



PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS

Programa de Pós-Graduação em Informática

Bootcamp: Titanic - Machine Learning from Disaster

.....

Escola de férias

André de Sousa Araújo

BH, 2018

Quem sou eu?

André de Sousa Araújo

Profissão?

Desenvolvedor de software e sócio da  **superbuy**.com.br

Atualmente

Entusiasta em Aprendizado de Máquina e Inteligência Artificial.
Mestrando em Análise de Dados, Descoberta de Conhecimento e Recuperação de Informação

Contato

github.com/dedeco

linkedin.com/in/dedeco

twitter.com/dedecu

<https://stackoverflow.com/users/2452792/andre-araujo>

dedeco@gmail.com

Bootcamp?

boot camp

noun [C usually sing] • **US**  /'but ,kæmp/

★ a place where new members of the US military receive their first training

“Na área de tecnologia, podemos dizer **que é um treinamento prático e imersivo** “

Teremos:

- Conceitos rápidos e enxutos;
- **Exercícios guiados;**
- E ao final um resultado simples projeto de data science executado;
- Ritmo médio (seguindo a média da turma)

NÃO terá:

- Explicações detalhadas e conceituais sobre tudo que for abordado;
- Atenção personalizada para 1 pessoa ou um pequeno grupo;
- Debates

Programação

Aquecimento:

- **Conceito:** Introdução a Redes neurais
- **Exercício:** Minha primeira rede neural - UCI ML hand-written digits

~ 1 hora

Intervalo

~ 15 minutos

Imersão:

- **Apresentando o Kaggle**
- **Dataset:** Titanic: Machine Learning from Disaster

Prática:

- **Etapas:**
 - Exploração dos dados
 - Pré-processamento
 - Mineração dos dados

~ 2 horas

Intervalo

~ 15 minutos

Conclusão:

- **Discussão dos resultados**
- **Submissão**
- **Perguntas e dúvidas**

~ 30 minutos

Introdução a redes neurais

- Resumidamente em uma frase:

"São modelos computacionais inspirados no cérebro humano, que a partir dos dados de entrada conseguem **aprender um padrão**; e após o treinamento, são capazes de reconhecer, classificar ou prever um dado, evento ou comportamento"

1 Aquecimento

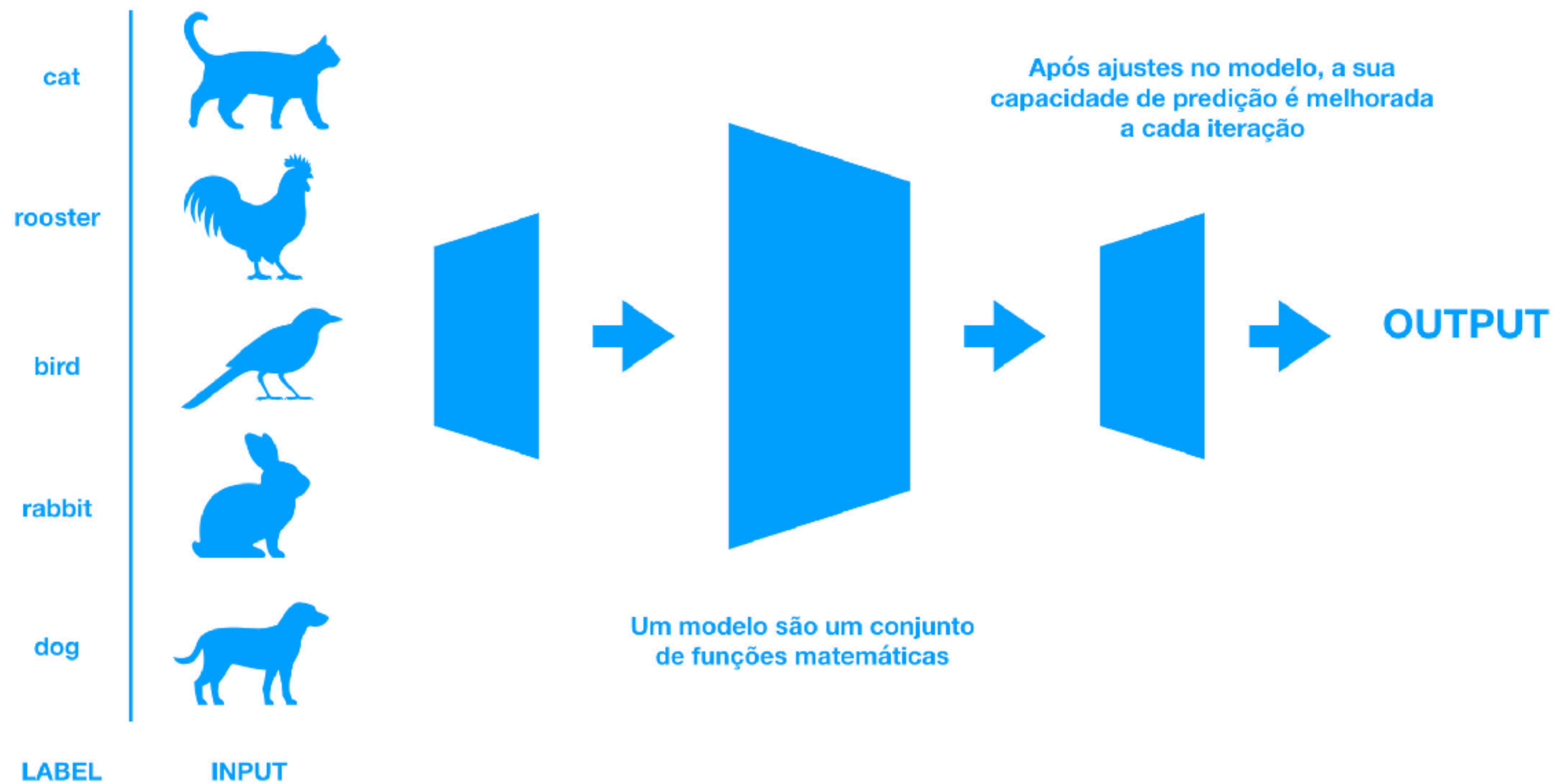
2 Imersão

3 Prática

4 Conclusão

Redes neurais

Passo 1: Treinamento



Redes neurais

- Após o modelo ser treinado.

Se apresentarmos uma nova imagem, sem rótulo, que nunca foi apresentada antes, o modelo deverá ser capaz de classificar esta nova imagem como um cachorro, por exemplo.

1 Aquecimento

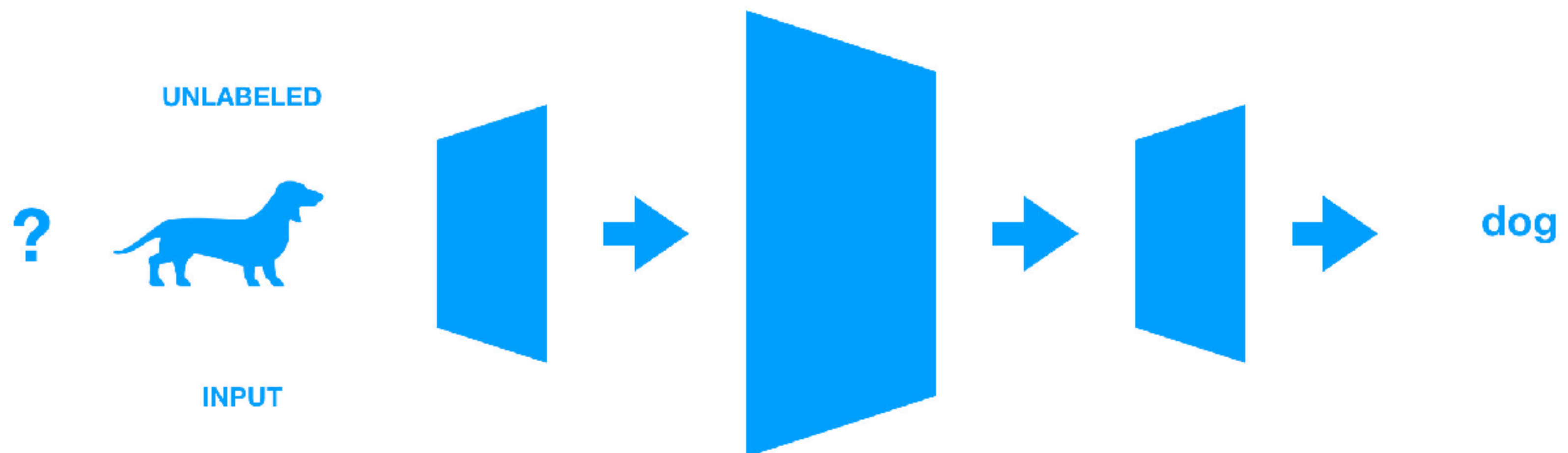
2 Imersão

3 Prática

4 Conclusão

Redes neurais

Passo 2: Classificar utilizando o modelo



1 Aquecimento

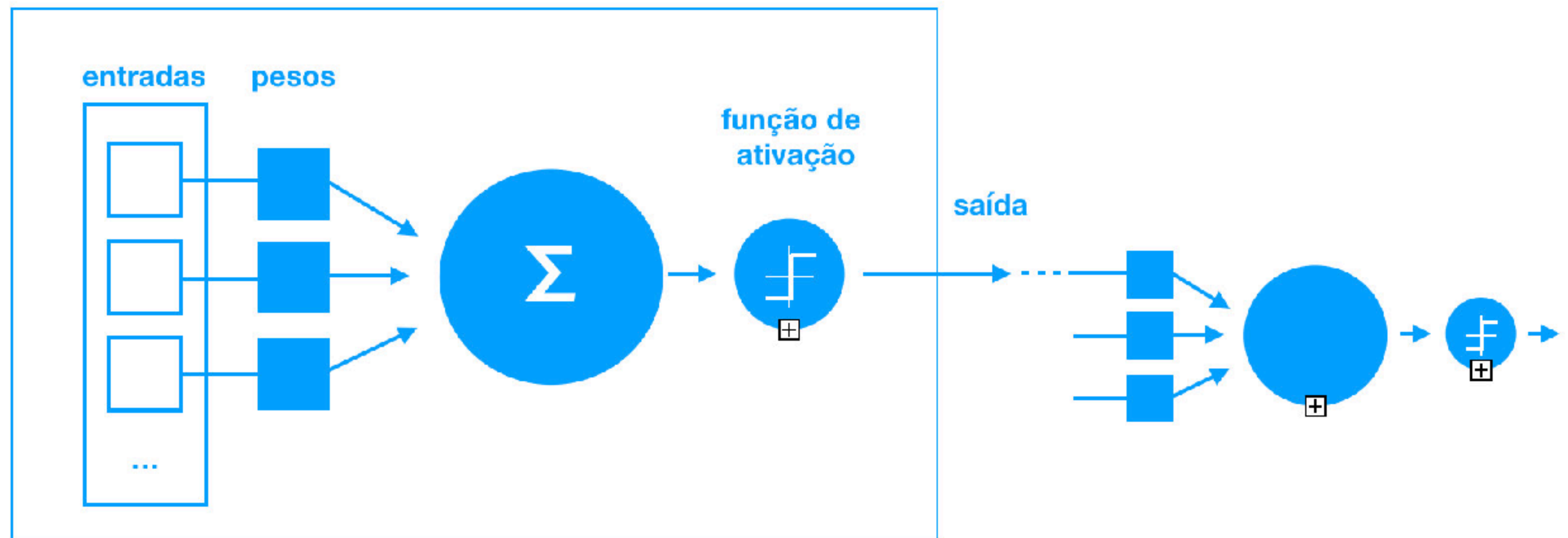
2 Imersão

3 Prática

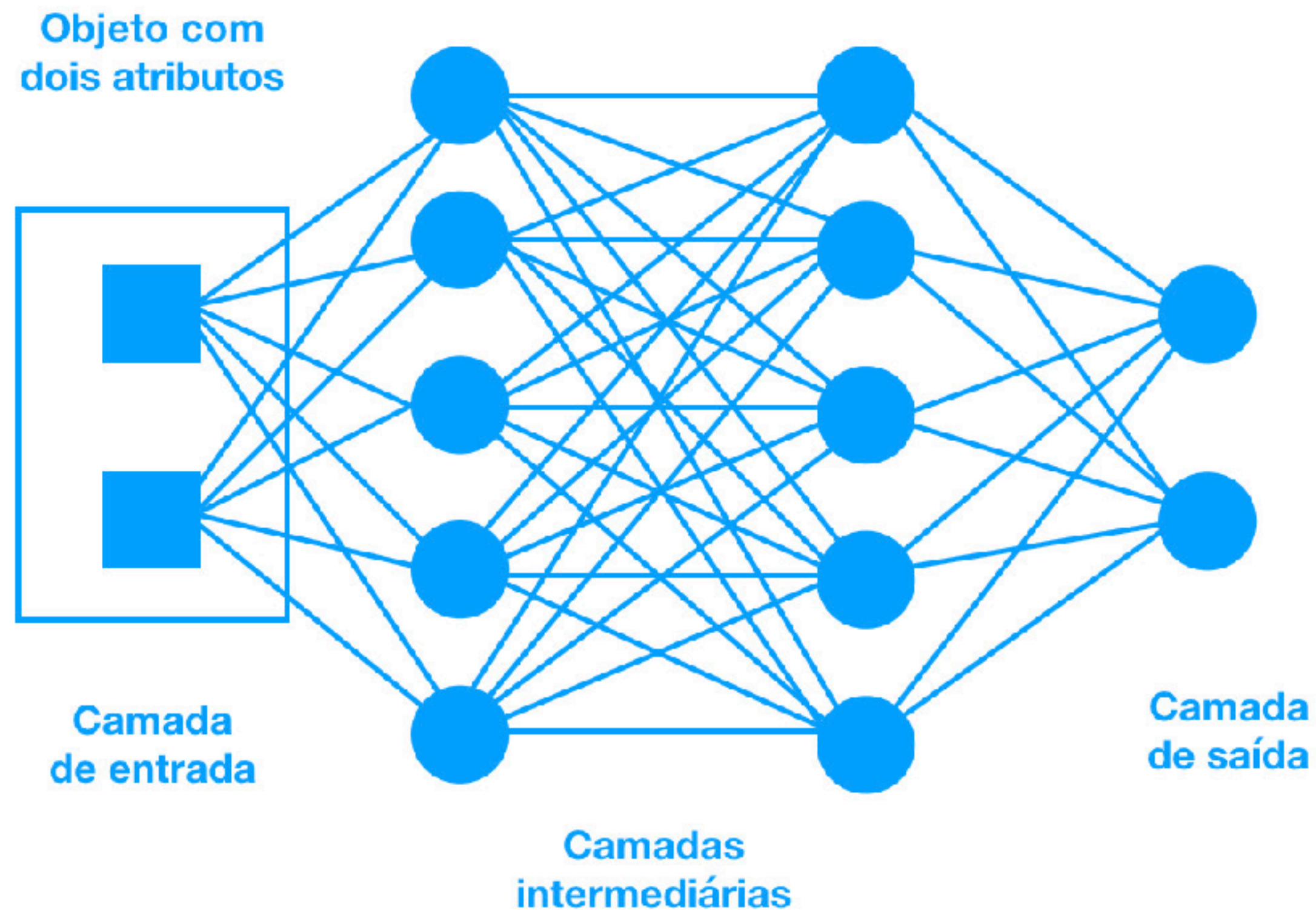
4 Conclusão

Neurônio artificial

NEURÔNIO ARTIFICIAL



Uma rede neural multicamada típica



Playground Tensorflow

- A proposta aqui não é aprofundar matematicamente, e sim apenas dar uma visão geral. Assim, uma boa opção é ver graficamente como funciona, o Google criou uma funcionalidade muito legal, na qual é possível **brincar e simular graficamente uma rede neural**, veja o vídeo a seguir:
- Link: <http://playground.tensorflow.org>



Exercício - Kernel 01

- Existe um dataset que é chamado do **Hello World** do Machine Learning, é o *Optical Recognition of Handwritten Digits Data Set* (NIST). Vamos usá-lo para mostrar basicamente como é.
- **Objetivo:**
 - Classificar cada imagem do dataset como um número 0 até 9.
- São 5620 imagens escritas a mão, o dataset possui 64 atributos para cada imagem. Na verdade, cada atributo é um pixel, assim cada imagem tem 8x8 pixels, e cada pixel valorado de 1 até 16 .



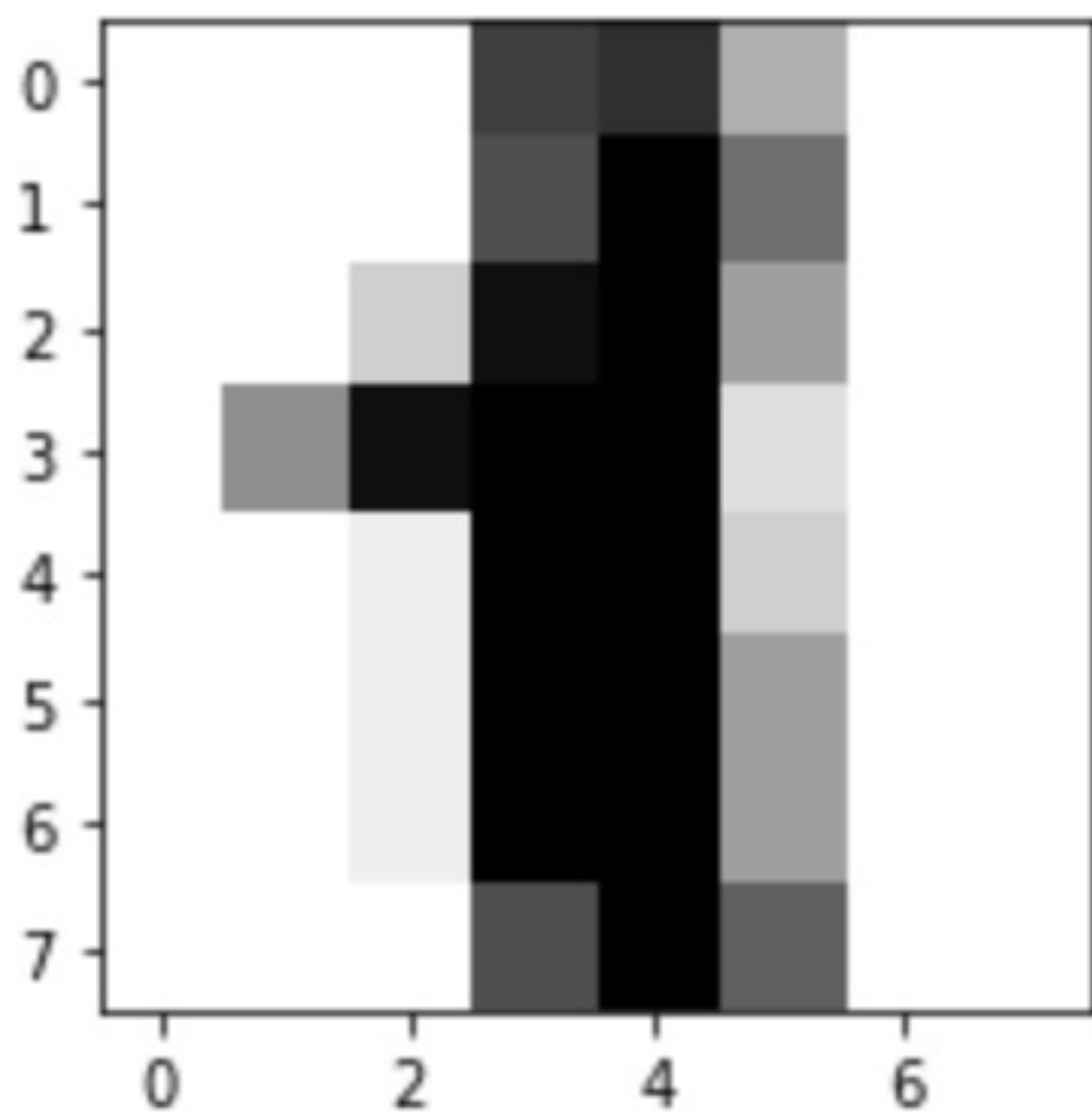
1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Exercício



```
[ 0.  0.  0. 12. 13.  5.  0.
 0.  0.  0.  0. 11. 16.  9.  0.
 0.  0.  0.  3. 15. 16.  6.  0.
 0.  0.  7. 15. 16. 16.  2.  0.
 0.  0.  0.  1. 16. 16.  3.  0.
 0.  0.  0.  1. 16. 16.  6.  0.
 0.  0.  0.  1. 16. 16.  6.  0.
 0.  0.  0.  0. 11. 16. 10.  0.
 0.]
```

Vá na pasta Exercicios >> Kernel 01



1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Kaggle

kaggle



1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Titanic

O naufrágio do RMS Titanic é um dos mais infames naufrágios da história. Em 15 de abril de 1912, durante sua viagem inaugural, o Titanic afundou depois de colidir com um iceberg, matando **1502 de 2224 passageiros e tripulantes**.

Uma das razões pelas quais o naufrágio causou tal perda de vida foi que não havia botes salva-vidas suficientes para os passageiros e a tripulação. Embora houvesse algum elemento de sorte envolvido na sobrevivência do naufrágio, **alguns grupos de pessoas tinham maior probabilidade de sobreviver do que outros, como mulheres, crianças e a classe alta.**

Neste desafio, pedimos que você conclua a análise de quais tipos de pessoas provavelmente sobreviveriam.

Em particular, pedimos que você aplique as ferramentas de aprendizado de máquina para prever quais passageiros sobreviveram à tragédia.

Link: <https://www.kaggle.com/c/titanic>



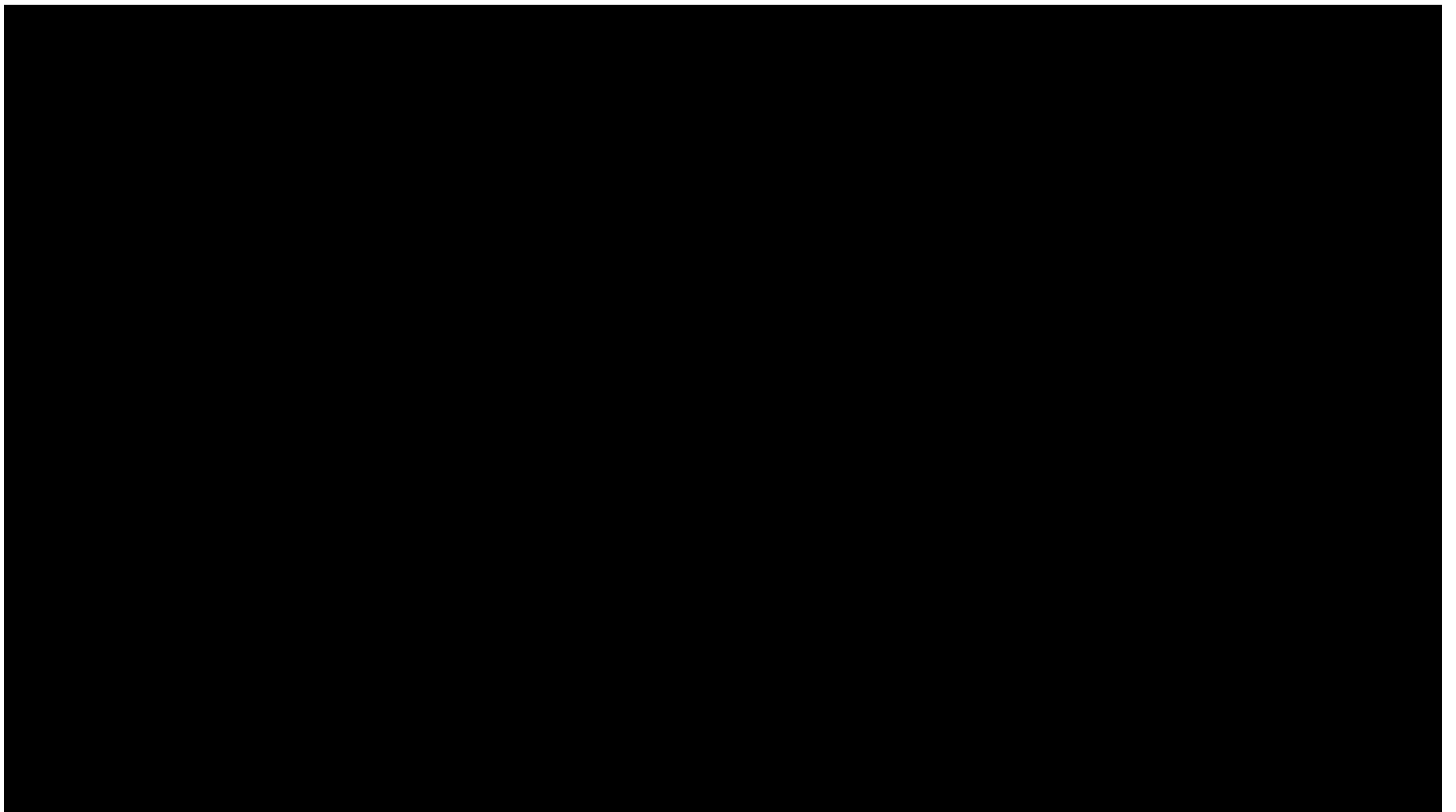
1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Titanic



1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Titanic Dataset

Problema: Classificação binária

Porque? Porque temos que classificar somente em 2 classes:

- Sobrevivente
- Não sobrevivente



Exercício - Kernel 02

A mais importante etapa:

A etapa de **Pré-processamento** vai variar de acordo com a metodologia. Como nosso foco aqui é mais prático, e estamos atacando um problema mais simples. Basicamente esta etapa consiste em toda preparação antes de aplicar o algoritmo.

Vale lembrar que normalmente esta etapa consome **a maior parte de um projeto de *data science***, e algumas metodologias (e deve ser assim) consideram uma etapa anterior de modelagem dos dados e entendimento do problema (não é caso, pois temos a base já modelada)

Vamos considerar nesta etapa:

- Tratamento de dados ausentes
- Criação de novas variáveis a partir das existentes (*Feature Engineering*)
- Limpeza dos dados

Vá na pasta Exercicios >> Kernel 02



1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Exercício - Kernel 03

Finalmente a aplicação do algoritmo:

O algoritmos a ser escolhido poderia ser qualquer algoritmo de classificação: arvores de decisão, florestas randômicas, clusterização pelos vizinhos mais próximos (KNeighborsClassifier), etc.

Geralmente esta escolha é embasada e faz parte da modelagem dos dados e entendimento do problema, mas aqui como foco é prático, vamos aplicar um rede neural multicamada (MLPClassifier).

Vá na pasta Exercicios >> Kernel 03



Matriz de confusão

Se um sistema de classificação foi treinado para distinguir entre gatos, cães e coelhos, uma matriz de confusão resumirá os resultados do teste do algoritmo para uma inspeção adicional.

Assumindo uma amostra de 27 animais - 8 gatos, 6 cães e 13 coelhos, a matriz de confusão resultante pode se parecer com a tabela abaixo:

		Actual class		
		Cat	Dog	Rabbit
Predicted class	Cat	5	2	0
	Dog	3	3	2
	Rabbit	0	1	11

Tabela de confusão

Assumindo a matriz de confusão acima, sua tabela de confusão correspondente, para a classe cat, seria:

		Actual class	
		Cat	Non-cat
Predicted class	Cat	5 True Positives	2 False Positives
	Non-cat	3 False Negatives	17 True Negatives

[True positive rate](#) (TPR), [Recall](#), [Sensitivity](#),
probability of detection =
$$\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$$

[False positive rate](#) (FPR), [Fall-out](#),
probability of false alarm =
$$\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$$

[False negative rate](#) (FNR), Miss rate =
$$\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$$

[True negative rate](#) (TNR), [Specificity](#) (SPC) =
$$\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$$

1 Aquecimento







2 Imersão

3 Prática

4 Conclusão

Submissão - Kaggle

Submetendo os resultados no Kaggle:

6266	new	m9zjl		0.77511	1	13h
6267	▲ 2323	martin31		0.77511	5	9h
Sun Apr 29 2018 16:35:04 GMT-0300 (-03)				0.77511	1	now
6268	new	Andre Araujo				
Your Best Entry ↑						
Your submission scored 0.77511, which is not an improvement of your best score. Keep trying!						
6269	▼ 630	Sanjiv Patel		0.77033	4	2mo
6270	▼ 630	callinew		0.77033	6	2mo



1 Aquecimento

2 Imersão

3 Prática

4 Conclusão

Obrigado

"O sucesso é ir de fracasso em fracasso sem perder entusiasmo."

Winston Churchill