THE DEVELOPER'S CONFERENCE

A Maldição da Dimensionalidade: como selecionar características em problemas complexos?

Msc. Felipe Teodoro
Head of Data Science

Felipe Teodoro

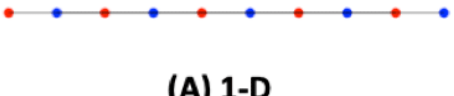
THE DEVELOPER'S CONFERENCE

- Mestre em Sistemas de Informação pela USP.
- MBA em Engenharia de Software pela FIAP.
- Tecnólogo em Análise e Desenvolvimento de Sistemas pela Faculdade de Tecnologia Termomecânica .
- Autor de artigos acadêmicos e entusiasta de Inteligência Artificial.
- Head de Data Science da empresa BuiltCode.

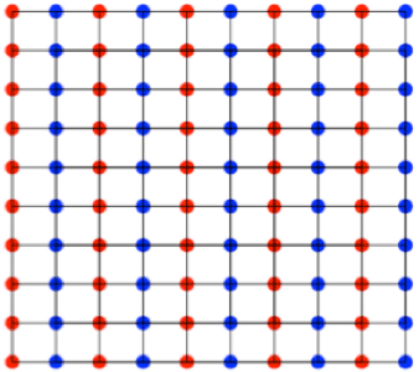
builtcode
THE COGNITIVE COMPANY.

O que é a maldição da dimensionalidade?

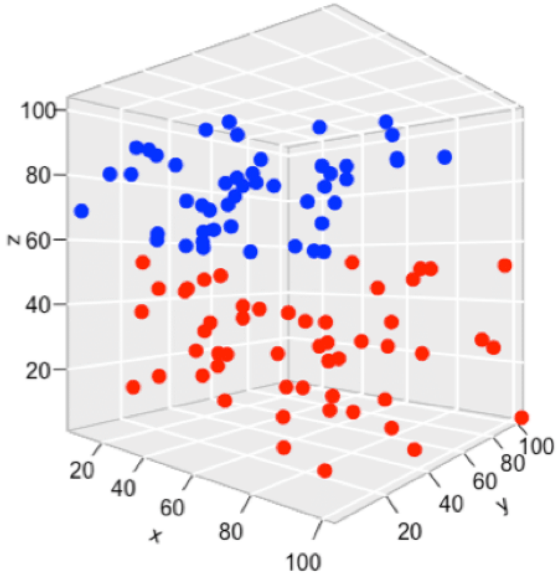
THE DEVELOPER'S CONFERENCE



(A) 1-D



(B) 2-D



(C) 3-D

builtcode
THE COGNITIVE COMPANY.

O que é a maldição da dimensionalidade?

THE DEVELOPER'S CONFERENCE

We select only useful features.

https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2

builtcode
THE COGNITIVE COMPANY.

O que é a maldição da dimensionalidade?

THE DEVELOPER'S CONFERENCE

Modelo **1**
2 Features

Acurácia: 57%

Modelo **2**
16 Features

Acurácia: 83%

Modelo **3**
128 Features

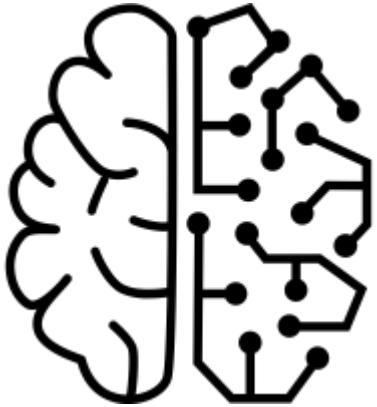
Acurácia: 85%

Modelo **4**
1024 Features

Acurácia: 79%

builtcode
THE COGNITIVE COMPANY.

O que queremos?

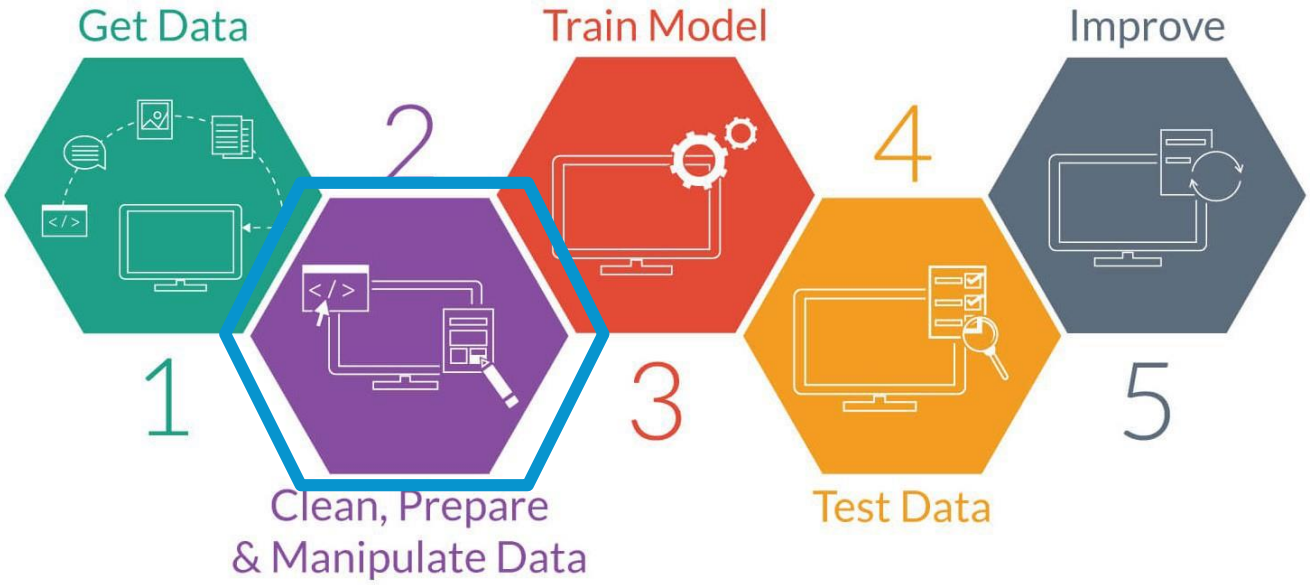


Modelo **5**
37 Features

Acurácia: 90%

Treinamento de Modelos de Machine Learning

THE DEVELOPER'S CONFERENCE



builtcode
THE COGNITIVE COMPANY.

Redução da Dimensionalidade

THE DEVELOPER'S CONFERENCE

Análise de Componentes Principais

Dataset 1

Seq	Host
ATGTTTGTTTTGCTTGTTGCATATGCCTTGTTGCATATTGCTGGTT...	human
ATGTTTTTGATACTTTTAATTTCCTTACCAATGGCTTTTGCTGTTA...	human
ATGTTTATTTTCTTATTATTTCTTACTCTCACTAGAGGTAGTGACC...	human
ATGACGCCTTTAATTTACTTCTGGTTGTTCTTACCAGTACTTCTAA...	porcine
ATGAAGTCTTTAACTTACTTCTGGTTGTTCTTACCAGTACTTTCAA...	porcine
ATGCAGAGAGCTCTATTGATTATGACCTTACTTTGTCTCGTTCGAG...	porcine
ATGTTTTTGATACTTTTAATTTCCTTACCAACGACTTTTGCTGTTA...	bovine
ATGAAACTTTTTATAGTTTTTTGTGCTCCTTTTTTAGGGTGTGTTATT...	bat
ATGTTGGTGAAGTCACTGTTTTTTAGTGACTCTTTTGTTTGCACTAT...	avian
ATGTTGGTAACACCTCTTTTTATTAGTGACTCTTTTGTTTGCACTAT...	avian

730 rows

builtcode
THE COGNITIVE COMPANY.

Redução da Dimensionalidade

THE DEVELOPER'S CONFERENCE

Análise de Componentes Principais

Feature Vector

Class

Dataset 1

	G	AG	GG	CG	TG	TT	Host
Sample → 0	0.665225	-0.304268	-0.281132	1.510045	2.153277	1.823231	human
36	-2.466542	-0.859019	0.530849	-0.118405	-1.293848	-1.017817	human
72	-0.254466	-0.433647	-0.162090	-0.915341	-0.778621	-1.161352	human
108	-1.376046	-1.309138	-0.711194	0.430361	-0.279186	-0.342964	human
144	-0.989784	-0.582103	0.818058	-0.970409	1.390465	1.075697	human
180	-0.388242	-0.212588	-0.247272	0.758464	-0.598655	-0.412604	human
216	1.176985	-0.129859	-1.289674	0.330971	-0.855900	-0.671098	porcine
252	1.072384	-0.110824	-1.039439	-0.308360	-0.846077	-0.705007	porcine
288	1.135144	-0.065912	-0.983338	-0.327244	-0.835327	-0.692845	porcine
324	1.093304	-0.050562	-1.063914	-0.351032	-0.852434	-0.618089	porcine
360	-1.194676	1.171019	1.793185	0.959133	2.159648	1.867804	porcine
396	-0.022611	-0.617248	0.096755	-1.050849	-0.861156	-1.175692	bovine
432	-0.043884	-0.613164	0.126130	-1.517958	-0.737744	-1.247607	bovine
468	-0.212750	0.626935	-0.250648	0.604421	1.079619	1.205546	bat
504	-0.229375	-0.973427	-0.557187	0.780419	0.258101	-0.322454	bat
540	1.798511	-0.072512	-1.286707	1.121890	-0.186046	-0.351052	murine
576	-0.382002	1.768964	1.532858	0.815515	0.521813	1.266507	avian
612	-0.288020	1.730763	1.722878	0.087307	0.462897	1.034374	avian
648	-0.335804	1.652483	1.683257	0.165865	0.520066	0.872854	avian
684	-0.238121	1.729452	1.214314	0.326928	0.544972	1.038252	avian
720	-2.321768	1.091406	2.558835	0.117995	1.634372	2.089170	avian

730 rows x 19 cols

n = 730 samples

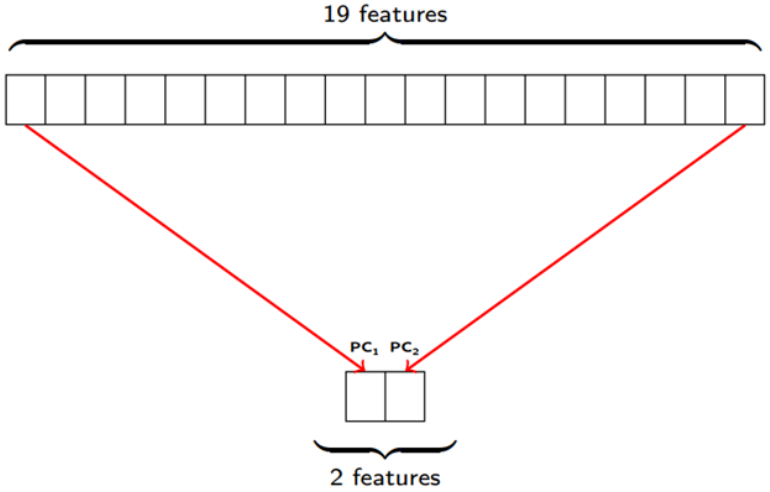
Dimension: d = 19 features

builtcode
THE COGNITIVE COMPANY.

Redução da Dimensionalidade

THE DEVELOPER'S CONFERENCE

Análise de Componentes Principais



Dataset 1

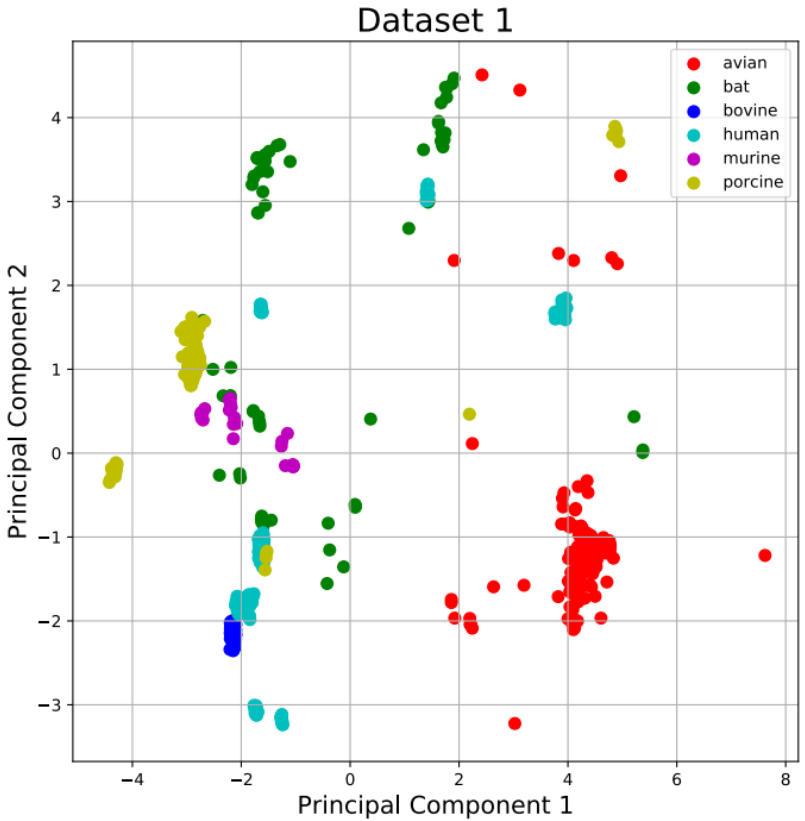
	PC 1	PC 2	Host
0	3.775624	1.613269	human
36	-1.254264	-3.114602	human
72	-1.845260	-1.941322	human
108	-1.657282	-0.982642	human
144	1.436796	3.099578	human
180	-1.643563	1.725470	human
216	-2.941701	1.510518	porcine
252	-2.935069	1.062321	porcine
288	-2.914471	0.908152	porcine
324	-2.912731	0.903012	porcine
360	4.935531	3.713215	porcine
396	-2.124629	-2.173583	bovine
432	-2.178364	-2.104020	bovine
468	1.622693	3.955135	bat
504	-1.699718	2.866013	bat
540	-1.049189	-0.162940	murine
576	4.694584	-1.157702	avian
612	4.440957	-1.328387	avian
648	4.228678	-0.941589	avian
684	4.290554	-1.297679	avian
720	4.907956	2.258447	avian

730 rows x 2 cols

Redução da Dimensionalidade

THE DEVELOPER'S CONFERENCE

Análise de Componentes Principais



builtcode
THE COGNITIVE COMPANY.

Redução da Dimensionalidade

Análise de Componentes Principais

- Mas e se temos características inúteis/ruins em nosso Dataset?

Seleção de Características

THE DEVELOPER'S CONFERENCE

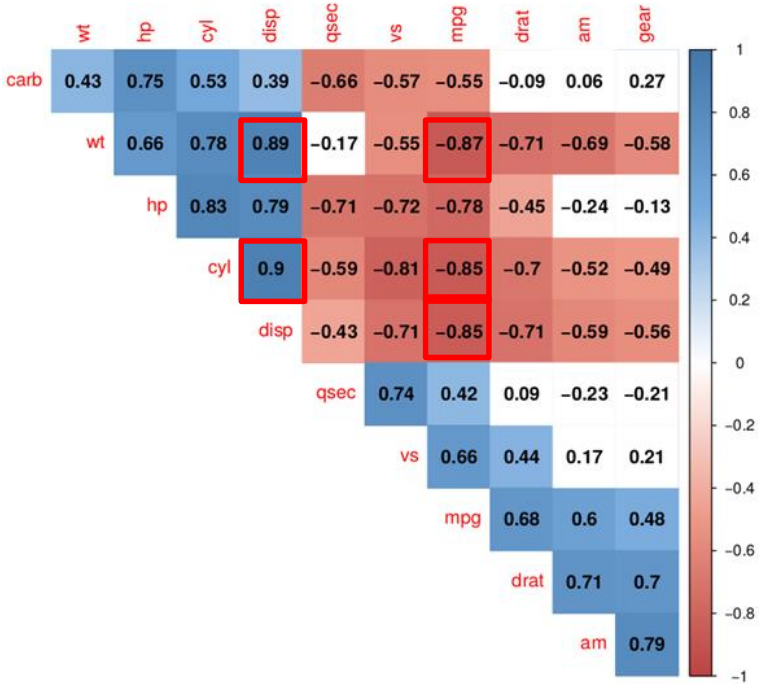
- Métodos de Filtro (*filter methods*)
- Métodos Embutidos (*embedded methods*)
- Métodos Invólucros (*wrapper methods*)

builtcode
THE COGNITIVE COMPANY.

Métodos de Filtros

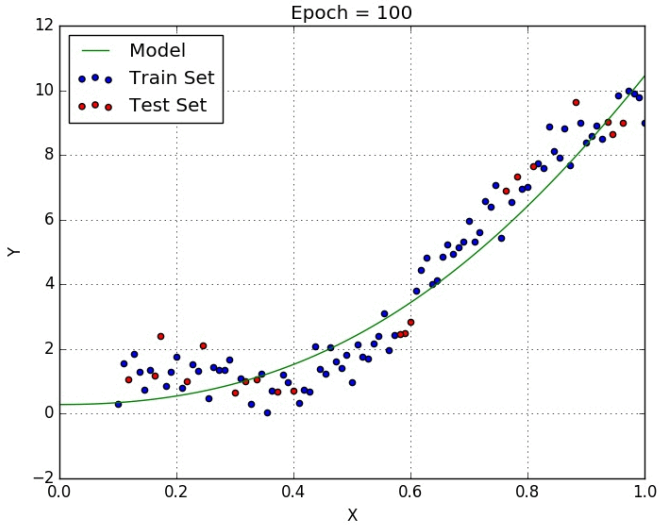
THE DEVELOPER'S CONFERENCE

Correlação de Pearson



builtcode
THE COGNITIVE COMPANY.

Métodos de Embutidos



LASSO, Elastic Net, Ridge Regression

THE DEVELOPER'S CONFERENCE

RANDOM FOREST

LightGBM

XGBoost

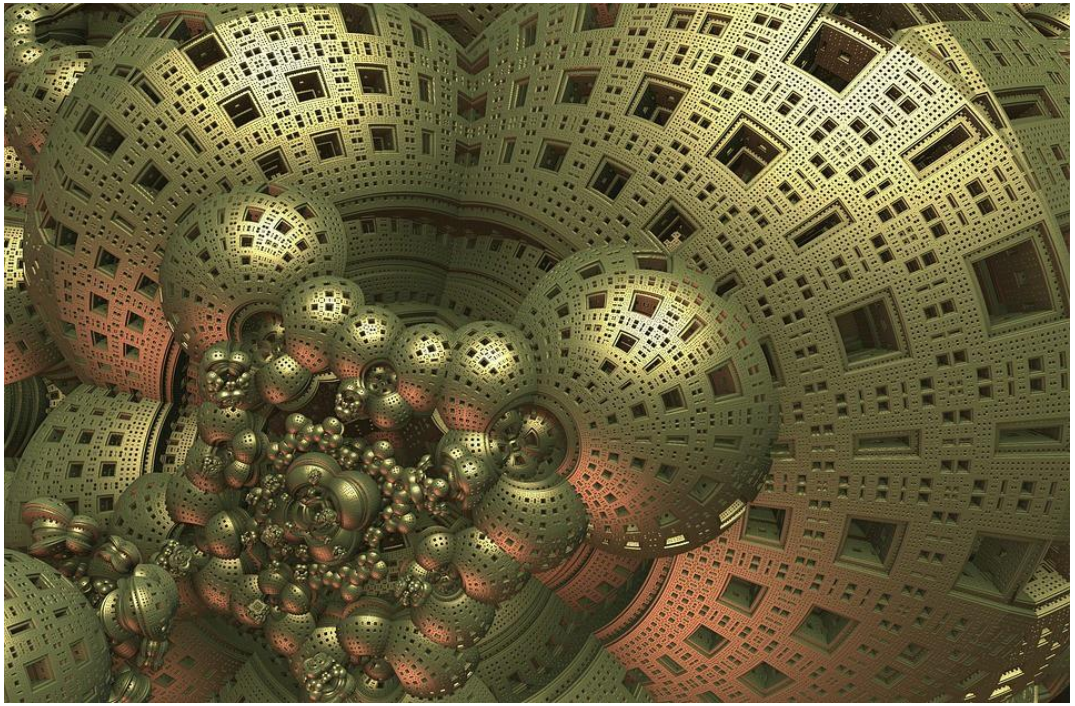
builtcode
THE COGNITIVE COMPANY.

Métodos Invólucros (*wrapper methods*)

THE DEVELOPER'S CONFERENCE

Recursive Feature Elimination

- Para *n* features temos 2^n combinações!



builtcode
THE COGNITIVE COMPANY.

Métodos Invólucros

THE DEVELOPER'S CONFERENCE

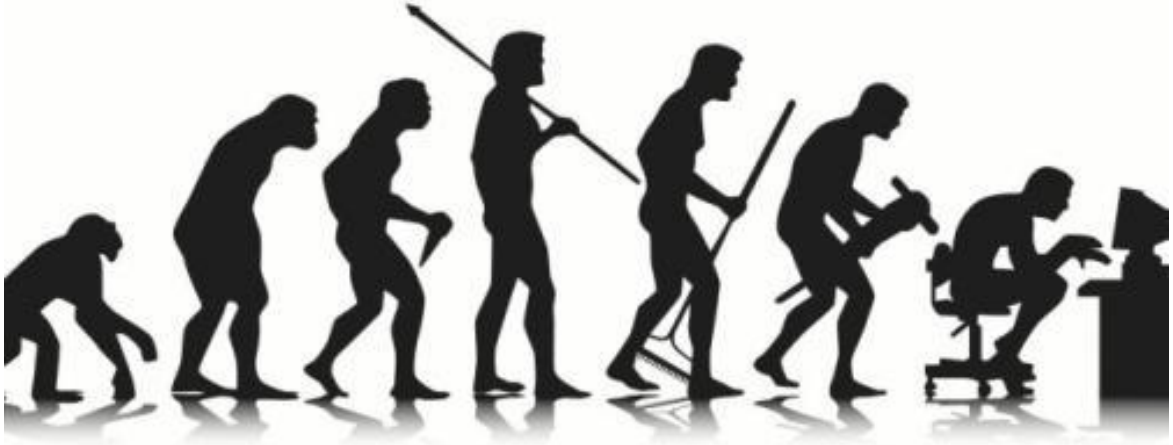
E agora?

Heurísticas!

builtcode
THE COGNITIVE COMPANY.

Algoritmos Genéticos

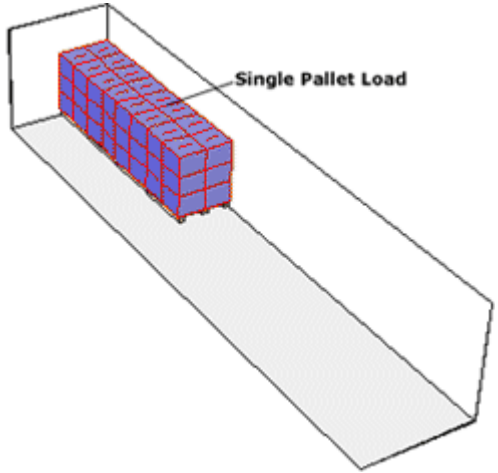
THE DEVELOPER'S CONFERENCE



builtcode
THE COGNITIVE COMPANY.

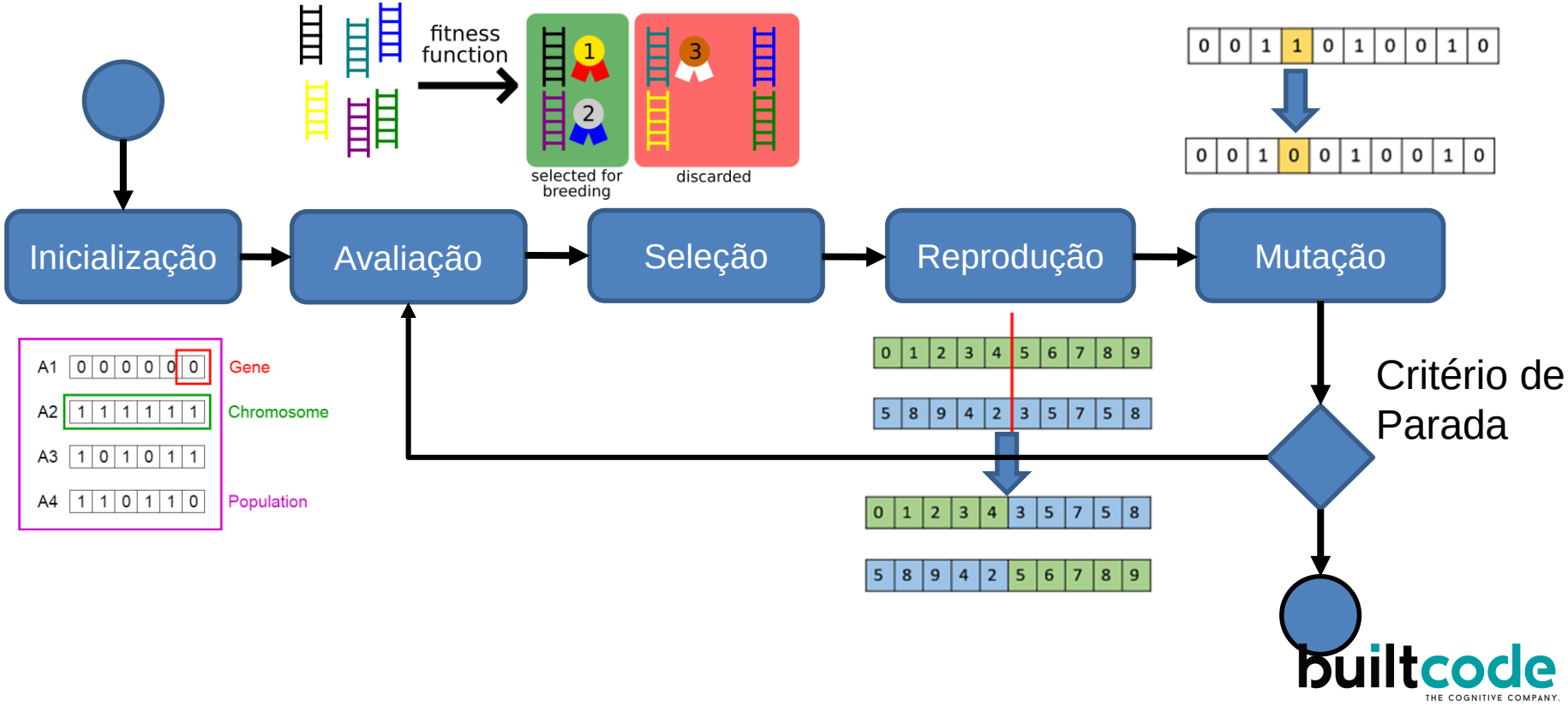
Algoritmos Genéticos

THE DEVELOPER'S CONFERENCE



builtcode
THE COGNITIVE COMPANY.

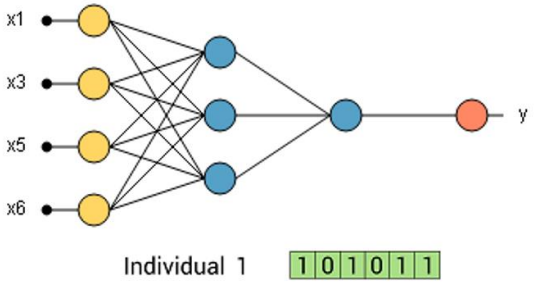
Algoritmos Genéticos

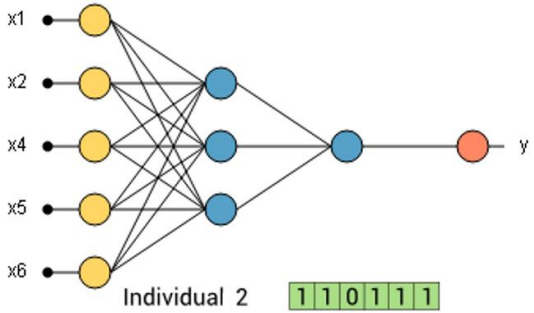


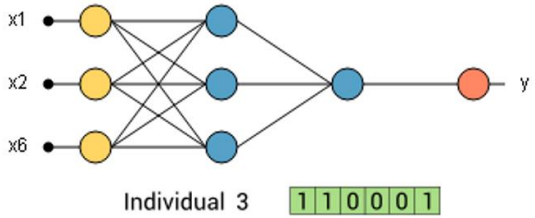
Algoritmos Genéticos

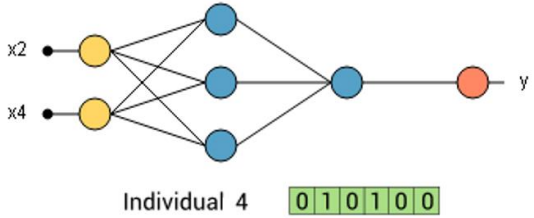
THE DEVELOPER'S CONFERENCE

- Codificação da Solução









Fonte: https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection

Algoritmos Genéticos

THE DEVELOPER'S CONFERENCE

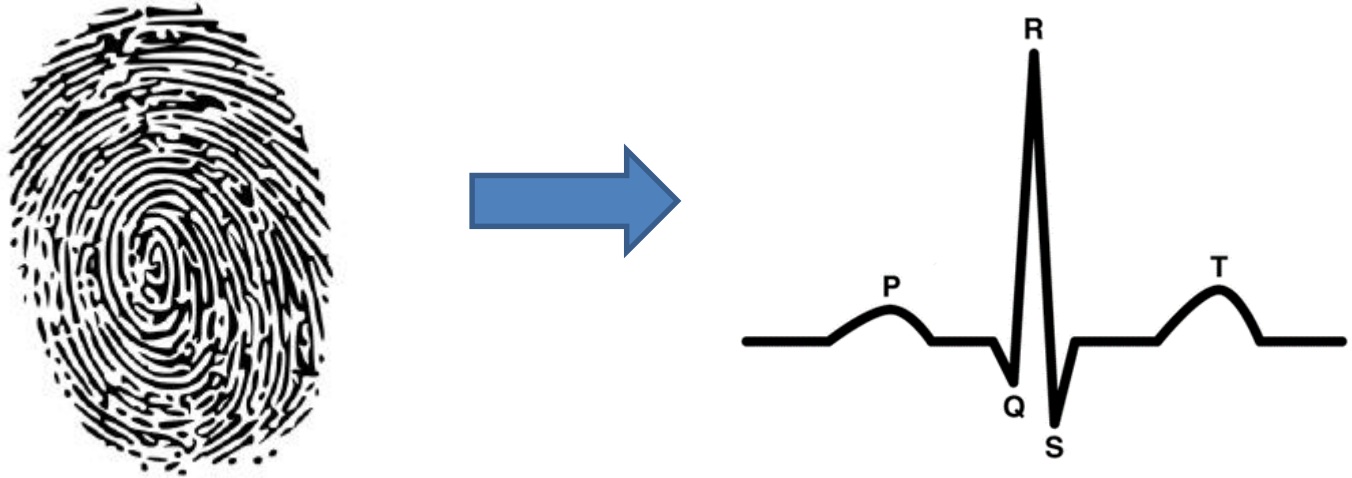
- Função de Avaliação:
  - Seleciona as características do cromossomo e faz validação cruzada (*cross validation*) e obtém a acurácia, usada como critério de avaliação do GA.

```
def calculate_fitness(individual):
        np_ind = np.asarray(individual)
        if np.sum(np_ind) == 0:
            return  (0.0,)
        else:
            feature_idx = np.where(np_ind==1)[0]
            x_temp = X[:,feature_idx]
        cv_set = np.repeat(-1.,x_temp.shape[0])
        skf = StratifiedKFold(n_splits = 5)
        for train_index,test_index in skf.split(x_temp,y):
            X_train,X_test = x_temp[train_index],x_temp[test_index]
            y_train,y_test = y[train_index],y[test_index]
            if X_train.shape[0] != y_train.shape[0]:
                raise Exception()
            classifier.fit(X_train,y_train)
            predicted_y = classifier.predict(X_test)
            cv_set[test_index] = predicted_y
        acc = accuracy_score(y, cv_set)
        return (acc,)
```

Seleção de Características com GA – Case I

THE DEVELOPER'S CONFERENCE

- Problema: Biometria utilizando batimentos cardíacos:

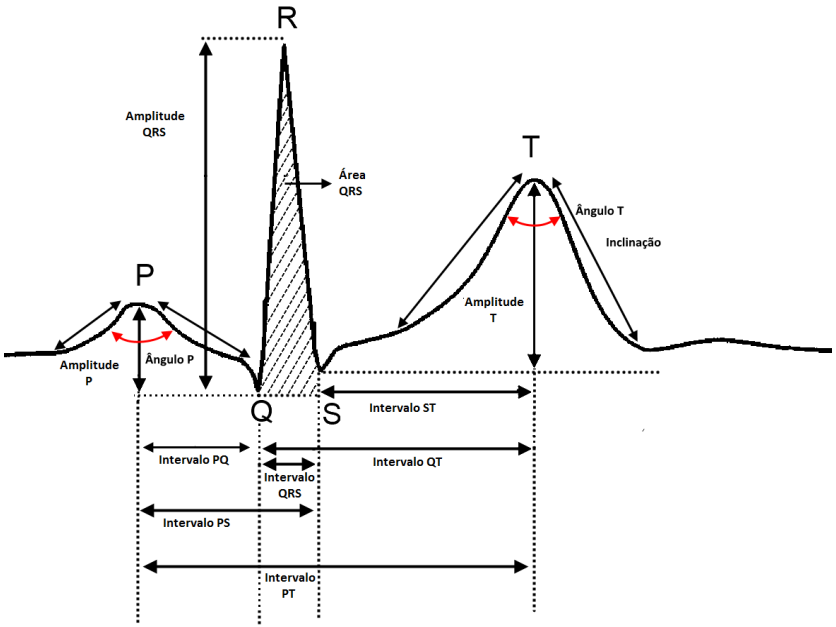


builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case I

THE DEVELOPER'S CONFERENCE

- Extração de Características:



- Características do domino do tempo
- Transformada Discreta Cosseno
- Transformada de Fourier
- Função de Autocorrelação
- Modelo Autoregressivo
- Codificação Linear Preditiva
- Transformada Pulso Ativo
- Representação Linear por partes
- Polinômios de Hermite
- Coecientes Mel-cepstrais
- Transformada Wavelet
- Métodos de estimativa da Dimensão Fractal
- Decomposição do Modo Empírico

754 características extraídas

builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case I

THE DEVELOPER'S CONFERENCE

- Conjunto de Dados:
  - PTB ECG Database;
  - 290 pessoas distintas (logo 290 classes);
  - 754 características extraídas;
  - O numero de gravações por pessoa varia de 2 a 20, aproximadamente 2000 amostras.

builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case I

THE DEVELOPER'S CONFERENCE

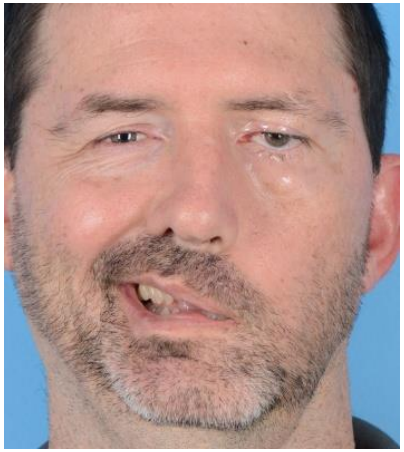
- Resultados:
  - Acurácia de **97,93%** no conjunto de teste, redução de **754** características para **31** após o término da execução;
  - Classificador Optimum-Path Forest utilizado na função fitness;
  - Artigo publicado em: https://ieeexplore.ieee.org/document/7966216

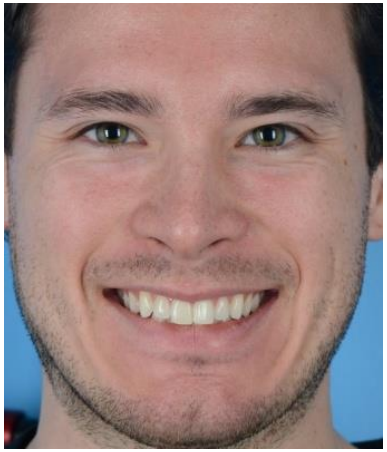
builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case II

THE DEVELOPER'S CONFERENCE

- Problema: Classificação de severidade de paralisia facial
  - Escala de House-Brackmann (de 1 a 5 em grau de severidade)



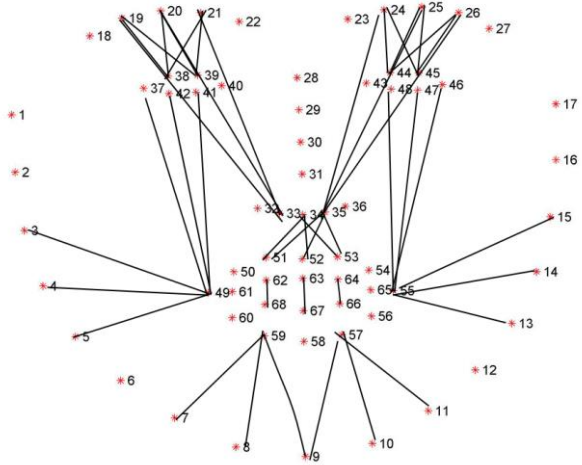


Imagens do Dataset Sir **Charles Bell Society** (SCBS) para paralisia facial

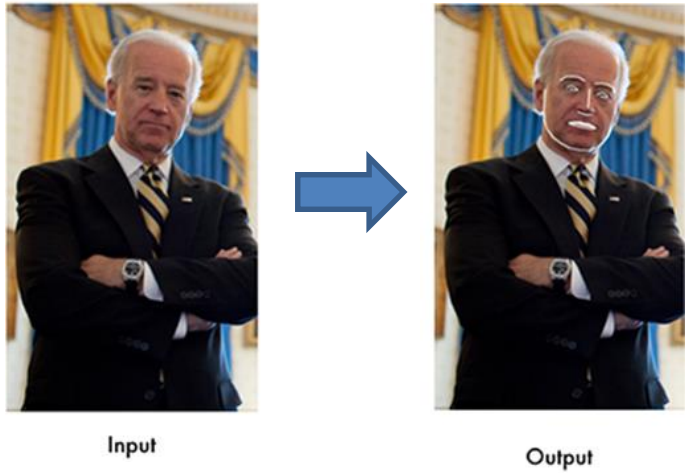
Seleção de Características com GA – Case II

THE DEVELOPER'S CONFERENCE

- Extração de Características



Características 2D

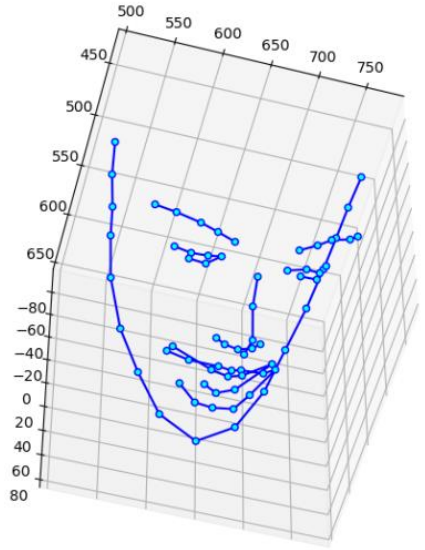


Características Biométricas de Face

Seleção de Características com GA – Case II

THE DEVELOPER'S CONFERENCE

- Extração de Características



Características 3D

builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case II

THE DEVELOPER'S CONFERENCE

- Conjunto de dados e configuração dos experimentos:
  - 1202 faces obtidas do dataset Sir **Charles Bell Society** (SCBS) e de pacientes do Hospital das Clínicas em São Paulo ;
  - 5 classes;
  - 1451 características extraídas.
  - Implementação utilizando a biblioteca DEAP para otimização.

builtcode
THE COGNITIVE COMPANY.

Seleção de Características com GA – Case II

THE DEVELOPER'S CONFERENCE

- Resultados preliminares:
  - 1451 características extraídas.
  - Acurácia inicial de ~=60% utilizando Correlação de Pearson e RFE (com Random Forest como classificador);
  - **78%** de acurácia após o término da execução com **162** características e classificador SVM-RBF.

builtcode
THE COGNITIVE COMPANY.

Conclusão

THE DEVELOPER'S CONFERENCE

- GA para seleção de características em problemas de alta dimensionalidade funcionam muito bem;
- Diversos cases no mercado e na literatura;
- Possibilidade de utilização de outras Meta-Heurísticas como otimização de exame de partículas, colônias de formigas e algoritmos meméticos.

builtcode
THE COGNITIVE COMPANY.

THE DEVELOPER'S CONFERENCE

Quer fazer parte de um time inquieto?

builtcode

THE COGNITIVE COMPANY.

Obrigado!

THE DEVELOPER'S CONFERENCE

https://www.linkedin.com/in/felipe-teodoro-87b25217/

https://github.com/felipesteodoro

builtcode
THE COGNITIVE COMPANY.

THE DEVELOPER'S CONFERENCE