Modelos de previsão para jogos de futebol

Felipe Viberti

Orientador: Marco Molinaro

Introdução

- Técnicas Quantitativas
- Gols Esperados (xG)
- ELO Rating
- Dados disponíveis

Objetivo da Previsão

A previsão a ser realizada é se o resultado de um jogo será vitória do mandante, empate ou vitória do visitante. É um problema de classificação com 3 classes possíveis.

Features

- Pontos no campeonato atual e anterior
- Pontos mandante em casa e visitante fora de casa
- Quantidade de títulos
- Momento Atual
- Confronto direto

Modelos Individuais x Modelos Universais

Diferenciação apenas na forma como os modelos são treinados

Individual : Dados de treinamento vêm de apenas 1 time. Aprender comportamento individual

Universal : Dados de treinamento vêm de vários times. Aprender comportamento geral

Estudo de Métodos Machine Learning

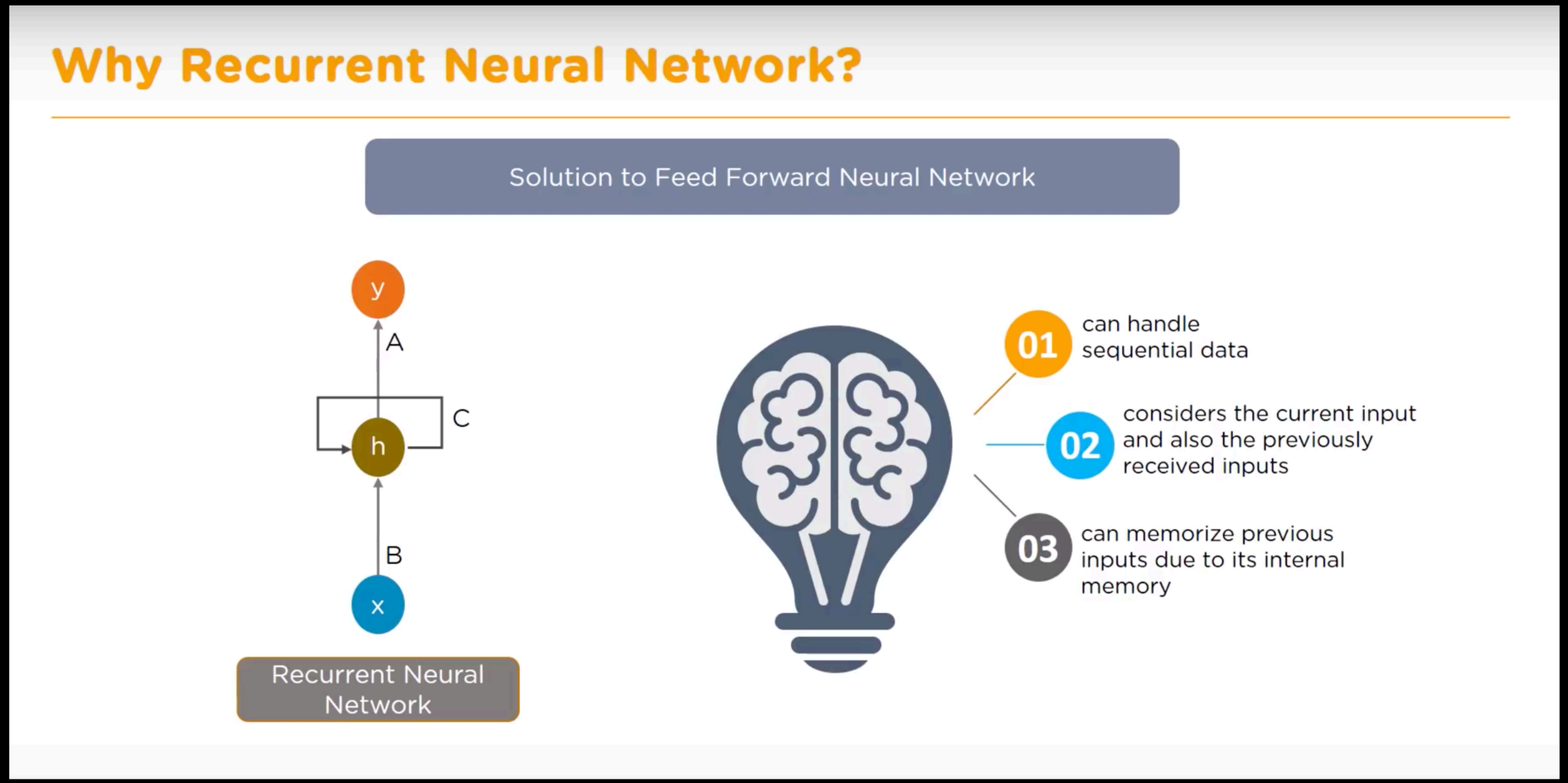
Métodos como k-Nearest Neighbors, SVM, Regressão Logística e Naive Bayes foram estudados mas aqui iremos focar no que foi usado na aplicação

Métodos Aplicados

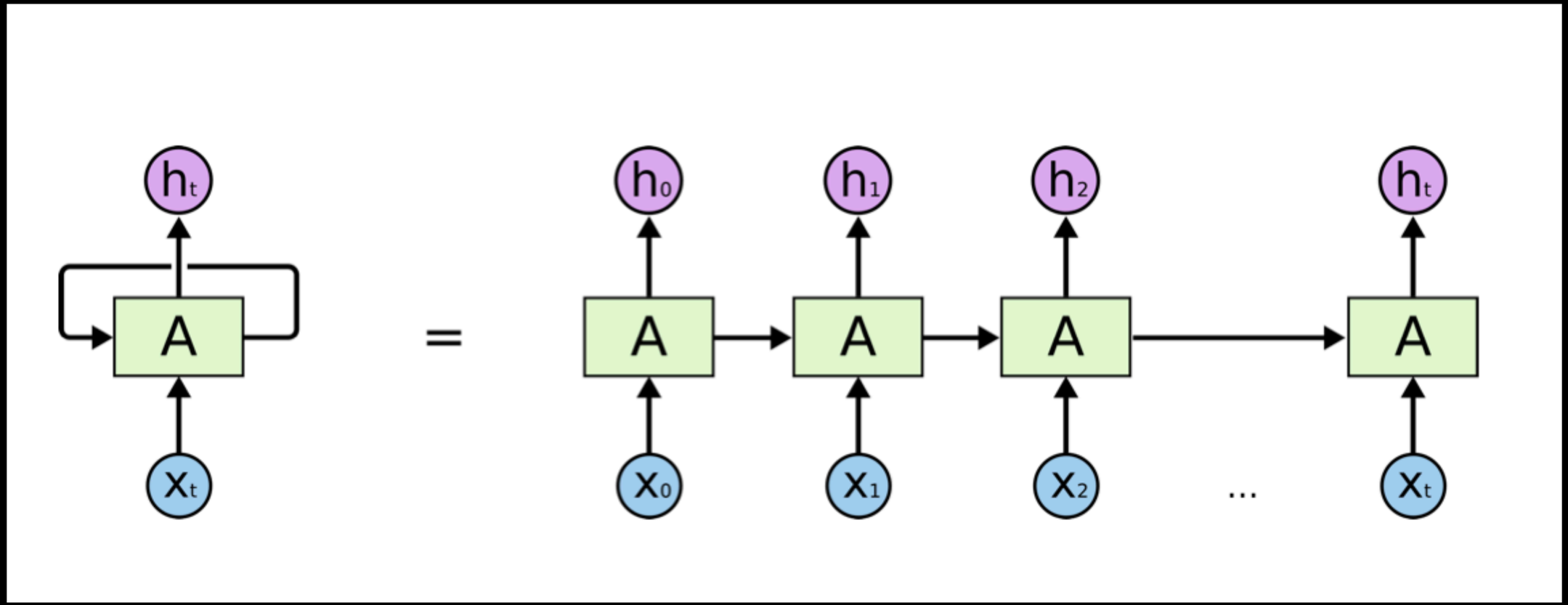
- Random Forest
- Rede Neural
- Rede Neural Recorrente (LSTM)

Rede Neural Recorrente

Similares à redes neurais, com algumas modificações que as melhor adequam para lidar com dados sequenciais, em especial, séries temporais.

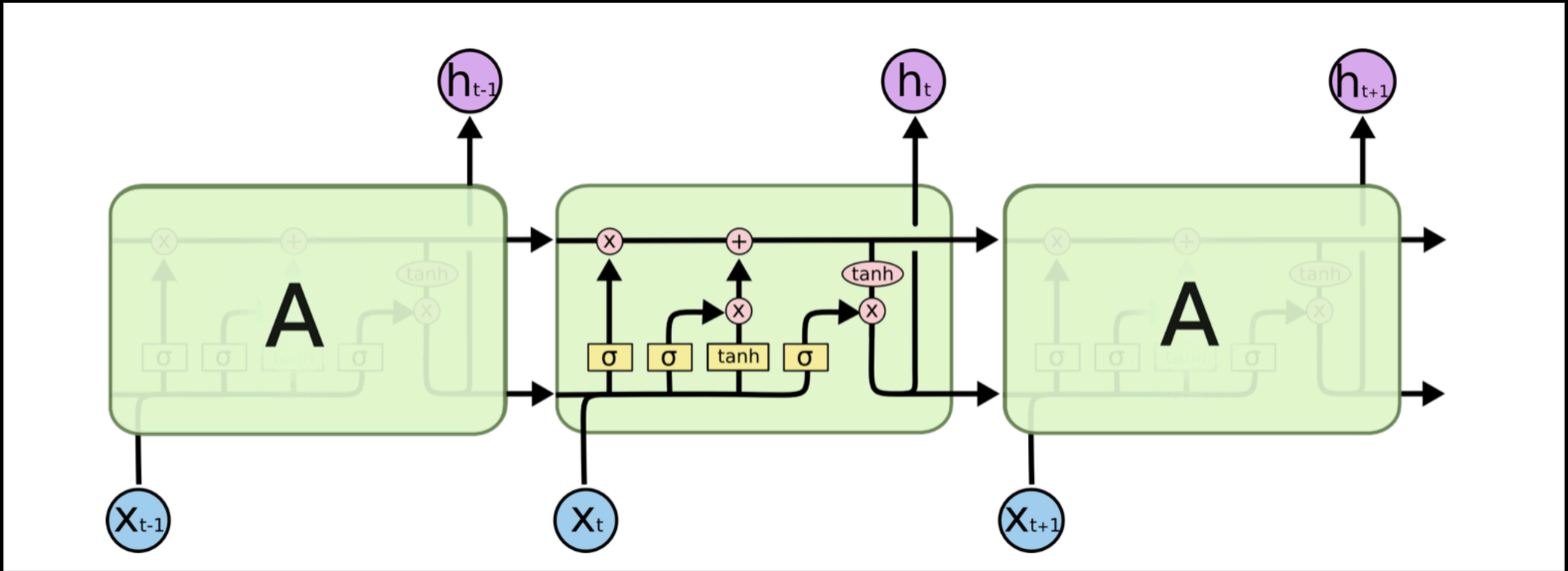


Exemplo RNN



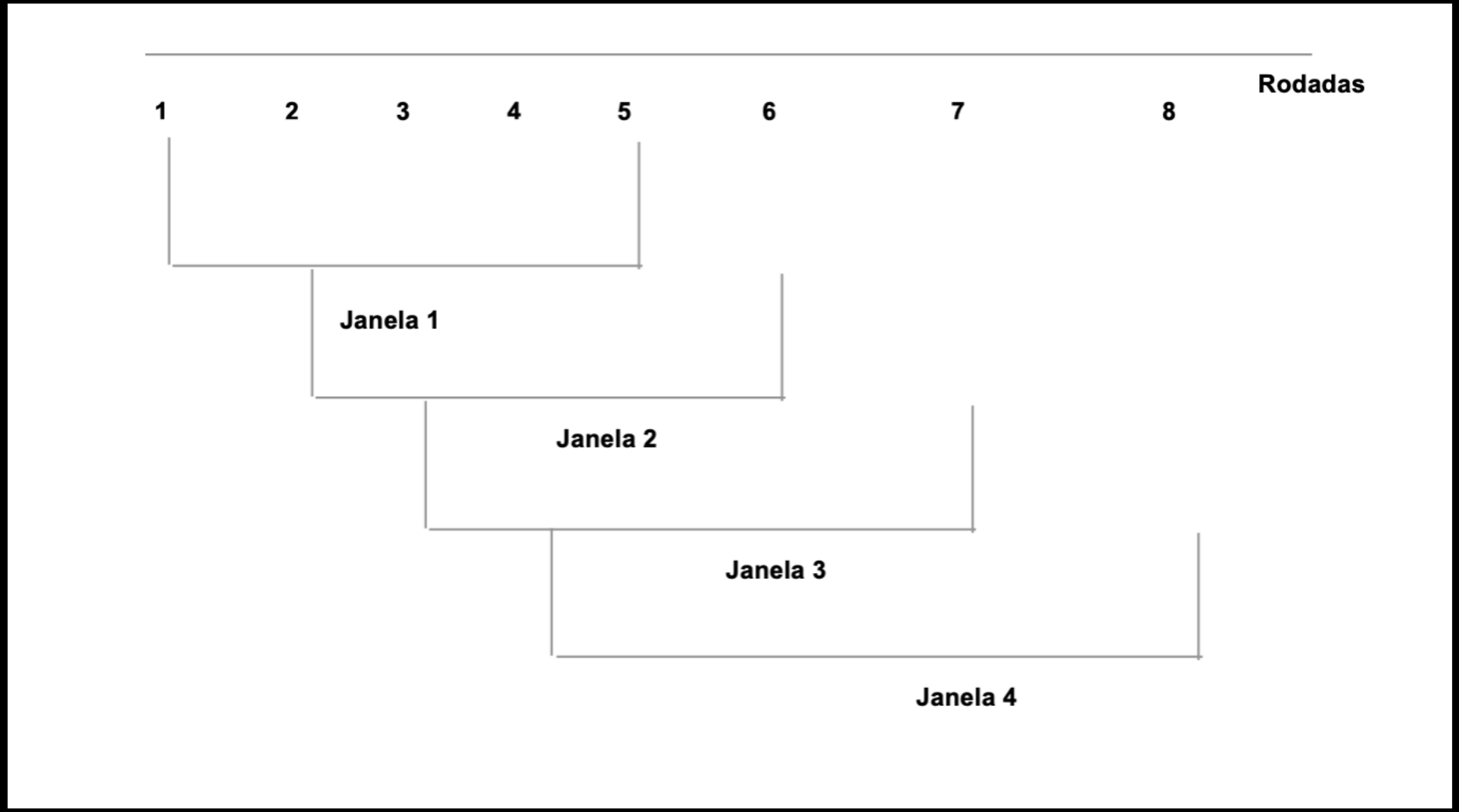
LSTM

LSTM é um tipo especial de rede neural recorrente capaz de aprender dependências por um longo período de tempo.



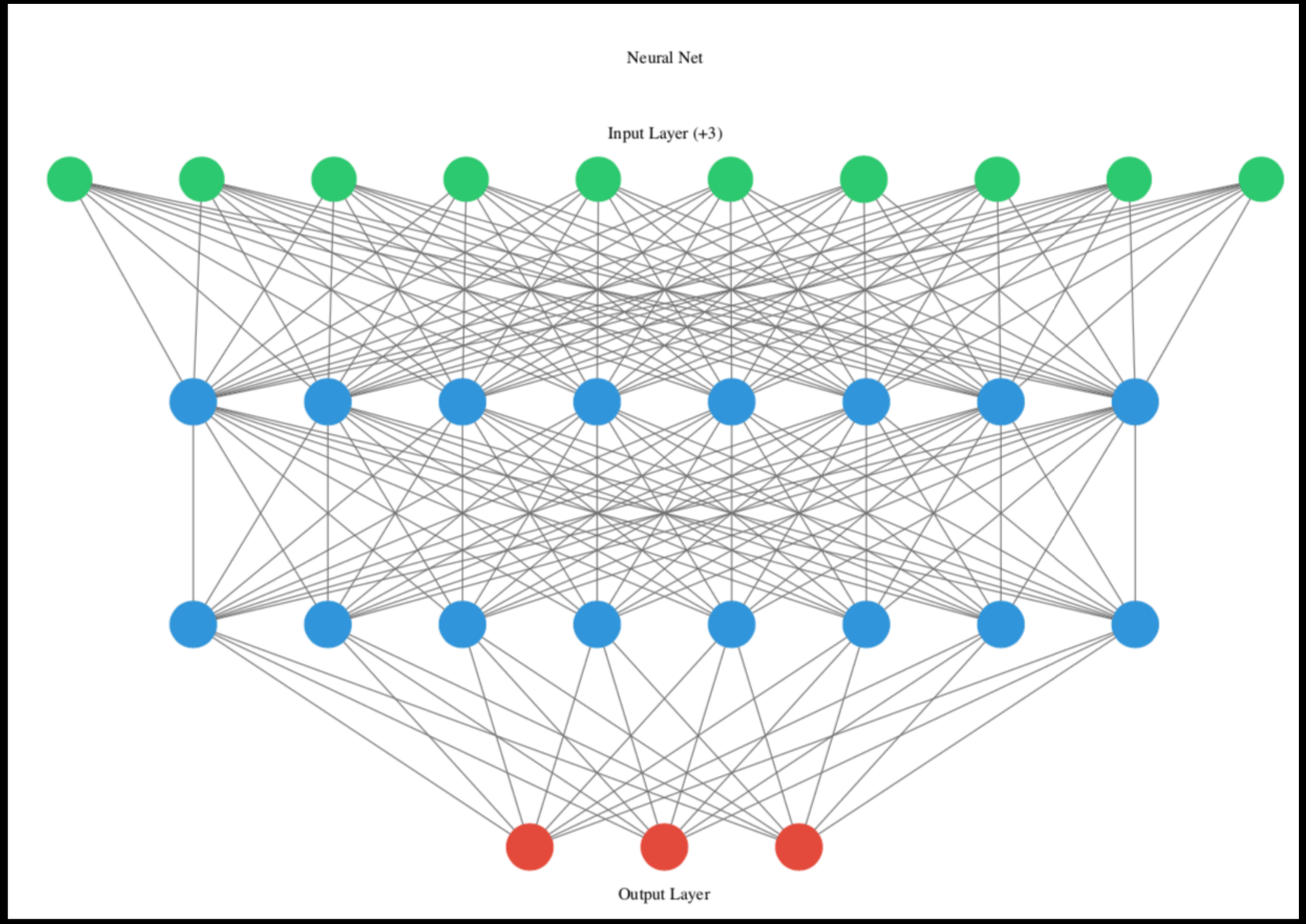
Porque LSTM para futebol?

- Histórico de Partidas é uma série temporal
- Janelas Temporais Dentro do Histórico
- Cada rodada é uma parte da sequência do LSTM
- Um só modelo copiado ao longo do tempo
- Modelo tenta aprender o “momento” de uma equipe ao longo do tempo

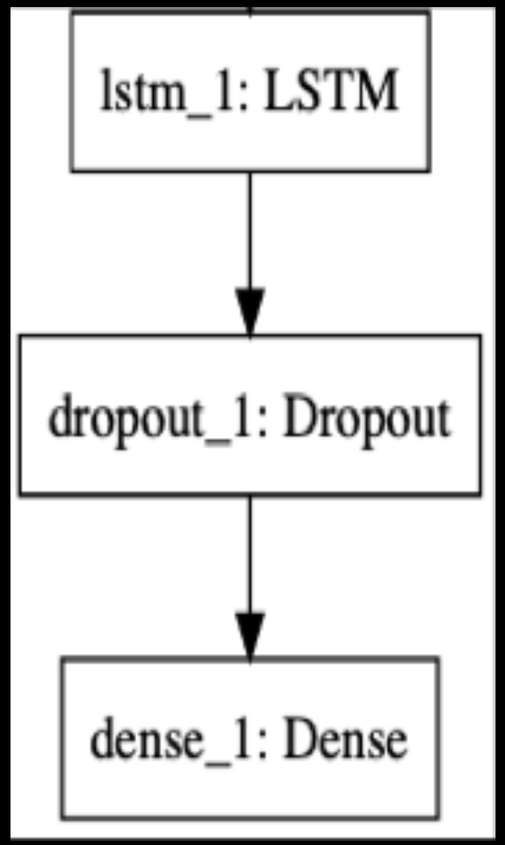

Rodadas
1
2
3
4
5
6
7
8
Janela 1
Janela 2
Janela 3
Janela 4

Modelos

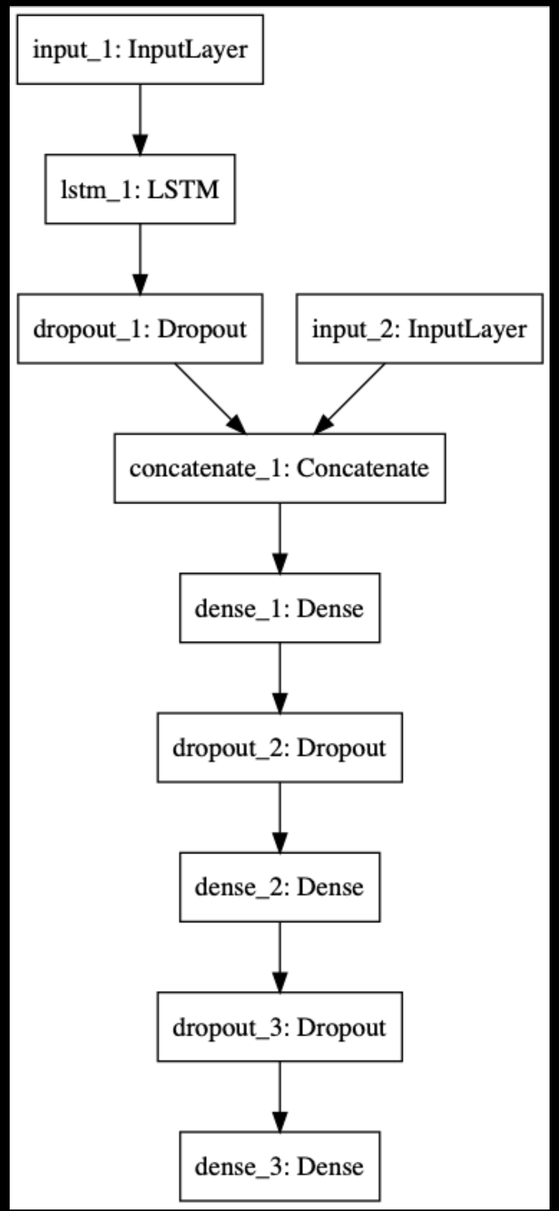
Modelo Básico



Modelo LSTM



Modelo Híbrido LSTM



Resultados

Dataset

Football Data UK: Dados dos campeonatos espanhol, inglês, alemão, italiano de 2005 à 2019

Equipes Escolhidas

Selecionamos os seguintes times do campeonato espanhol: **Barcelona, Real Madrid, Atlético de Madri, Valencia, Sevilla e Espanyol**. Estas equipes foram escolhidas por representarem uma mistura entre equipes fortes, médias e fracas e pelo fato de estarem presentes no campeonato em todos os anos de 2005 até 2019 (não foram rebaixadas nenhuma vez durante estes anos).

Modelos Individuais

	Geral	Barcelona	Real Madrid	Atlético Madrid	Valencia	Sevilla	Espanyol
Baseline	54,35%	72,89%	63,55%	59,81%	44,85%	47,66%	37,38%
Random Forest	**68,35%**	78,48%	**77,22%**	**68,35%**	**68,35%**	53,16%	64,56%
Rede Neural	61,94%	71,71%	65,65%	67,67%	41,41%	**60,60%**	**64,64%**
LSTM	51,64%	74,25%	62,37%	42,57%	43,56%	45,54%	41,58%
Híbrido LSTM	59,73%	**81,18%**	64,35%	64,35%	56,43%	50,49%	41,58%

Modelos Universais

	Geral	Barcelona	Real Madrid	Atlético Madrid	Valencia	Sevilla	Espanyol
Random Forest	73,98%	**85,85%**	**77,77%**	76,76%	66,66%	69,69%	67,67%
Rede Neural	**75,53%**	83,83%	75,75%	**80,80%**	67,67%	**73,73%**	**69,69%**
LSTM	56,60%	71,28%	58,41%	58,41%	44,55%	46,53%	37,40%
Híbrido LSTM	72,43%	84,15%	68,31%	77,22%	**68,31%**	63,36%	62,37%

Time fora dos dados de treino

	Modelo Individual	Modelo Universal
Random Forest	**63,64%**	**69,70%**
Rede Neural	62,62%	65,65%
LSTM	37,62%	41,58%
Híbrido LSTM	50,50%	60,39%

Discussão Resultados

- Modelos Universais superiores aos Modelos Individuais
- Dificuldade de não serem sempre as mesmas equipes no campeonato
- Random Forest e Rede Neural tiveram, em geral, um melhor desempenho
- Quantidade de parâmetros a serem aprendidos pelo Modelo Híbrido é muito grande. Possível motivo do desempenho ser pior é a baixa quantidade de dados para essa quantidade de parâmetros
- Híbrido LSTM superior ao LSTM