



Universidade Federal do Rio Grande - FURG
Centro de Ciências Computacionais - C3
Engenharia de Computação



Projeto de Graduação em Engenharia de Computação II

Previsão de cotações de Soja futura na bolsa de Chicago (CBOT) utilizando modelo LSTM e relacionando a dados climáticos das regiões mais produtivas dos EUA.

Marlon Rubio de Carvalho Franco
Orientador: Prof. Dr. Marcelo R. Pias

Rio Grande, 2019.

Sumário

1. Introdução
2. Trabalhos relacionados
3. Metodologia
4. Resultados
5. Conclusão

Sumário

1. Introdução

1.1. A Bolsa de Chicago (CBOT)

1.2. Análise de dados para a agricultura

2. Trabalhos relacionados

3. Metodologia

4. Resultados

5. Conclusão

Sumário

1. Introdução

2. Trabalhos relacionados

3. Metodologia

3.1. Obtenção dos dados

3.2. Pré-processamento dos dados

3.3. LSTM

4. Resultados

5. Conclusão

Introdução

Soja

Introdução

Soja

A ***Glycine max (L.) Merr.*** (M.M.P.N.D., 2000) é uma das ***commodities agrícolas*** mais econômicas e valiosas devido a sua composição química única.

Características

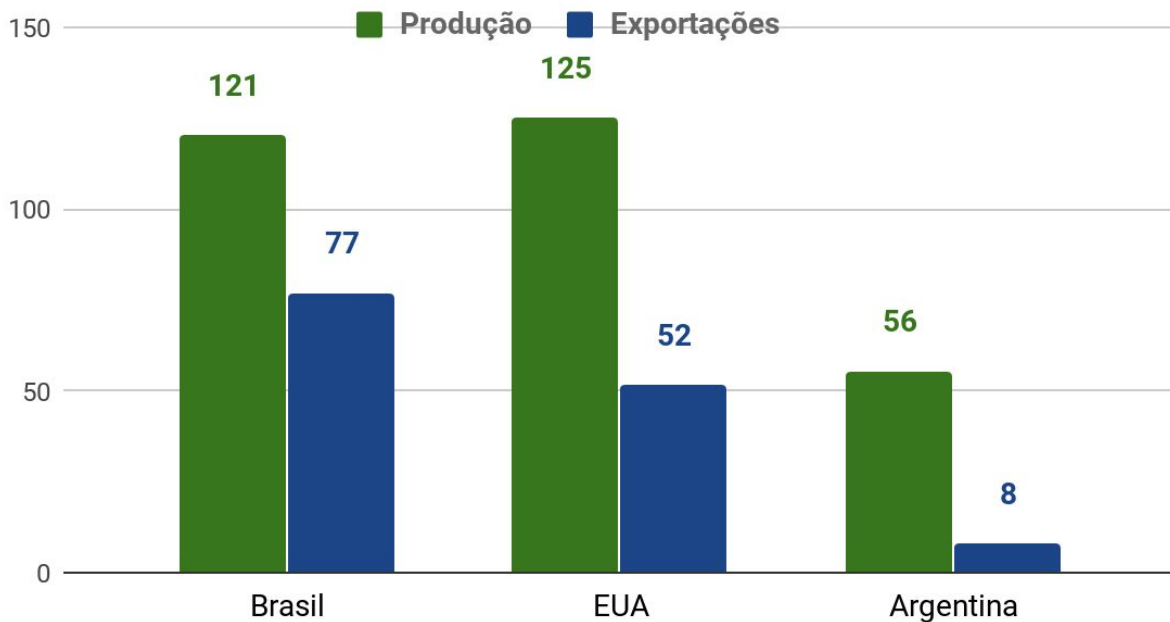
Possui o **maior teor de proteína** (cerca de 40%) entre os cereais e leguminosas, e possui a **segunda maior concentração de óleo** (cerca de 30%), perdendo apenas para o amendoim (48%) (LIU, 1997).

Relevância

Por suas características, a leguminosa se tornou a mais importante **fonte de alimento**, proteína e óleo, sendo cultivada em larga escala pelo mundo (PAGANO; MIRANSARI, 2016).

Introdução

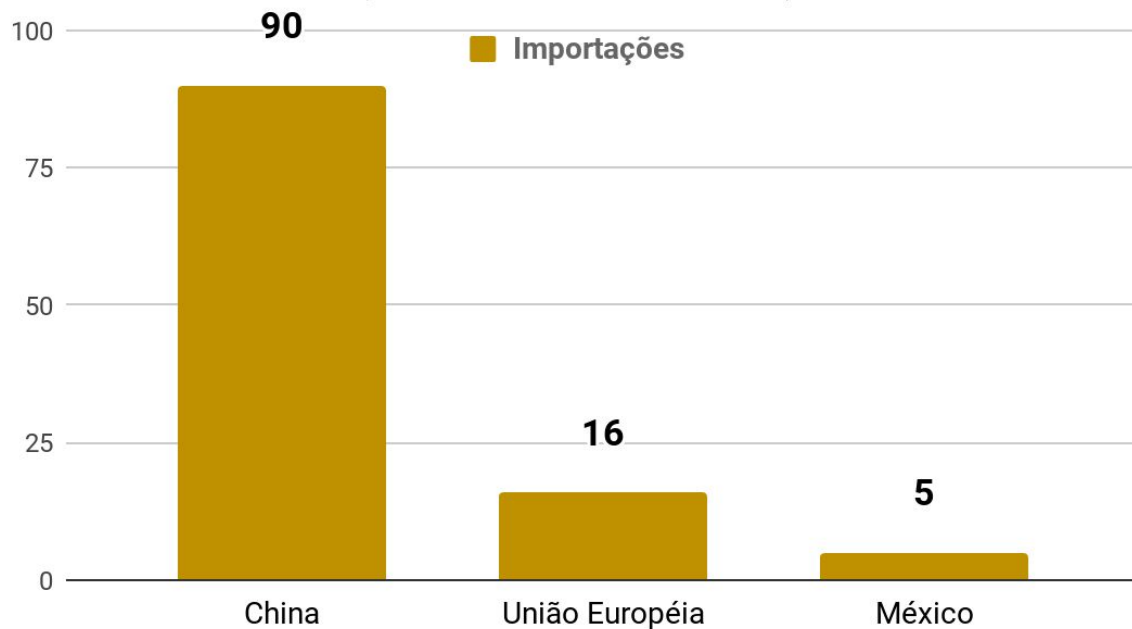
Soja: Abastecimento Mundial e Distribuição em Nov/2018
(milhões de toneladas)



Fonte: Adaptado de USDA, 2018a.

Introdução

Soja: Abastecimento Mundial e Distribuição em Nov/2018
(milhões de toneladas)



Fonte: Adaptado de USDA, 2018a.

Introdução

A Bolsa de Chicago (CBOT)

CHICAGO BOARD OF TRADE

Introdução

CBOT

A **Chicago Board of Trade** (CBOT) é a primeira bolsa de futuros do mundo, **fundada em 1848**.
(CMEGROUP, 2018)

Características

O **mercado futuro** é o tipo de mercado onde são realizadas negociações de compra e venda por meio de **contratos**, cuja entrega ou liquidação se dá em data **futura** e já estabelecida no contrato.

Relevância

É a **principal referência** para os **preços internacionais da soja**, por possuir uma alta concentração de ofertantes e demandantes dos principais países **produtores** e **importadores** da oleaginosa.
(IMEA, 2017)

Introdução

Preço da soja produzida no Brasil

$$(R\$/sc) = \frac{(US\$c/bushel) \times (cotação_do_dollar) \times 2,2046}{100}$$

Cotação de
Chicago

Fator de conversão
(bushel - saca de 60 kg)

O preço da soja no mercado interno depende também de descontos, ou acréscimos, do **prêmio de exportação** e dos custos de movimentação do produto na área produtora para o porto (**frete**). (IMEA, 2017)

Introdução

Evidências de um efeito persistente do clima sobre a dinâmica dos preços da safra de milho, trigo e soja.

“Dado esse caráter não-aleatório do **clima** e dado que os **cinturões de milho, trigo e soja** são **geograficamente concentrados** o suficiente para serem dominados por um fenômeno climático regional, supõe-se que seus **mercados futuros** reflitam essa assimilação de informações meteorológicas não-aleatórias como estruturas de preços não aleatórias.” (Stevens, 1991)

“Este estudo descobriu que esses eventos climáticos não aleatórios transferem uma influência não aleatória para os preços de **commodities** correspondentes durante suas respectivas estações de crescimento.” (Stevens, 1991)

Introdução

Análise de dados para a agricultura

Introdução

Antes de 1940

A primeira fase da Agricultura era caracterizada pela força de **trabalho humano** e baixa produtividade.

1940 a 2000

Agricultura Convencional

teve como principal característica o uso de **fertilizantes** e produtos agrícolas para aumentar a produtividade das lavouras.

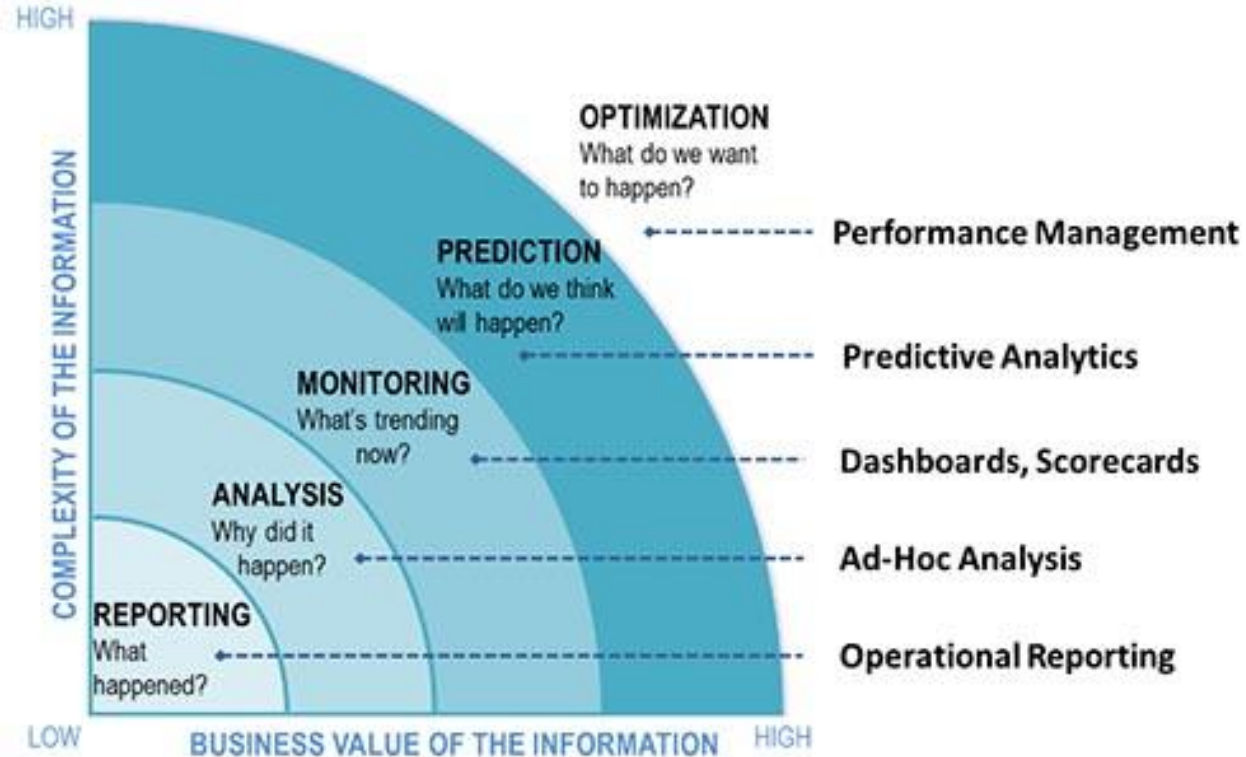
2000 - atual

Agricultura de Precisão (PA)

consiste na análise de **dados coletados** na agricultura para aumentar a **precisão** e auxiliar na tomada de decisões. (PHAM; STACK, 2018)

Introdução

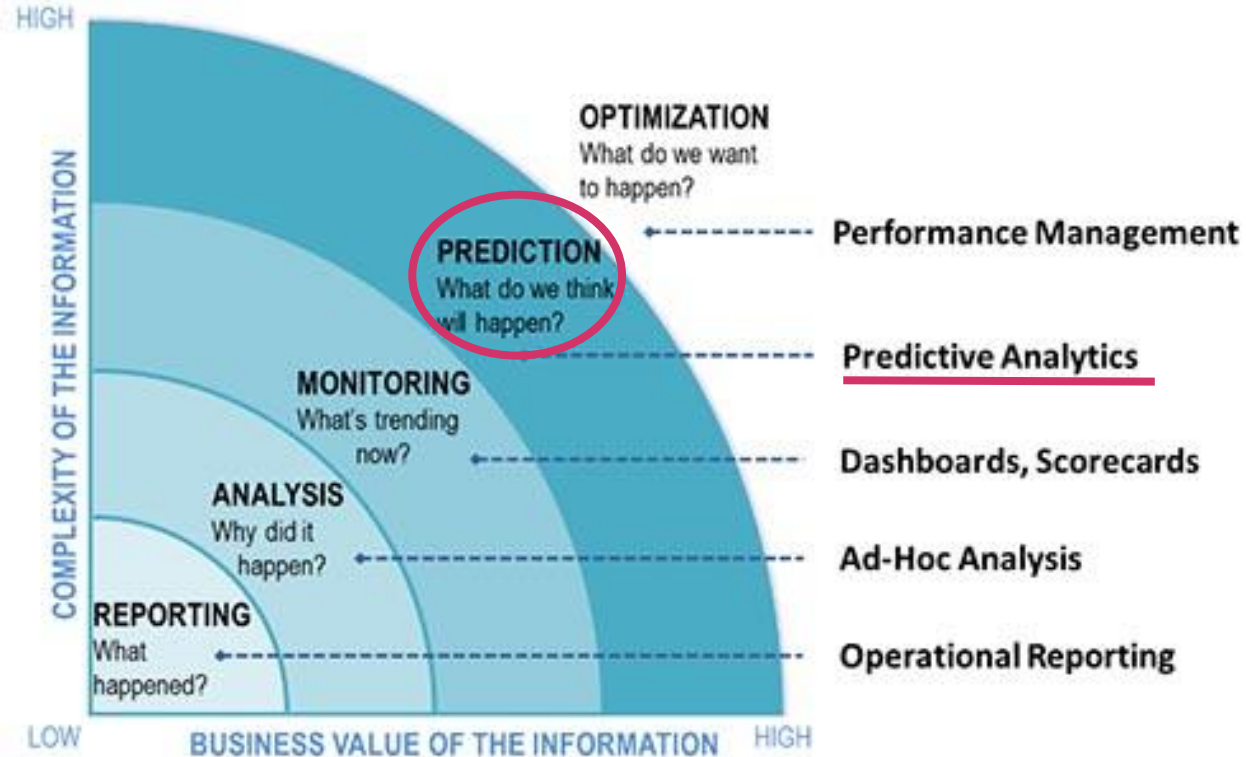
Complexidade e valor da informação



Fonte: DeVries, 2018.

Introdução

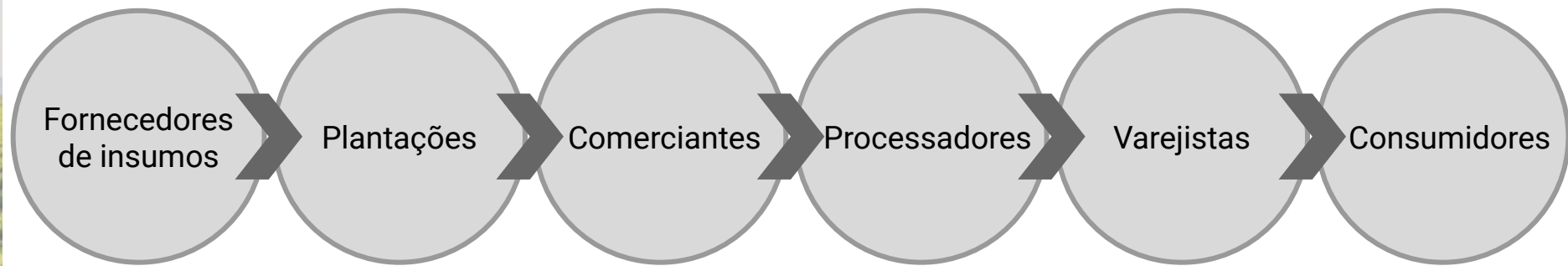
Complexidade e valor da informação



Fonte: DeVries, 2018.

Introdução

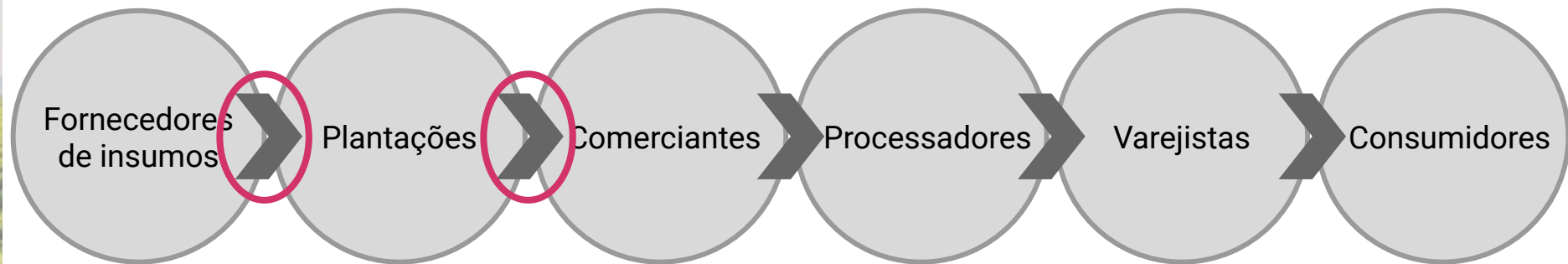
A cadeia de valor da agricultura



Fonte: Adaptado de PHAM e STACK, 2018.

Introdução

A cadeia de valor da agricultura



Fonte: Adaptado de PHAM e STACK, 2018.

Objetivo

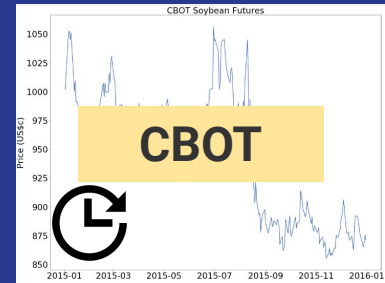
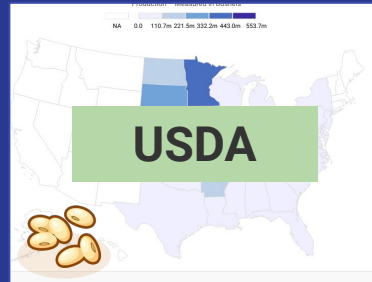
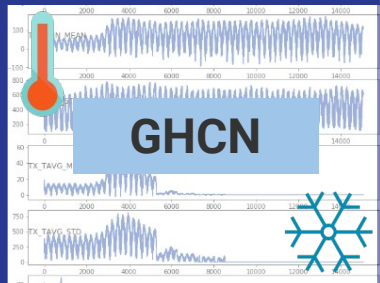
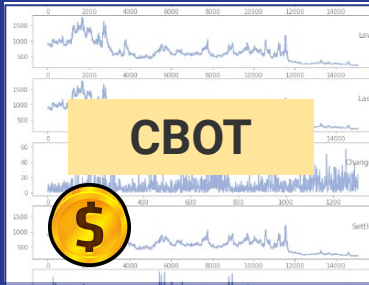
Prever cotações de soja na Bolsa de Chicago (CBOT)

Objetivo

Prever cotações de soja na Bolsa de Chicago (CBOT)

Como?

Dataset
Marlon



Trabalhos Relacionados

LI e TAM (2017)

Comparam o desempenho das técnicas SVM e LSTM para a predição do preço de ações na China.

Acurácia:

SVM > LSTM para dados com **alta** volatilidade.

LSTM > SVM para dados com **baixa** volatilidade.

McNally et. al. (2018)

Comparam o desempenho das redes RNN e LSTM para a predição da cotação do BitCoin.

LSTM 52% mais preciso que RNN.

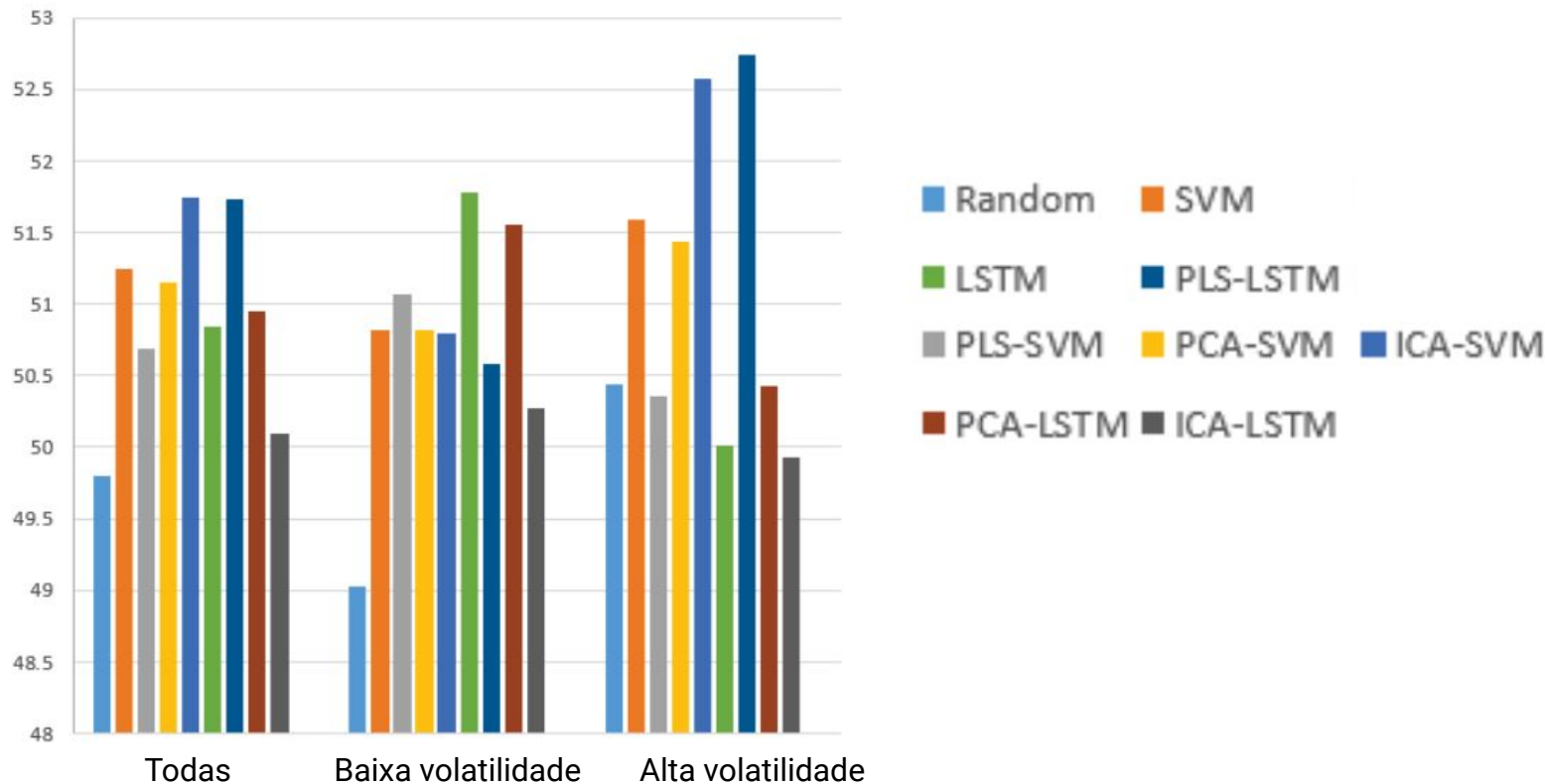
Tempo de treinamento da rede em **GPU** 67,7% mais rápido que em CPU.

Wang (2017)

Compara as técnicas SVM e regressão logística na predição das cotações do milho, soja e petróleo bruto. Importância de se alimentar o modelo com os **dados na forma sequencial**, para preservar a tendência no comportamento dos dados.

Trabalhos Relacionados

Acurácia média das predições de cotações na China: Shanghai Stock Exchange 50 (SSE 50)

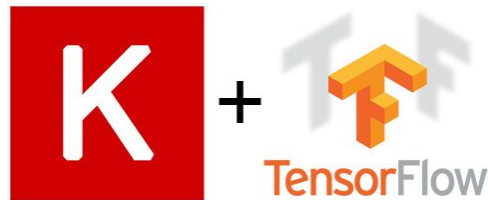
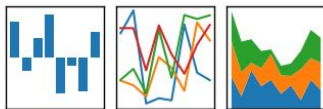


Metodologia



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Metodologia



Obtenção dos dados



Metodologia



Chicago Mercantile Exchange &
Chicago Board of Trade



National Oceanic and Atmospheric
Administration

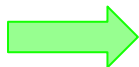


United States Department of Agriculture

Metodologia



Chicago Mercantile Exchange & Chicago
Board of Trade



JSON:

[https://www.quandl.com/api/v3/datasets/
CHRIS/CME_S1](https://www.quandl.com/api/v3/datasets/CHRIS/CME_S1)

```
{
  "dataset": {
    "id": 10922651,
    "dataset_code": "CME_S1",
    "database_code": "CHRIS",
    "name": "Soybean Futures, Continuous Contract #1 (S1) (Front Month)",
    "description": "Historical Futures Prices: Soybean Futures, Continuous Contract #1. Non-adjusted price based on spot-month continuous contract calculations. Raw data from CME. ",
    "refreshed_at": "2019-06-30 02:30:54 UTC",
    "newest_available_date": "2019-06-28",
    "oldest_available_date": "1959-07-01",
    "column_names": [
      "Date",
      "Open",
      "High",
      "Low",
      "Last",
      "Change",
      "Settle",
      "Volume",
      "Previous Day Open Interest"
    ],
  },
  ...
}
```

Metodologia

```
{
  "dataset": {
    "id": 10922651,
    "dataset_code": "CME_S1",
    "database_code": "CHRIS",
    "name": "Soybean Futures, Continuous Contract #1 (S1) (Front Month)",
    "description": "Historical Futures Prices: Soybean Futures, Continuous Contract #1. Non-adjusted price based on spot-month continuous contract calculations. Raw data from CME. ",
    "refreshed_at": "2019-06-30 02:30:54 UTC",
    "newest_available_date": "2019-06-28",
    "oldest_available_date": "1959-07-01",
    "column_names": [ ...
```

Metodologia

```
... "column_names": [  
    "Date",           Data no formato YYYY-MM-DD  
    "Open",           Abertura em US$c  
    "High",           Alta em US$c  
    "Low",            Baixa em US$c  
    "Last",           Ultimo valor em US$c  
    "Change",         ( Settle atual – Settle do dia anterior ) em US$c  
    "Settle",         Fechamento em US$c  
    "Volume",         Volume de contratos negociados no dia  
    "Previous Day Open Interest" Número total de contratos em aberto  
],  
"frequency": "daily",  
"type": "Time Series",  
...
```

Metodologia

Dataset CBOT: cotações de soja da Bolsa de Chicago

	Date	Open	High	Low	Last	Change	Settle	Volume	Previous Day	Open Interest
0	2018-11-21	882.00	889.00	876.00	884.00	2.00	883.00	62504.0		292755.0
1	2018-11-20	874.25	885.75	870.50	880.75	7.25	881.00	62089.0		300785.0
2	2018-11-19	892.00	892.25	871.25	873.75	18.50	873.75	89125.0		299897.0
3	2018-11-16	889.75	894.75	881.75	890.25	3.50	892.25	71310.0		298062.0
4	2018-11-15	884.25	897.50	883.75	889.00	5.25	888.75	80828.0		302082.0

15.105 linhas × 8 colunas

Registros diários desde **08/07/1959** até **21/06/2019**

Metodologia

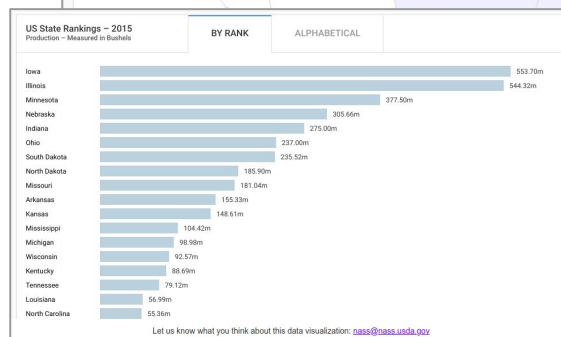
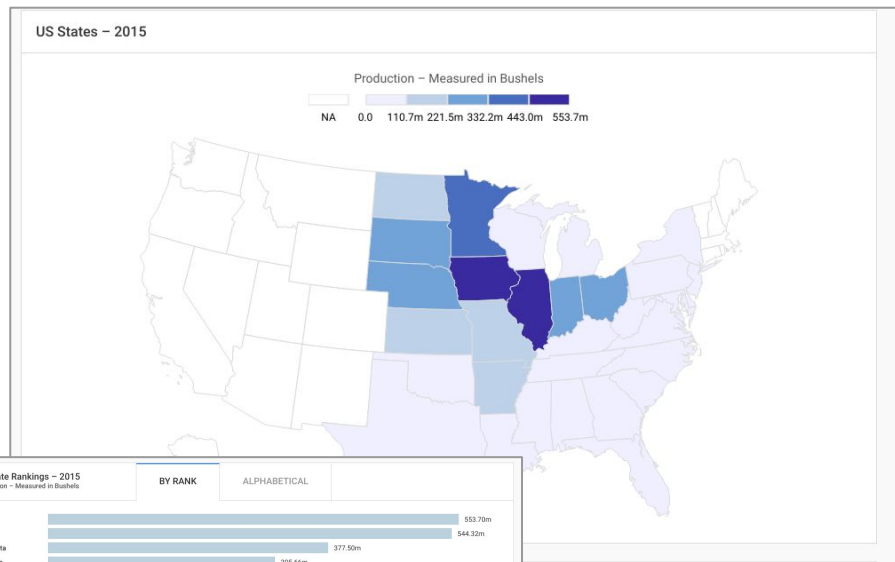


United States Department of Agriculture -
National Agricultural Statistics Service

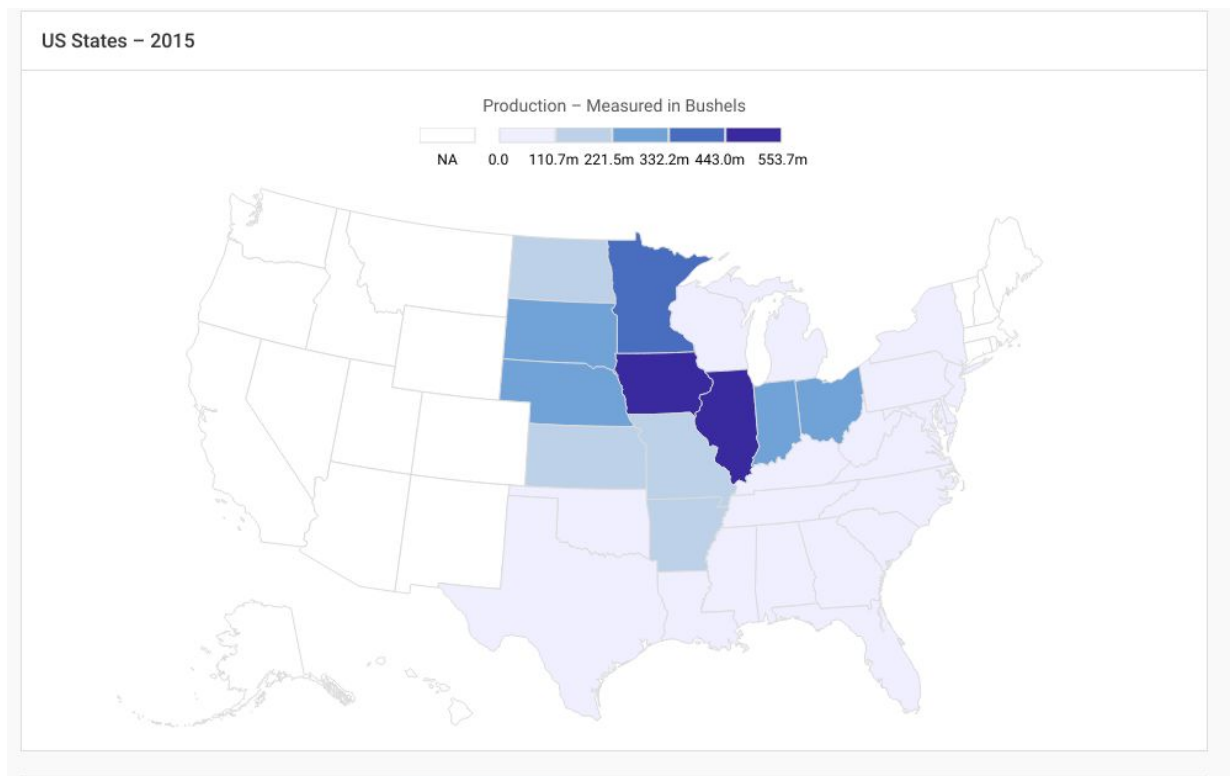
Data Visualization



Possui registros desde **2001** da produção
de soja (em *bushels*) por estado
norte-americano.

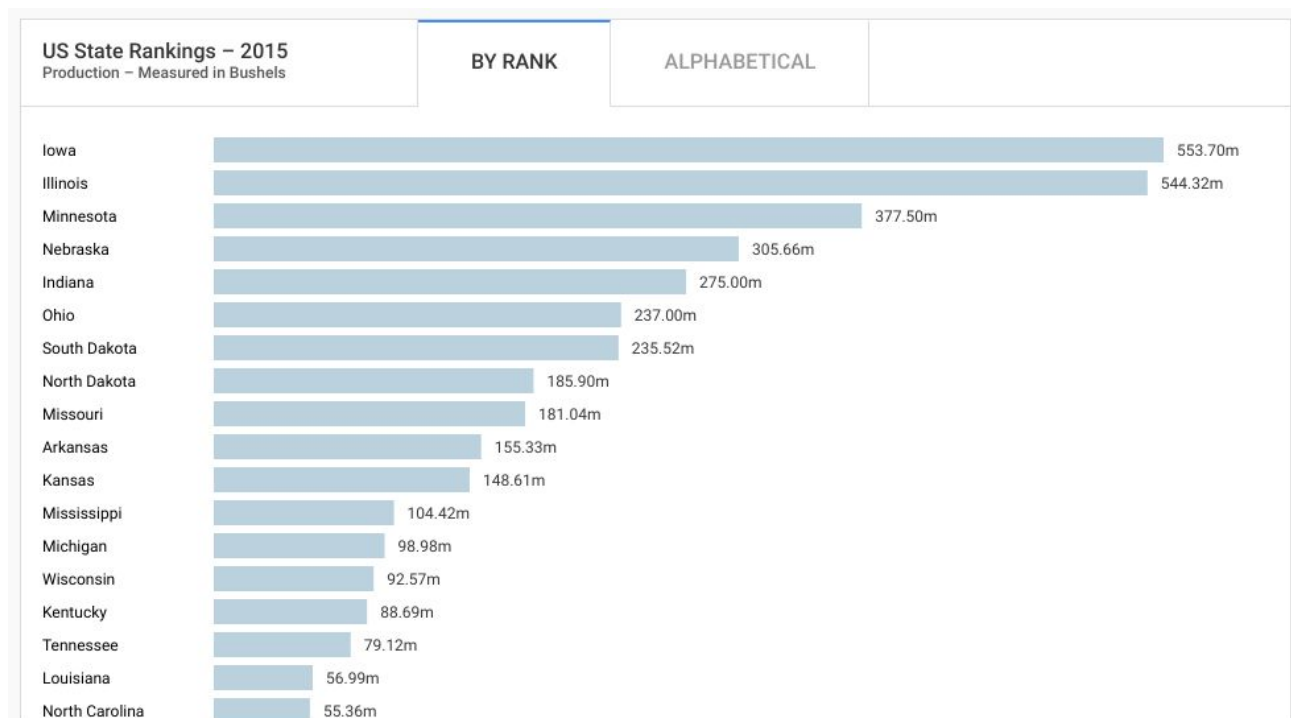


Metodologia



Fonte: USDA, 2018b.

Metodologia



Let us know what you think about this data visualization: nass@nass.usda.gov

Fonte: USDA, 2018b.

Metodologia

USDA-NASS: produção de soja nos estados norte-americanos em 2015.

	state_name	state_alpha	Value	unit_desc	commodity_desc	year
0	ALABAMA	AL	20090000	BU	SOYBEANS	2015
1	ARKANSAS	AR	155330000	BU	SOYBEANS	2015
2	DELAWARE	DE	6920000	BU	SOYBEANS	2015
3	FLORIDA	FL	1102000	BU	SOYBEANS	2015
4	GEORGIA	GA	13330000	BU	SOYBEANS	2015

Metodologia



NOAA

National Oceanic and Atmospheric
Administration -
Global Historical Climatology Network
(GHCN)



<ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/>

Contém dados diários de estações meteorológicas de diversos países num período maior a **175 anos**.


Index of /pub/data/ghcn/







[parent directory]

Name	Size	Date Modified
alaska-temperature-anomalies.txt	10.2 kB	3/24/08, 9:00:00 PM
alaska-temperature-means.txt	3.3 kB	3/11/08, 9:00:00 PM
anom/		11/9/11, 10:00:00 PM
blended/		6/18/19, 7:46:00 AM
daily/		6/30/19, 12:45:00 AM
forts/		12/14/09, 10:00:00 PM
grid_gpcp_1979-2002.dat	14.3 MB	9/10/03, 9:00:00 PM
Lawrimore-ISTI-30Nov11.ppt	3.7 MB	11/29/11, 10:00:00 PM
snow/		2/26/13, 9:00:00 PM
v1/		8/21/01, 9:00:00 PM
v2/		12/16/18, 10:00:00 PM
v3/		6/30/19, 6:34:00 AM
v4/		6/30/19, 9:42:00 AM

Metodologia

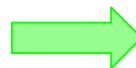
 [parent directory]

Name	Size	Date Modified
 USC00011084.dly	1.6 MB	6/30/19, 12:46:00 AM
 USC00012813.dly	2.6 MB	6/30/19, 12:46:00 AM
 USC00013160.dly	1.3 MB	6/30/19, 12:46:00 AM
 USC00013511.dly	2.0 MB	6/30/19, 12:46:00 AM

60.811 arquivos .csv, um para cada estação meteorológica dos EUA.



Penne (2019)



*“A tool to interface with and **download** Global Historical Climatology Network (GHCN) data into easily readable **CSVs**.”*

Metodologia

Estação climática "USS0017B04S" do estado de Washington.

				ID	TMAX	TMAX_FLAGS	TMIN	TMIN_FLAGS	TOBS	TOBS_FLAGS	TAVG	TAVG_FLAGS	PRCP	PRCP_FLAGS	WESD	WESD_FLAGS	SNWD	SNWD_FLAGS
MM/DD/YYYY	YEAR	MONTH	DAY															
1986-06-23	1986	6	23	USS0017B04S	268.0	__T	138.0	__T	163.0	__T	198.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-24	1986	6	24	USS0017B04S	270.0	__T	159.0	_IT	139.0	_IT	206.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-25	1986	6	25	USS0017B04S	246.0	__T	129.0	__T	138.0	__T	185.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-26	1986	6	26	USS0017B04S	206.0	__T	128.0	__T	142.0	__T	159.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-27	1986	6	27	USS0017B04S	269.0	__T	140.0	__T	260.0	__T	192.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-29	1986	6	29	USS0017B04S	83.0	__T	49.0	__T	191.0	_IT	70.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-30	1986	6	30	USS0017B04S	83.0	_IT	49.0	__T	248.0	__T	70.0	__T	NaN	NaN	NaN	NaN	NaN	NaN

Metodologia

Estação climática "USS0017B04S" do estado de Washington.

ID					TMAX	TMAX_FLAGS	TMIN	TMIN_FLAGS	TOBS	TOBS_FLAGS	TAVG	TAVG_FLAGS	PRCP	PRCP_FLAGS	WESD	WESD_FLAGS	SNWD	SNWD_FLAGS
MM/DD/YYYY	YEAR	MONTH	DAY															
1986-06-23	1986	6	23	USS0017B04S	268.0	__T	138.0	__T	163.0	__T	198.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-24	1986	6	24	USS0017B04S	270.0	__T	159.0	_IT	139.0	_IT	206.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-25	1986	6	25	USS0017B04S	246.0	__T	129.0	__T	138.0	__T	185.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-26	1986	6	26	USS0017B04S	206.0	__T	128.0	__T	142.0	__T	159.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-27	1986	6	27	USS0017B04S	269.0	__T	140.0	__T	260.0	__T	192.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-29	1986	6	29	USS0017B04S	83.0	__T	49.0	__T	191.0	_IT	70.0	__T	NaN	NaN	NaN	NaN	NaN	NaN
1986-06-30	1986	6	30	USS0017B04S	83.0	_IT	49.0	__T	248.0	__T	70.0	__T	NaN	NaN	NaN	NaN	NaN	NaN

Presente em todas as
tabelas

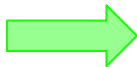
Colunas variáveis por estação

Metodologia

Pré-processamento dos dados

Metodologia

60.811 arquivos .csv
com colunas **distintas**.



Normalização das tabelas

TMAX (temperatura máxima em d°C)
TMIN (temperatura mínima em d°C)
TAVG (temperatura média em d°C)
PRCP (precipitação em dmm)
ID (identificação da estação climática)
MM/DD/YYYY (renomeada para "**Date**")
YEAR
MONTH
DAY



60.811 arquivos .csv
com colunas **iguais**.



Metodologia

Estação climática "USS0017B04S" do estado de Washington após a normalização das colunas.

				ID	TMAX	TMIN	TAVG	PRCP
Date	YEAR	MONTH	DAY					
1986-06-23	1986-01-01	6	23	USS0017B04S	268.0	138.0	198.0	NaN
1986-06-24	1986-01-01	6	24	USS0017B04S	270.0	159.0	206.0	NaN
1986-06-25	1986-01-01	6	25	USS0017B04S	246.0	129.0	185.0	NaN
1986-06-26	1986-01-01	6	26	USS0017B04S	206.0	128.0	159.0	NaN
1986-06-27	1986-01-01	6	27	USS0017B04S	269.0	140.0	192.0	NaN
1986-06-29	1986-01-01	6	29	USS0017B04S	83.0	49.0	70.0	NaN
1986-06-30	1986-01-01	6	30	USS0017B04S	83.0	49.0	70.0	NaN

Metodologia

Esquema da estrutura das tabelas de dados climáticos

Estação #N	Date					TAVG		PRCP	
	YEAR	MONTH	DAY
Estação #1	Date	YEAR	MONTH	DAY	...	TAVG	PRCP	...	
	1986-06-23	1986	06	23	...	198.0	NaN	...	
	1986-06-24	1986	06	24	...	206.0	NaN	...	
	1986-06-25	1986	06	25	...	185.0	NaN	...	
	1986-06-26	1986	06	26	...	159.0	NaN	...	
	1986-06-27	1986	06	27	...	192.0	NaN	...	
	1986-06-29	1986	06	29	...	70.0	NaN	...	
	

(60.811 tabelas)

Metodologia



Filtragem das tabelas por estados norte-americanos com produção de soja

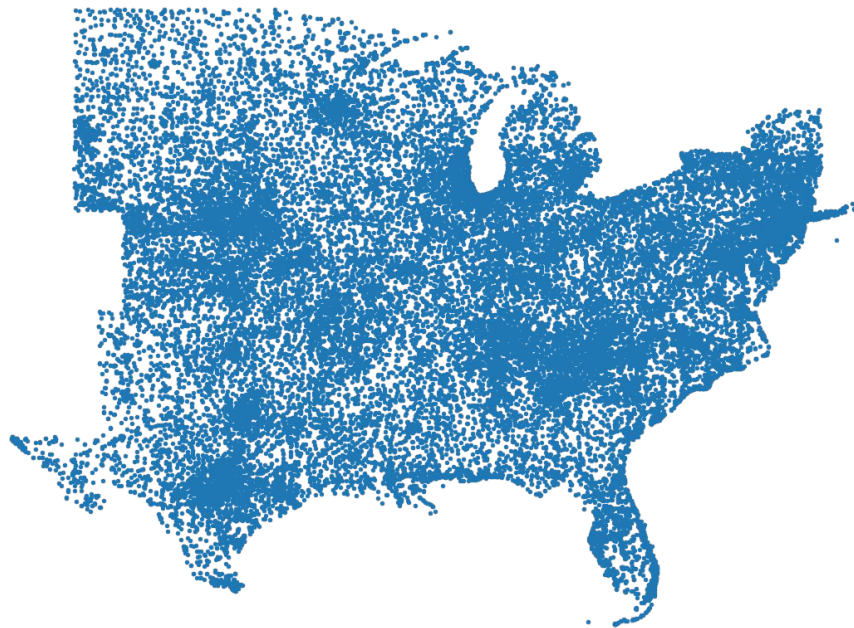
Segundo os dados da USDA-NASS, **31** estados norte-americanos apresentaram produção de soja desde 2001.

'AL', 'AR', 'DE', 'FL', 'GA', 'IL', 'IN', 'IA', 'KS', 'KY', 'LA', 'MD', 'MI', 'MN', 'MS', 'MO', 'NE', 'NJ', 'NY', 'NC', 'ND', 'OH', 'OK', 'PA', 'SC', 'SD', 'TN', 'TX', 'VA', 'WV' e 'WI'.

Com esta informação, foram selecionadas apenas as estações meteorológicas que estão localizadas nesses estados, resultando num total de **38.802 estações**.

Metodologia

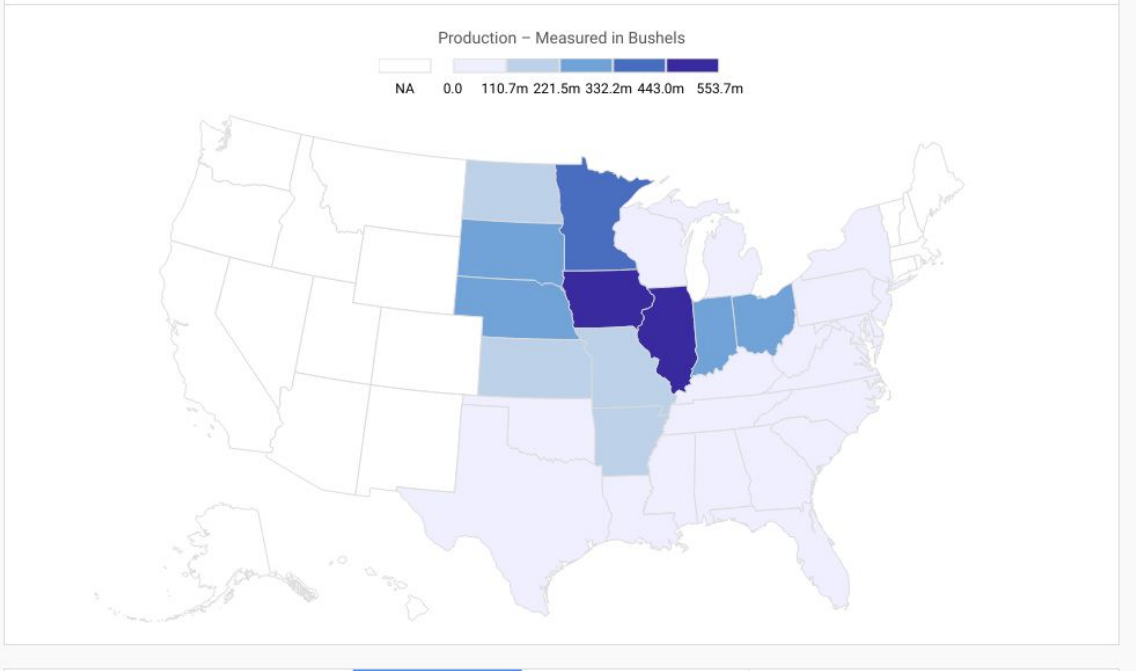
Mapa com a localização das estações meteorológicas para os estados norte-americanos com produção de soja.



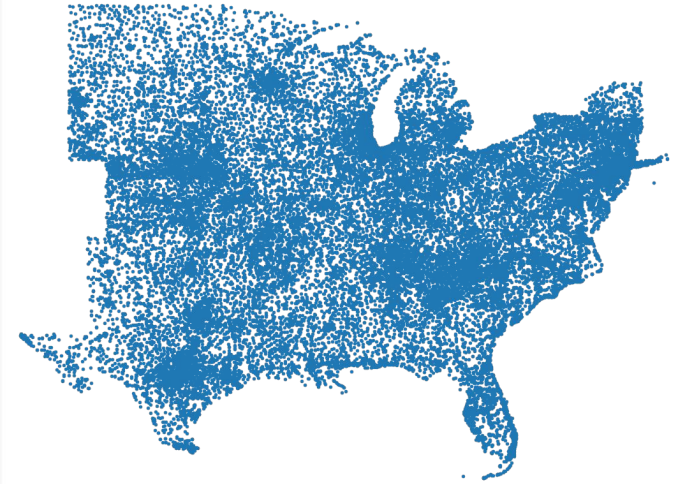
Cada ponto azul no mapa representa uma das 38.802 estações meteorológicas.

Metodologia

US States – 2015



Fonte: USDA, 2018b.



Metodologia

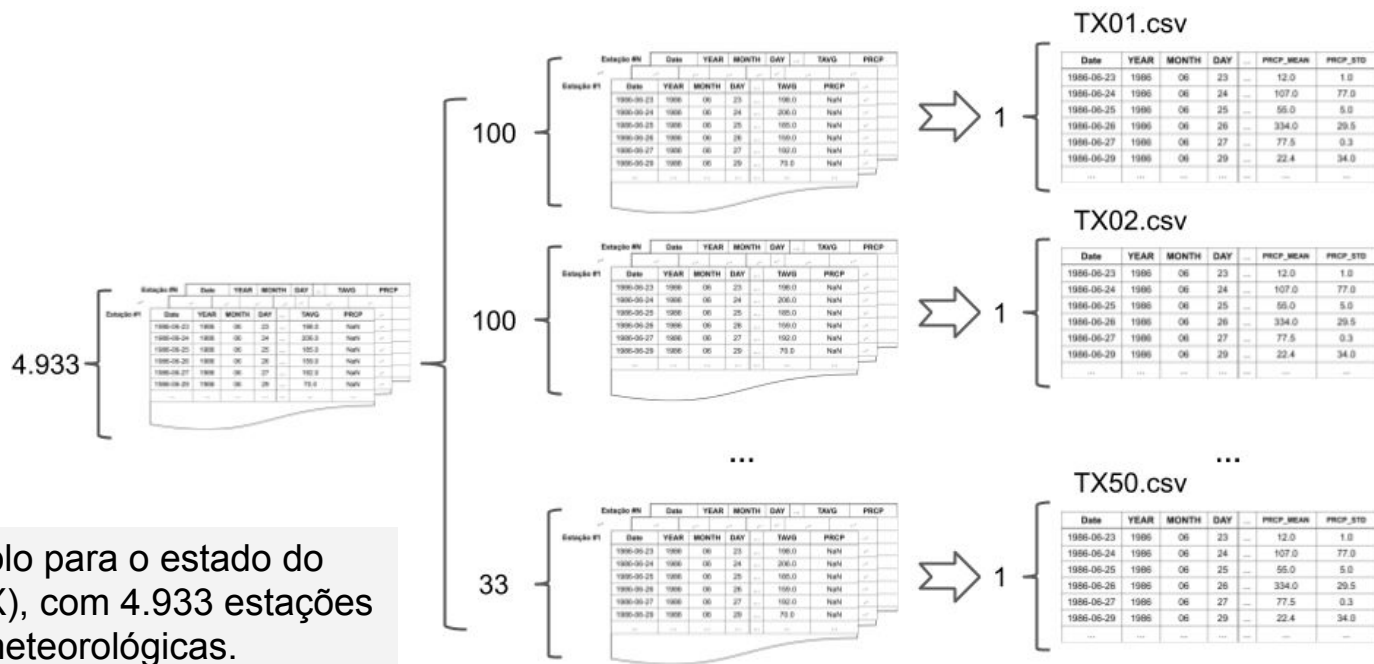
Fazendo um cálculo rápido:

38.802 tabelas de dados climáticos × **4 colunas** de interesse = **155.208 colunas**

Como a disponibilidade de memória RAM da infra-estrutura utilizada é limitada, optou-se por **agrupar estas 38.802 tabelas por estados norte-americanos**, calculando a **média e desvio padrão** de cada uma das 4 colunas de interesse ("TMAX", "TMIN", "TAVG" e "PRCP") entre as tabelas de cada estado.

Metodologia

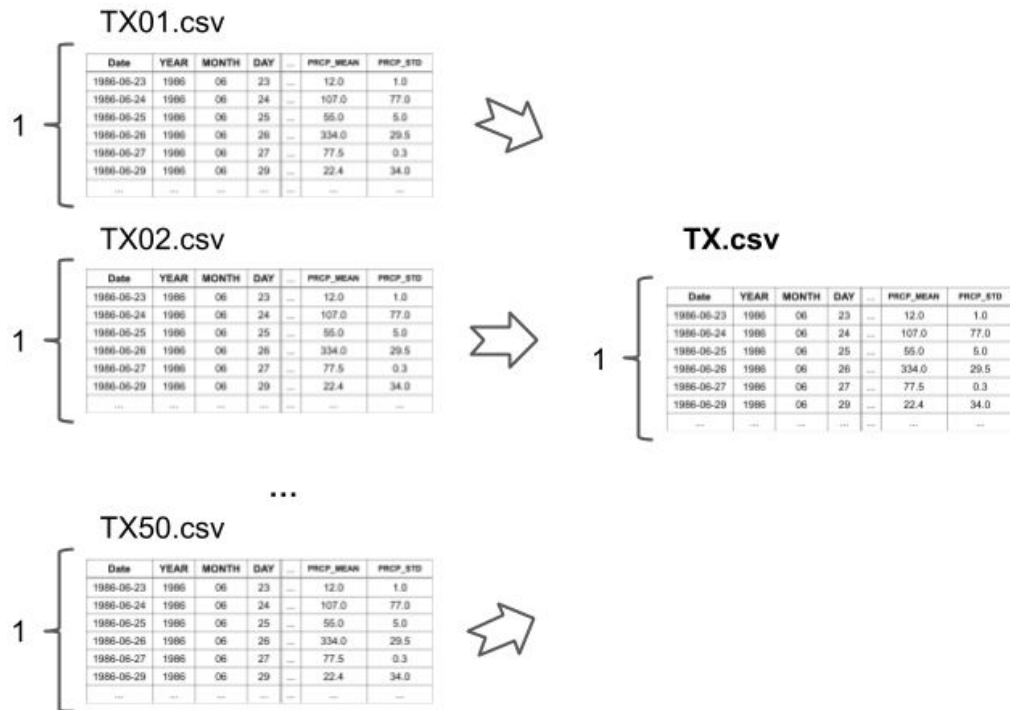
Processamento em lotes de 100 tabelas por vez, calculando a média e o desvio padrão para cada uma das quatro colunas de interesse ("TMAX", "TMIN", "TAVG" e "PRCP")



Exemplo para o estado do Texas (TX), com 4.933 estações meteorológicas.

Metodologia

Processamento das tabelas agrupadas anteriormente.



Metodologia

Após esta etapa, obtiveram-se **31 tabelas** referentes aos 31 estados.

Estas foram reunidas em uma única tabela para criar o *dataset climático*, adicionando o **prefixo com a sigla do estado** em cada uma das colunas de interesse.

Metodologia

Dataset climático após o pré-processamento.

				TX	TMAX_MEAN	TX_TMAX_STD	TX_TMIN_MEAN	TX_TMIN_STD	TX_TAVG_MEAN	TX_TAVG_STD	TX_PRCP_MEAN	TX_PRCP_STD	NC	TMAX_MEAN	NC_TMAX_S
Date	YEAR	MONTH	DAY												
2019-04-22	2019	4	22		72.627790	566.42426	44.583332	354.48105	18.318048	317.19855	0.000000	0.000000		NaN	N
2019-04-18	2019	4	18		54.506924	582.77580	27.069391	301.68390	8.202712	267.15765	77.998770	722.597100		53.402460	434.612
2019-04-17	2019	4	17		62.617880	641.38715	37.206060	412.67450	11.757029	343.05920	0.546157	26.782059		48.102604	403.454
2019-04-16	2019	4	16		65.986390	662.37030	30.363200	340.81598	11.515528	328.71072	0.009583	1.444365		39.248573	341.688
2019-04-15	2019	4	15		59.496593	612.05500	17.168463	219.85052	10.452809	296.80405	0.042482	2.736362		41.162914	356.904
2019-04-12	2019	4	12		54.676180	586.09040	18.959663	278.85687	8.596012	259.29416	0.423592	20.780325		44.908268	384.148
2019-04-11	2019	4	11		69.903590	715.77690	28.139763	354.06543	10.630163	314.85098	0.041998	4.000743		46.239610	391.194

TX

NC

Metodologia

Dataset climático após o pré-processamento.

				TX_TMAX_MEAN	TX_TMAX_STD	TX_TMIN_MEAN	TX_TMIN_STD	TX_TAVG_MEAN	TX_TAVG_STD	TX_PRCP_MEAN	TX_PRCP_STD	NC_TMAX_MEAN	NC_TMAX_S
Date	YEAR	MONTH	DAY										
2019-04-22	2019	4	22	72.627790	566.42426	44.583332	354.48105	18.318048	317.19855	0.000000	0.000000	NaN	N
2019-04-18	2019	4	18	54.506924	582.77580	27.069391	301.68390	8.202712	267.15765	77.998770	722.597100	53.402460	434.612
2019-04-17	2019	4	17	62.617880	641.38715	37.206060	412.67450	11.757029	343.05920	0.546157	26.782059	48.102604	403.454
2019-04-16	2019	4	16	65.986390	662.37030	30.363200	340.81598	11.515528	328.71072	0.009583	1.444365	39.248573	341.688
2019-04-15	2019	4	15	59.496593	612.05500	17.168463	219.85052	10.452809	296.80405	0.042482	2.736362	41.162914	356.904
2019-04-12	2019	4	12	54.676180	586.09040	18.959663	278.85687	8.596012	259.29416	0.423592	20.780325	44.908268	384.148
2019-04-11	2019	4	11	69.903590	715.77690	28.139763	354.06543	10.630163	314.85098	0.041998	4.000743	46.239610	391.194

64.044 linhas × 248 colunas

Registros diários desde **01/07/1836** até **22/04/2019**

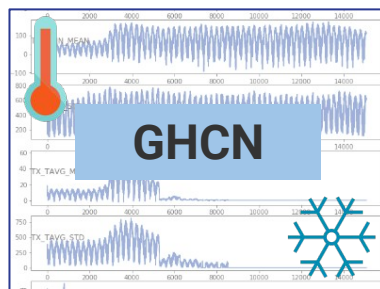
Metodologia

União dos dados climáticos e cotações para a criação do DatasetMarlon

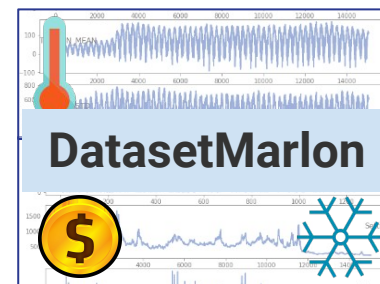


15.105 linhas
8 colunas
De 08/07/1959
Até 21/06/2019

(inner join)



64.044 linhas
248 colunas
De 01/07/1836
Até 22/04/2019



15.062 linhas
256 colunas
De **10/07/1959**
Até **22/04/2019**

Metodologia

DatasetMarlon, contendo dados diários de cotações de soja e dados climáticos dos estados norte-americanos produtores de soja

				Open	High	Low	Last	Change	Settle	Volume	Previous Day Open Interest	TX_TMAX_MEAN	TX_TMAX_STD	TX_TMIN_MEAN	TX_TMIN_STD	TX_TAVG_MEAN	TX_TAVG_STD	TX_PRCP_MEAN	TX_PRCP_STD
Date	YEAR	MONTH	DAY																
2019-04-22	2019	4	22	881.50	883.25	876.25	876.75	3.50	877.00	62527.0	205572.0	72.627790	566.42426	44.583332	354.48105	18.318048	317.19855	0.000000	0.000000
2019-04-18	2019	4	18	878.75	882.00	876.50	880.75	1.50	880.50	63485.0	214732.0	54.506924	582.77580	27.069391	301.68390	8.202712	267.15765	77.998770	722.597100
2019-04-17	2019	4	17	887.75	890.50	878.50	879.00	9.00	879.00	89706.0	219956.0	62.617880	641.38715	37.206060	412.67450	11.757029	343.05920	0.546157	26.782059
2019-04-16	2019	4	16	898.25	899.00	886.25	888.00	10.75	888.00	92852.0	221960.0	65.986390	662.37030	30.363200	340.81598	11.515528	328.71072	0.009583	1.444365
2019-04-15	2019	4	15	895.00	902.00	894.75	898.75	3.50	898.75	91118.0	232341.0	59.496593	612.05500	17.168463	219.85052	10.452809	296.80405	0.042482	2.736362
2019-04-12	2019	4	12	895.00	898.50	893.75	894.50	NaN	895.25	69411.0	242526.0	54.676180	586.09040	18.959663	278.85687	8.596012	259.29416	0.423592	20.780325
2019-04-11	2019	4	11	901.25	904.00	893.50	895.75	6.75	895.25	85699.0	253700.0	69.903590	715.77690	28.139763	354.06543	10.630163	314.85098	0.041998	4.000743

CBOT
(8 colunas)

GHCN
(248 colunas)

Metodologia

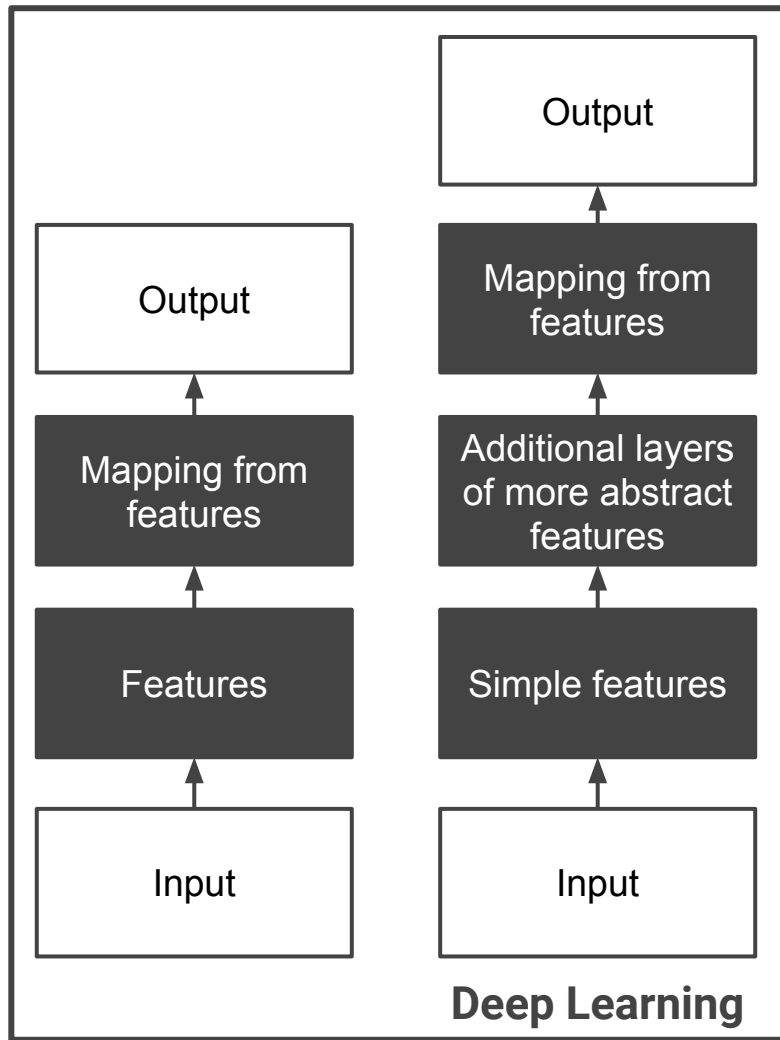
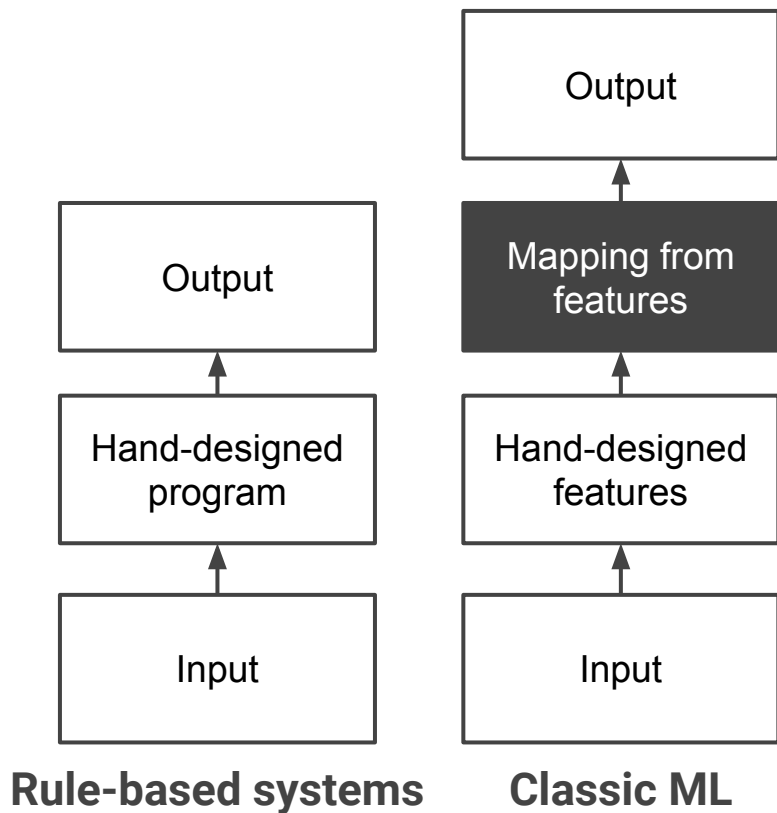
LSTM

Long Short-Term Memory

Proposto por Hochreiter et. al. (1997), é um tipo de rede neural pertencente à família de Redes neurais recorrentes (RNN).

Segundo Géron (2017), RNN's podem analisar dados de séries temporais, como preços de ações, e informar quando comprar ou vender.

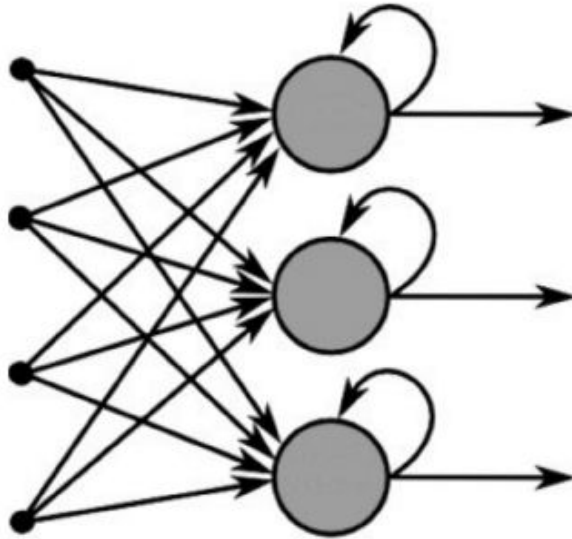
Por que LSTM?



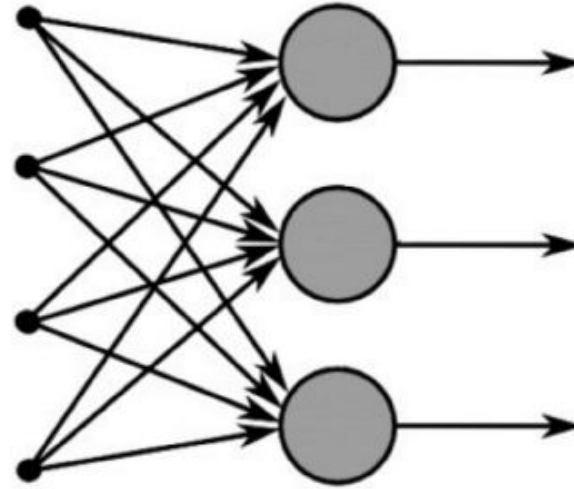
Representation Learning

Fonte: Adaptado de Goodfellow, 2016.

Metodologia



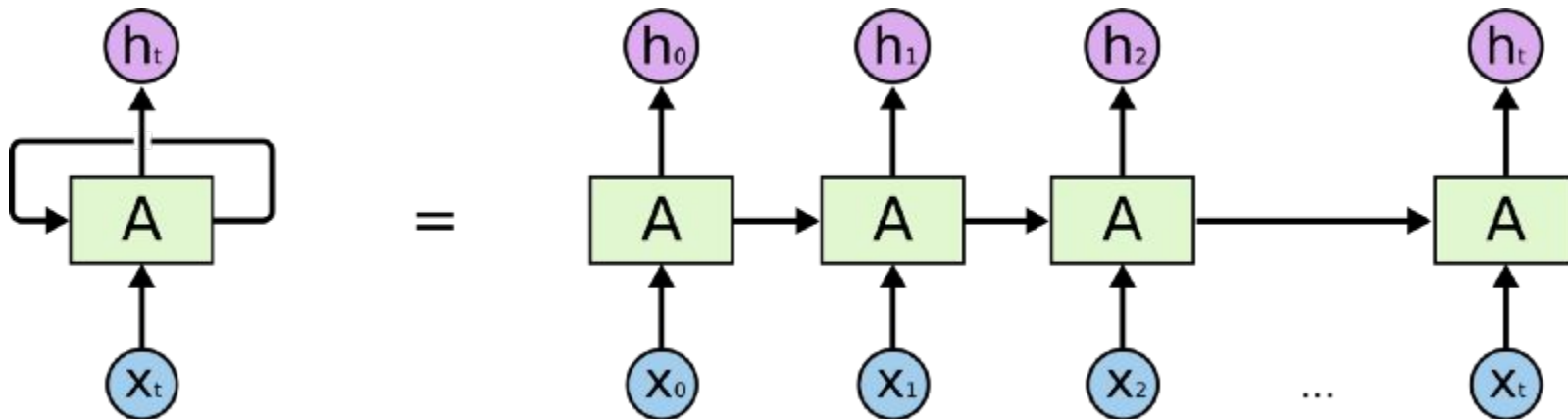
Recurrent Neural Network



Feed-Forward Neural Network

Fonte: Donges, 2018.

Metodologia



Fonte: Olah, 2015.

Metodologia

Legenda



Camada de
rede neural



Operação
ponto-a-
ponto



Transferência
de vetor

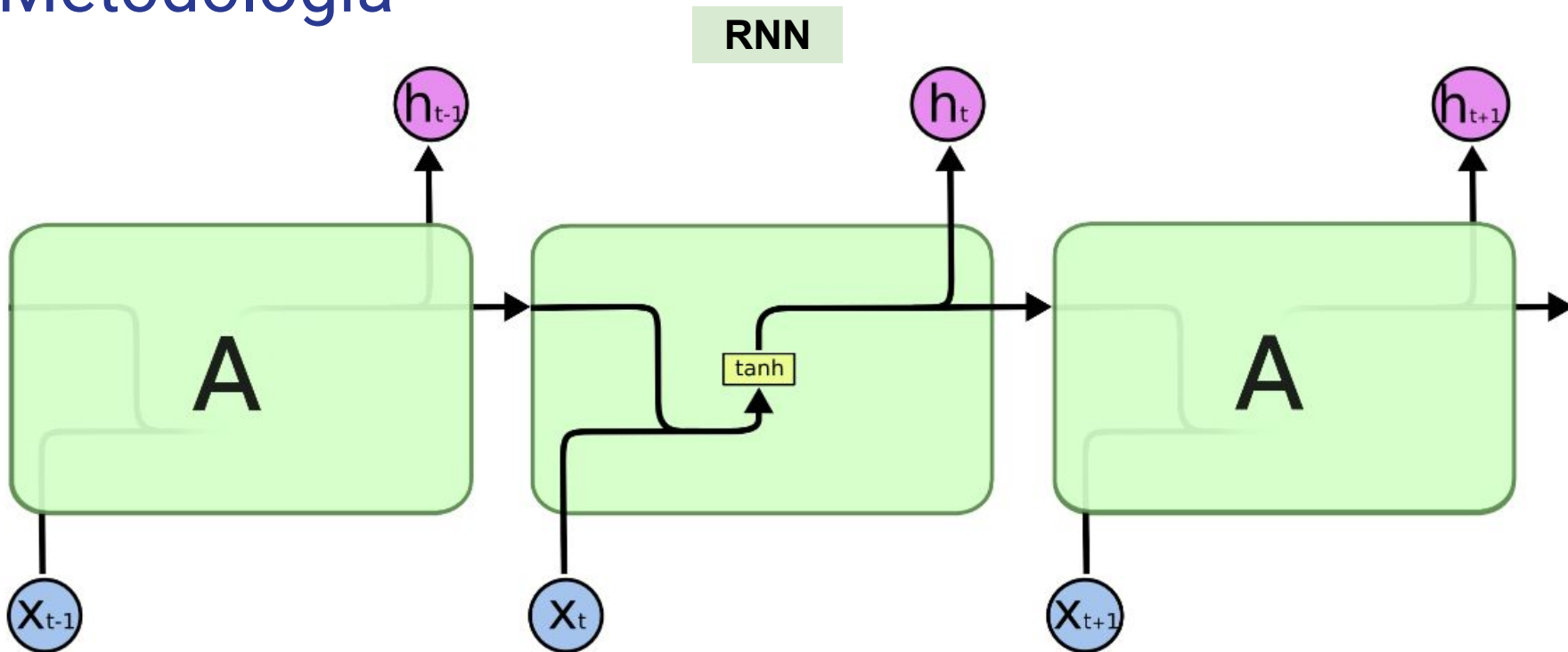


Concatenação



Cópia

Metodologia



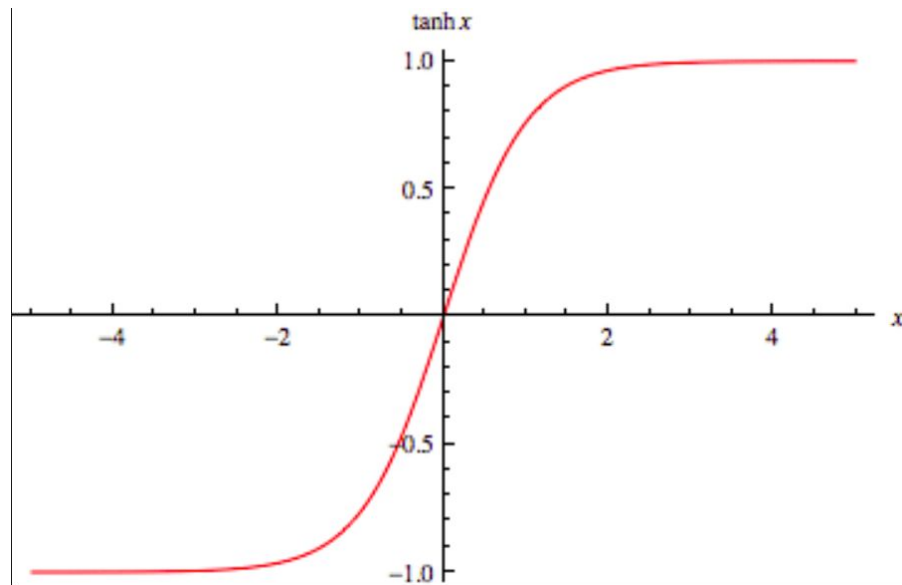
Fonte: Olah, 2015.

Metodologia

tanh

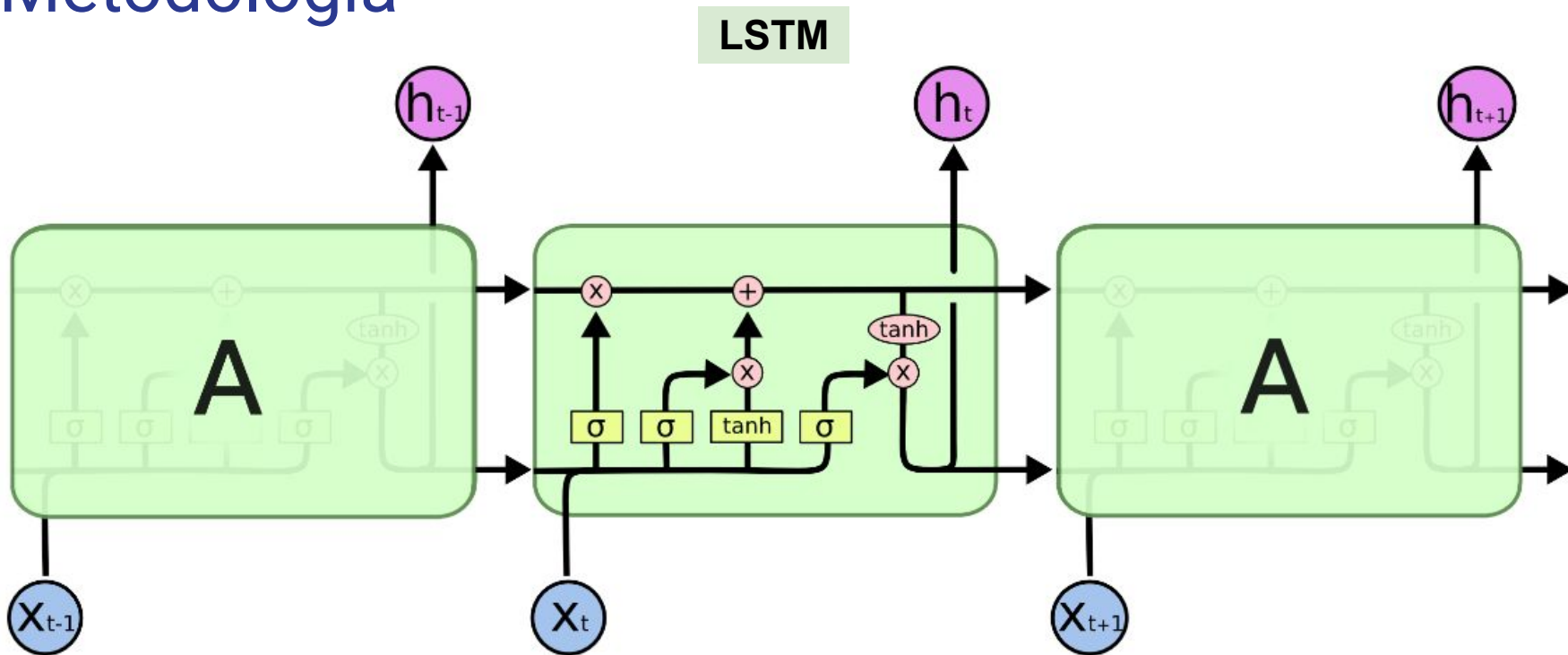
A função de ativação **tanh** possui a forma de "S" e seu valor varia de -1 a 1 , o que tende a fazer com que a saída desta camada seja centrada ao redor do zero (diferente da função *Sigmoid*) no início do treinamento *.

$$f(x) = \tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$$



* Segundo Géron (2017), isso geralmente ajuda a acelerar a convergência.

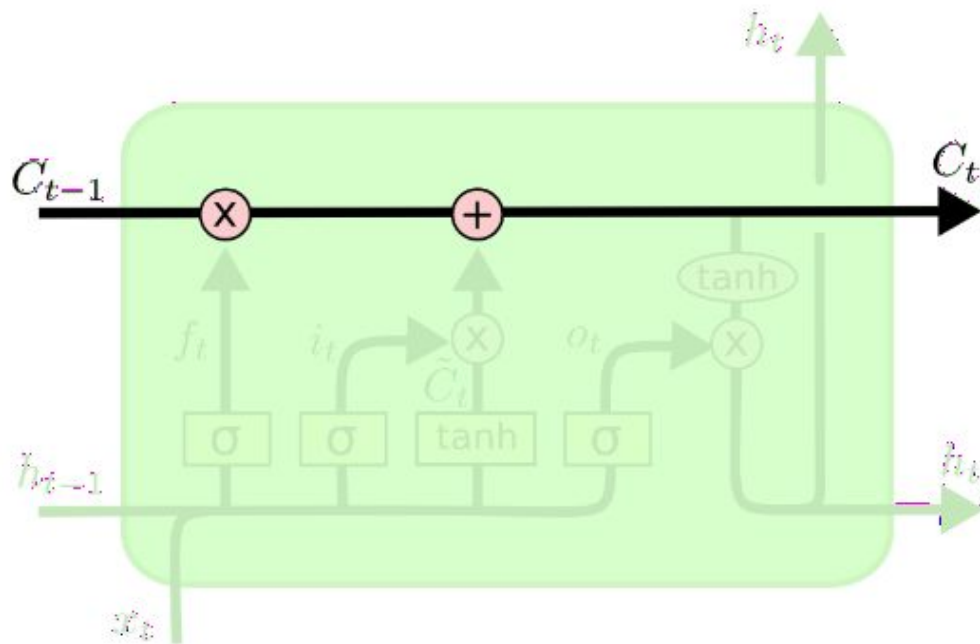
Metodologia



Fonte: Olah, 2015.

Metodologia

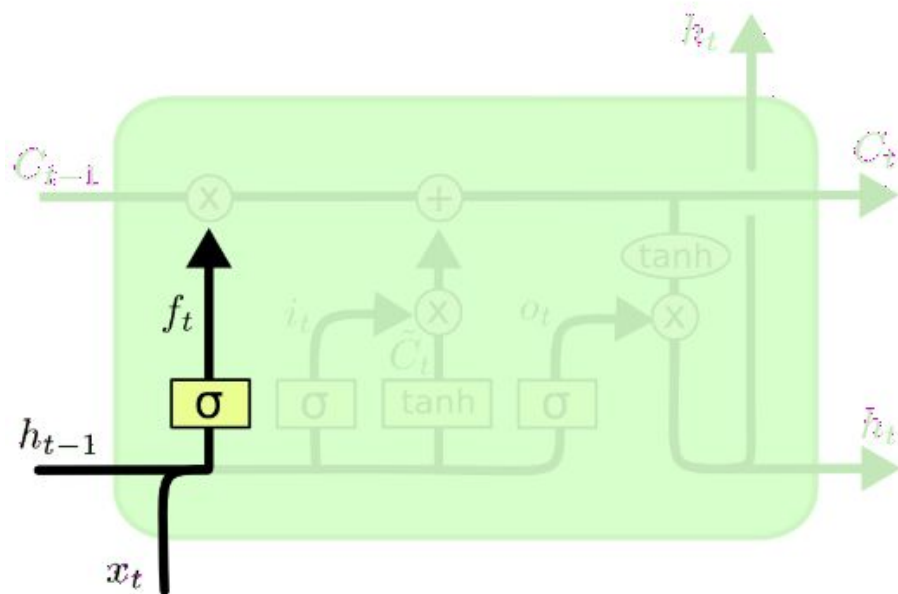
Cell state



O estado da célula (*cell state*) é a memória interna da célula LSTM.

Metodologia

Forget gate

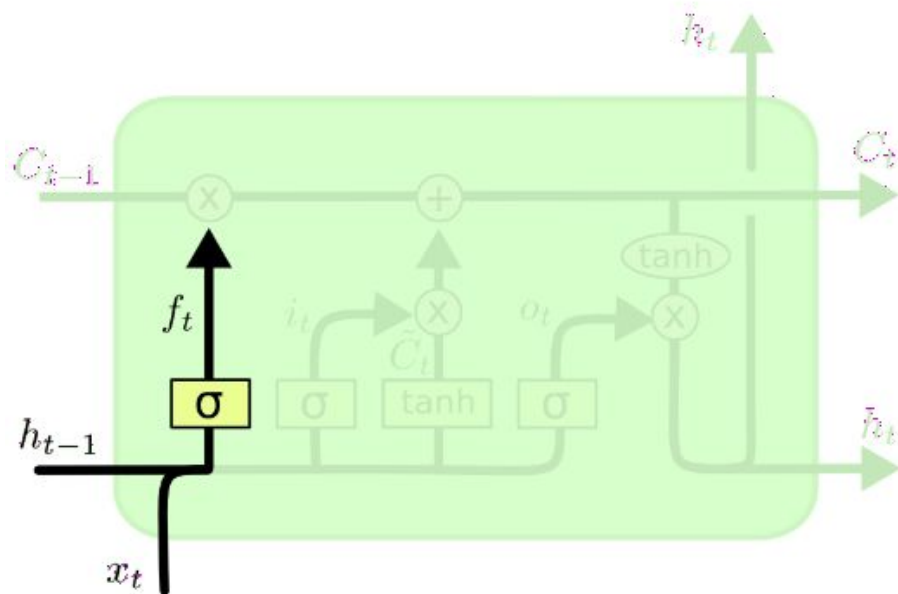


Decide qual informação deve ser descartada do *cell state*.

$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

Metodologia

Forget gate



função de ativação Sigmoid

forget gate

vetor de entrada do passo atual

$$f_t = \sigma (W_f \cdot [h_{t-1}, x_t] + b_f)$$

peso

bias

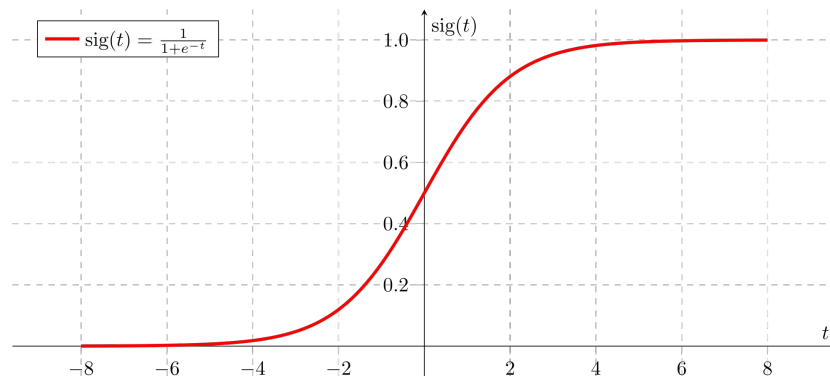
vetor de saída do passo anterior

Metodologia

σ

Segundo Olah (2015), a função **Sigmoid** retorna números entre 0 e 1, determinando o quanto cada sinal deve ser transmitido.

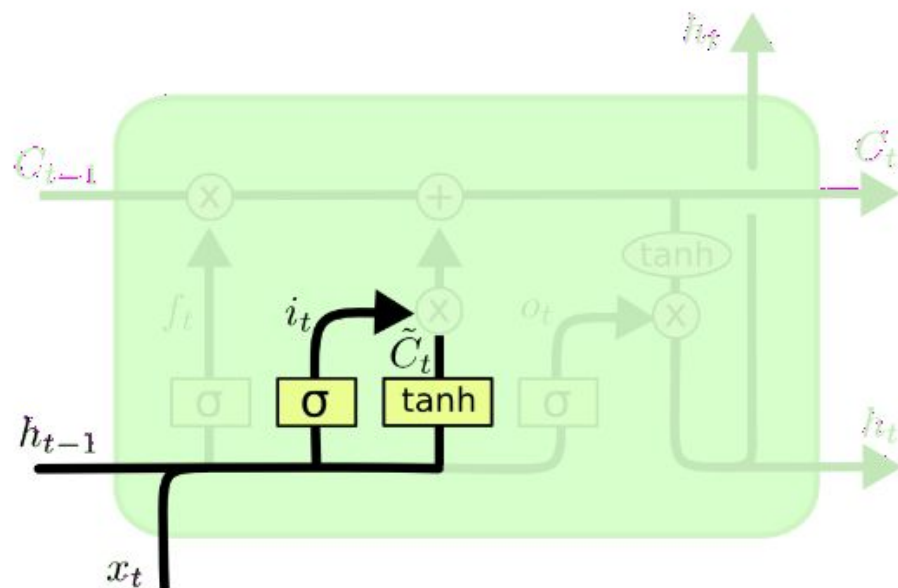
- **0** significa que o sinal não deve ser transmitido;
- **1** significa que deve ser transmitido totalmente.



$$f(x) = \sigma(x) = \frac{1}{1 + e^{-x}}$$

Metodologia

Input gate



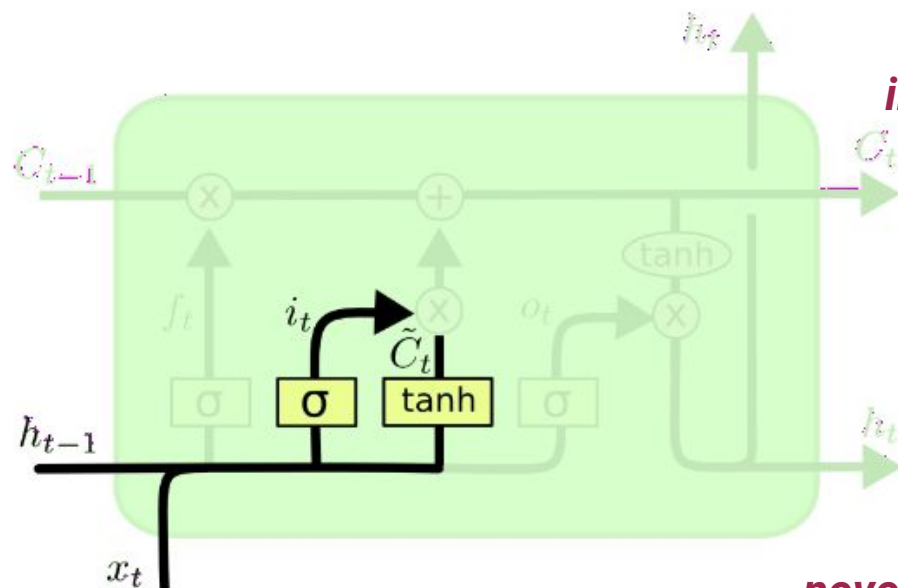
Sua função é definir o quanto cada informação deve ser atualizada no *cell state*.

$$i_t = \sigma (W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

Metodologia

Input gate



input gate

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

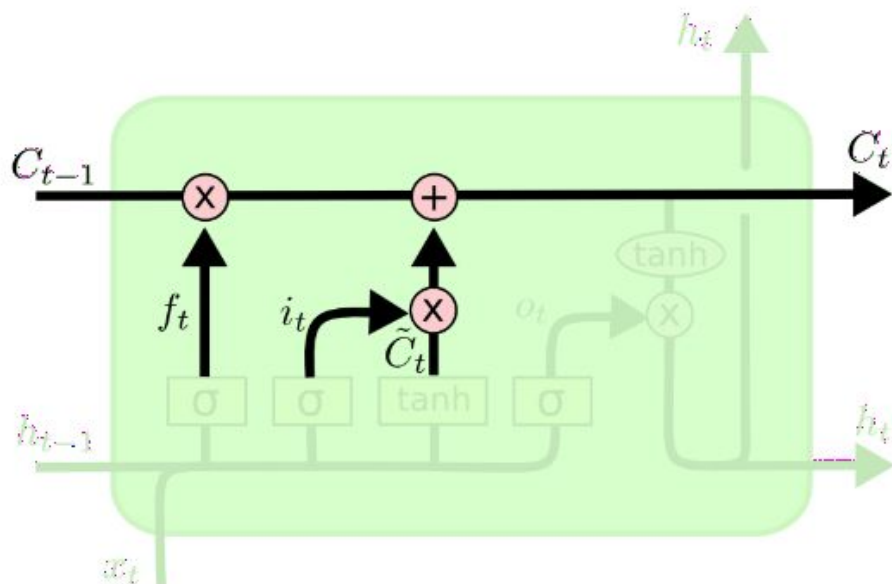
função de ativação tanh

**novos valores para
cell state**

Fonte: Olah, 2015.

Metodologia

Atualizando o estado da célula (Cell state)



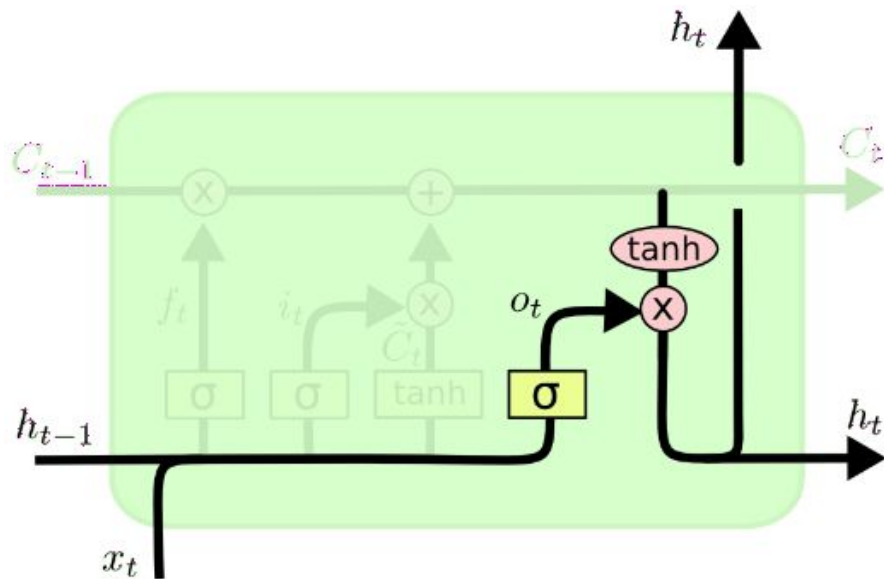
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

forget gate points to f_t
input gate points to i_t

cell state atual points to C_{t-1}
cell state anterior points to C_t
novos valores para cell state points to \tilde{C}_t

Metodologia

Output gate



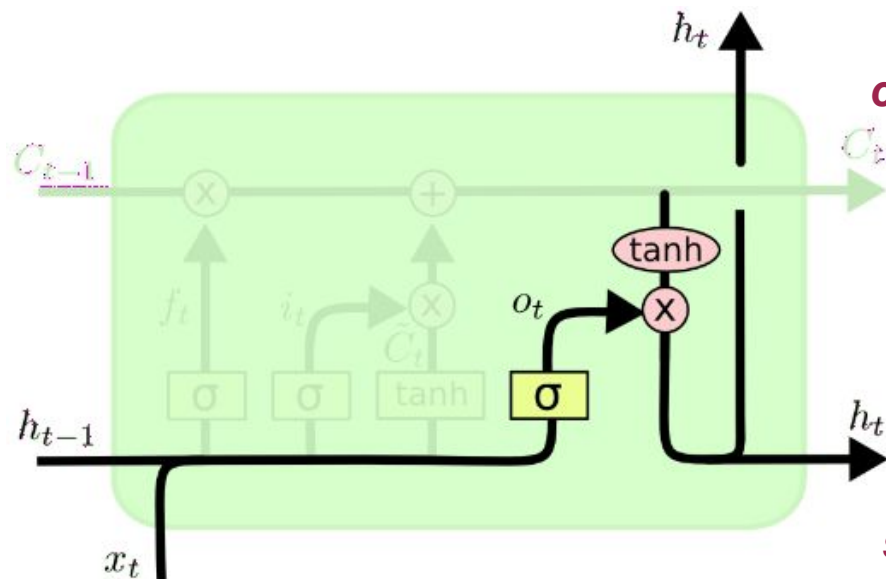
Sua função é definir o quais valores devem fazer parte da saída h_t .

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

Metodologia

Output gate



output gate

$$o_t = \sigma (W_o [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t * \tanh (C_t)$$

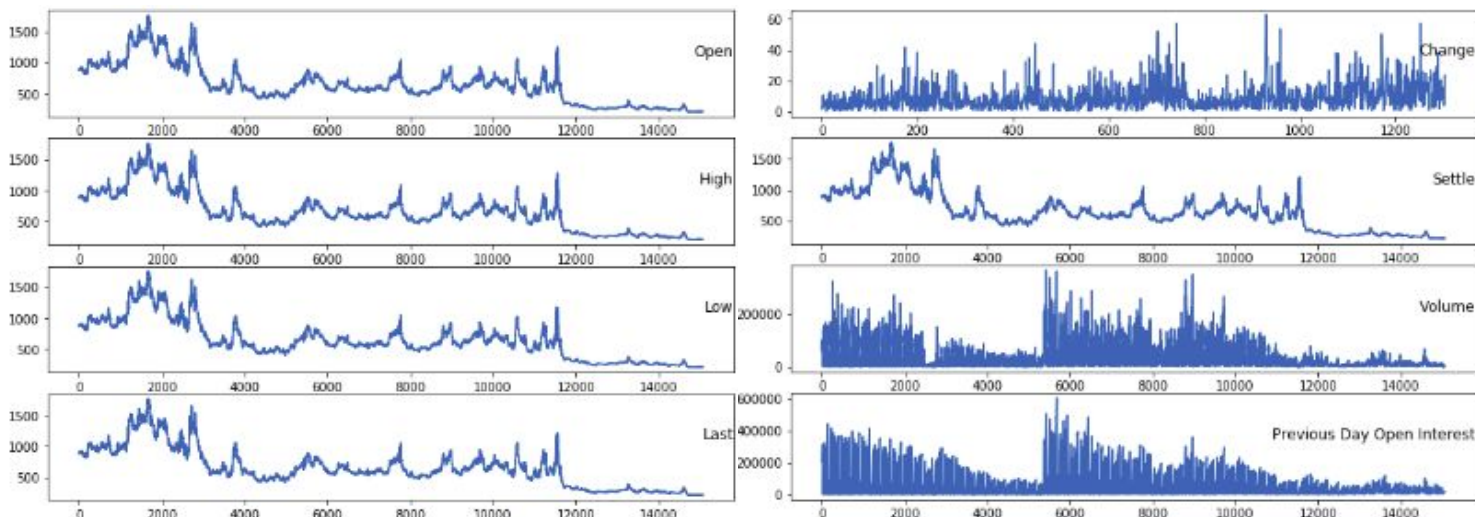
saída atual

**cell state
atual**

Resultados

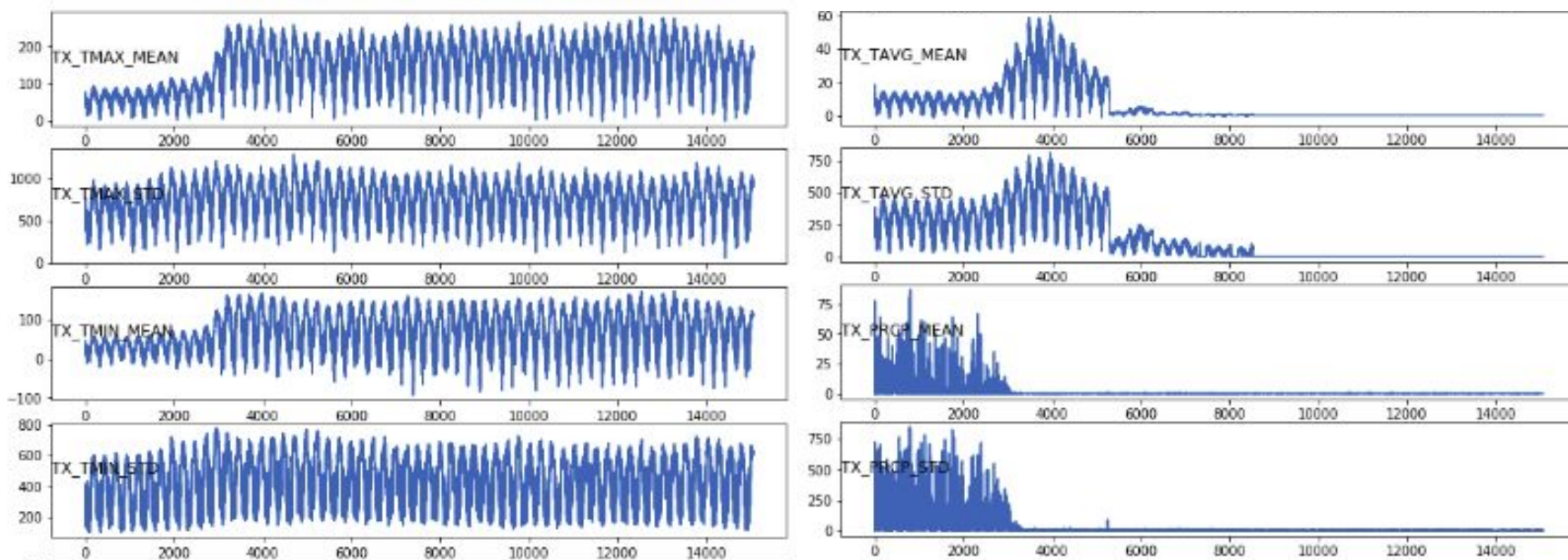
Resultados

DatasetMarlon: gráfico das variáveis vindas do dataset das cotações no período de 10/07/1959 à 22/04/2019, sendo Settle a variável alvo.



Resultados

DatasetMarlon: gráfico das variáveis vindas do dataset de dados climáticos no período de 10/07/1959 à 22/04/2019 filtrado para o estado do Texas.



Resultados

Normalização dos dados

Foram eliminados os registros que possuíam alguma das colunas com valor nulo (NaN), e normalizados os dados para estarem no intervalo entre **0 e 1**. Para a normalização, foi utilizada a classe *MinMaxScaler* da biblioteca *pandas*.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Resultados

Deslocamento dos dados

Foi realizado um deslocamento nos dados para que fossem criadas colunas referentes a cada um dos 7 dias anteriores.

	var1(t-7)	var2(t-7)	var3(t-7)	var4(t-7)	var5(t-7)	var6(t-7)	var7(t-7)		var253(t-1)	var254(t-1)	var255(t-1)	var256(t-1)	var6(t)
8	0.430163	0.431369	0.429128	0.431731	0.000359	0.431572	0.171927		0.442540	0.142258	0.037625	0.254125	0.443221
9	0.435911	0.436790	0.430404	0.430614	0.002942	0.430614	0.242938	...	0.470187	0.265425	0.000550	0.006923	0.443221
10	0.442617	0.442212	0.435350	0.436359	0.003516	0.436359	0.251458		0.477043	0.300327	0.036553	0.085632	0.443381
11	0.440542	0.444126	0.440775	0.443221	0.001148	0.443221	0.246762		0.452272	0.191550	0.024786	0.053430	0.448168
12	0.440542	0.441893	0.440137	0.440508	0.669824	0.440987	0.187976		0.443218	0.145499	0.070249	0.187655	0.443221



**Settle à 7 dias
atrás
(t-7)**



**Settle no
tempo atual
(t)**

Resultados

Após a normalização e o deslocamento, o DatasetMarlon passou a ter **10.991** linhas e **1.793** colunas.

Estes dados foram separados em dois *datasets* (um para treino e um para teste) numa proporção de **0,87**.

Dataset de treino:
9.562 linhas e 1.793 colunas.

Dataset de teste:
1.429 linhas e 1.793 colunas.

Resultados

Design do modelo preditivo

128 neurônios LSTM

lstm1: LSTM	input:	(None, 7, 256)
	output:	(None, 7, 128)

128 neurônios LSTM

lstm2: LSTM	input:	(None, 7, 128)
	output:	(None, 128)

1 neurônio com função
de ativação linear

output: Dense	input:	(None, 128)
	output:	(None, 1)

Resultados

Loss function (função de perda): RMSE

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

Root-Mean-Square Error
(RMSE) calcula a magnitude média do erro entre a curva de cotação estimada e a curva real.

Resultados

Função de otimização: ADAM

Adapted Moment Estimation (KINGMA; BA, 2014)

É a função de otimização escolhida para realizar a atualização do peso W e bias b ao término de cada época, com taxa de aprendizado $\alpha = 0.001$.

Resultados

Treinamento

100 épocas
(*epochs*)

Lotes de tamanho **77**
(*batch size*)

dataset de treino

dataset de teste

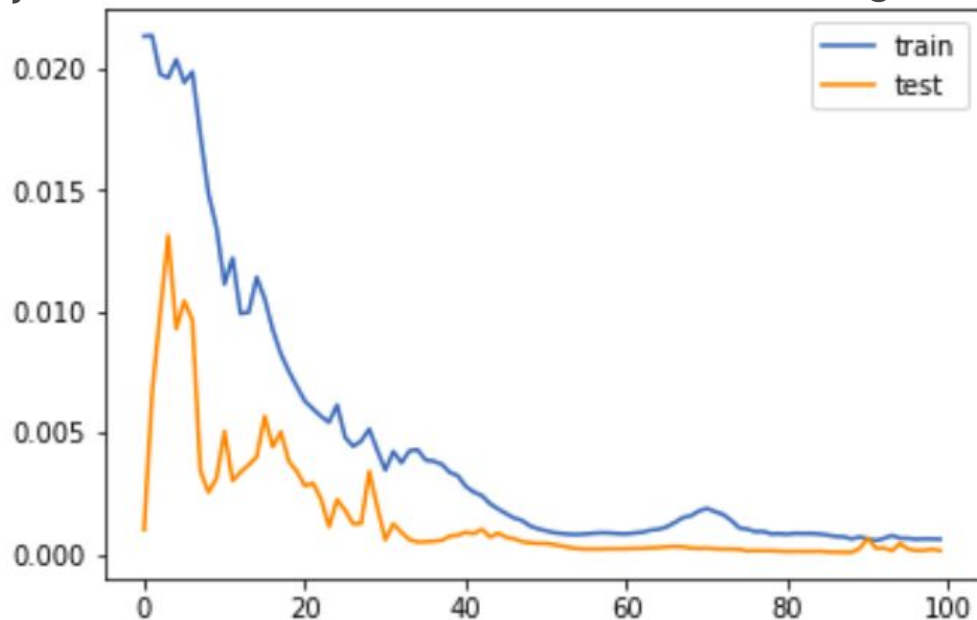
Train on 9562 samples, validate on 1429 samples *

* Fonte: <https://github.com/marlonrcfranco/soyforecast/blob/master/soyforecast.ipynb>

Resultados

Treinamento

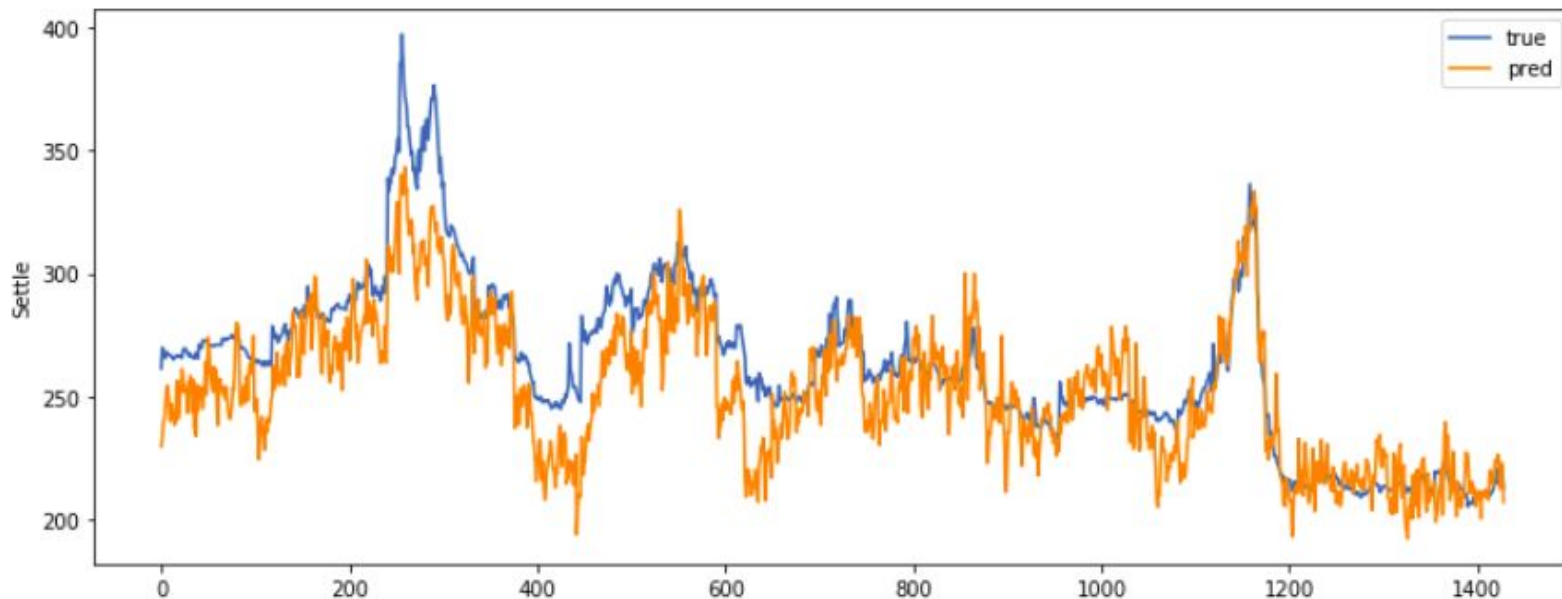
Evolução dos erros de teste e treinamento ao longo das épocas.



Resultados

Avaliação

Comparação entre a cotação predita para os dados de teste e a cotação real.



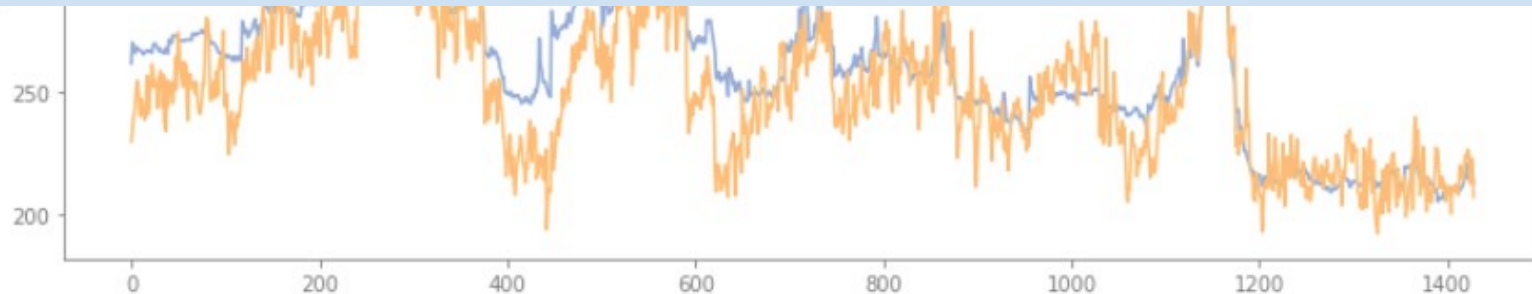
Resultados

Avaliação

Comparação entre a cotação predita para os dados de teste e a cotação real.



O RMSE calculado ao término da predição foi **18,789**, o que significa que o **erro médio** entre a cotação predita e a cotação real é de **US\$ 0,18789**.



Resultados

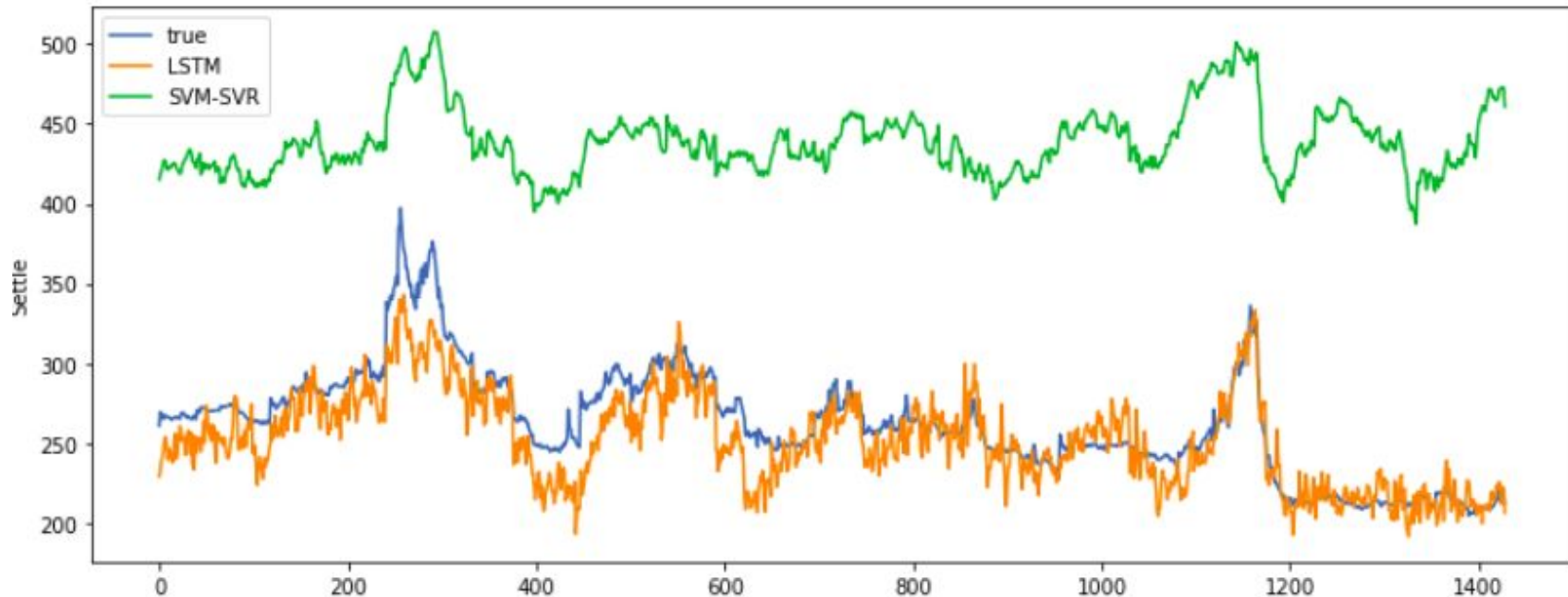
Comparação com um modelo SVM-SVR

Um modelo de *Machine Learning* clássico **SVM** para regressão (**SVR**) também foi treinado utilizando a biblioteca sklearn (SKLEARN. . . , 2019) e sobre o DatasetMarlon, com a mesma separação entre dados de treino e teste.

Resultados

Comparação com um modelo SVM-SVR

Comparação entre a cotação real e as cotações preditas pelos modelos LSTM e SVM-SVR.



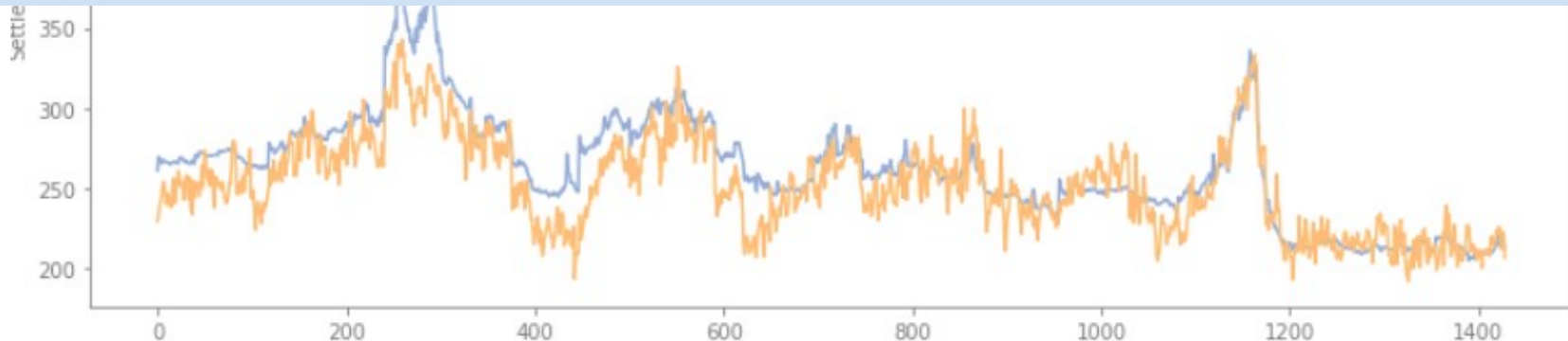
Resultados

Comparação com um modelo SVM-SVR

Comparação entre a cotação real e as cotações preditas pelos modelos LSTM e SVM-SVR.



O resultado da predição do modelo **SVM** foi avaliado seguindo a métrica RMSE, obtendo um erro médio de **178,381**.



Conclusão

Conclusão

A significância do Brasil no cenário mundial da produção e comercialização de soja assim como o fato das cotações de soja seguirem o padrão de séries temporais, motivou o desenvolvimento de um modelo preditivo capaz de prever cotações na Bolsa de Chicago (CBOT).

A relação entre o clima e a produção da soja, assim como a forte influência dos EUA nas cotações fez com que fosse explorada a união de dados históricos da Bolsa de Chicago com dados climáticos dos estados norte-americanos produtores de soja, a fim de ser criado um *dataset* reunindo estas informações denominado **DatasetMarlon**.

Conclusão

O modelo preditivo escolhido foi o **LSTM** (*Long Short-Term Memory*), um tipo de Rede Neural Recorrente (RNN) com uma memória interna que permite uma janela de aprendizado longa sobre uma série temporal.

A escolha pela utilização da LSTM se deu pelo fato de ser uma Rede Neural Profunda (**Deep Neural Network**), possuindo alto desempenho comparada a técnicas de Aprendizado de Máquina Clássico (**Machine Learning clássico**), conforme evidenciam os trabalho relacionados.

Conclusão

O modelo foi desenvolvido possuindo **duas camadas com 128** células (**neurônios**) **LSTM** cada, e uma camada de saída com apenas uma única célula com função de ativação linear, que não altera o valor.

Os dados de entrada foram separados em *dataset* de **treino** (com 9.562 amostras) e de **teste** (com 1.429 amostras), normalizados e organizados de modo a permitirem que o modelo realize a predição da cotação (variável "*Settle*") com **uma semana** de antecedência.

Conclusão

A predição realizada sobre os dados de teste foi avaliada utilizando a métrica *Root Mean-Square Error* (**RMSE**), que calculou um erro médio entre a curva de cotação estimada e a curva real de **US\$ 0,18789**.

Este resultado foi comparado ao resultado da predição realizada por um modelo de **Machine Learning clássico SVM para regressão**, que também foi treinado sobre o DatasetMarlon, com a mesma separação entre dados de treino e teste.

Conclusão

Comparando o RMSE entre os modelos **LSTM** e **SVM**, pôde-se verificar que o modelo **LSTM chegou mais próximo de prever a cotação da soja** do que o modelo SVM, evidenciando a melhor performance dos modelos de **Deep Learning** em relação aos de Machine Learning clássico.

O **DatasetMarlon**, o **modelo** treinado e a **documentação** deste trabalho se encontram disponíveis no GitHub: <https://github.com/marlonrcfranco/soyforecast/>

Referências

- CMEGROUP. Chicago Mercantile Exchange Chicago Board Of Trade. **Timeline of CME Achievements**. 2018. Disponível em: <<https://www.cmegroup.com/company/history/timeline-of-achievements.html>> Acesso em: 09 dez 2018.
- DEVRIES, Henry. **Breaking Down the Silos**. Ellucian. 2018. Disponível em: <<https://www.ellucian.com/emea-ap/Blog/Breaking-Down-the-Silos/>>. Acesso: 22 out. 2018.
- DONGES, Niklas. **Recurrent Neural Networks and LSTM**. Towards Data Science. 2018. Disponível em: <<https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5>>. Acesso: 22 out. 2018
- FRANCO, M. R. C. **marlonrcfranco/soyforecast**. 2019. Disponível em: <<https://github.com/marlonrcfranco/soyforecast>>. Acesso em: 23 jun 2019.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2017. ISBN 1491962291, 9781491962299.

Referências

GLOBAL HISTORICAL CLIMATOLOGY NETWORK. **Data File Access (FTP)**. 2019. Disponível em: <ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/>. Acesso em: 02 jun 2019.

GLOBAL HISTORICAL CLIMATOLOGY NETWORK. **Global Historical Climatology Network (GHCN)**. 2019. Disponível em: <https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/global-historical-climatology-network-ghcn>. Acesso em: 02 jun 2019.

GOOGLE COLABORATORY. **Welcome to Colaboratory!** 2019. Disponível em: <https://colab.research.google.com/notebooks/welcome.ipynb>. Acesso em: 02 jun 2019.

GOOGLE TENSORFLOW. **Recurrent Neural Networks**. 2018. Disponível em: <https://www.tensorflow.org/tutorials/sequences/recurrent#lstm>. Acesso em: 26 nov 2018.

GOODFELLOW, Ian. Bengio, Yoshua. Courville, Aaron. **Deep Learning**. MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Acesso: 22 out. 2018.

Referências

- HOCHREITER, S.; SCHMIDHUBER, J. **Long short-term memory**. Neural Computation, v. 9, n. 8, p. 1735–1780, 1997. Disponível em: <<https://doi.org/10.1162/neco.1997.9.8.1735>>.
- IMEA. Instituto Mato-grossense de Economia Agropecuária. **Entendendo o mercado da soja**. In: . 2017. v. 3, p. 01–48. Disponível em: <http://www.imea.com.br/upload/pdf/arquivos/2015_06_13_Paper_jornalistas_boletins_Soja_Versao_Final_AO.pdf>. Acesso em: 25 nov. 2018.
- KERAS. **Keras: The Python Deep Learning library**. 2019. Disponível em: <<https://keras.io/>>. Acesso em: 22 jun 2019.
- KINGMA, D. P.; BA, J. **Adam: A Method for Stochastic Optimization**. 2014.
- LI, Z.; TAM, V. **A comparative study of a recurrent neural network and support vector machine for predicting price movements of stocks of different volatilities**. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). [S.l.: s.n.], 2017. p. 1–8.

Referências

LIU, K. Chemistry and nutritional value of soybean components. In: Soybeans: Chemistry, Technology, and Utilization. Boston, MA: Springer US, 1997. p. 25–113. ISBN 978-1-4615-1763-4. Disponível em: <https://doi.org/10.1007/978-1-4615-1763-4_2>.

MCNALLY, S.; ROCHE, J.; CATON, S. **Predicting the price of bitcoin using machine learning.** In: 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). [S.l.: s.n.], 2018. p. 339–343. ISSN 2377-5750.

M.M.P.N.D. Multilingual Multiscript Plant Name Database. Mar., 2000. **Sorting Glycine names.** [S.l.]. Disponível em: <<http://www.plantnames.unimelb.edu.au/Sorting/Glycine.html#max>>. Acesso em: 25 nov. 2018.

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. **About our agency.** 2019. Disponível em: <<https://www.noaa.gov/about-our-agency>>. Acesso em: 02 jun 2019.

OLAH, Christopher. **Understanding LSTM Networks.** 2015. Colah's blog. Disponível em: <<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>>. Acesso: 11 dez. 2018.

Referências

PAGANO, M. C.; MIRANSARI, M. **1 - the importance of soybean production worldwide**.

In: MIRANSARI, M. (Ed.). Abiotic and Biotic Stresses in Soybean Production. San Diego: Academic Press, 2016. p. 1 – 26. ISBN 978-0-12-801536-0. Disponível em: <<http://www.sciencedirect.com/science/article/pii/B9780128015360000013>>.

(Pandas) MCKINNEY, W. **Python for Data Analysis**. 2th. ed. [S.l.]: O'Reilly Media, 2017. 550 p.

PENNE, A. **Get NOAA GHCN Data**. 2019. Disponível em:

<https://github.com/aaronpenne/get_noaa_ghcn_data>. Acesso em: 02 jun 2019.

PHAM, Xuan. Stack, Martin. **How data analytics is transforming agriculture**. Business Horizons, Volume 61, Issue 1, 2018. Pages 125-133. DOI: <<https://doi.org/10.1016/j.bushor.2017.09.011>>.

PROJECT JUPYTER. **The Jupyter Notebook**. 2018. Disponível em: <<http://jupyter.org/index.html>>. Acesso em: 26 nov 2018.

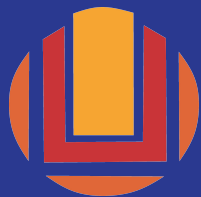
Referências

- QUANDL API FOR COMMODITY DATA. **API for Commodity Data**. 2013. Disponível em: <<https://blog.quandl.com/api-for-commodity-data>>. Acesso em: 26 nov 2018.
- SCIPY. **NumPy**. 2019. Disponível em: <<https://www.numpy.org/>>. Acesso em: 22 jun 2019.
- SKLEARN.**SVM.SVR**. 2019. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>>. Acesso em: 23 jun 2019
- STEVENS, Stanley C. **Evidence for a weather persistence effect on the corn, wheat, and soybean growing season price dynamics**. The Journal of Future Markets. Fev.,1991. DOI: <<https://doi.org/10.1002/fut.3990110108>>. Acesso: 24 out. 2018.
- USDA.a. United States Department Of Agriculture. **Oilseeds: World Markets and Trade, p. 15. 2018**. Foreign Agricultural Service/USDA - Office of Global Analysis. Disponível em: <<https://apps.fas.usda.gov/psdonline/circulars/oilseeds.pdf>>. Acesso em: 25 nov 2018.

Referências

USDA.b. United States Department of Agriculture - National Agricultural Statistics Service. **Data Visualization: 2017 – Production – Measured in Bushels**. Disponível em: <https://www.nass.usda.gov/Data_Visualization/index.php>. Acesso: 22 out. 2018.

WANG, F. **Forecasting agricultural commodity prices through supervised learning**. Month, v. 2016, p. 11–11, 2017.



Universidade Federal do Rio Grande - FURG
Centro de Ciências Computacionais - C3
Engenharia de Computação



Projeto de Graduação em Engenharia de Computação II

Previsão de cotações de Soja futura na bolsa de Chicago (CBOT) utilizando modelo LSTM e relacionando a dados climáticos das regiões mais produtivas dos EUA.

Obrigado!

Marlon Rubio de Carvalho Franco
Orientador: Prof. Dr. Marcelo R. Pias

Rio Grande, 2019.