

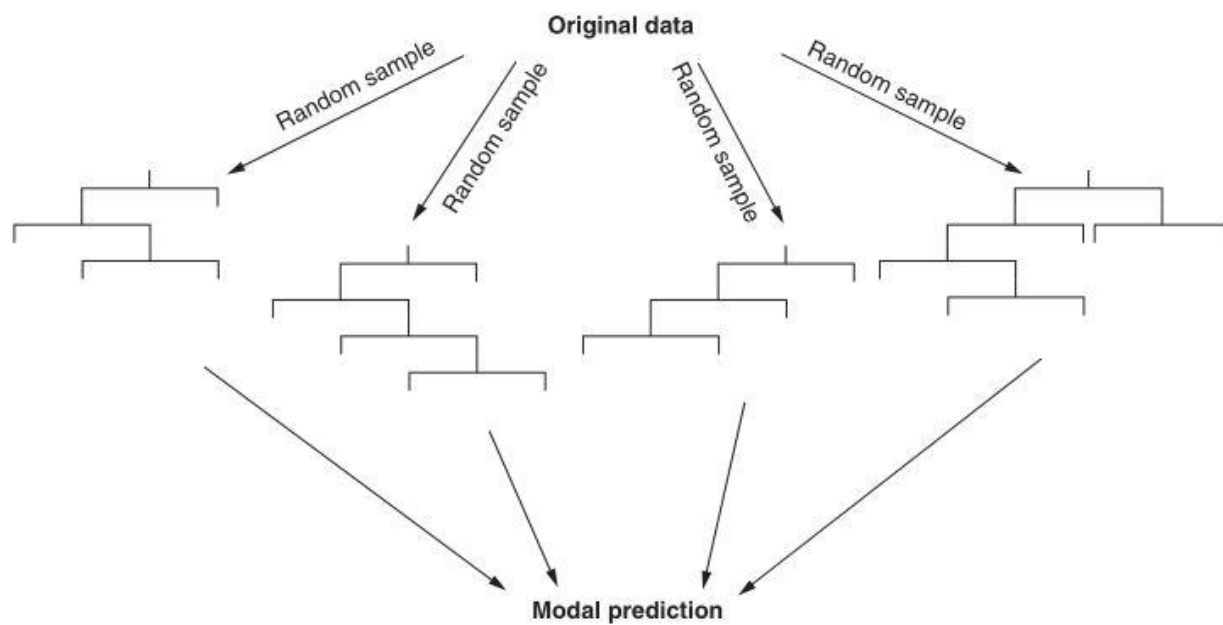
Classificação do uso e cobertura da terra utilizando o algoritmo Random Forest na Plataforma Google Earth Engine

Eduardo Ribeiro Lacerda

Fluminense Federal University (UFF)
International Institute for Sustainability (IIS)



Sobre o que vamos falar?

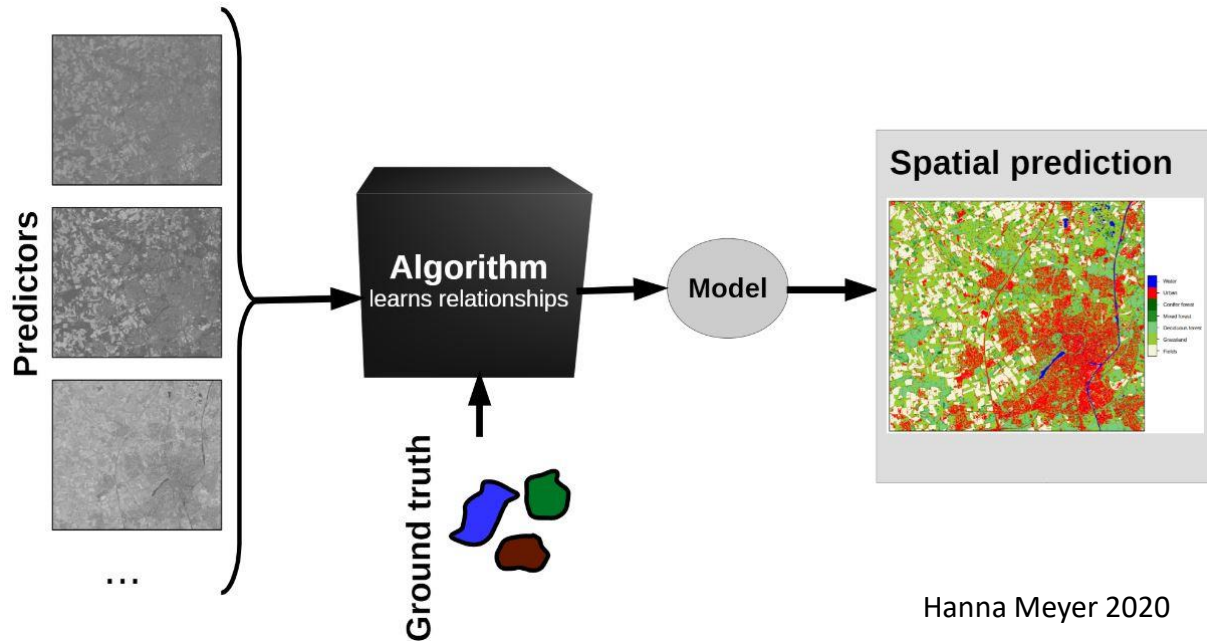


O que vamos aprender?

- Como trabalhar com dados espaciais no Google Earth Engine
- Como coletar boas amostras para o seu modelo
- Como limpar e preparar os seus dados
- Como aplicar algoritmos como o Random Forest para classificar imagens de satélite
- Como melhorar seu modelo
- Como validar o modelo e seus resultados
- Como visualizar e exportar seus dados



A ideia principal é...



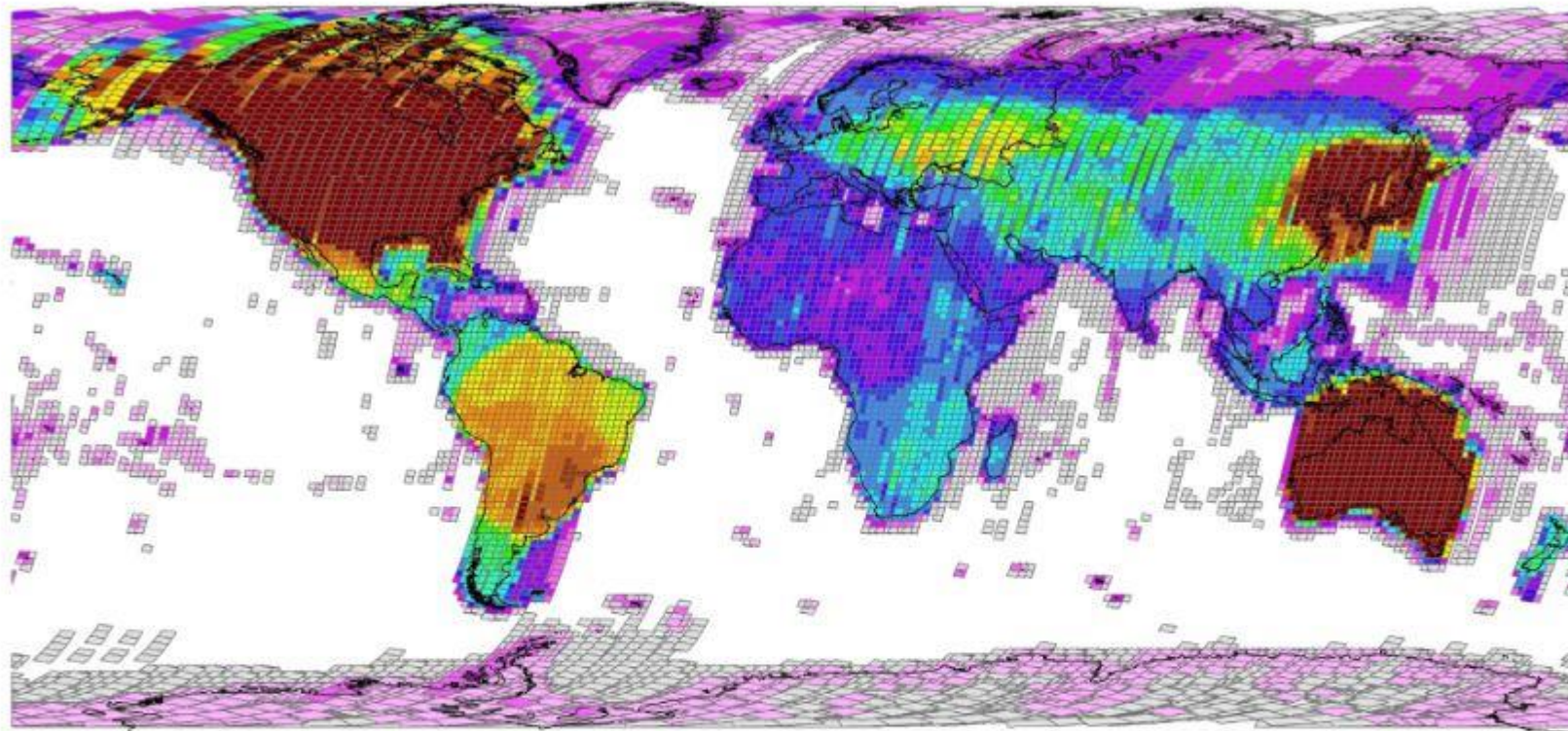
- Primeiro, selecionar imagens de satélite que sirvam como bons dados de entrada pro seu modelo
- Coletar boas amostras
- Treinar o modelo e fazer testes que melhorem seu desempenho
- Usar o modelo para classificar dados raster e criar bons resultados

Exemplos de dados de entrada

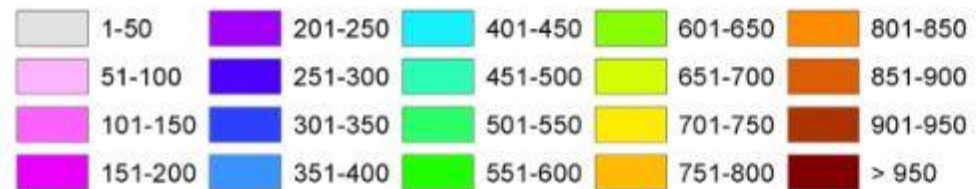
Platform/Sensor	Spatial Resolution	Temporal Resolution	Availability
Landsat MSS	79	16 days	started in 1972
Landsat TM	30	16 days	started in 1982
Landsat ETM+	30	16 days	started in 1999
Landsat 8 OLI	30	16 days	started in 2013
Landsat 9 OLI-2	30	16 days	mid 2021
Sentinel 2	10-20	5/10 days	started in 2014
MODIS	250-1000	4 per day	started in 2000



Acessibilidade do sensor Landsat OLI (Landsat 8)



Number of images



Wulder et al., 2015

Como ter acesso?

The screenshot displays the USGS EarthExplorer interface. At the top, the USGS logo and tagline "science for a changing world" are visible. Below the header, the "EarthExplorer - Home" page is shown with navigation links like "Home", "New System Message", "Save Criteria", "Load Favorite", and "Manage Criteria". A "Page Expires In 1:57:21" timer is also present.

The main content area is divided into two sections. On the left, the "Search Results" section shows a list of data sets. The "Data Set" dropdown is set to "Sentinel-2". The results list includes:

- Item 54: Acquisition Date: 2018/08/08, Platform: SENTINEL-2B, Tile Number: T32UMC. It includes a thumbnail and icons for download, share, and other actions.
- Item 55: ID: L1C_T32UMC_A016307_20180806T104340, Acquisition Date: 2018/08/06, Platform: SENTINEL-2A, Tile Number: T32UMC. It includes a thumbnail and icons.
- Item 56: ID: L1C_T32UMC_A016307_20180806T104340, Acquisition Date: 2018/08/06, Platform: SENTINEL-2A, Tile Number: T32ULC. It includes a thumbnail and icons.
- Item 57: ID: L1C_T32ULC_A007370_20180804T105022, Acquisition Date: 2018/08/04, Platform: SENTINEL-2B, Tile Number: T32ULC. It includes a thumbnail and icons.
- Item 58: ID: L1C_T32ULC_A016264_20180803T103239, Acquisition Date: 2018/08/03, Platform: SENTINEL-2A. It includes a thumbnail and icons.

At the bottom of the results list are buttons for "View Item Details" and "Submit Download Request".

On the right, the "Search Criteria Summary (Show)" section displays a map of the Netherlands. The map is overlaid with a large green rectangular area representing the search results. A red pin is located on the map. The map is labeled with "Map" and "Satellite" tabs. The Google logo is visible in the bottom left corner of the map area. A disclaimer at the bottom of the map states: "The up-to-date Google map is not for purchase or for download. It is to be used as a guide for reference and search purposes only."

Usando a base de dados do Google Earth Engine

Earth Engine Data Catalog

Search

English

Home View all datasets Browse by tags Landsat MODIS Sentinel API Docs

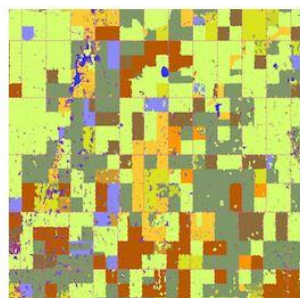
Earth Engine Data Catalog

Earth Engine's public data catalog includes a variety of standard Earth science raster datasets. You can import these datasets into your script environment with a single click. You can also upload your own [raster data](#) or vector data for private use or sharing in your scripts.

Looking for another dataset not in Earth Engine yet? Let us know by [suggesting a dataset](#).

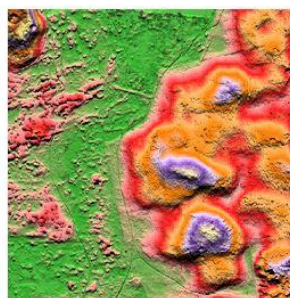
Filter list of datasets

Canada AAFC Annual Crop Inventory



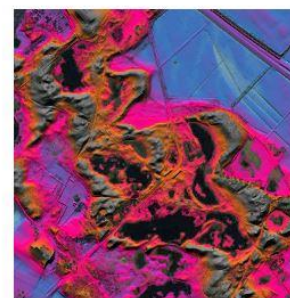
Starting in 2009, the Earth Observation Team of the Science and Technology Branch (STB) at Agriculture and Agri-Food Canada (AAFC) began the process of generating annual crop type digital maps. Focusing on the Prairie Provinces in 2009

AHN Netherlands 0.5m DEM, Interpolated



The AHN DEM is a 0.5m DEM covering the Netherlands. It was generated from LIDAR data taken in the spring between 2007 and 2012. It contains ground level samples with all other items above ground (such as buildings, bridges, trees etc.) removed.

AHN Netherlands 0.5m DEM, Non-Interpolated



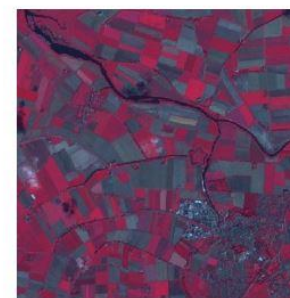
The AHN DEM is a 0.5m DEM covering the Netherlands. It was generated from LIDAR data taken in the spring between 2007 and 2012. It contains ground level samples with all other items above ground (such as buildings, bridges, trees etc.) removed.

AHN Netherlands 0.5m DEM, Raw Samples



The AHN DEM is a 0.5m DEM covering the Netherlands. It was generated from LIDAR data taken in the spring between 2007 and 2012. This version contains both ground level samples and items above ground (such as buildings, bridges, trees etc.).

ASTER L1T Radiance Samples

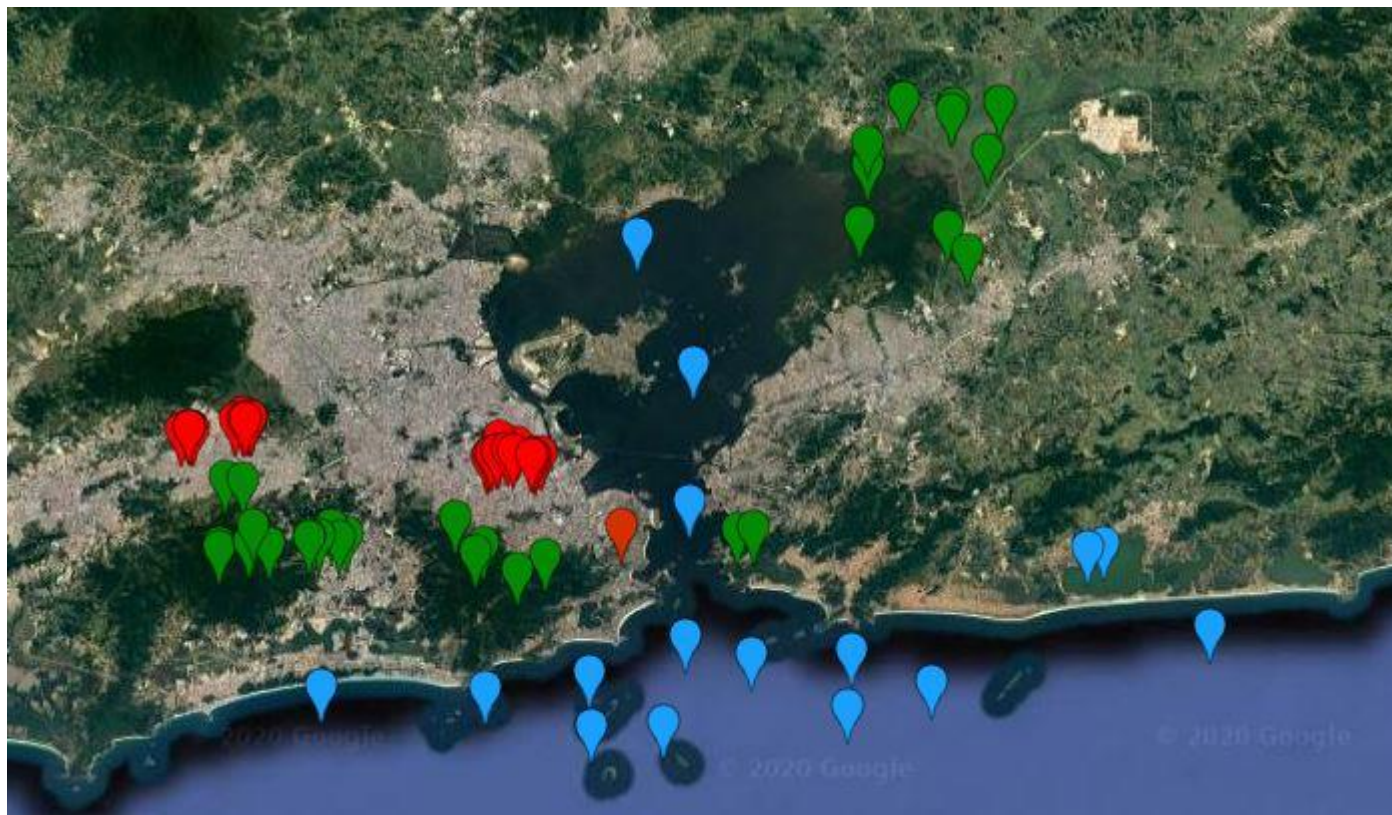


The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) is a multispectral imager that was launched on board NASA's Terra spacecraft in December, 1999. ASTER can collect data in 14 spectral bands from the



Fiz o download da imagem. E agora?

Pontos



Polígonos



Escolhendo o algoritmo a ser usado

Existem muitas opções...

ee.Classifier

```
ee.Classifier.cart(crossvalidationFactor, maxDepth, minLeafPopula...  
ee.Classifier.decisionTree(treeString)  
ee.Classifier.decisionTreeEnsemble(treeStrings)  
ee.Classifier.gmoMaxEnt(weight1, weight2, epsilon, minIterations, ...  
ee.Classifier.libsvm(decisionProcedure, svmType, kernelType, shri...  
ee.Classifier.minimumDistance(metric)  
ee.Classifier.naiveBayes(lambda)  
ee.Classifier.randomForest(numberOfTrees, variablesPerSplit, min...  
ee.Classifier.smileCart(maxNodes, minLeafPopulation)  
ee.Classifier.smileNaiveBayes(lambda)  
ee.Classifier.smileRandomForest(numberOfTrees, variablesPerSpl...  
ee.Classifier.spectralRegion(coordinates, schema)  
ee.Classifier.svm(decisionProcedure, svmType, kernelType, shrinki...
```



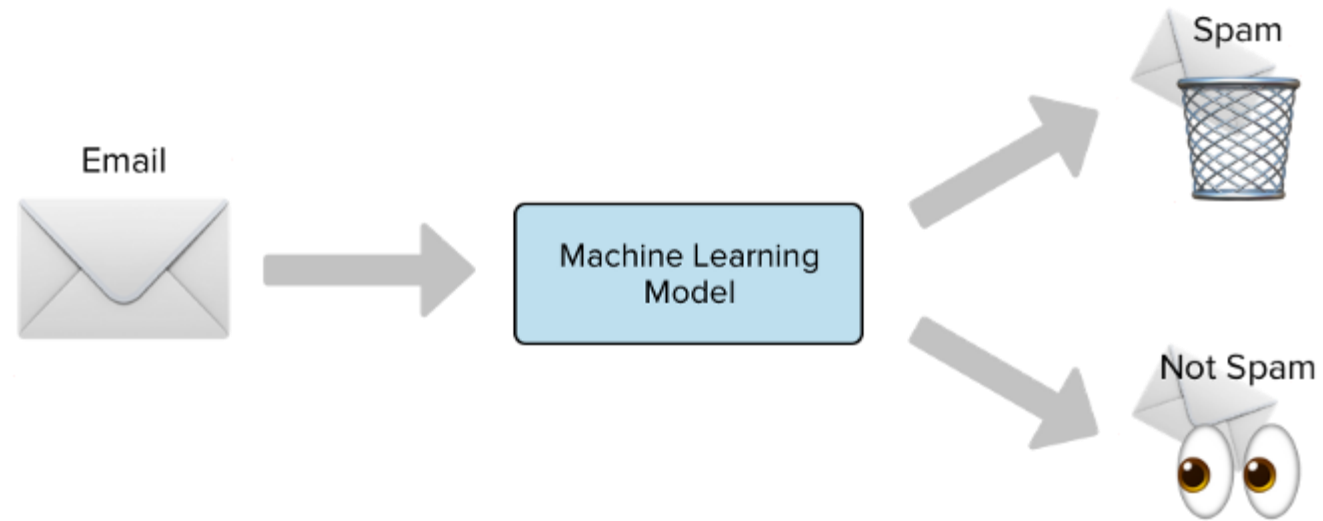
	class	name	short.name	package
1	classif.ada	ada Boosting	ada	ada,rpart
2	classif.adaboostm1	ada Boosting M1	adaboostm1	RWeka
3	classif.bartMachine	Bayesian Additive Regression Trees	bartmachine	bartMachine
4	classif.binomial	Binomial Regression	binomial	stats
5	classif.boosting	Adabag Boosting	adabag	adabag,rpart
6	classif.bst	Gradient Boosting	bst	bst,rpart
7	classif.C50	C50	C50	C50
8	classif.cforest	Random forest based on conditional inference trees	cforest	party
9	classif.clusterSVM	Clustered Support Vector Machines	clusterSVM	SwarmSVM,Liblinear
10	classif.ctree	Conditional Inference Trees	ctree	party
11	classif.cvglmnet	GLM with Lasso or Elasticnet Regularization (Cross Validated...	cvglmnet	glmnet
12	classif.dbnDNN	Deep neural network with weights initialized by DBN	dbn.dnn	deepnet
13	classif.dcSVM	Divided-Conquer Support Vector Machines	dcSVM	SwarmSVM,e1071
14	classif.earth	Flexible Discriminant Analysis	fda	earth,stats
15	classif.evtree	Evolutionary learning of globally optimal trees	evtree	evtree
16	classif.extraTrees	Extremely Randomized Trees	extraTrees	extraTrees
17	classif.fdausc.glm	Generalized Linear Models classification on FDA	fdausc.glm	fda.usc
18	classif.fdausc.kernel	Kernel classification on FDA	fdausc.kernel	fda.usc
19	classif.fdausc.knn	fdausc.knn	fdausc.knn	fda.usc
20	classif.fdausc.np	Nonparametric classification on FDA	fdausc.np	fda.usc
21	classif.featureless	Featureless classifier	featureless	mlr
22	classif.fnn	Fast k-Nearest Neighbour	fnn	FNN
23	classif.gamboost	Gradient boosting with smooth components	gamboost	mboost
24	classif.gaterSVM	Mixture of SVMs with Neural Network Gater Function	gaterSVM	SwarmSVM

Dado de saída



- Urbano
- Floresta
- Solo Exposto
- Gramíneas
- Pasto
- Outros

Bias-variance trade-off



Bias-variance trade-off

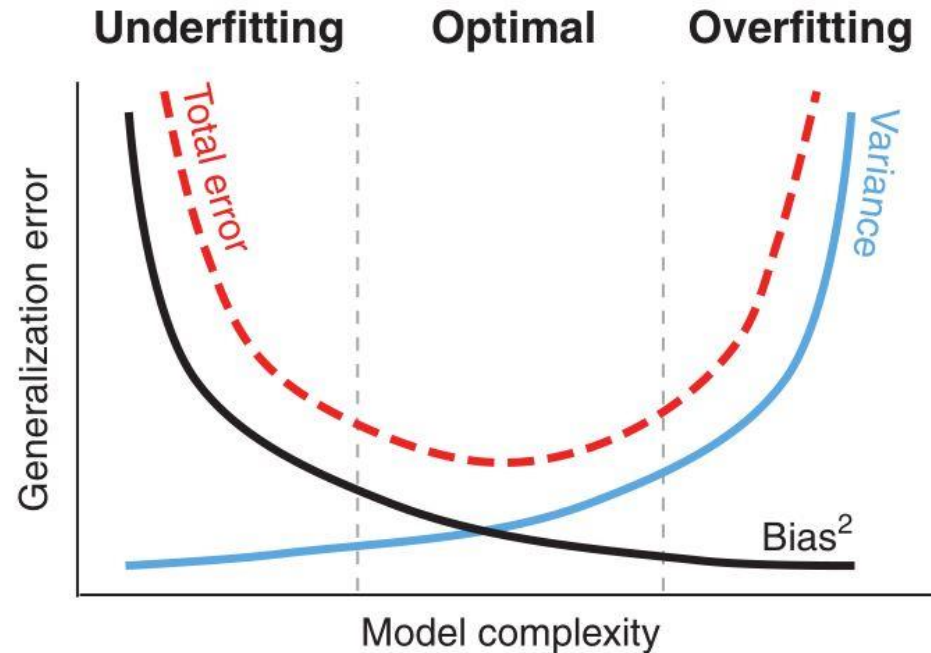
Underfitting e **Overfitting** são duas fontes de erro na construção de modelos.

Underfitted (subajuste) – Acontece quando seu modelo é simples demais e pode acabar gerando erros na classificação de algumas classes. Um modelo com essa característica terá uma performance ruim quando for usado para classificar novos dados.

Overfitted (sobreajuste) – O modelo é complexo demais e está considerando ruídos no dado que você utilizou para treinar o modelo. Um modelo com esta característica irá ter performance ruim quando for usado para classificar novos dados, mas se sairá muito bem classificando os dados que foram usados para treinar o modelo.

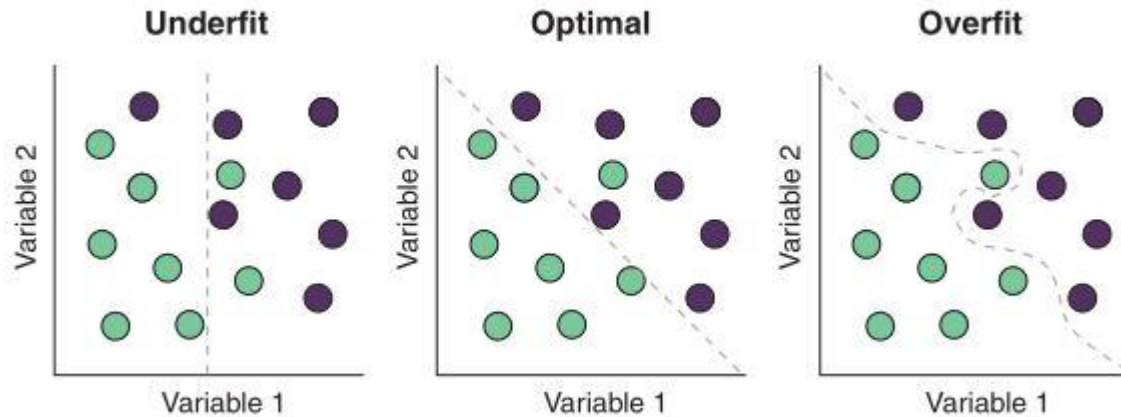


Bias-variance trade-off



- O “generalization error” é a proporção de previsões erradas que um modelo faz quando influenciado tanto em situações de subajuste (underfitting) como sobreajuste (overfitting).
- O erro associado ao sobreajuste acontece quando o modelo é complexo demais. Ou seja, quando utilizamos variáveis demais para treinar o modelo.
- O erro associado ao subajuste acontece quando o modelo é simples demais. Ou seja, quando utilizamos variáveis de menos para treinar o modelo.
- Um modelo ideal (optimal/balanced) equilibra esse trade-off.

Bias-variance trade-off



- A linha pontilhada representa o limite de decisão do modelo
- Então, como podemos resolver este problema? A resposta está em utilizar uma técnica chamada cross-validation (validação cruzada).

Cross-validation (Validação Cruzada)

A solução é avaliar o desempenho do nosso modelo usando dados que o modelo ainda não viu! Podemos realizar uma nova coleta de dados, mas também podemos apenas dividir o conjunto de dados entre amostras de treinamento e testes.

Fazendo isso, nós Podemos usar algumas métricas de desempenho para mostrar como o nosso modelo se irá se comportar com um novo conjunto de dados.

Tipos de cross-validation (validação cruzada)

- Holdout cross-validation
- K-fold cross-validation
- Leave-one-out cross-validation



Cross-validation (Validação Cruzada)

Holdout CV



1. The data is randomly split into a training and test set.
2. A model is trained using only the training set.
3. Predictions are made on the test set.
4. The predictions are compared to the true values.

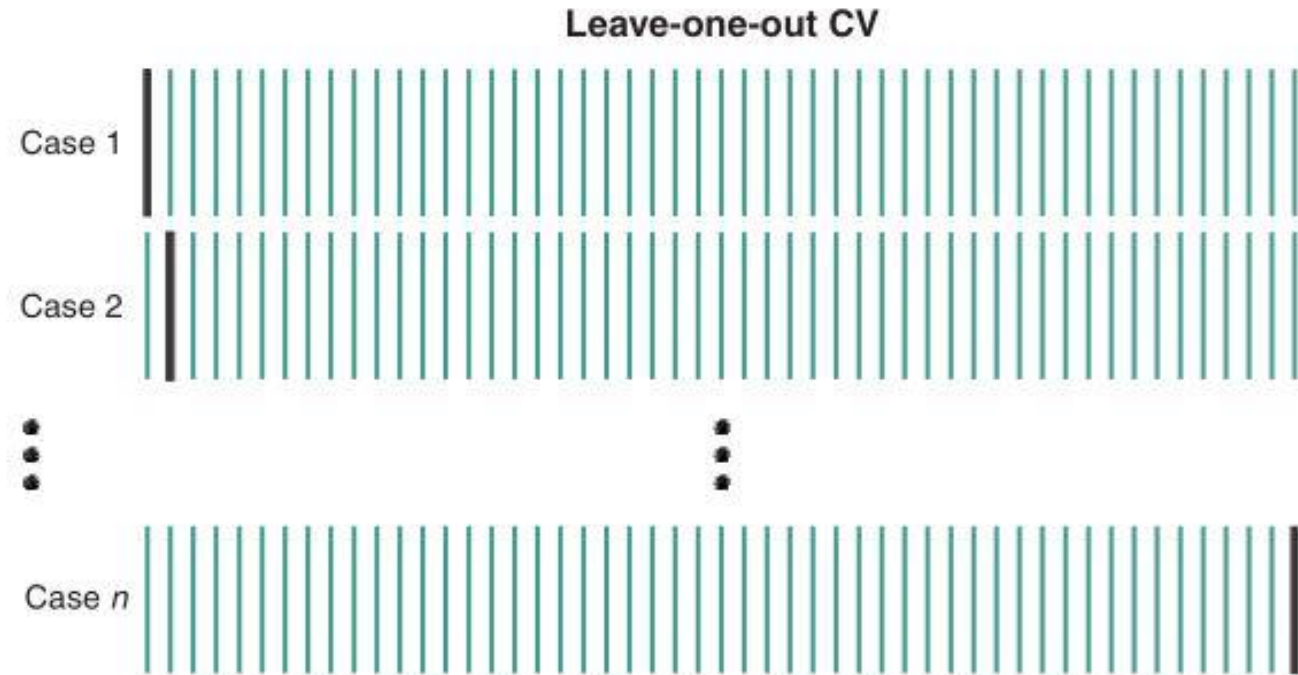
Cross-validation (Validação Cruzada)

K-fold CV

Fold 1		Training set		Test set
Fold 2			Test set	
Fold 3			Test set	
Fold 4		Test set		
Fold 5	Test set			

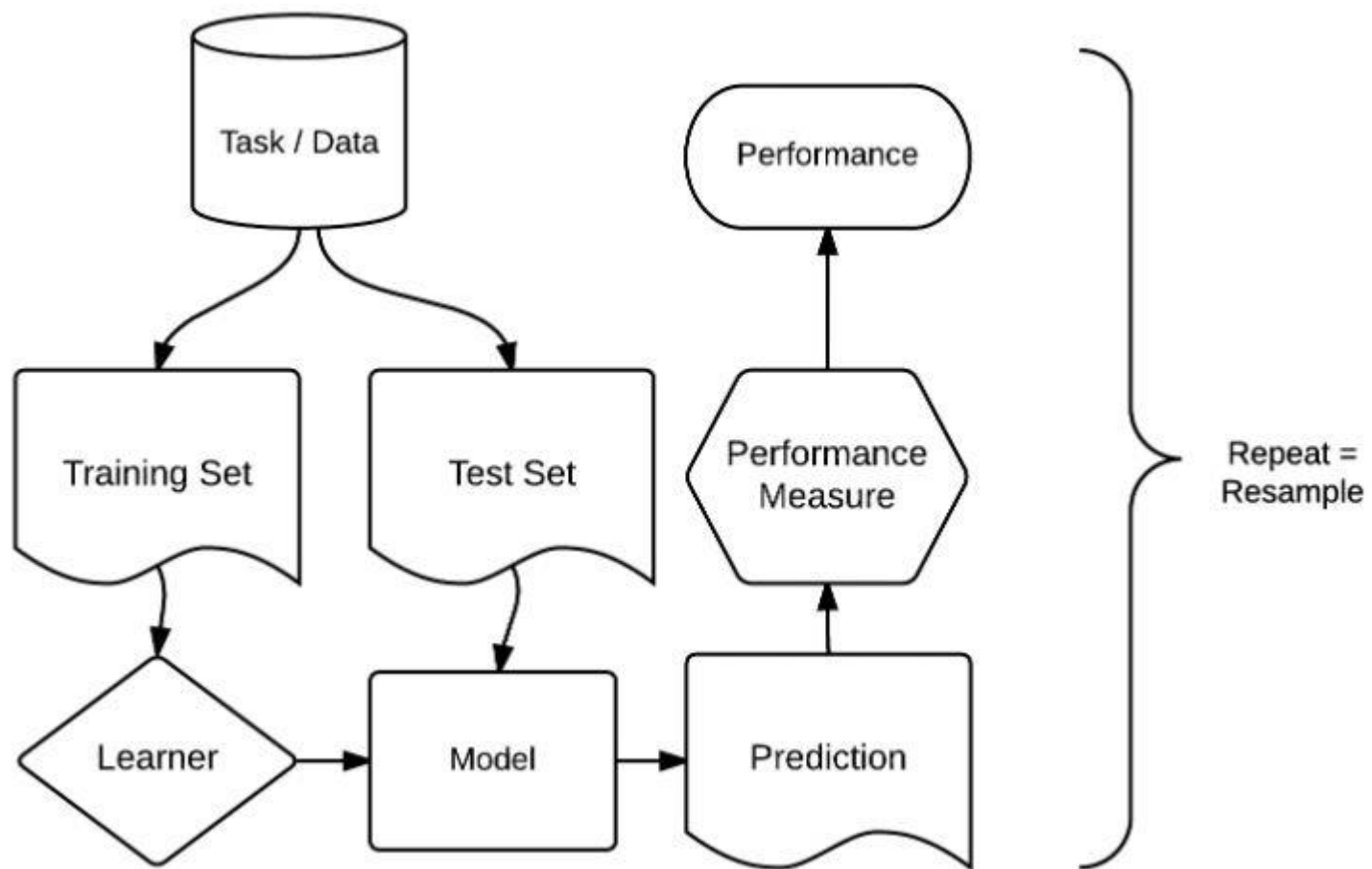
1. The data is randomly split into k equal-sized folds.
2. Each fold is used as the test set once, where the rest of the data makes the training set.
3. For each fold, predictions are made on the test set.
4. The predictions are compared to the true values.

Cross-validation (Validação Cruzada)



1. Use all of the data except a single case as the training set.
2. Predict the value of the single test case.
3. Repeat until every case has been the test case.
4. The predictions for each case are compared to the true values.

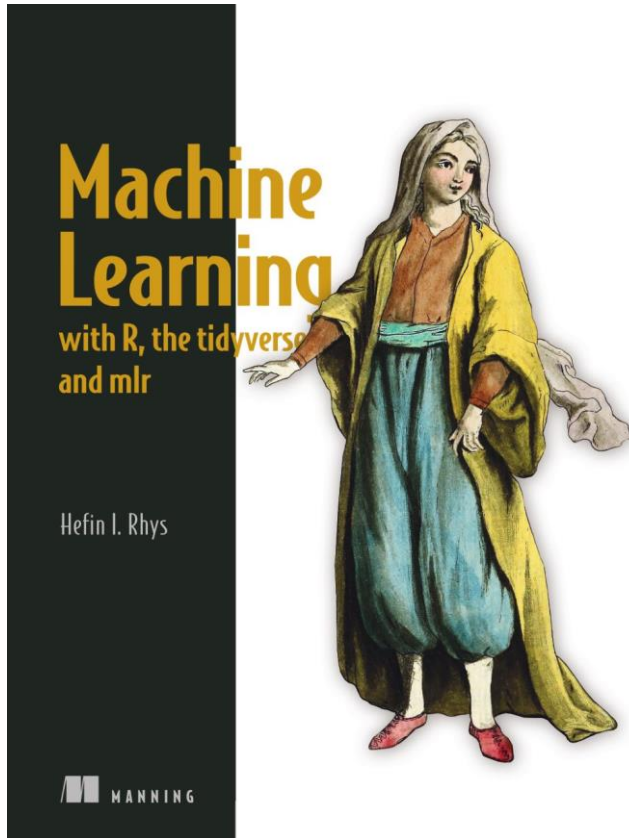
Típico fluxo de trabalho



Yay! :)



Material Extra



Decision Trees:

https://www.youtube.com/watch?v=7VeUPuFGJHk&ab_channel=StatQuestwithJoshStarmer

Random Forest:

https://www.youtube.com/watch?v=J4Wdy0Wc_xQ&t=7s&ab_channel=StatQuestwithJoshStarmer

https://www.youtube.com/watch?v=nyxTdL_4Q-Q&ab_channel=StatQuestwithJoshStarmer

Material Extra

Curso Google Earth Engine (Sadeck)

https://www.youtube.com/watch?v=Dqjtoj9AJak&list=PLNFvG6bTA4NReWtgC93Mh9Tw1RNG4EBMP&ab_channel=LuisSadeck

Google Earth Engine: Explaining all Classifiers

https://www.youtube.com/watch?v=D_KaouS3q20&ab_channel=ProgramSam

