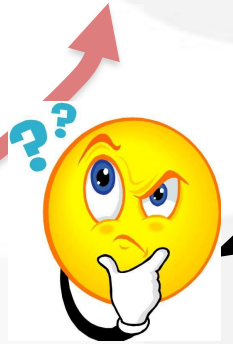
Trained **RandomForest**

Using the Testing Set

Print Accuracy Metrics

Step5

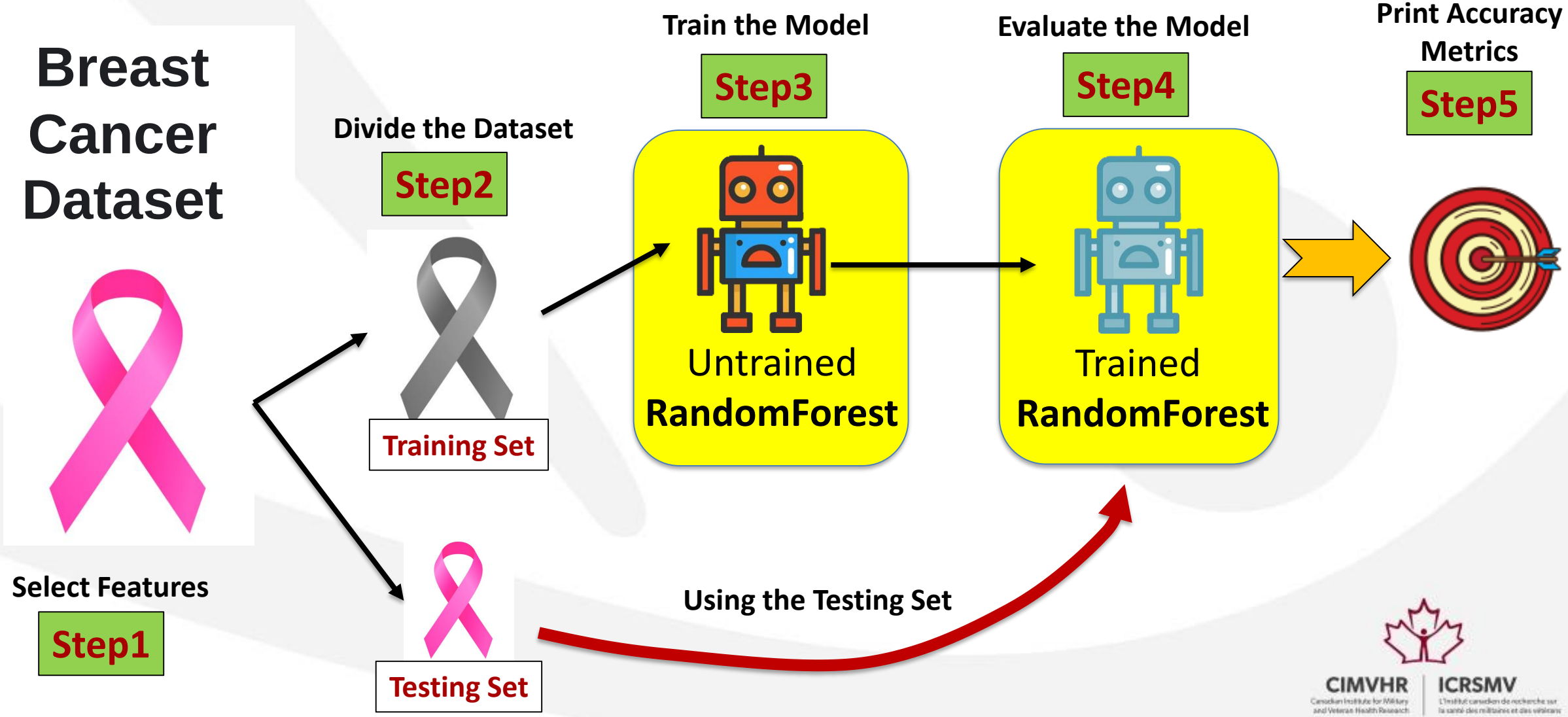




Lets try another Learning Algorithm…. **Random Forest?**

CIMVHR | ICRSMV

# Demo#3: Trying Another Classifier




CIMVHR | ICRSMV

# Demo#3: Trying Another Classifier

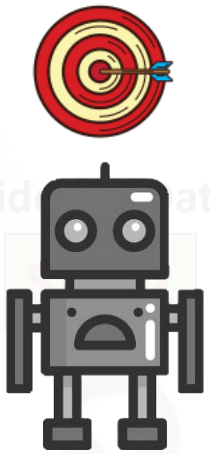

**Lets Do Some Coding!**

# Demo#3: Trying Another Classifier

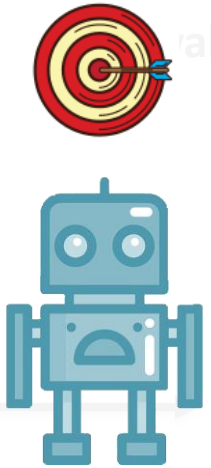

**Naïve Bayes** **RandomForest**

Based on the Accuracy Metrics.. Which Learning Algorithm is performing better?

# Demo#3: Which one is better?



Naïve Bayes



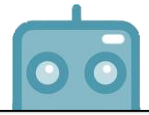
RandomForest

Based on the Accuracy Metrics.. Which Learning Algorithm is performing better?

What about trying Model Validation Techniques?Why?

# Demo#3: Which one is better?

| | | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| malignant | 0 | 0.93 | 0.76 | 0.83 | 98 |
| benign | 1 | 0.88 | 0.97 | 0.92 | 187 |
| accuracy | | | | 0.89 | 285 |
| macro avg | | 0.90 | 0.86 | 0.88 | 285 |
| weighted avg | | 0.90 | 0.89 | 0.89 | 285 |



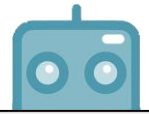
| | | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| malignant | 0 | 0.89 | 0.82 | 0.85 | 98 |
| benign | 1 | 0.91 | 0.95 | 0.93 | 187 |
| accuracy | | | | 0.90 | 285 |
| macro avg | | 0.90 | 0.88 | 0.89 | 285 |
| weighted avg | | 0.90 | 0.90 | 0.90 | 285 |

Based on the Accuracy Metrics.. Which Learning Algorithm is performing better?

What about trying Model Validation Techniques?Why?

# Demo#3: Which one is better?

| | | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| malignant | 0 | 0.93 | 0.76 | 0.83 | 98 |
| benign | 1 | 0.88 | 0.97 | 0.92 | 187 |
| accuracy | | | | 0.89 | 285 |
| macro avg | | 0.90 | 0.86 | 0.88 | 285 |
| weighted avg | | 0.90 | 0.89 | 0.89 | 285 |

| | | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| malignant | 0 | 0.89 | 0.82 | 0.85 | 98 |
| benign | 1 | 0.91 | 0.95 | 0.93 | 187 |
| accuracy | | | | 0.90 | 285 |
| macro avg | | 0.90 | 0.88 | 0.89 | 285 |
| weighted avg | | 0.90 | 0.90 | 0.90 | 285 |

What else RandomForest does? Descriptive analysis

# DEMO#4

## Exploring Model Validation

# Demo#4: Model Validation

Breast Cancer Dataset

Select Features

Step1

Divide the Dataset

Step2

Training Set

Testing Set

We change how data is splitted? And use K-fold validation

# Demo#4: Model Validation

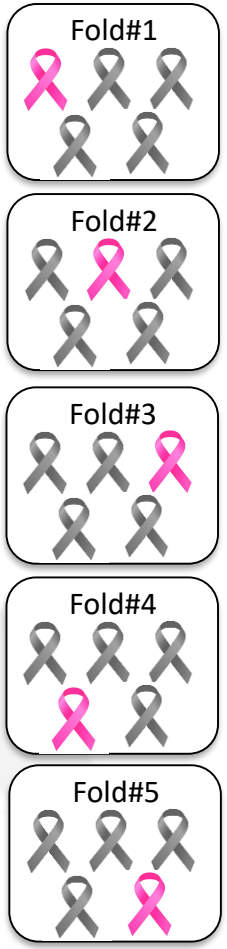
**Breast Cancer Dataset**

**Select Features**

**Step1**

**Step2**

**5 folds cross-validation**




Reduces the chances of overfitting.

CIMVHR | ICRSMV

# Demo#4: Model Validation

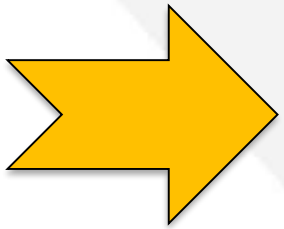
**Breast Cancer Dataset**

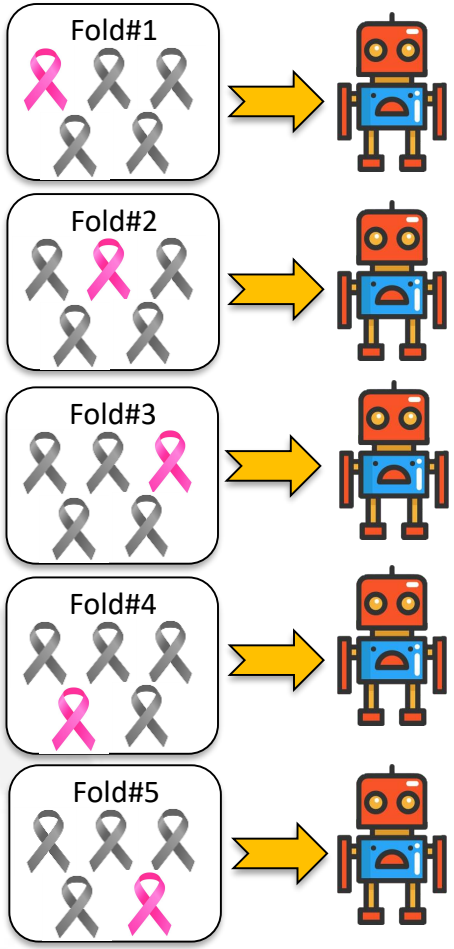
**Select Features**

**Step1**

**Step2**

**5 folds cross-validation**






Reduces the chances of overfitting.

CIMVHR | ICRSMV

# Demo#4: Model Validation

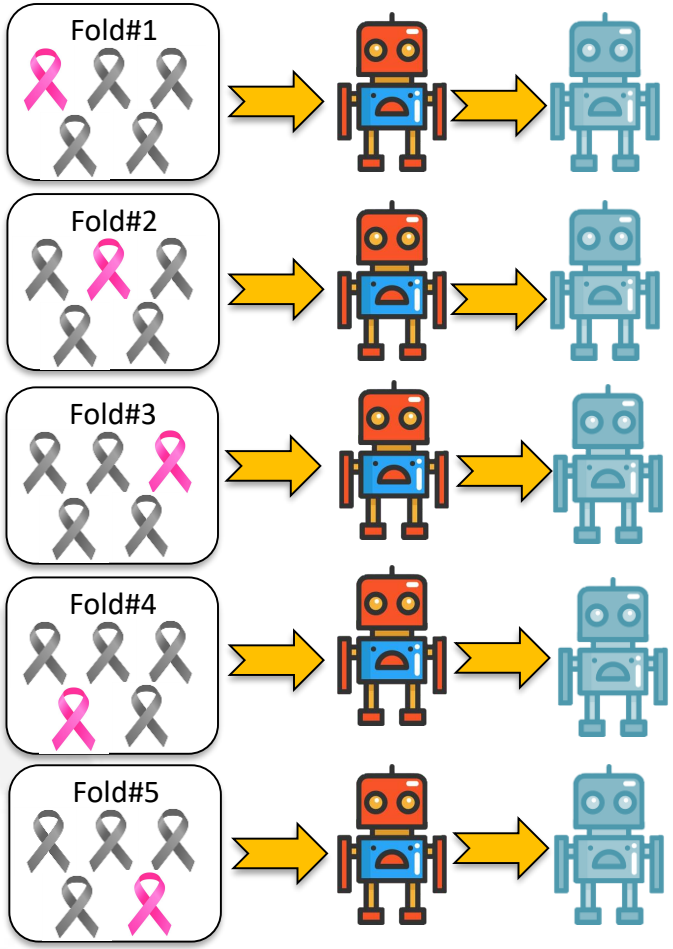
**Breast Cancer Dataset**

**Select Features**

**Step1**

**Step2**

**5 folds cross-validation**




Reduces the chances of overfitting.

CIMVHR | ICRSMV

# Demo#4: Model Validation

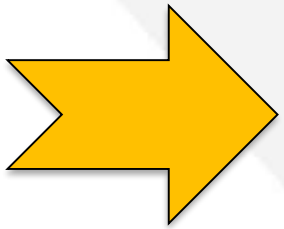
**Breast Cancer Dataset**

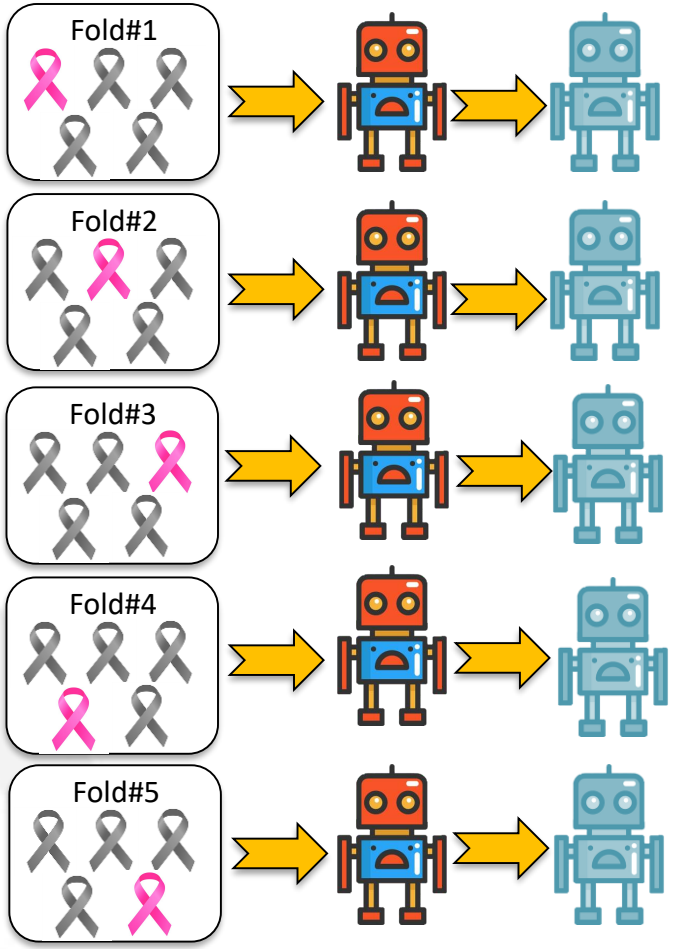
**Select Features**

**Step1**

**Step2**

**5 folds cross-validation**






**Step3**

**Report Average Accuracy Metrics**

Reduces the chances of overfitting.

CIMVHR | ICRSMV

# Demo#4: Model Validation

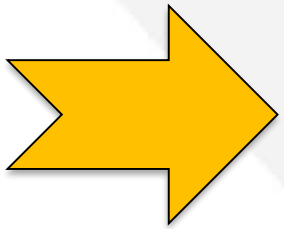
**Breast Cancer Dataset**

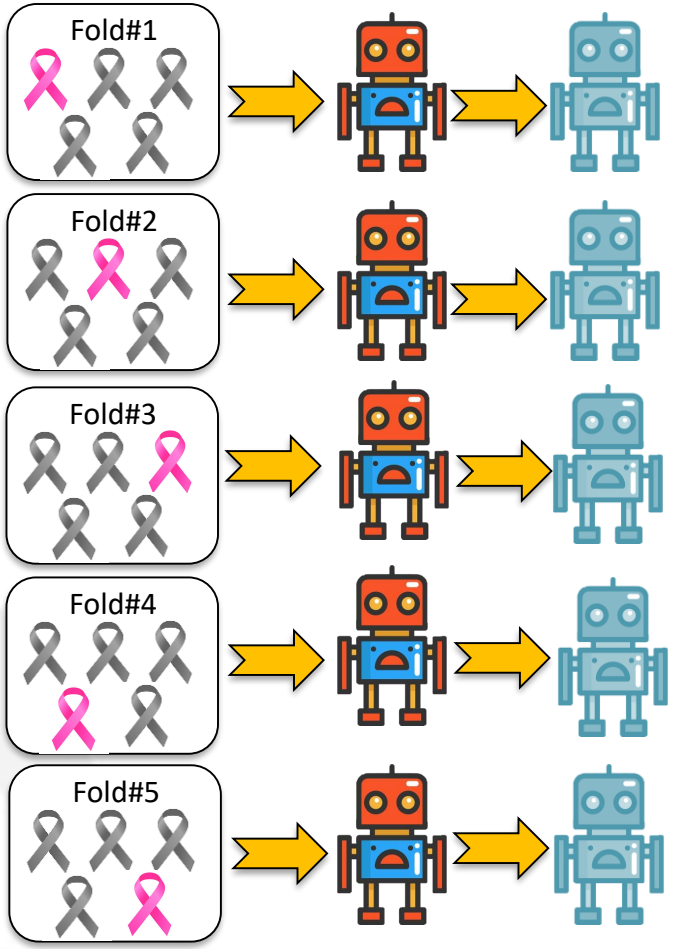
**Select Features**

**Step1**

**Step2**

**5 folds cross-validation**



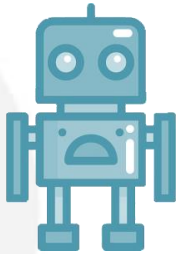



**Step3**

**Report Average Accuracy Metrics**

**Step4**

**Pick best one out of the five**



Reduces the chances of overfitting.

CIMVHR | ICRSMV

# Demo#4: Model Validation


**Lets Do Some Coding!**

# UnSupervised Learning Demos

## Clustering

CIMVHR | ICRSMV

# DEMO#5

## Clustering with Kmeans

# Demo#5: Clustering with K-means

**Breast Cancer Dataset**

**Select Features**

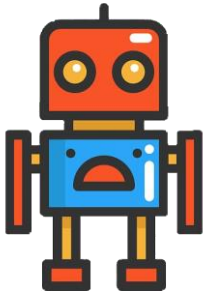
**Step1**

**Step2**

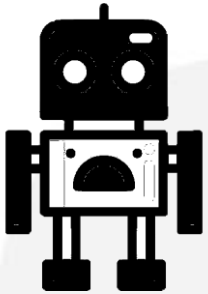
**Send data to Kmeans after ignoring the labels**

**Start Clustering**

**Step3**



Before Clustering Kmeans



Clustered Data Kmeans

**Print Accuracy Metrics**

**Step4**

**Compare the labelled data and Clustered data**

CIMVHR | ICRSMV

# Demo#5: Clustering with K-means


**Lets Do Some Coding!**

# Selected Topics

**Parameter Optimization**

**Automatic Feature Selection**

# Selected Topics

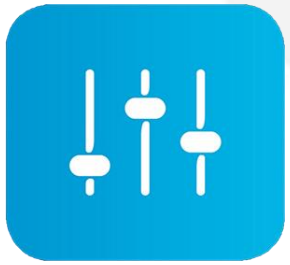
**Parameter Optimization**

**Automatic Feature Selection**

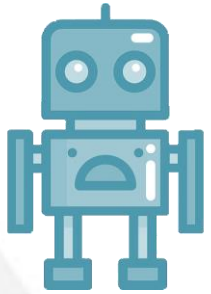
# DEMO#6

## Parameter Optimization

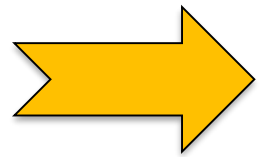
# Demo#6: Parameter Optimization



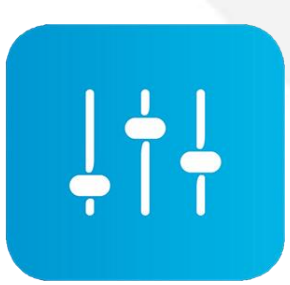
Parameter Optimization



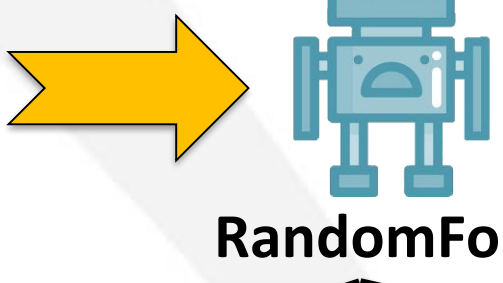
**RandomForest**



# Demo#6: Parameter Optimization



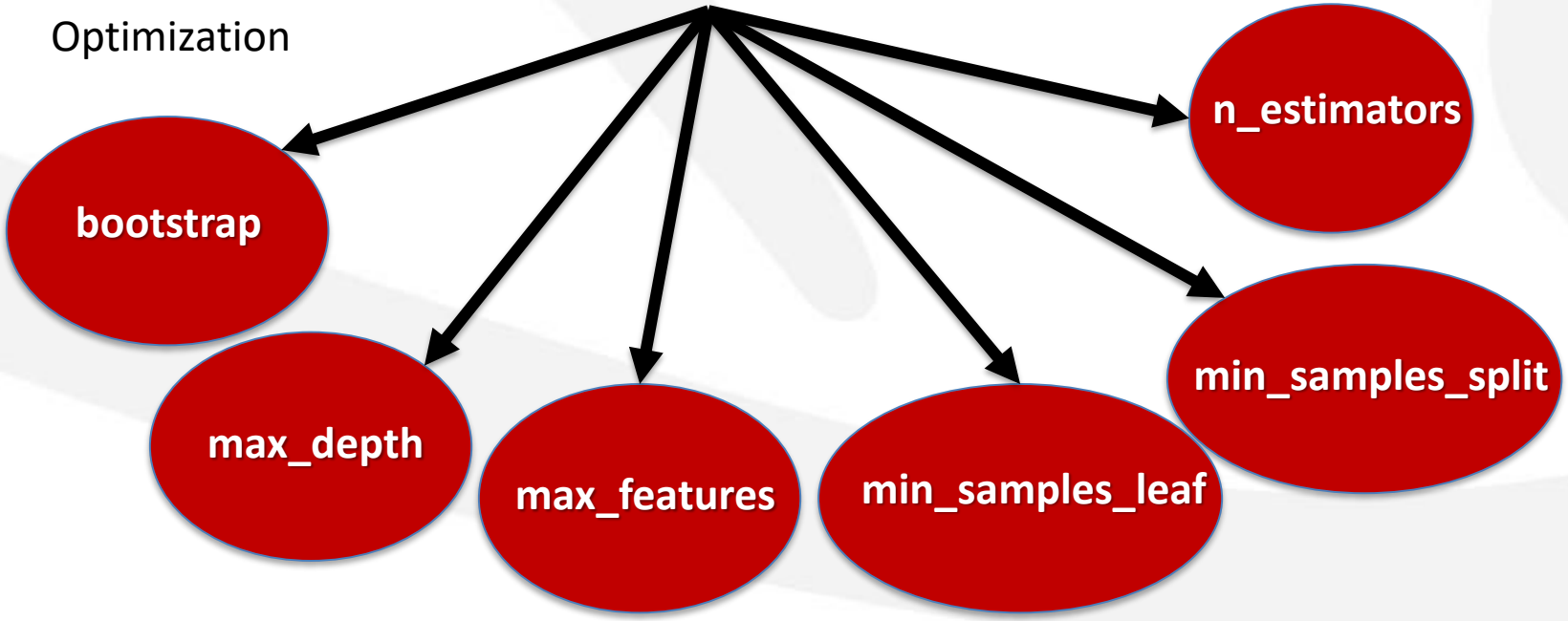
Parameter Optimization

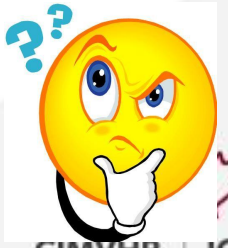




**RandomForest**



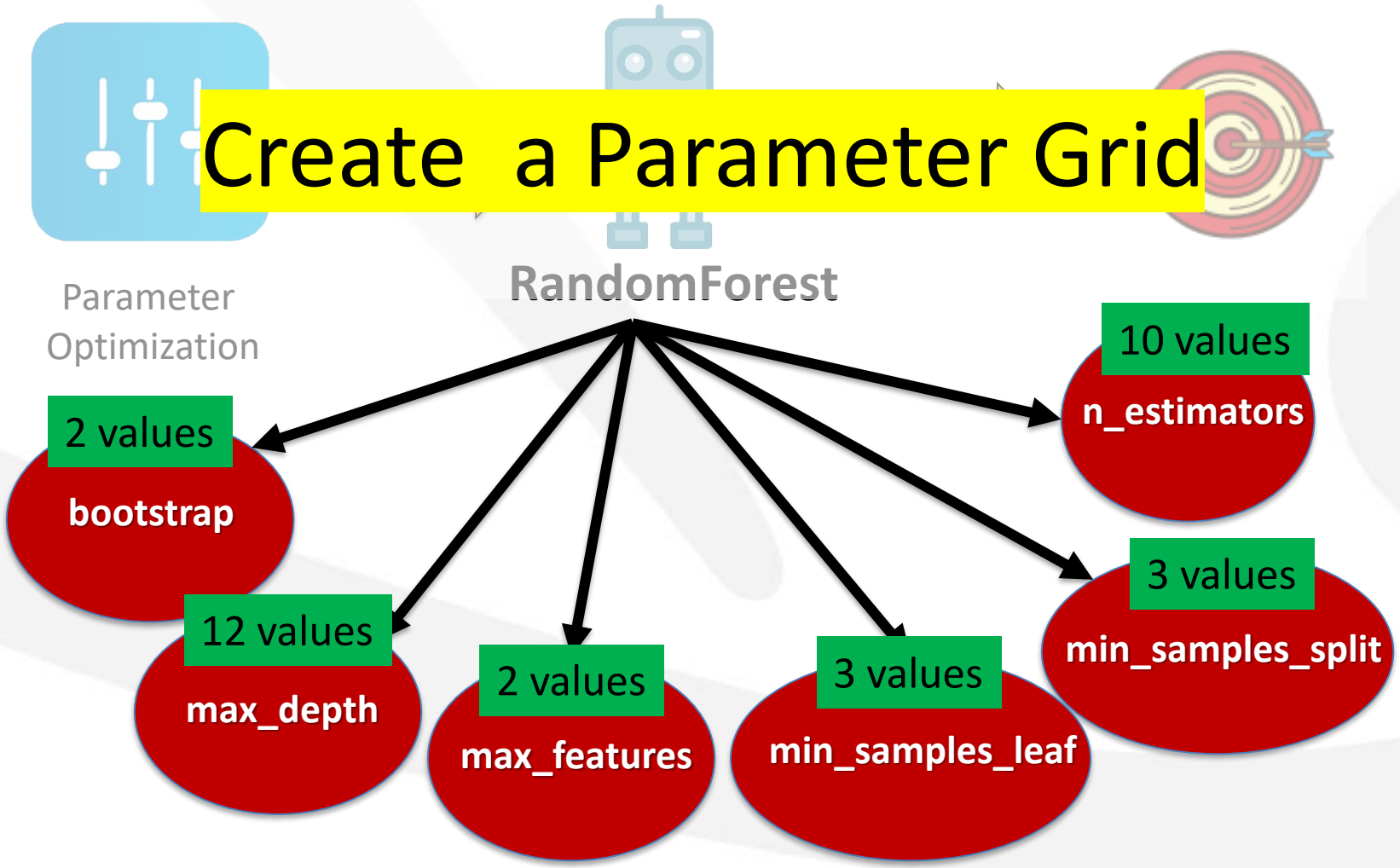


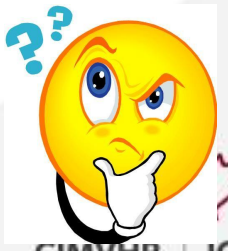



CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#6: Parameter Optimization

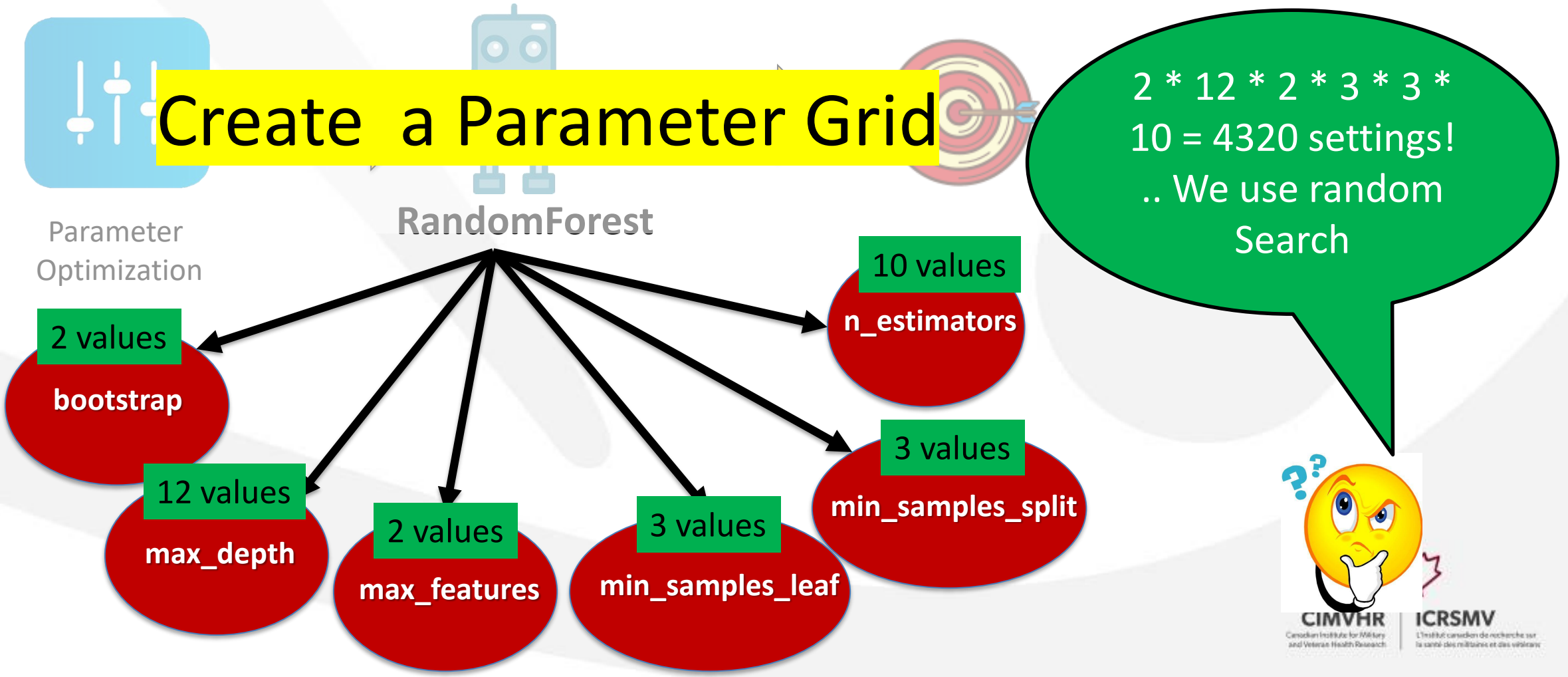


CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#6: Parameter Optimization



# Demo#7: Parameter Optimization


**Lets Do Some Coding!**

# Selected Topics Demos

**Parameter Optimization**

**Automatic Feature Selection**

# DEMO#7 [PART1]

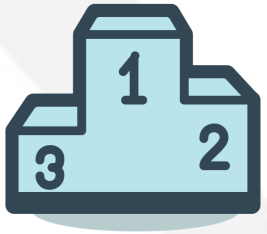
## Automatic Feature Selection

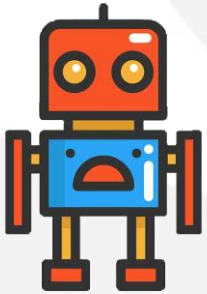
# Demo#7: Automatic Feature Selection

## Breast Cancer Dataset

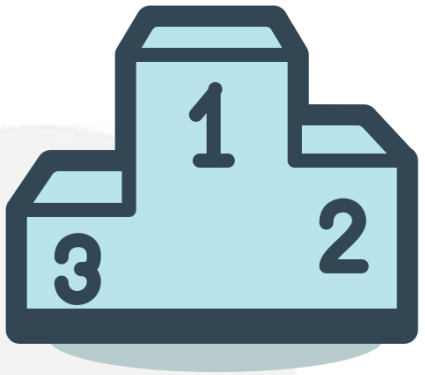
## 30 Features

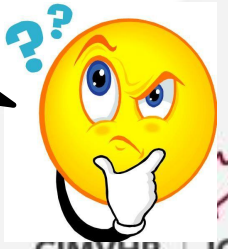
| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |



So many features… can automatically select the good features and ignore the noisy one



CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

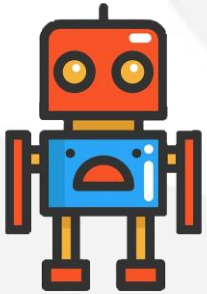
# Demo#7: Automatic Feature Selection

## Breast Cancer Dataset

## 30 Features

| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

**1. Univariate Selection**



One solution is studying the statistical relationship between the Features and the Target Class

CIMVHR Canadian Institute for Military and Veteran Health Research

ICRSMV L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#7: Automatic Feature Selection

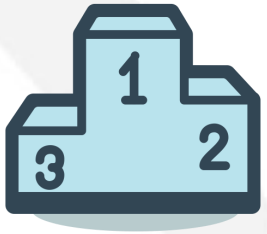
**Breast Cancer Dataset**

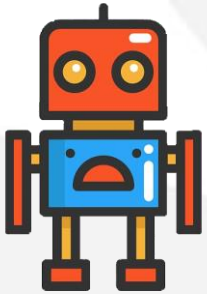
## 30 Features

| | | | | | |
|---|---|---|---|---|---|
| radius | compactness | radius | compactness | radius | compactness |
| texture | concavity | texture | concavity | texture | concavity |
| perimeter | concave points | perimeter | concave points | perimeter | concave points |
| area | symmetry | area | symmetry | area | symmetry |
| smoothness | fractal dimension | smoothness | fractal dimension | smoothness | fractal dimension |

**1. Univariate Selection**

**2. Recursive Feature Selection**



One solution is studying the statistical relationship between the Features and the Target Class

Another example is the **Recursive Feature Selection**

CIMVHR Canadian Institute for Military and Veteran Health Research

ICRSMV L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#7: Automatic Feature Selection


**Lets Do Some Coding!**

# DEMO#7 [PART2]

## Most Informative Features

CIMVHR
Canadian Institute for Military and Veteran Health Research

ICRSMV
L'Institut canadien de recherche sur la santé des militaires et des vétérans

# Demo#7: Most Informative Features