UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo

Edson da Silva Guimarães

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Edson da Silva Guimarães

Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo

Dissertação apresentada ao Instituto de Ciências
Matemáticas e de Computação – ICMC-USP,
como parte dos requisitos para obtenção do título
de Mestre – Mestrado Profissional em Matemática,
Estatística e Computação Aplicadas à Indústria.
EXEMPLAR DE DEFESA

Área de Concentração: Matemática, Estatística e
Computação

Orientador: Prof. Dr. Paulino Ribeiro Villas-Boas

USP – São Carlos
Novembro de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

G963a Guimarães, Edson da Silva
 Aprendizado de Máquina aplicado à predição da
 produtividade da cultura da soja utilizando dados
 de clima e solo / Edson da Silva Guimarães;
 orientador Paulino Ribeiro Villas-Boas. -- São
 Carlos, 2019.
 75 p.

 Dissertação (Mestrado - Programa de Pós-Graduação
 em Mestrado Profissional em Matemática, Estatística
 e Computação Aplicadas à Indústria) -- Instituto de
 Ciências Matemáticas e de Computação, Universidade
 de São Paulo, 2019.

 1. Previsão da Produtividade da Soja. 2.
 Aprendizado de Máquina aplicado à predição da
 produtividade. 3. Random Forest. 4. Multilayer
 Perceptron. 5. Extreme Gradient Boosting . I.
 Ribeiro Villas-Boas, Paulino, orient. II. Título.

Edson da Silva Guimarães

**Machine Learning applied to soybean crop yield prediction
using climate and soil data**

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Paulino Ribeiro Villas-Boas

**USP – São Carlos
November 2019**

Este trabalho é dedicado à fonte de todas as boas energias do universo, pela confiança, desafios e vitórias experimentadas ao longo de minha carreira. Agradeço a oportunidade de conviver e escrever este trabalho com grandes professores, cientistas e profissionais, com o apoio do orientador Prof. Dr. Paulino Ribeiro Villas-Boas. Agradeço a todos os professores, pesquisadores e profissionais do Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) pelo carinho, apoio e conhecimento compartilhado ao longo do mestrado. Aos meus pais Francisco e Angelina pelo esforço constante para promover equilíbrio emocional e estrutura educacional ao longo da minha vida. Ao meu sogro Luiz e sogra Maria do Carmo pelo carinho, fé e incentivos. Ao meu filho Luiz Francisco, que embora criança, entende minhas necessárias ausências, colabora com meu desenvolvimento e me motiva diariamente a seguir sempre em frente. À minha maravilhosa esposa Luciana, a quem tenho a grande sorte de conviver e com ela aprender. Obrigado de todo coração pelos incentivos, paciência e apoio. A todos que direta ou indiretamente participaram desse grande desafio, muitíssimo obrigado.

AGRADECIMENTOS

Ao professor Dr. Paulino Ribeiro Villas-Boas, pela paciência, orientação e oportunidade de crescimento acadêmico.

À empresa I3E Innovation pelo suporte, inspiração e desafios durante o desenvolvimento deste trabalho.

Aos professores do ICMC com os quais tive o prazer de conviver durante essa jornada.

À minha família pela compreensão, carinho, incentivo e pelas lições de vida.

*“O trabalho não é ver aquilo que ninguém viu,
más pensar o que ninguém ainda pensou
sobre aquilo que todo mundo vê.”
(Arthur Schopenhauer)*

RESUMO

GUIMARÃES. E. **Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo**. 2019. 75 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A previsão do rendimento das culturas durante o período de crescimento é útil para as práticas de planejamento e manejo agrícola, para companhias desenvolvedoras de sementes avaliar a performance de novas variedades, para a tomada de decisão dos produtores, para os serviços do governo que lidam com importação e exportação, para organizações internacionais responsáveis por monitorar a produção e o comércio mundial de alimentos, para os comerciantes de *commodities*, bem como para a elaboração de políticas alimentares. As abordagens comuns para a previsão de rendimento incluem o uso de pesquisas manuais caras ou dados de sensoriamento remoto. Técnicas de inteligência artificial e ciência de dados podem oferecer inúmeros benefícios tais como, permitir o desenvolvimento de modelos computacionais para tomada de decisões, ajudar a combater a escassez de alimentos e capacitar pequenos agricultores, gerenciamento de doenças e pragas nas culturas, investigação de nichos agrícolas, lidar com mudanças climáticas, dentre outros. Neste contexto onde envolve a comercialização, armazenamento e processamento de grãos, é comum deparar-se com cenários de incerteza, especialmente quando se deseja prever a produtividade da lavoura e avaliar riscos de perdas da produtividade. O objetivo deste trabalho consiste em desenvolver modelos de *Machine Learning* para realizar predições da produtividade da cultura da soja. Nesta análise foram avaliadas Redes Neurais *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, posteriormente os modelos foram comparados entre si e com o modelo de estimativa de produtividade adotado pela FAO. Os dados referentes ao clima e características do solo foram adquiridos de estações meteorológicas no período entre 2010 e 2018 de 27 cidades, representando os três biomas do estado do Mato Grosso, enquanto que os dados da produtividade real foram coletados do sítio do IBGE. Os resultados foram avaliados através da métrica *Mean Absolute Percentage Error* (MAPE), que leva em consideração valores reais alimentados em modelos e valores ajustados a partir do modelo e calcula a diferença absoluta entre os dois como porcentagem do valor real, com base neste cálculo é possível obter a acurácia de cada modelo. O modelo baseado no *Multilayer Perceptron* apresentou acurácia de 95.20%, o modelo baseado em *Random Forest* apresentou acurácia de 95.03% e o modelo *Extreme Gradient Boosting* alcançou acurácia de 95.54% quando utilizados para fazer a predição do rendimento da soja para as 27 cidades no ano de 2018.

Palavras-chave: Previsão da Produtividade; *Big Data*; Ciências de Dados; Redes Neurais; *Random Forest*; *Extreme Gradient Boosting*; *Multilayer Perceptron*.

ABSTRACT

GUIMARÃES. E. **Machine Learning applied to soybean crop yield prediction using climate and soil data**. 2019. 75 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Crop yield forecasting during the growing season is useful for agricultural planning and management practices, for seed companies to evaluate new variety performance, for farmers' decision-making, for import and export government services, for international organizations responsible for monitoring world food production and trade, for commodity traders, and for food policy-making. Common approaches to yield prediction include the use of expensive manual surveys or accessible remote sensing data. Artificial intelligence and data science techniques can offer numerous benefits such as allowing the development of computational models for decision making, helping fight food scarcity and empower small farmers, managing crop diseases and pests, investigating agricultural niches, to cope With climate change, among others . In this context where it involves the marketing, storage and processing of grain, it is common to come across uncertainty scenarios, especially when one wants to predict crop yield and assess risks of yield losses. The objective of this work is to develop Machine Learning models based on Neural Networks and Decision Tree to make predictions of soybean crop yield. In this analysis were evaluated Multilayer Perceptron Neural Networks, Random Forest and Extreme Gradient Boosting, later the models were compared with each other and with the productivity estimate model adopted by FAO. climate and soil data were obtained from weather stations between 2010 and 2018 from 27 cities, representing the three biomes of the state of Mato Grosso, while real productivity data were collected from the IBGE site. The results were evaluated using the Mean Absolute Percentage Error (MAPE) metric, which takes into account real values fed into models and values adjusted from the model and calculates the absolute difference between the two as a percentage of the real value, based on which calculation is possible. get the accuracy of each model. The Multilayer Perceptron based model was 95.20% accurate, the Random Forest based model was 95.03% accurate, and the Extreme Gradient Boosting model was declared winner with 95.54% accurate when used to predict soybean yield for the 27 cities in the year 2018.

Keywords: Crop Yield Prediction; Big Data; Data Science; Neural Network; Random Forest; Extreme Gradient Boosting; Multilayer Perceptron.

LISTA DE ILUSTRAÇÕES

Figura 1 – Metodologia aplicada neste trabalho - Diagrama de Atividades	36
Figura 2 – Estádios Fenológicos da Soja	38
Figura 3 – Estádios Fenológicos da Soja - Dias após a Semeadura	39
Figura 4 – Biomas do Estado de Mato Grosso	39
Figura 5 – Municípios produtores de soja	40
Figura 6 – Fases do Particionamento dos dados	42
Figura 7 – Estrutura <i>Multilayer Perceptron</i>	44
Figura 8 – Estrutura da Rede Neural	46
Figura 9 – Análise de Importância - Relação de variáveis importantes	52
Figura 10 – Correlação das variáveis do modelo com a variável Produtividade Real . . .	53
Figura 11 – Correlação entre as Variáveis do modelo - Mapa de calor	54
Figura 12 – Correlação de Variáveis do modelo por Ano	55
Figura 13 – Previsão da Produtividade por Cidades - Ano 2018	58
Figura 14 – Previsão da Produtividade Tangará da Serra - MLP	59
Figura 15 – Previsão da Produtividade por Cidades - Ano 2018	61
Figura 16 – Previsão da Produtividade Tangará da Serra - RF	62
Figura 17 – Previsão da Produtividade por Cidades - XGB - Ano 2018	64
Figura 18 – Previsão da Produtividade Tangará da Serra - XGB	65
Figura 19 – Distribuição dos Erros- MAPE - Tangará da Serra	66
Figura 20 – Comparação do desempenho dos modelos na previsão da Produtividade das 27 cidades	68
Figura 21 – Comparação do desempenho dos modelos - Previsão Tangará da Serra . . .	69
Figura 22 – Comparação do desempenho dos quatro modelos - Previsão Tangará da Serra	70

LISTA DE TABELAS

Tabela 1 – Desempenho de predição utilizando variáveis meteorológicas e de solo . . .	29
Tabela 2 – Variáveis do Conjunto de Dados	37
Tabela 3 – Perfil do Conjunto de Dados	41
Tabela 4 – Particionamento dos Dados - 1 Cidade	41
Tabela 5 – Particionamento dos Dados - 27 cidades	41
Tabela 6 – Produtividade Real de Soja por Cidade - Quilos/ha	50
Tabela 7 – Média das variáveis do conjunto de dados - Ano	51
Tabela 8 – Parâmetros Rede - <i>Multilayer Perceptron</i>	56
Tabela 9 – Previsão da Produtividade - <i>Multilayer Perceptron</i> - 2018	57
Tabela 10 – Previsão da Produtividade - Tangará da Serra - MLP	59
Tabela 11 – Previsão da Produtividade - <i>Random Forest</i> (RF) - 2018	60
Tabela 12 – Previsão da Produtividade - Tangará da Serra - RF	62
Tabela 13 – Previsão da Produtividade - <i>Extreme Gradient Boosting</i> - 2018	63
Tabela 14 – Previsão da Produtividade - Tangará da Serra - XGB	65
Tabela 15 – Comparativo dos Erros dos Modelos - Tangará da Serra	66
Tabela 16 – Comparativo dos Erros dos Modelos - 2018	67

LISTA DE ABREVIATURAS E SIGLAS

BHC	Balanço Hídrico Climatológico
CAD	Capacidade de Água Disponível
CIMMYT	<i>The International Maize and Wheat Improvement Center</i>
CONAB	Companhia Nacional de Abastecimento
GPUs	<i>Graphic Processing Unit</i>
MAE	<i>Mean Absolute Error</i>
MAPE	<i>Mean Absolute Percentage Error</i>
ME	<i>Mega Environment</i>
MLR	<i>Multiple Linear Regression</i>
MSE	<i>Mean Square Error</i>
NASS	<i>National Agricultural Statistical Service</i>
NESR	<i>Northeastern Seaboard Region</i>
ReLU	<i>Rectifier Linear Unit</i>
RMSE	Root Mean Square Error
RNA	Redes Neurais Artificiais
USDA	United States Department of Agriculture

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Contextualização	23
1.2	Objetivos	25
1.2.1	<i>Geral</i>	25
1.2.2	<i>Específicos</i>	26
1.3	Organização	26
2	TRABALHOS RELACIONADOS	27
2.1	Aplicação de Redes Neurais artificiais para estimação do volume de árvores	27
2.2	Previsão de rendimento de culturas usando redes neurais profundas	28
2.3	<i>Random Forest</i> para previsão global e regionais de produção agrícola	29
2.4	<i>Extreme Gradient Boosting</i> para predição do rendimento do milho nos EUA	30
2.5	Modelos agrometeorológicos de estimativa de produtividade	31
3	MATERIAIS E MÉTODOS	35
3.1	Fluxo das Atividades	35
3.2	Conjunto de Dados	37
3.2.1	<i>Descrição do Conjunto de Dados</i>	37
3.2.2	<i>Particionamento dos Dados</i>	41
3.2.3	<i>Pré-processamento</i>	42
3.3	Modelos	43
3.3.1	<i>Ferramentas</i>	43
3.3.2	<i>Algoritmo</i>	44
3.3.3	<i>Função de Ativação</i>	45
3.3.4	<i>Rede Neural - Multilayer Perceptron</i>	46
3.3.5	<i>Random Forest</i>	47
3.3.6	<i>Extreme Gradient Boosting</i>	47
3.4	Avaliação do desempenho dos modelos	48
3.4.1	<i>Funções de Erro</i>	48
4	RESULTADOS	49
4.1	Análise Exploratória de Dados	50

4.1.1	<i>Sumarização - Produtividade de soja por hectare</i>	50
4.1.2	<i>Média das variáveis do conjunto de dados - Ano</i>	51
4.1.3	<i>Análise de Variáveis Importantes</i>	52
4.1.4	<i>Correlação de Variáveis</i>	53
4.1.5	<i>Correlação de Variáveis - Mapa de Calor</i>	54
4.2	<i>Modelo Rede Neural Multilayer Perceptron</i>	56
4.3	<i>Modelo - Random Forest</i>	60
4.4	<i>Modelo Extreme Gradient Boosting</i>	63
4.5	<i>Comparação dos Erros entre os modelos - MAPE</i>	66
4.6	<i>Comparação da acurácia dos Modelos de Predição</i>	68
5	CONCLUSÃO	71
	REFERÊNCIAS	73

INTRODUÇÃO

1.1 Contextualização

De acordo com a [ONU \(2012\)](#), a população global deverá atingir 9,5 bilhões até 2050, e tais números representam crescimento de 34,90% entre 2012 e 2050 devendo ocorrer principalmente em países em desenvolvimento. Aumentar a área de plantio de culturas ou criar mais gado não é a solução para evitar a crise alimentar global ([ROSE *et al.*, 2016](#)). É necessário maior eficiência nos métodos agrícolas atuais. Além disso, cerca de 805 milhões de pessoas no mundo não têm comida suficiente para levar uma vida saudável e ativa ([FAO, 2015](#)). Assim, o crescimento populacional, o aumento do consumo e renda per capita e a expansão das cidades nas próximas décadas tornam mais relevantes o presente debate sobre a incapacidade de atender às novas necessidades humanas.

Dentre os maiores países produtores agrícolas, o Brasil é um dos que tem grande potencial de expansão da oferta, como destaca [Saath e Fachinello \(2018\)](#). Esse aumento de oferta de alimentos deve vir da melhoria da produtividade agrícola e não da expansão de áreas agricultáveis, que tem avançado em ritmo mais lento nos últimos anos. Em relação à safra anterior, a área plantada aumentou apenas 2%, atingindo cerca de 63 milhões de hectares em 2018.

Estimativas da Companhia Nacional de Abastecimento (CONAB) indicam que os agricultores brasileiros devem colher 241,3 milhões de toneladas de grãos durante a safra 2018/2019. Os números do 10º Levantamento da Safra de Grãos 2018/2019 apontam para uma produtividade 6% superior aos 227,6 milhões de toneladas colhidas na última safra de grãos. Caso a previsão se confirme, o setor registrará mais um recorde favorável de produtividade. O principal destaque é o milho. Mesmo considerando que ocorra uma redução de 2,1 % na primeira safra, a segunda safra deve atingir produção recorde de 73,1 milhões de toneladas. Resultado que, se confirmado, significará um acréscimo de 35,6% em comparação à safra de 2017/18 ([CONAB, 2018](#)).

Apesar desse aumento em produtividade, [Saath e Fachinello \(2018\)](#) observam que nas

próximas décadas enfrentaremos muitas restrições, muitas delas referentes ao modelo tecnológico atual, à dependência do fósforo derivado da rocha e aos aspectos climáticos. Embora o debate sobre atendimento das demandas futuras de alimentos tenha sido retomado, poucos estudos abordaram o Brasil sob o aspecto da restrição da fronteira agrícola. [Saywer \(2013\)](#) cita que para alguns pesquisadores a fronteira agrícola brasileira não existe. Neste contexto [Saath e Fachinello \(2018\)](#) utilizaram dados sobre a demanda por exportações e consumo interno do Brasil entre 2012 e 2024 e apresentou a hipótese de que para o Brasil atender à demanda por alimentos até 2024 é necessário expandir a produtividade e a área plantada. Entretanto, [Saath e Fachinello \(2018\)](#) constataram que utilizando os limites legais para o uso das terras agropecuárias estimado pela Embrapa (2014), em 2012, restavam no Brasil apenas 3%, ou 7,8 milhões de hectares de terras possíveis de exploração. Assim tal disponibilidade de terras agricultáveis, se mantida as produtividades obtidas em 2012 não é suficiente para atender às novas demandas mundiais de alimentos por produtos brasileiros até 2024.

Por estas razões, precisamos de maneiras mais eficientes de cultivar alimentos e fazer melhor uso da terra. Tecnologias como sistemas de irrigação automatizados, *drones*, tratores autônomos, monitoramento da saúde das culturas, reconhecimento facial de animais e técnicas de ciências de dados podem ajudar os produtores a serem mais eficientes. Dentre as técnicas de ciência de dados, podemos destacar o uso de modelos de rendimentos das colheitas. Esses modelos permitem avaliar a performance de novas variedades pelas empresas de sementes, melhorar a tomada de decisão pelos produtores, elaborar políticas nacionais de alimentos pelos governos e, especialmente, controlar a crise global de alimentos pelos países ([HORIE; YAJIMA; NAKAGAWA, 1992](#)).

As técnicas relacionadas às Ciências de Dados podem oferecer inúmeros benefícios neste nicho de mercado que precisam responder rapidamente aos anseios da população dado à necessidade de atender às políticas de segurança alimentar e garantir lucratividade aos produtores rurais. No entanto, dentre os diversos suportes que as técnicas relacionadas às Ciências de Dados podem oferecer, as relacionadas à previsão da produtividade de culturas ainda são tímidas.

A previsão da produtividade das culturas é extremamente desafiadora devido a vários fatores complexos. Por exemplo, informações genótípicas são geralmente representadas por dados multidimensionais e de grande volume, contendo bilhões de marcadores para cada planta individual. É necessário estimar os efeitos dos marcadores genéticos, que podem estar sujeitos a interações com várias condições ambientais, estrutura física e composição química do solo e práticas de manejo de campo.

Apesar da dificuldade, tais análises podem fortalecer a tomada de decisão, melhorar as políticas globais de segurança alimentar, gerenciar o armazenamento dos grãos (silos), otimizar a negociação com fornecedores e aumentar a lucratividade do produtor rural. Para alcançar os benefícios citados acima é necessária a exploração de um conjunto de técnicas de pré-processamento dos dados coletados, além da modelagem e avaliação das ferramentas de predição.

Muitos estudos se concentraram em explicar o fenótipo (por exemplo, rendimento) como uma função explícita do genótipo (G), ambiente (E) e suas interações ($G \times E$) (DELACY *et al.*, 1996). Crossa *et al.* (1995) e Crossa e Cornelius (1997) usaram a regressão de sites e os modelos multiplicativos deslocados para análise de interação $G \times E$ dividindo ambientes em grupos semelhantes. Burgueno *et al.* (2008) propuseram uma abordagem integrada de modelos analíticos de fatores (FA) e lineares mistos para agrupar ambientes e genótipos e detectar suas interações. Eles também afirmaram que o modelo de FA melhorou a previsibilidade em até 6% quando havia padrões $G \times E$ complexos nos dados (BURGUENO *et al.*, 2011).

Mais recentemente, técnicas de aprendizado de máquina foram aplicadas para prever a produtividade das culturas, incluindo regressão multivariada, árvore de decisão e redes neurais artificiais. Uma característica marcante dos modelos de aprendizado de máquina é que eles tratam o resultado (rendimento da colheita) como uma função implícita das variáveis de entrada (clima, solo e características da cultura), o que poderia ser uma função altamente não linear e complexa.

Liu J. e Tian (2013) empregaram uma rede neural com uma camada oculta para prever a produção de milho usando dados do solo, clima e manejo. Drummond *et al.* (2013) usaram regressão linear múltipla *stepwise*, regressão de busca de projeção e redes neurais para prever o rendimento das culturas, e descobriram que seu modelo de rede neural superou os outros dois métodos. Drummond *et al.* (2016) propuseram a regressão ponderada para prever o rendimento de diferentes variedades de soja, que demonstraram desempenho superior aos algoritmos de regressão convencionais. Romero *et al.* (2016) propuseram árvore de decisão aplicada e mineração de regras de associação para classificar os componentes da produção de trigo.

Nesse contexto, este trabalho busca desenvolver abordagens de predição da produtividade da cultura da soja baseadas em aprendizado de máquina com foco em Redes Neurais (*Multilayer Perceptron*), *Random Forest* e *Extreme Gradient Boosting* e comparar o desempenho dos três modelos propostos com o modelo agrometeorológico de estimativa de produtividade proposto por (DOORENBOS; KASSAN, 1979). A hipótese apresentada é que os modelos de predição baseados em aprendizado de máquina têm maior acurácia na previsão da produtividade das culturas e, com isto, possibilita que pequenos ajustes na produtividade destas culturas possam atender à demanda sem aumento da área plantada.

1.2 Objetivos

1.2.1 Geral

O objetivo geral deste trabalho consistiu em desenvolver modelos de predição baseados em Rede Neural (*Multilayer Perceptron*), *Random Forest* e *Extreme Gradient Boosting* para realizar previsões da produtividade da cultura da soja em 27 cidades do estado do Mato Grosso,

comparar o desempenho dos três modelos com o do modelo agrometeorológico de estimativa de produtividade proposto por [Doorenbos e Kassan \(1979\)](#).

1.2.2 Específicos

- Desenvolver modelos de predição: *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*;
- Aplicar os modelos em uma base de dados real para predição da produtividade do ano de 2018 para 27 cidades do Estado do Mato Grosso;
- Aplicar os modelos em uma base de dados real para predição da produtividade do período 2010-2018 para uma cidade do estado do Mato Grosso;
- Comparar entre si a eficiência preditiva dos modelos *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*;
- Comparar a eficiência preditiva dos modelos *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting* com o modelo agrometeorológico de estimativa de produtividade proposto por [Doorenbos e Kassan \(1979\)](#).

1.3 Organização

Este trabalho está organizado em cinco capítulos: Introdução, Trabalhos Relacionados, Materiais e Métodos, Resultados e Conclusão. No segundo capítulo serão apresentados trabalhos relacionados aos modelos de previsão baseados em técnicas de aprendizado de máquina. No terceiro capítulo será apresentado o fluxo das atividades para elaboração deste trabalho, a base de dados utilizada, os procedimentos adotados para a construção dos modelos, a etapa de pré-processamento e transformação dos dados, técnicas de análise exploratória, os modelos de previsão usados neste trabalho e o método de avaliação dos modelos, ou seja, serão apresentados os materiais e métodos utilizados no desenvolvimento deste trabalho, que consiste na predição do rendimento da soja utilizando técnicas de aprendizado de máquina. No quarto capítulo serão apresentados os resultados das previsões baseadas na Rede Neural (*Multilayer Perceptron*), *Random Forest* e *Extreme Gradient Boosting*, os três modelos serão comparados entre si e comparados com o modelo agrometeorológico de estimativa de produtividade. O quinto capítulo conclui os resultados deste estudo e apresenta propostas de trabalhos futuros.

TRABALHOS RELACIONADOS

Com o objetivo de entender e levantar modelos relacionados a predição, foram analisados iniciativas e modelos baseados em paradigmas de aprendizado de máquina, árvore de decisão, aumento do gradiente e modelo agrometeorológico. Os trabalhos relacionados foram divididos em quatro tipos de modelos: (i) redes neurais *Multilayer Perceptron*, (ii) *Random Forest*, (iii) aumento extremo de gradiente *Extreme Gradient Boosting* e (iv) modelo agrometeorológico. Esta divisão tem como objetivo facilitar o entendimento sobre as diversas abordagens existentes para cada um dos aspectos de implementação dos modelos de predição.

2.1 Aplicação de Redes Neurais artificiais para estimação do volume de árvores

Binoti, Binoti e Leite (2014) propuseram uma metodologia com a utilização de Redes Neurais Artificiais (RNA), com foco principal na redução das medições de diâmetro ao longo do fuste das árvores, possibilitando a cubagem das árvores em pé ao invés de derrubá-las. Neste trabalho os autores utilizaram dados de 27.000 árvores de eucalipto provenientes do sul da Bahia. Foram utilizadas as variáveis: diâmetro com casca a 1,30m do solo (dap) (cm) e altura total (m); medidas de 2,0 em 2,0 m a partir do fuste até o topo foram coletadas após determinado as posições 0,10; 0,5; 1,0; 1,50 e 2,0 m ao longo do fuste. Foram investigados diversos modelos de regressão linear e não linear para a estimação do volume de povoamentos florestais, no entanto, alguns modelos e situações foram tendenciosos e apresentaram níveis de erro inaceitáveis para os objetivos do manejo florestal.

A RNA foi treinada visando estimar o volume com e sem casca. Os dados foram particionados em 80% para treinamento e 20% para teste. As melhores RNAs foram selecionadas com base na correlação entre o volume observado e o volume estimado e na estabilidade dos índices de treinamento das redes fornecidos pelo software Statistica 7 nas fases de treinamento,

de seleção e de avaliação. A acurácia do modelo foi medida pelo procedimento estatístico proposto por [LEITE e OLIVEIRA \(2002\)](#). A metodologia proposta mostrou-se eficiente para a estimação do volume de árvores, sendo indicada para a obtenção do volume total com e sem casca de plantação de eucalipto, além disso, o modelo proposto exige somente a medição de posições da árvore que são facilmente mensuráveis em campo, sem derrubá-las. Em alguns casos, o número ideal de árvores-amostra para ajuste de um modelo volumétrico pode chegar a 100 ou 150 árvores.

2.2 Previsão de rendimento de culturas usando redes neurais profundas

[Saaed e Wang \(2019\)](#) apresentaram uma proposta de previsão da produtividade de milho híbridos baseado em aprendizado de máquina em resposta ao desafio proposto pela Syngenta em 2018. O conjunto de dados utilizado é um dos maiores e mais abrangentes disponíveis publicamente para pesquisa e previsão de rendimento, possuindo 2.267 variedades híbridas experimentais plantadas em 2.247 localidades entre 2008 e 2016 nos Estados Unidos e no Canadá. Os dados de treinamento eram compostos de três conjuntos: genótipo da cultura, desempenho produtivo e dados ambientais/climáticos. O conjunto de dados do genótipo continha informações genéticas para todos os híbridos experimentais, cada um com 19.465 marcadores genéticos.

O conjunto de dados de desempenho de rendimento continha o rendimento observado, o rendimento de verificação (rendimento médio em todos os híbridos do mesmo local) e a diferença de rendimento de 148.452 amostras para diferentes híbridos plantados em anos e locais diferentes, e o conjunto de dados ambientais continha 8 variáveis do solo e 72 variáveis climáticas (6 variáveis meteorológicas medidas durante 12 meses de cada ano).

O desafio era prever o rendimento do milho em 2017 e [Saaed e Wang \(2019\)](#) utilizaram redes neurais profundas para prever a produtividade e avaliar a performance da predição. A rede foi cuidadosamente projetada para aprender com o complexo e não linear relacionamento entre os genes, condições ambientais e os dados históricos. O modelo encontrado foi relativamente sensível à qualidade da previsão do tempo, sugerindo a importância desta variável na predição. Foram implementadas duas redes neurais profundas e outros três modelos de predição para comparação com as redes neurais.

A Tabela 1 compara o desempenho dos quatro modelos nos conjuntos de dados de treinamento e validação em relação ao Root Mean Square Error (RMSE) e ao coeficiente de correlação. Esses resultados sugerem que as redes neurais profundas superaram os outros três modelos em graus variados. O modelo de redes neurais profundas foi particularmente eficaz em prever e verificar o rendimento sendo aproximadamente 11% de seus respectivos valores médios. A precisão do rendimento de verificação foi um pouco maior que a do rendimento, porque o

primeiro é o rendimento médio em todas as variedades híbridas e todos os anos no mesmo local, o que é mais fácil de prever do que o rendimento das variedades híbridas individual em um local específico em um ano específico.

Tabela 1 – Desempenho de predição utilizando variáveis meteorológicas e de solo

Model	Response Variable	Training RMSE	Training correlation coefficient(%)	Validation RMSE	Validation correlation coefficient (%)
DNN	Yield	10.55	88.30	12.79	81.91
	Check yield	8.21	91.00	11.38	85.46
	Yield difference	11.79	45.87	12.40	29.28
Lasso	Yield	20.28	36.68	21.40	27.56
	Check yield	18.85	28.49	19.87	23.00
	Yield difference	15.32	19.78	13.11	6.84
SNN	Yield	12.96	80.21	18.04	60.11
	Check yield	10.24	71.18	15.18	60.48
	Yield difference	9.92	58.74	15.19	11.39
RT	Yield	14.31	76.70	15.03	73.80
	Check yield	14.55	82.00	14.87	69.95
	Yield difference	17.62	21.12	15.92	5.10

Fonte: Artigo *Crop Yield Prediction Using Deep Neural Networks*, (SAAED; WANG, 2019).

Apesar do bom desempenho do modelo baseado em redes neurais profundas, essa abordagem possui a limitação de ser uma caixa preta, assim como a maioria dos modelos baseados em aprendizado de máquina. O modelo conseguiu capturar interações $G \times E$, no entanto, sua complexa estrutura dificultou a produção de hipóteses testáveis que poderiam fornecer *insights* biológicos. No intuito de corrigir esta limitação, os autores realizaram a seleção das *features* utilizando uma rede neural profunda treinada com método *backpropagation*. Essa técnica possibilitou encontrar várias *features* importantes e revelou que o fator ambiental teve um efeito maior no rendimento da cultura que o genótipo.

2.3 *Random Forest* para previsão global e regionais de produção agrícola

Jeong *et al.* (2016) avaliaram o desempenho do algoritmo *Random Forest* usando *Multiple Linear Regression* (MLR) como referência para previsões globais e regionais do rendimento para três culturas básicas: trigo (*Triticum aestivum*), milho (*Zea mays*) e batata (*Solanum tuberosum*). Especificamente, aplicaram *Random Forest* para prever: i) a produção de grãos de trigo em escala global que inclui todas as regiões designadas como *Mega Environment* (ME) definidos pelo *The International Maize and Wheat Improvement Center* (CIMMYT); ii) produção de grãos de milho nos Estados Unidos por um período de 30 anos; e iii) produção de batata e milho de silagem na *Northeastern Seaboard Region* (NESR), composta por treze estados do Maine à Virgínia nos Estados Unidos. Os autores concluíram que o algoritmo de *Random Forest* tem muitas vantagens em modelar sistemas de cultivo complexos, mas ainda não está sendo amplamente utilizado.

Foi demonstrado que *Random Forest* fornece desempenho superior na previsão do rendimento de todas as culturas e regiões testadas. O resultado do estudo mostrou potencial

para a implementação de um algoritmo como método alternativo de modelagem estatística para previsões de produtividade. Deve-se observar que *Random Forest* tem o risco de *overfitting* nas condições em que os dados de treinamento estão bastante concentrados, enquanto sua precisão pode diminuir quando os dados de treinamento são escassos.

Da mesma forma, a aplicação da regressão *Random Forest* para extrapolar as dimensões externas dos dados de treinamento deve ser evitada. Em resumo, o resultado obtido por Jeong *et al.* (2016) sustentam que a regressão baseada em *Random Forest* pode ser uma ferramenta eficaz para prever o rendimento das culturas em escala global e regional, com a seleção cautelosa de um conjunto de dados de treinamento que inclui uma ampla gama de variabilidade de preditores.

No artigo *Implementation of Effective Crop Selection by Using the Random Forest Algorithm*, Karthikeyan *et al.* (2018) propuseram um sistema com uma abordagem para analisar grandes conjuntos de dados. Primeiramente, esta proposta fornece uma introdução à aplicação da análise massiva em conjunto de dados no campo da agronomia. Dados sobre clima, irrigação e rendimento da cultura das últimas décadas foram coletados de várias fontes (por exemplo, estação meteorológica e registros de irrigação) e analisados para produzir como saída a maior produtividade de cada cultura em suas respectivas condições geográficas.

A partir desses registros, o modelo de *Random Forest* foi treinado para avaliar a colheita perfeita para aquelas condições geográficas. Os autores observaram que o sistema proposto obteve maior acurácia, boa performance na predição comparado aos modelos baseados em classificadores *Naive Bayes*. Desta forma o sistema proporcionou aos agricultores prever o valor da colheita e assim plantar de forma mais eficientemente.

2.4 *Extreme Gradient Boosting* para predição do rendimento do milho nos EUA

Na dissertação *Forecasting U.S. Corn Yield Using Machine Learning*, Oliva (2016) desenvolveu e testou uma metodologia que utiliza *Machine Learning* (*Random Forest* e *Extreme Gradient Boosting*) para analisar e fazer predição do rendimento da colheita. Analisando as *features* mais significantes foi possível determinar como elas influenciavam o rendimento, e portanto, ajudar os agricultores a entender quais fatores deveriam prestar mais atenção durante o ciclo de vida da cultura. Foi considerado neste estudo a cultura do milho nos Estados Unidos e então, modelado uma rede hierárquica das várias regiões geográficas dos Estados Unidos e seus relacionamentos. Foi feito um piloto nos estados de Iowa e Texas.

Os dados utilizados no estudo foram principalmente dados públicos e provenientes da web, com a maioria deles obtida do *National Agricultural Statistical Service* (NASS). Os dados foram divididos em 5 categorias: i) Dados-alvo, consistindo em rendimento, produção, área colhida e área plantada; ii) Dados de composição do solo, consistindo em medições de campo

que indicam a concentração de potássio, nitrogênio, fósforo e outras substâncias; iii) Dados de manejo do solo, consistindo em dados de pesquisa sobre uso de fertilizantes e produtos químicos, bem como variáveis associadas à forma como os agricultores utilizam a terra; iv) Dados meteorológicos, consistindo em medidas diárias de temperatura, precipitação e outras variáveis climáticas; e v) Dados demográficos, consistindo em dados sobre a população e sobre a estrutura e o comportamento da economia.

O modelo proposto por (OLIVA, 2016) teve desempenho superior aos modelos desenvolvidos pelo United States Department of Agriculture (USDA), porém, teve acurácia menor. Esta diferença, pode em parte ser explicado pelo fato do governo americano basear suas previsões em levantamentos de campos com grande acurácia.

Nos últimos meses modelos baseados em *Extreme Gradient Boosting* tem crescido muito em número de implementações, assumindo o posto de modelo de predição mais utilizado entre os vencedores de competições em plataformas como Kaggle.

No artigo *Crop yield predictions - high resolution statistical model for intra-season forecasts applied to corn in the US* Yiqing et al. (2016) propuseram um *framework* para modelos gerais de produção agrícola, com capacidade de prever grandes quantidades de variação na produção com alta granularidade espacial e fazer previsões no início da temporada. A análise do rendimento de milho ficou concentrado no cinturão do milho dos Estados Unidos (Iowa, Illinois, Indiana, Minnesota, Dakota do Sul, Ohio, Nebraska, Wisconsin, Missouri e Kansas). Os autores modelaram explicitamente o rendimento no nível do condado no Cinturão do Milho, onde há mais dados disponíveis, e os combinaram com as previsões do USDA no nível estadual para o resto do país para obter uma previsão no nível nacional.

O modelo com melhor desempenho entre todos os que foram testados é o de 12 de outubro com um meta modelo formado pelo empilhamento dos algoritmos *xgboost* e *cubist*. No nível nacional, o modelo alcançou uma taxa de erro absoluta de 1,89%, em média, para os anos de 2001 a 2015. No nível local, 73% dos municípios do cinturão do milho foram previstos dentro de 15,8 *bushels*/acre de seus rendimentos reais. Usando esse método, observou que a previsão de produção de milho nos EUA para 2016 foi de 173,75 *bushels*/acre.

2.5 Modelos agrometeorológicos de estimativa de produtividade

Um dos fatores mais importantes para a produtividade de uma cultura é a disponibilidade hídrica do solo. Com base nesse requisito, THORNTHWAITE e MATHER (1955) desenvolveram um método para determinar o regime hídrico de um local, sem a necessidade de medidas diretas das condições do solo. O método do Balanço Hídrico Climatológico (BHC) leva em conta a precipitação, o Capacidade de Água Disponível (CAD) e a estimativa da evapotranspiração

potencial. Com essas informações, o BHC permite deduzir a evapotranspiração real, a deficiência ou o excedente hídrico, e o total de água retida no solo em cada período.

Esse modelo quantifica o efeito da falta de água disponível no solo sobre o decréscimo da produtividade potencial, e baseia-se na penalização da produtividade potencial da cultura em função da relação ETr/ETo , ocorrida durante o ciclo da cultura, de modo que:

$$\frac{Yr}{Yp} = 1 - ky \left(1 - \frac{ETr}{ETo} \right), \quad (2.1)$$

onde Yr é a produtividade estimada ($kg.ha^{-1}$); Yp , a produtividade potencial ($kg.ha^{-1}$); ETr , a evapotranspiração real (mm); ETo , a evapotranspiração potencial (mm) e ky , o coeficiente de penalização da produtividade por deficit hídrico, com valor de 0,85 para todo o ciclo.

O modelo aditivo desenvolvido por STEWART J. I.; HAGAN e PRUITT (1976) e proposto por Doorenbos e Kassan (1979) tem suas bases no método anterior, porém leva em conta o somatório das relações ETr/ETo em nível decendial, reduzindo a produtividade à medida que as necessidades hídricas da cultura da soja deixam de ser satisfeitas durante os quatro estádios fenológicos considerados. Os coeficientes de penalização da produtividade por deficit hídrico (ky), para cada estágio fenológico, de acordo com Doorenbos e Kassan (1979) foram: 1) desenvolvimento vegetativo (0,2); 2) florescimento (0,8); 3) enchimento de grãos (1,0) e 4) maturação (0,2), de forma que:

$$\frac{Yr}{Yp} = 1 - \left[\sum_{i=1}^4 ky_i \left(1 - \frac{ETr}{ETo} \right)_i \right], \quad (2.2)$$

onde o somatório é realizado nos quatro estádios fenológicos das plantas.

No artigo *Modelo agrometeorológico para estimativa da produtividade para a cultura da soja no estado de São Paulo*, Camargo, Brunini e Miranda (1986) desenvolveram um modelo agrometeorológico de estimativa de produção de grãos com base em dados fenológicos obtidos de experimentos conduzidos pelo Departamento de Leguminosas do Instituto Agrônomo, em Campinas, Ribeirão Preto e Pindamonhangaba, durante os anos agrícolas de 1977/78, 1978/79 e 1979/80 para os cultivares de soja UFV-1, Viçoja, Santa Rosa e Paraná. Foram aplicados os mesmos tratos culturais para as três localidades, sendo feito o controle fitossanitário quantas vezes necessário.

Durante o ciclo de cada cultura, foram assinalados os dados fenológicos: (a) data de plantio; (b) data de início de florescimento, quando 50% das plantas tinham pelo menos uma flor; (c) data de início de maturação, quando 50% das plantas tinham coloração palha-escura; (d) produção de grãos e os dados meteorológicos diários, relativos aos períodos considerados, foram obtidos nos postos meteorológicos da Seção de Climatologia Agrícola próximos à área do experimento.

O modelo proposto, semelhante ao de Doorenbos e Kassan (1979) e STEWART J. I.; HAGAN e PRUITT (1976), baseou-se na penalização da produtividade potencial, segundo

as condições meteorológicas predominantes no ciclo da cultura, ou seja, considerou-se que cada elemento climático exerce certo controle na produtividade da cultura, sendo um fator de eficiência. Assim, a produção final foi definida como uma função da produção potencial da região e da sua interação com os elementos meteorológicos, de modo que:

$$YEst = (YPot) \times (FTer) \times (FExc), \quad (2.3)$$

onde $YEst$ é a produtividade estimada em grãos (kg/ha); $YPot$, a produtividade potencial do cultivar, para cada localidade (kg/ha); $FTer$, o fator térmico indicativo do grau de desenvolvimento relativo da cultura e $FExc$, o fator referente à penalização para excedente hídrico.

Os autores aplicaram o método proposto por [THORNTHWAITE e MATHER \(1955\)](#) para cálculo do balanço hídrico e a determinação da evapotranspiração potencial pelo método proposto por [CAMARGO \(1962\)](#). Além disso, consideraram produtividade potencial dos cultivares como sendo a maior produtividade esperada para o cultivar na região, desde que não ocorra qualquer restrição.

[Camargo, Brunini e Miranda \(1986\)](#) concluíram que as produtividades potenciais estimadas, embora satisfatórias, podem alcançar maior acurácia se aplicados em experimentos específicos para cada região do estado; que a produtividade da soja tende a diminuir com o atraso do plantio a partir de outubro; que a relação ER/EP apresentou resultados bastante satisfatórios na quantificação da penalização dos défices hídricos e que nos anos com excedente hídricos elevados o fator excedente de penalização no modelo agrometeorológico de estimativa de produtividade se mostrou muito importante.

MATERIAIS E MÉTODOS

Neste capítulo são apresentados os materiais e métodos utilizados na elaboração desta dissertação de mestrado. Levantamento e coleta dos dados foram os maiores desafios deste trabalho. Há uma grande quantidade de dados disponíveis, no entanto, muitas estações meteorológicas são recentes, outras não são automáticas e isto implicou diretamente na série histórica de cada cidade pesquisada.

3.1 Fluxo das Atividades

A Figura 1 apresenta o fluxo das principais atividades adotadas no desenvolvimento dos modelos de predição baseado no aprendizado de máquina (Redes Neurais *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*), resumindo a metodologia da pesquisa. As principais atividades desta pesquisa são: (1) coleta da base de dados brutos (informações de clima, solo, cultura e produtividade); (2) pré-processamento dos dados, como limpeza e transformação dos dados; (3) análise exploratória através de estatística descritiva; e (4) análise de características (*features*) mais importantes para os modelos, separação em amostra de treinamento, validação e teste, desenvolvimento de modelos de previsão, avaliação e comparação da acurácia dos modelos. Observa-se que trata-se de um processo iterativo e interativo, portanto, embora as tarefas estejam em sequência, essas tarefas, eventualmente tem que se retomar a etapas anteriores.

Figura 1 – Metodologia aplicada neste trabalho - Diagrama de Atividades

Fonte: Elaborada pelo autor.

3.2 Conjunto de Dados

3.2.1 Descrição do Conjunto de Dados

As predições realizadas pelos modelos baseados em Redes Neurais, *Random Forest* e *Extreme Gradient Boosting* foram baseados em dados gerados por estações meteorológicas (INMET, 2019) no período de 2010 a 2018, como podemos ver na tabela 2. Os dados representam todos os três biomas presentes no estado do Mato Grosso (Amazônica, Pantanal e Cerrado), como podemos observar na Figura 4. A cultura da soja foi utilizada no estudo e seguimos as recomendações presentes na portaria numero 13 publicada em 20 de Julho de 2017 pelo Ministério da Agricultura. O objetivo desta portaria é aprovar o Zoneamento Agrícola de Risco Climático para a cultura da Soja no Estado do Mato Grosso, ano safra 2017-2018 (PORTARIA 13, 2017).

A Portaria 13 (2017) define quais os tipos de solo, quais municípios estão aptos para o cultivo da soja, indica as cultivares, o período de semeadura e os respectivos riscos associados, considerando (tipo de solo, cultivar e semana da semeadura). Objetivou-se, com o zoneamento agrícola de risco climático, identificar os municípios aptos e os períodos de plantio com menor risco climático para o cultivo da soja no Estado.

A Tabela 2 apresenta todas as variáveis presentes nos dados coletados.

Tabela 2 – Variáveis do Conjunto de Dados

Variável	Tipo de Variável	Descrição
ETo	Contínua	Evapotranspiração de referência
Kc	Contínua	Coefficiente da cultura
ETc	Contínua	Evapotranspiração da cultura
Precipitação	Contínua	Precipitação Pluvial
ARM	Contínua	Armazenamento de água no solo
Alteração	Contínua	Alteração do Armazenamento
ETr	Contínua	Evapotranspiração real
Perda	Contínua	Perda estimada
Produtividade	Contínua	Produtividade estimada
Temperatura	Contínua	Temperatura
CAD	Contínua	Capacidade de água disponível
ETr/ETc	Contínua	ETr/ETc
P-ETr	Contínua	Precipitação - ETr
Deficit	Contínua	Deficiência hídrica
Excedente	Contínua	Excedente hídrico
TMax	Contínua	Temperatura máxima
TMin	Contínua	Temperatura mínima

A Produtividade Relativa e a Perda de produtividade acumulada (%), considerando as

perdas apenas por deficiência hídrica, são estimadas de acordo com o modelo conceitual proposto por [Doorenbos e Kassan \(1979\)](#), que relaciona a queda relativa de rendimento com o déficit relativo de evapotranspiração ($1 - ETr/ETc$).

No decorrer do seu processo produtivo a soja passa por diferentes estádios. [FEHR e CAVINESS \(1977\)](#), visando unificar a descrição desses diferentes estádios, desenvolveram uma metodologia denominada de estágios fenológicos. O sistema proposto divide os estádios em: vegetativos, representados pela letra V e os reprodutivos pela letra R, exceto os estádios, VE (emergência) e VC (cotilédone), como podemos observar nas Figuras 2 e 3.

Figura 2 – Estádios Fenológicos da Soja

Fonte: Adaptado Bayer, 2015

A Figura 3 apresenta os estádios fenológicos e os respectivos dias após a semeadura.

Figura 3 – Estádios Fenológicos da Soja - Dias após a Semeadura

Fonte: Diogo Kersten

A Figura 4 apresenta os três principais biomas do estado do Mato Grosso, sendo eles: Amazônica, Pantanal e Cerrado.

Figura 4 – Biomas do Estado de Mato Grosso

Fonte: Elaborada pelo autor.

A Figura 5 mostra as 27 cidades selecionadas, que possuem dados gerados pelas estações meteorológicas desde 2010 e que tenham informações sobre a produtividade real por ha.

Figura 5 – Municípios produtores de soja

Fonte: Elaborada pelo autor.

As informações referentes à produtividade real foram coletadas do sítio do IBGE (IBGE, 2019). Na Tabela 3 podemos observar o perfil do conjunto de dados.

Tabela 3 – Perfil do Conjunto de Dados

Qte cidades	Qtde anos	Qtde amostras/Ano	Variáveis/Amostra	Total de amostras
27	9	120	17	29160

3.2.2 Particionamento dos Dados

Finalmente, o conjunto de dados foi particionado em três conjuntos de dados distintos usados para: treinamento, validação e teste. Antes do particionamento, as linhas no conjunto de dados foram embaralhadas (reorganizadas aleatoriamente) para evitar a distribuição desigual de dados semelhantes e distorcer o resultado.

Após estudar a literatura mencionada nesta dissertação e observar os esquemas de particionamento de dados comumente usados e visando atender às duas abordagens de predição chegamos as seguintes taxas de divisão observadas nas Tabelas 4 e 5.

Na Tabela 4 podemos observar o particionamento final dos dados para a predição dos 9 anos de uma única cidade.

Tabela 4 – Particionamento dos Dados - 1 Cidade

Partição	% Relativo	Qtde Entradas
Treinamento	96.2%	234
Validação	25%	59
Teste	3.8%	9

Na Tabela 5 podemos observar o particionamento final dos dados para a predição de 1 ano para cada cidade do modelo. Neste caso o modelo é executado 1 vez para cada cidade. Na Figura 6 podemos observar as fases do particionamento dos dados, a seleção das *features* e ajuste dos modelos propostos neste trabalho.

Tabela 5 – Particionamento dos Dados - 27 cidades

Partição	% Relativo	Qtde Entradas
Treinamento	99.589%	242
Validação	25%	61
Teste	0.411%	1

Figura 6 – Fases do Particionamento dos dados

Fonte: Elaborada pelo autor.

3.2.3 Pré-processamento

Foram realizadas transformações nos dados para adequar aos modelos propostos neste trabalho. O processo de transformação consistiu em 3 etapas:

1. Padronização das variáveis
2. Transformação das 120 variáveis/ano/cidade em 1 amostra/ano/cidade
3. Adição da Produtividade Real à amostra/ano/cidade

Foram removidas variáveis correlacionadas e ruidosas e posteriormente os dados foram normalizados utilizando o método *MinMaxScaler* da biblioteca *Scikit Learn*. O método re-escala os dados entre 0 e 1, conforme definido na equação 3.1 para normalização dos dados.

$$Y_i = \frac{X_i - X_{min}}{X_{max} - X_{min}}, \quad (3.1)$$

onde, Y_i é o i^{esimo} valor normalizado, X_i é o valor real e X é o conjunto dos valores daquela variável.

Após este processo, foram mantidas as variáveis: *ETo*, *Kc*, *Precipitação*, *ARM*, *Alteração*, *ETr*, *Temperatura*, *CAD*, *Deficit*, *Excedente*, *TMax*, *TMin*. O valor de CAD Final, que representa a Capacidade de Água Disponível quando a raiz estiver plenamente desenvolvida, é estimado em função da Cultura e do Solo selecionados. Para efeito de simulação, contudo, é possível alterar esse valor.

Após a padronização das variáveis, transformamos as 120 amostras/ano/cidade em uma amostra/ano/cidade. Desta cada amostra passou a ter 1440 variáveis, resultado das 12 variáveis resultantes da padronização multiplicado por 120 amostras. Ao final do processo de transformação das variáveis adicionamos a variável Produtividade Real, Produtividade (estimada pelo método proposto por [Doorenbos e Kassan \(1979\)](#)), Cidade e Ano, totalizando ao final 1444 variáveis.

3.3 Modelos

3.3.1 Ferramentas

O modelo de Rede Neural foi desenvolvido utilizando as bibliotecas *scikit learn*, *Keras* e *TensorFlow* na plataforma *Anaconda Navigator*.

Scikit-learn é um projeto *Python* de código aberto com um grande número de pessoas fazendo contribuições constantemente. No entanto, há um pequeno grupo de desenvolvedores principais ou mantenedores do código base e, antes que qualquer contribuição chegue à base de código, ela é verificada por dois desenvolvedores principais diferentes. Todos os diferentes modelos de aprendizado de máquina ou funções de suporte neste trabalho foram implementados no *scikit-learn* e estão prontos para uso imediato, ou foram desenvolvidos com a ajuda de diferentes funções do *scikit-learn*. Para facilitar o uso, todas as funções do *scikit-learn* que requerem entrada do usuário têm valores padrão que foram projetados para que a função seja tão útil em uma ampla variedade de situações quanto possível.

Keras é um invólucro de alto nível que é executado sobre o *Theano* ou sobre o *Tensorflow*, ao contrário do *Lasagne*, que sempre usa algum código *Theano*. Ele foi projetado para minimizar a sobrecarga, para permitir a prototipagem rápida e fácil de algoritmos de aprendizado de máquina.

O *Tensorflow* é uma biblioteca *Python* de código aberto criada em 2015 como parte do projeto Google Brain e possui algumas ferramentas integradas, como o *TensorBoard*, que permite ao usuário produzir facilmente gráficos visualizando itens como taxa de aprendizado, pesos do modelo, funções de perda e muito mais. O *Tensorflow* também é a única biblioteca capaz de distribuir a carga de trabalho nas *Graphic Processing Unit* (GPUs) de vários dispositivos conectados, o que pode ser uma grande vantagem computacional.

Anaconda é uma das mais populares plataformas para ciências de dados e possui mais de 15 milhões de usuários. *Anaconda Navigator* é uma aplicação desktop que traz embarcada a distribuição *Anaconda* e permite executar diversas ferramentas de forma fácil e rápida, gerenciar os pacotes, ambientes e canais conda.

As visualizações dos gráficos dos dados foram desenvolvidas utilizando os pacotes *matplotlib* e *seaborn* para linguagem *Python*.

3.3.2 Algoritmo

Na teoria matemática das redes neurais, o teorema de aproximação universal observa que uma rede *feed-forward* tal como as redes *Multilayer Perceptron* com uma única camada oculta contendo um número finito de neurônios é capaz de aproximar qualquer função contínua mensurável com qualquer grau de precisão desejado.

As redes *Multilayer Perceptron* são uma ferramenta poderosa usada para criar modelos preditivos a partir de um conjunto de dados de entrada contendo um determinado número de variáveis e, como resultado, prever uma ou mais variáveis de destino. A topologia de um *Multilayer Perceptron* é simples e direta, veja a Figura 7. A rede é dividida em camadas, que vêm em três tipos: camada de entrada, camadas ocultas e camadas de saída. Neste trabalho, um único nó de saída foi usado para produzir a estimativa da produtividade da soja (regressão).

Cada nó na camada inferior é conectado a todos os nós na camada acima, formando um gráfico bipartido completo. Cada sinal de entrada propaga-se no sentido para frente, camada a camada através da rede. Um peso está associado à conexão, como pode ser observado na Figura 7. A saída de um nó na rede é calculada com uma função de ativação cuja entrada é a soma ponderada das conexões de entrada.

Figura 7 – Estrutura *Multilayer Perceptron*

Fonte: Elaborada pelo autor.

Para a previsão do rendimento das 27 cidades em um ano específico, o algoritmo particiona o conjunto de dados com a amostra de 1 das 27 cidades do ano especificado para o conjunto de testes, e os demais anos da cidade selecionada juntamente com as outras 26 cidades para o conjunto de treinamento como podemos observar na Tabela 5. Este procedimento é repetido para todas as cidades.

Para a previsão do rendimento de uma cidade no período 2010 a 2018, o algoritmo seleciona uma cidade e particiona o conjunto de dados com todas as amostras da cidade selecionada para o conjunto de testes e as demais cidades para o conjunto de treinamento, como podemos observar na Tabela 4.

3.3.3 Função de Ativação

A função de ativação é a transformação não linear que fazemos ao longo do sinal de entrada. Elas basicamente decidem se um neurônio deve ser ativado ou não. Ou seja, se a informação que o neurônio está recebendo é relevante para a informação fornecida ou deve ser ignorada. Na equação 3.2 podemos observar que a função de ativação é mais uma camada matemática no processamento.

$$Y = Y = Activation(\sum(weights * inputs) + bias) \quad (3.2)$$

Em nosso modelo utilizamos a função *Rectifier Linear Unit* (ReLU) na camada de entrada e nas camadas ocultas. Atualmente a ReLU é a função de ativação mais utilizada ao projetar redes neurais. A função ReLU é definida como:

$$f(x) = x^+ = \max(0, x) \quad (3.3)$$

Com o objetivo de evitar *overfitting* adicionamos uma camada de *dropout*, que de forma simplificada ignorará alguns neurônios escolhidos aleatoriamente durante a fase de treinamento, ou seja, tais neurônios não serão considerados durante o processo. Mais tecnicamente, em cada estágio de treinamento, nós individuais são eliminados da rede com probabilidade $1 - p$ ou mantidos com probabilidade p , reduzindo o tamanho da rede.

3.3.4 Rede Neural - Multilayer Perceptron

A Figura 8 apresenta a estrutura da rede neural desenvolvida para este trabalho. Na camada de entrada temos 1.440 neurônios, possuindo 7 camadas ocultas densas, uma camada de *dropout* com *rate* de 0.2 (fração de unidades a serem ignoradas) e 1 camada de saída. Na camada de entrada e nas camadas ocultas foram aplicadas a função de ativação ReLu e na camada de saída uma função de ativação Linear.

Figura 8 – Estrutura da Rede Neural

Fonte: Elaborada pelo autor.

A rede foi compilada com os parâmetros (loss = “mape”, optimizer = “adam”, metrics=[“mae”, “mape”]) e posteriormente treinada com os parâmetros (x=train, y=target, epochs=500, verbose = 1, batch_size=8, validation_split = 0.25, callbacks=callbacks_list).

Para a camada de entrada e para cada uma das 7 camadas ocultas foram gerados 2.077.922 parâmetros, resultado do produto da quantidade de neurônios(1441) pela quantidade de pesos(1441) mais o *bias*(1).

3.3.5 *Random Forest*

O modelo de *Random Forest* foi treinado para prever o rendimento da colheita da soja usando as mesmas variáveis e dados utilizadas na rede neural para fins de *benchmarking*.

Os dados de origem para o treinamento do modelo são inicializados de forma a criar vários subconjuntos gerando um grande número de árvores aleatoriamente. As variáveis preditoras são avaliadas em função do quanto elas diminuem a impureza do nó quando selecionadas para as divisões ou com que frequência fazem previsões bem-sucedidas.

A impureza do nó é definida como *Mean Square Error* (MSE) do nó na regressão de *Random Forest*. Para predição foi usado o algoritmo *RandomForestRegressor* da biblioteca *Scikit Learn* com os parâmetros (`n_estimators = 100`, `criterio = MSE`, `min_samples_split = 2`, `bootstrap = True`) e os conjuntos de Treinamento e *Target*).

3.3.6 *Extreme Gradient Boosting*

O modelo de *XGBoost* foi treinado com os mesmos dados empregados no treinamento do modelo *Random Forest* e *Multilayer Perceptron*. O algoritmo automaticamente adiciona novos modelos para corrigir os erros cometidos pelos modelos existentes. Os modelos são adicionados sequencialmente até que não sejam feitas mais melhorias. Para controlar *overfitting*, o *XGBoost* incorpora o conceito de regularização e validação cruzada nativa para evitar modelos muito complexos. É chamado de aumento de gradiente porque usa um algoritmo de descida de gradiente para minimizar a perda ao adicionar novos modelos.

No método de *Boosting* a avaliação da performance dos modelos é feita sequencialmente, e os critérios na etapa seguinte podem ser alterados em função da avaliação na etapa anterior. Por exemplo, se o primeiro “avaliador” deu uma pontuação baixa para determinado critério, esta nota será considerada pelo próximo “avaliador” em sua avaliação. Assim, a avaliação final é mais precisa e o processo, mais dinâmico. No método *Gradient Boosting* os erros são minimizados pelo algoritmo de descida do gradiente. Isto equivale a evitar “mínimos locais” (onde o segundo melhor critério receberia a melhor avaliação) ajustando os critérios para capturar o melhor desempenho.

O método *Extreme Gradient Boosting* busca melhorar o desempenho do *gradient boosting* otimizando a utilização do software e hardware. Seria como criar um ambiente perfeito para que os avaliadores possam fazer o melhor trabalho possível: a nota mais factível no menor tempo.

Foi utilizado o método *XGBRegressor* importado da biblioteca *xgboost* com os parâmetros (`n_estimators=100`, `learning_rate = .1`, `max_depth = 6`, `random_state=42`, `early_stopping_rounds=10`) e a métrica de avaliação da performance foi utilizado MAPE.

3.4 Avaliação do desempenho dos modelos

3.4.1 Funções de Erro

Uma função de erro é usada para calcular a diferença entre o valor previsto e o valor alvo (desejado). Os modelos propostos utilizam *Mean Absolute Percentage Error* (MAPE) como função de erro. De acordo com a [OECD \(2003\)](#), acurácia (ou exatidão) indica o grau de concordância que há entre o resultado da medição e o dito valor verdadeiro (aquele que é aceito, desde que estabelecido por uma definição ou consenso) da grandeza. O MAPE considera valores reais alimentados em modelos e valores ajustados a partir do modelo e calcula a diferença absoluta entre os dois como porcentagem do valor real, com base neste cálculo, é possível obter a acurácia de cada modelo. Segundo [Barros \(2015\)](#), o MAPE pode ser definido pela equação 3.4:

$$MAPE = \frac{\sum_{t=1}^N \frac{Y_{(t)} - \hat{Y}_{(t)}}{Y_{(t)}} \cdot 100}{N}, \quad (3.4)$$

onde $Y_{(t)}$ é o valor real no período, e $\hat{Y}_{(t)}$ é o valor da previsão para o período t com N totais de observações.

Para efeitos comparativos também calculamos o *Mean Absolute Error* (MAE). O MAE tem dimensão igual à dimensão dos valores observados e preditos. Seu valor representa o desvio médio entre observado e predito, sendo considerado peso igual para todos os desvios. Segundo [Barros \(2015\)](#), o MAE pode ser definido pela equação 3.5:

$$MAE = \frac{1}{n} \sum_{i=1}^n |e_i|, \quad (3.5)$$

onde $|e_t| = |y_i - x_i|$, sendo y_i o valor predito no período, e x_i o valor real.

RESULTADOS

No capítulo 3 foi descrita a base de dados, como foi realizada a filtragem/seleção das variáveis e como foi implementado os modelos de aprendizado de máquina. Neste capítulo, são apresentados os resultados das previsões baseadas em aprendizado de máquina por meio dos algoritmos *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, assim como o comparativo com o modelo proposto por Doorenbos e Kassan (1979).

A Seção 4.1 apresenta a sumarização do rendimento das culturas, a análise das variáveis mais importantes do conjunto de dados e a correlação das variáveis do conjunto de dados. Nas Seções 4.2, 4.3 e 4.4, são apresentadas as previsões de produtividade das 27 cidades para o ano de 2018 e da cidade de Tangará da Serra no período 2010-2018 obtidas pelos modelos *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, respectivamente.

Para avaliar o desempenho de predição dos modelos foi utilizado a função de perda *Mean Absolute Percentage Error* (MAPE) como apresentado na Seção 4.5. A Seção 4.6 apresenta os resultados da previsão da produtividade para as 27 cidades e os compara com o modelo de predição de Doorenbos e Kassan (1979).

4.1 Análise Exploratória de Dados

Nesta Seção, são apresentadas análises exploratórias no conjunto de dados, considerando o período de Outubro de 2010 à Fevereiro de 2018, das 27 cidades produtoras de soja do estado de Mato Grosso avaliadas neste trabalho.

4.1.1 Sumarização - Produtividade de soja por hectare

Na Tabela 6, são apresentadas as cidades com maior produtividade no período 2010 a 2018. Pode-se observar que a cidade de **Cáceres**, foi a cidade com maior produtividade (safra 2018) e a cidade de **Nova Ubiratã** tem a maior média de produtividade das safras (2010-2018).

Tabela 6 – Produtividade Real de Soja por Cidade - Quilos/ha

Cidade	2010	2011	2012	2013	2014	2015	2016	2017	2018	Med.	Max
Água Boa	3059	3300	3180	2820	3240	3060	2400	3270	3300	3070	3300
Alto Taquari	2994	2928	3200	3180	3217	3120	3240	3280	3480	3182	3480
Cáceres	3300	3180	3180	3200	2956	3120	3000	3600	4080	3291	4080
Campo Verde	3180	3300	3300	3000	3120	3113	3100	3360	3300	3197	3360
Canarana	3000	3180	3120	2700	3180	3060	2580	3331	3300	3050	3331
Comodoro	2819	2820	3060	2520	3117	3060	3060	3300	3360	3013	3360
Confresa	3000	3030	3300	1668	3120	3300	3420	3300	3600	3082	3600
Diamantino	2939	3180	3120	2880	3082	3180	2940	3120	3240	3076	3240
Gaucha do Norte	3149	3180	3300	2934	3350	3300	2100	3030	3048	3043	3350
Guarantã	3000	3120	3300	3300	3000	2971	2820	3000	3000	3057	3300
Guiratinga	3001	3180	2900	3004	3120	3121	2880	3200	3300	3078	3300
Itiquira	3000	3120	3180	3000	3300	3085	3000	3380	3120	3132	3380
Juara	3000	3120	3200	3200	3000	2872	2880	3300	3100	3075	3300
Matupá	3000	3120	3300	3321	2946	2998	2820	3300	3100	3101	3321
Nova Ubiratã	3150	3420	3181	2998	2921	3168	3940	3450	3480	3301	3940
Nova Xavantina	2699	3000	2520	2340	3120	2400	1980	3000	3000	2673	3120
Novo Mundo	3000	3300	3300	3300	2967	3300	3240	3300	3280	3221	3300
Paratininga	2680	3120	2800	2964	3350	3120	2100	3000	3120	2917	3350
Pontes e Lacerda	3000	2800	3000	3000	3200	3300	3540	3540	3520	3211	3540
Poxoreu	3179	3440	3300	3120	3120	3120	3060	3300	3180	3202	3440
Querência	3300	3468	3180	3000	3180	3180	2760	3570	3360	3222	3570
Rondonópolis	3000	3120	2940	3000	3120	3240	2880	3120	3380	3089	3380
S. F. do Araguaia	3000	3300	3184	3028	3105	3360	2880	3242	3310	3157	3360
São J. Rio Claro	3180	2985	3000	2934	3060	3000	2700	3120	3250	3025	3250
Sinop	3071	3300	2862	2936	3046	3106	2880	3450	3480	3126	3480
Sorriso	2984	3480	3239	3124	3146	3148	2880	3480	3720	3245	3720
Tangará da Serra	3000	3120	3180	3060	2980	3180	3000	3060	3480	3118	3480

4.1.2 Média das variáveis do conjunto de dados - Ano

Na Tabela 7 podemos observar a média das variáveis por ano. Os valores em negritos representam as maiores médias e os valores em itálico representam as menores médias de cada variável do conjunto de dados. Utilizamos a soja com ciclo de 120 dias, com plantio iniciando-se no início de outubro e a colheita em fevereiro do ano seguinte. O valor da variável “Prod. Estimada” é em percentual e variável ano representa o ano da safra.

Tabela 7 – Média das variáveis do conjunto de dados - Ano

Variável	2010	2011	2012	2013	2014	2015	2016	2017	2018
ETo	2.34	2.31	2.39	2.25	2.29	2.56	<i>2.10</i>	2.24	2.19
ETc	1.81	1.76	1.86	1.75	1.79	1.94	<i>1.64</i>	1.71	1.76
Precipitação	7.61	8.12	8.13	7.78	<i>6.56</i>	6.180	7.95	7.56	6.82
ARM	27.65	30.37	29.58	30.05	28.69	27.76	29.55	29.50	29.25
Alteração	0.27	0.29	0.29	0.28	0.28	<i>0.26</i>	0.29	0.28	0.27
ETr	1.51	1.63	1.65	1.62	1.56	1.65	<i>1.50</i>	1.52	1.59
Prod. Estimada ¹	<i>87.44</i>	93.48	88.76	93.80	91.611	87.70	91.50	92.28	91.30
Temperatura	25.19	<i>24.96</i>	25.58	25.05	25.41	26.53	25.27	25.77	25.73
CAD	33.01	33.01	33.01	33.01	33.01	33.01	33.01	33.01	33.01
Deficit	0.30	0.14	0.21	<i>0.13</i>	0.23	0.28	0.15	0.18	0.17
Excedente	6.11	6.50	6.50	6.17	5.03	<i>4.56</i>	6.46	6.06	5.25
TMax	30.96	<i>30.79</i>	31.28	30.92	31.53	32.67	31.07	31.56	31.38
TMin	21.61	<i>21.36</i>	22.09	21.59	21.74	22.58	21.90	22.33	22.33
Prod. Real	3025.33	3170.78	3123.19	2945.59	3113.44	3110.44	<i>2891.85</i>	3274.19	3329.19

¹ Prd_Est. = Produtividade estimada pelo método proposto por (DOORENBOS; KASSAN, 1979)

4.1.3 Análise de Variáveis Importantes

As árvores de decisão, utilizam a estratégia de dividir para conquistar: um problema complexo é decomposto em sub-problemas mais simples e recursivamente esta técnica é aplicada a cada sub-problema. O critério utilizado para realizar as partições é o da utilidade do atributo. Aplica-se, por este critério, um determinado ganho de informação a cada atributo. O atributo escolhido como atributo teste para o corrente nó é aquele que possui o maior ganho de informação. A partir desta aplicação, inicia-se um novo processo de partição. A divisão é feita de tal forma que valores semelhantes da variável dependente acabem no mesmo conjunto após a divisão. A condição é baseada na impureza, que, no caso de problemas de classificação, é a impureza de *Gini* ou o ganho de informação (entropia), enquanto nas árvores de regressão sua variação.

Ao treinar uma árvore, podemos calcular quanto cada recurso contribui para diminuir a impureza ponderada. Para análise, foi utilizado *Scikit-feature_importances_* do pacote Scikit-Learn que é baseado nessa lógica, mas no caso da *Random Forest*, estamos falando sobre a média da diminuição da impureza sobre as árvores.

A Figura 9 apresenta a análise das variáveis importantes do conjunto de dados (*Random Forest* e *Extreme Gradient Boosting*), referente às 27 cidades do estado de Mato Grosso. Podemos observar que embora sejam as variáveis mais importantes, sua importância é muito baixa. A variável “TMAX80”, considerada a mais importante para o *Random Forest* teve importância de 0,05 a variável “ARM94” foi considerada a mais importante para o modelo *Extreme Gradient Boosting* com pontuação de 0.016.

Figura 9 – Análise de Importância - Relação de variáveis importantes

Fonte: Elaborada pelo autor.

4.1.4 Correlação de Variáveis

A Figura 10 apresenta a correlação das variáveis do modelo com a Produtividade Real. Podemos observar que não há relação forte entre as demais variáveis do conjunto de dados e a variável Produtividade, sugerindo complexidade e não linearidade do conjunto de dados. Além disso, na comparação entre a variável Ano e a variável Produtividade, pode observar que a maioria das cidades obtiveram produtividade superior à média do estado do Mato Grosso no ano 2018 - 3346 kg/ha.

Figura 10 – Correlação das variáveis do modelo com a variável Produtividade Real

Fonte: Elaborada pelo autor.

4.1.5 Correlação de Variáveis - Mapa de Calor

A Figura 11 apresenta a correlação entre todas as variáveis do conjunto de dados. Podemos observar que há baixa correlação entre a variável Produtividade obtida através do método proposto por [Doorenbos e Kassan \(1979\)](#) e as demais variáveis do modelo, o que justificaria a necessidade da previsão da produtividade utilizando algoritmos de aprendizado de máquina.

Figura 11 – Correlação entre as Variáveis do modelo - Mapa de calor

Fonte: Elaborada pelo autor.

A Figura 12 apresenta a correlação das variáveis do modelo com a variável Produtividade (Produtividade Real) nos anos entre 2010 e 2018.

Figura 12 – Correlação de Variáveis do modelo por Ano

Fonte: Elaborada pelo autor.

4.2 Modelo Rede Neural *Multilayer Perceptron*

A tabela 8 apresenta a distribuição dos parâmetros da rede neural construída para previsão da produtividade da cultura soja das 27 cidades do estado do Mato Grosso. Podemos observar que para a camada de entrada da rede e para cada uma das 7 camadas ocultas foram gerados 2.077.922 parâmetros [(1441 Pesos + 1 *bias*) × 1441 Entradas].

Tabela 8 – Parâmetros Rede - *Multilayer Perceptron*

Layer (type)	Output Shape	Param #
dense_1 (Dense)	(None, 1441)	2077922
dense_2 (Dense)	(None, 1441)	2077922
dense_3 (Dense)	(None, 1441)	2077922
dense_4 (Dense)	(None, 1441)	2077922
dense_5 (Dense)	(None, 1441)	2077922
dense_6 (Dense)	(None, 1441)	2077922
dense_7 (Dense)	(None, 1441)	2077922
dense_8 (Dense)	(None, 1441)	2077922
dropout_1 (Dropout)	(None, 1441)	0
dense_9 (Dense)	(None, 1)	1442

Aplicando a equação 4.1 no sumário apresentado na Tabela 8 encontramos 16.624.818 parâmetros treinados na rede.

$$Y = \left(\sum_{l=1}^L (w_l + b_l) \times i_l \right) + o, \quad (4.1)$$

onde L é o total de camadas ocultas + camada de entrada, w é a quantidade de pesos, i é o i^{esimo} parâmetro de entrada, b é o *bias* e o é a quantidade de parâmetros da camada de saída.

Na tabela 9 é possível observar que algumas cidades tiveram uma diferença maior entre os valores reais e previstos. Podemos considerar que existem fatores externos que podem ter impactado no modelo de previsão proposto, como a capacidade de água disponível em relação aos estádios fenológicos da cultura, bem como outros fatores. Desta forma, o modelo não conseguiu detectar cenários como esse, pois tais variáveis não foram consideradas na análise inicialmente. Ainda assim, o modelo apresentou no melhor caso acurácia de 99.95% para a cidade de Nova Ubitatã e no pior caso acurácia 82.52% para a cidade de Cáceres. No entanto, quando comparamos o valor previsto de 3367 kg/ha com 3291 kg/ha que é a produtividade média do período 2010 a 2018 para a cidade de Cáceres a acurácia é de 97.73%.

Tabela 9 – Previsão da Produtividade - *Multilayer Perceptron* - 2018

Cidade	Produtividade Kg/ha				
	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
Água Boa	3300	3106	194	5.86%	94.14%
Alto Taquari	3480	3172	308	8.86%	91.14%
Cáceres	4080	3367	713	17.48%	82.52%
Campo Verde	3300	3214	86	2.59%	97.41%
Canarana	3300	3422	-122	3.70%	96.30%
Comodoro	3360	3338	22	0.66%	99.34%
Confresa	3600	3345	255	7.08%	92.92%
Diamantino	3240	3094	146	4.51%	95.49%
Gaucha do Norte	3048	3069	-21	0.68%	99.32%
Guarantã	3000	3009	-9	0.29%	99.71%
Guiratinga	3300	3215	85	2.57%	97.43%
Itiquira	3120	3410	-290	9.30%	90.70%
Juara	3100	3338	-238	7.68%	92.32%
Matupá	3100	3297	-197	6.36%	93.64%
Nova Ubitatã	3480	3478	2	0.05%	99.95%
Nova Xavantina	3000	3035	-35	1.15%	98.85%
Novo Mundo	3280	3327	-47	1.43%	98.57%
Paratininga	3120	3011	109	3.48%	96.52%
Pontes e Lacerda	3520	3525	-5	0.16%	99.84%
Poxoreu	3180	3344	-164	5.15%	94.85%
Querência	3360	3545	-185	5.51%	94.49%
Rondonópolis	3380	3368	12	0.36%	99.64%
São Felix do Xingu	3310	3691	-381	11.51%	88.49%
São José do Rio Claro	3250	3110	140	4.31%	95.69%
Sinop	3480	3304	176	5.05%	94.95%
Sorriso	3720	3474	246	6.62%	93.38%
Tangará da Serra	3480	3230	250	7.18%	92.82%

A Figura 13 apresenta os valores reais e os previstos no ano de 2018 para as 27 cidades estudadas. Cáceres teve produtividade real de 4080 kg/ha e o modelo previu rendimento de 3367 kg/ha, ou seja, 17,48% abaixo. No entanto, quando comparamos a média histórica do período 3291 kg/ha com o rendimento real a diferença é de 24%. Analisando o rendimento real do período, podemos considerar o rendimento de 2018 como *outlier*, e este fator pode ter impactado no poder de predição do modelo.

Figura 13 – Previsão da Produtividade por Cidades - Ano 2018

Fonte: Elaborada pelo autor.

Na tabela 10 podemos conferir o resultado da predição do rendimento da soja para a cidade de Tangará da Serra no período 2010 a 2018, utilizando para treinamento os dados das demais 26 cidades do conjunto de dados. É possível observar que alguns anos tiveram uma diferença maior entre os valores reais e previstos. Podemos considerar que existem fatores externos que podem ter impactado no modelo de previsão proposto, como ter ou não ocorrido irrigação, bem como outros fatores. Apesar de não ter detectado tais mudanças, o modelo apresentou acurácia de 85.36% no pior caso (Ano 2012) e 99.96% no melhor caso (Ano 2013).

Tabela 10 – Previsão da Produtividade - Tangará da Serra - MLP

Produtividade Kg/ha					
Ano	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
2010	3000	2807	-193	6.42%	93.58%
2011	3120	2929	-191	6.12%	93.88%
2012	3180	3646	466	14.64%	85.36%
2013	3060	3059	-1	0.04%	99.96%
2014	2980	3030	50	1.68%	98.32%
2015	3180	3014	-166	5.23%	94.77%
2016	3000	2817	-183	6.09%	93.91%
2017	3060	3013	-47	1.54%	98.46%
2018	3480	3230	-250	7.18%	92.82%

Na Figura 14, temos o resultado do modelo de rede neural *Multilayer Perceptron*. O gráfico apresenta os valores reais e os previstos durante o período de 2010 a 2018 para a cidade de Tangará da Serra. A previsão do modelo acompanha a tendência da produtividade real: cresce em 2011 e 2012, decresce em 2013 e 2014, cresce em 2015, decresce em 2016, cresce em 2017 e 2018.

Figura 14 – Previsão da Produtividade Tangará da Serra - MLP

Fonte: Elaborada pelo autor.

4.3 Modelo - *Random Forest*

Na tabela 11 podemos observar que algumas cidades (Cáceres, Confresa e Sorriso) tiveram uma diferença maior entre os valores reais e previstos. A ausência de informações sobre outros fatores externos como a capacidade de água disponível em relação aos estádios fenológicos da planta, ausência ou presença de irrigação artificial, dentre outros pode ter influenciado no modelo. Assim, o modelo não conseguiu detectar tais cenários, dado que estas variáveis não foram consideradas na análise inicialmente. Desta forma, o modelo apresentou no melhor caso acurácia de 98.48% para a cidade de São José do Rio Claro e no pior caso acurácia 85.88% para a cidade de Cáceres. Entretanto, quando comparamos o valor previsto de 3504 kg/ha com a média do período 3291 kg/ha para a cidade de Cáceres a acurácia é de 93.91%.

Tabela 11 – Previsão da Produtividade - *Random Forest* (RF) - 2018

Cidade	Produtividade Kg/ha				
	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
Água Boa	3300	3150	150	4.55%	95.45%
Alto Taquari	3480	3314	166	4.77%	95.23%
Cáceres	4080	3504	576	14.12%	85.88%
Campo Verde	3300	3245	55	1.68%	98.32%
Canarana	3300	3164	137	4.14%	95.86%
Comodoro	3360	3198	162	4.82%	95.18%
Confresa	3600	3225	375	10.43%	89.57%
Diamantino	3240	3120	120	3.70%	96.30%
Gaucha do Norte	3048	3135	-87	2.85%	97.15%
Guarantã	3000	3099	-99	3.31%	96.69%
Guiratinga	3300	3228	72	2.18%	97.82%
Itiquira	3120	3386	-266	8.53%	91.47%
Juara	3100	3228	-128	4.13%	95.87%
Matupá	3100	3225	-125	4.03%	95.97%
Nova Ubiratã	3480	3314	166	4.76%	95.24%
Nova Xavantina	3000	3094	-94	3.12%	96.88%
Novo Mundo	3280	3228	52	1.59%	98.41%
Paratininga	3120	3061	59	1.88%	98.12%
Pontes e Lacerda	3520	3332	188	5.33%	94.67%
Poxoreu	3180	3236	-56	1.76%	98.24%
Querência	3360	3209	151	4.49%	95.51%
Rondonópolis	3380	3118	262	7.75%	92.25%
São Felix do Xingu	3310	3186	124	3.76%	96.24%
São José do Rio Claro	3250	3201	49	1.52%	98.48%
Sinop	3480	3235	245	7.04%	92.96%
Sorriso	3720	3347	373	10.02%	89.98%
Tangará da Serra	3480	3202	278	7.99%	92.01%

A Figura 15 apresenta os rendimentos reais e previstos no ano de 2018 para as 27 cidades do modelo. A maior diferença entre o rendimento real e o previsto foi de 14.12% para a cidade de Cáceres, seguido pela diferença de 10.42% para a cidade de Confresa e 10.02% para a cidade de Sorriso. Cáceres teve produtividade real de 4080kg/ha e o modelo previu rendimento de 3504 kg/ha. O rendimento previsto para a cidade de Nova Xavantina foi de 3094 kg/ha e o rendimento real 3000 kg/ha, apresentando diferença de 3.12% entre real e previsto. Comparando o rendimento previsto com a rendimento médio 2673 kg/ha, a diferença é de 15.73%.

Figura 15 – Previsão da Produtividade por Cidades - Ano 2018

Fonte: Elaborada pelo autor.

Na tabela 12 é possível observar que apenas no ano de 2016 houve uma diferença pouco maior entre os valores reais e os previstos pelo modelo baseado em árvore de decisão. Apesar desta diferença a acurácia do modelo foi de 90.54% no pior caso (Ano 2016) e 99.40% no melhor caso (Ano 2017).

Tabela 12 – Previsão da Produtividade - Tangará da Serra - RF

Produtividade Kg/ha					
Ano	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
2010	3000	3186	186	6.19%	93.81%
2011	3120	3158	38	1.21%	98.79%
2012	3180	2960	-220	6.91%	93.09%
2013	3060	2969	-91	2.98%	97.02%
2014	2980	3143	163	5.46%	94.54%
2015	3180	3042	-138	4.35%	95.65%
2016	3000	2716	-284	9.46%	90.54%
2017	3060	3042	-18	0.60%	99.40%
2018	3480	3221	-260	7.46%	92.54%

Na Figura 16, temos o resultado do modelo *Random Forest*. O gráfico apresenta os valores reais e os previstos durante o período de 2018 para as 27 cidades que foram estudadas neste trabalho. O modelo conseguiu acompanhar a tendência da produtividade real apenas a partir de 2015.

Figura 16 – Previsão da Produtividade Tangará da Serra - RF

Fonte: Elaborada pelo autor.

4.4 Modelo Extreme Gradient Boosting

Na tabela 13 podemos observar que algumas cidades (Cáceres, Confresa e Sorriso) tiveram uma diferença maior entre os valores reais e previstos. Podemos considerar que existem fatores externos tais como, necessidade hídrica da planta em cada estágio fenológico, correção artificial da capacidade de água disponível, dentre outros podem ter impactado no modelo de previsão proposto. Desta forma, o modelo não conseguiu detectar tais cenários, dado que estas variáveis não foram consideradas na análise inicialmente. Ainda assim, o modelo apresentou no melhor caso acurácia de 99.96% para a cidade de Pontes e Lacerda e no pior caso acurácia 83.49% para a cidade de Cáceres.

Tabela 13 – Previsão da Produtividade - *Extreme Gradient Boosting* - 2018

Cidade	Produtividade Kg/ha				
	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
Água Boa	3300	3157	143	4.33%	95.67%
Alto Taquari	3480	3206	274	7.88%	92.12%
Cáceres	4080	3406	674	16.51%	83.49%
Campo Verde	3300	3241	59	1.79%	98.21%
Canarana	3300	3321	-21	0.64%	99.36%
Comodoro	3360	3297	63	1.88%	98.12%
Confresa	3600	3254	346	9.60%	90.40%
Diamantino	3240	3124	116	3.59%	96.41%
Gaucha do Norte	3048	3050	-2	0.07%	99.93%
Guarantã	3000	3017	-17	0.58%	99.42%
Guiratinga	3300	3205	95	2.88%	97.12%
Itiquira	3120	3387	-267	8.55%	91.45%
Juara	3100	3289	-189	6.10%	93.90%
Matupá	3100	3286	-186	6.01%	93.99%
Nova Ubiratã	3480	3433	47	1.36%	98.64%
Nova Xavantina	3000	3008	-8	0.27%	99.73%
Novo Mundo	3280	3279	1	0.04%	99.96%
Paratininga	3120	3023	97	3.11%	96.89%
Pontes e Lacerda	3520	3519	1	0.04%	99.96%
Poxoreu	3180	3292	-112	3.51%	96.49%
Querência	3360	3556	-196	5.84%	94.16%
Rondonópolis	3380	3126	254	7.50%	92.50%
São Felix do Xingu	3310	3250	60	1.82%	98.18%
São José do Rio Claro	3250	3125	125	3.83%	96.17%
Sinop	3480	3402	78	2.23%	97.77%
Sorriso	3720	3334	386	10.37%	89.63%
Tangará da Serra	3480	3127	353	10.14%	89.86%

Na Figura 17, temos o resultado do modelo *Extreme Gradient Boosting*. O gráfico apresenta valores reais e os previstos durante o período de 2018 para as 27 cidades que foram utilizados no estudo. O modelo não teve boa acurácia na previsão da produtividade para a cidade de Cáceres, assim como aconteceu com os demais modelos. A maior diferença entre o rendimento real e o previsto foi de 16.51% para a cidade de Cáceres, seguido pela diferença de 10.14% para a cidade de Sorriso e 10.37% para a cidade de Tangará da Serra. Foi previsto produtividade de 3406 kg/ha para Cáceres, porém a produtividade real foi de 4080kg/ha.

Figura 17 – Previsão da Produtividade por Cidades - XGB - Ano 2018

Fonte: Elaborada pelo autor.

Na tabela 14 podemos observar que não houve diferenças significativas entre os valores previstos pelo algoritmo *Extreme Gradient Boosting* e os valores reais. O modelo apresentou acurácia 93.31% no pior caso (Ano 2017) e de 99.95% no melhor caso (Ano 2018).

Tabela 14 – Previsão da Produtividade - Tangará da Serra - XGB

Produtividade Kg/ha					
Ano	Real	Previsto	Dif. Kg/ha	Dif. %	Acurácia
2010	3000	3161.56958	161.5695801	5.39%	94.61%
2011	3120	3074.644287	-45.35571289	1.45%	98.55%
2012	3180	2982.222656	-197.7773438	6.22%	93.78%
2013	3060	3134.293701	74.29370117	2.43%	97.57%
2014	2980	3047.769775	67.76977539	2.27%	97.73%
2015	3180	3264.834961	84.83496094	2.67%	97.33%
2016	3000	2908.144287	-91.85571289	3.06%	96.94%
2017	3060	3264.834961	204.8349609	6.69%	93.31%
2018	3480	3478.321777	-1.678222656	0.05%	99.95%

Na Figura 18, temos o resultado do modelo *Extreme Gradient Boosting*. O gráfico apresenta valores reais e os previstos durante o período de 2018 para as 27 cidades que foram utilizados para teste. O modelo conseguiu acompanhar a tendência da produtividade real a partir de 2014.

Figura 18 – Previsão da Produtividade Tangará da Serra - XGB

Fonte: Elaborada pelo autor.

4.5 Comparação dos Erros entre os modelos - MAPE

Na tabela 15 podemos observar os valores do MAPE para os respectivos algoritmos. O *Extreme Gradient Boosting* apresentou a melhor média dentre os três modelos analisados.

Tabela 15 – Comparativo dos Erros dos Modelos - Tangará da Serra

Erros MAPE dos três modelos			
Ano	MLP	Random Forest	XG Boosting
2010	6.42%	6.19%	5.39%
2011	6.12%	1.21%	1.45%
2012	14.64%	6.91%	6.22%
2013	0.04%	2.98%	2.43%
2014	1.68%	5.46%	2.27%
2015	5.23%	4.35%	2.67%
2016	6.09%	9.46%	3.06%
2017	1.54%	0.60%	6.69%
2018	7.18%	7.46%	0.05%
Média	5.44%	4.96%	3.36%

Na Figura 19, temos a apresentação gráfica da distribuição dos erros dos modelos aplicados na predição da produtividade da soja para a cidade de Tangará da Serra no período 2010-2018. Podemos observar que o MAPE do *Multilayer Perceptron* do ano 2013 foi muito baixo 0.4%, enquanto que em 2012, 14.64%, foi o mais alto de todo o período analisado, sendo portanto o de maior amplitude.

Figura 19 – Distribuição dos Erros- MAPE - Tangará da Serra

Fonte: Elaborada pelo autor.

A tabela 16 apresenta os valores do MAPE para os respectivos modelos aplicados à predição da produtividade das 27 cidades do conjunto de dados. Novamente o *Extreme Gradient Boosting* apresentou a melhor média, 4.46% e o menor erro, 0.04%, dentre os três modelos analisados. O modelo *Multilayer Perceptron* apresentou MAPE de 0.05% no melhor caso e 17.48% no pior caso, o modelo *Random Forest* apresentou MAPE de 1.52% no melhor caso e 14.12% no pior caso e o modelo *Extreme Gradient Boosting* apresentou MAPE de 0.04% no melhor caso e 16.51% no pior caso. O MAPE do modelo *Multilayer Perceptron* apresentou amplitude de 17.43, do modelo *Random Forest* 12.6 e do modelo *Extreme Gradient Boosting* 16.48.

Tabela 16 – Comparativo dos Erros dos Modelos - 2018

Erros MAPE - 27 cidades - Ano 2018			
Ano	MLP	Random Forest	XG Boosting
Água Boa	5.86%	4.55%	4.33%
Alto Taquari	8.86%	4.77%	7.88%
Cáceres	17.48%	14.12%	16.51%
Campo Verde	2.59%	1.68%	1.79%
Canarana	3.70%	4.14%	0.64%
Comodoro	0.66%	4.82%	1.88%
Confresa	7.08%	10.43%	9.60%
Diamantino	4.51%	3.70%	3.59%
Gaucha do Norte	0.68%	2.85%	0.07%
Guarantã	0.29%	3.31%	0.58%
Guiratinga	2.57%	2.18%	2.88%
Itiquira	9.30%	8.53%	8.55%
Juara	7.68%	4.13%	6.10%
Matupá	6.36%	4.03%	6.01%
Nova Ubiratã	0.05%	4.76%	1.36%
Nova Xavantina	1.15%	3.12%	0.27%
Novo Mundo	1.43%	1.59%	0.04%
Paratininga	3.48%	1.88%	3.11%
Pontes e Lacerda	0.16%	5.33%	0.04%
Poxoreu	5.15%	1.76%	3.51%
Querência	5.51%	4.49%	5.84%
Rondonópolis	0.36%	7.75%	7.50%
São Felix do Xingu	11.51%	3.76%	1.82%
São José do Rio Claro	4.31%	1.52%	3.83%
Sinop	5.05%	7.04%	2.23%
Sorriso	6.62%	10.02%	10.37%
Tangará da Serra	7.18%	7.99%	10.14%
Média	4.80%	4.97%	4.46%
Desvio Padrão	0.040	0.031	0.041
Variância	0.002	0.001	0.002

4.6 Comparação da acurácia dos Modelos de Predição

A figura 20 apresenta o desempenho de predição dos modelos *Multilayer Perceptron*, *Random Forest*, *Extreme Gradient Boosting* comparados à produtividade real para as 27 cidades. Podemos observar que nas cidades de Cáceres, Confresa, Itiquira, Sorriso e Tangará os 03 modelos não tiveram bom desempenho na predição. Diversos fatores externos podem ter interferido na predição, tais como: Capacidade de Água Disponível, uso de Irrigação, dentre outros.

Figura 20 – Comparação do desempenho dos modelos na previsão da Produtividade das 27 cidades

Fonte: Elaborada pelo autor.

A figura 21 apresenta o desempenho de predição dos modelos baseados em aprendizado de máquina: *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, comparados à produtividade real para a cidade de Tangará da Serra durante o período 2010-2018. Podemos observar que em 2012 os três modelos não tiveram bom desempenho na predição, possivelmente influenciado por variáveis externas não tratadas no modelo, tais como: irrigação, cultivares plantadas, período do plantio, dentre outras. Apesar disso, o modelo *Multilayer Perceptron* acompanha melhor a tendência da produtividade real: cresce em 2011 e 2012, decresce em 2013 e 2014, cresce em 2015, decresce em 2016, cresce em 2017 e 2018.

Figura 21 – Comparação do desempenho dos modelos - Previsão Tangará da Serra

Fonte: Elaborada pelo autor.

A figura 22 compara o desempenho do modelo agrometeorológico de estimativa de produtividade com a produtividade real e com os modelos baseados em aprendizado de máquina: *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*. Podemos observar que o modelo agrometeorológico não consegue simular a tendência da produtividade real. Em 2013, o modelo previu uma queda que não ocorreu. Depois ele prevê um aumento em 2016, mas foi queda nos dados reais. Finalmente, ele prevê redução em 2017 e 2018, mas houve aumento.

Figura 22 – Comparação do desempenho dos quatro modelos - Previsão Tangará da Serra

Fonte: Elaborada pelo autor.

CONCLUSÃO

Neste trabalho foram analisados comparativamente aplicações dos algoritmos de *Machine Learning* aplicados a previsão da produtividade da soja para 27 cidades representando os 3 biomas do estado do Mato Grosso. Foram avaliados *Multilayer Perceptron*, *Random Forest* e *Extreme Gradient Boosting*, posteriormente foram comparados com o modelo Agrometeorológico de estimativa de produtividade proposto por (DOORENBOS; KASSAN, 1979).

Nos resultados observou-se que o modelo *Extreme Gradient Boosting*, apresentou desempenho superior aos modelos *Multilayer Perceptron* e *Random Forest*. Isto corrobora com o crescente aumento nos últimos meses das implementações deste algoritmo, assumindo o posto de modelo de predição mais utilizado entre os vencedores de competições em plataformas como Kaggle. O modelo *Multilayer Perceptron* foi o que melhor acompanhou a tendência da produtividade real para a cidade de Tangará da Serra no período 2010-2018: cresce em 2011 e 2012, decresce em 2013 e 2014, cresce em 2015, decresce em 2016, cresce em 2017 e 2018.

Tanto *Extreme Gradient Boosting* quanto *Random Forest* estimaram com baixo *score* as variáveis mais importantes do modelo preditivo. O modelo baseado no *Extreme Gradient Boosting* conseguiu prever a produtividade com acurácia de até 99.96% no melhor caso e no pior caso com 83.49% de acurácia. A média da acurácia da predição da produtividade das 27 cidades foi de 95.54% para o *Extreme Gradient Boosting*, 95.03% para *Random Forest* e 95.20% para *Multilayer Perceptron*(MLP). Assim podemos observar que os três modelos tiveram ótima performance na predição da produtividade da soja utilizando dados climáticos e de solo.

Os modelos *Multilayer Perceptron* e *Random Forest*, embora tenham apresentado desempenhos inferior em relação aos valores previstos comparado ao *Extreme Gradient Boosting*, os modelos conseguiram realizar estimativas acima de 95% de acurácia.

O desempenho da Rede Neural *Multilayer Perceptron* foi penalizado especialmente em função da dificuldade de calibração dos seus pesos. A ausência de informações relativas às cultivares da soja das amostras, precisão da data do plantio, identificação dos respectivos estádios

fenológicos de cada cultivar, mapeamento da necessidade hídrica da cultura em cada estágio fenológico e o possível emprego de irrigação artificial, informações físico/químicas do solo, dentre outros podem ter prejudicado a acurácia do modelo. No entanto, o modelo *Multilayer Perceptron* se mostrou muito flexível para tratar novas variáveis, o que facilita a inclusão de tais informações no modelo quando disponíveis para coleta.

Assim, o modelo de aprendizado de máquina, baseado no algoritmo *Extreme Gradient boosting* foi escolhido como o melhor modelo. Importante salientar que tanto *Extreme Gradient Boosting* quanto *Random Forest* são algoritmos muito mais rápidos que o *Multilayer Perceptron*, além de consumirem menos recursos computacionais. No entanto, o modelo de rede neural está preparado para receber informações de previsão climática de algoritmos dedicados à esse fim e com base nestas informações e informações do solo fazer a previsão da produtividade da cultura alvo. Neste trabalho utilizamos a cultura da soja, em função da disponibilidade de informações, todavia, o modelo também foi testado com dados da cultura do milho, porém, em menor escala. Desta forma, entendemos que o modelo suporta a predição da produtividade de outras culturas.

Uma contribuição deste trabalho está na análise comparativa do uso dos algoritmos de aprendizado de máquina para prever com maior acurácia a produtividade da cultura da soja, possibilitando com isso que os produtores/agricultores consigam melhorar a tomada de decisão, contribuir com questões agrônômicas, melhorar o planejamento e estratégia de políticas nacionais de alimentos e ajuda no controle da crise global de alimentos.

Em trabalhos futuros, é sugerido o uso de redes convolucionais no intuito de reduzir as atividades de pré-processamento, o mapeamento da necessidade hídrica para cada estágio fenológico, além da análise e possibilidade de inclusão de novas variáveis externas que possam influenciar os modelos.

Em síntese, este projeto foi muito relevante, pois, possibilitou ao autor interagir com produtores de soja, cooperativas de produtores e processadoras de grãos, utilização de dados reais no estudo e avaliar diferentes abordagens de predição. Além dos desafios anteriormente mencionados no desenvolvimento deste projeto, há desafios culturais a serem superados junto e em conjunto com a comunidade de produtores.

REFERÊNCIAS

- BARROS, M. F. de. **Análise e Previsão de Séries Temporais Utilizando Amortecimento Exponencial com Múltiplos Ciclos e Técnicas de Simulação na Produção de Energia Eólica**. Tese (Doutorado) — PUC-Rio, 2015. Citado na página 48.
- BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G. Configuração de redes neurais artificiais para estimação do volume de árvores. **Ciência da Madeira**, Braz. J. Wood Sci., v. 05, p. 58–67, 2014. Citado na página 27.
- BURGUENO, J.; CROSSA, J.; CORNELIUS, P. L.; YANG, R.-C. Using factor analytic models for joining environments and genotypes without crossover genotype environment interaction. **Crop Sci**, Crop Sci 2008, v. 48, p. 1291–1305, 2008. Citado na página 25.
- BURGUENO, J.; CROSSA, J.; COTES, J. M.; VICENTE, F. S.; DAS, B. Prediction assessment of linear mixed models for multienvironment trials. **Crop Sci**, Crop Sci 2011, v. 51, p. 944–954, 2011. Citado na página 25.
- CAMARGO, A. P. Contribuição para a determinação da evapotranspiração potencial no estado de são paulo. **Bragantia**, Bragantia, v. 21, n. 12, p. 163–213, 1962. Citado na página 33.
- CAMARGO, M. B. P.; BRUNINI, O.; MIRANDA, M. A. C. Modelo agrometeorológico para estimativa da produtividade para a cultura da soja no estado de são paulo. **SciELO**, SciELO, 1986. Citado nas páginas 32 e 33.
- CONAB. **Conab estima safra recorde de grãos para safra 2018/2019**. 2018. <<http://agenciabrasil.ebc.com.br/economia/noticia/2019-08/conab-estima-safra-recorde-de-graos-para-safra-20182019>>. Accessed: 2019-09-29. Citado na página 23.
- CROSSA, J.; CORNELIUS, P. L. Sites regression and shifted multiplicative model clustering of cultivar trial sites under heterogeneity of error variances. **Crop Sci**, Crop Sci 1997, v. 37, p. 406–415, 1997. Citado na página 25.
- CROSSA, J.; CORNELIUS, P. L.; SAYRE, K.; ORTIZ-MONASTERIO R. AND IVÁN, J. A shifted multiplicative model fusion method for grouping environments without cultivar rank change. **Crop Sci**, Crop Sci 1995, v. 35, p. 54–62, 1995. Citado na página 25.
- DELACY, I. H.; BASFORD, K. E.; COOPER, M.; BULL, J. K.; MCLAREN, C. G. Analysis of multi-environment trials, an historical perspective. **Plant Adaptation and Crop Improvement**, Wallingford: CABI, 1996. Citado na página 25.
- DOORENBOS, J.; KASSAN, A. Yield response to water. **FAO Irrigation and Drainage Paper**, FAO, n. 33, p. 193, 1979. Citado nas páginas 25, 26, 32, 38, 43, 49, 51, 54 e 71.
- DRUMMOND, S. T.; SUDDUTH, K. A.; JOSHI, A.; BIRRELL S. J. AND KITCHEN, N. R. Statistical and neural methods for site specific yield prediction. **ASAE**, ASAE, v. 46, 2013. Citado na página 25.

- _____. Soybean varieties portfolio optimisation based on yield prediction. **Comput. Electron. Agric.**, j. compag, v. 127, p. 467–474, 2016. Citado na página 25.
- FAO. **The state of food insecurity in the world 2014**. 2015. <<http://www.fao.org/publications/sofi/en/>>. Accessed: 2019-11-15. Citado na página 23.
- FEHR, W.; CAVINESS, C. Stages of soybean development. **Special Report**, Iowa State University of Science and Technology, n. 80, p. 11, 1977. Citado na página 38.
- HORIE, T.; YAJIMA, M.; NAKAGAWA, H. Yield forecasting. **Agric. Syst.**, PubMed, v. 40, p. 211–236, 1992. Citado na página 24.
- IBGE, S. **Produção Agrícola Municipal**. 2019. <<https://sidra.ibge.gov.br/tabela/1612>>. Accessed: 2019-09-20. Citado na página 41.
- INMET. **Balanço Hídrico de Cultivo**. 2019. <<http://sisdagro.inmet.gov.br/sisdagro/app/monitoramento/bhc>>. Accessed: 2019-09-29. Citado na página 37.
- JEONG, J. H.; RESOP, J. P.; MUELLER, N. D.; FLEISHER, D. H.; YUN, E. E. B. K.; TIMLIN, K.-M. S. D. J.; GERBER, V. R. R. J. S.; KIM, S.-H. Random forests for global and regional crop yield predictions. **Plos One**, 2016. Accessed: 2019-09-30. Citado nas páginas 29 e 30.
- KARTHIKEYAN, R.; GOWTHAMI, M.; ABHISHHEK, A.; KARTHIKEYAN, P. Implementation of effective crop selection by using the random forest algorithm. **International Journal of Engineering and Technology**, International Journal of Engineering and Technology, v. 7, p. 287–290, 2018. Citado na página 30.
- LEITE, H. G.; OLIVEIRA, F. H. T. Statistical procedure to test the identity of analytical methods. **Communications in Soil Science and Plant Analysis**, Communications in Soil Science and Plant Analysis, v. 33, n. 7/8, p. 1105–118, 2002. Citado na página 28.
- LIU J., G. C. E.; TIAN, L. A neural network for setting target corn yields. **ASAE**, ASAE, v. 44, 2013. Citado na página 25.
- OECD. **Glossary of Statistical Terms**. 2003. <<https://stats.oecd.org/glossary/detail.asp?ID=21>>. Accessed: 2019-11-10. Citado na página 48.
- OLIVA, G. **Forecasting U.S. Corn Yield Using Machine Learning**. 2016. Citado nas páginas 30 e 31.
- ONU. United nations, department of economic and social affairs the united nations, population division, population estimates and projections section. **ONU**, ONU, 2012. Citado na página 23.
- PORTARIA 13, M. d. A. Portaria nº 13, de 20 de julho de 2017. **Ministério da Agricultura, Pecuária e Abastecimento**, <http://www.agricultura.gov.br/legislacao/portarias>, 2017. Citado na página 37.
- ROMERO, J. R.; RONCALLO, P. F.; AKKIRAJU, P. C.; PONZONI, I.; ECHENIQUE, V. C.; CARBALLIDO, J. A. Using classification algorithms for predicting durum wheat yield in the province of buenos aires. **Comput. Electron. Agric.**, j. compag, v. 96, p. 173–179, 2016. Citado na página 25.
- ROSE, D. C.; J., S. W.; CAROLINE, P.; MATT, L.; MICHAEL, W.; CAROL, M.; SUSAN, T.; CHARLES, F.; TATSUYA, A.; V., D. L. Decision support tools for agriculture: Towards effective design and delivery. **Agricultura System**, v. 40, p. 165–174, 2016. Citado na página 23.

SAAED, K.; WANG, L. Random forests for global and regional crop yield predictions. **Front. Plant Sci**, v. 10, n. 621, p. 51–59, 2019. Citado nas páginas 28 e 29.

SAATH, K. C. de O.; FACHINELLO, A. L. **Crescimento da demanda mundial de alimentos e restrições do fator terra no Brasil**. 2018. <-https://http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-20032018000200195#aff1>. Accessed: 2019-11-16. Citado nas páginas 23 e 24.

SAYWER, D. R. Fluxo e refluxo da fronteira agrícola no brasil: ensaio de interpretação estrutural e espacial. **Revista Brasileira de Estudos de População**, Revista Brasileira de Estudos de População, v. 1, p. 3–34, 2013. Citado na página 24.

STEWART J. I.; HAGAN, R. M.; PRUITT, W. Production functions and predicted irrigation programs for principal crops as required for water resources planning and increased water use efficiency. **Final report. Washington DC**, Department of Interior, p. 80, 1976. Citado na página 32.

THORNTHWAITE, C.; MATHER, J. The water balance. **Publications in Climatology**, Center-ton, Drexel Institute of Technology - Laboratory of Climatology, v. 8, n. 1, p. 104, 1955. Citado nas páginas 31 e 33.

YIQING, C.; MOORE, K.; PELLEGRINI, A.; ELHADDAD, A.; LESSEL, J.; TOWNSEND, C.; SOLAK, H.; SEMRET, N. A scalable tree boosting system. **KDD**, KDD, 2016. Citado na página 31.