Análise de dados em R

Regression

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Antes de começar

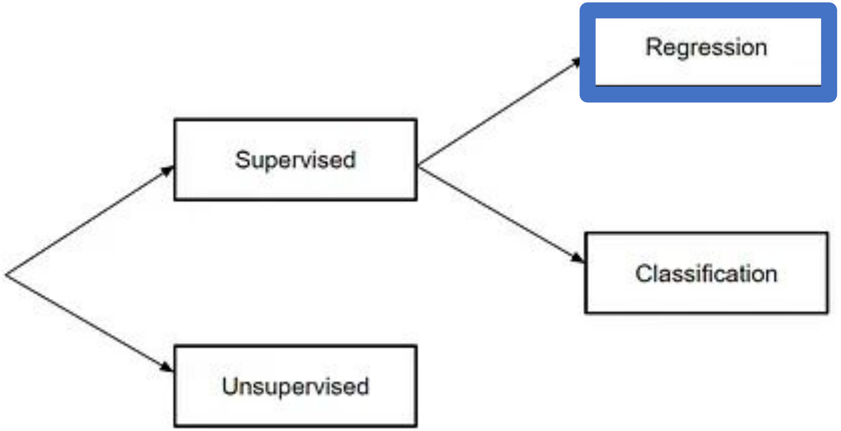
Instalar packages:

```
install.packages(c(“Metrics",”e1071”,
“randomForest”,”xgboost”,”caret”,”mlbench”,”pROC”,”rpart”))
```

Techniques - Statistical Modelling

Relate one or more characteristics (independent variables) to another characteristic of a individual (dependent variables)

Find a estimator, that given input (X) gives a estimation for dependent variable



- 💡 Relate CO2 with Earth temperature
- 💡 Relate the characteristics of flowers to their species
- 💡 Group patients with similar symptoms or diseases

Tasks - Predictive

Make predictions of a dependent variable using other independent variables.

- 💡 Forecast sales for next week
- 💡 Predict measure of progression in patients with Parkinson's disease
- 💡 Identifying the bird specie using bird sound
- 💡 Just go see more predictions tasks in https://www.kaggle.com/competitions

Regression - Example

Can we relate the music attributes with the popularity for Metallica songs?

Regression - Step by step

1) Problem definition
2) Data preparation
3) Select regression algorithms
4) Evaluation and refine algorithm hyperparameter

Regression - Problem definition

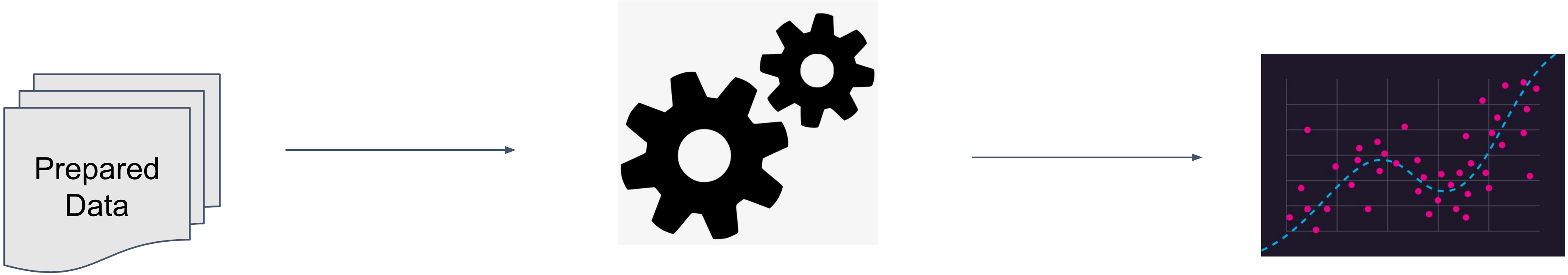
1. **Problem definition:**
    a. **Is this a regression problem?**
    b. **Which granularity of our target and the input data?**
    c. **Which attributes of our object of study we should consider?**

Regression - Step by step

2) Data preparation:

- **Transform data to the right granularity**
- **Categorical features to numeric (depending the algorithms)**
- **Deal with missing values**
- **Deal with outliers**
- **Select features**

Regression - Modelling



Prepared Data

Regression Algorithm

Model

Regression - Evaluation of Performance

The MSE, MAE, RMSE, and R-Squared metrics are mainly used to evaluate the prediction error rates and model performance in regression analysis.

- **MAE** (Mean absolute error) represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.
- **MSE** (Mean Squared Error) represents the difference between the original and predicted values extracted by squared the average difference over the data set.
- **RMSE** (Root Mean Squared Error) is the error rate by the square root of MSE.
- **R-squared** (Coefficient of determination) represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages. The higher the value is, the better the model is.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

Where,

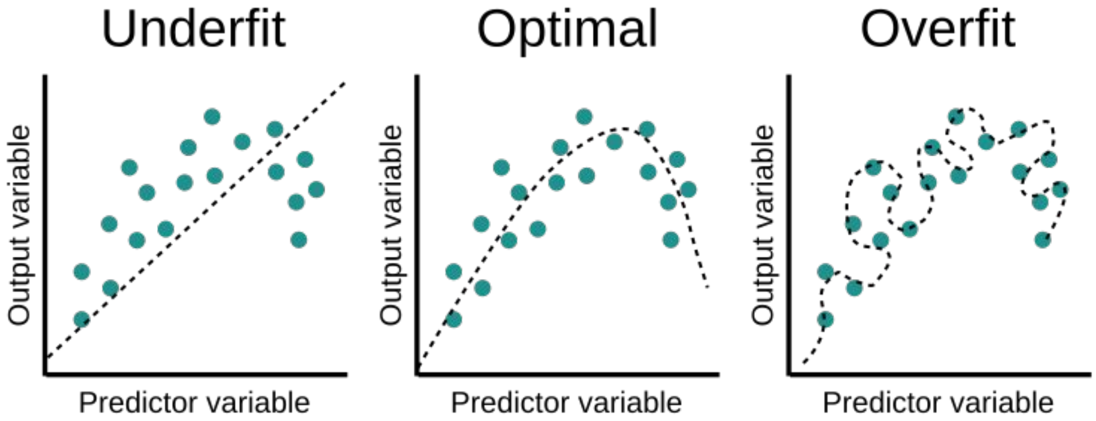
$\hat{y}$ $-$ *predicted value of y*

$\bar{y}$ $-$ *mean value of y*

Regression - Evaluation of Performance

Why

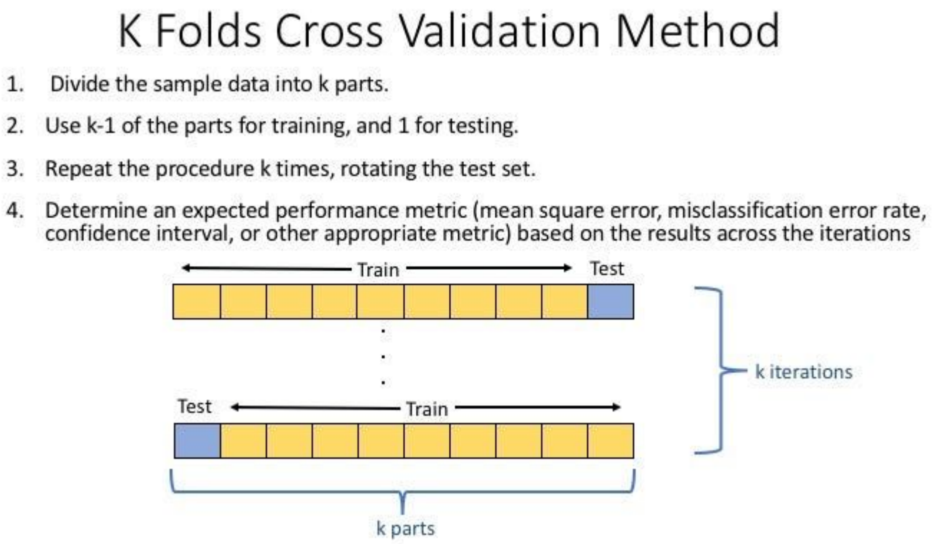
- **Capability of prediction in new data**
- **Generalization capability**
- **Avoid coincidences**



How

- **Split validation**
- **k Cross-Validation**

Regression - Evaluation of Performance



K Folds Cross Validation Method

1. Divide the sample data into k parts.
2. Use k-1 of the parts for training, and 1 for testing.
3. Repeat the procedure k times, rotating the test set.
4. Determine an expected performance metric (mean square error, misclassification error rate, confidence interval, or other appropriate metric) based on the results across the iterations

Split Validation is just one iteration (when you have a lot of data)

Regression - Evaluation of Performance

	Overfit	Right Fit	Underfit
Train Evaluation	High	High	Low
Test Evaluation	Low	High	Low

Regression - Evaluation of Performance

When you evaluate a new machine learning model and end up with an accuracy number or other metric, you need to know if it is meaningful.

Baseline:

- Train target average
- Homologous period

Regression - Evaluation of Performance

Exercise

Use the Metallica regression example and :

1. **Criar o modelo rpart**
   a. **avaliar MSE, RMSE para o rpart no treino e teste**
   b. **Visualizar a arvore**
2. **Add an algorithm: SVM and Random Forest and evaluations**

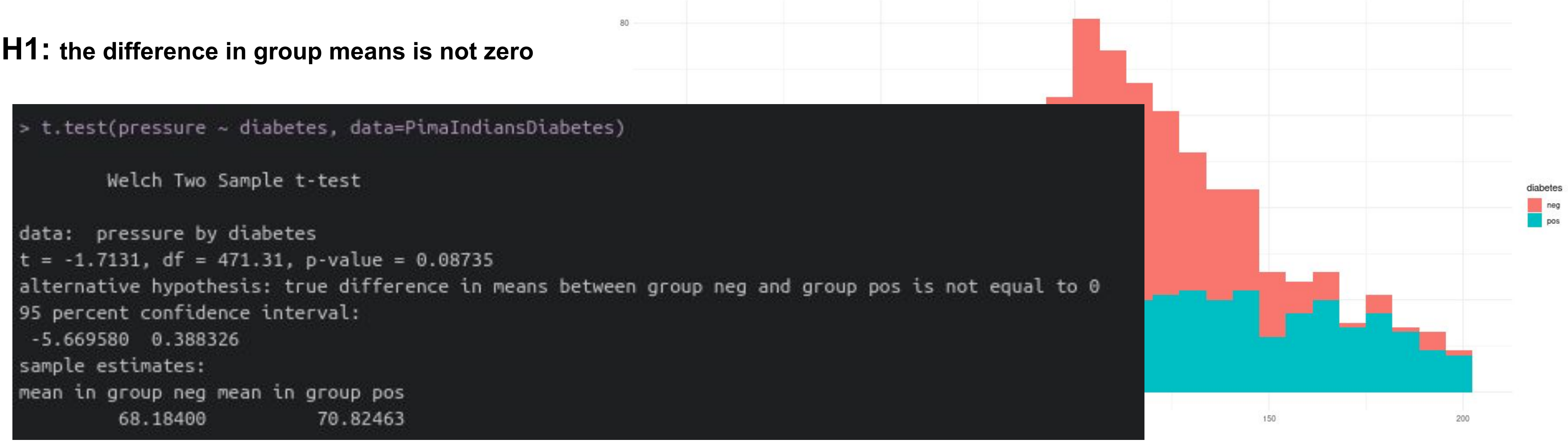
Make a summary table which compare the results of the algorithms and should which one shield be use

Classification - Example

The null and alternative hypothesis of an T-test are:

H0: the difference in group means is zero

H1: the difference in group means is not zero



```
> t.test(pressure ~ diabetes, data=PimaIndiansDiabetes)

	Welch Two Sample t-test

data:  pressure by diabetes
t = -1.7131, df = 471.31, p-value = 0.08735
alternative hypothesis: true difference in means between group neg and group pos is not equal to 0
95 percent confidence interval:
 -5.669580  0.388326
sample estimates:
mean in group neg mean in group pos
         68.18400          70.82463
```

A statistically significant test result (P > 0.05) means that the test hypothesis should not be rejected.