UNIVERSIDADE FEEVALE

KLAUS BENETTI KICH

APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA PREVISÃO DA SAFRA DE SOJA NO ESTADO DO RIO GRANDE DO SUL

Novo Hamburgo

2020

KLAUS BENETTI KICH

APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NA PREVISÃO DA SAFRA DE SOJA NO ESTADO DO RIO GRANDE DO SUL

Trabalho de Conclusão de Curso apresentado como requisito parcial à obtenção do grau de Bacharel em Ciências da Computação pela Universidade Feevale

Orientador: Prof. Dr. Paulo Ricardo Muniz Barros

Novo Hamburgo

2020

AGRADECIMENTOS

Gostaria de agradecer a todos que, de alguma maneira contribuíram para que este trabalho pudesse se tornar realidade.

RESUMO

A soja está entre as principais *comodities* agrícolas do mundo, e é o principal produto de exportação brasileiro, tendo somado no ano de 2018 a quantia de mais de quarenta bilhões de dólares exportados. O estado do Rio Grande do Sul é o terceiro maior produtor nacional, com 14% do total produzido no país. Sistemas de previsão da safra podem auxiliar as pessoas que estão envolvidas nesse mercado a se planejarem com as oscilações futuras na produção. Entre as técnicas que podem ser usadas nesta previsão, destaca-se o uso de RNAs, que vem apresentado grandes avanços ao longo dos anos, e tendo bons resultados em previsões baseadas em séries históricas, em especial com o emprego de redes do tipo LSTM. No trabalho foram desenvolvidos quatro modelos com o emprego de redes LSTM, onde foi possível demonstrar que há resultados satisfatórios na previsão dos valores da safra no próximo ano, com exceção à anos em que ocorrem oscilações climáticas fora do padrão histórico, apresentando valores similares aos modelos atualmente utilizados pela CONAB em suas previsões.

Palavras-chave: Redes Neurais Artificiais; Previsão; Soja; Agricultura de Precisão.

ABSTRACT

Soybean is one of the main agricultural commodities in the world, and is the main Brazilian export product, having in 2018 the amount of more than forty billion dollars exported. The state of Rio Grande do Sul is the third largest national producer, with 14% of the total produced in the country. Crop forecasting systems can help people involved in this market to plan future fluctuations in production. Among the techniques that can be used in this forecast, the use of RNAs stands out, which has shown great advances over the years and having good results in forecasts based on historical series in particular with the use of LSTM-type networks. In the work, four models were developed based on LSTM-type network, where it was possible to demonstrate that the technique presents satisfactory results in the prediction of next year's crop values, except for years in which climatic fluctuations occur outside the historical standard, presenting similar values to the models currently used by CONAB in its forecasts.

Keywords: Artificial neural networks, Forecast, Soybean, Precision agriculture

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

APE	*Absolute percentual error*
ARIMA	Auto Regressivo Integrado Média móvel
CONAB	Companhia Nacional de Abastecimento
ENN	*Ensemble neural network*
GRMSE	*Geometric root mean squared error*
INMET	Instituto Nacional de Meteorologia
LSTM	*Long Short Term Memory*
MAE	*Mean absolute error*
MCP	*McCulloch-Pitts*
ME	*Mean error*
MPE	*Mean percentual error*
MPL	M*ultilayer perceptron*
MSE	*Mean squared error*
NAR	*Nonlinear autoregressive neural network*
RMSE	*Root mean squared error*
RNA	Redes Neurais Artificiais
RNN	*Recurrent Neural Network*
API	*Application Programming Interface*
RELU	*Rectified Linear Unit*
ELU	*Exponential linear unit*

SUMÁRIO

1 INTRODUÇÃO

A Soja está entre as principais *comodities* agrícolas no mundo, sendo negociada em várias bolsas de valores ao redor do planeta, como, por exemplo, a Bolsa de Chicago e a Bovespa em São Paulo. Também é o principal produto brasileiro de exportação, tendo somado no ano de 2018 a quantia de mais de quarenta bilhões de dólares exportados, somadas as exportações do grão in natura e derivados (MINISTÉRIO DA ECONOMIA, 2019). Além da sua importância no comércio internacional, a soja que fica no país é processada de várias maneiras, sendo usada para o consumo humano, elaboração de biodiesel e na alimentação animal (APROSOJA BRASIL, 2019), demostrando sua valia também para o mercado nacional.

A produção mundial de soja é liderada pelos Estados Unidos da América e pelo Brasil, onde é possível observar que na safra de 2017/18 foram produzidos um total de 122 milhões de toneladas do grão (UNITED STATES DEPARTMENT OF AGRICULTURE, 2019) . Dentro do contexto nacional, o estado do Rio Grande do Sul tem grande influência na produção do grão, sendo o terceiro maior estado produtor, com cerca de 14% do total produzido neste mesmo período (CONAB, 2019a).

Dentro do estado, a soja também tem uma grande importância para a economia. Em 2017, foi o produto mais exportado, somando mais de quatro bilhões de dólares, o equivalente a 26% do total de exportações do estado (SEPLAG RS, 2019a). Tendo em vista o valor da produção do grão para a economia do estado, ações governamentais para o controle da balança comercial, garantia do abastecimento interno e criação de políticas de financiamento da produção precisam de informações constantes sobre a previsão de produção da safra. Tais informações também são importantes para auxiliar agricultores a planejar a armazenagem, estoque, transporte e a comercialização do grão. (FIGUEIREDO, 2005).

Além de todos os benefícios que a previsão pode trazer para as pessoas diretamente envolvidas na safra, em 2018, cerca de 811 milhões de pessoas no mundo (FAO *et al*., 2019) viviam sem a quantidade adequada de alimentos. Sistemas de previsão podem ajudar na constituição de ações que visam diminuir este número (YOU *et al*., 2017), além de ajudar no planejamento de outros setores que dependem indiretamente do resultado da safra.

Existem atualmente diversas maneiras de realizar previsões de valores, entre elas, uma das alternativas que podem ser utilizadas é a previsão baseada em séries temporais. Neste tipo de modelo de previsão são utilizados dados históricos colhidos durante um determinado período, eles são empregados na tentativa de identificar padrões de comportamento nos valores ao longo do tempo, e assim, poder prever o seu valor no futuro. Entre as técnicas utilizadas nestas previsões podemos citar os modelos ARIMA (Auto - Regressivo – Integrado – Média móvel), os modelos Lineares Dinâmicos e as Redes Neurais Artificiais (RNA) (BRESSAN, 2004).

Segundo GURNEY (2014), Redes Neurais Artificiais são um conjunto de elementos, unidades ou nós de processamento interconectados, cuja a funcionalidade é baseada vagamente em um neurônio animal como os de humanos. Esse neurônios artificiais recebem como entrada diversos valores que podem ou não fazer este neurônio acionar outro neurônio e assim por diante, este neurônios também podem ser treinados para se adaptar conforme a entrada que for dada para ele, fazendo ele assim “aprender” sobre as informações que transitam por ele. Após este processamento espera-se ter como resultado uma rede que seja capaz de classificar corretamente o padrão das informações que foram apresentadas para ela, mas que não estavam em um padrão visível de fácil verificação.

As RNAs tiveram grandes avanços e vem sendo aplicadas com bons resultados em diversas áreas, como no trabalho realizado por MUNTASER; SILVA; PENEDO (2017), onde foi possível prever com uma margem de erro baixa o preço das ações de empresas brasileiras que atuam no setor de óleo e gás na bolsa de valores. No trabalho de ABEYRATHNA; GRANMO; GOODWIN (2018) os autores desenvolveram um modelo de previsão para futuros surtos de dengue nas Filipinas, onde foi possível observar bons resultados em seu estudo.

O uso de RNAs empregado na estimativa de produção de cultivos agrícolas, também vem sendo usado com bons resultados como no estudo apresentado por KUNG *et al*. (2016), onde se procura estimar a produção de alimentos na ilha de Taiwan. Para isto foram utilizados dados referentes a fatores meteorológicos, ambientais e econômicos para a confecção da rede. Como resultado do experimento, ele conseguiu estimar a produção de tomates tendo em boa parte dos casos um erro menor do que 2% do valor real da produção. Em outro estudo apresentado por HAIDER *et al*., (2019) procurou-se prever a safra de trigo no Paquistão, para isso os

autores utilizaram dados históricos da produção do grão no país, e compararam o emprego de redes neurais do tipo LSTM, com redes RNN, ARIMA, e a aplicação de funções de suavização de dados. O resultado mostrou que a rede do tipo LSTM com dados suavizados apresentou um melhor desempenho em comparação as outras.

No que diz respeito à soja, a geração de modelos de previsão pode ser considerada complexa devido aos diversos fatores que podem influenciar na sua produtividade ao final da safra. Entre eles podemos citar a área total plantada, diversidade do solo, ataque de pragas e doenças, e principalmente, variações de condições climáticas onde o aumento ou diminuição das temperaturas e da quantidade de precipitações, principalmente no estado do Rio Grande do Sul, pode ter um impacto sobre a produtividade da lavoura (CASTRO, 2015). No estudo apresentado por COX; JOLLIFF, (1986) os autores indicam que, quando submetida a um déficit hídrico, a produção de soja pode sofrer perdas de até 87% da produção do grão em casos mais severos.

Tendo em vista estas informações, no trabalho é proposto a elaboração de quatro modelos de redes neurais artificiais do tipo LSTM, onde com a utilização de dados sobre a safra de soja fornecidos pela CONAB, e com dados meteorológicos fornecidos pelo INMET, buscam prever a quantidade de soja colhida no estado do Rio Grande do Sul durante a próxima safra.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Desenvolvimento de um modelo de RNA, capaz de prever a quantidade total de soja produzida no estado do Rio Grande do Sul ao final de uma safra, fazendo-se uso de dados provenientes de órgão públicos estaduais e nacionais.

1.1.2 Objetivos específicos

- Buscar levantamento bibliográfico sobre modelos de previsão;
- Realizar pesquisa bibliográfica sobre RNAs;

- Realizar levantamento de dados referentes ao cultivo da soja no estado do Rio Grande do Sul;
- Desenvolver um modelo de previsão com o uso de RNA;
- Analisar o funcionamento do modelo desenvolvido com os dados informados;
- Avaliar o resultado gerado pelo modelo através de um modelo estatístico para comparação

1.2 ESTRUTURA DO TRABALHO

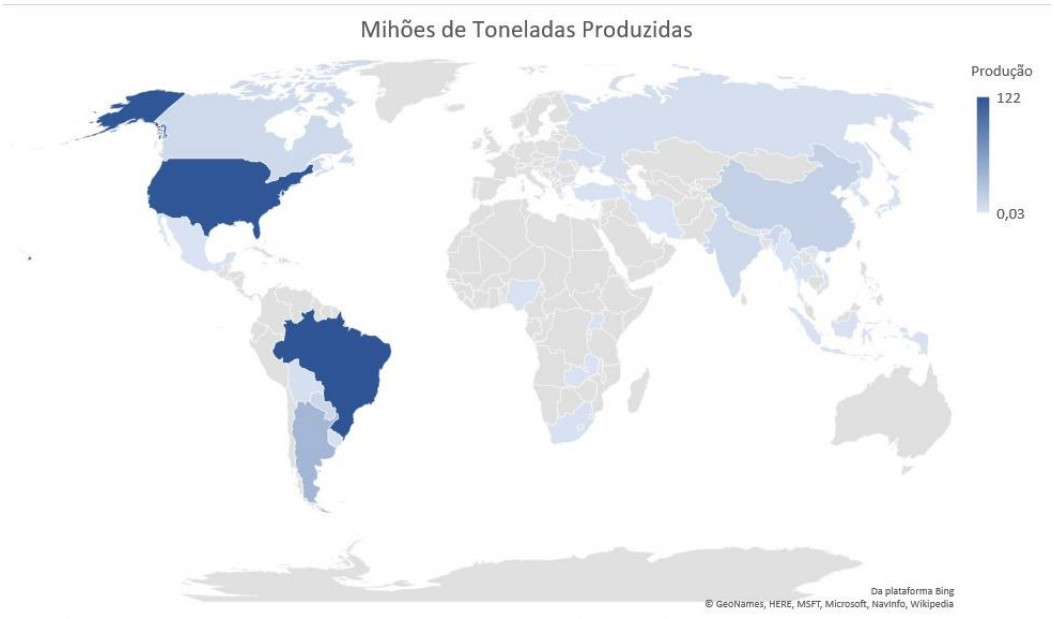
O trabalho encontra-se estruturado da seguinte maneira, primeiramente, no capítulo 2, são apresentadas informações e características relacionadas a produção de soja, no 3 capítulo é feita uma revisão sobre métodos usados para previsão de valores. Após no capítulo 4 é abordado sobre redes neurais artificiais e sua aplicação na previsão de valores com base em séries históricas, então no capítulo 5 são apresentados os trabalhos correlatos. Já no capítulo 6 são expostas informações sobre os conjuntos de dados utilizados no trabalho, e descrita a estrutura utilizada nos modelos de previsão gerados. No capítulo 7 são discutidos os resultados apresentados nos experimentos dos modelos propostos, e por fim, no capítulo 8 são apresentadas as conclusões do trabalho.

2 PRODUÇÂO DE SOJA

As primeiras menções ao uso da soja, também conhecida pelo nome científico *Glycine max,* como alimento, datam para mais de cinco mil anos atrás, a planta é nativa da costa leste da Ásia e se desenvolvia principalmente ao longo do rio *Yangtse* na China, lá ela era considerada um grão sagrado e um dos seus primeiros registros históricos está no livro "*Pen Ts'ao Kong Mu*" (EMBRAPA SOJA, 2019).

Até o final do século XIX a produção e utilização da soja como alimento ficou restrita ao leste Asiático, porém no começo do século XX, o teor de óleo e a proteína do grão chamou a atenção de diversas indústrias mundiais, iniciando assim o seu plantio comercial nos Estados Unidos da América. Neste mesmo período outros países como a Rússia, Inglaterra e Alemanha também tentaram realizar o plantio comercial, porém fracassaram, provavelmente divido as plantas não estarem adaptadas as condições climáticas destes países (APROSOJA/MT, 2019).

Gráfico 1 - Mapa com os maiores países produtores de soja no mundo. Produção em milhões de toneladas.

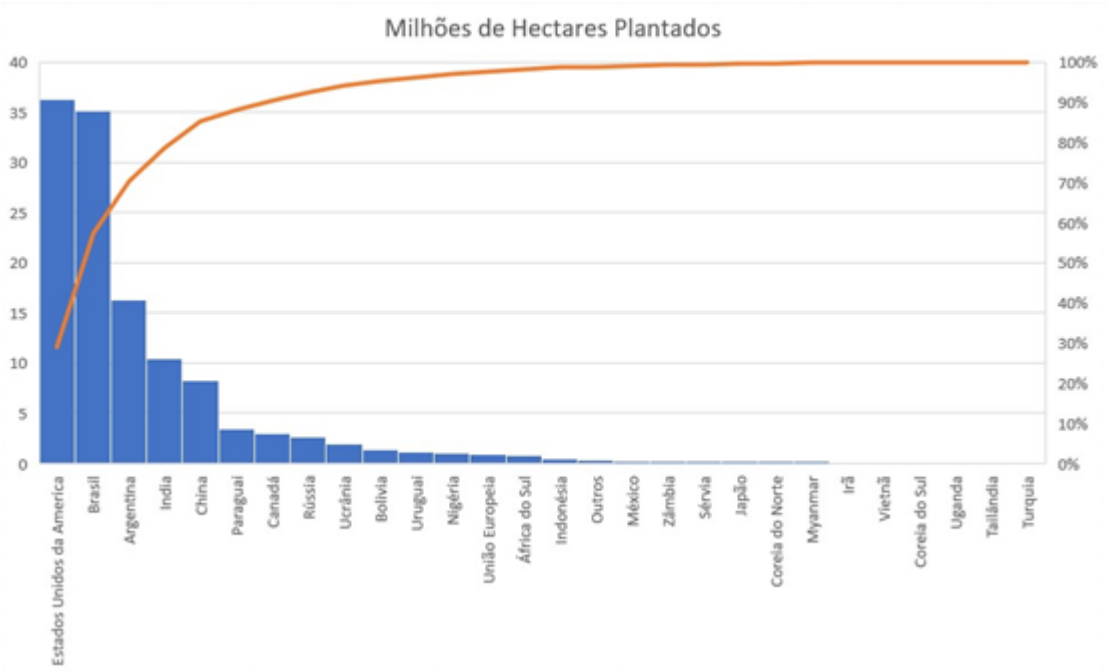


Fonte: Adaptado de UNITED STATES DEPARTMENT OF AGRICULTURE (2019)

Atualmente a soja se tornou uma das principais comodities agrícolas do mundo sendo produzida e negociada em diversos países ao redor do mundo. Durante a safra de 2017/18 foram cultivados mais de 124 milhões de hectares do grão, o que resultou

em uma produção de mais de 341 milhões de toneladas. Dentre os principais produtores mundiais, podemos dar um destaque especial para o Brasil e os Estados Unidos da América. Conforme exibido no gráfico 1 é possível observar a coloração azul escuro atribuída a esses países indicando uma ampla margem sobre a produção de outros países, tendo eles produzindo 122 e 120,07 milhões de toneladas respectivamente durante a safra de 2017/18 (UNITED STATES DEPARTMENT OF AGRICULTURE, 2019).

Gráfico 2 - Hectares de soja plantados por país.



Fonte: Adaptado de UNITED STATES DEPARTMENT OF AGRICULTURE (2019)

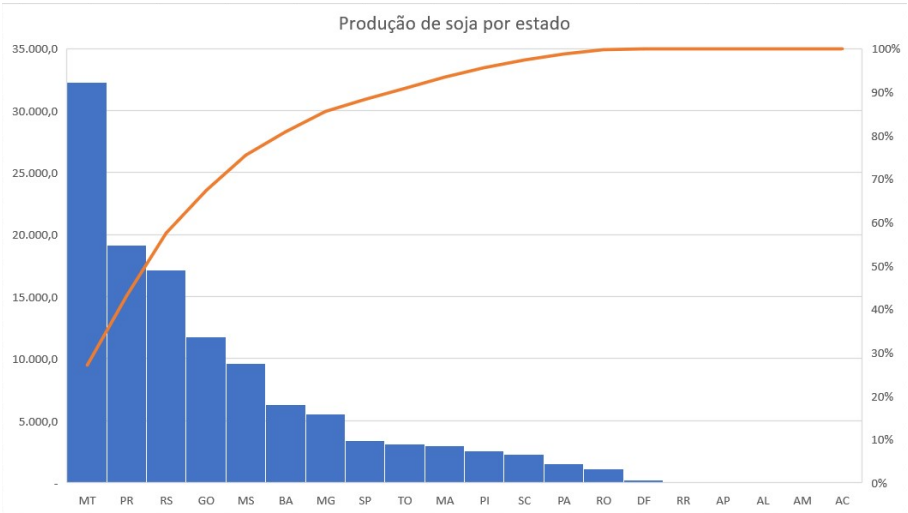
O gráfico 2 acima, traz a quantidade de hectares de soja plantados na safra 2017/18 por país, onde também podemos observar que os Estados Unidos da América e o Brasil, lideram com ampla margem sobre os países subsequentes. Uma análise interessante que pode ser realizada comparando-se o gráfico 2 com o gráfico 1, é que mesmo apresentando uma área plantada ligeiramente maior que o Brasil, os Estados Unidos da América apresentaram uma produção final do grão menor que o Brasil nesta safra, evidenciando uma maior produtividade das lavouras brasileiras.

2.1 PRODUÇÃO DE SOJA NO BRASIL

A introdução da soja no Brasil se deu no início do século XX, na Estação Agropecuária de Campinas onde ocorreram os primeiros cultivos e distribuição de sementes para produtores. Entretanto a expansão e consolidação da soja no Brasil veio a ocorrer somente no começo da década de 70, devido a diversos fatores locais e internacionais, entre os quais podemos citar: o aumento da demanda no pais pelo farelo de soja para uso em alimentação animal, a expansão internacional da indústria do óleo, e pela vantagem competitiva que o pais apresenta devido ao fato do escoamento da safra ocorrer no período de entressafra da produção americana, o que elevava o preço do produto no mercado internacional (APROSOJA/MT, 2019).

A expansão da soja no Brasil iniciou-se pelos estados ao sul do pais, devido as características climáticas serem mais propensas para o plantio, porém no decorrer dos anos o investimento em tecnologia e melhoramento genético da planta realizados no país possibilitou que a planta fosse cultivada em regiões tropicais, o que possibilitou a expansão do cultivo para os estados do centro-oeste, norte e nordeste do país, além de um aumento na quantidade de grãos produzidos por hectare (EMBRAPA SOJA, 2019).

Gráfico 3 - Produção nacional de soja por estado em mil toneladas.



Fonte: adaptado de CONAB (2019a)

O gráfico 3 nos apresenta a quantidade de soja produzida por cada estado brasileiro na safra 2017/18. Nele podemos verificar que os maiores produtores

nacionais são os estados do Mato Grosso, Paraná e Rio Grande do Sul, tendo eles produzido, 32.306, 19.170 e 17.150 mil toneladas do grão respectivamente (CONAB, 2019a).

2.1.1 Destino da soja produzida no Brasil

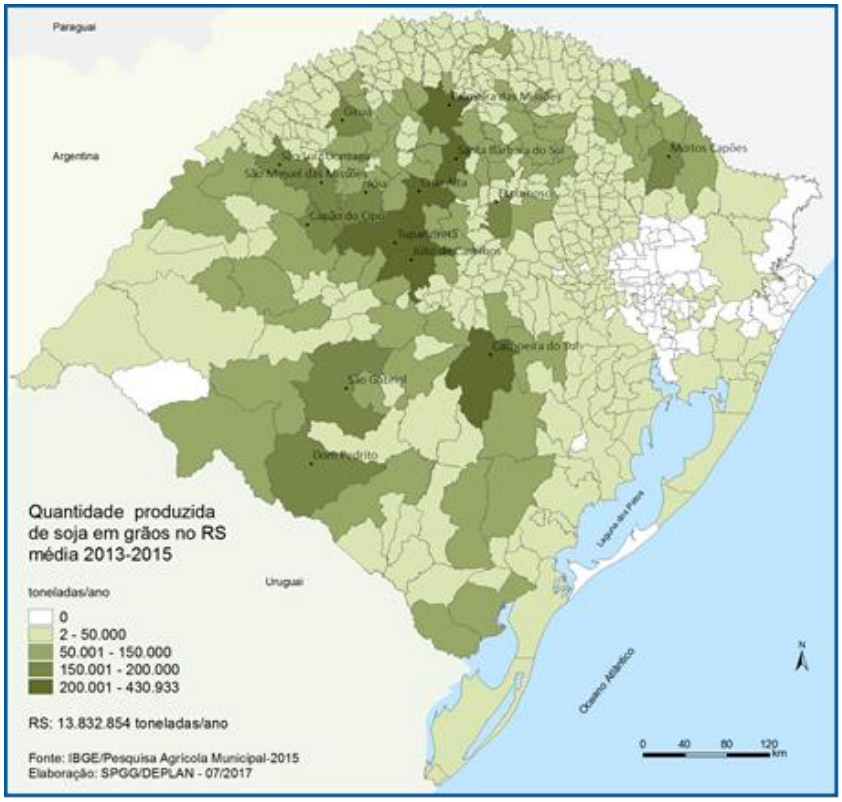
O principal destino da soja produzida no país é a exportação com cerca de 44% dos grãos produzidos exportados in natura, outros 49% são processados no país, gerando deste processamento 21% de óleo e 79% de farelo. Do óleo produzido 23% é destinado à exportação e os 77% restantes são utilizados no consumo doméstico para a alimentação e na elaboração de biodiesel. O farelo tem 52% do seu total produzido exportado e o restante é utilizado internamente para a alimentação animal (APROSOJA BRASIL, 2019).

Estes números tornam a soja o principal produto de exportação brasileiro tendo somando no ano de 2018, entre o grão in natura e processados o total de 40.704.436.117,00 (quarenta bilhões e setecentos e quatro milhões e quatrocentos e trinta e seis mil e cento e dezessete) dólares exportados, tornando ela essencial para a balança comercial do país (MINISTÉRIO DA ECONOMIA, 2019).

2.2 PRODUÇÃO DE SOJA NO RIO GRANDE DO SUL

O cultivo de soja no estado apresentou forte expansão no final da década de 60 início da década de 70, expandindo-se das regiões do Alto Uruguai, Planalto Médio e Missões onde era originalmente cultivada, para quase todas as regiões do estado (AUGUSTO C. CONCEIÇÃO, 1986). O gráfico 4 mostra a quantidade média de soja produzida pelos municípios do estado entre os anos de 2013 a 2015. É possível destacar que os principais municípios produtores estão localizados na região norte-noroeste do estado com destaque para os municípios de Tupanciretã, Cachoeira do Sul, Palmeira das Missões, Júlio de Castilhos, Cruz Alta e Santa Bárbara do Sul que no período produziram em média mais de 200 mil toneladas por safra (SEPLAG RS, 2019b).

Gráfico 4 - Quantidade média de soja produzida por município no estado no período de 2013 a 2015



Fonte: (SEPLAG RS, 2019b)

Durante os últimos anos as áreas plantadas também tiveram um crescimento substancial indo de 3.030.556 (três milhões trinta mil quinhentos e cinquenta e seis) hectares no ano de 2000 para 5.263.899 (cinco milhões duzentos e sessenta e três mil oitocentos e noventa e nove) hectares no ano de 2015, um crescimento de mais de 73% na área plantada. Neste mesmo período, a quantidade de grão produzido aumentou mais de 228%, indo de uma safra de 4.783.895 (quatro milhões setecentos e oitenta e três mil oitocentos e noventa e cinco) de toneladas no ano 2000 para 15.700.264 (quinze milhões setecentos mil duzentos e sessenta e quatro) de toneladas no ano de 2015, o que demonstra um forte investimento por parte dos produtores em tecnologias para o aumento da produtividade do grão (SEPLAG RS, 2019b).

Tabela 1 – Exportações do estado do Rio Grande do Sul no ano de 2017

PRODUTOS	US$ 1000 FOB	%
Soja, mesmo triturada, exceto para semeadura	4,634,051	26.05
Tabaco n/manufaturado, total ou parcialmente destalado, em folhas secas em secador de ar quente do tipo Virgínia	1,329,293	7.48
Automóveis com motor explosão, de cilindrada superior a 1.000 cm3, mas não superior a 1.500 cm3	642,569	1.47
Bagaços e outros resíduos sólidos, da extração do óleo de soja	605,799	1.18
Carnes de galos/galinhas, não cortadas em pedaços, congelala	547,849	1.17
Pedaços e miudezas, comestíveis de galos/galinhas, congelados	543,469	1.17
Outras carnes de suíno, congeladas	433,231	1.14
Pastas quím. de mad, à soda ou ao sulfato, exceto pastas para dissolução, semibranqueadas ou branqueadas, de n/coníferas	427,134	0.98
Outros polietilenos sem carga, densidade >= 0.94, em formas primárias	260,685	0.96
Polietileno linear, densidade < 0.94, em forma primária	209,820	0.94
Óleo de soja, em bruto, mesmo degomado	207,963	0.91
Polietileno sem carga, densidade < 0.94, em forma primária	207,533	0.90
Outras partes e acessórios para tratores e veículos automóveis	202,415	0.85
Tabaco não manufaturado, total ou parcialmente destalado, em folhas secas (light air cured), do tipo Burley	174,164	0.81
Outros calçados sola exterior borracha/plástico, de couro/natural	171,114	0.75
Outros couros e peles inteiros, de bovinos (incluindo os búfalos), divididos, com o lado flor	167,185	0.66
Carrocerias para veículos automóveis com capacidade de transporte => 10 pessoas, ou para carga	161,194	0.61
Copolímeros de propileno, em formas primárias	159,248	0.59
Polipropileno sem carga, em forma primária	151,039	0.58
Outras espingardas e carabinas de caça ou de tiro ao alvo	143,988	0.58
Buta-1, 3-dieno não saturado	132,711	0.57
Outros calçados cobrindo o tornozelo, parte superior de borracha, plástico	118,193	0.66
Preparações alimentícias e conservas, da espécie bovina	108,641	0.61
Consumo de bordo - combustíveis e lubrificantes para embarcações	104,275	0.59
Partes de outras máquinas e aparelhos para colheita, debulha, etc.	102,920	0.58
Outros trigos e misturas de trigo com centeio, exceto para semeadura	102,752	0.58
Benzeno	102,229	0.57
Outros produtos	5,630,795	31.67
Total	17,782,259	100.00

Fonte: (SEPLAG RS, 2019a)

No contexto econômico a soja possui uma grande importância para a balança comercial do estado, na tabela 1 estão listados os produtos mais exportados pelo estado durante o ano de 2017, nela é possível observar o papel de destaque que a soja possui, sendo ela responsável por 26% de todas as exportações realizadas.

Aprofundando-se a análise nos dados, e adicionado os valores referentes ao beneficiamento do óleo de soja, ela se torna responsável por 28% do total exportado no estado, o que equivale a um montante de mais de 5 bilhões de dólares, valor consideravelmente superior aos cerca de 1 bilhão e 300 mil dólares exportados de tabaco o segundo colocado na lista (SEPLAG RS, 2019a).

2.3 CARACTERÍSTICAS DA PRODUÇÃO DE SOJA

A soja caracterizasse por ser um plantio de verão, então no estado do Rio Grande do Sul o seu período de semeadura varia entre os meses de outubro a dezembro, e a sua colheita costuma acontecer entre os meses de fevereiro a abril dependendo da época de plantio e da cultivar escolhida. A soja produzida no estado

pode ter a sua produtividade afetada por diversos fatores ambientais entre os quais podemos citar a área total plantada, diversidade do solo, ataque de pragas e doenças e principalmente fatores relacionados ao clima, como a deficiência hídrica, a insuficiência térmica e a falta de uma estação seca na época da colheita (ROCCA DA CUNHA *et al*., 2001).

A precipitação fluvial é indicada como sendo dentre todos os fatores o de maior relevância para a oscilação da produtividade entre os anos. Conforme o estágio de desenvolvimento da planta, ela tem necessidades de água distintas, sendo que tanto o excesso quando a sua falta pode influenciar negativamente na colheita final. Em seu estudo COX; JOLLIFF (1986), observaram o desenvolvimento da soja em 3 situações distintas de irrigação do solo, sendo elas: solo irrigado, com déficit hídrico e seco. O solo com déficit hídrico apresentou uma redução de 27% na produção de sementes quando comparado ao solo irrigado, já o solo seco chegou a uma perda de produtividade no montante de 87% quando comparado ao solo irrigado.

Os períodos mais sensíveis ao estresse hídrico são os da germinação à emergência e no da floração-enchimento de grãos. O estágio de geminação e emergência corresponde ao começo da vida da planta e vai do seu plantio até a planta encontrar-se com os cotilédones[1] sobre a superfície do solo e eles formarem um ângulo igual ou maior de 90º ao hipocótilo[2] da planta (NEUMAIER *et al*., 2000), neste período a semente da soja precisa absorver no mínimo 50% do seu peso em água. O estágio de floração-enchimento corresponde por quase todo estágio reprodutivo da planta, ele inicia-se com o aparecimento da primeira flor aberta no pé, conhecido como estágio R1 e entende-se até o estágio R6, onde a planta encontra-se contendo grãos verdes preenchendo totalmente a cavidade da vagem em um dos quatro nós superiores da haste principal da planta e conta ainda com uma folha completamente desenvolvida (NEUMAIER *et al*., 2000). Neste período a planta atinge o seu maior consumo de água e déficits hídricos neste período provocam alterações fisiológicas na planta o que acaba resultado em uma redução no rendimentos dos grãos (EMBRAPA SOJA, 2013).

Outro fator climático relevante para o desenvolvimento da planta é a temperatura do ar. A soja se adapta melhor a temperaturas entre 20ºC a 30ºC, sendo a temperatura ideal para o seu crescimento próxima aos 30ºC. Temperaturas

[1] Cotilédones são as primeiras folhas que surgem dos embriões germinados.
[2] Hipocótilo é a parte do caule do embrião germinado.

extremas impactam em diversos ciclos da planta, onde temperaturas abaixo de 10ºC impedem o crescimento vegetativo da planta, abaixo de 13ºC impendem a floração da planta, e temperaturas acima de 40ºC provocam distúrbios na floração, o que diminui a retenção de vagens e consequentemente a produtividade da planta (EMBRAPA SOJA, 2013).

2.4 PREVISÃO DE RENDIMENTO DAS SAFRAS

A realização oficial de estimativas e previsões das safras de soja no Brasil desde a safra 1976/77 são de responsabilidade da CONAB (Companhia Nacional de Abastecimento). Até o começo dos anos 2000 este órgão utilizava como metodologia para a geração dos modelos a consulta direta ao setor produtivo, um método subjetivo e que poderia causar diversos erros na estimativa gerada, devido à falta de confiabilidade das informações repassadas.

A partir do ano de 2004, a CONAB passou por uma modernização nos seus métodos de previsão e desde então foram incluídos dados de sensoriamento remoto, posicionamento por satélite (GPS), sistemas de informações geográficas e modelos agro meteorológicos, trazendo mais confiabilidade para o sistema e precisão para os resultados (FIGUEIREDO, 2005).

A realização do trabalho de estimativas é de suma importância para os governos federal, estaduais e municipais, proporcionado a eles aprimorar ações de políticas públicas do agronegócio, trazendo previsibilidade e os possibilitando estabelecer uma melhor logística em diversas situações que envolvem principalmente as pontas extremas da cadeia produtiva, onde estão localizados os produtores e os consumidores.

Na iniciativa privada diversos setores podem se aproveitar destas informações como o setor produtivo, de armazenagem e estoque, transporte, industrialização, comercialização e comércio exterior, além de setores ligados indiretamente ao agronegócio mas que dependem dele para manter os seus negócios aquecidos como por exemplo o comercio de cidades pequenas, proporcionado assim a eles dados importantes para o planejamento empresarial, abrindo a possibilidade de aumentarem os seus lucros e a sua competividade (FIGUEIREDO, 2005).

2.5 DADOS HISTÓRICOS DAS SAFRAS

A companhia nacional de abastecimento disponibiliza dados históricos da safra de soja desde a safra de 1976/77 (CONAB, 2019a), nesta série histórica estão presentes dados de: área plantada, produtividade por hectare, e produção ao final da safra. Eles estão organizados por estado e região do país.

Além dos dados referentes a safra, a CONAB também disponibiliza ao final da safra dados referentes ao custo de produção por estado (CONAB, 2019b). Para o estado do Rio Grande do Sul os dados estão disponíveis a partir do ano de 1999 e contam com as seguintes informações: despesas de custeio da lavoura, despesas pós-colheita, despesas financeiras, depreciações, outros custos fixos e renda de fatores.

O Instituto Nacional de Meteorologia (INMET), também disponibiliza em seu site séries históricas sobre as condições meteorologias registradas em suas estações (INMET, 2019). No Rio Grande do Sul eles possuem estações instaladas nas seguintes cidades: Bagé, Bento Gonçalves, Bom Jesus, Caixas do Sul, Cruz Alta, Encruzilhada do Sul, Irai, Lagoa Vermelha, Passo Fundo, Pelotas, Porto Alegre, Rio Grande, Santa Maria, Santa Vitoria do Palmar, Santana do Livramento, São Luiz Ganzaga, Torres e Uruguaiana.

O INMET disponibiliza essas séries em três tipos de periodicidade diferentes sendo elas por hora, dia e mês. Todas as estações das cidades listadas acima possuem dados a partir do ano de 1976 (ano de início das séries de soja da CONAB), com exceção da estação da cidade de Pelotas, onde os primeiros registros datam do mês de setembro de 1977. Entre os dados apresentados nelas estão disponíveis informações referentes ao vento, evaporação, insolação, nebulosidade, precipitação, temperatura e umidade.

3 MÉTODOS DE PREVISÃO

Métodos de previsão podem ser classificados basicamente em dois tipos: métodos qualitativos e métodos quantitativos. Os métodos qualitativos também podem ser chamados intuitivos ou subjetivos, eles dependem da experiencia de especialistas para prever o valor de um evento, os métodos quantitativos utilizam dados coletados sobre o objeto de previsão para a realização da previsão de um evento.

3.1 MÉTODO QUALITATIVOS

Métodos qualitativos buscam realizar previsões com base no conhecimento de especialistas no campo de estudo desejado, eles são baseados em planos, metas, expectativas, conhecimento técnico e intuição destes profissionais. Devido a esta característica subjetiva este método é frequentemente usado para previsões de médio e longo prazo, ou em situações em que não haja um precedente histórico ou que os dados para a análise sejam limitados.

Este método tende a apresentar tendências no seu processo preditivo que podem acabar distorcendo os resultados dela, entre os quais podemos citar, o otimismo, inconsistência, novidades, disponibilidade, correlações ilusórias, conservadorismo e percepção seletiva. Devido a estes fatores é importante a aplicação de técnicas para reduzir o impacto negativo destas tendências, e assim poder alcançar um melhor resultado final.

Entre os modelos que fazem uso do modelo previsão quantitativo podemos citar, o Jogo de Representação, Pesquisa de Intenções e Delphi (LEMOS, 2006).

3.2 MÉTODO QUANTITATIVO

Métodos quantitativos podem também ser conhecido por matemáticos, buscam realizar previsões com o emprego de modelagem matemática, são baseados no emprego de séries de dados históricos e buscam através das análises encontrar padrões de comportamento nesses, partindo do pressuposto que estes comportamentos deles tendem a se repetir no futuro. Este método pode ser divido em três modelos sendo eles: casuais, séries temporais e modelos de aprendizado de máquina (BRANCO; SAMPAIO, 2008).

Os modelos casuais podem ser utilizados para a realização que qualquer tipo de previsão de demanda, porém são mais indicados para previsões de médio e longo prazo. Eles pressupõem uma relação de causa e efeito entre as entradas e as saídas do sistema, então eles buscam encontrar a relação entre as duas através da análise dos resultados gerados na saída. Entre os modelos que fazem uso desta técnica podemos citar os de correlação, regressão simples, regressão múltipla e modelos econométricos (LIN, 2000).

Modelos baseados em séries temporais são mais indicados para previsões de curto prazo e são fundamentados na teoria de que o padrão antecedente da variável dependente persistirá no futuro, com isso podemos dizer que o método busca encontrar padrões nos valores vistos no passado, para assim proporcionar dados satisfatórios para a previsão deles no futuro (ARTHUS *et al*., 2017).

Uma série temporal pode ser classificada de acordo com quatro comportamentos ou efeitos relacionados a ela (PEINADO; GRAEML, 2007), sendo eles:

- Tendência: é o comportamento que os dados apresentam ao longo da série, podendo ser ele de crescimento, queda ou estabilidade, podendo ser ou não linear.
- Sazonalidade: é quando uma determinada variável apresenta um padrão de variação que se repete durante o tempo.
- Ciclicidade: é quando um padrão de flutuação de médio prazo da série afeta a sua tendência global.
- Componente Irregular: são flutuações imprevisíveis que apresentam um comportamento aleatório sem correlação temporal.

O reconhecimento destes fatores na série histórica é de grande importância para a seleção de um método adequado e melhora da acuracidade dos resultados, entre os diversos métodos atualmente disponíveis na literatura o quadro 1 a seguir expõe um resumo de alguns destes modelos.

Quadro 1 - Modelos de previsão de séries temporais

MODELO	RESUMO
MÉDIA MÓVEL	Aplicabilidade se dá em demanda sem tendência e sem sazonalidade. É usada para previsão a partir da média aritmética de n números seguidos de demanda real, anteriores à previsão.
ALISAMENTO EXPONENCIAL SIMPLES	Aplicabilidade em demanda sem tendência e sem sazonalidade. Esse método assume a previsão do período anterior e a complementa com um ajuste para ter a previsão do período seguinte.
MODELO DE HOLT	Aplicabilidade em demanda com tendência e sem sazonalidade. É pertinente sua utilização quando a demanda tem nível e tendência no componente sistemático. A demanda e o tempo têm relação linear.
MODELO DE WINTER	Aplicabilidade em demanda com tendência e sazonalidade. O modelo de Winter é utilizado em organizações cuja sua demanda apresenta variação no seu aspecto de nível, tendência e sazonalidade.
MODELO ARIMA	A forma mais geral do modelo não considera a sazonalidade. O modelo ARIMA ajusta os valores observados para que a diferença entre esses valores e os valores produzidos no modelo seja próximo de zero. O modelo avalia a autocorreção e a auto correlação parcial entre os dados dentro dos valores críticos e distingue os padrões aleatórios (ruído branco).

Fonte: adaptado de ARTHUS *et al*. (2017)

Os modelos de aprendizado de máquina têm as suas primeiras pesquisas datadas da década de 80 e vem se estabelecendo atualmente como candidatos alternativos a modelos clássicos de séries temporais como o ARIMA. Atualmente existem diversas técnicas capazes de trabalhar com séries históricas sendo mais frequentemente encontrados estudos empregando RNAs (AHMED *et al*., 2010). Contudo, além de RNAs, outras técnicas que também podem ser empregadas dentre

as quais podemos citar: sistemas de inferência *Fuzzy*, programação genética, e máquinas de vetor de suporte (WANG et al., 2009). No capítulo 4 será abordado mais detalhadamente o uso de aprendizado de máquina como técnica para realização de previsões, com enfoque na utilização de RNAs.

3.3 ERRO DE PREVISÃO

Para um método de previsão ser considerado bom, é necessário ele apresentar um erro estatístico semelhante a característica de imprevisibilidade da demanda (ARTHUS *et al*., 2017). A literatura normalmente apresenta este valor de erro em medidas que estão na forma de percentuais de erro absoluto ou erros quadráticos. Os valores são obtidos realizando a análise de medidas de tendência central como medianas, médias aritméticas e médias geométricas (LEMOS, 2006).

A seleção de um método de medida de erro varia de acordo com a situação de uso do mesmo e do número de séries temporais analisadas, também podem ser utilizados mais de um método afim de compensar erros das diferentes equações. A seguir o quadro 2 apresenta algumas das medidas de acurácia disponíveis para serem utilizadas em processos preditivos, onde Y_i é o valor de demanda atual no período i, $\hat{Y}_i$ é a previsão de demanda no período i e n é o número de períodos considerados para o cálculo da medida de acurácia (LEMOS, 2006).

Quadro 2 - Equações de medida de acurácia

SIGLA	SIGNIFICADO	EQUAÇÃO
ME	*Mean error* Erro médio	$ME = \frac{1}{n}\sum_{i=1}^{n} Y_i - \hat{Y}_i$
MAE	*Mean absolute error* Erro absoluto médio	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert Y_i - \hat{Y}_i \rvert$
MPE	*Mean percentual error* Erro percentual médio	$MPE = \frac{1}{n}\sum_{i=1}^{n} \left[(Y_i - \hat{Y}_i) \,/\, Y_i \right]$
APE	*Absolute percentual error* Erro percentual absoluto	$APE = \left\lvert \frac{Y_i - \hat{Y}_i}{Y_i} \right\rvert$

SIGLA	SIGNIFICADO	EQUAÇÃO
MSE	*Mean squared error* Erro quadrático médio	$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$
RMSE	*Root mean squared error* Raiz do erro quadrático médio	$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right)^{\frac{1}{2}}$
GRMSE	*Geometric root mean squared error* Raiz da média geométrica do erro quadrático	$GRMSE = \left(\prod_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2\right)^{\frac{1}{2n}}$

Fonte: Adaptado de LEMOS (2006)

Entre as equações de cálculo de erro listadas acima, podemos dar um destaque especial para os métodos MAE e RMSE. Eles são comumente utilizados em trabalhos onde há o uso de RNAs como nos trabalhos realizados por HAIDER *et al*. (2019) e por KUNG *et al*. (2016).

A equação MAE, mede o desvio absoluto médio entre um valor previsto e o valor alvo. Para calculá-lo é preciso subtrair o valor real da série do valor previsto, e então transformar este valor em absoluto. Após deve ser calculada a média te todos os valores de erro absolutos (GARMSIRI, 2018).

É importante destacar que o como o MAE trabalha com valores absolutos a direção dos dados não tem importância para o cálculo, e devido ele ser uma pontuação linear todos os erros individuais são ponderados igualmente na média. Quando mais próximo de 0 for o valor do MAE, melhor foi o desempenho de previsão do sistema (SILVA, 2017).

A equação RMSE, mede o erro quadrático médio dos valores previstos. Para calculá-lo é preciso subtrair o valor real da série do valor previsto e elevar este número ao quadrado. Após deve ser calculada a média destes valores e calculada a raiz quadrada da média para que a escala do erro calculado seja igual a escala do destino (DRAKOS, 2018).

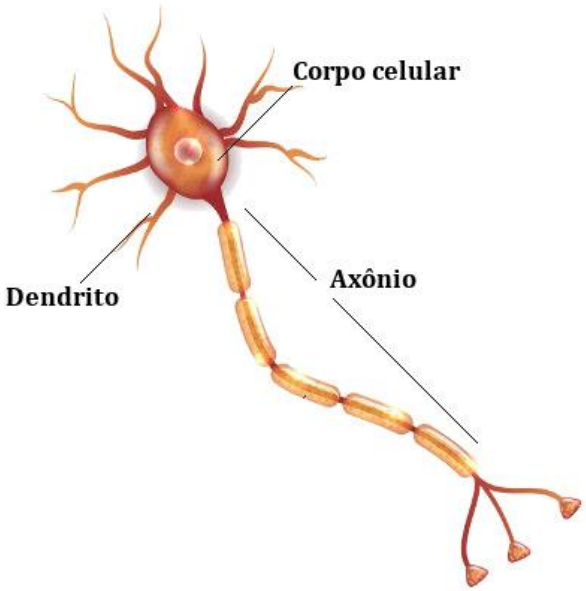
A principal diferença do RMSE, quando comparado ao MAE é que devido ele elevar o valor da diferença calculada ao quadrado ele penaliza mais sistemas onde há uma discrepância maior nos dados. O Valor de RMSE também pode ser chamado de desvio padrão (GARMSIRI, 2018).

4 REDES NEURAIS ARTIFICIAIS

Segundo GURNEY (2014), uma RNA pode ser preliminarmente descrita por ser um conjunto interconectado de elementos, unidades ou nós de processamento simples, cuja a funcionabilidade é vagamente baseada em um neurônio animal. A capacidade de processamento da rede é armazenada nos seus pontos fortes ou nos pesos de conexão entre as unidades, obtidos por um processo de adaptação ou aprendizado de um conjunto de dados de treinamento.

Um cérebro humano possui cerca de 100 bilhões de neurônios, e cada um desses neurônios se comunica com milhares de outros neurônios continuamente e em paralelo. Os neurônios são divididos em 3 seções, conforme pode ser observado na figura 1: o corpo da célula, os dendritos e o axônio. A função dos dendritos é receber as informações (impulsos nervosos), transmitidos de outros neurônios e conduzi-las até o corpo celular. Nele essa informação recebida e processada e novos impulsos são gerados, esses impulsos são transmitidos pelo axônio até o dendrito do próximo neurônio, o ponto onde ocorre esta conexão entre o axônio de um neurônio e o dendrito de outro é chamada de sinapse. O conjunto de diversas dessas ligações forma uma rede neural (BRAGA; CARVALHO; LUDERMIR, 2000).

Figura 1 - Representação gráfica de um neurônio animal



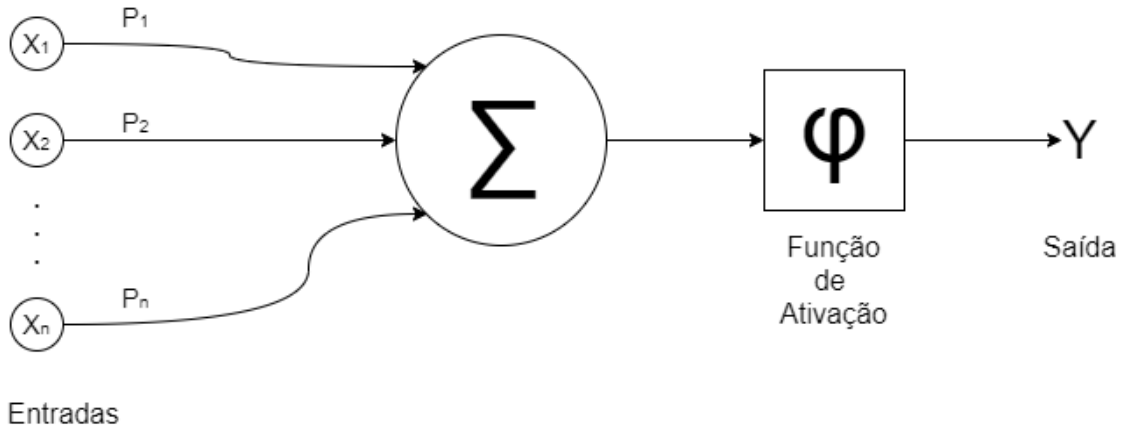
Fonte: Adaptado de SANTOS (2019a)

A busca pelo desenvolvimento de um modelo computacional que simule o funcionamento de um neurônio data da década de 40 com o trabalho desenvolvido por McCulloch e Pitts em 1943. Os estudos tiveram grandes avanços até a década de 70, onde acabaram tendo o seu desenvolvimento freado devido a problemas relacionados a limitações metodológicas e tecnológicas, no entanto na década de 80 devido ao avanço nos recursos computacionais e a avanços metodológicos importantes o estudo das redes neurais artificiais retornou com força (FERNEDA, 2006).

Um neurônio artificial é composto por três elementos básicos conforme demonstrado na figura 2 a seguir:

- Um conjunto de n conexões de entrada $(x_1, x_2, x_3, \dots, x_n)$, caracterizadas por pesos $(p_1, p_2, p_3, \dots, p_n)$.
- Um somador (∑) para acumular os sinais de entrada.
- Uma função de ativação (φ) que limita o intervalo permissível de amplitude do sinal de saída (y) a um valor fixo.

Figura 2 - Diagrama de um neurônio artificial



Fonte: Adaptado de HAYKIN (2001)

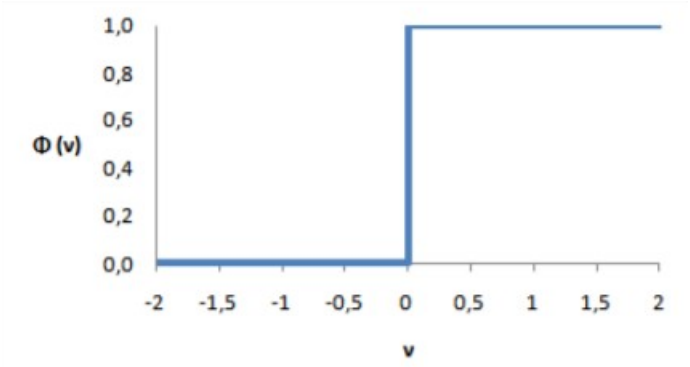
A simulação do comportamento das conexões de um neurônio animal, é realizada pelos pesos associados as entradas, eles podem ser negativos ou positivos, dependendo se o seu tipo for inibitório ou excitatório. O fluxo da informação no neurônio começa com ele recebendo em suas entradas um sinal (valor) oriundo de um outro neurônio. Esse valor é multiplicado pelo peso da entrada correspondente

$(x_n\ X\ p_n)$, os valores de resultantes dessa multiplicação são então somando e enviados para a função de ativação, que então define com base no valor recebido qual será a saída (y) do neurônio (FERNEDA, 2006).

As funções de ativação representadas pelo símbolo φ na figura 2, podem ser divididas em 3 tipos básicos:

Função de ativação Limiar: conforme descrito na figura 3, apresenta a saída de um neurônio k que assume o valor 1, se o valor de entrada da função de ativação não for negativo e 0 em caso contrário. A função pode ser expressa pela equação $y_k = \begin{cases} 1\ se\ v_k \geq 0 \\ 0\ se\ v_k < 0 \end{cases}$ onde o valor de v_k é a entrada da função de ativação.

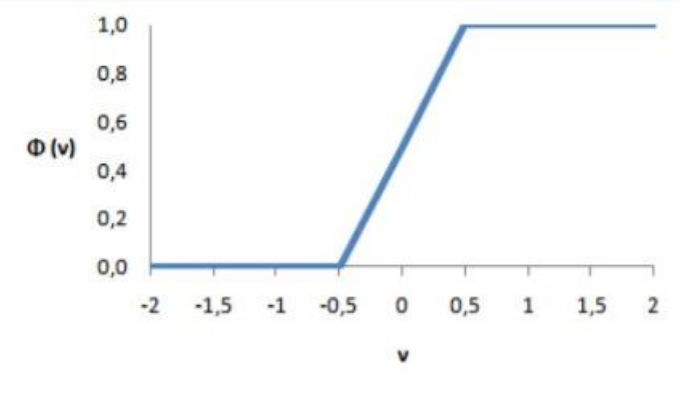
Figura 3 - Representação de uma função limiar



Fonte: Adaptado de HAYKIN (2001)

Função Linear por Partes: conforme descrito na figura 4, esta função de ativação pode ser vista como uma aproximação de um amplificador não-linear. Ela pode ser expressa pela seguinte função, $\varphi(v) = \begin{cases} 1\ se\ v \geq 0{,}5 \\ v/2\ se\ +0{,}5 > v > -0{,}5 \\ 0\ se\ v \leq -0{,}5 \end{cases}$

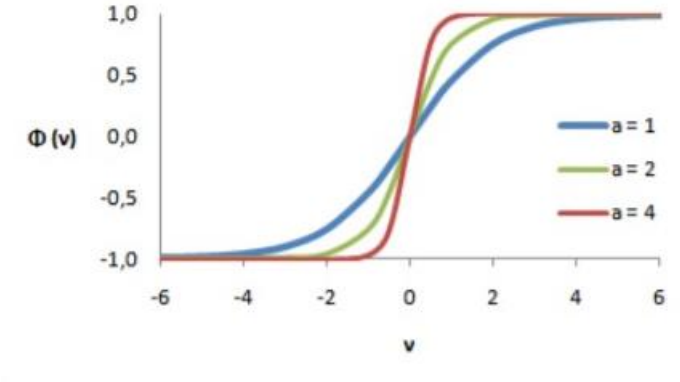
Figura 4 - Representação de uma função linear por partes



Fonte: Adaptado de HAYKIN (2001)

Função Sigmoide: conforme descrito pela figura 5, possui um gráfico em forma de S, é a forma mais utilizada na construção de redes neurais artificiais. É definida como uma função crescente, que apresenta um balanço entre o comportamento linear e não-linear, um exemplo de função sigmoide é a função de logística definida pela fórmula $\varphi(v) = \frac{1}{1 + exp(-av)}$ onde a é o parâmetro de inclinação da função sigmoide.

Figura 5 - Representação de uma função sigmoide com 3 inclinações distintas



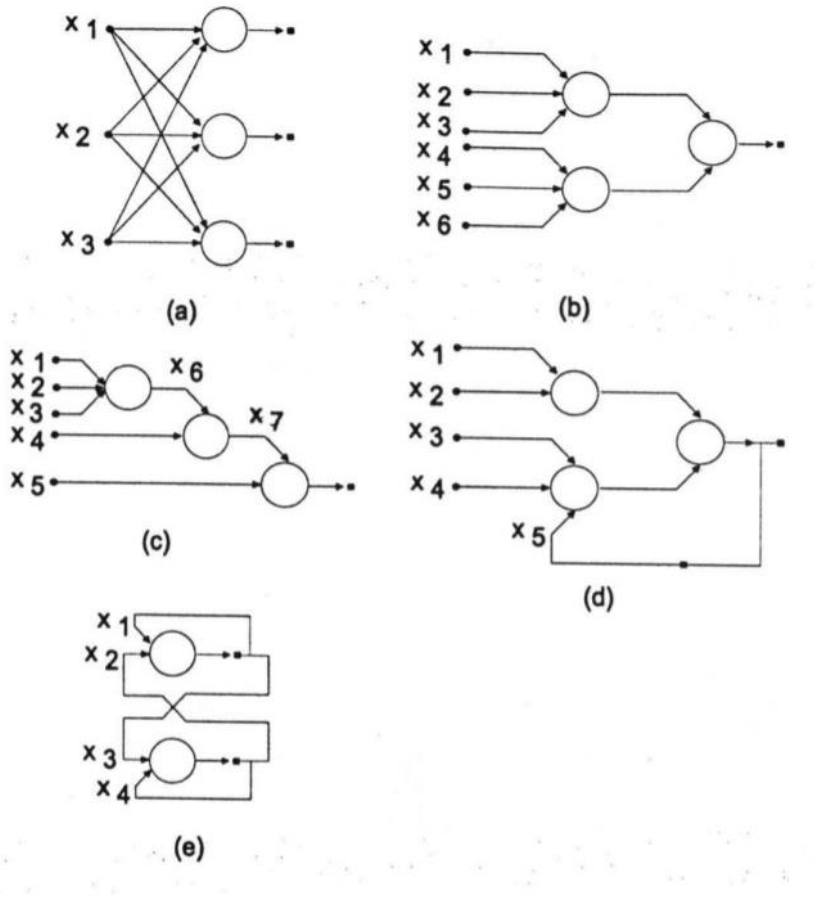
Fonte: adaptado de HAYKIN (2001)

Fazendo-se a combinação de vários desses neurônios temos uma RNA, diferentes formas de combinação destes neurônios formam diversos tipos diferentes de RNAs, sendo cada uma indicada para algum tipo de problema. Por exemplo, redes com uma camada única de nodos MCP (McCulloch-Pitts) só conseguem resolver problemas linearmente separáveis. Já redes recorrentes são mais indicadas para

resolver problemas que necessitem de processamento temporal (BRAGA; CARVALHO; LUDERMIR, 2000).

Como pode ser observado na figura 6, a arquitetura da RNA pode ser classificada segundo alguns parâmetros apresentados a seguir:

Figura 6 - Exemplos de arquiteturas de RNAs



Fonte: (BRAGA; CARVALHO; LUDERMIR, 2000)

- Número de camadas
    - Redes de camada única: Só existe um nó entre qualquer entrada e qualquer saída da rede. Exemplo figura 6 letras a, e.
    - Redes de múltiplas camadas: existe mais de um neurônio entre alguma entrada e alguma saída da rede. Exemplo figura 6 letras b, c, d;
- Tipo de conexões

- *Feedforward* ou acíclica: a saída de um neurônio na i-ésima camada da rede não pode ser usada como entrada de nodos em camadas de índice menor ou igual a i. Exemplo figura 6 letras a, b, c.
- *Feedback*, cíclica ou recorrente: a saída de um neurônio na i-nésima camada da rede é usada como entrada de nodos em camadas de índice menor ou igual a i. Exemplos figura 6 letras d, e.
- Redes cuja saída final (única) é ligada às entradas e comportam-se como autômatos reconhecedores de cadeia, onde a saída que é realimentada fornece o estado do autômato. Exemplo figura 6 letra d.
- Auto associativas: nesta rede todas as ligações são cíclicas, elas associam um padrão de entrada com ele mesmo, e são particularmente úteis para a recuperação ou "regeneração" de um padrão de entrada. Exemplo figura 6 letra e.

- Conectividade
  - Rede fracamente ou parcialmente conectada. Neste tipo de rede um neurônio não está conectando a todos os outros da próxima camada. Exemplo figura 6 letras b, c, d.
  - Rede completamente conectada. Neste tipo de rede o neurônio está ligado a todos os neurônios da próxima camada. Exemplo figura 6 letras a, e.

4.1 APRENDIZADO

A sua capacidade de aprender por meio de exemplos é uma das características mais importantes das RNAs. Esse processo de aprendizado ocorre através de um processo interativo de ajustes dos pesos entre as conexões da rede, que ao final do processo, guardam o conhecimento que a rede adquiriu do ambiente externo (BRAGA; CARVALHO; LUDERMIR, 2000).

Existem diversos algoritmos de aprendizado, que basicamente diferem na forma como os ajustes que devem ser aplicados aos pesos é calculado. Esses algoritmos são separados em dois paradigmas principais, o de aprendizado supervisionado e o de aprendizado não supervisionado.

4.1.1 Aprendizado Supervisionado

Este paradigma implica na necessidade de obrigatoriamente haver um supervisor ou professor externo, o qual deve apresentar para rede os padrões de entrada e observar as saídas geradas pela mesma, comparando-as com as saídas desejadas. O supervisor então fornece para a rede informações sobre a direção em que deve ser realizados os ajustes dos pesos. Este processo é feito incrementalmente diversas vezes de forma que estes valores caminhem para se possível uma solução.

O aprendizado supervisionado se aplica a problemas em que se deseja obter um mapeamento entre padrões de entrada e saída. Ele pode ser implementado de maneira *off-line* onde uma vez obtida uma solução para a rede os dados do conjunto de treinamento de não mudam, ou de maneira *on-line* onde o conjunto de dados muda de maneira contínua, e a rede deve estar continuamente em adaptação (BRAGA; CARVALHO; LUDERMIR, 2000).

4.1.2 Aprendizado Não Supervisionado

Neste paradigma não há a presença de um professor ou supervisor externo acompanhando o processo de aprendizagem. Para o treinamento da rede somente os valores de entrada estão disponíveis, estes valores são apresentados continuamente à rede, e a existência de regularidade nesses dados faz com que seja possível realizar o aprendizado desta rede, sendo a existência desta característica imprescindível para haver o aprendizado.

Este paradigma aplica-se a problemas que tem por objetivo a descoberta de características estatisticamente relevantes nos dados de entrada, como por exemplo a descoberta de agrupamentos ou classes (BRAGA; CARVALHO; LUDERMIR, 2000).

4.2 REDES *FEEDFORWARD* DE MULTIPLAS CAMADAS

Tipicamente esta variante de rede neural consiste em um conjunto de unidades sensoriais (nós de fonte) que constituem a camada de entrada, uma ou mais camadas ocultas de nós de processamento computacionais e uma camada de saída de nós computacionais. Nela o sinal de entrada se propaga para frente através da rede

camada por camada. Estas redes costumam ser chamadas de perceptrons de múltiplas camadas (MPL, *multilayer perceptron*) (HAYKIN, 2001).

Uma rede MPL possui três características importantes, sendo elas:

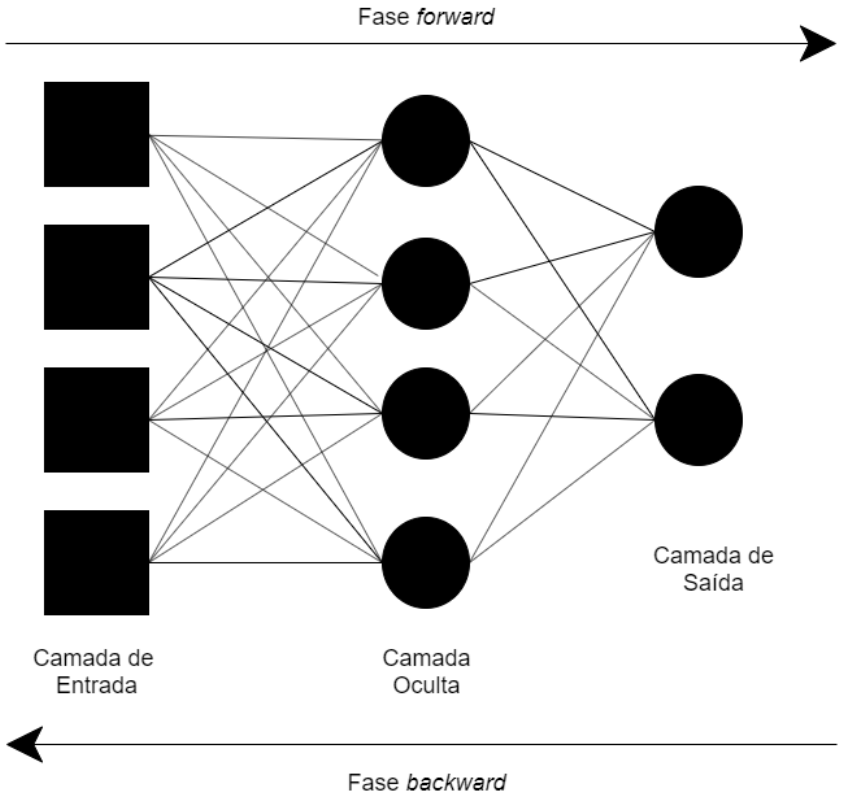
- Cada neurônio faz uso de uma função de ativação não linear, normalmente utilizando a não linearidade sigmoide definida pela função logística. Essa presença é importante para evitar que a relação de entrada e saída da rede se reduza à de um *perceptron* de camada única, além de ter a motivação biológica de levar em conta a fase refratária dos neurônios reais.
- A rede deve conter uma ou mais camadas de neurônios oculta, que não são parte da entrada ou da saída da rede. Estes neurônios são os responsáveis pelo aprendizado de tarefas complexas, extraindo as características mais significativas dos valores de entrada.
- A rede exibe um elevado grau de conectividade, determinado pelas sinapses da rede. Uma mudança na conectividade da rede requer uma mudança na população de conexões sinápticas ou de seus pesos.

4.2.1 Algoritmo *Back-propagation*

O algoritmo *back-propagation* é o mais popular para o treinamento de RNAs do tipo MPL, e quando não utilizado diretamente é utilizada normalmente alguma variante sua. Nele o treinamento da rede ocorre em duas fazes, sendo elas a de *forward* onde a rede é abastecida por um vetor de entrada e as saídas dos neurônios da primeira camada oculta são calculados, esses neurônios proverão os valores de entrada da camada oculta seguinte e assim por diante até chegarem a camada de saída, nela as saídas produzidas pela rede são comparadas as saídas desejadas para o vetor de entrada e calculado o seu erro correspondente.

Durante a fase *backward,* o erro calculado na camada de saída é utilizado para atualizar os seus pesos com o uso do gradiente descendente do erro. Esse erro então é propagado para a camada oculta anterior, onde é calculada a medida de influência de cada neurônio na camada de saída, tendo assim um valor estimado e erro de cada neurônio. Este erro é então utilizado para realizar o ajuste dos pesos do neurônio, da mesma forma que ocorre na camada de saída. Esse processo se repete até o momento em que se chega na primeira camada oculta e todos os pesos da rede estejam atualizados. A figura 7 a seguir traz uma representação visual do processo.

Figura 7 - Fases do algoritmo *back-propagation*



Fonte: Adaptado de BRAGA; CARVALHO; LUDERMIR (2000)

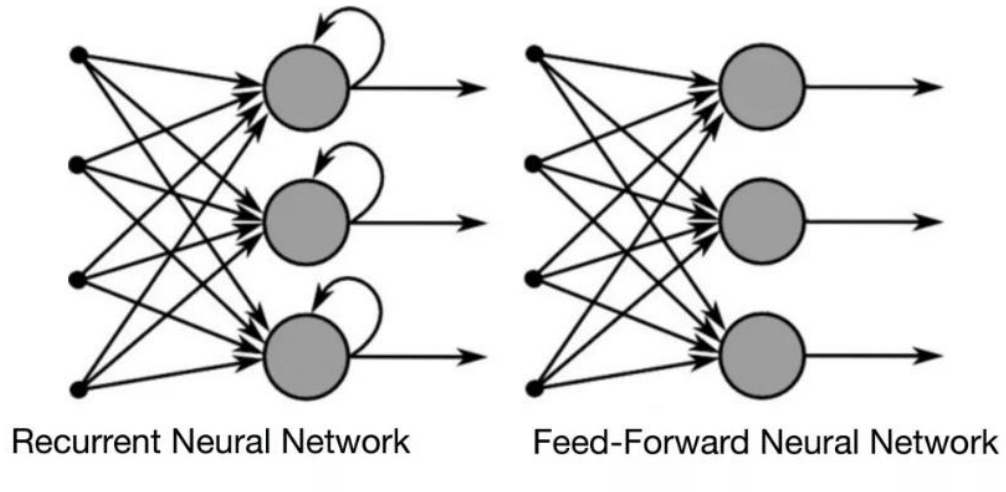
O algoritmo *back-propagation* é capaz de resolver boa parte dos problemas relacionados a classificação, regressão e previsão (BRAGA; CARVALHO; LUDERMIR, 2000). Estudos que o comparam à outros algoritmos trazem bons resultados. O trabalho realizado por TSAI; SHIUE, (2004) onde utilizando RNAs com o algoritmo de *back-propagation*, foi possível observar que o algoritmo teve um melhor desempenho em prever a produção de capim-elefante em Taiwan, quando comparado ao uso de regressão linear múltipla. Já no trabalho de KUNG *et al*., (2016), eles empregaram um conjunto de RNAs com o algoritmo de *back-propagarion* e conseguiram prever com melhor desempenho a safra de tomates em Taiwan, quando comparado com o algoritmo de regressão por etapas.

4.3 REDES RECORRENTES

As redes neurais recorrentes mais conhecidas como RNNs (*Recurrent Neural Networks*), são uma classe de redes neurais artificiais onde a saída gerada pelos

neurônios na camada oculta podem ser usadas posteriormente como entrada para esse neurônio, trazendo para esse neurônio uma espécie de “memória”, com isso os pesos da rede são compartilhados a atualizados ao longo do tempo (AMIDI; AMIDI, 2019).

Figura 8 - Comparação de uma rede neural recorrente com uma de *Feedforward*



Fonte: (CAPÍTULO 48 - REDES NEURAIS RECORRENTES, 2019)

A figura 8 acima compara o diagrama de uma rede do tipo *Feedforward* abordada no tópico anterior com o de uma rede neural recorrente. Podemos perceber que a principal diferença entre os dois diagramas está nos *loops* de *feedback* conectados aos neurônios da camada oculta da rede recorrente. Devido a esse processo de *feedback*, uma decisão tomada por um neurônio em um momento $t - 1$ afeta a decisão que este neurônio tomará mais tarde no tempo t e assim sucessivamente.

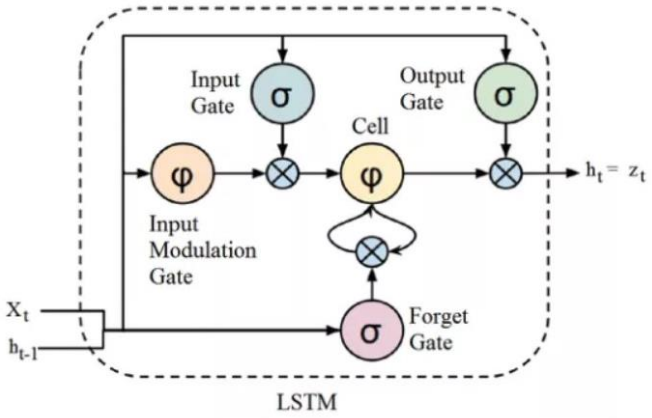
4.3.1 Redes Neurais *Long Short Term Memory*

As redes neurais *Long Short Term Memory* (LSTM), são uma variante de RNNs, elas foram propostas por HOCHREITER; SCHMIDHUBER, (1997), tendo como objetivo aprender padrões complexos de estruturas com dependências temporais e vários estágios de processamento. O diferencial das redes LSTM em comparação as

redes recorrentes convencionais está na capacidade de armazenar informações por longos períodos ao processar uma sequência temporal (SANTOS, 2019b).

A arquitetura das redes LSTM são baseadas em portas, cujas suas funções são verificar se os dados serão mantidos ou descartados. Ao todo uma célula LSTM possui 3 portas, sendo elas, conforme pode ser observado na figura 9 abaixo: Porta de entrada, é a responsável por avaliar a informação que está chegando no momento t na célula e informar se o valor deve ou não ser memorizado pela célula. Porta de saída, tem a tarefa de extrair as informações úteis do estado atual da célula e disponibilizá-las para a próxima célula de processamento. E a porta de esquecer que é a responsável por apagar informações que não são mais uteis naquele momento para a célula (HAIDER *et al*., 2019).

Figura 9 - Diagrama de uma célula LSTM



Fonte: (CAPÍTULO 51 - ARQUITETURA DE REDES NEURAIS LONG SHORT TERM MEMORY (LSTM), 2019)

Trabalhos empregando o uso de redes LSTM, em diversas áreas do conhecimento, tem conseguido obter bons resultados, como no trabalho proposto por SANTOS, (2019b), onde a utilização de redes LSTM apresentou um desempenho superior em relação ao modelo de simulação *Bootstrap*, ao prever os preços da energia elétrica no mercado brasileiro, mostrando-se uma ferramenta com bons resultados na análise de movimentações dos preços. Já no trabalho realizado por HAIDER *et al*., (2019), as redes LSTM, foram empregas para a previsão da produção de trigo no Paquistão, foi possível observar que elas apresentaram melhores

resultados quando comparados a redes neurais recorrentes tradicionais e ao modelo ARIMA, o capítulo a seguir irá abordar este trabalho de forma mais detalhada.

5 TRABALHOS CORRELATOS

HAIDER *et al*. (2019), propuseram em seu trabalho realizar a previsão da safra de trigo no Paquistão devido a sua grande importância para o país. O trigo é a segunda maior colheita do país, para isto eles fizeram o uso de redes neurais LSTM, os modelos de previsão foram desenvolvidos com o software MATLAB 2018 ®, juntamente com o seu kit de ferramentas para o desenvolvimento de aprendizado profundo. No estudo também foram desenvolvidos modelos utilizando RNNs convencionais e com o método ARIMA, a fim de comparar os seus resultados com os das redes LSTM.

No desenvolvimento do modelo foram utilizados dados de produção de trigo compreendido entre os anos de 1902 a 2018, os dados foram obtidos do *Federal Bureau of Statistics, Pakistan*, e do *Economic Survey of Pakistan*. Esses dados foram então divididos pelos autores em dois grupos, um com dados compreendidos entre os anos de 1902 a 2008, utilizados para treinar os modelos, e outro grupo com os dados entre os anos de 2009 a 2018, utilizados para fins de validação dos modelos.

Devido ao fato de os dados brutos apresentarem grandes flutuações, os autores propuseram então o uso da função de suavização conhecida como *Robust-LOWESS*, tendo em vista que valores muito discrepantes no modelo podem afetar negativamente os resultados da previsão realizada. Os três modelos então foram treinados de duas maneiras, uma com os dados suavizados e outra com os dados brutos, tendo como resultado a elaboração de seis modelos. Estes modelos foram comparados no estudo utilizando as fórmulas MAE, RMSE e com o valor de confidente de relação (valor R).

No estudo eles utilizaram o modelo ARIMA, denominado ARIMA (p, d, q), onde p, d e q são, ordem dos modelos de regressão automática, diferenciação e média móvel, respectivamente, que foram preenchidos com os seguintes valores (1, 2, 2). Já para a elaboração da RNN foi feito o uso do modelo NAR (*Nonlinear autoregressive neural network*) com os parâmetros apresentados na tabela 2 abaixo. Da mesma maneira que o modelo RNN, os parâmetros utilizados para a elaboração do modelo LSTM são apresentados abaixo na tabela 3.

Tabela 2 - Parâmetros utilizados na confecção da rede RNN

PARÂMETRO	VARIAÇÃO
ATRASOS NO *FEEDBACK*	1 a 5
Nº DE CAMADAS OCULTAS	1 a 3
Nº DE NEURÔNIOS EM CADA CAMADA OCULTA	1 a 40
FUNÇÕES DE TREINAMENTO	*Bayesian regularization* e *Levenberg–Marquardt*

Fonte: Adaptado de HAIDER *et al*. (2019)

Tabela 3 - Parâmetros utilizados na confecção da rede LSTM

PARÂMETRO	VARIAÇÃO
UNIDADES OCULTAS	1 a 2000
TAXA INICIAL DE APRENDIZADO	0,001 a 0,009
PERÍODO DE QUEDA DA TAXA DE APRENDIZAGEM	1 a 100
SOLVER	*Adam optimizer* e *stochastic gradient descent with momentum (SGDM)*

Fonte: Adaptado de HAIDER *et al*. (2019)

Como resultados do estudo foi possível apontar que o uso do modelo LSTM, com os dados que passaram pela função de suavização apresentaram melhores resultados, com um ganho de 14% em relação ao modelo de *benchmark* utilizando a fórmula de verificação de erro MAE, já com a forma RMSE esse ganho chega a 25%.

Na tabela 4 é exibida uma comparação dos valores de erro dos diversos métodos estudados. Nela é possível observar como os modelos que empregam o LSTM com dados brutos sem tratamento e com LSTM pré-processado apresentam valores menores de MAE e RMSE, em comparação aos outros modelos, trazendo como conclusão para o estudo que a aplicação destas técnicas é indicada devido a sua maior eficácia, para o problema proposto.

Tabela 4 - Valores de erro encontrados no estudo

MODELO	RMSE	MAE	VALOR R
ARIMA (1,2,2)	1065	847	0,80
ARIMA (1,2,2) PRÉ-PROCESSADO	1420	1248	0,81
RNN	1754	1313	0,58
RNN PRÉ-PROCESSADO	1379	997	0,79
LSTM	1002	808	0,81
LSTM PRÉ-PROCESSADO	792	729	0,81

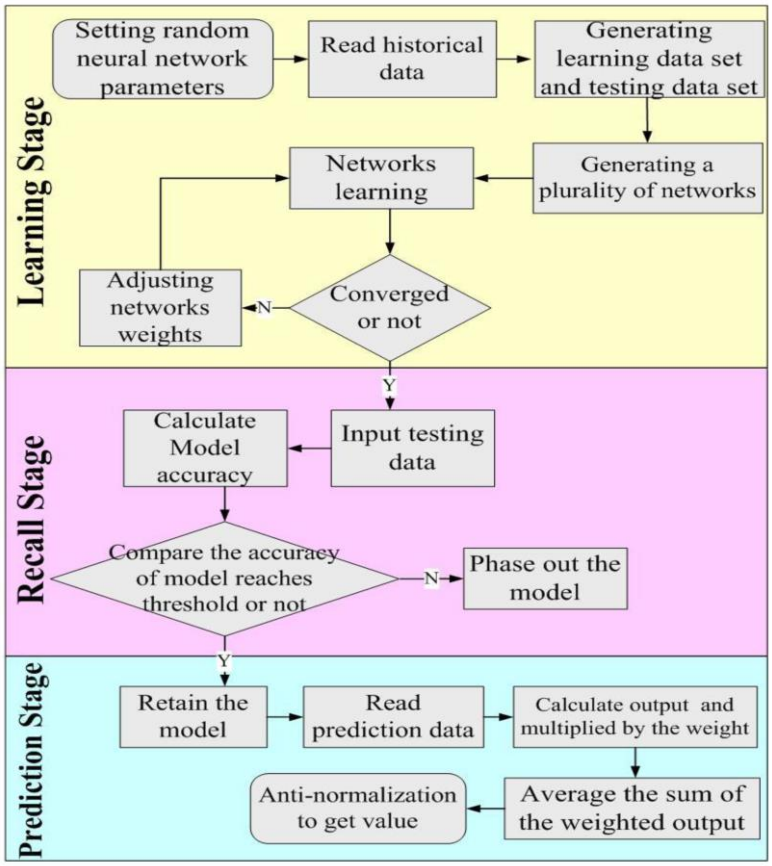
Fonte: Adaptado de HAIDER *et al*. (2019)

Em outro estudo, apresentado por KUNG *et al*. (2016), no qual foi proposto uma alternativa aos métodos de previsões agrícolas usado pelo Conselho de Agricultura de Taiwan, para isso eles desenvolveram um estudo utilizando *ensemble neural network* (ENN), baseado em redes do tipo *back-propagation*.

Para fazer isto eles acumularam dados referentes a fatores meteorológicos como por exemplo umidade relativa do ar, precipitação, temperatura, fatores ambientais como por exemplo área de plantio, área de colheita e fatores econômicos como custo de produção e valor de venda dos produtos. Devido esses dados terem como origem diversas fontes, eles foram integrados em uma única base de dados, limpos para evitar informações incompletas ou incorretas e posteriormente normatizados.

Esses dados então são inseridos no modelo ENN desenvolvido. Após nele são gerados aleatoriamente diversas redes neurais do tipo *back-propagation*, cada uma dessas redes criadas possui uma quantidade distinta de neurônios e camadas ocultas. Essas redes então são treinadas com os dados inseridos nelas e cada uma delas tem seus pesos ajustados conforme for ocorrendo os processos de aprendizagem. Ao final do processo de aprendizagem, as redes são testadas, a fim de se obter a precisão do modelo, os que não atingirem o limite de precisão definido são descartados. Após esse processo todos os dados inseridos para previsão no sistema são então avaliados pelas redes restantes, seus resultados são somados e calculada a média ponderada. A figura 10 mostra de maneira visual o funcionado da rede ENN.

Figura 10 - Esquema de funcionamento da ENN



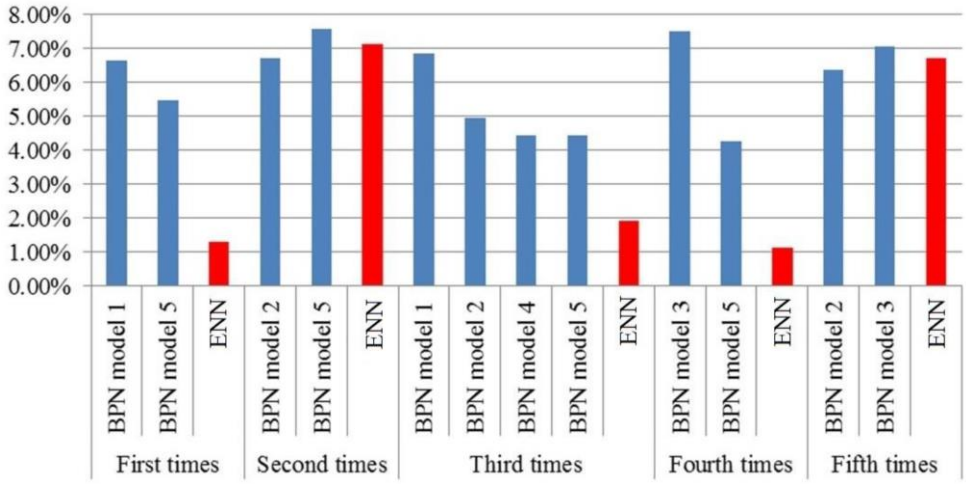
Fonte: (KUNG *et al.*, 2016)

Como ambiente experimental desde modelo proposto pelos autores, foram utilizados 9953 registros sobre o cultivo de tomates em Taiwan, que compreendem os anos de 1997 a 2014. O sistema então gerou 5 modelos de redes neurais distintos com até 5 camadas ocultas de até 5 neurônios, dos dados inseridos nestas redes 60% deles foi utilizado para treinamento e os 40% restantes para validação e definido um limiar de precisão de 90%. No primeiro experimento realizado essas redes geraram taxas de precisão de 90,81%, 86,70%, 88,10%, 89,87%, 93,30%, então somente as redes 1 e 5 com mais de 90% de precisão foram mantidas.

A taxa de erro ponderada no primeiro experimento foi de 1,30%, tendo como valor calculado 191.240 kg, frente aos 191.500 kg do rendimento real. Esse percentual de erro é menor quando comprado ao dos modelos *back-progagation* gerados que individualmente que obtiveram 6,64% e 5,47% de erro respectivamente. O experimento então foi repetido por mais 4 vezes totalizando 5 modelos ENN gerados para os dados de entrada. No gráfico 5 é possível ver uma comparação entre o erro

que as redes *back-propagation* obtiveram individualmente em cada modelo e com o valor resultante do modelo ENN proposto no estudo.

Gráfico 5 - Comparação dos erros resultantes de cada experimento



Fonte: (KUNG *et al*., 2016)

Como conclusão final dos estudos os autores afirmam que aplicação dos modelos com o método de ENNs, na maioria dos experimentos realizados apresentou uma menor taxa de erro e uma maior precisão, quando comparado com a utilização tradicional das redes *back-propagation* e com a análise de regressão múltipla que foi realizada para a validação dos modelos de teste, tendo esta apresentado um erro 12,4% superior aos algoritmos propostos.

6 MODELO DE PREVISÃO

Para a elaboração do estudo foram desenvolvidos quatro modelos de previsão com a utilização de redes neuras artificiais do tipo LSTM. Também foram elaborados dois tipos de conjuntos de dados que foram utilizados para alimentar os modelos de previsão com as informações necessárias para o treinamento deles. Neste capítulo será descrito os métodos e processos utilizados na estruturação e elaboração destes conjuntos de dados e modelos de previsão.

6.1 DESENVOLVIMENTO DO CONJUNTO DE DADOS

Para a elaboração dos dois conjuntos de dados foram utilizados dados provenientes da Companhia Nacional de Abastecimento e do Instituto Nacional de Meteorologia. Os dados brutos, oriundos desses órgãos, foram processados e beneficiados através de scripts em Python[3] e com o auxílio do software Microsoft Excel[4]. O procedimento detalhado será listado nos tópicos a seguir.

6.1.1 Conjunto de dados da CONAB

A CONAB disponibiliza os dados históricos das safras nacionais através do seu site na sessão de informações agropecuárias (CONAB, 2019a). As informações referentes a safra de soja são disponibilizadas em um arquivo no formato .xls (*Excel Binary File Format*), agrupadas no formato de safras que se repetem a cada período de um ano. Os dados são frequentemente atualizados pelo órgão onde ao canto esquerdo do site é possível verificar a data de quando o arquivo foi atualizado.

O arquivo contém dados históricos desde a safra 1976/77 até a previsão para a próxima safra, segmentados por estado e por região, e fornece informações sobre:

- Área plantada, em escala de mil hectares.
- Produtividade, em escala de quilograma por hectare.
- Produção, em escala de mil toneladas.

[3] Python é uma linguagem de programação de alto nível, com um modelo de desenvolvimento comunitário e aberto. Mais informações sobre a linguagem podem ser obtidas pelo link www.python.org

[4] O Microsoft Excel é um software de edição de planilhas eletrônicas, com diversas ferramentas para trabalhar com conjunto de dados. Mais informações sobre o programa podem ser obtidas através do link www.microsoft.com/pt-br/microsoft-365/excel

Para a elaboração do conjunto de dados do estudo foram utilizados os valores de área plantada, produtividade e produção para o estado do Rio Grande do Sul de todo o período disponível no arquivo, exceto a previsão para o ano de 2020.

6.1.2 Conjunto de dados do INMET

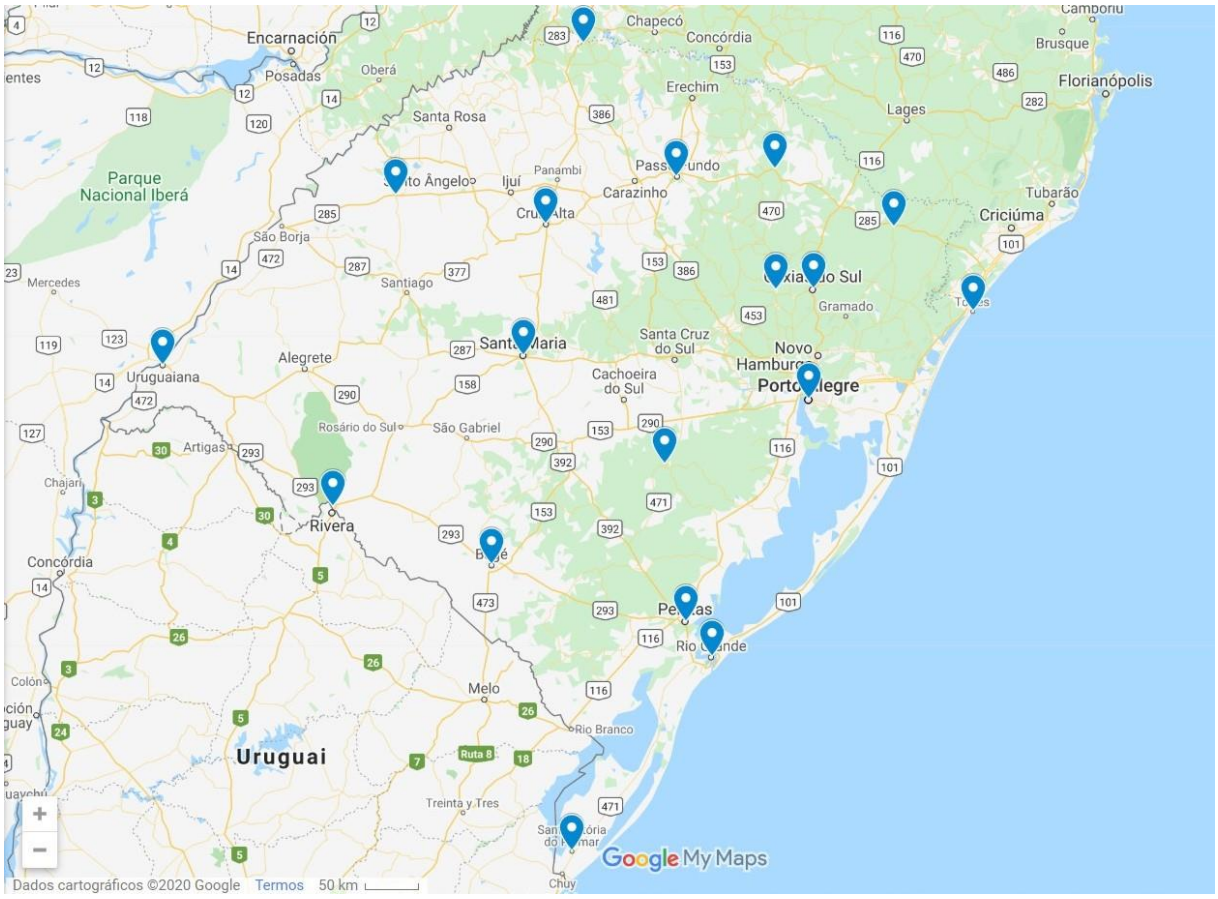
O INMET (INMET, 2019) disponibiliza em seu site, um banco de dados de informações meteorológicas para ensino e pesquisa. Para acessá-lo é necessário realizar um cadastro no site, que após ativo lhe dá acesso as séries históricas do INMET em três formatos distintos: dados horários, diários e mensais. Para a elaboração do trabalho foram selecionados os dados no formato diário, pelo fato de eles apresentarem um conjunto de dados mais completo quando comparados a opção mensal.

Na base de dados diária são disponibilizadas as seguintes variáveis climáticas que foram utilizadas no estudo:

- Precipitação pluviométrica: acumulada nas últimas 24 horas em escala de milímetros.
- Temperatura máxima: em escala de graus celsius.
- Temperatura mínima: em escala de graus celsius.
- Insolação: em escala de horas.
- Evaporação do Piche: em escala de milímetros.
- Temperatura compensada média: em escala de graus celsius.
- Umidade relativa média: em escala percentual.

O INMET disponibiliza esses dados para as suas estações meteorológicas localizadas no estado do Rio Grande do Sul presentes nas cidades de: Bage, Bento Gonsalves, Bom Jesus, Caxias do Sul, Cruz Alta, Encruzilhada do Sul, Irai, Lagoa Vermelha, Passo Fundo, Pelotas, Porto Alegre, Rio Grande, Santa Maria, Santa Vitoria do Palmar, Santana do Livramento, São Luiz Gonzaga, Torres e Uruguaiana, a localização destas cidades no mapa pode ser observada na figura 11. Os dados são disponibilizados para download no formato .csv (*Comma-separated values*), e para o estudo foram utilizados os dados de todas as estações compreendidos entre os anos de 1976 a 2019.

Figura 11 – Localização das estações meteorológicas no estado do Rio Grande do Sul



Fonte: elaborado pelo autor

6.1.3 Pré-processamento do conjunto e dados

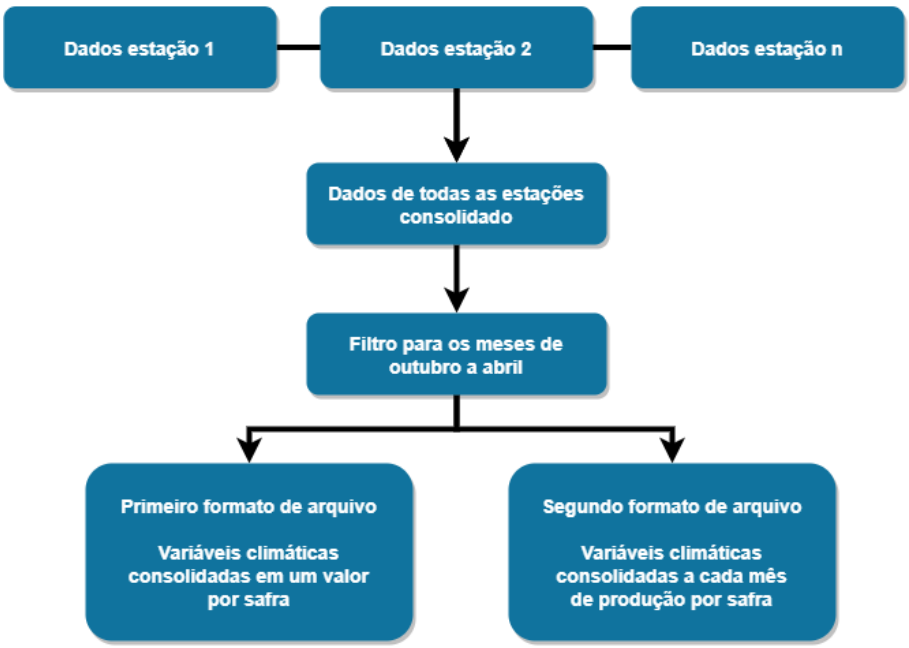
Para tornar estes dados utilizáveis para o aprendizado de máquina, houve a necessidade de realizar o pré-processamento deles, consolidando os dados das estações meteorológicas em um único arquivo e posteriormente o combinando com os dados da safra oriundos da CONAB.

Para o pré-processamento dos dados meteorológicos, foi desenvolvido um código em Python, onde é efetuada a leitura dos 18 arquivos .csv referente a cada uma das estações situadas no Rio Grande do Sul. Esses dados são então combinados em um único *dataframe pandas* (posteriormente será detalhado mais sobre este pacote Python), e realizado um filtro para buscar os valores que estão compreendidos entre os meses de outubro a abril, período em que ocorre a safra de soja no estado.

Então os dados de todas as estações, que se apresentam em formato diário são combinados de duas maneiras diferentes, cada uma delas gerando ao final um

arquivo .csv. Na figura 12 é exemplificado todo processo realizado, e as versões dos arquivos são detalhadas posteriormente.

Figura 12 – Esquema de pré processamento dos arquivos climáticos



Fonte: elaborado pelo autor

No primeiro formato todos os dados que compreendem o período de cada safra, por exemplo, na safra de 1977 os valores que estão entre os meses de outubro de 1976 a abril de 1977, são consolidados em um único valor por variável, através do cálculo da média dos valores diários, resultando ao final em um arquivo com as seguintes variáveis:

- ano (safra),
- precipitação,
- temperatura máxima,
- temperatura mínima,
- insolação,
- evaporação piche,
- temperatura compensada média
- umidade relativa média

Tendo para cada uma das variáveis, uma entrada por safra.

No segundo formato, os dados que compreendem o período de cada safra são consolidados a cada mês em uma variável, através do cálculo da média dos valores, gerando para cada período de safra uma entrada, com as seguintes variáveis, onde os números correspondem ao mês de referência dos valores:

- ano (safra),
- precipitação 1, precipitação 2, precipitação 3, precipitação 4, precipitação 10, precipitação 11, precipitação 12
- temperatura máxima 1, temperatura máxima 2, temperatura máxima 3, temperatura máxima 4, temperatura máxima 10, temperatura máxima 11, temperatura máxima 12,
- temperatura mínima 1, temperatura mínima 2, temperatura mínima 3, temperatura mínima 4, temperatura mínima 10, temperatura mínima 11, temperatura mínima 12,
- insolação 1, insolação 2, insolação 3, insolação 4, insolação 10, insolação 11, insolação 12,
- evaporação piche 1, evaporação piche 2, evaporação piche 3, evaporação piche 4, evaporação piche 10, evaporação piche 11, evaporação piche 12,
- temperatura compensada média 1, temperatura compensada média 2, temperatura compensada média 3, temperatura compensada média 4, temperatura compensada média 10, temperatura compensada média 11, temperatura compensada média 12,
- umidade relativa média 1, umidade relativa média 2, umidade relativa média 3, umidade relativa média 4, umidade relativa média 10, umidade relativa média 11, umidade relativa média 12,

Após gerados os dois arquivos pelo código em Python, com o auxílio do software Microsoft Excel, são adicionados em ambos os arquivos, as variáveis sobre a safra fornecidas pela CONAB, sendo elas:

- Produção
- Área plantada
- Produtividade

A figura 13 a seguir exibe uma visualização parcial do conjunto de dados de primeiro formato.

Figura 13 – Visualização parcial do conjunto de dados de primeiro formato

	A	B	C	D	E	F	G	H	I	J	K
1	Ano	Produção	Produtividade	Area	Precipitacao	TempMaxima	TempMinima	Insolacao	EvaporacaoPiche	TempCompMedia	UmidadeRelativaMedia
2	1977	5650	1618,91	3490	4,405683737	24,92396007	15,40008338	6,791318958	2,711894766	19,54490608	76,44473242
3	1978	4676	1245,6	3754	3,401598837	25,76689877	15,74929012	6,779052721	3,13432795	20,11073251	73,38081873
4	1979	3600	911,39	3950	3,741917227	24,77365347	14,83980019	6,97362299	3,264628975	19,43269039	72,50841602
5	1980	5581,8	1400	3987	3,927553648	25,11235161	15,26102578	6,918737504	3,20331004	20,28446141	73,62867237
6	1981	6139	1594,96	3849	4,682273342	25,10158858	15,4746126	6,512486457	2,96323378	19,82133489	75,08373494
7	1982	4251,5	1179,99	3603	2,949329318	25,57021858	15,38942006	7,048479035	3,43942782	19,76655336	72,73533289
8	1983	5200,5	1457,95	3567	5,655864198	24,63716677	15,47488545	6,132277061	2,93856799	19,34961252	76,94987941
9	1984	5404	1515	3567	5,244005902	24,96478102	15,54791199	5,928940472	3,066730548	19,22913684	76,91228911

Fonte: elaborado pelo autor

Já a figura 14 também exibe uma visualização parcial do conjunto de dados, porém do segundo formato. Nela é possível observar os dados referentes ao mês de janeiro, visível através do número que acompanha a variável.

Figura 14 - Visualização parcial do conjunto de dados de segundo formato

	A	B	C	D	E	F	G	H	I	J	K
1	Ano	Produção	Produtividade	Area	Precipitacao - 1	TempMaxima - 1	TempMinima - 1	Insolacao - 1	EvaporacaoPiche - 1	TempCompMedia - 1	UmidadeRelativaMedia - 1
2	1977	5650	1618,91	3490	5,631477927	25,93838951	17,66026365	5,938212928	2,604761905	21,1909542	79,43142857
3	1978	4676	1245,6	3754	4,98618677	27,13879781	16,94381818	7,070566038	3,227504554	21,38932358	72,93613139
4	1979	3600	911,39	3950	1,540431267	26,54515419	16,3536105	8,264304462	4,634355828	21,29839009	64,60339506
5	1980	5581,8	1400	3987	3,344972578	26,38049645	15,69307958	8,038025594	3,994562648	21,55267677	69,32294118
6	1981	6139	1594,96	3849	4,172400756	26,47847866	16,78815081	6,98147448	3,113953488	21,26312169	74,57714617
7	1982	4251,5	1179,99	3603	1,495660377	27,60897196	16,27363636	8,00510397	4,660352423	21,37807487	66,3388521
8	1983	5200,5	1457,95	3567	5,027560521	25,66561181	17,64031621	5,236312849	2,966044776	20,99444954	79,48227612
9	1984	5404	1515	3567	6,53659306	25,64723404	16,40962343	5,343729904	3,231290323	20,42972727	79,44677419

Fonte: elaborado pelo autor

Os dois modelos de conjunto de dados apresentaram algumas variáveis meteorológicas com o seu valor ausente, devido a uma deficiência nos valores oriundos dos arquivos de séries históricas do INMET, e necessitavam ter os seus valores calculados ou a linha deveria ser descartada, tendo em vista que redes LSTM não aceitam valores ausentes.

Existem diversas abordagens que podem ser implementadas nestes casos como listou PRATAMA *et al.* (2016) em seu trabalho, desde abordagens consideradas mais simples como o cálculo da média, mediana, moda e deleção dos dados, a algoritmos mais complexos e modernos como algoritmos genéticos, interpolação de dados, entre outros. A opção com melhor resultado depende muito das características do conjunto de dados que está sendo estudado, embora opções mais modernas como

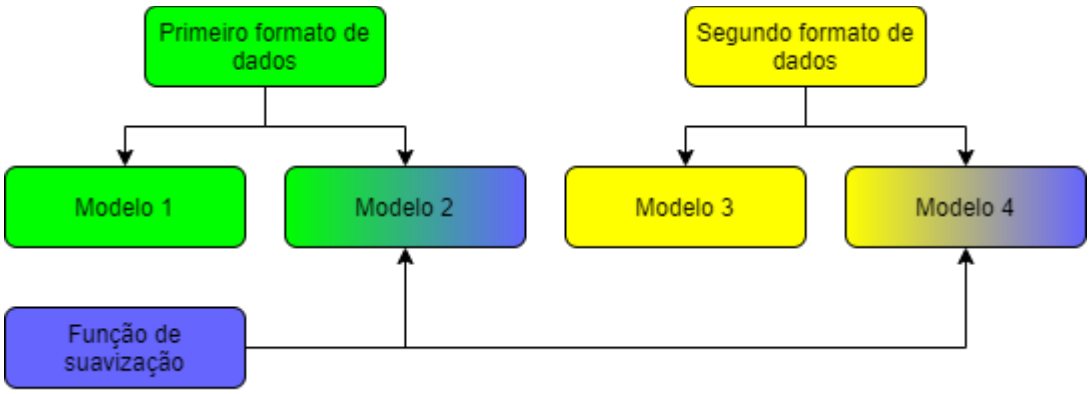
algoritmos genéticos tendem a apresentar melhores resultados, por outro lado, eles tendem a ter uma dificuldade de implantação maior, quando comparado a métodos mais simples.

Então no estudo ainda com o auxílio do software Excel, optou-se por calcular os valores faltantes com base na média das duas leituras anteriores e das duas leituras posteriores a ela, tendo em vista a implementação mais simplificada deste método, frente a outros métodos mais complexos, o que abre a possibilidade para um novo estudo para a avaliação de qual método traria melhores resultados para o conjunto de dados do estudo. Ao término deste processo os arquivos estão finalizados e prontos para serem utilizados nos modelos de previsão dos valores.

6.2 DESENVOLVIMENTO DOS MODELOS DE REDES NEURAIS

Os quatro modelos de rede LSTM desenvolvidos no estudo foram estruturados da seguinte maneira, cada um com suas características únicas: Modelo 1 - faz uso do primeiro formato de dados apresentado anteriormente. Modelo 2 - também faz uso do primeiro formato de dados, porém neste modelo os dados são suavizados com o algoritmo de suavização Lowess (*locally weighted scatterplot smoothing*), na busca de melhorar o resultado final do modelo, de maneira semelhante ao proposto no trabalho elaborado por HAIDER *et al*. (2019). Modelo 3 - neste modelo se faz o uso do segundo formato de dados proposto no capítulo anterior e o modelo 4 - que também faz uso do segundo formato de dados, porém aplicando o algoritmo de suavização de maneira idêntica ao modelo 2. Na figura 15 é possível visualizar o esquema descrito acima.

Figura 15 – Esquema de uso do formato de dados e função de suavização por modelo



Fonte: elaborado pelo autor

A elaboração destes modelos foi realizada utilizando-se a linguagem de programação Python, em conjunto com diversos pacotes que auxiliam na manipulação dos dados, realização de cálculos e abstração de funções complexas, dentre os quais é importante dar um destaque especial ao TensorFlow e ao Keras, pacotes dedicados a realização do aprendizado de máquina e que são utilizados em estudos semelhantes como no realizado por MUTHUSINGHE *et al*. (2019). No próximo capítulo esses pacotes são abordados mais abrangentemente.

6.2.1 Pacotes Python utilizados nos modelos

O primeiro pacote/biblioteca selecionado para ser usado no estudo foi o TensorFlow, ele é uma biblioteca de código aberto desenvolvido pela equipe do Google denominada Google Brain. É um sistema dedicado ao aprendizado de máquina capaz de operar desde sistemas em larga escala em *datacenters*, até a execução local em dispositivos moveis. Suporta uma variedade de aplicações tendo como foco o treinamento e inferência de redes neurais profundas. Ele é empregado em vários produtos que o Google oferece, sendo que no ano de 2016 era utilizado por mais de 150 times de desenvolvimento dentro da empresa (ABADI *et al*., 2016).

Com o pacote de aprendizado de máquina selecionado, buscou-se por alguma ferramenta que fornecesse uma abstração na modelagem das redes, tornando este processo mais intuitivo e fácil, para isso foi selecionado o pacote Keras. Ele é uma API (*Application Programming Interface*) de alto nível para o desenvolvimento de redes neurais, capaz de operar sobre diversas bibliotecas como TensorFlow (utilizada no estudo), CNTK e Theano. Foi desenvolvido com o intuito de permitir a experimentação e prototipagem rápida e fácil de redes neurais, tem como princípios orientadores, a facilidade de uso, modularidade, fácil extensibilidade e trabalho exclusivo com Python.

Em 2018 a ferramenta possuía mais de 250.000 usuários individuais, com uma forte adoção nas comunidades de pesquisa, além de ser usado por grandes empresas como Netflix, Uber, Yelp, Instacart, Zocdoc, Square (WHY USE KERAS, 2020).

O Keras necessita que os dados informados nele para a realização dos treinamentos e previsões de valores estejam em formatos específicos para funcionarem, para isto no trabalho foram utilizados os pacotes Pandas e Numpy. O Pandas é uma ferramenta de análise e manipulação de dados, desenvolvida em

código aberto. Ela facilita a manipulação dos dados de diversas maneiras como por exemplo, na leitura e gravação de arquivos de dados .csv, na remodelagem do conjunto de dados, no tratamento de dados ausentes, entre outras funcionalidades (PANDAS - PYTHON DATA ANALYSIS LIBRARY, 2020).

Já o Numpy é um pacote usado principalmente para a realização de cálculos em *arrays* multidimensionais. Facilita as operações de concatenação, adição e subtração de *arrays,* além de diversas outras funcionalidades voltadas a computação cientifica (NUMPY, 2020).

Outro pacote que se fez necessário no estudo foi o Scikit-learn, ele é um pacote que implementa uma ampla gama de algoritmos de aprendizado de máquina distribuídos pela licença BSD (PEDREGOSA *et al*., 2011). No estudo foi realizado o uso de seus algoritmos de métricas para a avaliação dos modelos gerados.

E por último, para poder gerar melhores visualizações dos dados o pacote Matplotlib foi selecionado. Ele é um pacote voltado para a geração de gráficos em 2D, desenvolvido por John D. Hunter, atualmente é suportado nos sistemas operacionais Windows, Mac OS X e nas principais distribuições Linux. Ele suporta os principais tipos de gráficos 2D e gráficos interativos incluindo gráficos XY, barras, torta e linhas que foi utilizado mais extensamente no estudo (HUNTER, 2007).

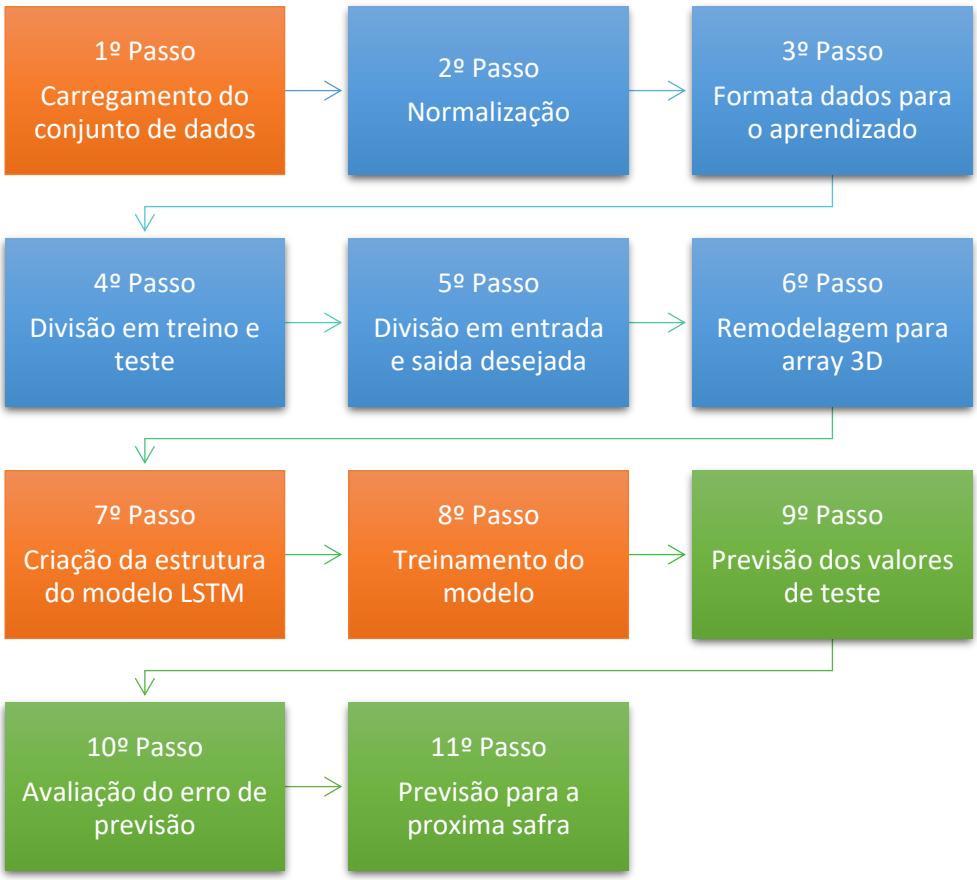
6.2.2 Estrutura dos modelos

Os quatro modelos desenvolvidos fazem o uso de redes neurais do tipo LSTM, devido a sua boa aplicabilidade com séries temporais e por suas demais características apresentadas no capítulo 4. Ambos os modelos foram desenvolvidos com a intenção de prever a safra do próximo ano com base nas informações de safra e clima do ano atual, então, por exemplo, para se prever os valores de produção da safra de 2020 são informados para o modelo as informações sobre, produtividade, área plantada, produção e condições climáticas, referentes a safra do ano de 2019.

Todos os modelos seguem a mesma estrutura básica, tendo como diferenças: o conjunto de dados importado, a aplicação ou não de uma função de suavização nos dados e os parâmetros utilizados nas funções. Na figura 16 são exibidos de maneira resumida os passos que foram necessários para a montagem das estruturas dos modelos, os quadros em laranja representam os passos aonde ocorre a alteração de parâmetros entre os modelos propostos, em azul são os passos onde ocorre o pré-

processamento dos dados e em verde os passos onde ocorre a avaliação do modelo gerado. Após a figura a estrutura do modelo é explicada detalhadamente.

Figura 16 – Resumo da estrutura dos modelos



Fonte: elaborado pelo autor

1º passo, carregamento do conjunto de dados: Os dados são importados dos arquivos .csv de origem e carregados para dentro de um conjunto de dados pandas, tendo os modelos 1 e 2 carregado o arquivo de primeiro formato e os modelos 3 e 4 o arquivo de segundo formato. Após os modelos número 2 e 4, são suavizados utilizando a função Lowess. Está função cria uma linha de dados mais suave, eliminando deles grandes picos ou vales que em alguns casos pode prejudicar os resultados dos algoritmos de aprendizado de máquina. A função utilizada no trabalho foi desenvolvida por Alexandre Gramfort e Dan Neuman e distribuída pelo GitHub com licença BSD, o seu uso ocorreu devido ela apresentar um funcionamento integrado com conjuntos de dados no formato do pacote pandas, o que auxiliou na sua

implementação no estudo (LOWESS SMOOTHING FUNCTION FOR PYTHON USING PANDAS AND NUMPY, 2020).

2º passo, normalização: os dados são então normalizados, com valores em uma escala de zero a um, com o objetivo de melhorar o aprendizado e a convergência da rede, já que em alguns casos entradas muito grandes podem impedir o aprendizado efetivo da rede.

3º passo, formata dados para o aprendizado supervisionado: Após os valores se encontrarem normatizados, são processados pela função *series_to_supervised*, ela tem como função formatar os dados para o aprendizado supervisionado. Como no trabalho optou-se por prever a safra do ano seguinte, com base nos valores de produtividade e climáticos do ano anterior, a função pega os valores por exemplo da safra de 1977 e os combina com os valores da safra de 1978, após são excluídas todas as variáveis da safra de 1978, exceto o valor referente a produção que é o valor que queremos prever posteriormente. Assim a linha do conjunto de dados ficará com todos valores referentes a safra de 1977 mais o valor de produção da safra do ano de 1978. Este processo é então repetido para todas as entradas do conjunto de dados.

4º passo, divisão entre treino e teste: O conjunto de dados é então dividido em dois conjuntos, um para treino e outro para teste. Em todos os modelos o conjunto de treino possui 38 registros, equivalente a 90,48% do total de registros, e o conjunto de testes 4 registros valor equivalente a 9,52% do total. Optou-se por estas quantidades devido à similaridade com a quantidade empregada por HAIDER *et al*. (2019) que em seu trabalho utilizou 8,62% dos dados para testes. Outros fatores que contribuíram para a escolha foi o conjunto de dados ter uma quantidade relativamente pequena de valores e as características necessárias para treinamento de redes LSTM com dados históricos.

5º passo, divisão entre entrada e saída desejada: Os conjuntos de dados de treino e de teste são divididos em entrada e saída desejada.

6º passo, remodelagem para *array* 3D: Os conjuntos de dados de entrada, são então remodelados para o formato de *array* 3D, onde o primeiro valor equivale a quantidade de amostras, o segundo as unidades de tempo e o terceiro a quantidade de variáveis de entrada. Este processo é necessário pois o treinamento de redes LSTM com o Keras, somente funcionará se os dados estiverem neste formato.

7º passo, criação da estrutura do modelo LSTM: É montada a estrutura do modelo LSTM, neste passo são definidas as camadas ocultas dos modelos, quantidade de neurônios, função de ativação e otimizador de aprendizado.

8º passo, treinamento do modelo: Neste passo é realizado o treinamento efetivo do modelo, informando-se a quantidade de épocas de treinamento e o conjunto de dados de treino e teste, após o término do treinamento é impresso um gráfico, onde é realizada a comparação do valor de perda entre o conjunto de testes e de treino, afim de poder-se avaliar se o modelo possui um bom desempenho ou não, ou se modificações nele serão necessárias.

9º passo, previsão dos valores de teste: É realizado a previsão dos valores de teste com o modelo gerado no passo anterior. Após os valores passam pelo processo inverso da normatização, deixando-os novamente no formato original.

10º passo, avaliação dos erros de previsão: Com os valores em formato original são aplicados os cálculos de avaliação de erro RMSE, MAE e APE, conforme explicitado no capítulo 3, para validar a qualidade do modelo gerado. Também é impresso um gráfico de linhas comparando os valores originais com os valores gerados pelos modelos.

11º passo, previsão para próxima safra: É realizada a previsão da safra do ano de 2020, com base nos valores da safra de 2019, após a previsão os valores sofrem o processo inverso da normalização e é exibida a previsão no formato original dos dados.

Com a estrutura dos modelos prontas, ambos modelos foram experimentados através de testes empíricos cada um com as suas características únicas e com a realização de alterações nos parâmetros, que acontecem nos passos 1, 7 e 8 descritos acima. O próximo capítulo irá exibir os resultados destes experimentos e avaliar os resultados encontrados no experimento de cada um dos modelos.

7 RESULTADOS OBTIDOS

Com a conclusão da elaboração da estrutura dos modelos, iniciou-se testes empíricos com a inserção de diversas combinações de parâmetros nas funções, a fim de se obter um melhor resultado para a rede. Os parâmetros alterados entre os testes realizados incluem, número de camadas ocultas, número de unidades ocultas, função de ativação, otimizador, taxa de aprendizado do otimizador, quantidade de épocas de treinamento, e nos modelos 2 e 4 em que se faz uso da função de suavização, também foram feitos testes alterando a fração dos dados que são usados na suavização e no número de interações. A tabela 5 exibe os intervalos inferiores, superiores e algoritmos utilizados durante os testes.

Tabela 5 – Intervalo de valores utilizados nos testes empíricos

PARÂMETRO	INTERVALO
CAMADAS OCULTAS	1 a 5
NEURÔNIOS EM CADA CAMADA	5 a 2800
FUNÇÔES DE ATIVAÇÃO	ELU, SOFTMAX, SELU, SOFTPLUS, SOFTSING, TANH, SIGMOID, HARD SIGMOID, LINEAR e RELU
OTIMIZADORES	Adam e SGD
TAXA DE APRENDIZADO OTIMIZADOR	0,1 a 0,00001
ÉPOCAS DE TREINAMENTO	20 a 3000
FRAÇÃO DE DADOS SUAVIZAÇÃO	0,1 a 0,3
NÚMERO DE INTERAÇÕES	2 a 8

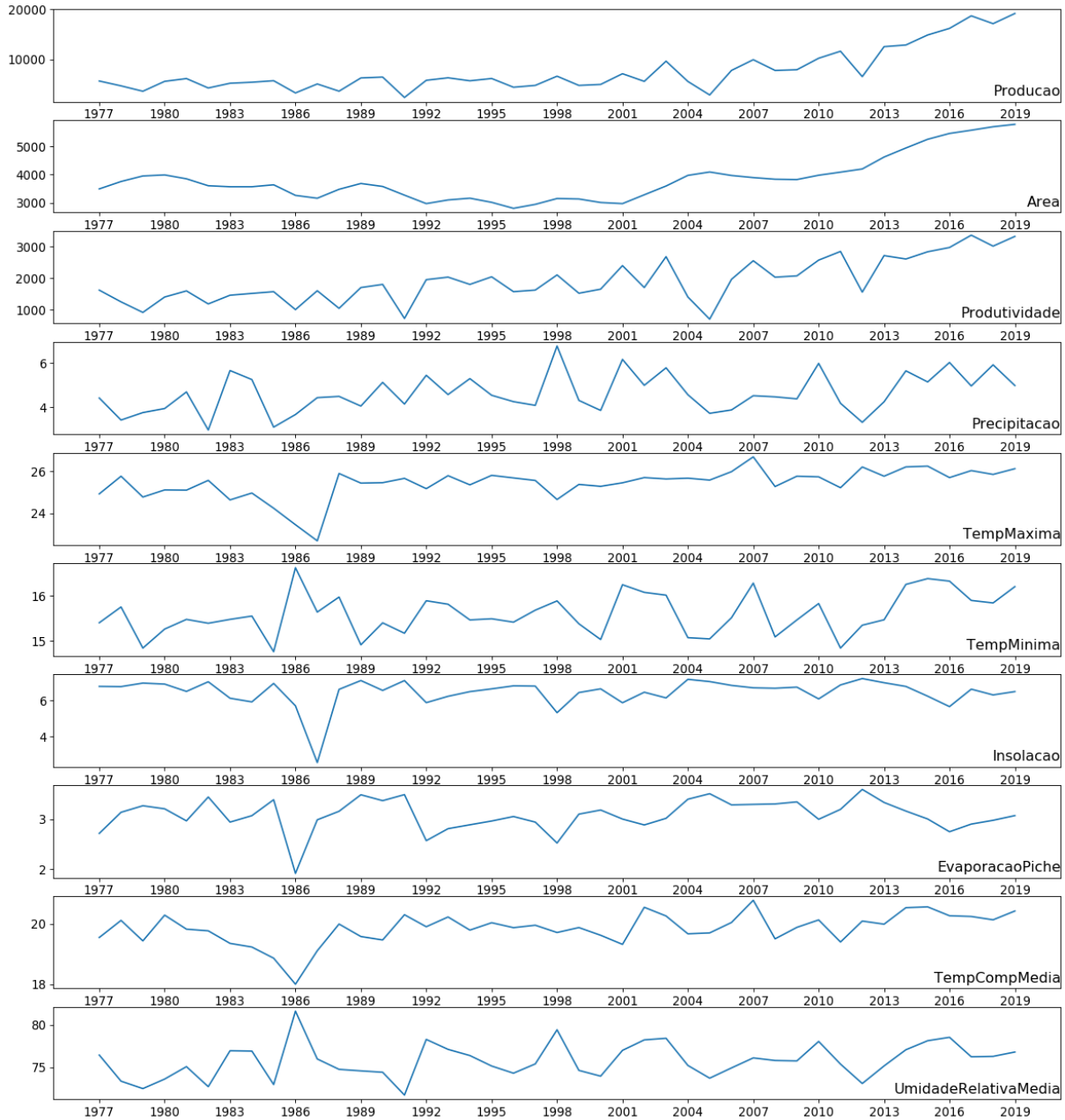
Fonte: elaborado pelo autor

O melhor resultado encontrado para cada um dos modelos nos testes realizados é exibido nos próximos capítulos.

7.1 RESULTADOS DO EXPERIMENTO COM O MODELO 1

Na elaboração do experimento com o modelo 1, foi realizado o uso do conjunto de dados de primeiro formato, que apresenta um menor conjunto de variáveis, conforme especificado no capítulo 6. No gráfico 6, é possível visualizar os dados importados onde cada ano representa uma safra.

Gráfico 6 – Conjunto de dados importados no modelo 1



Fonte: elaborado pelo autor

Ao fazermos uma análise do seu comportamento ao longo do tempo, é possível observar que nas variáveis referentes a safra, há uma tendência de aumento dos valores ao longo do tempo, com exceção há algumas quedas que ocorrem na produção e na produtividade, como ocorre por exemplo na safra de 2005, onde segundo dados da Fundação de Economia e Estatística do Rio Grande do Sul (FEE RS), houve uma grande estiagem em todo o estado, causando muitas perdas nas safras de grãos (FÜRSTENAU, 2005).

O modelo então segue a estrutura descrita no capítulo anterior, onde no sétimo passo é realizada a configuração da estrutura do modelo LSTM. Para este modelo encontrou-se como melhores parâmetros após a realização de diversos testes a utilização de duas camadas ocultas, tendo cada uma 10 neurônios, e a utilização da função de ativação RELU (*Rectified Linear Unit)*, uma variante das funções de ativação elencadas no capítulo 4*.* Na estrutura também é informado o algoritmo de otimização, na qual o algoritmo Adam com uma taxa de aprendizado de 0.00023 apresentou um melhor desempenho entre as opções testadas para o conjunto de dados proposto.

No oitavo passo é informada a quantidade de épocas de treinamento, neste modelo optou-se pela quantidade de 300 épocas. Visto que elas se mostraram suficientes para chegarmos em um melhor aprendizado sem haver excesso de treino.

A tabela 6 exibe um resumo dos parâmetros utilizados nos passos sete e oito do algoritmo.

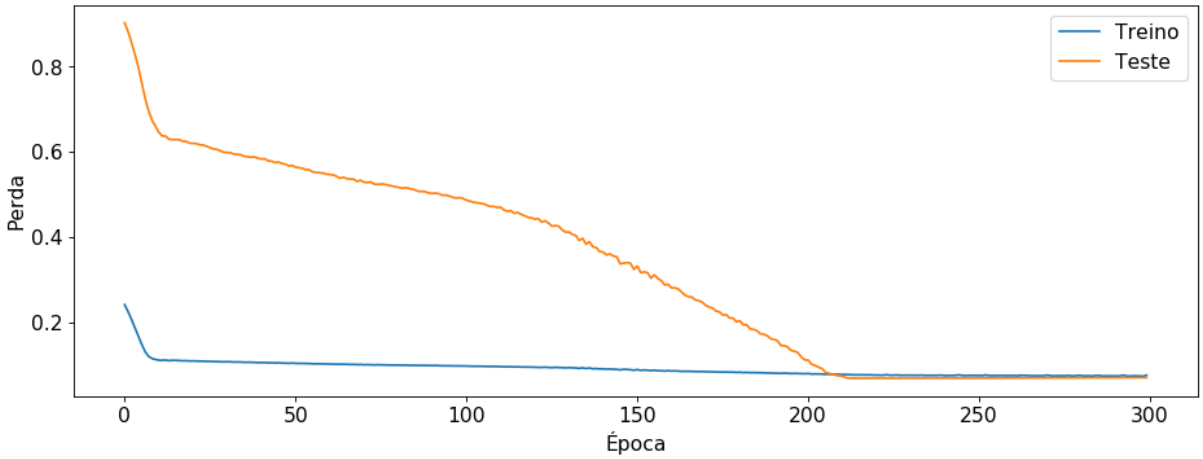
Tabela 6 – Resumo dos parâmetros utilizados no modelo 1

PARÂMETRO	VALOR
CAMADAS OCULTAS	2
NEURÔNIOS EM CADA CAMADA	10
FUNÇÃO DE ATIVAÇÃO	RELU
OTIMIZADOR	Adam
TAXA DE APRENDIZADO OTIMIZADOR	0,00023
ÉPOCAS TREINAMENTO	300

Fonte: elaborado pelo autor

O gráfico 7, exibe o histórico de treinamento da rede, onde a linha na cor azul representa os valores de treinamento da rede e a linha na cor laranja representa os valores de teste, quanto mais próximas as duas linhas ficarem, melhor tende a ser desempenho de previsão da rede. No gráfico é possível verificar que essa maior aproximação entre as linhas ocorre após cerca 210 épocas de treinamento.

Gráfico 7 – Valores de treinamento da rede modelo 1



Fonte: elaborado pelo autor

Com o término do treinamento do modelo, os valores dedicados para teste que correspondem as safras dos anos de 2016 a 2019, são então previstos conforme apresentado no nono passo, os resultados são expostos na tabela abaixo.

Tabela 7 – Valores projetados no teste do modelo 1

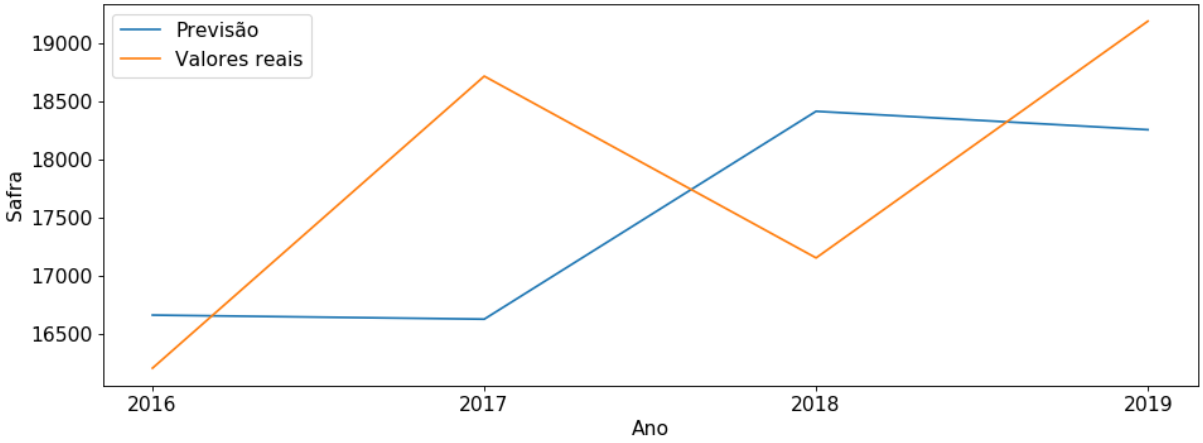
ANO	PREVISÂO	REAIS	DIFERENÇA	APE
2016	16658,1	16201,4	456,7	2,81%
2017	16623,2	18713,9	-2090,6	11,17%
2018	18411,7	17150,3	1261,4	7,35%
2019	18253,2	19187,1	-933,8	4,86%

Fonte: elaborado pelo autor

Na tabela 7 é possível verificar que o melhor resultado apresentado na projeção foi o do ano de 2016 que apresentou uma diferença positiva de 456,7 mil toneladas ou 2,81% a mais que o valor real, e o pior resultado foi o ano subsequente que

apresentou um valor de 2090,6 mil toneladas menor que o valor real ou 11,17% a menos que o valor real. No gráfico 8 abaixo é possível observar visualmente a diferença entre os valores reais em laranja e a previsão gerada em azul.

Gráfico 8 – Comparação entre previsão e valores reais do modelo 1



Fonte: elaborado pelo autor

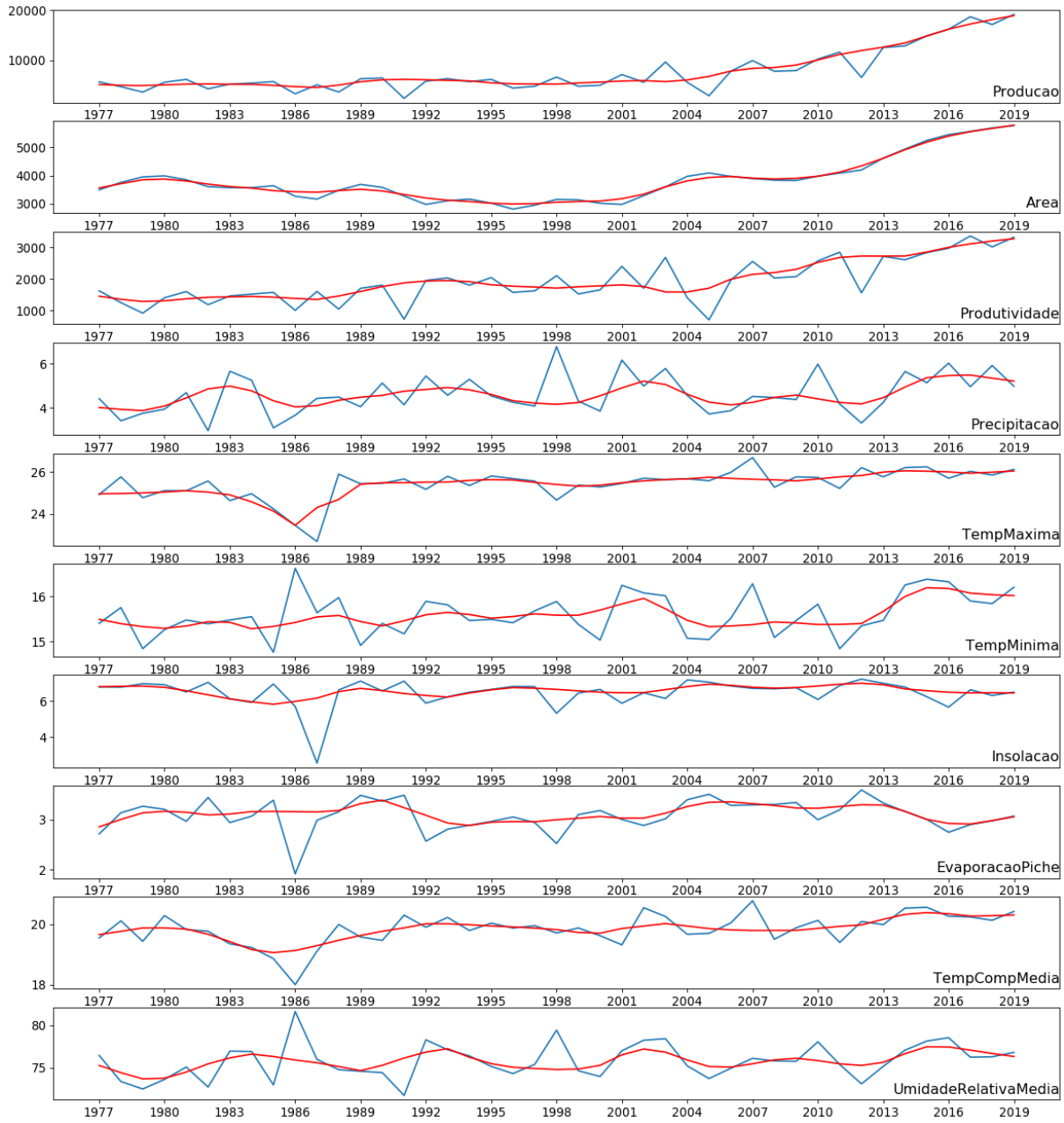
Além da comparação direta entre cada um dos valores reais e previstos foram aplicados os cálculos de mensuração de erro, RMSE que apresentou o valor de 1326,9 mil toneladas, e o MAE que apresentou o valor de 1185,6 mil toneladas.

Com o término das avaliações foi realizado o décimo primeiro passo do modelo onde é gerada a previsão para a safra do ano de 2020, a qual apresentou o valor previsto foi de 20932,0 mil toneladas de soja produzidas durante a safra.

7.2 RESULTADOS DO EXPERIMENTO COM O MODELO 2

Para o experimento com o modelo 2, o conjunto de dados de primeiro formato, com uma menor quantidade de variáveis, foi utilizado, da mesma maneira que o modelo 1, porém neste modelo os dados passaram pela função de suavização *Lowess* conforme especificado no capítulo 7. Como parâmetros para a função *Lowess*, foram informados os valores de 0,1 para a fração de dados usados na suavização, e 6 para a quantidade de interações. No gráfico 9 é possível ver uma comparação dos dados originais em azul e os dados suavizados pela função em vermelho. Pode-se observar que o algoritmo acaba eliminando do conjunto de dados locais onde há grandes vales e picos nos dados.

Gráfico 9 – Conjunto de dados importados no modelo 2



Fonte: elaborado pelo autor

O modelo então segue a estrutura descrita no capítulo anterior, onde no sétimo passo é realizada a configuração da estrutura do modelo LSTM. O modelo utiliza os mesmos valores empregados no modelo 1, sendo eles, a utilização de duas camadas ocultas, tendo cada uma 10 neurônios, a utilização da função de ativação RELU, e o algoritmo de otimização Adam com a taxa de aprendizado de 0,00023. Já neste modelo foram necessárias apenas 200 épocas de aprendizado para que houvesse

uma melhor conversão possível dos valores. Na tabela 8 um resumo dos parâmetros informados é exibido.

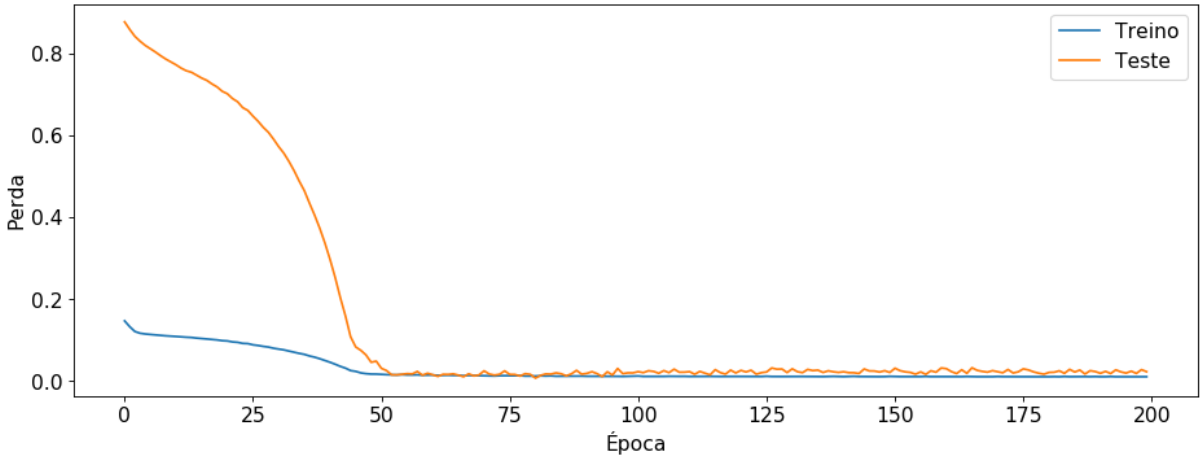
Tabela 8 – Resumo dos parâmetros usados no modelo 2

PARÂMETRO	VALOR
CAMADAS OCULTAS	2
NEURÔNIOS EM CADA CAMADA	10
FUNÇÃO DE ATIVAÇÃO	RELU
OTIMIZADOR	Adam
TAXA DE APRENDIZADO OTIMIZADOR	0,00023
ÉPOCAS TREINAMENTO	200

Fonte: elaborado pelo autor

O gráfico 10, exibe o histórico de treinamento da rede, onde a linha na cor azul representa os valores de treinamento da rede e a linha na cor laranja representa os valores de teste, quanto mais próximas as duas linhas ficarem, melhor tende a ser o desempenho de previsão da rede. No gráfico é possível verificar que para o modelo 2 a aproximação entre as linhas ocorre após 50 épocas de treinamento e ele apresenta um nível de ruído um pouco maior quando ao comparado ao modelo 1.

Gráfico 10 - Valores de treinamento da rede modelo 2



Fonte: elaborado pelo autor

Na tabela 9 abaixo são exibidas as previsões realizadas com os valores dedicados aos testes que compreendem os anos de 2016 a 2019. Nela é possível observar que os anos de 2016 e 2018 apresentam a menor diferença entre os valores da previsão e reais, com 31,4 mil toneladas a menos e 427,6 mil toneladas a mais, o que dá um erro percentual de 0.19% para o ano de 2016 e de 2,49% para 2018. Na contramão o ano de 2017 apresenta a maior diferença, com uma previsão de 1728,9 mil toneladas a menos que o valor real, o que equivale a uma diferença de um pouco mais de 9%.

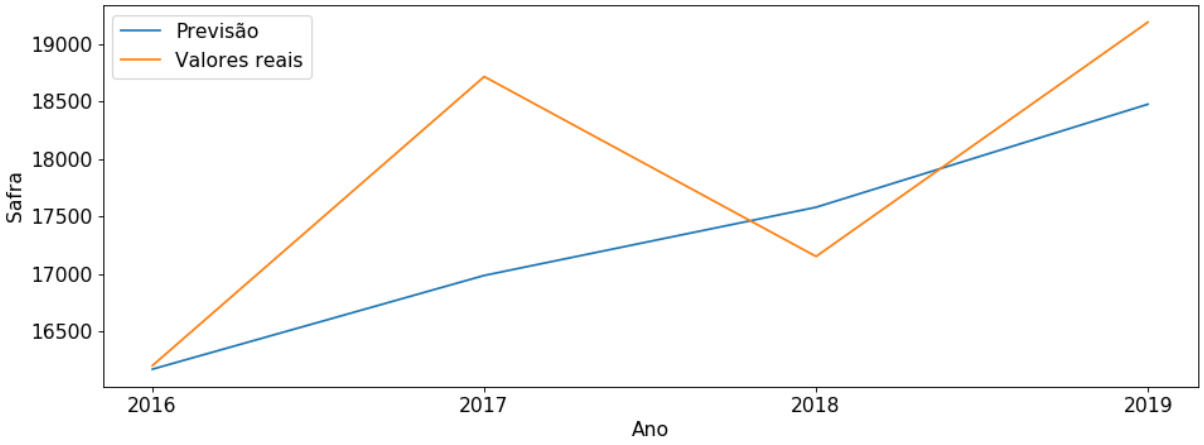
Tabela 9 - Valores projetados no teste do modelo 2

ANO	PREVISÂO	REAIS	DIFERENÇA	APE
2016	16169,9	16201,4	-31,4	0,19%
2017	16984,9	18713,9	-1728,9	9,23%
2018	17577,9	17150,3	427,6	2,49%
2019	18473,5	19187,1	-713,5	3,71%

Fonte: elaborado pelo autor

No gráfico 11 é possível observar visualmente a diferença entre os valores reais em laranja e a previsão gerada em azul.

Gráfico 11 - Comparação entre previsão e valores reais do modelo 2



Fonte: elaborado pelo autor

No modelo 2, os cálculos de mensuração de erro apresentaram o valor de 959,4 mil toneladas, para o RMSE, e de 725,3 mil toneladas para o MAE, valores menores dos que os apresentados pelo modelo 1.

Com o término das avaliações foi realizado o décimo primeiro passo do modelo onde é gerada a previsão para a safra do ano de 2020, a qual apresentou o valor previsto de 19321,4 mil toneladas de soja produzida durante a safra.

7.3 RESULTADOS DO EXPERIMENTO COM O MODELO 3

Para o modelo 3, o experimento contou com a utilização do conjunto de dados de segundo formato, este conjunto de dados é mais complexo e possui 52 variáveis de entrada, onde cada variável climática, tem o seu valor retratado por mês, conforme explicitado no capítulo 6.

O modelo então segue a estrutura descrita no capítulo anterior, onde no sétimo passo é realizada a configuração da estrutura do modelo LSTM. Para este modelo encontrou-se como melhores parâmetros a utilização de três camadas ocultas, tendo cada uma delas 100 neurônios, e a utilização da função de ativação ELU (*Exponential Linear Unit*), uma variante das funções de ativação elencadas no capítulo 4. Na estrutura também é informado o algoritmo de otimização, na qual o algoritmo Adam com uma taxa de aprendizado de 000023 apresentou um melhor desempenho entre as opções testadas para o conjunto de dados proposto. Além destes parâmetros, no oitavo passo foi informada a quantidade de 200 épocas de treinamento para o modelo, na tabela 10 é possível visualizar um resumo dos parâmetros informados.

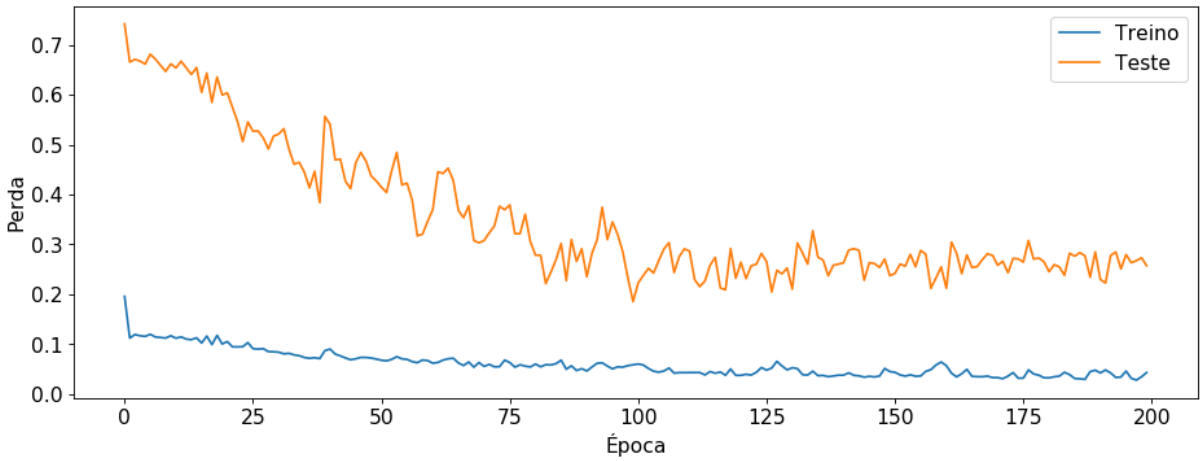
Tabela 10 - Resumo dos parâmetros utilizados no modelo 3

PARÂMETRO	VALOR
CAMADAS OCULTAS	3
NEURÔNIOS EM CADA CAMADA	100
FUNÇÃO DE ATIVAÇÃO	ELU
OTIMIZADOR	Adam
TAXA DE APRENDIZADO OTIMIZADOR	0,00023
ÉPOCAS TREINAMENTO	200

Fonte: elaborado pelo autor

No gráfico abaixo é exibido o histórico de treinamento do modelo, nele é possível verificar que não há uma conversão muito alta dos dados com o modelo, vide a distância e a tendência que as linhas laranja (teste) e azul (treinamento) apresentam não virem a convergir.

Gráfico 12 - Valores de treinamento da rede do modelo 3



Fonte: elaborado pelo autor

Após o término do treinamento, o modelo foi testado com os valores do conjunto de dados de teste, com dados dos anos de 2016 a 2019. Os resultados são apresentados na tabela abaixo, nele podemos observar que o modelo apresentou uma grande taxa de erro, principalmente se comparado aos dois modelos anteriores. O valor previsto mais próximo ao valor real foi o do ano de 2018, que ficou 1978,9 mil toneladas a menos que o valor real, o que equivale a uma diferença de 11,53%. No pior caso a diferença chegou a um percentual de 38,17%, no ano de 2017 equivalente a 7144,7 mil toneladas a menos que o valor real.

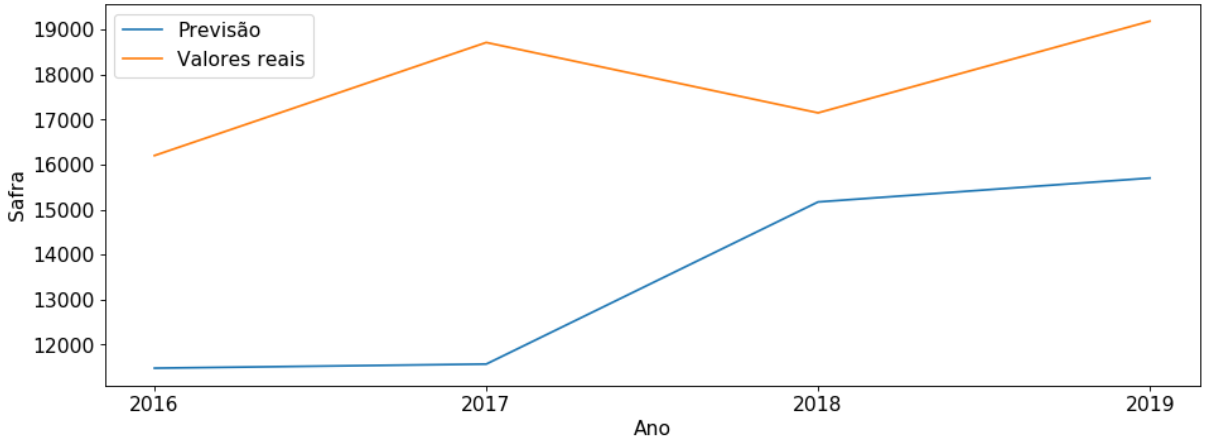
Tabela 11 - Valores projetados no teste do modelo 3

ANO	PREVISÂO	REAIS	DIFERENÇA	APE
2016	11478,1	16201,4	-4723,2	29,15%
2017	11569,1	18713,9	-7144,7	38,17%
2018	15171,3	17150,3	-1978,9	11,53%
2019	15701,6	19187,1	-3485,4	18,16%

Fonte: elaborado pelo autor

No gráfico 13 é possível observar visualmente a diferença entre os valores reais em laranja e a previsão gerada em azul.

Gráfico 13 - Comparação entre previsão e valores reais do modelo 3



Fonte: elaborado pelo autor

Os cálculos de mensuração de erro RMSE apresentou o valor de 4728,1 mil toneladas, enquanto o MAE apresentou o valor de 4333,1 mil toneladas, valores que evidenciam ainda mais a baixa performance do modelo, frente aos modelos listados anteriormente.

Com o término das avaliações foi realizado o décimo primeiro passo do modelo onde é gerada a previsão para a safra do ano de 2020, a qual apresentou o valor previsto de 16615,4 mil toneladas de soja produzida durante a safra.

7.4 RESULTADOS DO EXPERIMENTO COM O MODELO 4

Para a o modelo 4, foi realizado no experimento a utilização do conjunto de dados de segundo formato do mesmo modo que o modelo 3, porém, neste modelo os dados também passaram pela função de suavização *Lowess,* de maneira idêntica a aplicada no modelo 2. Como parâmetros para a função, foram informados os valores de 0,1 para a fração de dados usados na suavização e 6 para a quantidade de interações.

O modelo então segue a estrutura descrita no capítulo anterior, onde no sétimo passo é realizada a configuração da estrutura do modelo LSTM. Para este modelo

encontrou-se como melhores parâmetros a utilização de duas camadas ocultas, tendo cada uma delas 30 neurônios, e a utilização da função de ativação ELU. Na estrutura também é informada o algoritmo de optimização, na qual o algoritmo Adam com uma taxa de aprendizado de 0,00023 apresentou um melhor desempenho entre as opções testadas para o conjunto de dados proposto. No treinamento da rede foram necessárias 300 épocas de treinamento. Na tabela 12 abaixo temos um resumo dos valores informados na elaboração da rede.

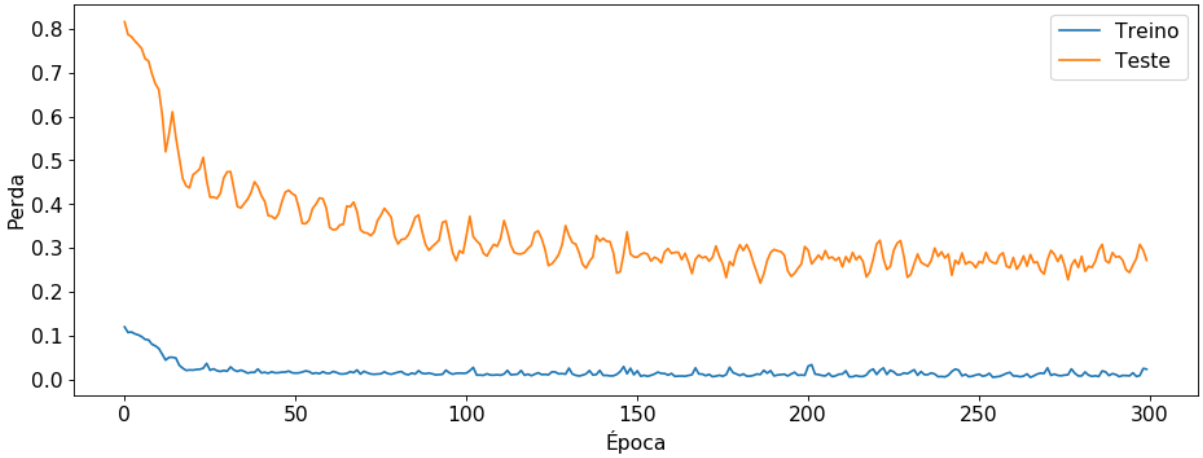
Tabela 12 - Resumo dos parâmetros utilizados no modelo 4

PARÂMETRO	VALOR
CAMADAS OCULTAS	2
NEURÔNIOS EM CADA CAMADA	30
FUNÇÃO DE ATIVAÇÃO	ELU
OTIMIZADOR	Adam
TAXA DE APRENDIZADO OTIMIZADOR	0,00023
ÉPOCAS TREINAMENTO	300

Fonte: elaborado pelo autor

O gráfico 14, exibe o histórico de treinamento da rede, onde a linha na cor azul representa os valores de treinamento da rede e a linha na cor laranja representa os valores de teste, nele é possível observar que não há uma conversão muito alta dos dados com o modelo, vide a distância entre as linha e o fato de elas não apresentarem tendência de convergir.

Gráfico 14 - Valores de treinamento da rede do modelo 4



Fonte: elaborado pelo autor

Na tabela 13 são exibidas as previsões realizadas com os valores dedicados aos testes que compreendem os anos de 2016 a 2019. Nela é possível observar que os valores previstos, apresentam números muito inferiores aos reais, sendo o pior resultado o do ano de 2019, no qual a previsão subestimou em 5792,6 mil toneladas a produção de soja do ano, o que equivale a uma diferença de 30,19%, por outro lado o melhor resultado apresentado pelo modelo foi o do ano de 2016 onde a diferença ficou em 9,91% o que equivale a 1607,1 mil toneladas a menos que o valor real.

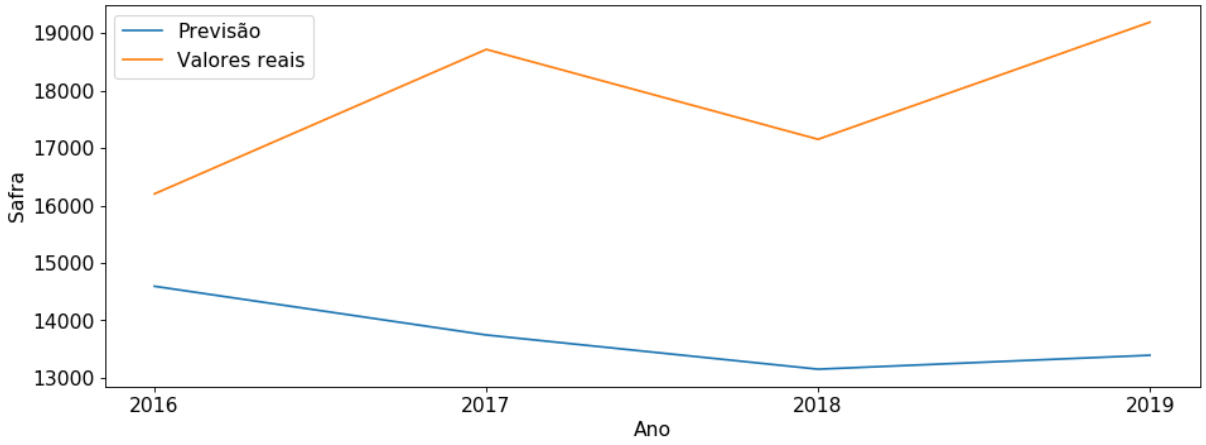
Tabela 13 - Valores projetados no teste do modelo 4

ANO	PREVISÂO	REAIS	DIFERENÇA	APE
2016	14594,2	16201,4	-1607,1	9,91%
2017	13747,1	18713,9	-4966,7	26,54%
2018	13152,4	17150,3	-3997,8	23,31%
2019	13394,4	19187,1	-5792,6	30,19%

Fonte: elaborado pelo autor

No gráfico 15 é exibida a diferença entre os valores reais em laranja e a previsão gerada em azul, nele é possível verificar a grande distância entre as linhas principalmente no ano de 2019.

Gráfico 15 - Comparação entre previsão e valores reais do modelo 4



Fonte: elaborado pelo autor

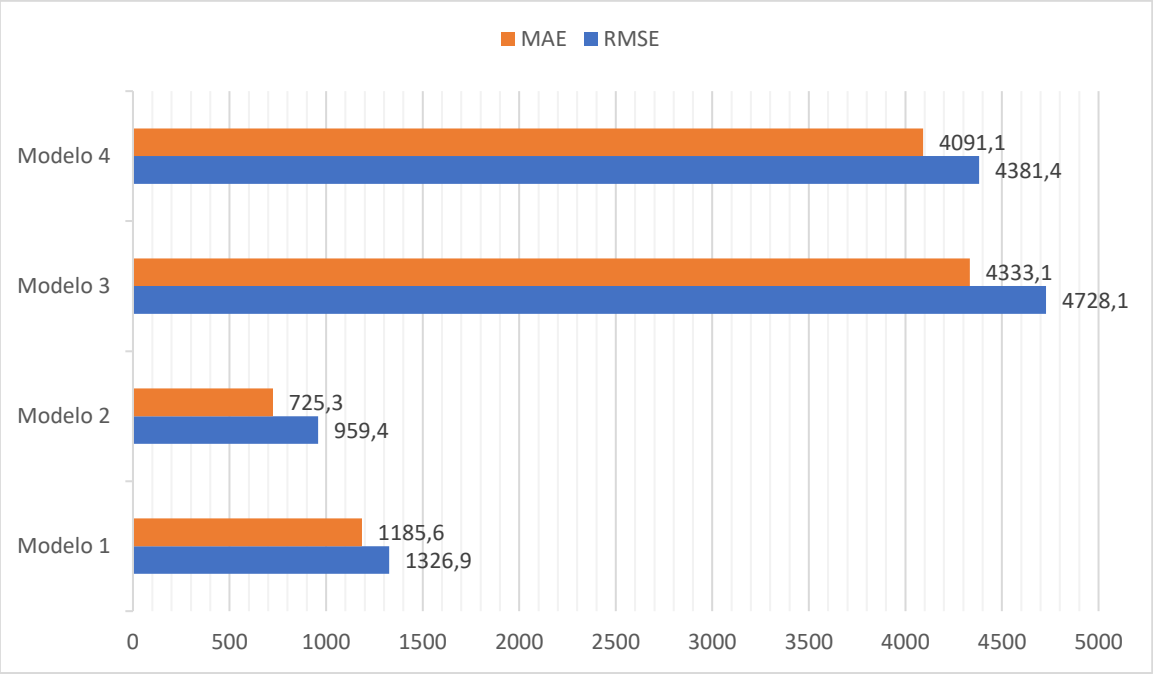
No modelo 4, os cálculos de mensuração de erro apresentaram o valor de 4381,4 mil toneladas, para o RMSE, e de 4091,1 mil toneladas para o MAE, valores ligeiramente menores do que os apresentados no modelo 3, mas ainda assim muito piores do que os dois primeiros modelos.

Com o término das avaliações foi realizado o décimo primeiro passo do modelo onde é gerada a previsão para a safra do ano de 2020, a qual apresentou o valor previsto de 13685,8 mil toneladas de soja produzida durante a safra.

7.5 ANÁLISE DOS RESULTADOS

Os quatro experimentos desenvolvidos no estudo apresentaram resultados bastante distintos evidenciando que diferenças na formatação do conjunto de dados e a aplicação de funções para tratamento dos dados brutos, podem trazer diferentes resultados em redes de mesma estrutura básica. No gráfico 16 abaixo, podemos verificar a diferença entre os valores de erro RMSE em azul e MAE em laranja em cada modelo. É possível observar que os modelos 1 e 2, que fazem uso do primeiro conjunto de dados mais enxuto, com um total de 11 variáveis, teve um desempenho superior, quando comparados aos modelos 3 e 4 que utilizam o segundo conjunto de dados com 52 variáveis. Também é possível verificar que o modelo 2 com dados suavizados apresentou uma taxa de erro menor, frente ao modelo 1 que utilizou os dados originais.

Gráfico 16 – Comparação entre os valores de erro dos modelos



Fonte: elaborado pelo autor

A tabela 14 abaixo exibe um comparativo entre os valores encontrados nos experimentos de modelo 1 e 2, que apresentaram os melhores resultados no estudo, com os valores encontrados por HAIDER *et al*. (2019), no seu trabalho realizado na previsão da safra de trigo no Paquistão.

Tabela 14 – Comparação entre valores de erro do estudo com o trabalho de Haider *et al*

Estudo	Variante	RMSE	MAE
Autor	Dados Brutos – (Modelo 1)	1326,9	1185,6
	Dados Suavizados – (Modelo 2)	959,4	725,3
Haider *et al*	Dados Brutos	1002	808
	Dados Suavizados	792	729

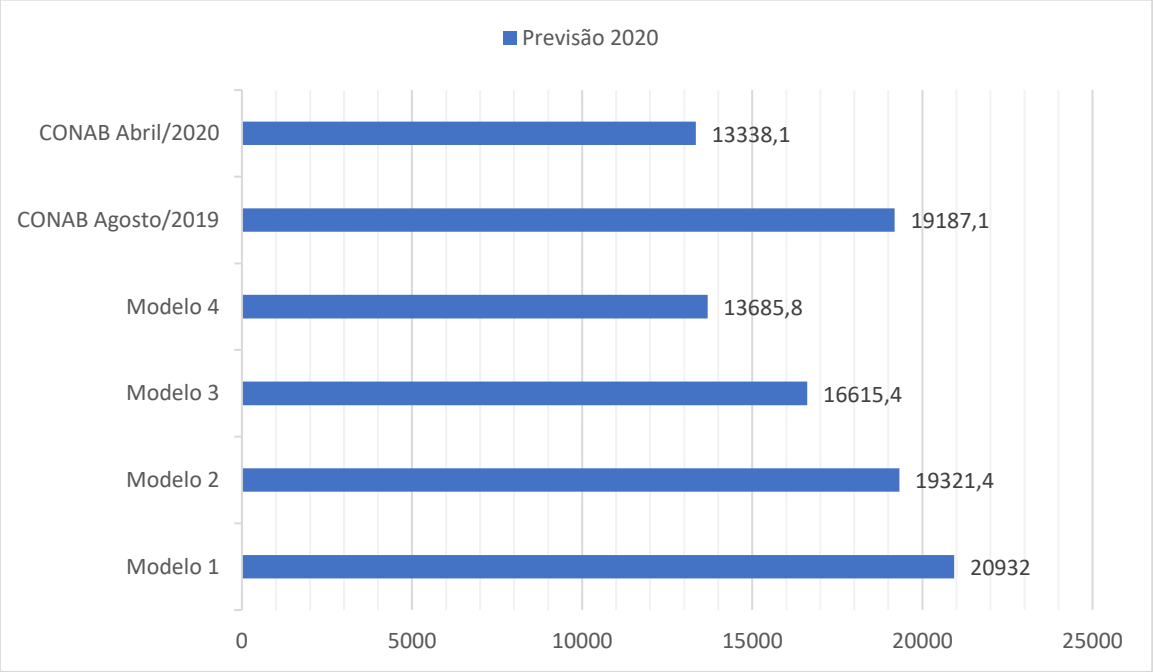
Fonte: elaborado pelo autor

Podemos verificar que os valores encontrados em ambos os estudos são relativamente próximos e que aplicação da função de suavização Lowess em ambos os estudos trouxe benefícios na diminuição dos indicadores de erro.

Quanto aos dados gerados referentes a estimativa de produção da safra para o ano de 2020, no gráfico 17 podemos ver uma comparação entre os 4 experimentos

gerados no estudo, com os valores estimados pela CONAB, no mês de agosto de 2019 e do mês de abril de 2020 (última disponível na data em que este texto está sendo escrito).

Gráfico 17 - Comparação entre estimativas de safra para o ano de 2020



Fonte: elaborado pelo autor

As estimativas geradas pela CONAB, encontram-se localizada junto com o arquivo de séries históricas disponibilizadas em seu site e que foi utilizado como fonte de dados do estudo conforme abordado em capítulos anteriores. Este arquivo é atualizado periodicamente, sempre com a última previsão realizada pela CONAB para o ano corrente.

É possível verificar no gráfico que a previsão gerada pela CONAB no mês de agosto de 2019, é bem próxima a previsão gerada pelo modelo 2, que nos testes de avaliação de erro foi o que apresentou menores resultados. A diferença entre as duas previsões é de 134,3 mil toneladas ou uma diferença de 0,69%, isso demostra que o modelo 2 possui uma eficiência similar com o modelo atualmente adotado oficialmente pela CONAB, visto que ambos além de encontrarem valores próximos, foram gerados com dados de antes do início da safra. No caso específico do modelo 2 elaborado no estudo, com dados referentes a safra de 2019.

Porém, podemos perceber que na última previsão gerada pela CONAB no mês de abril de 2020, já com informações referentes a safra de 2020, os valores caíram para 13338,1 mil toneladas, isto se explica pela forte estiagem que ocorre no estado do Rio Grande do Sul no ano de 2020, que é a pior dos últimos 8 anos e já se encontra na lista das piores estiagens já verificadas no estado, o que acabou gerando uma grande quebra na safra, conforme informações emitidas pela própria CONAB (CONAB, 2020).

Com base nestas informações podemos apontar que os experimentos desenvolvidos no estudo, em especial os dos modelos 1 e 2, que apresentaram melhores resultados nos testes, são capazes de prever a safra de soja no estado do Rio Grande do Sul, com uma assertividade razoável em cenários de “normalidade” climática. Porém em casos atípicos, onde possa haver chuvas ou secas em excesso ou em padrões pouco comuns e que se repetem raramente, os modelos não são capazes de prever essa flutuação com um ano de antecedência.

7.6 TRABALHOS FUTUROS

Com base em uma revisão da literatura é possível observar que existe uma grande quantidade de trabalhos que utilizam aprendizado de máquina, e que o relacionam diversas metodologias, áreas e abordagens diferentes, no entanto ainda há uma baixa quantidade de trabalhos relacionando o aprendizado de máquina com a agricultura no Brasil, com isso ficam como sugestão de trabalhos futuros: avaliar a aplicação de diferentes técnicas para preenchimento de valores faltantes no conjunto de dados climáticos, testes com outras variáveis climáticas que possam trazer uma relação mais explicita com a probabilidade de haver chuvas ou secas em excesso no estado do Rio Grande do Sul, como por exemplo a ocorrência ou não de *El Niño* ou *La Niña*, e a exploração de diferentes combinações de tempo para previsão, como, por exemplo, usar os dados climáticos referentes aos três primeiros meses de safra, para fazer a previsão da safra final.

8 CONCLUSÃO

A soja é uma das principais culturas do estado do Rio Grande do Sul, e durante o levantamento de dados sobre o cultivo foi possível identificar como ela se destaca no cenário internacional, sendo uma das suas principais comodities agrícolas, além de sua extrema importância para a economia do Brasil como um todo, e em especial no Rio Grande do Sul, onde sozinha corresponde por mais de 26% das exportações. Foi possível identificar também a forte relação que existe entre o crescimento e desenvolvimento do grão com as questões relacionadas ao clima, onde a temperatura ambiente, precipitações, nível de radiação solar podem impactar no desenvolvimento da planta e conseguintemente na produção final de grãos.

Durante o levantamento de modelos de previsão, foram investigadas técnicas de previsão qualitativas e quantitativas, onde observou-se que técnicas quantitativas estavam mais alinhadas com o objetivo do estudo. Dentre as técnicas quantitativas observou-se que modelos baseados em séries temporais e aprendizado de máquina se encaixam melhor nos objetivos do estudo e com os dados de safras passadas que estão disponíveis para a realização das previsões.

Entre as técnicas de aprendizado de máquina encontradas na revisão bibliográfica, foi possível observar que existe uma grande quantidade de estudos que fazem uso de técnicas de RNA, que vem sendo desenvolvida a vários anos e recebendo grandes avanços metodológicos, o que a torna em muitos estudos mais precisa do que previsões realizadas através de métodos clássicos de séries temporais como do modelo ARIMA.

Na busca por trabalhos relacionados com a aplicação de RNAs em cultivos agrícolas, pode-se identificar que redes ENN, conforme utilizado por KUNG *et al*. (2016), que obtiveram bons resultados na previsão da produção de tomates em Taiwan, e que redes do tipo LSTM, uma variante de RNA com memória, também apresenta bons resultados na previsão de safras agrícolas conforme demonstrado no trabalho de HAIDER *et al*. (2019), onde foi possível prever com boa precisão a safra de trigo no Paquistão.

Com base no levantamento bibliográfico, foi possível perceber que o uso de redes LSTM poderiam trazer benefícios para o estudo, tendo em vista a sua boa aplicabilidade em séries temporais demostrada em outros estudos, o que levou ao desenvolvimento de quatro modelos de redes neurais artificiais do tipo LSTM, tendo

como objetivo avaliar qual configuração de entrada de dados poderia fornecer um melhor resultado.

Para a utilização nestes modelos foram desenvolvidos dois conjuntos de dados distintos com dados obtidos da CONAB e do INMET, um deles apresentando 10 variáveis e o outro 52. Os conjuntos de dados foram divididos em duas variantes: uma que utiliza os dados brutos e outra em que os dados passam pela função de suavização Lowess, totalizando quatro tipos de conjunto de dados distintos, cada um deles utilizado no experimento dos modelos que também receberam valores distintos nos parâmetros de construção da rede, para se adequarem melhor as características de cada um dos conjuntos de dados carregados.

Os quatro modelos desenvolvidos no estudo apresentaram resultados bastante distintos, sendo possível observar nos experimentos realizados que os modelos que utilizaram o conjunto de dados com uma menor quantidade de variáveis apresentaram um desempenho expressivamente superior quando comparados aos que utilizaram o conjunto de dados com mais variáveis.

Outro ponto importante que foi possível observar no estudo é que a função de suavização Lowess, quando aplicada no conjunto de dados, diminuiu o ruído existente nos dados e gerou previsões com uma taxa de erro menor, quando comparadas ao conjunto de dados que não passou pela função.

Os resultados apresentados no estudo mostraram que o experimento com melhores resultados, apresentou um erro variando entre 0,19% a 9,23% no conjunto de dados de teste e uma diferença de 0,69% o que equivale a 134,3 mil toneladas quando comparado a previsão gerada para o mesmo ano pela CONAB e realizada com base de informações similar.

A previsão da safra de soja no estado do Rio Grande do Sul mostrou-se viável com a técnica de rede neural artificial utilizada no estudo, tendo encontrado valores similares ao padrão atualmente utilizado pela CONAB, e apresentando valores de erros similares a outros estudos como no elaborado por HAIDER *et al*. (2019). Este fato demonstra o potencial inicial relevante do estudo e que com o aprofundamento em estudos complementares, que podem buscar a melhora dos resultados em anos com condições climáticas extremas, um refinamento maior no preenchimento de dados faltantes no conjunto de dados e a exploração de outros períodos de tempo para previsão, podem tornar a técnica ainda mais eficaz e capaz de melhorar a performance das previsões hoje presentes no mercado.

9 REFERÊNCIAS BIBLIOGRÁFICAS

ABADI, Mart'in *et al*. TensorFlow: A System for Large-Scale Machine Learning. In: 12TH SYMPOSIUM ON OPERATING SYSTEMS DESIGN AND IMPLEMENTATION (16) 2016, Savannah, GA. **Anais**... Savannah, GA: USENIX Association, 2016. Disponível em: <https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi>

ABEYRATHNA, Kuruge Darshana; GRANMO, Ole-Christoffer; GOODWIN, Morten. Effect of Data from Neighbouring Regions to Forecast Dengue Incidences in Different Regions of Philippines Using Artificial Neural Networks. **Norsk Informatikkonferanse**, [s. l.], n. 2018: Norsk Informatikkonferanse, 2018. Disponível em: <https://ojs.bibsys.no/index.php/NIK/article/view/505>. Acesso em: 2 set. 2019.

AHMED, Nesreen K. *et al*. An empirical comparison of machine learning models for time series forecasting. **Econometric Reviews**, [s. l.], v. 29, n. 5, p. 594–621, 2010.

AMIDI, Shervine; AMIDI, Afshine. **CS 230 - Recurrent Neural Networks Cheatsheet**. 2019. Disponível em: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks#>. Acesso em: 7 nov. 2019.

APROSOJA/MT. **A história da soja**. 2019. Disponível em: <http://www.aprosoja.com.br/soja-e-milho/a-historia-da-soja#targetText=O processo de "domesticação" da,se cultivava trigo de inverno.>. Acesso em: 4 out. 2019.

APROSOJA BRASIL. **Economia - Aprosoja Brasil**. 2019. Disponível em: <https://aprosojabrasil.com.br/a-soja/economia/>. Acesso em: 17 ago. 2019.

ARTHUS, Murilo Gattás *et al*. Planejamento da safra de soja no Oeste do Paraná. **Produto & Produção**, [s. l.], v. 17, n. 4, 2017.

AUGUSTO C. CONCEIÇÃO, Octavio. **A expansão da soja no Rio Grande do Sul 1950-75**. PORTO ALEGRE, RS. Disponível em: <http://cdn.fee.tche.br/digitalizacao/teses-fee/expansao-soja-rio-grande-do-sul-teses-6/expansao-soja-rio-grande-do-sul-teses-6-texto.pdf.pdf>. Acesso em: 8 out. 2019.

BRAGA, Antônio de Pádua; CARVALHO, André Ponce de Leon F. De; LUDERMIR, Teresa Bernarda. **Redes Neurais Artificiais: Teoria e Aplicações**. 1. ed. Rio de Janeiro.

BRANCO, Sacha Tadeu; SAMPAIO, Raimundo José Borges De. APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS EM MODELOS DE PREVISÃO DE DEMANDA PARA EQUIPAMENTOS DE INFRA-ESTRUTURA DE TELECOMUNICAÇÕES. In: 2008, Rio de Janeiro. **Anais**... Rio de Janeiro Disponível em: <http://www.abepro.org.br/biblioteca/enegep2008_tn_sto_074_529_10851.pdf>. Acesso em: 26 out. 2019.

BRESSAN, Aureliano Angel. Tomada de decisão em futuros agropecuários com modelos de previsão de séries temporais. **RAE eletrônica**, [s. l.], v. 3, n. 1, p. 0–0, 2004. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1676-56482004000100005&lng=pt&tlng=pt>. Acesso em: 31 ago. 2019.

Capítulo 48 - Redes Neurais Recorrentes. **Deep Learning Book**, [s. l.], 2019. Disponível em: <http://deeplearningbook.com.br/redes-neurais-recorrentes/>. Acesso em: 7 nov. 2019.

Capítulo 51 - Arquitetura de Redes Neurais Long Short Term Memory (LSTM). **Deep Learning Book**, [s. l.], 2019. Disponível em: <http://deeplearningbook.com.br/arquitetura-de-redes-neurais-long-short-term-memory/>. Acesso em: 8 nov. 2019.

CASTRO, Nicole Rennó. **O impacto de variáveis climáticas sobre o valor da produção agrícola - análise para alguns estados brasileiros**. 2015. Biblioteca Digital de Teses e Dissertações da Universidade de São Paulo, Piracicaba, 2015. Disponível em: <http://www.teses.usp.br/teses/disponiveis/11/11132/tde-22042015-151044/>. Acesso em: 29 ago. 2019.

CONAB. **Conab - Série Histórica das Safras**. 2019a. Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras?start=20>. Acesso em: 15 ago. 2019.

CONAB. **Conab - Planilhas de Custos de Produção - Séries Históricas**. 2019b. Disponível em: <https://www.conab.gov.br/info-agro/custos-de-producao/planilhas-de-custo-de-producao/itemlist/category/414-planilhas-de-custos-de-producao-series-historicas?start=10>. Acesso em: 19 out. 2019.

CONAB. **ACOMPANHAMENTO DA SAFRA BRASILEIRA GRÃOS**. [s.l: s.n.]. Disponível em: <https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos/item/download/31188_59a3ca776bb30fa3764094b3acad2b1c>.

COX, W. J.; JOLLIFF, G. D. Growth and Yield of Sunflower and Soybean under Soil Water Deficits1. **Agronomy Journal**, [s. l.], v. 78, n. 2, p. 226, 1986. Disponível em: <https://www.agronomy.org/publications/aj/abstracts/78/2/AJ0780020226>. Acesso em: 5 set. 2019.

DRAKOS, Georgios. **How to select the Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics**. 2018. Disponível em: <https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0>. Acesso em: 11 nov. 2019.

EMBRAPA SOJA. **Tecnologias de produção de soja – Região Central do Brasil 2014.** 1a edição ed. Londrina. Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/95489/1/SP-16-online.pdf>. Acesso em: 15 out. 2019.

EMBRAPA SOJA. **História**. 2019. Disponível em: <https://www.embrapa.br/soja/cultivos/soja1/historia>. Acesso em: 4 out. 2019.

FAO *et al*. **The State of Food Security and Nutrition in the World 2019.** [s.l: s.n.]. Disponível em: <http://www.fao.org/3/ca5162en/ca5162en.pdf>. Acesso em: 12 set. 2019.

FERNEDA, Edberto. **Redes neurais e sua aplicação em sistemas de recuperação de informação**. [s.l: s.n.]. Disponível em: <http://www.scielo.br/pdf/ci/v35n1/v35n1a03.pdf>. Acesso em: 28 out. 2019.

FIGUEIREDO, Divino Cristino. Projeto GeoSafras Sistema de Previsão de Safras da Conab. **Revista de Política Agrícola**, [s. l.], v. 14, n. 2, p. 110–120, 2005. Disponível em: <https://seer.sede.embrapa.br/index.php/RPA/article/view/543/492>. Acesso em: 22 ago. 2019.

FÜRSTENAU, Vivian. A quebra da safra gaúcha 2004/05. **Carta de Conjuntura FEE RS**, [s. l.], 2005. Disponível em: <http://carta.fee.tche.br/article/a-quebra-da-safra-gaucha-200405/>. Acesso em: 25 abr. 2020.

GARMSIRI, Saeed. **Art of Choosing Metrics in Supervised Models Part 1**. 2018. Disponível em: <https://towardsdatascience.com/art-of-choosing-metrics-in-supervised-models-part-1-f960ae46902e>. Acesso em: 11 nov. 2019.

GURNEY, Kevin. **An Introduction to Neural Networks**. [s.l.] : CRC Press, 2014. Disponível em: <https://www.taylorfrancis.com/books/9781482286991>. Acesso em: 26 ago. 2019.

HAIDER, Sajjad *et al*. LSTM Neural Network Based Forecasting Model for Wheat Production in Pakistan. **Agronomy**, [s. l.], v. 9, n. 2, p. 72, 2019. Disponível em: <http://www.mdpi.com/2073-4395/9/2/72>. Acesso em: 8 nov. 2019.

HAYKIN, Simon. **Redes Neurais Princípios e práticas**. 2. ed. PORTO ALEGRE, RS: ARTMED Editora, 2001.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, [s. l.], v. 9, n. 8, p. 1735–1780, 1997.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science & Engineering**, [s. l.], v. 9, n. 3, p. 90–95, 2007.

INMET. **INMET - Instituto Nacional de Meteorologia**. 2019. Disponível em: <http://www.inmet.gov.br/portal/index.php?r=bdmep/bdmep>. Acesso em: 19 out. 2019.

KUNG, Hsu-Yang *et al*. Accuracy Analysis Mechanism for Agriculture Data Using the Ensemble Neural Network Method. **Sustainability**, [s. l.], v. 8, n. 8, p. 735, 2016. Disponível em: <http://www.mdpi.com/2071-1050/8/8/735>. Acesso em: 26 ago. 2019.

LEMOS, Fernando de Oliveira. **Metodologia para seleção de métodos de previsão de demanda**. 2006. Universidade Federal do Rio Grande do Sul, [s. l.], 2006. Disponível em: <https://www.lume.ufrgs.br/handle/10183/5949>

LIN, Tamy Ymei. **ESTUDO DE MODELOS DE PREVISÃO DE DEMANDA**. 2000. Fundação Getúlio Vargas, [s. l.], 2000. Disponível em: <https://pesquisa-eaesp.fgv.br/publicacoes/pibic/estudo-de-modelos-de-previsao-de-demanda>. Acesso em: 26 out. 2019.

Lowess Smoothing Function for Python using Pandas and Numpy. **GitHub**, 2020. Disponível em: <https://gist.github.com/dneuman/a688299c3050ad5c561b1ae680f56b61>. Acesso em: 18 abr. 2020.

MINISTÉRIO DA ECONOMIA. **Séries Históricas - Ministério da Economia**. 2019. Disponível em: <http://www.mdic.gov.br/index.php/comercio-exterior/estatisticas-de-comercio-exterior/series-historicas>. Acesso em: 17 ago. 2019.

MUNTASER, João Gonçalves Silva; SILVA, Valter Pereira Da; PENEDO, Antonio Sergio Torres. Aplicação de Redes Neurais na Previsão das Ações do Setor de Petróleo e Gás da Bm&FBovespa. **Revista FSA**, [s. l.], v. 14, n. 6, p. 49–71, 2017. Disponível em: <http://www4.fsanet.com.br/revista/index.php/fsa/article/view/1456/1350>. Acesso em: 31 ago. 2019.

MUTHUSINGHE, M. R. S. *et al*. Towards smart farming: Accurate prediction of paddy harvest and rice demand. In: IEEE REGION 10 HUMANITARIAN TECHNOLOGY CONFERENCE, R10-HTC 2019, **Anais**... : Institute of Electrical and Electronics Engineers Inc., 2019.

NEUMAIER, Norman *et al*. Estádios de desenvolvimento da cultura de soja. In: **Estresse em Soja**. [s.l: s.n.]. p. 19–44.

NumPy. 2020. Disponível em: <https://numpy.org/>. Acesso em: 18 abr. 2020.

Pandas - Python Data Analysis Library. **Pandas**, [s. l.], 2020. Disponível em: <https://pandas.pydata.org/about/index.html>. Acesso em: 15 abr. 2020.

PEDREGOSA, F. *et al*. Scikit-learn: Machine Learning in {P}ython. **Journal of Machine Learning Research**, [s. l.], v. 12, p. 2825–2830, 2011.

PEINADO, Jurandir; GRAEML, Alexandre Reis. **Administração da Produção (Operações Industriais e de Serviços)**. Curitiba.

PRATAMA, I. *et al*. A review of missing values handling methods on time-series data. In: 2016 INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY SYSTEMS AND INNOVATION (ICITSI) 2016, **Anais**... [s.l: s.n.]

ROCCA DA CUNHA, Gilberto *et al*. Zoneamento agrícola e época de semeadura para soja no Rio Grande do Sul. **Revista Brasileira de Agrometeorologia**, [s. l.], v. 9, p. 446–459, 2001. Disponível em: <https://www.researchgate.net/profile/Joao_Pires12/publication/294688659_Zoneamento_agricola_e_epoca_de_semeadura_para_soja_no_Rio_Grande_do_Sul/links/576bf8a008aedb18f3eb0489/Zoneamento-agricola-e-epoca-de-semeadura-para-soja-no-Rio-Grande-do-Sul.pdf>. Acesso em: 14 out. 2019.

SANTOS, Vanessa Sardinha Dos. **O que é neurônio? - Brasil Escola**. 2019a. Disponível em: <https://brasilescola.uol.com.br/o-que-e/biologia/o-que-e-neuronio.htm#>. Acesso em: 28 out. 2019.

SANTOS, Guilherme. **UMA APLICAÇÃO DE REDES NEURAIS RECORRENTES DO TIPO LSTM À PREVISÃO DOS PREÇOS DE CURTO PRAZO DO MERCADO DE ENERGIA ELÉTRICA BRASILEIRO**. [s.l: s.n.].

SEPLAG RS. **Participação nas Exportações e Produtos - Atlas Socioeconômico do Rio Grande do Sul**. 2019a. Disponível em: <https://atlassocioeconomico.rs.gov.br/participacao-nas-exportacoes-e-produtos>. Acesso em: 11 out. 2019.

SEPLAG RS. **Soja - Atlas Socioeconômico do Rio Grande do Sul**. 2019b. Disponível em: <https://atlassocioeconomico.rs.gov.br/soja>. Acesso em: 8 out. 2019.

SILVA, DIEGO FELIPE REIDEL DA. **PROPOSTA DE UM SISTEMA DE RECOMENDAÇÃO PARA O HEALTH SIMULATOR**. Novo Hamburgo.

TSAI, Chih-Yung; SHIUE, Yih-Chearing. **Predicting the Productions of Napier-grass Based on Back-Propagation Neural Network**. [s.l: s.n.]. Disponível em: <http://daa.knjc.edu.tw/ezfiles/6/1006/img/217/7.pdf>. Acesso em: 4 nov. 2019.

UNITED STATES DEPARTMENT OF AGRICULTURE. **Approved by the World Agricultural Outlook Board**. [s.l: s.n.]. Disponível em: <https://apps.fas.usda.gov/psdonline/circulars/production.pdf>. Acesso em: 15 ago. 2019.

WANG, Wen Chuan *et al*. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. **Journal of Hydrology**, [s. l.], v. 374, n. 3–4, p. 294–306, 2009.

Why use Keras. **Keras Documentation**, [s. l.], 2020. Disponível em: <https://keras.io/why-use-keras/>. Acesso em: 15 abr. 2020.

YOU, Jiaxuan *et al*. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. **Thirty-First AAAI Conference on Artificial Intelligence**, [s. l.], 2017. Disponível em: <https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14435>. Acesso em: 9 set. 2019.