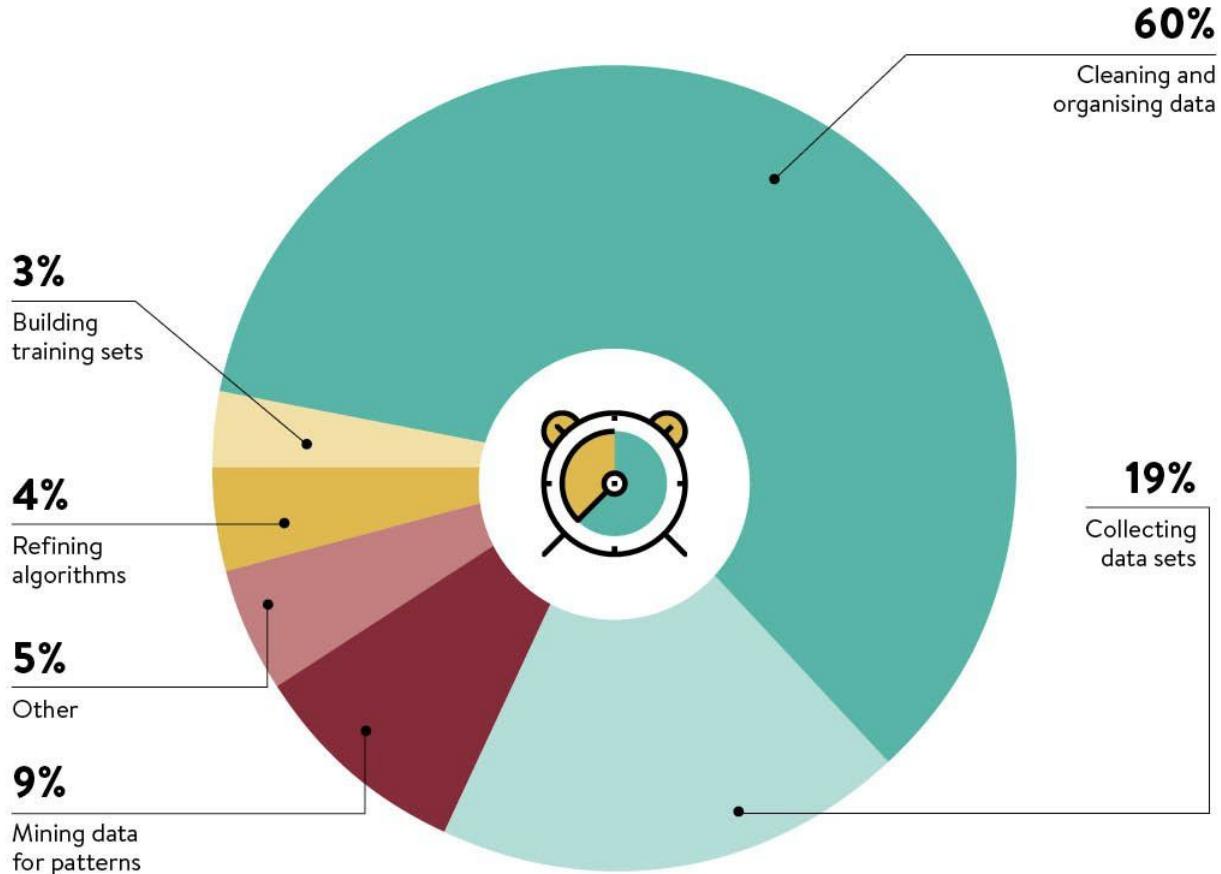




Feature engineering



WHAT DATA SCIENTISTS SPEND THE MOST TIME DOING



Source: CrowdFlower 2016

Escalas de Dados

- ◊ **Nominal:** nessa escala os valores **valores são não numéricos e não ordenados**. Por exemplo, cor, marca de carro, etc.
- ◊ **Ordinal:** Nessa escala os valores não são numéricos, mas são **ordenados**. Uma amostra pode apresentar um valor comparativamente maior do que uma outra. Ex: Função no trabalho

Escalas de Dados

- ◆ **Intervalar:** escala onde valores são numéricos, existindo uma ordem entre os valores e uma diferença entre esses valores. O zero é relativo.
- ◆ **Proporcional:** nessa escala de valores numéricos, além da diferença, tem sentido calcular a proporção entre valores.

Os atributos podem ser:

- ◊ **Qualitativo**:
escalas
nominais ou
ordinais
 - Variáveis Discretas
 - Binárias

- ◊ **Quantitativo**:
escalas
intervalar ou
proporcional
 - Variáveis **contínuas**

Ausentes ou
inaplicáveis

Pré-processamento do dado



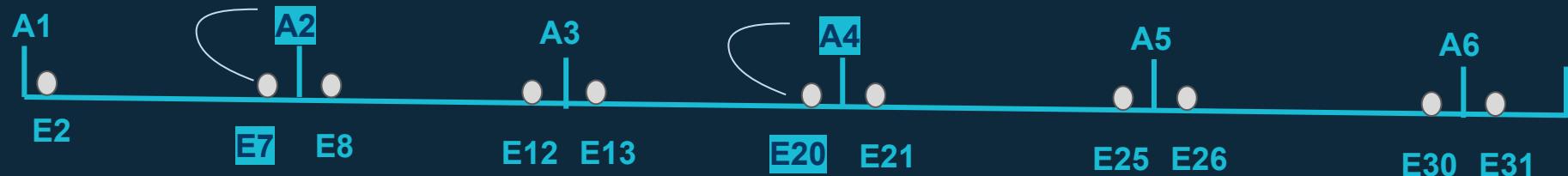


Coletar o dado

- ◊ Dados Públicos
- ◊ Dados no DBpedia
- ◊ Plataformas de ensino (e.g. Kaggle, UCI)
- ◊ Crowler
- ◊ REST API
- ◊ Acesso direto as fontes de dados
- ◊ Dado estruturado
- ◊ Dado não estruturado



Modelagem da amostra





Compreender e Integrar



Limpeza dos dados

- ◊ Preencher dados ausentes
 - Como preencher valores numéricos?
 - Como preencher valores nominais?
 - Aplicar ML
- ◊ Remover dados ausentes
 - Quando eliminar uma amostra?
 - Quando eliminar uma coluna?
- ◊ Identificar outlier
 - Qual a melhor fórmula?

Criação de features

- ◊ Aplicar Fórmula (e.g.: Faixa salarial)
- ◊ Valores proporcionais (e.g.: IMC)
- ◊ Opcional: Eliminar features originais

“É necessário para obter os dados em uma forma apropriada para a aplicar data science com machine learning”

Transformação

- ◊ Label encoding
 - Sexo (F, M) → Sexo (F: 0), (M: 1)
- ◊ One Hot Encoding
 - Resulta em uma matriz esparsa

	Idade	Sexo_M	Sexo_F
Amastra _1	10	1	0
Amastra _2	30	0	1

Agregação

- ◊ Combinar dois ou mais atributos (ou objetos) em apenas um atributo (ou objeto), e.g.: cargo e função)
- ◊ Objetivo:
 - Reduzir o número de atributos ou amostras
 - Mudar escala (e.g: cidade em estado)
 - Possuir dados mais estáveis devido a menor variabilidade

Feature scaling

Normalização: o propósito é **minimizar os problemas oriundos do uso de unidades e dispersões distintas entre as variáveis.**

Normalização segundo a amplitude: unidades diferentes ou dispersões muito heterogêneas.

◆ Min e max norm:

$$Y = \frac{X - \text{min}}{\text{Max} - \text{Min}}$$

◆ Média norma.:

$$Y = \frac{X - \text{media}}{\text{Max} - \text{Min}}$$

◆ Standardization

$$Y = \frac{X - \text{media}}{\text{std}}$$

Feature scaling

Normalização distribucional: é interessante nas situações em que há distorção nos valores aberrantes, obtenção de simetria etc. Por exemplo: salário dos brasileiros

Exemplo mais comum:
◊ Log X
Salários (1000, 10000)



Pré-processamento do dado

Merging



MERGE



Seleção de features



Importância

- ◊ Otimizar modelo
- ◊ Facilitar a interpretação
- ◊ Obter *insights*
- ◊ Eliminar atributos insignificantes
- ◊ ...



Filter method



- ◊ É independente do modelo de aprendizagem
 - ◊ Pode ser feito com base no conhecimento do negócio
- Exemplos:
- ◊ Seleção manual
 - ◊ Correlação de Pearson
 - ◊ Chi Square

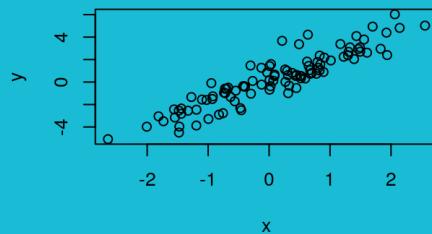
Filter method

Feature/Response	Contínua	Categórica
Contínua	Correlação de Pearson	LDA
Categórica	Anova	Chi-Square

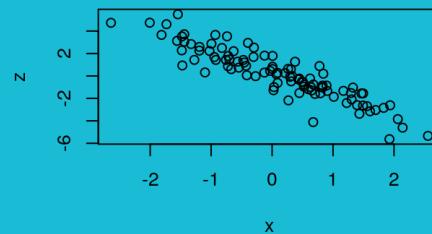
O que fazer com variáveis que são
fortemente correlacionadas?

Correlação de pearson

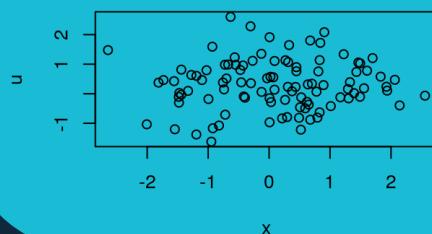
Relação linear positiva



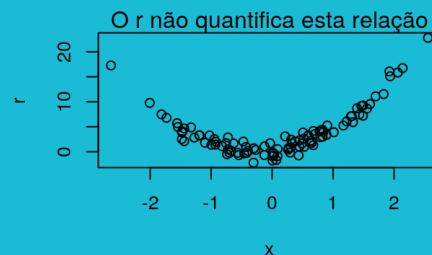
Relação linear negativa



Ausência de relação

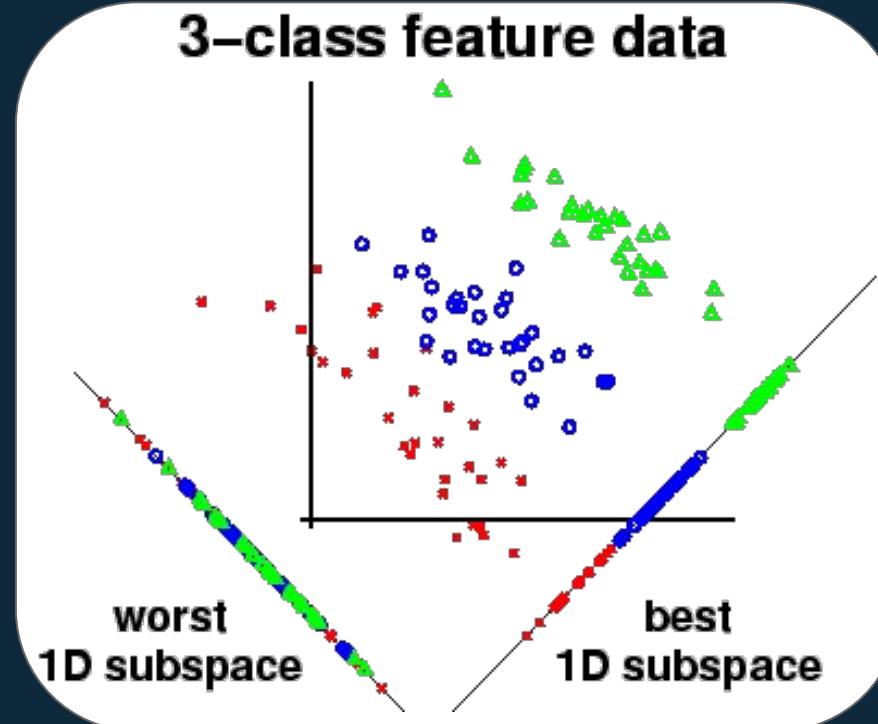


Relação não-linear

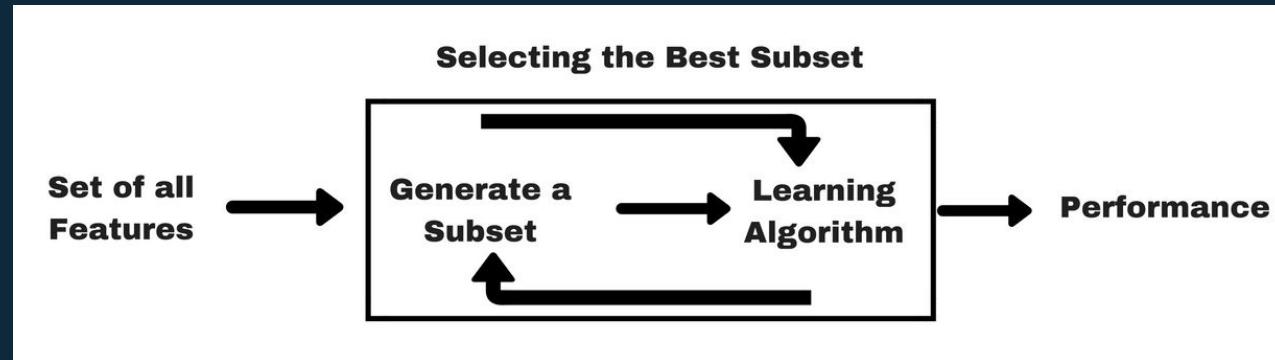


Filter method

LDA - Análise Discriminante Linear



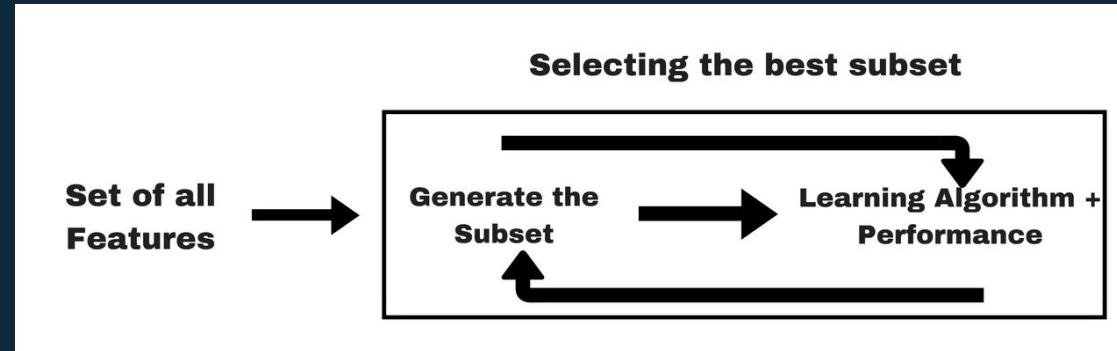
Wrapper method



- ◊ Forward Selection
- ◊ Backward Elimination
- ◊ Recursive Feature elimination

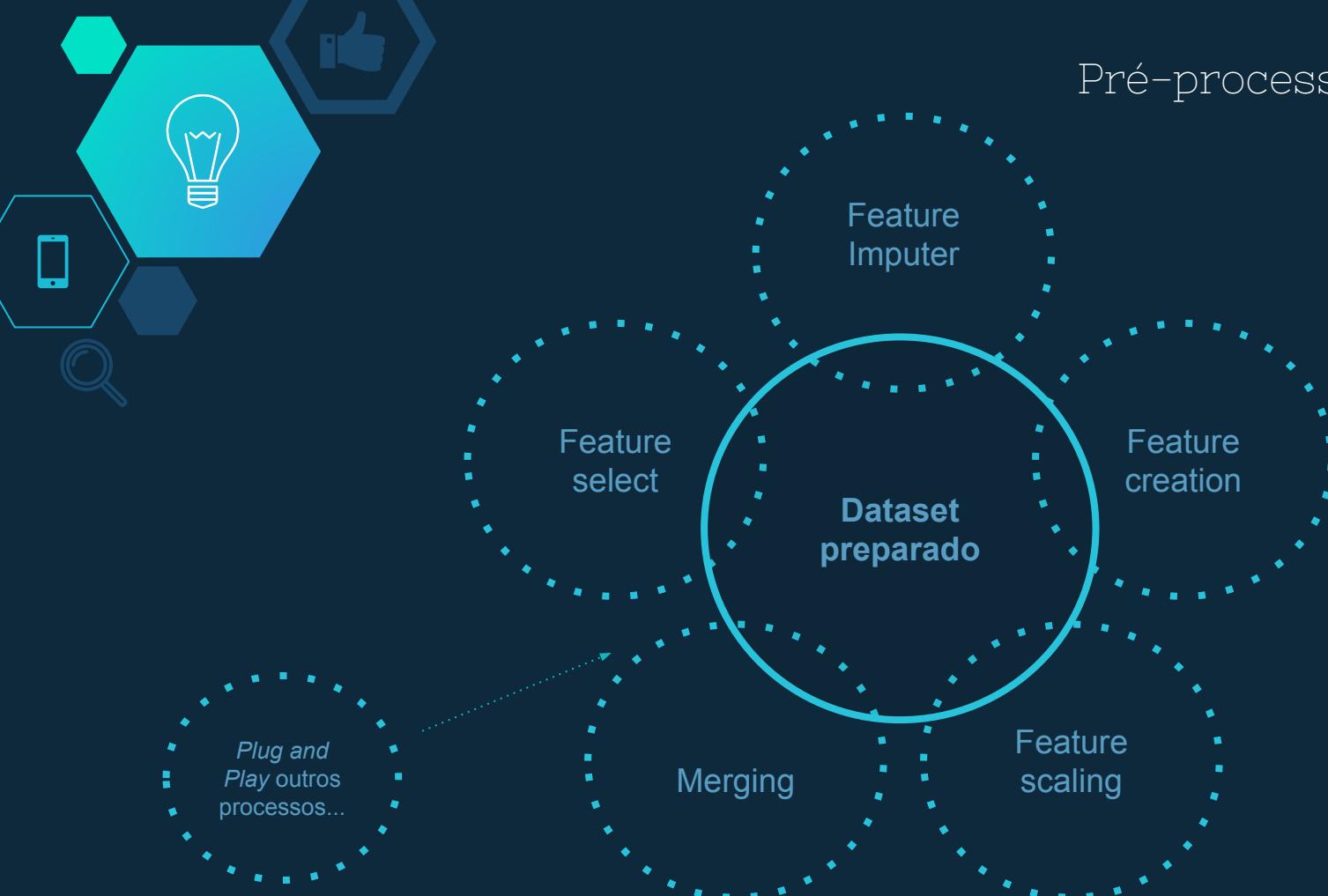
Qual a complexidade de todos os testes possíveis?

Embedded method



- ◊ Ganho da informação
 - Modelos baseado em árvore
- ◊ Lasso regression performs L1
- ◊ Ridge regression performs L2

Pré-processamento do dado





Qual o melhor
pré-processamento
do dado?

http://dontpad.com/kdd_uni7



Hands on