

DATA MASTERS

GUIA DE ESTUDOS
CIENTISTA DE DADOS



GUIA DATA MASTER





CIENTISTA DE DADOS

Olá, este é o **Guia Data Masters de Cientista de Dados!** Aqui, você encontrará uma seleção de conteúdo para auxiliá-lo na sua certificação.

A função do/a Cientista de Dados é solucionar problemas de negócios usando dados. Mais do que um simples `.fit()/.predict()`, o cientista de dados deve saber como limpar os dados, fazer uma boa análise descritiva, selecionar variáveis, dentre outras atribuições.

Sabemos que cada pessoa tem sua preferência em como estudar. Por isso, buscamos neste guia, disponibilizar conteúdos em formatos variados como vídeos, artigos, cursos, livros etc. Esta diversidade de materiais lhe proporcionará um aprendizado incrível.



Antes de começar, sugerimos que faça uma autoavaliação de seus conhecimentos, veja na próxima página a lista de temas que abordaremos ao longo deste guia.



Com estes conhecimentos, você será capaz de exercer a função de um Cientista de Dados!

Mas você sabe qual a principal função deste cargo?

É analisar e solucionar problemas complexos, e atrelado a isso existe uma série de pré requisitos, dentre eles estão: a limpeza, a estruturação, a organização e a preparação de dados.

Vamos aos estudos então?

Esperamos que aproveite esta jornada ao máximo e conquiste sua Certificação!

SUMÁRIO

- 04** Fundamentos de Big Data
- 06** Regressão
- 08** Classificação
- 10** Agrupamento
- 12** Estatística
- 14** Programação
- 16** Teoria do Aprendizado
- 18** Materiais Complementares
- 19** Roteiro de Avaliação das Bancas
- 20** Dicas dos Experts

TRILHA DE FORMAÇÃO

Na Academia Santander, temos trilhas disponíveis para impulsionar seus estudos e potencializar o seu conhecimento sobre a carreira de Ciência de Dados. Aproveite essa oportunidade!



[Trilha de aprendizagem Cientista de dados - Badge Driven \(Oficial\)](#)

[Trilha de aprendizagem Cientista de dados - Badge Driven](#)



[Trilha de aprendizagem Cientista de dados - Badge Advanced \(Oficial\)](#)

[Trilha de aprendizagem Cientista de dados - Badge Advanced](#)



[Trilha de Aprendizagem Cientista de Dados Badge Expert \(Oficial\)](#)


[Trilha de Aprendizagem Cientista de Dados Badge Expert](#)


FUNDAMENTOS DE BIG DATA

- HDFS
- Cloud Computing x Hadoop
- YARN
- Hadoop
- Map Reduce
- Data Structures
- Data Storage
- Data Processing
- Big Data
- ETL





[Entenda o que é a modelagem de Banco de Dados](#) 

[Introdução a SQL: Consulta e gerenciamento de dados](#) 

[O que é Big Data - Conceitos básicos](#) 
18min


[Comandos básicos em SQL - INSERT, UPDATE, DELETE e SELECT](#) 


[SQL Tutorial](#) 

[Big Data](#) 
6min


[O que é NoSQL?](#) 


[Guia completo de SQL](#) 

[The Complete SQL Bootcamp 2022: Go from Zero to Hero](#) 

[Modelagem de Dados - Conceitos de Bancos de Dados](#) 
21min

[Top 6 NoSQL Databases](#) 

[Spark and Python for Big Data with PySpark](#) 

[Big Data - O que é e qual sua importância?](#) 

REGRESSÃO

-
- Regressão linear
 - GLM
 - Árvores de Regressão / Bagging / Boosting / Outliers
 - Otimização de Erros: Principais diferenças entre eles
 - Redes Neurais: Funções de Ativação
 - Gradiente Descendente
 - SVR
 - KNN






[Model Evaluation - Classification](#) 

[Model Evaluation - Regression](#) 

[Métodos de Shrinkage](#) 


[Variable selection](#) 

<https://machinelearningmastery.com/probabilistic-model-selection-measures/>

[Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition](#) 
by Aurélien Géron
Released October 2022
Publisher(s): O'Reilly Media, Inc.
ISBN: 9781098125974

[Cap 4. Training Models](#) 

- [Linear Regression](#)

[Regression Metrics | MSE, MAE & RMSE | R2 Score & Adjusted R2 Score](#) 

[44min](#)

[Gradiente descendente, passo a passo](#) 

[24min](#)

[Stochastic Gradient Descent, Clearly Explained!!!](#) 

[11min](#)

[StatQuest: Random Forests Parte 1 - Construindo, Usando e Avaliando](#) 


[10min](#)

[Random Forest Algorithm Clearly Explained!](#) 

[8min](#)

[What is Random Forest?](#) 

[6min](#)

[The Main Ideas of Fitting a Line to Data \(The Main Ideas of Least Squares and Linear Regression.\)](#) 

[10min](#)



[Modelos de Regressão com Apoio Computacional](#) 


[Statistical Learning: 7.4 Generalized Additive Models and Local Regression](#) 

[10min](#)


[Happy Halloween \(Neural Networks Are Not Scary\)](#) 

[1min](#)



[Gradient Boost Part 1 \(of 4\): Regression Main Ideas](#) 


[15min](#)

[XGBoost Part 1 \(of 4\): Regression](#) 

[26min](#)

[Visual Guide to Gradient Boosted Trees \(xgboost\)](#) 

[4min](#)

[Gradient Boost Machine Learning|How Gradient boost work in Machine Learning](#) 


[14min](#)

[KNN Exemplo Completo](#) 


[34min](#)

[StatQuest: K-nearest neighbors, Clearly Explained](#) 


[5min](#)

[Minha Primeira Rede Neural \(Teoria\) - Redes Neurais e Deep Learning 01](#) 

[11min](#)

[Gradient Boost Part 1 \(of 4\): Regression Main Ideas](#) 

[15min](#)

[XGBoost Part 1 \(of 4\): Regression](#) 

[26min](#)

[Gradient Boosting with Regression Trees Explained](#) 

[5min](#)

[K Nearest Neighbour Easily Explained with Implementation](#) 

[18min](#)

[Support Vector Regression SVR](#) 

[9min](#)



[An Introduction to Statistical Learning: with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor](#) 


3.5 Comparison of Linear Regression with K-Nearest Neighbors

[The Elements of Statistical Learning, Second Edition, Trevor Hastie Robert Tibshirani Jerome Friedman](#) 

12.3.6 Support Vector Machines for Regression

[An Introduction to Statistical Learning: with Applications in Python, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, Jonathan Taylor](#) 

7.7.1 GAMs for Regression Problems

[Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition by Aurélien Géron](#) 

Released October 2022

Publisher(s): O'Reilly Media, Inc.

ISBN: 9781098125974

10. Introduction to Artificial Neural Networks with Keras 

- Regression MLPs

CLASSIFICAÇÃO

-
- Regressão Logística
 - Gradiente Descendente
 - Análise Discriminante Linear e Quadrática
 - Naive Bayes
 - SVM: Kernels e Fronteiras de Decisão
 - Árvores de Decisão
 - Ensemble: Votação e Stacking
 - Classificação Multiclasse
 - Calibração de Probabilidade
 - Redes Neurais: Multi-Layer Perceptron
 - Gaussian Process Classification
 - Aprendizado Semi-Supervisionado
 - Eventos Raros





[Árvores de Decisão](#) 


[Algoritmo de classificação Naive Bayes](#) 

[Regressão Logística](#) 


12min

[Árvore Binária de Busca](#) 


15min

[Naive Bayes - Georgia Tech - Machine Learning](#) 


8min

[Naive Bayes Theorem | Introduction to Naive Bayes Theorem | Machine Learning Classification](#) 

10min

[Uma gentil introdução ao aprendizado de máquina](#) 

13 min

[Stanford CS229: Machine Learning | Summer 2019 | Lecture 21 - Evaluation Metrics](#) 

1h47min

[Evaluation Metrics in Classification](#) 

7 min

[Calculating the Gini Coefficient](#) 

7 min

[ROC and AUC, Clearly Explained!](#) 

16 min


[Kolmogorov-Smirnov test](#) 

5 min

[Kolmogorov Smirnov | KS for business analytics](#) 

10 min

[Decision boundaries](#) 

[3.1.3 Decision Boundary by Andrew Ng](#) 

[Decision and Classification Trees, Clearly Explained!!!](#) 

18 min

[Decision Tree Classification Clearly Explained!](#) 

10 min




[Let's Write a Decision Tree Classifier from Scratch - Machine Learning Recipes #8](#) 

16 min

[Máquinas de Vetores de Suporte, Claramente explicadas!!!](#) 

21min

[SVM - Support Vector Machines: Fundamentos e prática](#) 

01hora

[Naive Bayes, Clearly Explained!!!](#) 

15 min

[Naïve Bayes com Python](#) 

19 min

[StatQuest: Random Forests Parte 1 - Construindo, Usando e Avaliando](#) 

10 min



[Random Forest Regression Explained in 8 Minutes](#) 

8 min

[Random Forest Regression Introduction and Intuition](#) 

8 min

AGRUPAMENTO

- Maldição da Dimensionalidade
- Métricas de Avaliação para Clusterização
- Distância entre pontos e Geometria
- Tipos de algoritmos de Clusterização
- K-means
- GMM
- DBSCAN
- Hierárquicos





[Entenda o Algoritmo K-Means](#) 

[Métodos de Agrupamento de Dados](#) 


[O que é análise de Cluster](#) 

[Latent Dirichlet Allocation](#) 

[A Text Mining Research Based on LDA Topic Modelling](#) 


[Topic Modeling and Latent Dirichlet Allocation \(LDA\) in Python](#) 

[Introdução ao agrupamento hierárquico](#) 


[Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd Edition by Aurélien Géron](#) 
[9. Unsupervised Learning Techniques](#)
[* Gaussian Mixtures](#)

[K-Means com Python Parte 1/2 Exemplo Visual](#) 

15min

[StatQuest: K-means clustering](#) 
8min

[K-Means com Python Parte 2/2 Exemplo Visual](#) 
13min

[Stanford CS229: Machine Learning | Summer 2019 | Lecture 16 - K-means, GMM, and EM](#) 
2h

[LDA Algorithm Description](#) 
10min

[Vídeo 49 - Agrupamento Hierárquico de Dados](#) 
16min

[O Modelo GMM \(Gaussian Mixture Models\)](#) 
18min

[StatQuest: K-means clustering](#) 
8min

[Clustering with DBSCAN, Clearly Explained!!!](#) 

10min

[Análise de Cluster na Mineração de Dados](#)

Curso do Coursera, em inglês com legendas em PT-BR. Duração aproximada de 16h.



ESTATÍSTICA



-
- Tipos de variáveis
 - Medidas Resumo
 - Probabilidade
 - Distribuições de probabilidade
 - Amostragem
 - Inferência



[Probabilidade Condicional](#) 

[Estimadores Pontuais](#) 

[Probability Events Conditional](#) 


[Binomial Distribution](#) 

[Distribuição Gaussiana](#) 

[Análise Exploratória de Dados - UEL](#) 

[Medidas de Centro](#) 


[Amostragem Estatística](#) 


[Revisão dos métodos de amostragem](#) 

[Curso completo de Probabilidade](#) 

[Curso Completo de Estatística](#) 

[Econometria Básica Aplicada](#) 

[Probabilidade: Conceitos Básicos](#) 
13min


[Cálculo de Probabilidades](#) 
10min

[Probabilidade da União de Dois Eventos](#) 
13min


[Probabilidade Condicional](#) 
9min

[Independent Events \(Basics of Probability: Independence of Two Events\)](#) 
25min

[Variáveis Aleatórias Independentes](#) 
11min

[O que são e como fazer Distribuição de Probabilidades](#) 
9min


[Binomial Distribution](#) 
12min


[Montando uma distribuição de probabilidades para variável discreta](#) 
8min


[Processo de Poisson 1](#) 
12min

[Processo de Poisson 2](#) 
11min


[Exploratory Data Analysis](#) 
20min

[Introdução à estatística: média, mediana e moda](#) 
9min


[O que são Variáveis Aleatórias Discretas e Contínuas](#) 
14min

[Tipos de amostragens - Introdução à Estatística](#) 
3min



[Noções de Probabilidade e Estatística \(Volume 1\), Edusp, Marcos Nascimento Magalhães e Antonio Carlos Pedroso de Lima, caps 1, 2 e 4](#) 

[An Introduction to Statistical Learning](#) 

[Histogramas](#) 
 4min

PROGRAMAÇÃO


- Pseudolinguagem (noções de algoritmo)
- Sintaxe – Python
- SQL
- Pyspark
- UDF / função lambda
- Performance e comparação de linguagens
- Lazy evaluation






[Lógica de Programação com Python](#) 

[Lógica de Programação](#) 

[Curso Python 01 - Introdução - Aprenda Programar do ZERO](#) 


[DESIGN PATTERNS](#) 

[Lógica de Programação e Algoritmos](#) 


[Deployment of Machine Learning Models](#) 

[Design Patterns Python](#) 

[An introduction to Apache Spark Architecture](#) * 

[10 ferramentas e bibliotecas para trabalhar com data mining e Big Data](#) 

1h

[Optimizing Apache Spark on Databricks](#) * 

6h

[Apache Spark Programming With Databricks](#) * 

12h

[Design Patterns in Python by Peter Ullrich](#) 
30min

[Introdução à programação com Python](#) 

[Learning Python](#) 


*Esses três cursos são pagos, mas o Santander possui um convênio com o Databricks. Assim, ao fazer a conta na plataforma, usar o e-mail corporativo

TEORIA DO APRENDIZADO


- Funções de Custo
- Gradiente Descendente
- Métricas de Avaliação
- Viés-Variância
- Validação
- Ensemble
- Redução de Dimensionalidade
- Data Augmentation
- Otimização de Hiperâmetros






[Stanford CS229: Machine Learning | Summer 2019 | Lecture 12 - Bias and Variance & Regularization](#) 

[StatQuest: Análise de Componentes Principais \(PCA\), Passo a Passo](#) 
22 min


[Redução de dimensionalidade e mudança de representação](#) 
44 min

[How to Evaluate the Performance of Clustering Algorithms in Python?](#) 
20min



[Structuring Machine Learning Projects - The Summer Edition!](#) 



NG, <https://www.deeplearning.ai/resources/#books>,  Caps 24, 25, 26 e 27

<https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation> 

MATERIAIS COMPLEMENTARES

Nesta seção, separamos alguns Cursos e Livros gerais para complementar o seu processo de aprendizagem.



[Introdução à Ciência de Dados versão 3.0](#)



[Big Data Fundamentos 2.0](#)



[Formação Cientista de Dados](#)



[The Data Science Course 2022: Complete Data Science Bootcamp](#)



[Data Scientist with Python](#)



[Web Scraping and API Fundamentals in Python](#)



[Bootcamp Completo em Data Science com Python 2022](#)



[Credit Risk Modeling in Python 2022](#)



[Python Fundamentos para Análise de Dados](#)



[Python for Data Science and Machine Learning Bootcamp](#)



[Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow](#)



[An Introduction to Statistical Learning: With Applications in R.](#) Springer



[Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy e IPython](#)



[Probabilidade - Aplicações à Estatística](#)



[The Elements of Statistical Learning: Data Mining, Inference, and Prediction](#)



[Estatística Básica](#)

ROTEIRO DE AVALIAÇÃO DAS BANCAS

Sabemos que a banca é um momento que causa muita ansiedade e, por isso, construímos este **roteiro** para te ajudar na preparação:

- Momento inicial: quebra gelo
- 1h30: apresentação do case e sabatina.
Durante a apresentação do case, serão feitas perguntas teóricas sobre as técnicas utilizadas e demais temas que podem ou não ter a ver com o case

Critérios avaliados em bancas:

- Case
- Elaboração do Material Apresentado
- Apresentação



DICAS IMPORTANTES PARA A BANCA



- Comece se apresentando, quebre “o gelo”. Você se sentirá mais confortável;
- Crie uma conexão com os avaliadores, eles estão ali para te ajudar também;
- Seja objetivo e assertivo;
- Faça associações entre modelos e técnicas;
- Não é necessário repassar a base/case pois os Experts já conhecem;
- Controle o tempo para que consiga apresentar toda sua resolução;
- Caso precise consultar algum material, avise os avaliadores da banca;
- Se você não entendeu algo, sinta-se à vontade para perguntar aos avaliadores e pedir que se aprofundem mais.
- E a dica final: a vivo, o seu case precisa funcionar!

DATA MASTERS

QUE A FORÇA ESTEJA COM VOCÊ...

E OS ESTUDOS TAMBÉM!