Ciência de Dados:
da teoria à prática

Prof. Dr. Felipe de Morais

UNISINOS

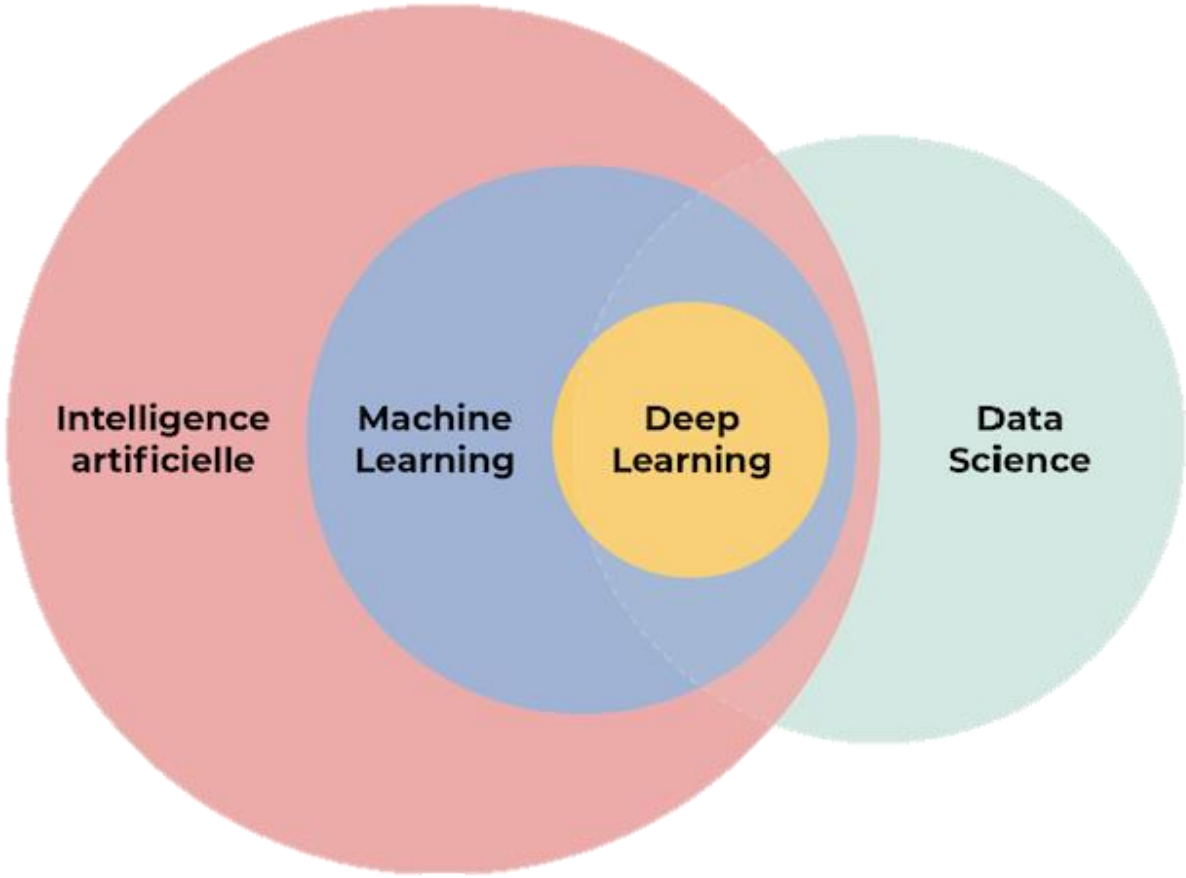
Parte teórica

UNISINOS

Conteúdo

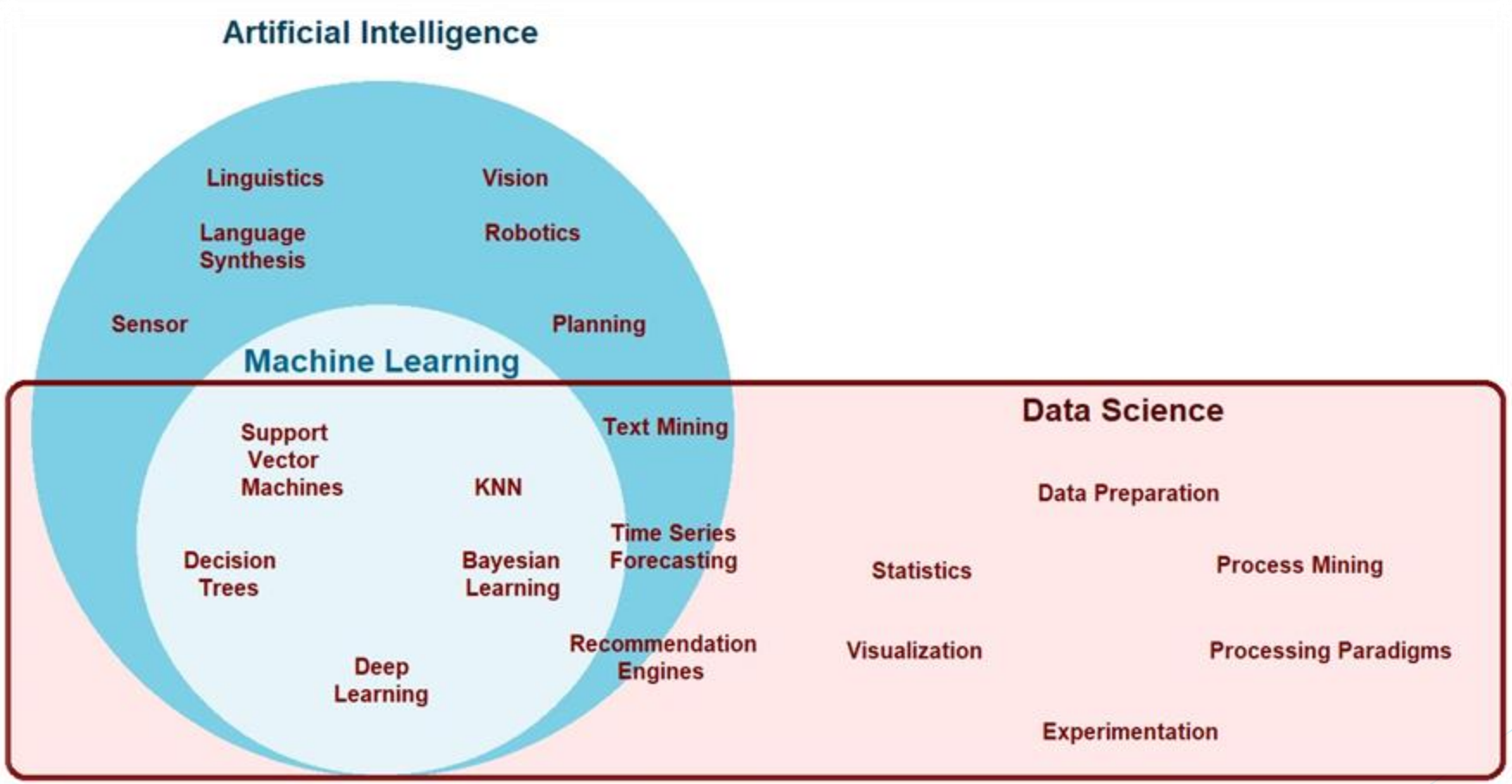
- Bases de dados
- Carregamento de bases de dados
- Valores inconsistentes
- Valores faltantes
- Escalonamento de atributos
- Transformações de variáveis categóricas
- Divisão de dados em treino e teste
- Recursos do pandas (localizar, remover linhas ou colunas, alterar valores, filtrar,...)

Definição



UNISINOS

Definição

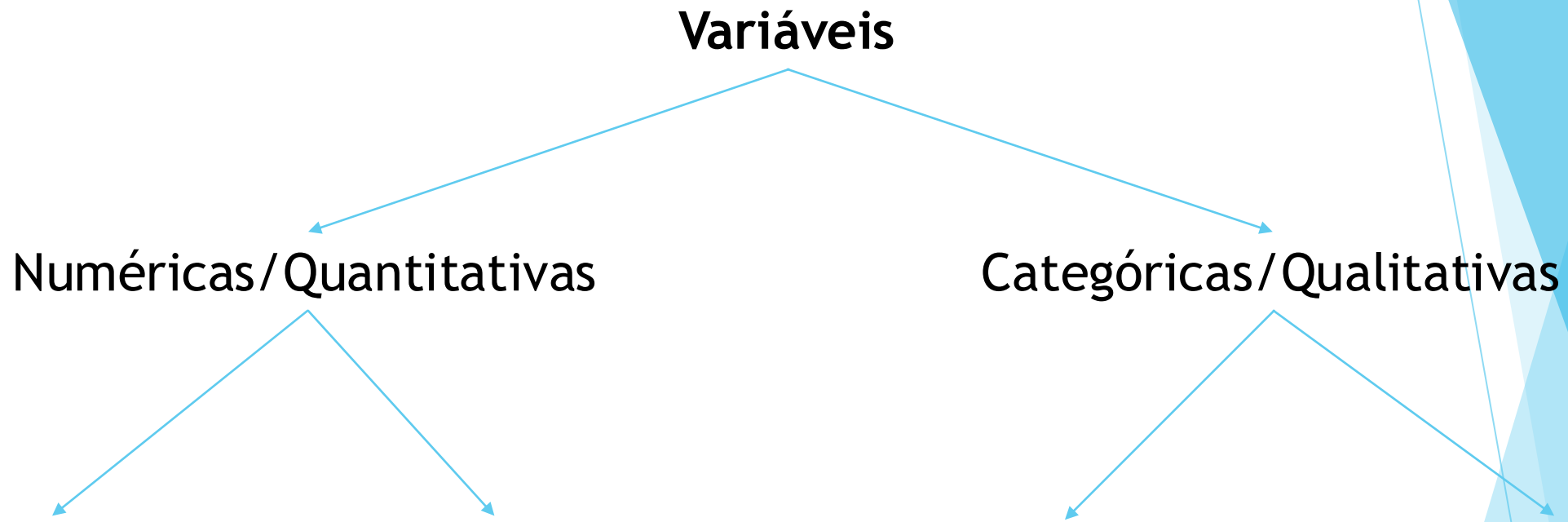


UNISINOS

Processo

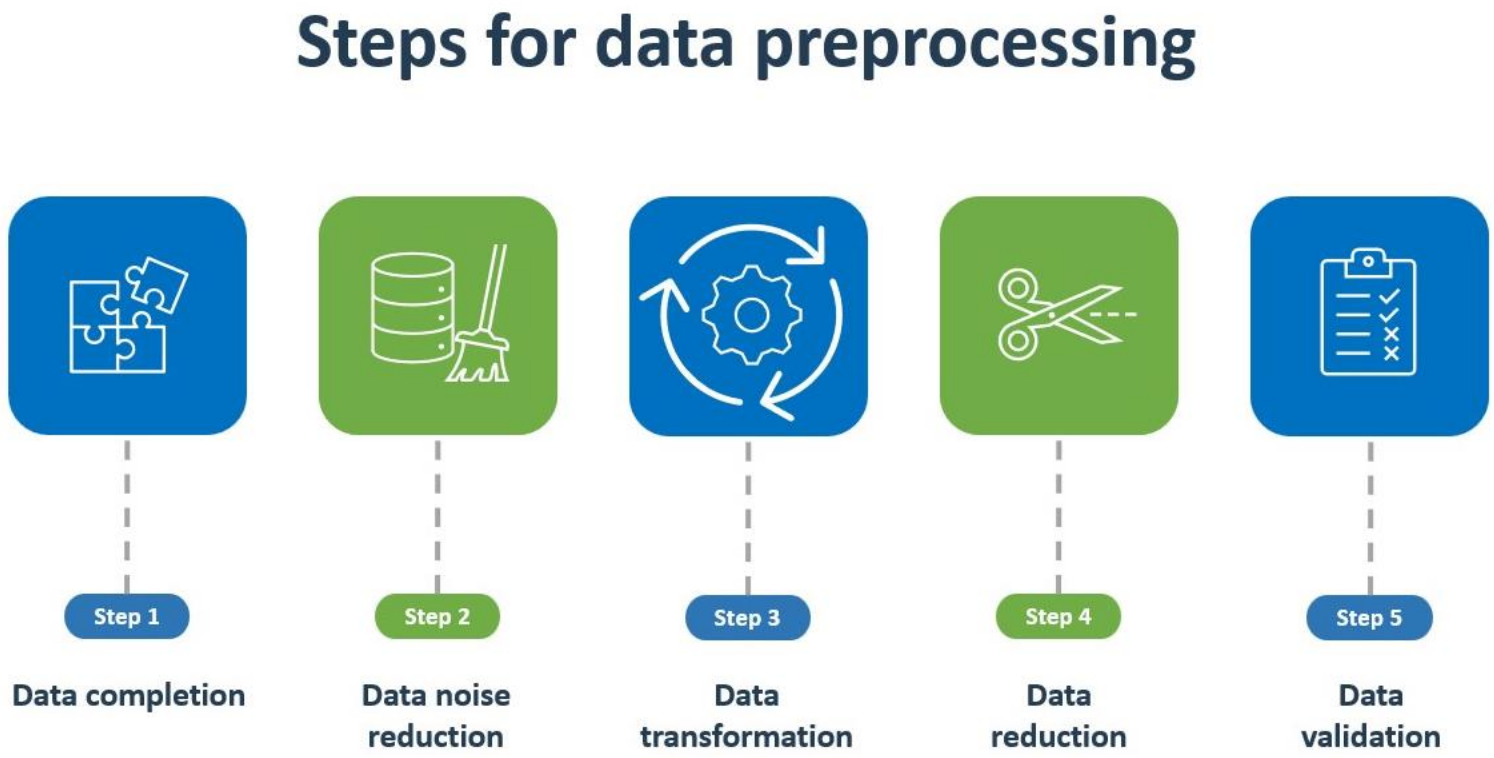
UNISINOS

Tipos de Variáveis



UNISINOS

Etapas



UNISINOS

Dataset – Análise de Crédito

- Kaggle
  - https://www.kaggle.com/datasets/laotse/credit-risk-dataset

Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

UNISINOS

Dataset – Análise de Crédito

- Kaggle
  - https://www.kaggle.com/datasets/laotse/credit-risk-dataset

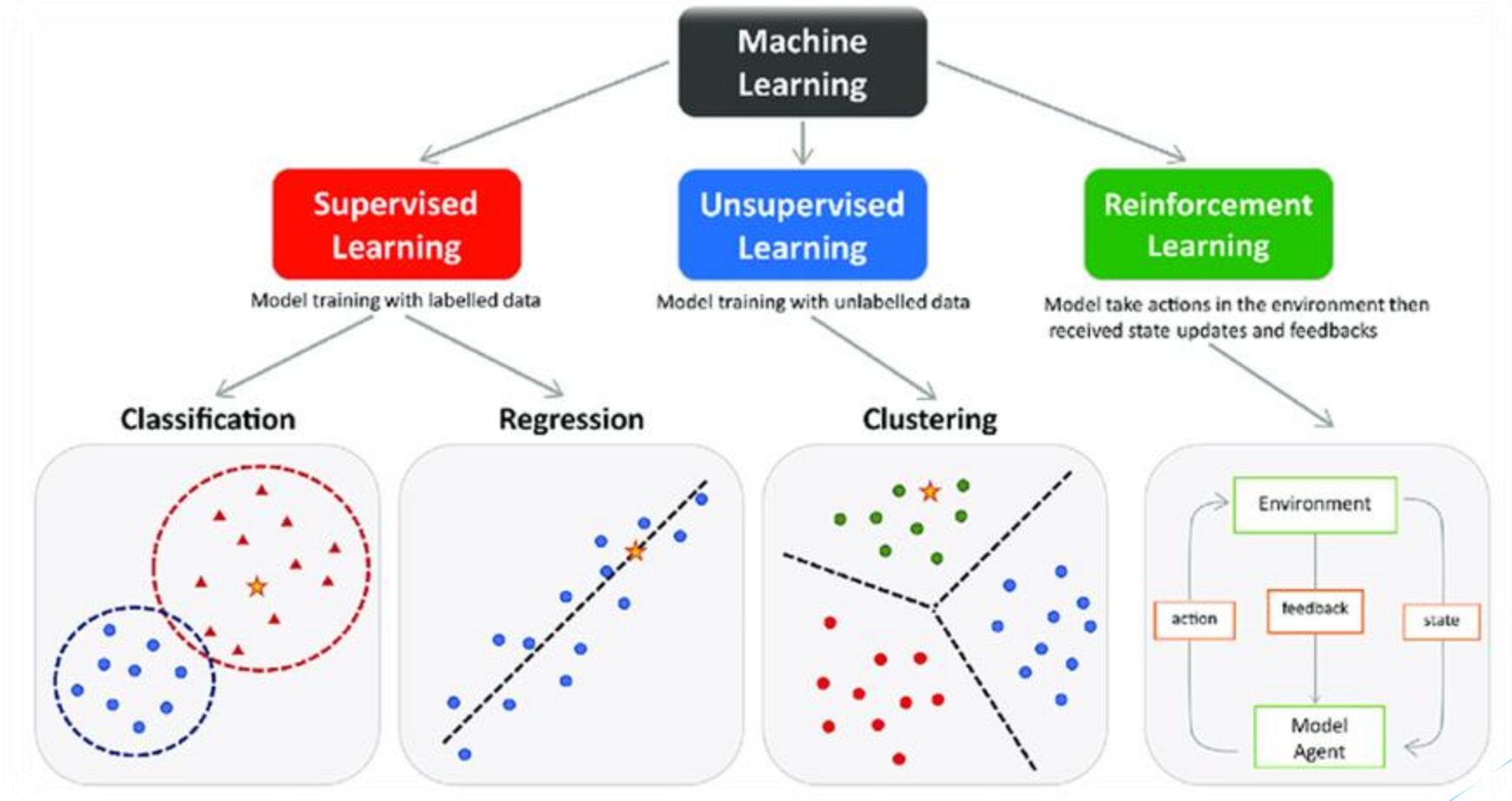
Feature Name	Description
person_age	Age
person_income	Annual Income
person_home_ownership	Home ownership
person_emp_length	Employment length (in years)
loan_intent	Loan intent
loan_grade	Loan grade
loan_amnt	Loan amount
loan_int_rate	Interest rate
loan_status	Loan status (0 is non default 1 is default)
loan_percent_income	Percent income
cb_person_default_on_file	Historical default
cb_preson_cred_hist_length	Credit history length

UNISINOS

O que fazer com a idade negativa?

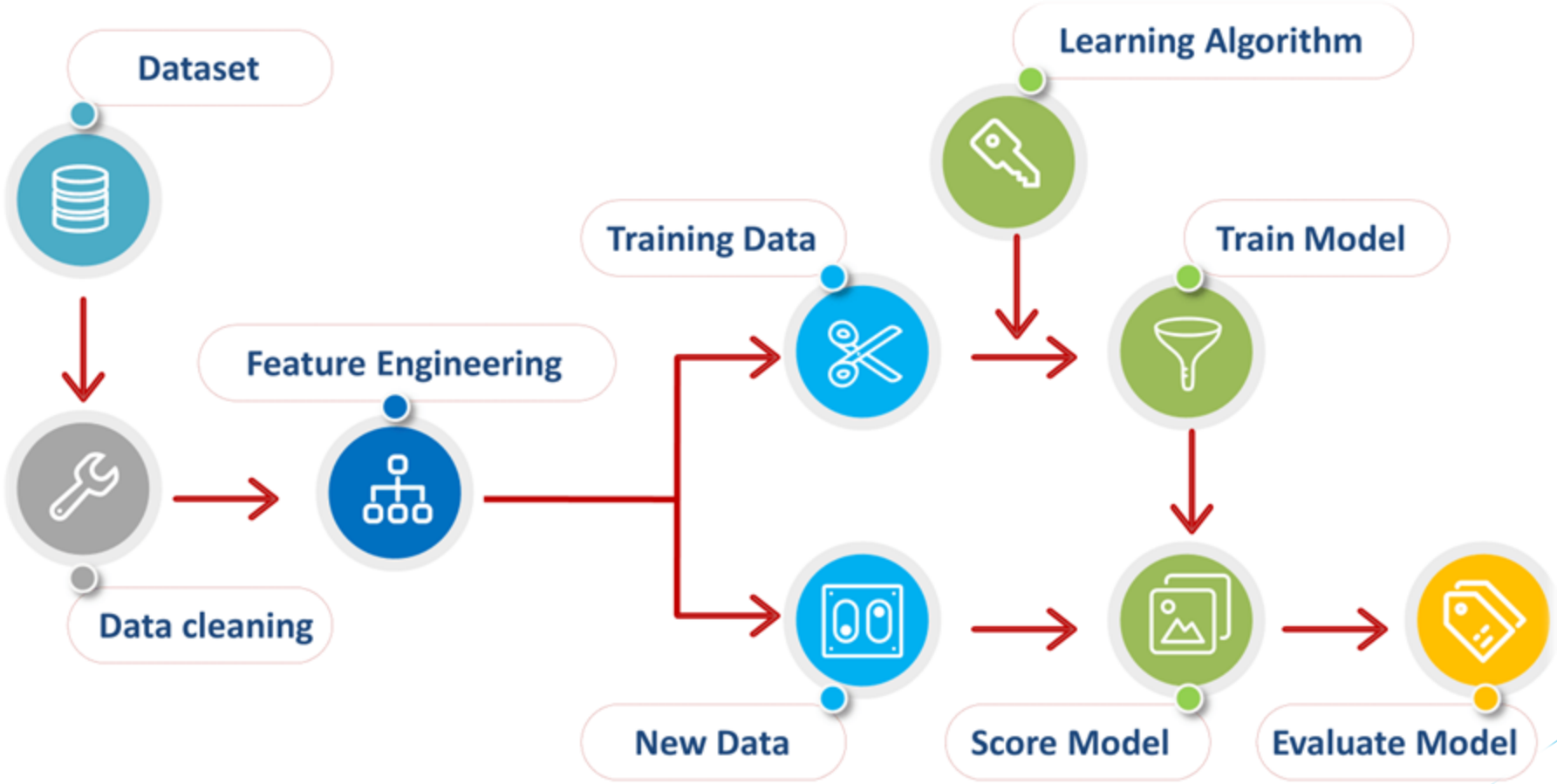
- Desconsidera a linha inteira
- Desconsidera a coluna inteira
- Colocou o – sem querer
- Usar a média
- Usar a mediana
- Usar a moda
- Prever a idade da pessoa com demais features
- Perguntar aos donos dos dados

Machine Learning



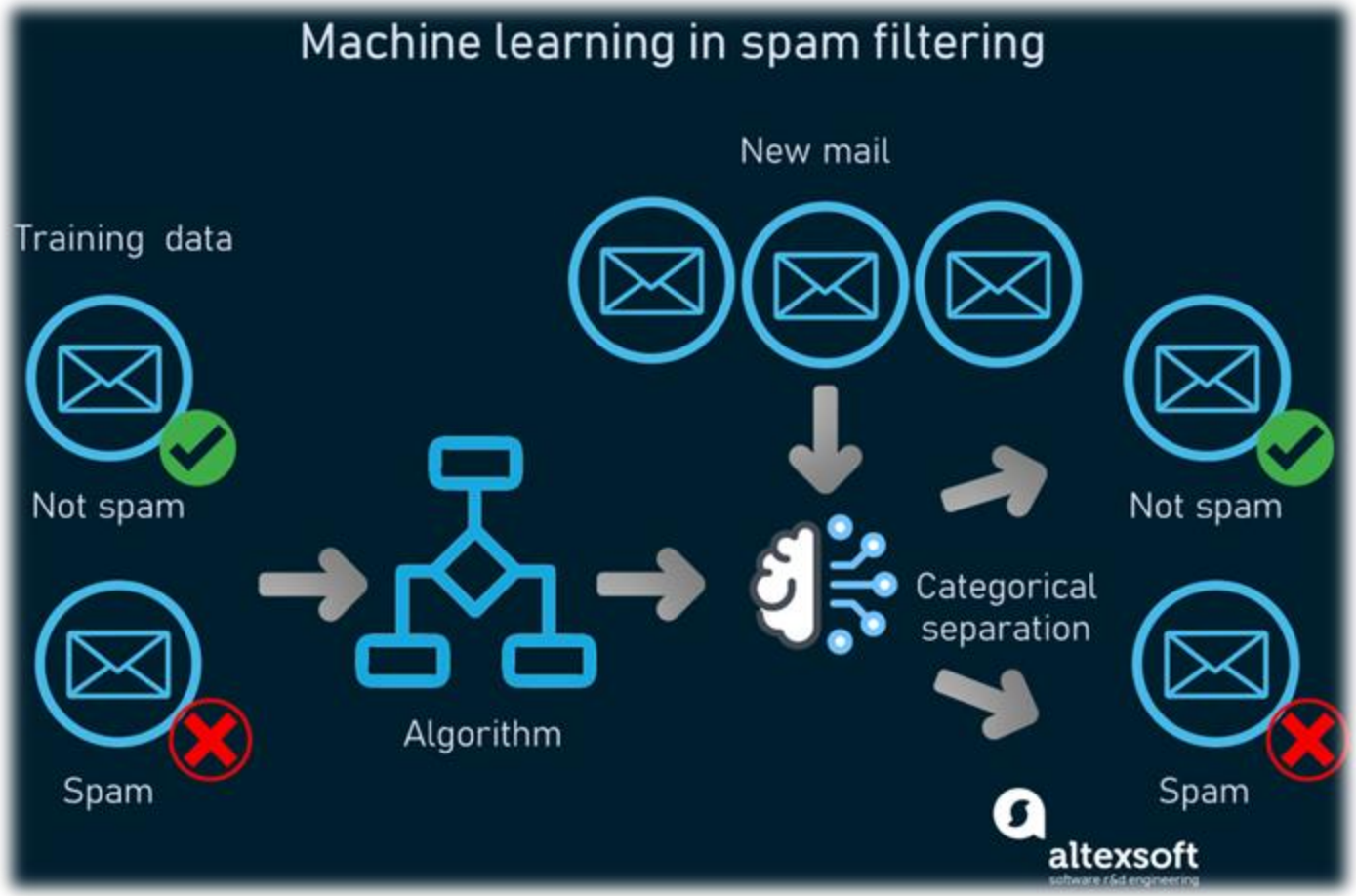
UNISINOS

Treinamento



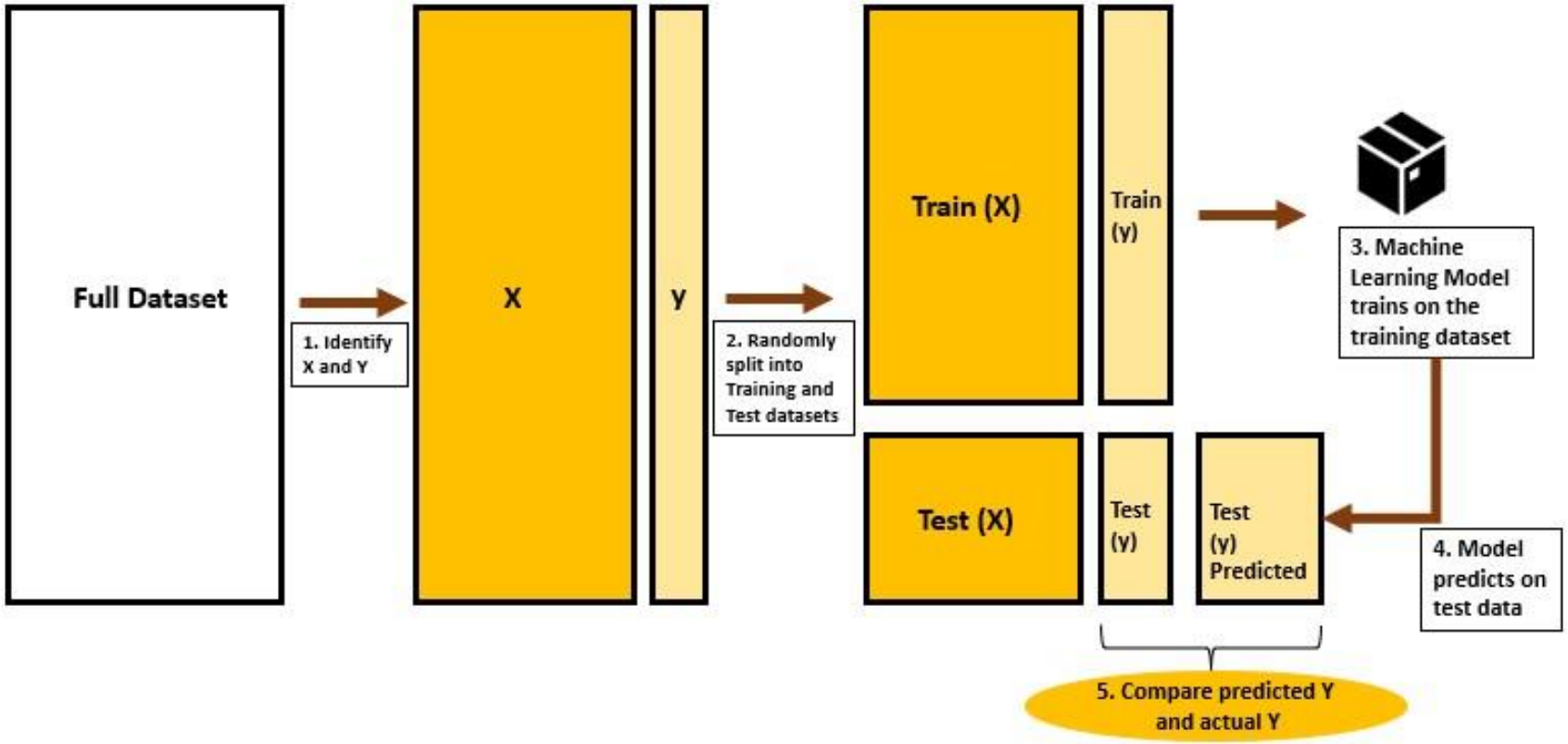
UNISINOS

Exemplo



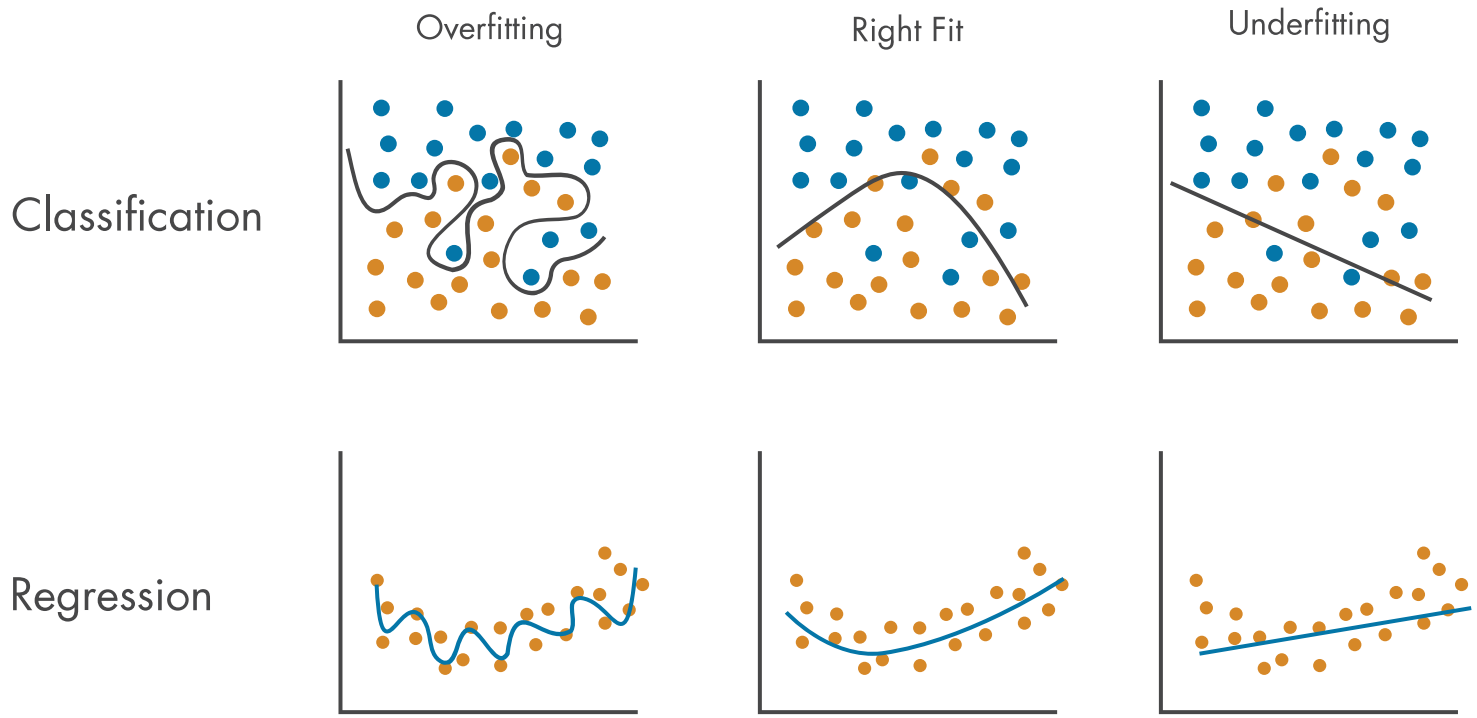
UNISINOS

Divisão dos dados



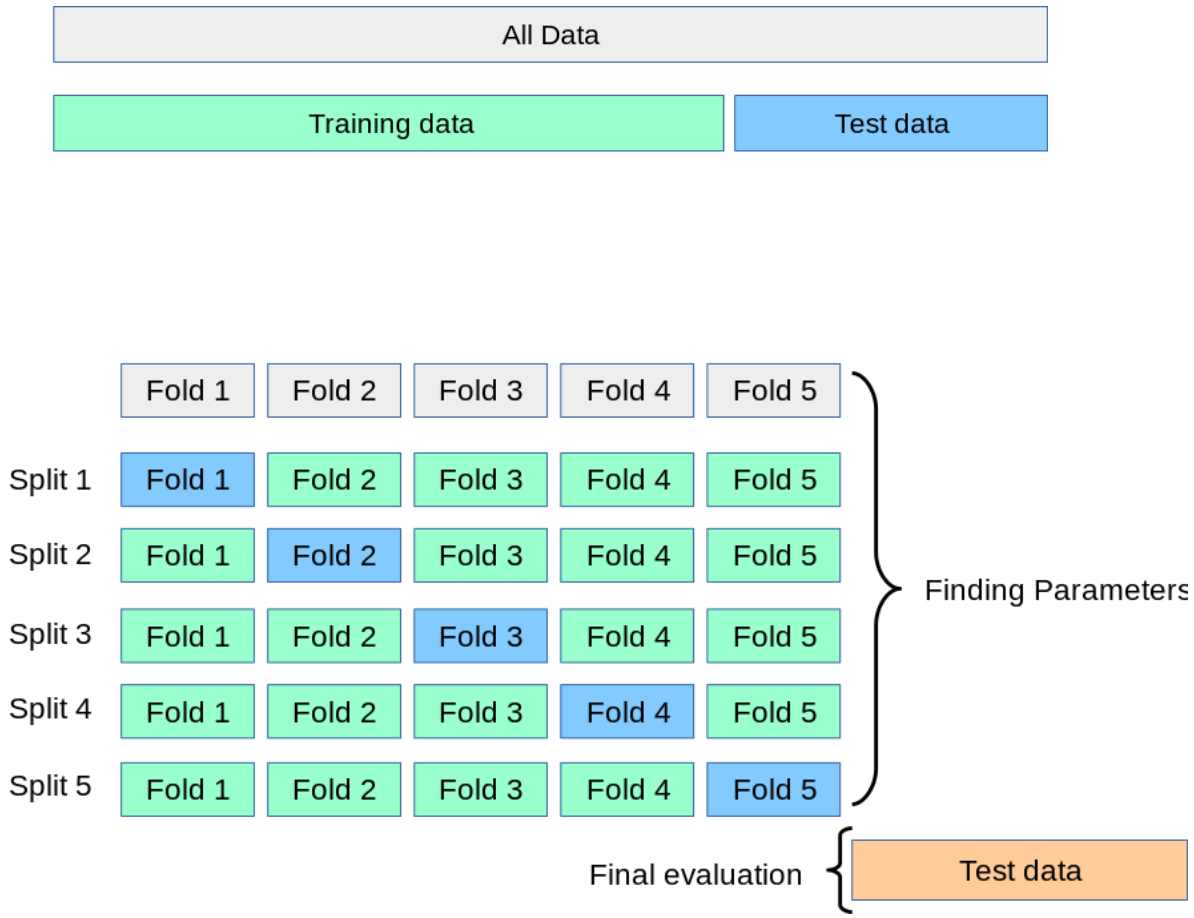
UNISINOS

Overfitting e Underfitting



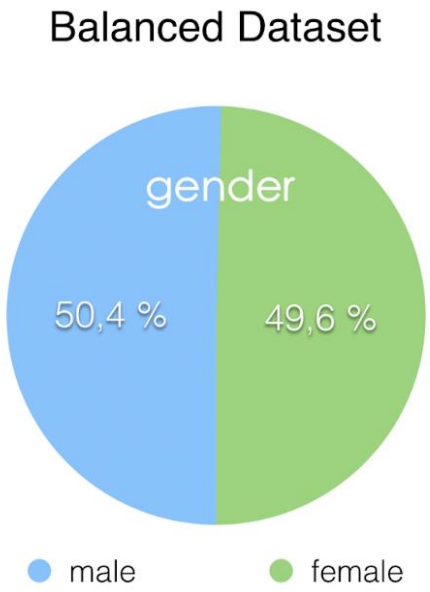
UNISINOS

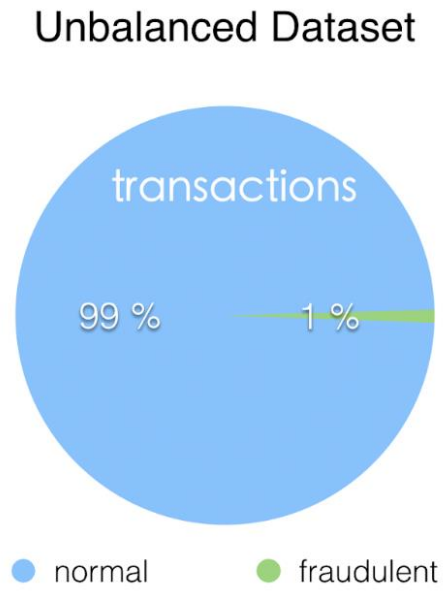
Cross-validation



UNISINOS

Distribuição de classes





UNISINOS

Avaliação de modelos de ML

Classificação de e-mails:
True = Spam
False = Normal

Real	Predito
True	True
True	False
False	False
False	True
True	True
False	False
...	...

UNISINOS

Matriz de confusão

Valor predito

	Positivo	**Negativo**
Valor real **Positivo**	True Positive (TP)	False Negative (FN)
Negativo	False Positive (FP)	True Negative (TN)

UNISINOS

Matriz de confusão – Classificação de Spam

Valor predito

Valor real		**Spam**	**Normal**
	Spam	200 (TP)	40 (FN)
	Normal	60 (FP)	700 (TN)

Total de e-mails: 1000
Total de Spam: 240 – 24%
Total de Normal: 760 – 76%

UNISINOS

Matriz de confusão – Accuracy

Valor predito

Valor real	**Spam**	**Normal**
Spam	200 (TP)	40 (FN)
Normal	60 (FP)	700 (TN)

Total de e-mails: 1000
Total de Spam: 240 – 24%
Total de Normal: 760 – 76%

Accuray: TP + TN / Total de amostras

UNISINOS

Matriz de confusão – Accuracy

Valor predito

Valor real		**Spam**	**Normal**
	Spam	200 (TP)	40 (FN)
	Normal	60 (FP)	700 (TN)

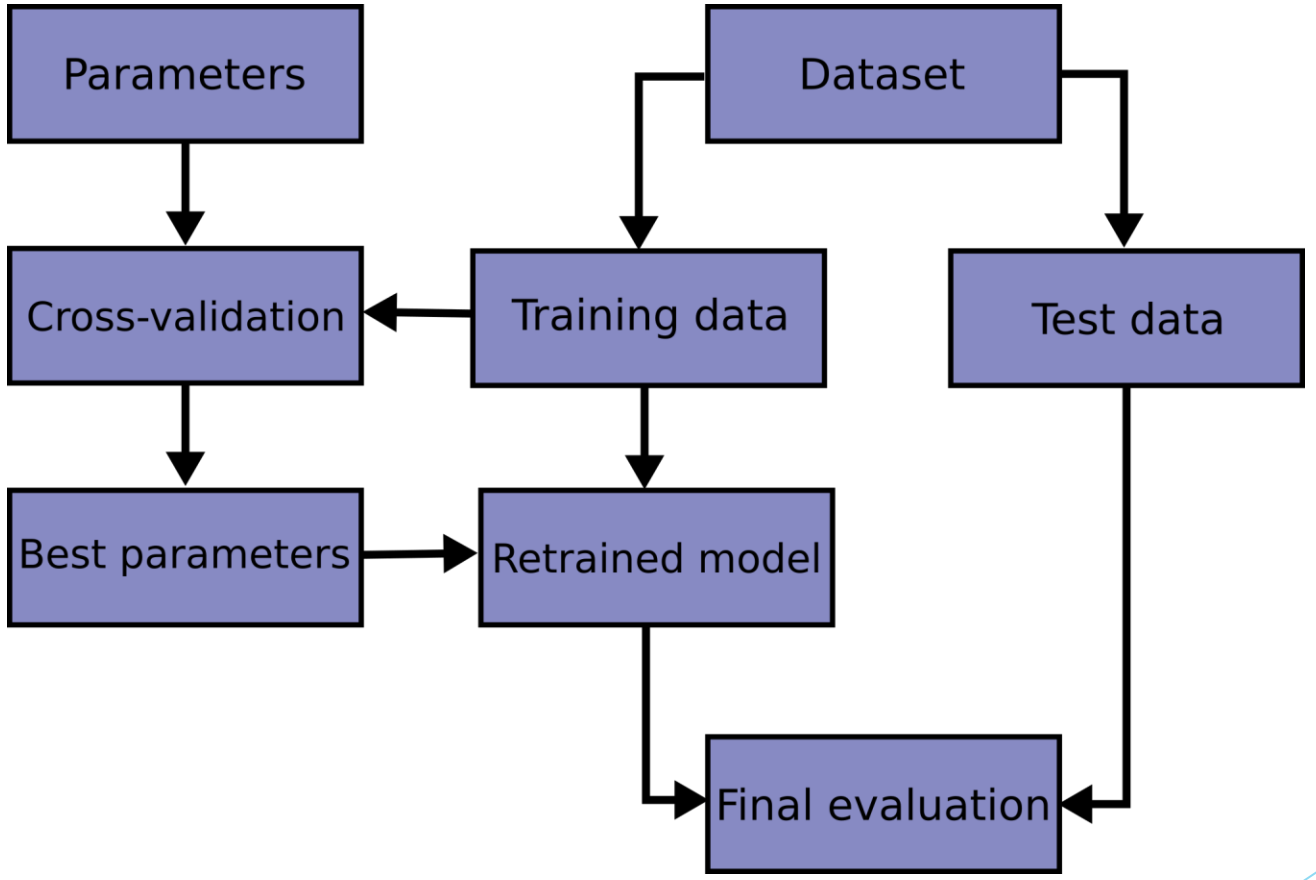
Total de e-mails: 1000
Total de Spam: 240 – 24%
Total de Normal: 760 – 76%

Accuray: TP + TN / Total de amostras

Accuray: 900 / 1000 = 0,9 = 90%

UNISINOS

Pipeline



UNISINOS

Parte prática

https://github.com/felipmorais/ml_pipeline

Dataset - Censo

- UCI
  - https://archive.ics.uci.edu/dataset/2/adult

Features

Attribute Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age	N/A		false
workclass	Feature	Categorical	Income	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.		true
fnlwgt	Feature	Integer				false
education	Feature	Categorical	Education Level	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.		false
education-num	Feature	Integer	Education Level			false
marital-status	Feature	Categorical	Other	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.		false
occupation	Feature	Categorical	Other	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.		true
relationship	Feature	Categorical	Other	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.		false
race	Feature	Categorical	Race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.		false
sex	Feature	Binary	Sex	Female, Male.		false

UNISINOS

Parte Prática

1. Achar e Baixar os dados
2. Importar os dados
3. Analisar os dados
4. Entender e explicar os dados
5. Utilizar gráficos para visualizar
6. Corrigir dados inválidos
7. Corrigir dados faltantes
8. Decidir quais atributos serão úteis
9. Classificar as variáveis/atributos
8. Separar entre previsores e classe
9. Transformar colunas
10. Padronizar os dados
11. Dividir entre conjuntos de treino e teste
12. Treinar um modelo de machine learning
13. Reportar os resultados estatísticos

Muito Obrigado!

https://www.linkedin.com/in/felipedemoraisphd/

UNISINOS