UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA AGRÍCOLA

CAROLINA LOBELLO LORENSINI

METODOLOGIA PARA CLASSIFICAÇÃO DA APTIDÃO AGRÍCOLA DE MUNICÍPIOS

CAMPINAS
2019

CAROLINA LOBELLO LORENSINI

METODOLOGIA PARA CLASSIFICAÇÃO DA APTIDÃO AGRÍCOLA DE MUNICÍPIOS

Dissertação apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para obtenção do título de Mestra em ENGENHARIA AGRÍCOLA na área de concentração Água e Solo.

Orientador: Prof. Dr. Stanley Robson de Medeiros Oliveira
Coorientador: Dr. Daniel de Castro Victoria

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DISSERTAÇÃO DEFENDIDA PELA ALUNA CAROLINA LOBELLO LORENSINI, E ORIENTADA PELO PROF. DR. STANLEY ROBSON DE MEDEIROS OLIVEIRA.

CAMPINAS
2019

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Lorensini, Carolina Lobello, 1991-
L886m Metodologia para classificação da aptidão agrícola de municípios / Carolina Lobello Lorensini. – Campinas, SP : [s.n.], 2019.

Orientador: Stanley Robson de Medeiros Oliveira.
Coorientador: Daniel de Castro Victoria.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Mineração de dados. 2. Solos - Aptidão agrícola. 3. Análise de algoritmos. 4. Sistemas de Informação Geográfica. 5. Aprendizado de máquina. I. Oliveira, Stanley Robson de Medeiros. II. Victoria, Daniel de Castro. III. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: Methodology for classification of cities's agricultural suitability
Palavras-chave em inglês:
Data mining
Soils - Agricultural suitability
Analysis of Algorithms
Geographic Information System
Machine Learning
Área de concentração: Água e Solo
Titulação: Mestra em Engenharia Agrícola
Banca examinadora:
Stanley Robson de Medeiros Oliveira
Célia Regina Grego
Jurandir Zullo Junior
Data de defesa: 21-02-2019
Programa de Pós-Graduação: Engenharia Agrícola

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0003-1592-0670
- Currículo Lattes do autor: http://lattes.cnpq.br/1790698085034855

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Carolina Lobello Lorensini**, aprovada pela Comissão Julgadora em 21 de Fevereiro de 2019, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

FEAGRI

_____

Dr. Stanley Robson de Medeiros Oliveira – Presidente e Orientador(a)

FEAGRI/UNICAMP

_____

Dra. Célia Regina Grego – Membro Titular

Faculdade de
Engenharia Agrícola
Unicamp

_____

Prof. Dr. Jurandir Zullo Junior – Membro Titular

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no processo de vida acadêmica da discente.

AGRADECIMENTOS

Agradeço aos meus pais, Paulo e Elisete, por tudo que me ensinaram, pela compreensão em todos os momentos, por lutarem para minha formação e pelo amor incondicional. Às minhas irmãs, Laura e Elisa, pela amizade, companheirismo e por serem tão presentes na minha vida mesmo com a distância. Ao meu avô, Valdir, por toda sabedoria e conselhos preciosos. Sem vocês não teria chegado até aqui.

Ao Prof. Dr. Stanley Robson de Medeiros Oliveira, meu orientador, pelo suporte, correções, atenção em me acompanhar em cada etapa e oportunidade de crescimento profissional e pessoal.

Ao meu coorientador, Dr. Daniel de Castro Victoria, pelos ensinamentos para desenvolver esse trabalho, pela paciência, dedicação e disposição em ajudar ao longo dos anos trabalhando juntos.

Aos amigos e colegas, pelas discussões, troca de conhecimentos e pelo equilíbrio entre a vida social e acadêmica.

A todos os professores, pesquisadores e funcionários da Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas pelo suporte e ajuda, direta ou indiretamente.

À minha família e amigos de Catanduva e Piracicaba, pelo carinho, apoio e pela compreensão nos meus momentos ausentes.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A classificação da aptidão agrícola das terras relaciona o recurso solo e o nível tecnológico com a produtividade, apoiando a gestão territorial no agronegócio, buscando a conservação e a sustentabilidade agroambiental. Essa metodologia enquadra-se na modalidade de classificações técnicas ou interpretativas, nas quais as terras são agrupadas de acordo com suas potencialidades, relacionadas com o tipo de utilização que se quer dar. Mapas de aptidão agrícola fornecem informações objetivas que podem ser aplicadas no planejamento agrícola e na avaliação do uso das terras, permitindo apontar áreas com uso adequado, sobreutilizadas ou que suportariam a intensificação do uso. Em particular, a região denominada MATOPIBA (Maranhão, Tocantins, Piauí e Bahia), ganhou importância no setor agropecuário brasileiro, apresenta forte crescimento desde os anos 2000 e atualmente é considerada como fronteira em expansão e intensificação do Brasil. Neste contexto, o objetivo deste trabalho foi desenvolver uma metodologia baseada em técnicas de mineração de dados para classificar municípios com aptidão agrícola, na região do MATOPIBA. Para isto, foram utilizados dados sociais (IDHM, taxa de analfabetismo com 25 anos ou mais, renda per capita e a porcentagem da população que trabalham no setor agropecuário com 18 anos ou mais, PIB e fluxo migratório), econômicos (área plantada, quantidade produzida e rendimento médio) e físicos (precipitação, temperatura, umidade relativa, declividade, relevo e tipo de solo) dos municípios, no período de 17 anos, entre os anos de 2000 a 2016. A interação recursos naturais e sociedade é importante para a preservação do meio ambiente e constitui-se em um desafio não considerado em metodologias anteriores. Para selecionar os atributos mais relevantes do conjunto de dados foram utilizados os seguintes métodos de seleção de atributos: Qui-Quadrado, CFS (Correlation Feature Selection), Ganho de Informação e Wrapper. Foram gerados modelos preditivos por meio de indução de árvores de decisão (J48), Random Forest e SVM (Support Vector Machine) juntamente com métodos de seleção de atributos. Os resultados encontrados revelaram o potencial das técnicas de mineração de dados para a classificação de áreas com aptidão agrícola, resultando em uma metodologia simplificada, através de ferramentas de Sistemas de Informações Geográficas (SIGs) e dados obtidos pelo IBGE, com isso, não há necessidade de trabalho de campo e uma equipe grande reduzindo custo, tempo e esforço em projetos de gestão territorial.

Palavras-Chave: Mineração de dados, seleção de atributos, MATOPIBA, sustentabilidade

ABSTRACT

The classification of agricultural land suitability is based on technological level, which supports agribusiness territorial management. It also relates natural resource with productivity, pursuing conservation and agri-environmental sustainability. This methodology fits in the modality of technical or interpretative classifications, where lands are not only grouped according to their potential, but also related to the type of land used that needs to be given. Agricultural suitability maps provide objective information that can be applied in agricultural planning and land use assessment, enabling areas to be properly used, over-utilized or that would support intensification of usage. The region called MATOPIBA (Maranhão, Tocantins, Piauí and Bahia), in particular, gained importance in the Brazilian agricultural sector. The region has experienced strong growth since the 2000s and its frontier is currently considered on expansion and intensification. In this context, the goal of this research is to develop a methodology based on data mining techniques to classify cities with agricultural suitability in the region of MATOPIBA. In order to develop such a methodology, it has been taken in consideration the following data: social (IDHM, illiteracy rate with 25 years or more, per capita income, the percentage of the population working in the agricultural sector within 18 years or more, GDP and Migratory Flow), economy (planted area, quantity produced and average yield) and physical (rainfall, temperature, relative humidity, slope, relief and soil type) of each city in the period of 17 years, from 2000 until 2016. The interaction between natural resources and society is important for the preservation of the environment and constitutes a challenge not considered in previous methodologies. With the purpose of selecting the most relevant attributes of the dataset, attribute selection techniques such as Chi-Square, CFS, Information Gain and Wrapper were applied. Additionally, the predictive models were generated through the induction of decision trees (J48), Random Forest and Support Vector Machines (SVM) along with attribute selection methods. The results showed the potential of data mining techniques for the classification of areas with agricultural suitability and a simplified methodology was elaborated by using tools of a Geographic Information System (GIS) and data obtained by IBGE. In doing so, there is no need of field work nor a large team, resulting in cost, time and effort reduction in territorial management projects.

Keyword: Data mining, attribute selection, Matopiba, sustainability

LISTA DE ILUSTRAÇÕES

Figura 1 - O processo de descoberta de conhecimento em bancos de dados (KDD). 15
Figura 2 - Tarefas de mineração de dados 16
Figura 3 - Exemplo de uma árvore de decisão 18
Figura 4 - Exemplo da geração de uma floresta aleatória. 19
Figura 5- Esquema da classificação SVM 20
Figura 6 - Configuração básica de um SIG 24
Figura 7 Aquisição de dados 27
Figura 8 - Localização da área de estudo 35
Figura 9 - Estações Meteorológicas dos estados Maranhão, Tocantins, Piauí e Bahia 37
Figura 10 - Classes dos municípios 43
Figura 11 - Algoritmo da técnica Random Forest (BREIMAN, 2001). 44
Figura 12 - Mapa dos municípios com aptidão agrícola, da região do MATOPIBA, com o classificador Árvore de decisão e seleção de atributos Wrapper. 54
Figura 13 - Mapa dos municípios com aptidão agrícola, mesorregião do Extremo Oeste Baiano, gerado a partir do classificador J48 e o método Wrapper. 55

LISTA DE TABELAS

Tabela 1- Simbologia correspondente às classes de aptidão agrícola das terras. 33
Tabela 2 - Classes de declividade 38
Tabela 3 - Agrupamento das classes de aptidão agrícola 40
Tabela 4 - Quadro-guia de avaliação da aptidão agrícola das terras 41
Tabela 5 - Matriz de confusão 45
Tabela 6 - Desempenho de modelos de classificação a partir do Kappa 47
Tabela 7 - Variáveis escolhidas pelos métodos de seleção de atributos. 48
Tabela 8 - Classificação dos atributos para os métodos Ganho de Informação e Qui-quadrado. 51
Tabela 9 - Classificação dos atributos para os métodos Ganho de Informação e Qui-quadrado. 52
Tabela 10 - Matriz de confusão da variável resposta “aptidão agrícola” utilizando o algoritmo J48 com método de seleção de atributos Wrapper para validação do modelo. 56
Tabela 11 - Comparação das técnicas de seleção de atributos para o algoritmo Random Forest 57
Tabela 12- Matriz de confusão da variável resposta “aptidão agrícola” utilizando o algoritmo Random Forest com método de seleção de atributos CFS para validação do modelo. 58
Tabela 13 - Comparação das técnicas de seleção de atributos para o algoritmo SVM 58
Tabela 14 - Comparação dos dados utilizados nas metodologias 60

SUMÁRIO

1. INTRODUÇÃO 11
2. HIPÓTESE 14
3. OBJETIVO 14
3.1 Objetivo Geral 14
3.2 Objetivos Específicos 14
4. REVISÃO BIBLIOGRÁFICA 15
4.1 Mineração de Dados 15
4.1.1 Descrição das Técnicas de Mineração de Dados Utilizadas 17
4.1.2 Métodos de seleção e atributos 20
4.1.2.1 Qui-quadrado (χ^2) 21
4.1.2.2 CFS (*Correlation Based Feature Selection*) 21
4.1.2.3 Ganho de informação 22
4.1.2.4 Wrapper 23
4.2 Geotecnologias 23
4.3 Sensoriamento Remoto 26
5. MATERIAL E MÉTODOS 35
5.1 Área de Estudo 35
5.2 Base de Dados 36
5.2.1 Dados Econômicos 36
2.2.2 Dados Sociais 36
2.2.3 Dados Físicos 37
5.3 Preparação dos Dados 39
5.3.1 Especificação do Atributo Meta 39
5.4 Modelagem 43
5.5 Validação 44
6. RESULTADOS 48
6.1 Seleção de Atributos 48
6.2 Classificação 52
6.3 Comparação da metodologia proposta e a metodologia de Ramalho Filho e Beek (1995) 59
7. CONCLUSÃO 62
8. REFERÊNCIAS BIBLIOGRÁFICAS 64

1. INTRODUÇÃO

A questão ambiental vem assumindo grande importância no contexto nacional e internacional, com envolvimento direto das instituições de pesquisa e da sociedade em geral. Por outro lado, é percebido que alguns aspectos se tornam cada vez mais explícitos e necessários. Dentre eles, com grande relevância, surge a necessidade de metodologias que forneçam subsídios para o planejamento e tomada de decisões mais precisas, adequadas e ágeis, porém, com visão mais efetiva quanto à incorporação da componente ambiental neste processo (PEREIRA; LOMBARDI NETO, 2004).

O Sistema de Avaliação da Aptidão Agrícola é uma metodologia de classificação de terras para uso agrícola (PEREIRA et al., 2006). No Brasil, iniciou-se na década de sessenta (BENNEMA et al., 1964), para classificar o potencial das terras para agricultura tropical. Essa metodologia pode ser classificações técnicas ou interpretativas, nas quais as terras são agrupadas de acordo com suas potencialidades, relacionadas com o tipo de utilização que se quer dar (PEREIRA et al., 2006).

A classificação da aptidão agrícola das terras, baseada no nível tecnológico, apoia a gestão territorial do agronegócio por relacionar o recurso natural solo com a produtividade, buscando a conservação e a sustentabilidade agroambiental (VALLADARES et al., 2007). Mapas de aptidão agrícola podem ser aplicadas no planejamento agrícola e na avaliação do uso das terras, permitindo apontar áreas com uso adequado, subutilizadas ou que suportariam a intensificação do uso (RAMALHO FILHO; BEEK, 1995; VALLADARES et al., 2007).

A expansão de áreas agrícolas, as mudanças na localização espacial dos cultivos e a intensificação do uso das terras são fontes potenciais de degradação do ambiente, principalmente se não respeitadas as potencialidades e limitações da capacidade produtiva dos solos (QUARTAROLI et al., 2006).

O uso inadequado do solo vem acarretando diversos problemas ambientais, tais como, o comprometimento de fontes e mananciais de água, degradação de *habitats* naturais, erosão, enchentes e assoreamento dos rios. Estes estão diretamente ligados a um mau gerenciamento do espaço, que por estar desprovido de algum critério técnico ou projeto ambiental, tal como um zoneamento do solo, não respeitam a capacidade de suporte da natureza (LEITE et al., 2012).

Para que ocorra o desenvolvimento de políticas agrícolas no Brasil, é necessário conhecimento sobre as potencialidades de uso do território e seus recursos naturais, bem como

o mapeamento do uso atual das terras. Assim, é possível avaliar a adequabilidade do uso atual e avaliar alternativas para desenvolvimento futuro. O mapeamento das áreas agrícolas também é de grande interesse para o monitoramento e as estimativas de safras. As estimativas de safras são importantes para a sociedade, pois fornecem informações ao mercado sobre a oferta de produtos agrícolas. Por conta da alta escala temporal dos cultivos agrícolas é necessário o uso correto de tecnologias para realizar a análise e interpretação de dados (ADAMI et al., 2005; BERNARDES et al., 2011).

A análise do uso e cobertura da terra, por meio de informações obtidas pelo sensoriamento remoto, é de grande utilidade ao planejamento e administração da ocupação do solo, além de possibilitar avaliar, monitorar a preservação de áreas de vegetação natural e também áreas submetidas ao processo de antropismo (RODRIGUES, 2000 citado por DAINESE, 2001).

Considerando que o Brasil é um país agrícola, onde o agronegócio é responsável por 25% do PIB nacional (PLANO Agrícola e Pecuário, 2015), e os municípios crescem por conta da atividade agrícola, o uso de métodos e técnicas de mineração e dados e sensoriamento remoto podem apresentar uma contribuição significativa em atividades de classificação do uso de terra e monitoramento de safras.

O sensoriamento remoto é uma tecnologia essencial no monitoramento agrícola, para estimativas de parâmetros biofísicos de interesse utilizados em modelos agrometeorológicos e outros fins. A extração de dados é feita a partir dos produtos gerados pelos sensores remotos (FORMAGGIO; SANCHES, 2017). E, para processar esse volume de dados são utilizadas técnicas de mineração de dados, etapa principal do processo de descoberta de conhecimento em grandes bases de dados, que é utilizada para extração de padrões, possibilitando ganho de tempo com processos de análise automáticos e o esforço na obtenção de informação de interesse para melhorar tomadas de decisões (REZENDE et al., 2005).

Um desafio dos últimos anos é migrar dados para informação, por meio do desenvolvimento dos Sistemas de Informação, e criar sistemas capazes de representar e processar o conhecimento para responder a diferentes necessidades dos indivíduos. O conjunto de dados, sozinho, não oferece embasamento para o entendimento da situação. A informação é o dado analisado e contextualizado. Com a compreensão, análise e síntese das informações são tomadas decisões inteligentes, a partir do nível do conhecimento (REZENDE et al., 2005).

Em particular, a região denominada MATOPIBA, que compreende o estado do Tocantins (TO) e porções dos estados do Maranhão (MA), Piauí (PI) e Bahia (BA), ganhou importância no setor agropecuário brasileiro. A região apresenta forte crescimento desde os

anos 2000 e, atualmente, é considerada fronteira em expansão e intensificação do Brasil (GUERRA et al. 2017). Números do IBGE mostram o desenvolvimento agrícola na região. De acordo com a Produção Agrícola Municipal (IBGE, 2018), a área plantada de algodão, milho e soja, na região do MATOPIBA, era de 21.470.086 ha em 2003. No ano de 2010, a área plantada era 27.486.222 ha e, em 2016, aumentou para 39.140.994 ha.

A topografia plana, os solos profundos e o clima favorável ao cultivo das principais culturas de grãos e fibras possibilitaram o crescimento vertiginoso da região (BORGHI et al., 2014). Outros atrativos incluem a proximidade de terminais portuários, terras comparativamente de baixo valor comercial, possibilidade de escoamento da produção por meio de outros modais de transporte – alternativos ou complementares ao modal rodoviário – e perspectiva de diminuição de custos de produção em função de melhorias de capacidade logística a serem implantadas nos portos do Arco Norte (BELCHIOR et al., 2017). Mapas referentes à infraestrutura da região do MATOPIBA podem ser encontrados no trabalho de Bolfe et al. (2015).

Alguns fatores diferenciam a fronteira agrícola da região do MATOPIBA das demais do Brasil. Primeiramente, a ocupação do Cerrado. Não é mais necessário desmatar vastas áreas para levantar novas cidades. A ocupação recente tem se dado principalmente em espaços já consolidados demograficamente, de menor valor econômico, mas propícios à exploração agropecuária (BELCHIOR et al., 2017). Lorensini et al. (2015) afirmam que grande parte da expansão da agricultura, entre 2006 a 2012, ocorreu em locais previamente desmatados, indicando a utilização de áreas já antropizadas.

Outro fator é a tecnologia disponível para os produtores e o desenvolvimento de pesquisa agropecuária no Brasil, que hoje não é mais um importador de alimentos. A grande demanda internacional por grãos também faz com que a região do MATOPIBA tenha se tornado um dos principais polos de expansão da agricultura brasileira. Com domínio tecnológico, a região é capaz de produzir em grande escala (BELCHIOR et al., 2017; GUERRA et al., 2017).

O terceiro fator diz respeito ao aspecto ambiental. A região até a década de 1980 praticava a pecuária de forma extensiva, porém, atualmente, há necessidade de investir em outro modelo de desenvolvimento que englobe aspectos sociais, ambientais e econômicos (BELCHIOR et al., 2017). Por isso, a análise de dados econômicos e sociais são importantes para avaliação das condições do solo, principalmente quando destinado à agricultura.

Diante desse contexto, pode-se afirmar que existe um potencial agrícola no MATOPIBA, mas é preciso realizar pesquisas para identificar e mapear: os recursos disponíveis, a capacidade da região de produção e compreender a interação de aspectos sociais

e econômicos que caracterize a comunidade local. A avaliação da aptidão agrícola é um processo interpretativo e pode sofrer alterações com avanços tecnológicos e, por isso, deve ser atualizada levando em consideração todos os aspectos (meio ambiente, sociedade e economia) em relação à necessidade local atual. A avaliação do potencial da terra constitui um estágio muito importante nos estudos ambientais para fins de zoneamento e planejamento.

1. HIPÓTESE

A utilização de técnicas de mineração de dados aplicadas a um conjunto de dados composto por atributos físicos, sociais e econômicos dos municípios da região do MATOPIBA, utilizando como atributo resposta a aptidão agrícola, fornecerá modelos preditivos, com elevada acurácia, capazes de subsidiar a elaboração de uma metodologia para classificação de municípios com aptidão agrícola.

2. OBJETIVO

2.1 Objetivo Geral

O objetivo deste trabalho é desenvolver uma metodologia baseada em técnicas de mineração de dados para classificar municípios com aptidão agrícola, na região do MATOPIBA.

2.2 Objetivos Específicos

i) Avaliar o potencial da integração dos dados físicos, sociais e econômicos para apoiar a geração de modelos para classificar municípios com aptidão agrícola, na região do MATOPIBA.

ii) Identificar quais são os atributos físicos, sociais e econômicos mais importantes para identificação de municípios com aptidão agrícola;

iii) Elaborar um conjunto de modelos preditivos capazes de classificar municípios com aptidão agrícola, na região do MATOPIBA.

3. REVISÃO BIBLIOGRÁFICA

3.1 Mineração de Dados

Avanços rápidos na tecnologia de coleta e armazenamento de dados permitem que as organizações acumulem vasta quantidade de dados. Às vezes, ferramentas e técnicas tradicionais de análise de dados não podem ser usadas devido ao grande volume de dados. Esses conjuntos de dados, justamente por serem grandes, podem fornecer estimativas para os modelos, porém, não podem ser analisados pelos métodos tradicionais. Neste contexto, a mineração de dados é a tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados**.** Ela também abriu oportunidades interessantes para explorar e analisar novos tipos de dados e para se analisar tipos antigos de novas maneiras (TAN et al., 2009; WITTEN et al., 2011).

A definição amplamente utilizada por diversos pesquisadores de mineração de dados foi elaborada por Fayyad et al. (1996) como: "Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados".

A mineração de dados é uma parte integral da descoberta de conhecimentos em bancos de dados (KDD- Knowledge Discovery in Databases), que é o processo geral de conversão de dados brutos em informações úteis, conforme a Figura 1. Este processo consiste de uma série de passos de transformação, do pré-processamento dos dados até o pós-processamento dos resultados da mineração de dados (TAN et al., 2009).

Figura 1 - O processo de descoberta de conhecimento em bancos de dados (KDD).

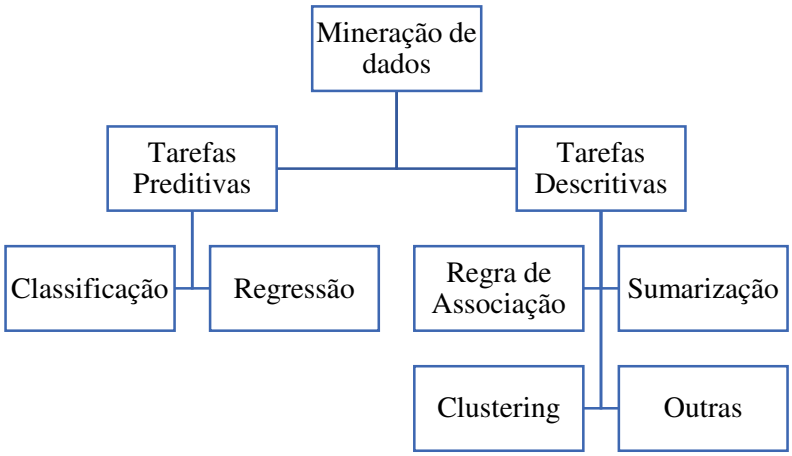
Fonte: Adaptado de TAN et al. (2009).

O pré-processamento de dados pode ser considerado o passo mais trabalhoso e demorado no processo geral de descoberta de conhecimento. Essa etapa consiste na transformação dados brutos em um formato apropriado para análises subsequentes. Os passos envolvidos no pré-processamento de dados incluem a fusão de dados de múltiplas fontes, a limpeza dos dados para remoção de ruídos e observações duplicadas, a seleção de registros e características que sejam relevantes à mineração de dados (TAN et al., 2009).

O pós-processamento é a validação dos resultados para que sejam incorporados ao sistema de apoio a decisões. A visualização é exemplo de pós-processamento, a qual permite que os analistas interpretem os dados e os resultados da mineração a partir de uma diversidade de pontos de vista (TAN et al., 2009).

As tarefas de mineração de dados podem ser divididas em: descritivas e preditivas. Uma tarefa descritiva consiste em identificar comportamentos pré-definidos do conjunto de dados, portanto não há um atributo-meta que seja obtido em função dos demais. Exemplos dessas tarefas incluem agrupamento de dados (ou clusterização), associação e sumarização. Diferentemente das tarefas descritivas, as tarefas preditivas buscam determinar o valor do atributo, chamado de atributo-meta, em função dos outros atributos. Na tarefa de classificação, o atributo-meta é categórico e quando o valor for numérico contínuo, a tarefa é a regressão (WITTEN et al., 2011). A Figura 2 representa as tarefas de mineração de dados.

Figura 2 - Tarefas de mineração de dados



Fonte: REZENDE et al. (2005)

A tarefa de associação é definida por Agrawal e Srikant (1994) como identificação de atributos relacionados de alguma uma implicação da forma $X \rightarrow Y$, em que X e Y podem ser um conjunto de atributos, tal que $X \cap Y = \varnothing$. É uma das tarefas mais conhecidas devido

aos bons resultados obtidos, principalmente nas análises de "Cestas de Compras"(Market Basket), onde identificamos quais produtos são levados juntos pelos consumidores.

Já a tarefa de agrupamento de dados ou clustering (clusterização) tem como função segmentar um conjunto de dados num número de subgrupos homogêneos ou clusters, de forma que os objetos de cada cluster sejam similares aos objetos do mesmo cluster e dissimilares aos objetos de outros clusters (HAN et al., 2011).

Rezende et al. (2005) descrevem a sumarização como método para encontrar uma descrição compacta para um subconjunto de dados. Um exemplo de sumarização é a definição da média e do desvio padrão de todos os atributos de um conjunto de dados.

3.1.1 Descrição das Técnicas de Mineração de Dados Utilizadas nesta pesquisa

Neste trabalha são utilizadas técnicas de classificação de dados. Em particular, são utilizados os seguintes classificadores: árvores de decisão, Random Forest (Florestas Aleatórias) e SVM (Support Vector Machines ou Máquinas de Vetores Suporte), cujas descrições são encontradas a seguir:

Han et al. (2011) definem classificação como um processo que é realizado em etapas:

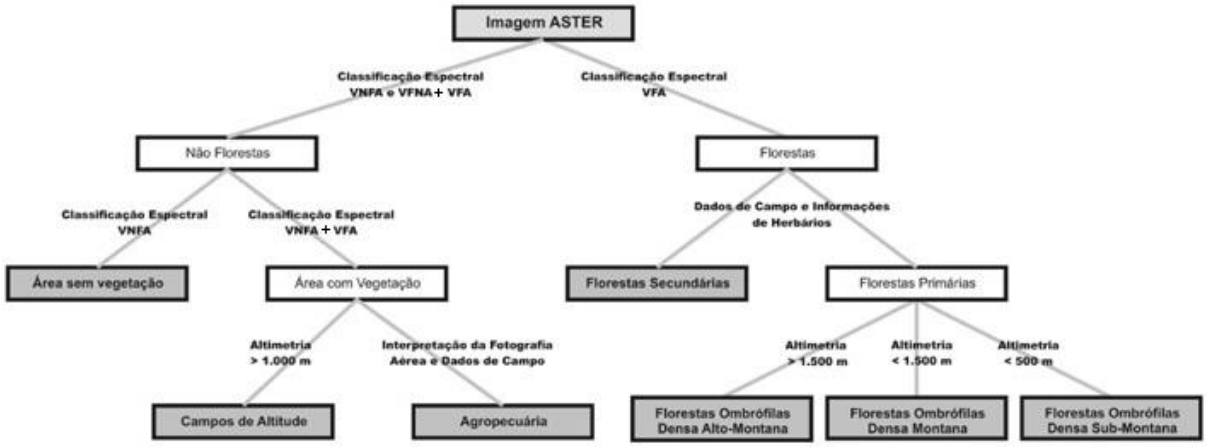
i) A primeira etapa é a criação do modelo de classificação, onde são constituídas regras para o modelo que permitem classificar a sequência do conjunto de dados em um número de classes pré-determinadas. O modelo é criado a partir de um conjunto de dados de treinamento e os elementos são chamados de amostras ou exemplos.

ii) A segunda etapa é a de classificação ou verificação do modelo, onde as regras são testadas sobre um conjunto de dados independente do utilizado na fase de construção e treinamento, chamado de conjunto de dados de testes. A qualidade do modelo é medida em termos da porcentagem de sequência do conjunto de dados de testes em que as regras do modelo conseguem classificá-la de forma satisfatória. Se as regras forem testadas no próprio conjunto de dados de treinamento, elas terão alta probabilidade de estarem corretas, uma vez que este banco foi usado para extraí-las. Por isso, há a necessidade de um conjunto de dados independente para evitar o sobreajuste ou *overfitting*. Diversas técnicas são empregadas na construção

de classificadores. Neste trabalho são utilizadas 3 técnicas: Árvore de Decisão, Floretas Aleatórias e Máquina de Vetores de Suporte.

O método de classificação por Árvore de Decisão é um fluxograma (flow-chart) semelhante a uma árvore, em que cada nó interno indica um teste em um atributo. O ramo (sub-árvore) representa o resultado do teste feito e as folhas representam a distribuição dos registros (HAN et al., 2011). Após gerar a árvore de decisão, para classificar um novo registro, o fluxo na árvore deve ser seguido, começando pelo nó da raiz até a folha. A análise deve ser feita nos testes feitos nos nós, não pelo das folhas. Essa é uma técnica simples, não necessita de parâmetros de configuração e apresenta elevado grau de assertividade. Por isso, a técnica é considerada muito eficiente, sendo necessária a análise detalhada dos dados para garantir que os resultados sejam gerados com sucesso (CAMILO; SILVA, 2009). Um exemplo de árvore de decisão pode ser visto na Figura 3.

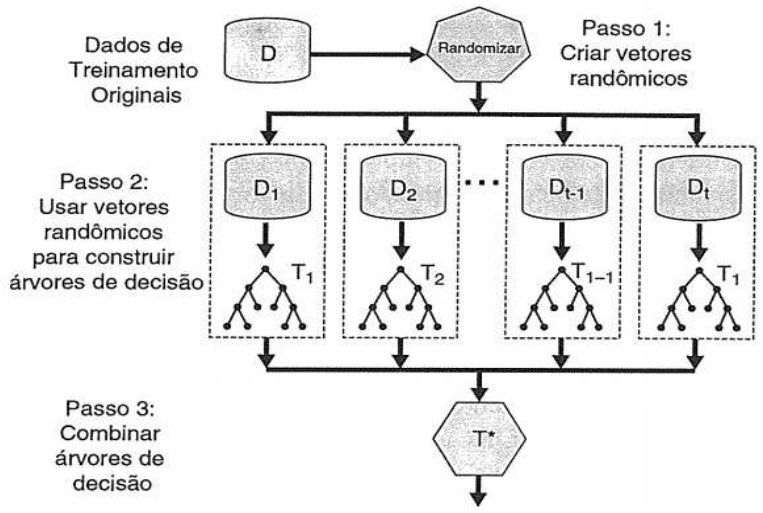
Figura 3 - Exemplo de uma árvore de decisão



Fonte: Carvalho Júnior et al. (2008)

As florestas aleatórias (*Random Forest*) são uma técnica de classificação desenvolvida por Breiman (2001). É uma classe de métodos de conjunto projetada especialmente para classificadores de árvore de decisão. Ela combina as previsões feitas por múltiplas árvores de decisões, onde cada árvore é gerada baseada nos valores de um conjunto independente de vetores aleatórios (TAN et al., 2009). Cada um destes conjuntos é criado por um tipo de amostragem chamado de *bootstrap* (HAN et al., 2011). A Figura 4 ilustra o processo de geração de uma floresta aleatória.

Figura 4 - Exemplo da geração de uma floresta aleatória.



Fonte: TAN et al. (2009)

Cada árvore de decisão usa um vetor aleatório que é gerado a partir de alguma distribuição de probabilidade fixa. Um vetor aleatório pode ser incorporado no processo de desenvolvimento da árvore de diversas maneiras. A primeira abordagem é selecionar aleatoriamente F características de entrada para dividir em cada nodo da árvore de decisão. Como consequência, em vez de examinar todas as características disponíveis, a decisão de dividir um nodo é determinada a partir dessas F características selecionadas. A árvore é então desenvolvida integralmente sem qualquer poda. Isto pode auxiliar a reduzir a tendência presente na árvore resultante. Assim que as árvores tiverem sido construídas, as previsões são combinadas usando um esquema de votação (TAN et al., 2009).

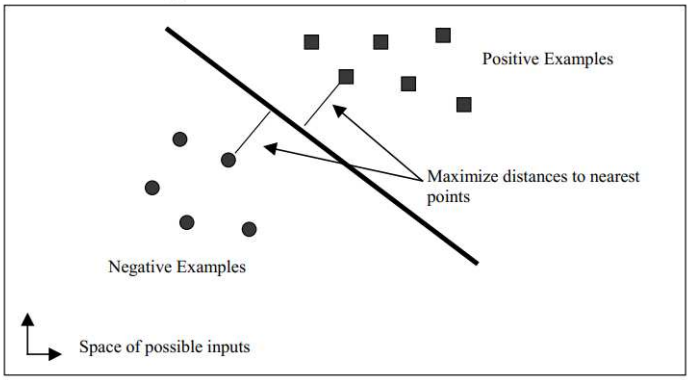
Para cada voto existe um "peso", que e é afetado pela similaridade entre as árvores, quanto menor a similaridade das duas árvores melhor. Pela força que cada árvore tem individualmente, então, quanto mais precisa uma árvore for, melhor será sua nota. O ideal é manter a precisão das árvores sem aumentar sua similaridade (HAN et al., 2011).

Máquina de Vetores de Suporte (em inglês, *Support Vector Machine* - SVM), desenvolvida por Vapnik (1995), é um método de classificação para dados lineares e não lineares. A classificação possui seus fundamentos na teoria de aprendizagem estatística e tem mostrado resultados empíricos promissores em muitas aplicações práticas. O SVM funciona muito bem com dados de alta dimensionalidade e evita problema da dimensionalidade (TAN et al., 2009).

Em sua forma linear mais simples, o SVM é um hiperplano que separa um conjunto de exemplos positivos de um conjunto de exemplos negativos com margem máxima (Figura 5).

Nesse caso, a margem é definida pela distância do hiperplano para o mais próximo dos exemplos positivos e negativos (PLATT, 1998).

Figura 5- Esquema da classificação SVM



Fonte: Platt (1998).

Há muitos casos em que não é possível dividir os dados de treinamento por um hiperplano (VAPNIK, 1995). Então, utiliza-se um mapeamento não linear para transformar os dados de treinamento originais em uma dimensão maior. Dentro dessa nova dimensão, ele procura pelo hiperplano de separação ótimo linear (isto é, um "limite de decisão" que separa uma classe da outra). Com o mapeamento não linear apropriado para uma dimensão suficientemente alta, os dados de duas classes sempre podem ser separados por um hiperplano. O SVM encontra este hiperplano usando vetores suporte (treinamento “essenciais”) e margens (definidas pelos vetores de suporte). Eles são menos propensos a *overfitting* do que outros métodos (HAN et al., 2011).

3.1.2 Métodos de seleção e atributos

Segundo GUYON e ELISSEEFF (2003), a seleção de atributos é o processo de refinamento do conjunto de treinamento pela escolha dos atributos mais importantes, sendo que a ideia por trás destes métodos é selecionar os atributos mais importantes do conjunto de dados e ignorar os demais, que pouco contribuem para um modelo de classificação. Os métodos de seleção de atributos utilizados neste trabaho são: Qui-quadrado (χ^2), Ganho de Informação, CFS (*Correlation Feature Selection*) e Wrapper.

3.1.2.1 Qui-quadrado (χ^2)

O teste do qui-quadrado é um teste global para detectar relações entre variáveis categóricas. Se o teste é significante, devem-se examinar os dados para compreender a natureza da relação. O uso mais comum e mais importante da estatística qui-quadrado é testar a hipótese nula (H_0) de que não há relação entre as variáveis categóricas (MOORE, 2005). Este método avalia o valor de um atributo calculando o valor da estatística qui-quadrado em relação à classe (atributo-meta). Se o valor da estatística qui-quadrado for maior que um certo valor crítico (com nível de significância a 5%), H0 é rejeitada e H1 (hipótese de que há associação entre os atributos) é aceita. Neste caso, o atributo que tem associação com o atributo-meta é mantido (selecionado).

O teste do χ^2 é apresentado na equação 1:

$$\chi^2 = \frac{\sum(Obervada - Esperada)^2}{Esperada} \quad (1)$$

Onde:

As frequências observadas são obtidas diretamente dos dados das amostras, enquanto que as frequências esperadas são calculadas a partir destas.

3.1.2.2 CFS (*Correlation Based Feature Selection*)

O método CFS é baseado nas relações entre os atributos. Primeiro ele calcula uma matriz de correlações entre atributo-atributo e entre atributo-classe (WITTEN et al., 2011). Um peso (Score) de um conjunto de atributos é associado utilizando a seguinte equação:

$$Mérito(S) = \frac{k \times r'_{ac}}{\sqrt{k + k(k-1) r'_{aa}}} \quad (2)$$

Mérito (S) é o mérito de um subconjunto de atributos S contendo k atributos;

r'_{ac} é a média da correlação entre atributo-classe;

r_{aa} é a média da correlação entre atributo-atributo.

O CFS é um algoritmo simples que classifica subconjuntos de acordo com a correlação baseada em avaliação heurística. A avaliação é feita por subconjuntos de atributos que contém características altamente correlacionadas com a classe (atrinuto-meta) e não correlacionadas entre si. As características irrelevantes do atributo são ignoradas pois têm baixa correlação com a classe. O objetivo desse algoritmo é escolher atributos com maior poder de predição e diminuir a dimensionalidade retirando atributos com informações redundantes (HALL, 1999).

3.1.2.3 Ganho de informação

O Ganho de informação é uma das medidas baseadas em impurezas e usa a entropia como medida da impureza. Para o cálculo da entropia utiliza-se a equação 3

$$Entropia(D) = -\sum_{i=1}^{m} p_i log_2 (p_i) \quad (3)$$

Onde:
D é a distribuição de probabilidade das m mensagens possíveis;
p_i é a probabilidade de ocorrência da i-ésima mensagem

Se a entropia é pequena, próxima de zero, significa que as classes têm a menor impureza e maior ganho de informação. Por outro lado, quanto maior a entropia de um atributo, menor o ganho de informação e maior a desordem do atributo (HAN e t al., 2011).

Este método avalia o valor de um atributo medindo o ganho de informação em relação ao atributo-meta.

O ganho de informação é a diferença entre as entropias (Equação 4) e a análise é feita atributo-meta. Considerando um conjunto de atributos, o *Gain (A)* é o atributo com maior ganho de informação e é escolhido como atributo *splitting* do nó. Isso equivale a dizer que atributo A que faria a "melhor classificação", de modo que a quantidade de informação ainda necessária para concluir a classificação das sequências seja mínima (mínimo $Entropia_A(D)$) (HAN et al., 2011).

$$Gain\ (A) = Entropia\ (D) - \ Entropia_A(D) \quad (4)$$

Para calcular a $Entropia_A(D)$ (Equação 5):

$$Entropia_A(D) = \sum_{j=1}^{V} \frac{|Dj|}{|D|} \times Entropia\ (Dj) \quad (5)$$

O termo $\frac{|Dj|}{|D|}$ atua no peso da partição j. $Entropia_A(D)$ é a informação esperada necessária para classificar uma sequência de D.

3.1.2.4 Wrapper

O método de seleção de atributos, Wrapper, utiliza um algoritmo de indução para estimar subconjuntos de atributos usando um esquema de aprendizado. A validação cruzada é usada para estimar a precisão do esquema de aprendizado para um conjunto de atributos (KOHAVI; JOHN, 1997).

Esse método geralmente alcança melhores resultados devido ao fato de estarem sintonizados na interação específica entre um algoritmo de indução e seus dados de treinamento. No entanto, eles tendem a ser mais lentos do que os métodos de seleção de atributos, pois precisam repetir o algoritmo de indução, com validação cruzada, como mencionado anteriormente. O método deve ser executado novamente quando um algoritmo de indução diferente for utilizado (HALL, 1999).

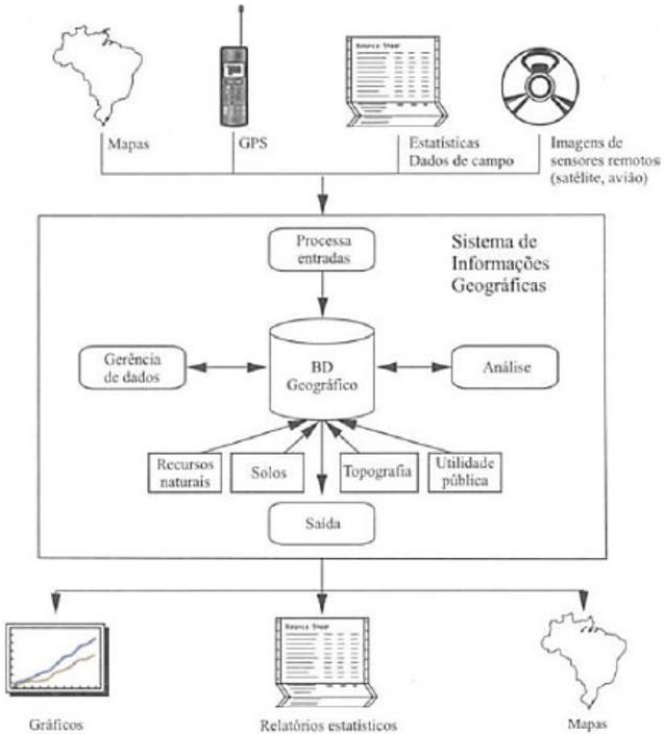
3.2 Geotecnologias

O desenvolvimento de sistemas computacionais para aplicação gráficas e de imagens vem influenciando de maneira crescente as áreas de cartografia, de mapeamento, de análise de recursos naturais e de planejamento urbano e regional. Essa tecnologia auxilia automatiza tarefas e facilita a realização de análises complexas, através da integração de dados de diversas fontes e da criação de um banco de dados geocodificado. Os sistemas para tal fim são denominados Sistemas de Informações Geográficas (SIG's) (FELGUEIRAS; CÂMERA, 1993).

O SIG (Sistema de Informação Geográfica), segundo Furtado et al. (2014), "é um exemplo de geotecnologia e representa a união de hardware e de software capazes de armazenar, analisar e processar dados georreferenciados", sendo a tecnologia usada para a análise de dados de sensoriamento remoto.

O especialista pode enxergar quadro diferentes subdivisões no SIG: entrada, gerência dos dados, análise e saída como mostra a Figura 6. Essas subdivisões não existem na implementação do sistema, e sim, a concepção do projeto lógico (MIRANDA, 2005).

Figura 6 - Configuração básica de um SIG



Fonte: Miranda (2005)

A entrada de dados compreende as funções de coleta, pré-processamento de dados espaciais ou não espaciais de várias fontes e transformação de diferentes tipos de dados espaciais. Essa etapa é da captura de dados espaciais como mapas existentes, observações em campo, imagens captadas por sensores remotos, informações coletadas por GPS (*Global Positioning System*). Neste momento começa-se montar um banco de dados georreferenciado para desenvolver o projeto (MIRANDA, 2005).

A etapa de consiste em organizar dados e permitir a sua recuperação, atualização e edição. O armazenamento e a recuperação referem-se à maneira da estruturação e organização de uma coleção de mapas e informações associadas (localização, topologia e atributos de elementos geográficos). Isso interfere no modo que os dados devem ser processados pelo usuário do sistema (MIRANDA, 2005).

A etapa mais importante é a da análise de dados. Sua função é agregar, desagregar dados, estimar parâmetros, restrições e realizar funções de modelagem. Esta etapa, implantam vários métodos de processamento de dados fornecendo respostas às requisições feitas pelo SIG.

As transformações são de domínio espacial, topológico e não espacial e esses dados podem ser trabalhados juntos ou separadamente. Exemplos que são encontrados no SIG: mudanças de escalas, ajuste de dados a novas projeções, recuperação da lógica de dados, cálculo de área e perímetro (MIRANDA, 2005).

A última etapa é de saída e sua função é mostrar toda sua base de dados ou parte dela em forma de tabela, gráfico ou mapa (MIRANDA, 2005). São os resultados das análises feitos pelo usuário que servirão de base para pesquisas, órgãos públicos e empresas, assim atingindo assim seus objetivos em diferentes áreas (planejamento do uso da terra, gestão ambiental analises, análise social, marketing empresarial) para melhorar a tomada de decisões (PRICE, 2016).

Os SIGs têm sido aplicados no gerenciamento em diversas áreas, tais como, urbana, regional, floresta, infraestrutura, utilidade pública, utilidade privada, além de serem utilizados também na otimização de tráfego, na atualização de dados cartográficos e no monitoramento agrícola, entre outros (CÂMARA, 1995).

Pela abrangência, os SIGs podem integrar, um banco de dados único, informações espaciais provenientes de diferentes áreas do conhecimento e oferecer mecanismos para combinar essas várias informações, por meio de algoritmos de manipulação e análise, promovendo ajustes e cruzamentos simultâneos de um grande número de informações; e para consultar, recuperar, visualizar, atualizar e permitir a impressão do conteúdo da base de dados geográficos para obtenção de novos mapas com rapidez e precisão. Portanto, é uma ferramenta importante para estudos de potencialidades do meio ambiente (THOME, 1998; ASSAD, 1993).

Sistemas de informações geográficas podem auxiliar a avaliação de impactos ambientais e fazer simulações de cenários alternativos. Estes sistemas são aplicados com sucesso no monitoramento e detecção de condições de mudança em áreas urbanas ou no campo, previsão de potencial de uso de recursos naturais e solos, avaliação de relações causa-e-efeito, modelagem da interação de vários componentes do ambiente terrestre para predição dos efeitos de mudanças, entre outros (MIRANDA, 2005).

Para Rigaux et al. (2000), citado por Adami (2004), os SIGs devem ter as características de um sistema de banco de dados para evitar inconsistência e redundância dos dados; facilitar o acesso (manipulação e consulta) e a atualização; manter a integridade e a atomicidade; permitir o acesso concorrente e garantir a segurança dos dados.

A utilização de SIGs possibilita a geração de bancos de dados codificados espacialmente, promovendo ajustes e cruzamentos simultâneos de um grande número de

informações, podendo, ainda, atualizar bancos de dados para obtenção de novos mapas com rapidez e precisão (ASSAD, 1993).

Os SIGs são considerados uma ferramenta essencial para representação gráfica das classes de aptidão agrícola e para atualizar as informações. Sua maior contribuição é minimizar a complexidade e o grau de subjetividade das estimativas feitas a partir de cruzamentos realizados manualmente. Para avaliar a aptidão agrícola das terras deve-se considerar aspectos ambientais como: solos, clima, recursos hídricos, vegetação, infraestrutura, ocupação humana, etc. que frequentemente são encontrados em diferentes formatos (textos, mapas, tabelas, fotos aéreas, imagens de satélite). O cruzamento simultâneo dessas informações pode tornar-se complexo se for realizado de forma manual (ASSAD, 1993).

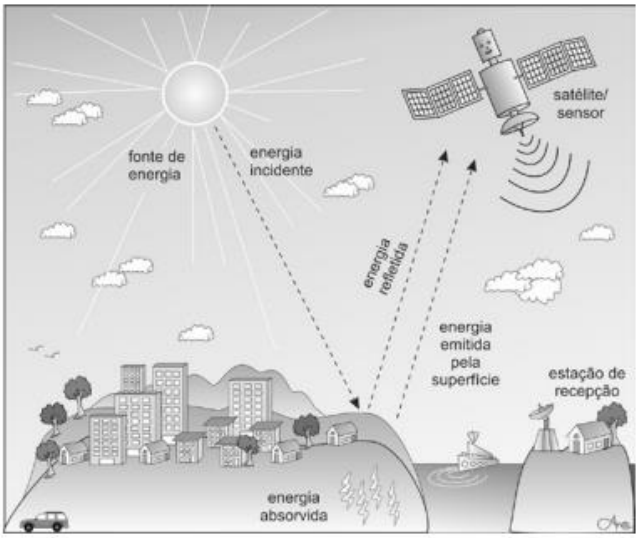
3.3 Sensoriamento Remoto

Sensoriamento Remoto, definido por Jenner (2011), é a arte e a ciência de obter informações sobre um objeto sem estar em contato físico direto com o mesmo, e podendo ser utilizado para medir e monitorar características biofísicas e atividades humanas na Terra.

Novo (2010) define o Sensoriamento Remoto como um conjunto de sensores, equipamentos para o processamento de dados e equipamentos de transmissão de dados colocados a bordo de aeronaves, espaçonaves, ou outras plataformas, que a partir de interações entre a radiação eletromagnética e as substâncias que compõem o planeta Terra pode-se estudar eventos, fenômenos e processos que ocorrem na superfície do planeta. Portanto, é a ciência e arte de coletar dados ou informações de um objeto a partir da energia refletida ou emitida por esse (FURTADO et al. 2014).

Os sensores remotos (passivos ou ativos) têm a função de captar a radiação eletromagnética (REM) refletida ou emitida por objetos ou alvos presentes na Terra, transformá-la em um sinal elétrico e transmitir esse sinal em uma estação recepção (FURTADO et al., 2014; MOREIRA, 2011). A interação energia-matéria ocorre quando a energia atinge o alvo (JENNER, 2011). A energia (radiação solar) que chega à superfície da Terra interage com os objetos (vegetação, recursos hídricos, edificações e áreas agrícolas) podendo ser refletida, absorvida ou transmitida, dependente da propriedade físico-química e biológica de cada objeto (FURTADO et al. 2014; MOREIRA, 2011). Após interagir com a radiação eletromagnética incidente (radiância), alvos ou objetos refletem essa radiação (irradiância) de forma diferente, e isso representa sua assinatura espectral (FURTADO et al., 2014). A Figura 7 ilustra a aquisição dos dados de sensores remotos.

Figura 7 - Aquisição de dados



Fonte: QUARTAROLI et al. (2014)

Os alvos da superfície da Terra podem ter sua resposta espectral alterada com o tempo por causa de modificações externas ao alvo (iluminação, alteração antrópicas) ou de modificações própria da sua natureza (NOVO, 2010). Com isso, o registro do fluxo radiante proveniente de um alvo permite adquirir informações importantes sobre suas características físicas (dimensão, forma, temperatura, cor, etc.) e químicas (composição) (QUARTAROLI el. al., 2014).

Os sistemas de sensoriamento remoto disponíveis atualmente fornecem dados repetidos e consistentes da superfície da Terra, os quais são de grande utilidade para várias aplicações como: urbanas, agrícolas, geológicas, ecológicas, florestais, cartográficas, oceanográficas, hidrológicas, limnológicas, militares, entre outras (NOVO, 2010).

No caso de culturas agrícolas, o comportamento espectral da vegetação se modifica ao longo do ciclo vegetativo. Alterações fenológicas e morfológicas sofridas pelas plantas modificam a sua resposta espectral e com isso, pode ser um meio que auxilie na identificação a cultura plantada e estimar a produção da cultura agrícola. Além, das características espectrais das imagens do satélite, o especialista em sensoriamento remoto utiliza também elementos da fotointerpretação tais como: forma, sombreamento e textura (RUDORFF; MOREIRA, 2002; NOVO, 2010)

Com isso em mente, o sensoriamento remoto pode contribuir de modo significativo no fornecimento de dados precisos sobre o setor agrícola, e a geotecnologia é o melhor meio para coletar as informações detalhadas e confiáveis sobre grandes áreas e com bom custo-

benefício. Algumas contribuições são: informações de área plantada, avaliações de eventos climáticos sobre a produção agrícola, informações para auxiliar o entendimento de possíveis efeitos de mudanças climáticas, mapeamentos e estimativas de produtividade, entre outros (FORMAGGIO; SANCHES, 2017)

3.4 Aplicações de Técnicas de Mineração de Dados e Sensoriamento Remoto na Agricultura

O uso de técnicas de mineração de dados é uma alternativa promissora para explorar grande quantidade de dados e informações dos sensores remotos. Lopes (2013) utilizou técnicas de mineração de dados para desenvolver modelos para classificar culturas temporárias, no estado de Mato Grosso, considerando séries de NDVI. A autora utilizou técnicas para seleção de atributos e as regras de classificação foram geradas por meio dos algoritmos J48, Random Forest e PART.

Nonato e Oliveira (2013) avaliaram método de seleção de atributos Wrapper e técnica de árvore de decisão binária na identificação de áreas cultivadas com cana-de-açúcar, no estado de São Paulo, por meio de imagens do satélite Landsat 5/TM. Para a classificação automática de áreas com cana-de-açúcar foi utilizado cinco alvos diferentes: solo exposto, áreas urbanas, corpos de água, florestas e áreas cultivadas com cana-de-açúcar, em três fases fenológicas diferentes (crescimento, perfilhamento e maturação). Também foram utilizados índices de vegetação para o treinar o modelo. Em seguida, os pixels dos alvos foram convertidos em valores de refletância de superfície, através de cálculos da radiância e correção atmosfera com auxílio de imagens do sensor MODIS. Feito isto, a imagem foi classificada através do algoritmo J48 (árvore de decisão) e com a seleção de atributos, Wrapper, o índice de acerto foi satisfatório na identificação de áreas cultivadas com cana, nas diferentes fases fenológicas.

Martins e Fonseca (2009) apresentam um trabalho com a proposta de utilizaram técnicas de mineração de dados para auxiliar na classificação orientada a objeto em uma região com atividade agrícola. Definens (2008), citado pelos autores, explica que a classificação orientada a objetos é interpretada a partir dos objetos das imagens e nas relações entre si, esta interpretação de imagem não está presente no pixel. Com isso, eleva o número de atributos utilizados para a distinção dos diferentes alvos da cena e para evitar atributos redundantes são utilizadas técnicas de mineração de dados para melhorar a classificação de alvos agrícolas em áreas heterogêneas. O algoritmo utilizado neste trabalho foi a árvore de decisão que mostrou eficiência para determinar melhores atributos e classificar as imagens agrícolas.

Manabe et al. (2015) diferenciaram áreas de cana de açúcar e pastagem a partir da análise do comportamento temporal dos índices de vegetação NDVI do sensor Terra/MODIS utilizando técnicas de mineração de dados. Para a seleção de atributos os autores utilizaram os métodos: CFS (*Correlationbased feature selection*), GainRatio (*Information Gain Ratio*), InfoGain (*Information Gain*), PCA (*Principal Components Analysis*), Wrapper e Qui-quadrado (x^2) e para realizar a classificação os algoritmos: J48 (Árvore de decisão), MLP – Multilayer Perceptron (Redes Neurais) e SMO – Sequential Minimlal Optimization (SVM). O trabalho comparou a classificação com as classes balanceadas e desbalanceadas, feito o balanceamento de classes houve uma melhora nos classificadores. A seleção de atributos também contribui para a melhora no desempenho dos modelos de classificação, sendo o método GainRatio melhor para os algoritmos J48 e SMO e o método chi-quadrado para o MLP. O algoritmo MPL apresentou melhor resultado neste trabalho depois da seleção de atributos e balanceamento de classes.

Girolamo Neto et al. (2015) avaliaram os algoritmos *Random Forests* e árvores de decisão para o mapeamento automático de áreas cafeeiras. Para realizar este trabalho foram utilizadas uma imagem com 11 bandas do sensor OLI (*Operational Land Imager*) do satélite Landsat-8 e outra imagem temática com classificação de áreas de café na região estudada. O atributo meta foi constituído das classes café ou não-café e os atributos preditivos partiram do valor de cinza de cada pixel Landasat-8 e a relação entre as bandas gerou os índices de vegetação como EVI e NDVI. Os dados foram balanceados e reduzidos com métodos de seleção de atributos. Os resultados estatísticos do trabalho não foram satisfatórios, a cultura cafeeira presenta complexidade, onde fatores ambientais, fenológicos e de manejo interferem na resposta espectral. A sugestão dos autores é ampliar o número de classes a fim de treinar um modelo com menor variabilidade espectral ou trabalhar com uma classificação baseada em segmentação ou avaliar outros algoritmos de classificação.

Johann et al. (2013) tiveram como objetivo agrupar, por semelhança, perfis temporais do produto NDVI decendial obtido pelo sensor SPOT Vegetation, para os municípios produtores de soja no estado do Paraná. Os algoritmos de clusterização utilizados no trabalho foram K-Means, MAXVER e DBSCAN. O melhor resultado foi o algoritmo K-Means, agrupando os 227 municípios produtores de soja em seis clusters, com base no perfil temporal médio de NDVI, utilizando o período de início de outubro ao final de março, equivalente ao ciclo da cultura. Com isso, foi possível gerar "máscaras" agrícolas regionalizadas para estimativa de área plantada de soja.

Carvalho Júnior et al. (2008) produziram um mapa de vegetação para o Parque Nacional da Serra dos Órgãos (PARNASO) através da classificação com o algoritmo J48 (árvore de decisão), utilizando dados do satélite ASTER, fotografias aéreas de alta resolução e modelo digital de elevação (MDE). O MDE indica estruturas de vegetação em ambiente com alta variabilidade espacial porque se correlaciona com fatores ambientais (microclima, umidade, solo e processos geomorfológicos). O resultado deste trabalho foi bastante satisfatório, o algoritmo mostrou eficiência quando utilizado dados de diferentes fontes (imagem de satélite e atributo do terreno) e com isso, foi possível elaborar um mapa de vegetação com sete fisionomias: agropecuária, campos de altitude, floresta ombrófila densa alto-montana, floresta ombrófila densa montana, floresta ombrófila densa sub-montana, floresta secundária e áreas sem vegetação.

Dias et al. (2016) avaliaram diferentes estratégias (detalhamento da legenda, o treinamento por observações em campo, a ampliação do conjunto de treinamento e o uso de diferentes algoritmos de mineração de dados) para a predição da distribuição de classes de solo em mapas pedológicos digitais de áreas sem dados de referência, na bacia do São Francisco. Para isto, foi analisado o desempenho dos algoritmos Random Forest, J48 e MLP com associados a procedimentos de discretização, balanceamento de classes, seleção de variáveis e expansão do conjunto de treinamento. O algoritmo Random Forest apresentou o melhor desempenho para predição de classes do solo a partir de variáveis de relevo, geologia e sensoriamento remoto.

4.5 Aptidão Agrícola das Terras

A classificação da aptidão agrícola das terras pode ser técnica ou interpretativa, os solos são agrupados de acordo com o interesse prático e específico, relacionado com seu comportamento. Há vários agrupamentos de terras, em termos agronômicos, como: aptidão agrícola para cada cultura, risco de erosão, necessidade de calagem, com finalidades de irrigação ou drenagem e a capacidade máxima de uso da terra. As classificações são importantes para fins como: geotecnia, construção de aeroportos, engenharia sanitária, taxação de impostos, engenharia rodoviária e ferroviária (RAMALHO FILHO; BEEK, 1995; LEPSCH et al., 1991; FREIRE, 1984 citado por PEREIRA, 2002). Aptidão agrícola pode ser definida como a adaptabilidade da terra para um tipo específico de utilização agrícola das terras, pressupondo-se um ou mais diferentes níveis de manejo (CURI et al., 1993).

A metodologia desenvolvida por Ramalho Filho e Beek (1995), para avaliar a aptidão agrícola das terras, considera cinco fatores limitantes quanto à utilização das terras: deficiência de fertilidade, deficiência de água, excesso de água, suscetibilidade à erosão e impedimentos à mecanização, avaliados a partir de atributos de solo, relevo e clima.

1) Deficiência de fertilidade: A fertilidade é a disponibilidade de macro e micronutrientes no solo. O fator inclui, também, a presença ou ausência de substâncias tóxicas solúveis (alumínio e manganês) e de sais solúveis (sódio). Para o cálculo do índice de fertilidade são usados dados obtidos na análise dos perfis do solo. Os dados são: saturação de bases (V%), saturação com alumínio (100 Al/Al+S), soma de bases trocáveis (S), capacidade de troca de cátions (CTC), relação C/N, fósforo assimilável, saturação com sódio, condutividade elétrica e pH.

   Outras indicações para análise da fertilidade natural: profundidade efetiva do solo, condições de drenagem, atividade biológica, tipos de vegetação, uso da terra, qualidade da pastagem, comportamento de culturas, rendimentos, etc. Esses fatores auxiliam na classificação da aptidão agrícola.
2) Deficiência de água: É a qualidade de água armazenada no solo, possível de ser aproveitada pelas plantas. O fator analisa dados de condições climáticas, principalmente precipitação e evapotranspiração, solo, características da vegetação natural e comportamento das culturas.
3) Excesso de água: Está relacionado com a classe de drenagem natural do solo. Os dados analisados para esse fator são: precipitação, evapotranspiração, relevo local e propriedades do solo.
4) Suscetibilidade à erosão: Diz respeito ao desgaste que a superfície do solo poderá sofrer, quando submetida a qualquer uso. O fator analisa dados de condições climáticas (principalmente o regime pluviométrico), solo, relevo e a cobertura vegetal.
5) Impedimento à mecanização: Esse fator analisa condições de solo (drenagem, profundidade, textura, tipo de argila, pedregosidade e rochosidade) e relevo.

A estrutura básica da metodologia de aptidão agrícola das terras, constitui-se de:

a) Níveis de manejo:

- Nível de manejo A (primitivo): baseado em práticas agrícolas que refletem um baixo nível técnico-cultural. Não há aplicação de capital para o manejo e as práticas agrícola dependem do trabalho braçal, podendo ser utilizada alguma tração animal;
- Nível de manejo B (pouco desenvolvido): baseado em práticas agrícolas de nível tecnológico médio. Modesta aplicação de capital e resultados de pesquisas para manejo, melhorando e conservação das condições das terras e das lavouras. Mecanização com base na tração animal ou na tração motorizada, apenas para desbravar e preparar o solo inicial; e
- Nível de manejo C (desenvolvido): Baseado em práticas agrícola que refletem um alto nível tecnológico. Aplicação intensiva do capital e de resultados de pesquisas para o manejo. A moto mecanização está presente em diversas fases da operação agrícola.

b) Grupos: representam as possibilidades de utilização das terras. Os grupos 1, 2 e 3 para o uso de lavouras; grupo 4 para o uso com pastagem plantada; grupo 5 para o uso com silvicultura e/ou pastagem natural e o grupo 6 para a preservação da flora e fauna (sem aptidão para uso agrícola)

c) Classes: resultam da interação das condições agrícolas, do nível de manejo e das exigências dos diversos tipos de utilização. As classes são definidas como boa, regular, restrita ou inapta (PEREIRA, 2002).

d) Subgrupos: referem-se às variações dos grupos. Representa a interação da classe com o nível de manejo, sendo descrito por uma legenda de identificação (PEREIRA, 2002).

Dos graus de limitação atribuídos a cada uma das unidades das terras, resulta a classificação de sua aptidão agrícola. As letras indicativas das classes de aptidão, de acordo com os níveis de manejo, podem aparecer nos subgrupos em maiúsculas, minúsculas ou minúsculas entre parênteses, com indicação de diferentes tipos de utilização, conforme pode ser observado na Tabela 1. A ausência de letras representativas das classes de aptidão agrícola, na simbolização dos subgrupos, indica não haver aptidão para uso mais intensivo. Essa situação exclui, necessariamente, o uso da terra com um tipo de utilização menos intensivo (RAMALHO FILHO; BEEK, 1995).

Tabela 1- Simbologia correspondente às classes de aptidão agrícola das terras.

Classe de aptidão agrícola	Tipo de utilização					
	Lavoura			Pastagem plantada	Silvicultura	Pastagem natural
	Manejo A	Manejo B	Manejo C	Manejo B	Manejo B	Manejo A
Boa	A	B	C	P	S	N
Regular	a	b	c	P	s	n
Restrita	(a)	(b)	(c)	(p)	(s)	(n)
Inapta	-	-	-	-	-	-

Fonte: Ramalho Filho e Beek (1995)

Exemplos:

1ABc

1 = grupo de aptidão

A = aptidão boa no nível de manejo A

B = aptidão boa nível de manejo B

c = aptidão regular no nível de manejo C

2 ab(c)

2= grupo de aptidão

a = aptidão regular no nível de manejo A

b = aptidão regular b nível de manejo B

(c) = aptidão restrita no nível de manejo C

O método da Avaliação da Aptidão Agrícola desenvolvido por Ramalho Filho e Beek (1995) mostra algumas vantagens, como: a) maior utilização ao nível nacional; b) considera, na sua estrutura, diferentes níveis de manejo; c) considera a viabilidade de redução de limitações, pelo uso de capital e tecnologia, distinguindo o pequeno e o grande agricultor; d) permite seu ajustamento e atualizações frente a novos conhecimentos; e) aceita adaptações e aplicações em diferentes escalas de mapeamento (PEREIRA; LOMBARDI NETO, 2004).

Larach (1990), citado por Pereira (2002), ressalta que, embora a concepção da metodologia de aptidão agrícola tenha sido desenvolvida para interpretação de levantamentos generalizados, ela é suficientemente elástica para permitir reajustamentos, no caso de ser aplicada em levantamentos mais detalhados.

Como a classificação da aptidão agrícola das terras (todas as formas de utilização agronômica das terras) é um processo interpretativo, seu caráter é efêmero, podendo sofrer

variações com a evolução tecnológica. A classificação da aptidão das terras não é precisamente um guia para obtenção do máximo benefício das terras, e sim uma orientação de como devem ser utilizados seus recursos no planejamento regional e nacional (RAMALHO FILHO; BEEK, 1995).

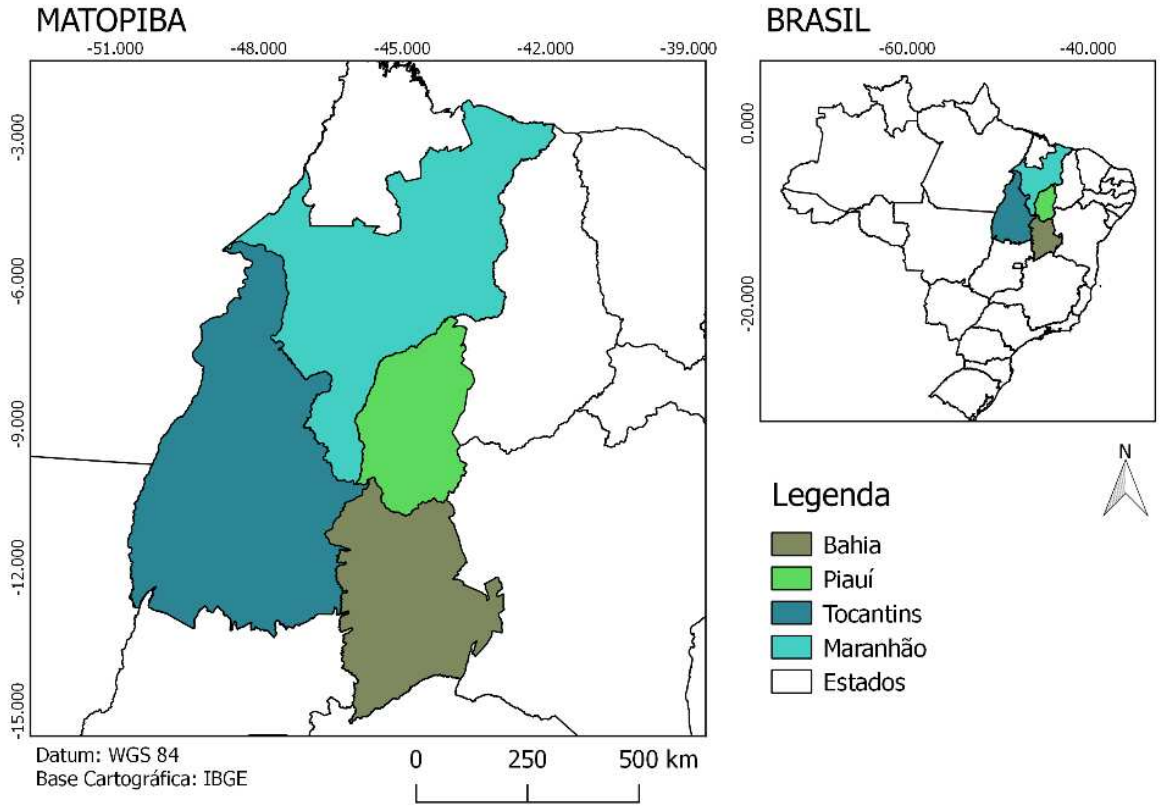
4. MATERIAL E MÉTODOS

4.1 Área de Estudo

A repartição territorial aproximada do MATOPIBA entre os quatro estados se dá da seguinte foma: 33% no Maranhão (15 microrregiões, 135 municípios, 23.982.346 ha); 38% no Tocantins (8 microrregiões, 139 municípios e 27.772.052 ha); 11% no Piauí (4 microrregiões, 13 municípios e 8.204.588 ha) e 18% na Bahia (4 microrregiões, 30 municípios e 13.214.499 ha) (EMBRAPA, 2017), como mostra a Figura 8. A região abriga 324.326 estabelecimentos agrícolas, 46 unidades de conservação, 35 terras indígenas, 781 assentamentos de reforma agrária e 34 áreas quilombolas (BELCHIOR et al., 2017).

Quanto ao clima, tem-se a ocorrência de duas estações bem definidas: uma seca, de maio a setembro, e outra chuvosa, de outubro a abril, com precipitação média anual de 800 mm a 2.000 mm (BOLFE et al., 2015).

Figura 8 - Localização da área de estudo



4.2 Base de Dados

Na etapa inicial do trabalho, foram obtidos dados sociais, econômicos e físicos dos municípios, no período de 17 anos, entre os anos de 2000 a 2016.

5.2.1 Dados Econômicos

Os dados econômicos compreendem: área plantada (hectares), quantidade produzida (toneladas) e rendimento médio (quilogramas por hectare) que estão disponíveis no banco de dados da Produção Agrícola Municipal (PAM - IBGE). Nesse estudo foram consideradas as culturas com maior destaque na região do MATOPIBA: soja, milho (1ª e 2ª safra) e algodão

O Instituto Brasileiro de Geografia e Estatística (IBGE) oferece recursos para consultas e pesquisas, dentre eles pode-se encontrar o Sistema IBGE de Recuperação Automática (SIDRA), onde é possível obter dados de diversos temas e fazer consultas específicas pelo endereço: https://sidra.ibge.gov.br. O tema selecionado "foi produção agrícola municipal" onde foi possível determinar as variáveis: área plantada, quantidade produzida e rendimento médio para os anos e culturas em estudo.

2.2.2 Dados Sociais

Os dados sociais englobam: Índice de desenvolvimento humano municipal (IDHM), Taxa de analfabetismo com 25 anos ou mais, renda per capita e a porcentagem da população que trabalham no setor agropecuário com 18 anos ou mais. Esses foram obtidos no Atlas do Desenvolvimento Humano no Brasil.

O Atlas do Desenvolvimento Humano no Brasil engloba o Atlas do Desenvolvimento Humano nos Municípios e o Atlas do Desenvolvimento Humano nas Regiões Metropolitanas. O Atlas é uma plataforma de consulta ao Índice de Desenvolvimento Humano Municipal (IDHM) e mais de 200 indicadores de demografia, educação, renda, trabalho, habitação e vulnerabilidade, com dados extraídos dos Censos Demográficos de todo o território brasileiro (ATLAS BRASIL, 2018).

Dados do Produto Interno Bruto dos Municípios (PIB) foram obtidos pelo o Sistema IBGE de Recuperação Automática (SIDRA), para isto foi selecionado o tema sistema de conta

nacionais para obter a variável. Para a análise do fluxo migratório interno e externo foram utilizados apenas dos censos demográficos, seguindo a metodologia de Maria (2018).

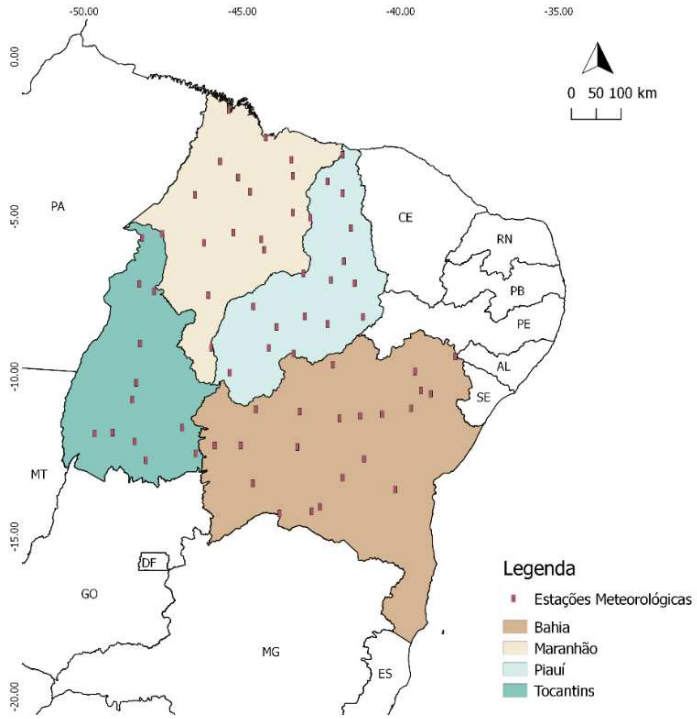
2.2.3 Dados Físicos

Os dados de caracterização física incluem: precipitação, temperatura, umidade relativa do ar, declividade, relevo e tipo de solo. Esses dados são referentes aos cincos fatores de aptidão agrícola utilizados na metodologia de Ramalho Filho e Beek (1995): deficiência de fertilidade, deficiência de água, excesso de água, suscetibilidade à erosão e impedimento à mecanização.

Os dados do clima foram obtidos a partir da plataforma computacional de controle de qualidade, preenchimento de falhas e espacialização de dados meteorológicos, denominada CONPREES (MONTEIRO et al., 2016), da Embrapa Informática Agropecuária.

Através do CONPRESS foram selecionados e extraídos dados históricos de temperatura máxima, mínima e média, precipitação acumulada e umidade relativa do ar para cada ano, dos estados do Maranhão, Piauí, Tocantins e Bahia. A plataforma ofereceu dados de todas as estações meteorológicos dos estados, porém não há uma estação em cada município. Então com o posicionamento geográfico (latitude, longitude e altitude) foi feita uma interpolação. A Figura 9 ilustra as 66 estações meteorológicas dos estados.

Figura 9 - Estações meteorológicas dos estados Maranhão, Tocantins, Piauí e Bahia



A interpolação espacial envolve o processo de determinar valores desconhecidos, ou não amostrados, de um atributo contínuo usando valores conhecidos ou amostrados (MIRANDA, 2005). Essa ferramenta é muito utilizada para elaboração do mapeamento de precipitação, elevação, zona da saturação de água subterrânea e densidade populacional (QGIS, 2018). O método de interpolação utilizado nesse trabalho foi o Inverso da Distância à Potência (IDW), esse método estima um valor para um local não amostrado como uma média dos valores dos dados dentro de uma vizinhança (MIRANDA, 2005). Com isso, foi possível determinar os valores de temperatura máxima, mínima e a média, precipitação e umidade relativa de cada município da região.

A declividade e o relevo foram obtidos do produto SRTM (*Shuttle Radar Topography Mission*). A base SRTM consiste no modelo digital de elevação (MDE), obtido por um sensor remoto que esteve a bordo do ônibus espacial americano em missão realizada. A resolução espacial do produto SRTM é de 30 metros, ou seja, para cada área de 30 por 30 metros do território nacional, dispõe-se de uma medida precisa da altitude do terreno (USGS, 2018). As condições de relevo foram determinadas de acordo com a Embrapa (1979) (Tabela 2). Para o conjunto de dados, neste trabalho, foram criadas seis variáveis indicando a porcentagem do munícipio ocupada por cada uma das classes de relevo.

Tabela 2 - Classes de declividade

Declividade (%)	**Relevo**
0 – 3	Plano
3 – 8	Suave ondulado
8 – 20	Ondulado
20 – 45	Forte ondulado
45 – 75	Montanhoso
> 75	Escarpado

Fonte: EMBRAPA (1979)

Para o solo, o IBGE disponibiliza o mapa de pedologia com a escala cartográfica de 1:250.000 que permite um nível de detalhamento regional. As informações estão em arquivos vetoriais do tipo “shapefile”, e podem ser editadas em softwares de Sistemas de Informações Geográficas (SIG), com isso, permitindo o cruzamento com diferentes temas (IBGE, 2018). A classe do solo do município foi determinada pela predominância na área total do município.

4.3 Preparação dos Dados

Na fase de preparação dos dados, entendeu-se a atribuição dos dados, para isso foi utilizada uma planilha do Excel, para visualizar as tabelas. O conjunto de dados tem 327 instâncias (linhas) e 149 atributos (colunas). As instâncias são os 327 municípios da região do MATOPIBA. A região é composta por 337 municípios, mas 10 foram retirados do conjunto de dados pois sua área total está dentro de reserva indígena ou unidade de conservação.

Os atributos incluem dados físicos (temperatura mínima, temperatura média, temperatura máxima, precipitação acumulada anual, umidade relativa, declividade e solo), sociais (IDHM, taxa de analfabetismo, renda per capita, porcentagem da população que trabalham no setor agropecuário, pib e o fluxo migratório) e o econômico (área plantada, quantidade produzida e rendimento médio das culturas).

Após organizar o conjunto de dados na planilha, prosseguiu-se com a preparação dos dados, notadamente consistência dos dados, verificação da existência de valores faltantes, seleção dos atributos mais relevantes e a possibilidade da aplicação de procedimentos para balanceamentos de classes. Com isso, alguns anos foram retiradas variáveis do conjunto de dados pois apresentavam 60% de dados faltantes. Não houve a necessidade do balanceamento de classes.

4.3.1 Especificação do Atributo Meta

O atributo meta, também conhecido por classe ou variável dependente, é um atributo especial para problemas de classificação, e deve ser categórico, ou seja, não numérico (LOPES, 2013). Neste caso, ele assume valores de duas classes: apta e não apta. Essa variável foi obtida de acordo com a classificação de aptidão agrícola feita por Lumbreras et al. (2015), baseada na metodologia de Ramalho Filho e Beek (1995). O mapa de aptidão agrícola produzido por Lumbreras et al. (2015) é bastante detalhado, com a escala de 1:500.000.

Para realizar esse trabalho foram consideradas aptas as áreas que, no mapeamento produzido por Lumbreras et al. (2015), apresentavam aptidão boa ou regular para lavouras, em qualquer um dos três níveis de manejo (A, B e C). Já as classes restrita e inapta para lavoura, áreas pastagem plantada, silvicultura e pastagem natural, representam áreas não aptas para a agricultura. E assim, foi possível definir a classe de cada município da região do MATOPIBA. A Tabela 3 exemplifica o agrupamento das classes de aptidão agrícola usada neste trabalho.

Tabela 3 - Agrupamento das classes de aptidão agrícola

Classe (atributo meta)	Apta	Não apta
Classes de aptidão agrícola	1(ab)C, 1(b)C , 1Abc, 1aBC, 1aBc, 1Abc, 1(ab)C 2(a)b(c), 2(a)bc, 2(ab)c, 2(b)c, 2ab(c), 2abc, 2c	3(a), 3(ab), 3(abc), 3(b), 3(bc), 3(c) 4(p), 5(n), 6, Corpo d' água, Unidade de Conservação

A união das classes boa e regular deve-se ao fato das duas classes apresentarem condições para a produção agricultura sustentada e foi feita a partir da Tabela 4 elaborada por Pereira e Lombardi Neto (2004), baseada na metodologia de Ramalho Filho e Beek (1995).

Tabela 4 - Quadro-guia de avaliação da aptidão agrícola das terras

Aptidão Agrícola			Graus de limitação* das condições agrícolas das terras para os níveis de manejo A, B e C														Tipo de Utilização Indicado
Grupo	Subgrupo	Classe	Deficiência de Fertilidade			Deficiência de Água			Excesso de Água			Suscetibilidade à Erosão			Impedimento à Mecanização		
			A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	
1	1ABC	Boa	0	1	0	1	1	1	1	1	1	2	1	0	2	1	0	
2	2abc	Regular	1	1	1	2	2	2	2	2	1	2	2	1	3	2	1	Lavouras
3	3(abc)	Restrita	2	2	2	3	2	2	3	2	2	3	2	1	3	3	2	
4	4P	Boa		2			2			3			2			3		Pastagem Plantada
	4p	Regular		3			3			3			3			3		
	4(p)	Restrita		3			3			3			4			3		
5	5S	Boa		2			2			1			3			3		Silvicultura e/ou Pastagem Natural
	5s	Regular		3			3			1			3			3		
	5(s)	Restrita		4			3			2			4			4		
	5N	Boa	2			2			3			3			4			
	5n	Regular	3			3			3			3			4			
	5(n)	Restrita	4			4			3			3			4			
6	6FF 6f f	Sem apt. Agrícola	Restrição de ordem Legal (áreas de proteção por Lei) Restrição por condições agroambientais (relevo e/ou solo e/ou clima)															Preservação da Fauna e Flora

Fonte: Quadro-guia de Ramalho e Beek (1995) elaborado por Pereira e Lombardi Neto (2004)

* Graus de limitação: 0 = Nulo, 1 = Ligeiro, 2= Moderado, 3 = Forte e 4 = Muito Forte

As diferenças entre as classes boa e regular são pelos fatores de limitação, descritas por Ramalho e Beek (1995):

i) O fator deficiência de fertilidade na classe boa é nula, as terras possuem elevadas reservas de nutrientes e na classe regular é ligeira, terras com boa reserva de nutrientes para as plantas.

ii) Deficiência de água para classe boa é ligeira, terras com deficiência de água pouco acentuada e com a classe regular é uma acentuada deficiência de água com o grau de limitação moderado.

iii) No fator excesso de água, a classe boa tem um grau de limitação ligeiro e a regular é moderado, as terras são moderadamente e imperfeitamente drenadas, respectivamente.

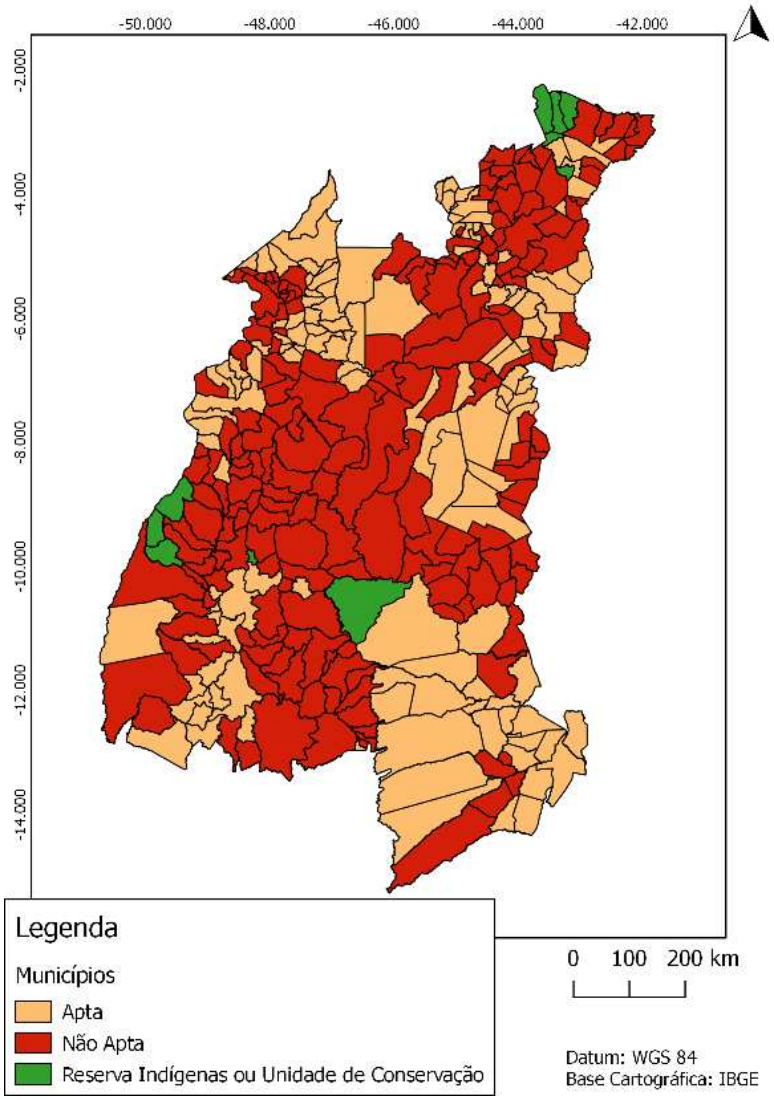
iv) O grau de limitação por suscetibilidade varia de nulo, ligeiro e moderado, as terras não suscetíveis, pouca e moderada suscetibilidade, respectivamente. A variação da declividade é de 0 a 13% requerendo práticas de controle à erosão desde o início da sua utilização agrícola.

v) Impedimento à mecanização, como o próprio nome indica, refere-se às condições apresentadas pelas terras para o uso de máquinas e implementos agrícolas. Esse fator é relevante para o nível de manejo C, o mais avançado, no qual é previsto o uso de máquinas e implementos agrícolas nas diversas fases da operação agrícola. Por isso no manejo C, o grau de limitação é nulo e ligeira, as terras permitem, em qualquer época do ano e durante quase todo ano, respectivamente, o emprego de máquinas agrícolas.

A classe restrita foi nomeada como “não apta”, pois apresenta limitações fortes para produção agrícola sustentada. Essas limitações reduzem a produtividade e aumentam o consumo de insumo. E a classe inapta, também foi nomeada como “não apta”, por apresentar condições que excluem a produção agrícola sustentada e usos menos intensivos (pastagem plantada, silvicultura ou pastagem natural), para fins não agrícolas (RAMALHO; BEEK, 1995).

Por esses fatores, formou-se o conjunto de dados de 327 instâncias, onde 196 representaram a classe “não apta”, 131 a classe “apta”. A Figura 10 mostra a classe que predomina no município de acordo com o agrupamento da Tabela 3.

Figura 10 - Classes de aptidão agrícola definida por Lumbreras et al. (2015) baseado na metodologia de Ramalho Filho e Beek (1995).



4.4 Modelagem

As técnicas de modelagem escolhidas para realizar este trabalho foram classificação, por meio do algoritmo J48, amplamente conhecido como C4.5 (QUINLAN, 1993) (árvore de decisão), *RandomFlorest* e SVM (Support Vector Machines ou Máquinas de Vetores Suporte). Para seleção de atributos foram utilizados os métodos: Qui-quadrado (χ^2), Ganho de Informação, CFS *(Correlation Feature Selection)* e Wrapper.

Com o conjunto de dados final, técnicas de classificação e seleção de atributos foram selecionadas e aplicadas ao conjunto de dados de atributos físicos, sociais e econômicos da região do MATOPIBA. Para atingir o objetivo deste trabalho, foi preciso fazer vários

experimentos e ajustes no número de “fold” em cada uma das técnicas, a fim de fornecer o melhor resultado em termos de acurácia, índice kappa e precisão das classes (descritos no item 5.5 – Validação).

A elaboração da árvore de decisão está relacionada com a etapa de seleção do atributo associado ao teste de divisão de um nó interno. O método padrão utilizado no algoritmo C4.5 para essa seleção é a avaliação do ganho de informação baseado no cálculo da entropia do atributo, antes e depois da seleção do mesmo. A entropia remete à aleatoriedade ou desordem ou seja, quanto maior a entropia menor a informação da amostra. Essa é a base da teoria da informação (QUINLAN, 1993).

A técnica de Random Forest, desenvolvida por Breiman (2001), consiste na combinação de várias árvores de decisão formadas a partir de amostras aleatórias do conjunto de dados de treinamento. O esboço do algoritmo Random Forest pode ser visto na Figura 11.

Figura 11 - Sequência básica de passos do algoritmo Random Forest (BREIMAN, 2001).

Dado um conjunto de dados $X = x_1, x_2, \ldots, x_j$ e $Y = y_1, y_2, \ldots y_k$.

Para $b = 1, 2, 3, \ldots B$, repita:

(a) Cria uma amostra *bootstrap* (X_b, Y_b) com n exemplos de (X, Y).

(b) Ajusta uma árvore de decisão f^b para o conjunto de treinamento (X_b, Y_b), utilizando m atributos para a escolha de cada nó.

Fim de repetição.

Gera o modelo final: $\hat{f}(x) = \sum_{b=1}^{B} f^b(x)$, que calcula os votos obtidos por cada modelo f^b, resultando uma classificação final de acordo com a votação majoritária.

4.5 Validação

Os modelos gerados foram avaliados por meio de validação cruzada, que funciona da seguinte forma: o conjunto de dados é dividido aleatoriamente em r partições mutuamente exclusivas (folds) de tamanho de exemplos semelhantes. Os exemplos nos (r - 1) folds são usados para treinamento e a hipótese induzida é testada no folds remanescente. Este processo é repetido r vezes, cada vez considerando um fold diferente para teste (MONARD; BARANAUSKAS, 2005; WITTEN et al., 2011). No trabalho, a validação cruzada foi com 5-folds, já que o conjunto de dados continha um número pequeno de instâncias. A taxa de erro é calculada no conjunto de validação. Assim, o procedimento de foi executado um total de cinco vezes em diferentes conjuntos de treinamento (cada um dos quais tem muito em comum).

Finalmente, as cinco estimativas de erro são calculadas para gerar um erro geral (WITTEN et al., 2011).

A matriz de confusão é um dos resultados que o classificador gera para avaliação do desempenho do modelo e é baseada nas contagens de registros de testes previstos correta e incorretamente pelo modelo (TAN et al., 2009).

A Tabela 5 representa a matriz de confusão para um problema de duas classes, denominadas C+ (classe positiva) e C- (classe negativa) (MONARD e BARANAUSKAS, 2002 citado por MEGETO, 2012). Existem quatro possíveis resultados em relação à classificação de exemplos.

Tabela 5 - Matriz de confusão

		Predita		
		C+	C-	Total
Verdade	C+	VP	FN	P
	C-	FP	VN	N
Total		P'	V'	P+N

Baseado nas entradas da matriz de confusão, o número total de previsões corretas (VP e VN) feita pelo modelo é apresentado na diagonal principal, o que significa bons modelos quando apresentados altos valores e o número de previsões incorretas é (FN e FP), na diagonal secundária deve apresentar valores baixos para um bom modelo de classificação.

A coluna Total da Tabela 4 é a soma dos casos positivos (P) e a soma dos casos negativos (N) existentes no conjunto de treinamento. A linha Total, P' é a soma de casos positivos e N' soma de casos negativos do modelo de classificação.

A matriz de confusão fornece informações necessárias para determinar o desempenho do modelo. A taxa de acerto ou acurácia é a porcentagem de exemplos classificados corretamente pelo classificador. A fórmula é representada na Equação 6.

$$Taxa\ de\ acerto = \frac{VP+VN}{P+N} \quad (6)$$

O desempenho do modelo pode ser expresso também pela taxa de erro que é 1- a taxa de acerto. Representa na Equação 7.

$$Taxa\ de\ erro = \frac{FP+FN}{P+N} \quad (7)$$

A sensitividade (taxa de verdadeiros positivos) e a especificidade (taxa de verdadeiros negativos) são outras medidas que a matriz de confusão pode oferecer. Representa nas Equações 8 e 9, respectivamente.

$$\text{sensitividade} = \frac{VP}{P} \tag{8}$$

$$\text{especificidade} = \frac{VN}{N} \tag{9}$$

Para medir o desempenho do classificador também é usado o índice Kappa (COHEN, 1960). O valor do índice Kappa varia entre 0 e 1, quanto maior o valor melhor a medida de concordância entre as classes preditivas e observadas.

O cálculo do índice Kappa é feito a partir da matriz de confusão. Representado na Equação 10, onde P_o é a proporção de acertos das classes, calculado na Equação 11, e P_e é a probabilidade de concordância esperada considerando eventos independentes, Equação 12.

$$\mathrm{K} = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}} \tag{10}$$

$$\mathrm{Po} = \frac{VP + VN}{P + N} \tag{11}$$

$$\mathrm{Pe} = \frac{(P'P) + (N'N)}{(P+N)^2} \tag{12}$$

A interpretação do desempenho dos modelos foi elaborada por Landis e Koch (1977) a partir do índice Kappa e é representado na Tabela 6.

Tabela 6 - Desempenho de modelos de classificação a partir do Kappa, segundo Landis e Koch (1977)

Índice Kappa	**Qualidade**
< 0,00	Péssima
0,00 – 0,20	Ruim
0,21 – 0,40	Razoável
0,41 – 0,60	Boa
0,61 – 0,80	Muito Boa
0,81 – 1,00	Excelente

4.5 Software utilizados

O software ArcGis (ESRI, 2018), uma ferramenta de Sistema de Informações Geográficas (SIG), foi utilizado na fase de preparação de dados bem como na elaboração de mapas temáticos. O ArcGis varia em termos de custo, uso e nível de funcionalidade, mas deve oferecer um conjunto mínimo de funções: Data entry, Data management, Thematic mapping, Data Analysis e Map layout (PRICE, 2016).

Os Sistemas de Informação Geográficas são utilizados para coletar, gerenciar e analisar dados com o objetivo de fornecer informações em diferentes áreas (planejamento do uso da terra, gestão ambiental analises, análise social, marketing empresarial) para dar suporte à tomada de decisão (PRICE, 2016).

O programa computacional utilizado para elaboração dos modelos preditivos e para seleção dos atributos foi o software WEKA 3.8 (*Waikato Environment for Knowledge Analysis*) (WEKA, 2018). O software foi desenvolvido na Universidade de Waikato, na Nova Zelândia, foi escrito em Java e é distribuído sob os termos de licença GNU que tem como característica ser livre e gratuito, de fácil instalação (WITTEN et al., 2011). Ele fornece uma interface uniforme para muitos algoritmos de aprendizado de máquina, junto com métodos de pré e pós-processamento. O ambiente de trabalho inclui tarefas para os problemas de mineração de dados: regressão, classificação, agrupamento, associação e seleção de atributos (WITTEN et al., 2011) e os parâmetros podem ser ajustados conforme a necessidade do conjunto de dados.

5. RESULTADOS

5.1 Seleção de Atributos

Os algoritmos de seleção de atributos foram utilizados para eliminar aqueles que não apresentam um grande poder predito para as classes. A Tabela 7 apresenta os atributos selecionados por meio dos quatros métodos em estudo (CFS, Wrapper, Ganho de Informação e Qui-quadrado (χ^2)). De modo geral, foram mantidas as configurações padrões do software WEKA para executar essa etapa, apenas para o Wrapper foi executado conjuntamente com o algoritmo J48.

Ao analisar os resultados da seleção de atributos, o solo foi aquele que apresentou a maior incidência na seleção. O atributo apareceu em todos os métodos, sendo assim considerado aquele com maior poder preditivo. Isso mostra que para determinar modelos preditivos de aptidão agrícola é indispensável o uso do atributo, considerando-o o mais importante.

O CFS, Ganho de Informação e o Qui-quadrado destacaram os atributos de área plantada, rendimento médio, quantidade produzida das culturas de soja e milho. E também, os atributos físicos de temperatura máxima, média e mínima e precipitação acumulada. O Wrapper selecionou um número menor de atributos. No entanto, também destacou os atributos: umidade relativa mínima (2009) e renda per capita (2000).

Tabela 7 – Atributos físicos, econômicos e sociais destacados pelos métodos de seleção de atributos como os mais relevantes do conjunto de dados.

Atributo	Método			
	CFS	Wrapper	Ganho de Informação	Qui-quadrado (χ^2)
Área Plantada -2008 (Milho 1ªSafra)			X	X
Área Plantada -2011 (Milho 1ªSafra)		X	X	X
Área Plantada -2012 (Milho 1ªSafra)			X	X
Área Plantada -2013 (Milho 1ªSafra)			X	X

Área Plantada -2014 (Milho 1ªSafra)			X	X
Área Plantada -2015 (Milho 1ªSafra)			X	
Área Plantada -2014 (Soja)	X		X	X
Área Plantada -2015 (Soja)		X		
Precipitação 2015				X
Precipitação 2008	X		X	X
Precipitação 2007			X	X
Precipitação 2005			X	
Quantidade Produzida - 2008 (Milho 1ªSafra)	X		X	X
Quantidade Produzida - 2010 (Milho 1ªSafra)			X	X
Quantidade Produzida - 2012 (Milho 1ªSafra)			X	X
Quantidade Produzida - 2014 (Milho 1ªSafra)			X	X
Quantidade Produzida - 2014 (Soja)			X	X
Quantidade Produzida - 2015 (Soja)		X		
Quantidade Produzida - 2016 (Soja)			X	X
Renda per Capita (2000)		X		
Rendimento Médio -2006 (Milho 1ªSafra)		X		
Rendimento Médio -2011 (Milho 1ªSafra)	X		X	X
Solo	X	X	X	X
Temperatura Máxima 2006			X	X
Temperatura Máxima 2008			X	X
Temperatura Máxima 2010			X	X
Temperatura Máxima 2014			X	X

Temperatura Máxima 2013		X		
Temperatura Média 2006	X		X	X
Temperatura Média 2007			X	X
Temperatura Média 2009		X	X	X
Temperatura Média 2011			X	X
Temperatura Mínima 2005			X	X
Temperatura Mínima 2006			X	X
Temperatura Mínima 2007			X	X
Temperatura Mínima 2008	X		X	X
Temperatura Mínima 2012	X		X	X
Temperatura Mínima 2013			X	X
Umidade Relativa 2009		X		

Ao analisar os resultados da Tabela 6, os dados sociais de IDHM, Taxa de analfabetismo, porcentagem da população que trabalham no setor agropecuário, PIB e o Fluxo migratório foram os únicos que não apareceram em nenhum dos métodos, podendo ser considerados os atributos com menor poder preditivo. A exceção foi o atributo renda per capita, único atributo dos dados sociais, selecionado apenas método Wrapper.

Outro atributo que também não apareceu foi o relevo. A condição do relevo está relacionada à suscetibilidade e à erosão, ou seja, ao desgaste que a superfície do solo poderá sofrer quando submetida ao uso e sem medidas conservacionistas. Além da condição do relevo, condições do solo, climáticas e da cobertura vegetal são importantes para esse fator (RAMALHO FILHO; BEEK, 1995).

O relevo também está relacionado diretamente o impedimento à mecanização que, junto às condições de drenagem, profundidade, textura e tipo de argila impactam o rendimento médio da máquina agrícola em uma determinada área (RAMALHO FILHO; BEEK, 1995).

O Ganho de Informação e o Qui-quadrado são métodos em que cada atributo é avaliado com relação à classe (atributo-meta). No final, foi feito uma classificação que privilegia os atributos com menor entropia e maior valor do Qui-quadrado para os métodos Ganho de Informação e Qui-quadrado, respectivamente. A Tabela 8 apresenta essa classificação dos dois métodos.

O atributo solo ocupa a primeira colocação da tabela de classificação, sendo o

atributo com maior relevância para determinar a aptidão agrícola de municípios. Em seguida, os atributos climáticos, temperatura, rendimento médio, precipitação, área plantada, quantidade produzida. Nas atividades humanas por ser a base de sustentação de um grande número delas.

Tabela 8 - Classificação dos atributos pelos métodos Ganho de Informação e Qui-quadrado.

Classif.	Ganho de Informação	Qui-quadrado
1	Solo	Solo
2	Temperatura mínima 2008	Temperatura mínima 2009
3	Temperatura média 2009	Temperatura mínima 2012
4	Temperatura mínima 2012	Temperatura média 2009
5	Temperatura média 2006	Rend. Médio - 2011 (Milho 1ª safra)
6	Rend. Médio - 2011 (Milho 1ª safra)	Temperatura média 2011
7	Temperatura média 2011	Temperatura média 2006
8	Temperatura mínima 2013	Temperatura mínima 2013
9	Precipitação 2008	Temperatura máxima 2014
10	Temperatura mínima 2006	Temperatura mínima 2006
11	Temperatura máxima 2014	Temperatura média 2007
12	Temperatura média 2007	Precipitação 2008
13	Temperatura mínima 2007	Precipitação 2015
14	Precipitação 2015	Temperatura máxima 2008
15	Temperatura máxima 2008	Temperatura máxima 2006
16	Temperatura máxima 2006	Temperatura mínima 2007
17	Temperatura mínima 2005	Temperatura máxima 2010
18	Quantidade Produzida - 2008 (Milho 1ª safra)	Temperatura mínima 2005
19	Temperatura máxima 2010	Área Plantada - 2012 (Milho 1ª safra)
20	Precipitação 2007	Quantidade Produzida - 2012 (Milho 1ª safra)
21	Quantidade Produzida - 2012 (Milho 1ª safra)	Quantidade Produzida - 2008 (Milho 1ª safra)
22	Área Plantada - 2012 (Milho 1ª safra)	Precipitação 2007
23	Quantidade Produzida - 2010 (Milho 1ª safra)	Quantidade Produzida - 2014 (Milho 1ª safra)
24	Área Plantada - 2013 (Milho 1ª safra)	Área Plantada - 2014 (Milho 1ª safra)
25	Área Plantada - 2011 (Milho 1ª safra)	Quantidade Produzida - 2010 (Milho 1ª safra)
26	Quantidade Produzida - 2014 (Milho 1ª safra)	Área Plantada - 2013 (Milho 1ª safra)
27	Área Plantada - 2008 (Milho 1ª safra)	Área Plantada - 2011 (Milho 1ª safra)
28	Área Plantada - 2014 (Milho 1ª safra)	Área Plantada - 2008 (Milho 1ª safra)
29	Área Plantada - 2015 (Milho 1ª safra)	Área Plantada - 2015 (Milho 1ª safra)
30	Área Plantada - 2014 (Milho 1ª safra)	Área Plantada - 2015 (Soja 1ª safra)
31	Quantidade Produzida - 2016 (Soja)	Quantidade Produzida - 2016 (Soja)
31	Quantidade Produzida - 2014 (soja)	Quantidade Produzida - 2014 (Soja)

A metodologia de Ramalho e Beek (1995), para interpretação de levantamentos de solos, seguiu as orientações do “Soil survey manual” (Estados Unidos, 1951) e a metodologia da FAO (1976), onde recomendaram que a avaliação da aptidão agrícola das terras seja baseada em resultados de levantamentos sistemáticos, com o suporte de vários atributos das terras: solo, clima, vegetação e geomorfologia. Isso explica a classificação dos atributos mais relevantes feito pelos métodos de seleção de atributo, relacionados aos cincos fatores da avaliação de aptidão agrícola, pela metodologia de Ramalho Filho e Beek (1995):

5.2 Classificação

Para a etapa de modelagem dos classificadores foram feitos testes com três algoritmos: Árvore de Decisão (J48), *Random Forest* e SVM *(Support Vector Machine).* Nesta etapa, foram feitos testes sem seleção de atributos para cada algoritmo e em seguida, com a análise dos atributos mais importantes, aplicou-se os algoritmos de classificação. Os modelos gerados foram avaliados por meio do método de validação cruzada (*cross-validation*) com 5-folds, além das métricas: acurácia, índice Kappa, precisão por classe, taxa verdadeiro positivo e falso positivo por classe.

A Tabela 9 apresenta os resultados das métricas para o classificador, Árvore de Decisão (J48), com e sem a seleção de atributos.

Tabela 9 - Classificação dos atributos por árvore de decisão (J48) para diferentes métodos de seleção de atributos

Métodos	Acurácia	Índice Kappa	Taxa de Verdadeiro Positivo		Taxa de Falso Positivo		Precisão		N° de regras
			Classe “não apta"	Classe "apta"	Classe "não apta"	Classe "apta"	Classe "não apta"	Classe "apta"	
Sem Seleção	78,28	0,54	0,852	0,679	0,321	0,148	0,799	0,754	37
Qui-quadrado	79,82	0,58	0,821	0,763	0,237	0,179	0,839	0,741	35
Ganho de Informação	79,82	0,58	0,821	0,763	0,237	0,179	0,839	0,76	35
Wrapper	**84,40**	**0,67**	**0,878**	**0,794**	**0,206**	**0,122**	**0,864**	**0,813**	**31**
CFS	74,92	0,48	0,776	0,71	0,29	0,224	0,8	0,679	31

Os métodos Qui-quadrado e Ganho de Informação obtiveram valores semelhantes como pode ser visto na Tabela 9. A única diferença foi na precisão da classe “apta”, isto é, 0,74

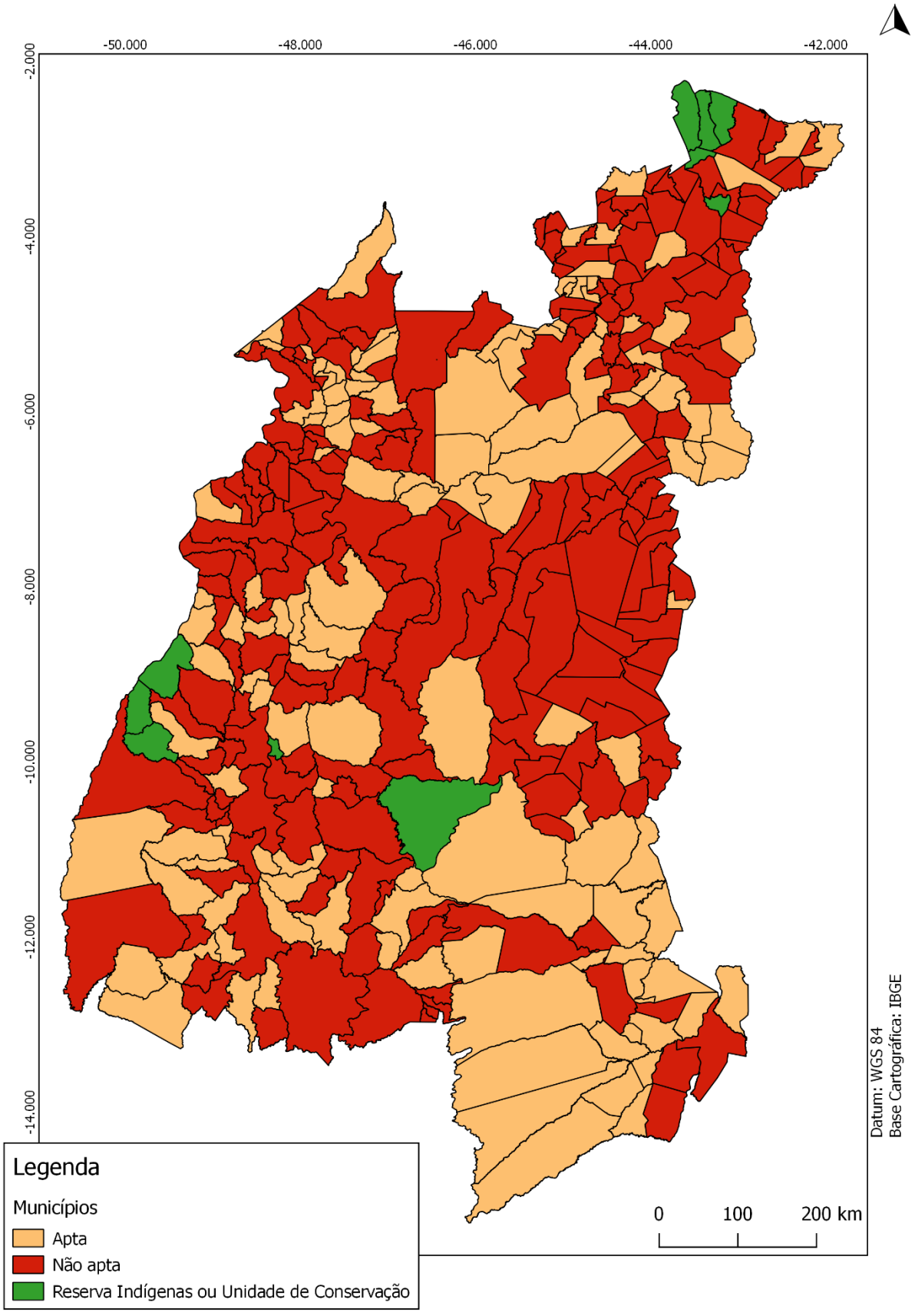
e 0,76, para os métodos Qui-quadrado e Ganho de Informação, respectivamente.

Por outro lado, o algoritmo J48, sem seleção e com a seleção de atributos CFS, gerou modelos preditivos com performance inferior aos demais. Sem seleção de atributos, a acurácia é de 78,28% e o índice Kappa com 0,54. A precisão da classe "não apta" é de 0,79 e classe "apta", 0,75. Neste caso, o número de regras foi a maior com 37. Sem a seleção de atributos, o banco de dados pode conter atributos redundantes e irrelevantes que reduzem a exatidão da classificação e a qualidade dos grupos encontrados. Atributos redundantes não agregam informação ao modelo gerado, por isso são eliminados pelos métodos de seleção em estudo (TAN et. al., 2009).

Para o método CFS a acurácia foi de 74,92% e o índice Kappa de 0,48, mesmo com os valores inferiores ao demais, o desempenho do índice Kappa é considerado bom, segundo Landis e Koch (1977). O CFS avalia a capacidade preditiva de cada atributo individualmente e o grau de importância entre eles, escolhendo os conjuntos de atributos altamente correlacionados com a classe, mas com baixa inter-correlação entre eles (WITTEN et al., 2011).

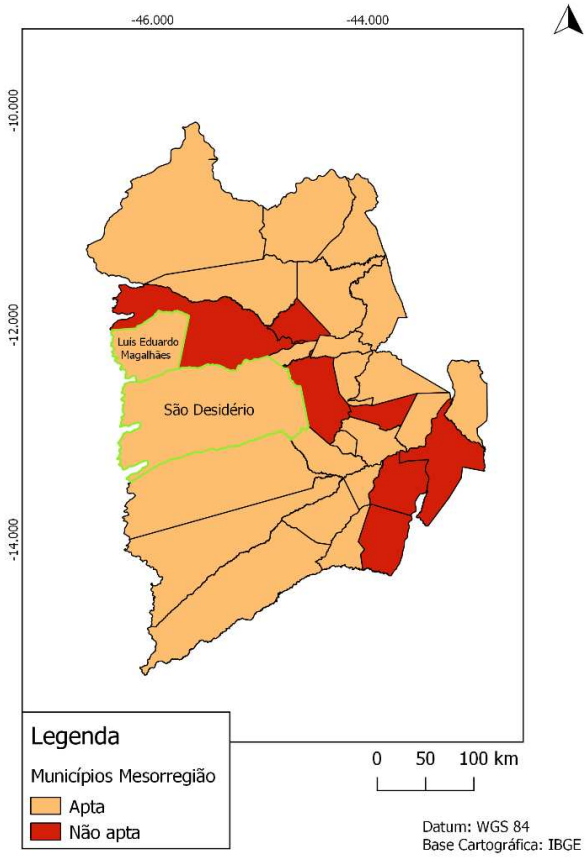
O Wrapper alcançou melhores resultados das métricas por ser executado com algoritmo indução e seus dados de treinamento (HALL, 1999). Para este caso, em particular, foram geradas 31 regras a partir da árvore de decisão. O método selecionou subconjuntos de atributos baseado no algoritmo indicado, J48, e avaliou a qualidade de cada subconjunto com o mesmo algoritmo que foi utilizado na classificação. Com isso, encontrou o melhor conjunto de atributos que se adequa ao algoritmo. A Figura 12 mostra a classificação obtida pelo algoritmo J48 com os atributos selecionados pelo método Wrapper.

Figura 12 - Mapa de aptidão agrícola dos municípios, da região do MATOPIBA, com o classificador árvore de decisão e seleção de atributos Wrapper.



O modelo classificou 41,32% (128 municípios) da área total do MATOPIBA como apta para agricultura, parte dessa área está concentrada na mesorregião do Extremo Oeste Baiano, umas das mesorregiões que mais crescem e se desenvolve no cenário agrícola brasileira (8ª posição na classificação apresentada por Freitas e Maciente (2015)). Em 2000, a área plantada das principais culturas da mesorregião (soja, milho e algodão) era de 861.518 ha, em 2017, esse número passa a ser de 2.031.700, isso representa um aumento de 135,82%. Sano et al. (2011) destacam dois municípios, Luís Eduardo Magalhães, onde 55% da sua área total era utilizada para produção de grãos, e São Desidério com 32%, em 2005. O crescimento agrícola, da mesorregião, pode ser visualizado com as máscaras agrícolas elaboradas por Lorensini et al. (2015). Na Figura 13, a mesorregião do Extremo Oeste Baiano com 23 municípios (82,13% da área total) com aptidão agrícola e 7 municípios sem (17,87 % da área total), com destaque a municípios citados.

Figura 13 - Mapa de aptidão agrícola dos municípios, da mesorregião do Extremo Oeste Baiano com destaque dois municípios que são grandes produtores de grãos, gerado a partir do classificador J48 e o método Wrapper.



Outros parâmetros importantes para a análise são a Taxa de Verdadeiro Positivo (isto é, quando o valor real era "verdade" e foi indicado corretamente) e a Taxa de Falso Positivo (quando o valor real era "verdade" e foi indicado o contrário), ambos apresentados na Tabela 9, gerados pelo software Weka. Os valores da Taxa de Verdadeiro Positivo para a classe "não apta" foram maiores na comparação com a "apta" para os métodos de seleção de atributos, da mesma forma que a Taxa de Falso Positivo. Isso indica que os modelos gerados tenderam a classificar a classe "não apta". Exemplo é mesorregião Sul Maranhense, composta por 19 municípios (IBGE, 2018), considera a 5ª mesorregião em expansão agrícola do Brasil por Freitas e Maciente (2015), com apenas sete municípios (32,74% da área total) com aptidão agrícola.

Entretanto, na comparação Taxa de Verdadeiro Positivo e Taxa de Verdadeiro Negativo do método Wrapper, que obteve a melhor acurácia, os valores da Taxa de Verdadeiro Positivo são superiores, com 0,88 para classe "não apta" e 0,79 para classe "apta", em relação a Taxa de Verdadeiro Negativo, 0,20 e 0,12, respectivamente. Isso significa que o modelo gerado pelo classificador com a o método de seleção de atributo Wrapper está sendo executado corretamente.

A avaliação da matriz de confusão, gerada com os valores das Verdadeiro Positivo e Verdadeiro Negativo, reproduzida na Tabela 10, mostra de forma complementar o comportamento do classificador, confirmando o que foi visto na Tabela 9, sendo que o classificador é preciso e tem taxas muito baixas de falsos positivos.

Tabela 10 - Matriz de confusão da variável resposta "aptidão agrícola" utilizando o algoritmo J48 com método de seleção de atributos Wrapper para validação do modelo.

		Classe Predita		Total Classe
		Não Apta	Apta	Verdadeira
Verdade	Não Apta	172	24	196
	Apta	27	104	131
Total Classe Predita		199	128	327

Com relação ao algoritmo *Random Forest,* foram realizados experimentos alterando o número de árvores a serem construídas e o número de atributos selecionados aleatoriamente para minimizar o erro. Foram utilizadas 500 árvores neste procedimento. Em geral, o erro estabiliza com menos de 500 árvores geradas de acordo com Witten et al. (2011).

A Tabela 11 mostra os resultados das métricas para ao classificador *Random Forest* com e sem seleção de atributos. O método de seleção de atributos, CFS (*Correlation Feature Selection)* apresentou o melhor resultado com a acurácia de 80,42 % e índice Kappa de 0,59, considerando o desempenho do classificador muito bom, segundo Landis e Koch (1977). A precisão da classe "não apta" é de 0,842 e "apta" 0,748. O método de seleção de atributos e o classificador tem uma base semelhante. O CFS começa com o conjunto vazio de atributos e usa a heurística *best-first-search* com um critério de parada de 5 consecutivos subconjuntos e o *Random Forest* combina as previsões feitas por múltiplas árvores de decisões, onde cada árvore é gerada baseada nos valores de um conjunto independente de vetores aleatório (TAN et al., 2009).

O método Wrapper apresentou resultado próximo ao CFS, com acurácia de 80,12% e com índice Kappa 0,58, o que também é considerado muito bom. A precisão das classes "não apta" e "apta" foi de 0,82 e 0,77, respectivamente.

Analisando os métodos Ganho de Informação e Qui-quadrado, os valores das métricas foram semelhantes com a acurácia de 77,06% e o índice Kappa de 0,52. Ao avaliar os erros, o método Qui-quadrado foi mais propenso a indicar falsos positivo para classe "não apta" (isto é, quando o valor real era "positivo" e foi indicado o contrário), com 0.328 e com a precisão da classe mais baixa com 0,792 e modelo com o Ganho de Informação também erra mais entre os positivos (isto é, indica haver mais valores negativos do que realmente existem) com a taxa da classe "apta" com 0,672.

Tabela 11 - Comparação das técnicas de seleção de atributos para o algoritmo *Random Forest*

Métodos	Acurácia	Índice Kappa	Taxa de Verdadeiro Positivo		Taxa de Falso Positivo		Precisão	
			Classe "não apta"	Classe "apta"	Classe "não apta"	Classe "apta"	Classe "não apta"	Classe "apta"
Sem Seleção	78,28	0,53	0,827	0,718	0,282	0,173	0,814	0,734
Qui-quadrado	77,06	0,52	0,837	0,672	0,328	0,163	0,792	0,733
Ganho de Informação	77,04	0,52	0,821	0,763	0,237	0,179	0,839	0,741
Wrapper	80,12	0,58	0,857	0,718	0,282	0,143	0,820	0,77
CFS	**80,42**	**0,59**	**0,842**	**0,748**	**0,252**	**0,158**	**0,842**	**0,748**

A Tabela 12 é a matriz de confusão para o método de seleção de atributos CFS, apresentada abaixo. Os valores das Taxa de Verdadeiro Positivo das classes “apta” e “não aptas”, 0,748 e 0,842, e as Taxas de Falso Positivo, 0,252 e 0,158, respectivamente.

Tabela 12- Matriz de confusão da variável resposta “aptidão agrícola” utilizando o algoritmo *Random Forest* com método de seleção de atributos CFS para validação do modelo.

		Classe Predita		Total Classe
		Não Apta	Apta	Verdadeira
Verdade	Não Apta	165	31	196
	Apta	33	98	131
Total Classe Predita		198	129	327

Por último, os resultados obtidos utilizando o algoritmo SVM, mostram que os quatro métodos de seleção de atributos têm proporções similares de acertos (77,06%) e o índice Kappa (0,53). O que varia entre eles, método Qui-quadrado e os demais, são a proporção de falsos positivos, verdadeiros positivos e a precisão, como pode ser visto na Tabela 13.

Tabela 13 - Comparação das técnicas de seleção de atributos para o algoritmo SVM

Métodos	Acurácia	Índice Kappa	Taxa de Verdadeiro Positivo		Taxa de Falso Positivo		Precisão	
			Classe “não apta"	Classe "apta"	Classe "não apta"	Classe "apta"	Classe "não apta"	Classe "apta"
Sem Seleção	**80,73**	**0,60**	**0,821**	**0,779**	**0,221**	**0,179**	**0,847**	**0,761**
Qui-quadrado	77,06	0,53	0,837	0,672	0,328	0,163	0,792	0,733
Ganho de Informação	77,06	0,53	0,765	0,779	0,221	0,235	0,838	0,689
Wrapper	77,06	0,53	0,765	0,779	0,221	0,235	0,838	0,689
CFS	77,06	0,53	0,765	0,779	0,221	0,235	0,838	0,689

De acordo com a Tabela 13, nota-se que o modelo gerado sem seleção de atributos tem acurácia superior aos demais gerados com seleção de atributos. A razão é que o algoritmo SVM apresenta bons resultados com dados de alta dimensionalidade e evita problema da dimensionalidade (TAN et. al., 2009).

5.3 Comparação da metodologia proposta e a metodologia de Ramalho Filho e Beek (1995)

Na elaboração da metodologia de Ramalho Filho e Beek (1995), os autores pontuaram que a metodologia seguirá se aprimorando ao longo do tempo e é passiva às alterações com novas pesquisas realizadas como neste trabalho.

> Todas as interpretações são elaboradas com base em classificações técnicas, com finalidades bem definidas que retratam o nível tecnológico do momento em que são feitas. Por isso, tanto a metodologia como as classificações em que são baseadas as interpretações podem ser substituídas e atualizadas à medida que os conhecimentos científicos e tecnológicos evoluem. Entretanto, os levantamentos de solos, baseados em classificações naturais, são de caráter bem mais duradouro, servindo de base a novas interpretações fundamentadas nos resultados mais atuais da pesquisa (Ramalho Filho e Beek, 1995)

A metodologia consolidada de Ramalho Filho e Beek (1995) leva em consideração cinco fatores: deficiência de fertilidade, deficiência de água, excesso de água ou deficiência de oxigênio, suscetibilidade à erosão e impedimentos à mecanização. Em suma, esses fatores estão relacionados a dados físicos. Por outro lado, a metodologia proposta neste trabalho considera não somente dados físicos, mas também econômicos e sociais. A Tabela 14 apresenta uma comparação dos dados utilizados nas duas metodologias. Na coluna, Metodologia Proposta encontra-se o conjunto de variáveis selecionadas pelo modelo preditivo (algoritmo J48) juntamente com o método de seleção de atributos Wrapper.

Tabela 14 - Comparação dos dados utilizados nas metodologias

Dados Analisados	Metodologia consolidada (Ramalho Filho e Beek (1995))	Metodologia Proposta
Textura	X	
Estrutura	X	
Profundidade efetiva	X	
Capacidade de permuta de cátions	X	
Saturação de bases	X	
Teor de matéria orgânica	X	
Condutividade elétrica	X	
pH	X	
Solo	X	X
Drenagem	X	
Pedregosidade	X	
Rochosidade	X	
Relevo	X	
Temperatura	X	X
Umidade	X	X
Pluviosidade	X	X
Evapotranspiração	X	
Cobertura vegetal	X	
Quantidade Produzida		X
Área Plantada		X
Rendimento Médio		X
Renda per Capita		X

Na metodologia de Ramalho Filho e Beek (1995) são necessárias dezoito variáveis, dentro dos cinco fatores, para determinar a aptidão agrícola em unidade de terra e apresenta maior detalhamento dentro do município. O método consolidado apresenta seis grupos (lavoura com aptidão restrita, regular e boa, pastagem plantada, silvicultura e/ou pastagem natural e preservação da flora e a fauna), três subgrupos (nível de manejo A, B e C) e quatro classes (boa, regular, restrita e inapta) para cada tipo de utilização indicado.

Para determinar os grupos, subgrupos e classes foram analisadas as dezoito variáveis, obtidas diretamente de levantamento do solo, com necessidade de ir a campo, coletar amostra de solo e fazer uma análise laboratorial, assim foi possível fazer uma representação cartográfica de cada unidade de terra através de um sistema de símbolos (números e letras) e cores e também a visualização da distribuição espacial.

Com esse método, a avaliação da terra exige estudos longo, como a participação de

equipes multidisciplinares e integrada. Sua precisão depende da escala do trabalho e do detalhe adequado da caracterização de cada um dos componentes do meio ambiente. Trata-se de um trabalho difícil de ser executado em áreas com pouco acesso e/ou com escassez de recursos financeiros e uma equipe especializada para a execução de estudos ambientais precisos (ASSAD, 1993).

Já na metodologia proposta é possível classificar a aptidão agrícola, com apenas oito variáveis, através de técnicas de mineração de dados, ferramentas do Sistema de Informação Geográfica (SIGs) e dados obtidos pelo IBGE. Isso torna a metodologia mais simples, com bom resultado e fácil interpretação. Não há necessidade de uma equipe grande, quando todos os dados estiverem disponíveis, apenas a participação de um especialista em ciência da computação e outro em meio ambiente, que tenha experiência em geotecnologias e técnicas de mineração de dados.

Diferentemente da metodologia de Ramalho Filho e Beek (1995), a metodologia proposta neste trabalho não apresenta tantos detalhes dentro do município e diversos grupos de classificação. Apenas determina se o município como um todo tem aptidão agrícola ou não, sem a análise do manejo da terra, auxiliando gestores públicos, pequenos e grandes produtores.

Outra diferença entre as metodologias são atributos econômicos e social: renda per capita, área plantada, rendimento médio e quantidade produzida são considerados importantes na metodologia proposta. A região do MATOPIBA vem ganhando destaque pela grande expansão agrícola nos últimos anos e consequentemente ocorre uma mudança na paisagem. Bertrand (1970), citado por Assad (1993), define paisagem “como a porção do espaço caracterizada por um tipo de combinação dinâmica, portanto instável, de elementos geográficos diferenciados - físicos, biológicos e antrópicos – que reagindo uns sobre os outro, fazem a paisagem ”. Portanto a paisagem natural que são os fatores ambientais (solo, clima, organismos vivos) e a paisagem antropizada refletem as interações dos aspectos socioeconômicos que caracterizam a comunidade local (ASSAD, 1993).

A preservação do meio ambiente deve incluir tomada de decisão que envolva interações entre recursos naturais e a sociedade. Problemas que podem causar pela falta de interação são: processos de desertificação, erosão (principalmente em solos agrícolas), falta monitoramento de espécies em extinção, locais com ocorrência de chuva ácida, poluição de corpos d’água e mudanças climáticas globais, entre outros. Para mapear e monitorar tais mudanças, bem como planejar respostas apropriadas são necessários programas que possam capturar e armazenar informações sobre os recursos naturais existentes, utilizando a tecnologia espacial disponível (imagens e dados de observação da Terra) e SIG (MIRANDA, 2005).

6. CONCLUSÃO

Conclui-se que foi possível gerar modelos preditivos para classificação da aptidão agrícola em municípios, na região do MATOPIBA, por meio de algoritmos de aprendizado de máquina, considerando dados físicos (temperatura máxima, media e mínima, precipitação acumulada e umidade relativa mínima e solo) já utilizados na metodologia consolidada.

O trabalho apresentou uma proposta de análise de outros atributos para classificação dos municípios, dados econômicos e sociais, e por meio de métodos de seleção de atributos foi possível verificar a importância de variáveis como: área plantada, rendimento médio, quantidade produzida das culturas de soja e milho e renda per capita.

A identificação desses atributos ajuda a reduzir custo, tempo e esforço em projetos de gestão territorial, já que a metodologia foi otimizada e requer apenas o conhecimento dos atributos que apresentaram melhores contribuições para o ajuste dos modelos preditivos. Além disso, uma vez que não há necessidade do trabalho em campo, apenas dados fornecidos pelo Instituto Brasileiro de Geografia e Estatística (IBGE) e dados de sensores remotos, a utilização de técnicas de mineração de dados foi o suficiente para se obter a avaliação do perfil de aptidão agrícola dos municípios.

A metodologia proposta não substitui o método consolidado pois é preciso especificar o atributo meta e dos atributos físicos. Esse método utilizando técnicas de mineração de dados é importante para realizar uma avaliação rápida da adequabilidade de áreas que estão sendo exploradas, porém não pode ser aplica para avaliar locais aptos para a expansão agrícola, uma vez que precisa de informações previa da agricultura local.

O modelo preditivo que apresentou a melhor acurácia (84.4%) foi o algoritmo J48 com o método de seleção de atributos Wrapper, com a estatística Kappa de 0.67 e um total de 31 regras, sendo considerado um modelo adequado para predizer municípios com aptidão agrícola.

Com isso, foi possível atingir o objetivo neste trabalho, a elaboração de uma metodologia simplificada para classificar a aptidão agrícola de municípios da região do MATOPIBA, e essa metodologia pode ser facilmente adaptada e aplicada em outras regiões do Brasil. Outro objetivo que foi atingido foi a geração de um conjunto de modelos preditivos, baseados em técnicas de mineração de dados, podendo assim auxiliar gestores públicos na identificação de municípios com elevada aptidão agrícola.

O mapeamento de municípios com aptidão agrícola é considerado um aliado para o município e auxilia na elaboração de uma política ambiental, rural e urbanística mais eficiente.

Com isso, torna-se um instrumento para identificar conflitos de uso e ocupação do solo e auxilia no planejamento de restauração e/ou recuperação de usos inadequados para cumprir a legislação.

7. REFERÊNCIAS BIBLIOGRÁFICAS

ADAMI, M. **Estimativa de áreas agrícolas por meio de técnicas de sensoriamento remoto, geoprocessamento e amostragem**. 183 f. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004.

ADAMI, M.; MOREIRA, M. A.; RUDORFF, B. F. T.; FREITAS, C. da C.; FARIA, R. T. de. Expansão direta na estimativa de culturas agrícolas por meio de segmentos regulares. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 12., 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. p. 9-15

AGRAWAL, R, SRIKANT, R: Fast Algorithms for Mining Association Rules in Large Databases. In: 20th International Conference on Very Large Data Bases, 478-499, 1994.

Atlas Brasil. Atlas do Desenvolvimento Humano no Brasil. Disponível em: <http://atlasbrasil.org.br/2013/pt/> Acesso em: 08 julho 2018

ASSAD, M.L.L. Sistema de informações geográficas na avaliação da aptidão agrícola de terras. In: ASSAD, E.D.; SANO, E.E., [ed.]. **Sistema de informações geográficas: aplicações na agricultura**. [Planaltina]: EMBRAPA-CPAC, 1993. cap.8. p.173-199.

BELCHIOR, E. B.; ALCANTARA, P. H. R.; BARBOSA, C. F. **Perspectivas e desafios para a região do Matopiba**. Palmas: Embrapa Pesca e Aquicultura, 2017. 3 p.

BENNEMA, J.; BEEK, K.J.; CAMARGO, M.N. **Um sistema de classificação de capacidade de uso da terra para levantamento de reconhecimento de solos.** Rio de Janeiro: Ministério da Agricultura. FAO, 1964. 49p. Mimeografado.

BERNARDES, T.; ADAMI, M.; FORMAGGIO, A. R.; MOREIRA, M. A.; FRANÇA, D. de A.; NOVAIS, M. R. de. **Imagens mono e multitemporais Modis para estimativa da área com soja no Estado de Mato Grosso**. Pesquisa Agropecuária Brasileira [online]. v.46, n. 11, p. 1530-1537, 2011.

BOLFE, E. L.; VICTORIA D. DE C.; ARAUJO, L. S.; BATISTELLA, M., COUTINHO, A.; MINGOTI, R.; MANZATTO, C.; LUMBRERAS J. F. **Relatório técnico**: organização de um banco de dados geoespaciais e elaboração de planos de informações sobre aspectos físico-bióticos, de logística e infraestrutura, conservação, unidades de paisagem e aptidão, expansão e intensificação agrícola da região do Matopiba. Projeto especial: plano estratégico de atuação da Embrapa na região do Matopiba. Campinas: Embrapa Monitoramento por Satélite, 2015.

BORGHI, E.; BORTOLON, L.; AVANZI, J. C.; BORTOLON, E. S. O.; UMMUS, M. E.; GONTIJO NETO, M. M.; COSTA, R. V. da Desafios das novas fronteiras agrícolas de produção de milho e sorgo no Brasil: desafios da região do MATOPIBA. In: KARAM, D.; MAGALHÃES, P. C. (Ed.). Eficiência nas cadeias produtivas e o abastecimento global. Sete Lagoas: Associação Brasileira de Milho e Sorgo, 2014. cap. 25, p. 263-278. P. 263-278.

BREIMAN, L. Random forests. Machine learning, Springer, v. 45, n. 1, p. 5–32, 2001.

CÂMARA, G. **Modelos, Linguagens e Arquiteturas para banco de dados geográficos**. São José dos Campos: Inpe, 1995. Tese de Doutorado. < http://marte3.sid.inpe.br/col/sid.inpe.br/iris@1905/2005/07.29.07.25/doc/index.htm>Acesso 15 jan 2018

CARVALHO JUNIOR, O. A. de.; COELHO, M. A. N.; MARTINS, E. de S., GOMES, R. A. T.; COUTO JUNIOR, A. F.; OLIVEIRA, S. N. de, SANTANA, O. A.; Mapeamento da vegetação na floresta atlântica usando o classificador de árvore de decisão para integrar dados de sensoriamento remoto e modelo digital de terreno. Rev. Bras. Geof., São Paulo , v. 26, n. 3, p. 331-345, Sept. 2008

CAMILO, C. O.; SILVA, J. C. da. **Mineração de dados: Conceitos, tarefas, métodos e ferramentas.** Universidade Federal de Goiás (UFC), p. 1-29, 2009. Disponível em <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf> Acesso 30 jan 2018

COHEN, J. A coefficient of agreement for nominal scales. **Educational and Psychological Measurement**. v. 20, 1960. p.37-46.

CORTES, S. C.; PORCARO, R. M.; LIFSCHITZ S. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens.** PUC-Rio Inf. MCC10/02, Rio de Janeiro, 2002. Disponível em <ftp://obaluae.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf> Acesso em 31 jan 2018

CURI, N.; LARACH, J. O. I.; KÄMPF, N.; MONIZ, A. C.; FONTES, L. E. F. **Vocabulário de ciência do solo**. Campinas: SBCS, 1993. 90p.

DAINESE, R. C**.; Sensoriamento remoto e geoprocessamento aplicado ao estudo temporal do uso da terra e na comparação entre classificação não-supervisionada e análise visual**. 2001. 185 F. Dissertação (Mestrado em Agronomia - Área de Concentração em Energia na Agricultura) – Universidade Estadual Paulista "Júlio Mesquita Filho", Faculdade de Ciências Agronômicas, Campus de Botucatu, Botucatu, 2001

DEFINIENS. Ecognition: user guide 3. Germany. Disponível em http://www.definiensimaging.com/down/ecognition. Acesso em janeiro de 2008. Apud ARTINS, V. A.; FONSECA, L. M. G. Classificação de uso de solo baseada na análise orientada a objeto e mineração de dados utilizando imagens SPOT/HRG-5 In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal. **Anais...** São José dos Campos: INPE, 2009. p. 3542-3548.

DIAS, L. M. da S.; COELHO, R. M.; Valladares, G. S.; ASSIS, A. C. C. de; FERREIRA, E. P.; Silva, R. C. da. **Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco**. Pesq. agropec. bras., Brasília , v. 51, n. 9, p. 1396-1404, Setembro 2016 .

EMBRAPA. **Empresa Brasileira de Pesquisa Agropecuária.** Disponível em <www.embrapa.br/gite/projetos/matopiba/150317_MATOPIBA_WEBSITE.pdf> Acesso em : 17 out. 2017

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Serviço Nacional de Levantamento e Conservação de Solos (Rio de Janeiro, RJ). Súmula da 10. Reunião Técnica de Levantamento de Solos. Rio de Janeiro, 1979. 83p. (EMBRAPA-SNLCS. Micelânea, 1)

ESRI. **Environmental Systems Research Institute**. Disponível em < www.esri.com> Acesso em 15 março 2019

ESTADOS UNIDOS. Department of Agriculture. Soil Conservation Service. Soil Survey Staff. **Soil survey manual**. Washington, 1951. 503p. (USDA Agriculture Handbook, 18).

FAO (Roma, Itália). A framework for land evaluation. Rome, 1976. 72p. (FAO Soil Bulletin, 32).

FAYYAD, U., G. PIATETSKY-SHAPIRO, & P. SMYTH. **The KDD process for extracting useful knowledge from volumes of data**. Communtications of the ACM 39(11), 27-34, 1996

FELGUEIRAS, C.A.; CAMARA, G. Sistema de informações geográficas do INPE In: ASSAD, E.D.; SANO, E.E. **Sistema de informaçõe geográficas: aplicações na agricultura.** Brasilia: EMBRAPA-CPAC, 1993. p.41-59

FIGUEIREDO, D. **Conceitos Básicos de Sensoriamento Remoto**. Brasília, DF: Companhia Nacional de Abastecimento - CONAB, 2005. Disponível em <http://www.conab.gov.br/conabweb/download/SIGABRASIL/manuais/conceitos_sm.pdf> Acesso 21 jan 2018

FREIRE, O**. Apontamentos de edafologia**. Piracicaba-SP: s. ed. 1984. 317p.

FREITAS, R. E.; MACIENTE, A. N. **Mesorregiões brasileiras com expansão de área agrícola**. Radar: tecnologia, produção e comércio exterior, n. 41, p. 7-18, 2015.

FORMAGGIO, A. R.; SANCHES, I. D.; **Sensoriamento Remoto em agricultura**. São José dos Campos: Editora Oficina de Textos, 2017. 283 p.

FRANK, E ; HALL, M. A.; WITTEN, I. (2009). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

FURTADO, A. L. dos S.; RODRIGUES, C. A. G.; TOSTO, S. G. Satélites e produtos. In: TÔSTO, S. G.; RODRIGUES, C. A. G.; BOLFE, E. L.; BATISTELLA, M. (Ed.). **Geotecnologias e Geoinformação.** Brasília, DF: Embrapa, 2014. p. 16-35.

GIROLAMO NETO, C. Di.; FONSECA, L. M. G.; KORTING, T. S.; SANCHES, I. D.; EBERART, I. D. R.; BENDINI, H. do N.; MARUJO, R. de F. B.; TRANBAQUINI, K. Classificação automática de áreas cafeeiras utilizando imagens de sensoriamento remoto e técnicas de mineração de dados. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17., 2015, João Pessoa. **Anais...** São José dos Campos: INPE, 2015.

GUERRA, B. J.; SCHULTZ, B.; SANCHES, I. D. Mapeamento automático da expansão da agricultura anual no MATOPIBA entre 2002 e 2015 utilizando a plataforma Google Earth Engine. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 18., 2017, Santos. **Anais...** São José dos Campos: INPE, 2017.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. In Journal of Machine Learning Research, v.3, pp.1157-1182, 2003.

HALL, M. **Correlation based feature selection for machine learning**. Doctoral dissertation, University of Waikato, Dept. of Computer Science. 1999

HAN, J.; KAMBER, M.; PEI, J. **Data mining**: concepts and techniques. 3ed. San Francisco: Morgan Kaufmann Publishers, 2011.

IBGE. **Produção Agrícola Municipal**. Disponível em: <https://sidra.ibge.gov.br/pesquisa/pam/tabelas >. Acesso em: 25 jan 2018.

JENSEN, J. R**. Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestre.** [s.l.] : São José dos Campos, SP: Parêntese, c2011., 2011.

JOHANN, J. A.; ROCHA, J. V.; OLIVEIRA, S. R. de M.; RODRIGUES, L. H. A.; LAMPARELLI, R. A. C.. Data mining techniques for identification of spectrally homogeneous areas using NDVI temporal profiles of soybean crop. Eng. Agríc., Jaboticabal , v. 33, n. 3, p. 511-524, June 2013 .

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. Artificial Intelligence, Elsevier, v. 97, n. 1-2, p. 273–324, 1997

LANDIS, J.R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**. Michigan, v. 33, n. 1, p.159-174, 1977.

LARACH, J. O. I. Estudos integrados e ordenamento territorial. B. Inf. da SBCS, v.15, n.3, p. 100-101, set./dez. 1990.

LARACH, J. O. I. Estudos integrados e ordenamento territorial. B. Inf. da SBCS, v.15, n.3, p. 100-101, set./dez. 1990. Apud PEREIRA, L. C. Aptidão agrícola das terras e sensibilidade ambiental: proposta metodológica. 122p. Tese (Doutorado em Planejamento e Desenvolvimento Rural Sustentável) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, 2002

LEITE, T. A.; NETO, J. DE O. M.; NASCIMENTO, A. F. DE J.; CHAGAS, R. M.; JÚNIOR, A. V. M. Delimitação das Áreas de Preservação Permanente (APP) com uso de geoprocessamento como subsídio à gestão dos recursos hídricos na bacia do Rio Jacaré. In: III Encontro de Recursos Hídricos, 2012, Sergipe. **Anais...** Aracaju, 2012. Artigos, p.34-38. 2012. CD-ROM

LEPSCH, I. F.; BELLINAZZI JR., R.; BERTOLINI, D.; ESPÍNDOLA, Carlos Roberto.**Manual para levantamento utilitário do meio físico e classificação de terras no sistema de capacidade de uso.** 4a aproximação. Campinas: SBCS, 1991. 175p.

LOPES, K. M. O. **Modelos Baseados em Data Mining Para Classificação Multitemporal de Culturas No Mato Grosso Utilizando Dados de NDVI/MODIS**. 2013. 108 f. Dissertação (Mestrado em Matemática Aplicada)- Instituto de Matemática, Estatística e Computação Científica, Universidade Estatual de Campinas, Campinas, 2013.

LORENSINI, C. L.; VICTORIA, D. de C.; VICENTE, L. E.; MAÇORANO, R. P. **Mapeamento e identificação da época de desmatamento das áreas de expansão da agricultura no MATOPIBA**. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17., 2015, João Pessoa. Anais... São José dos Campos: INPE, 2015. p. 3542-3548.

LUMBRERAS, J. F.; CARVALHO FILHO, A. de; MOTTA, P. E. F. da; BARROS, A. H. C.; AGLIO, M. L. D.; DART, R. de O.; SILVEIRA, H. L. F. da; QUARTAROLI, C. F.; ALMEIDA, R. E. M. de; FREITAS, P. L. de **Aptidão agrícola das terras do Matopiba.** Rio de Janeiro: Embrapa Solos, 2015. 48 p. il. color. (Embrapa Solos. Documentos, 179).

MANABE, V. D.; ROCHA, J. V.; LAMPARELLI, R. A. C.; Diferenciação de áreas cana de açúcar e pastagem através de técnicas de mineração de Dados In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17., 2015, João Pessoa. **Anais...** São José dos Campos: INPE, 2015. p. 3542-3548.

MARIA, P. F. De. **Diferenciais socioedemográficos e espaciais da pobreza no Estado de São Paulo (1991-2015).** 2018. 1 recurso online (222 p.). Tese (doutorado) - Universidade Estadual de Campinas, Instituto de Filosofia e Ciências Humanas, Campinas, SP.

MARTINS, V. A.; FONSECA, L. M. G. Classificação de uso de solo baseada na análise orientada a objeto e mineração de dados utilizando imagens SPOT/HRG-5 In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, Natal. **Anais...** São José dos Campos: INPE, 2009. p. 3542-3548.

MEGETO, G. A. S**. Avaliação Da Influência Da Temperatura E Da Precipitação Na Ocorrência Da Ferrugem Asiática Da Soja Por Meio Da Técnica De Árvore De Decisão**. 2012. 78 p. Dissertação (Mestrado em Engenharia Agrícola) - Faculdade de Engenharia Agrícola, Universidade Estatual de Campinas, CAMPINAS, 2012.

MINISTÉRIO DA AGRICULTURA, PECUÁRIA E ABASTECIMENTO (MAPA). Parâmetros de Risco Climático. ANEXO I. [s.l: s.n.]. Disponível em: <http://www.agricultura.gov.br/assuntos/riscos-seguro/risco-agropecuario/documentos/Parametrosderiscoclimatico.pdf>. Acesso em: 03 fev 2018.

MIRANDA, J. I. Fundamentos de sistemas de informações geográficas. [S.l.]: Brasília, DF: EMBRAPA, 2005.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 89-114

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Org.). **Sistemas inteligentes**: fundamentos e aplicações. Barueri: Editora Manole, 2002. p. 89-114 Apud MEGETO, G. A. S. Avaliação Da Influência Da Temperatura E Da Precipitação Na Ocorrência Da Ferrugem Asiática Da Soja Por Meio Da Técnica De

Árvore De Decisão. 2012. 78 p. Dissertação (Mestrado em Engenharia Agrícola) - Faculdade de Engenharia Agrícola, Universidade Estatual de Campinas, CAMPINAS, 2012

MONTEIRO, J. E. B. de A.; OTAVIAN, A. F.; NAKAI, A. M.; OLIVEIRA, A. F. de; CARVALHO, J. R. P. de; ROMANI, L. A. S.; MACIEL, R. J. S. CONPREES. Versão 1.0. Campinas: Embrapa Informática Agropecuária, 2016.

MOORE, David S. A estatística básica e sua pratica. 3. ed. Rio de Janeiro, RJ: Livros Técnicos e Científicos, 2005. 637p.

MOREIRA, M. A. **Fundamentos do sensoriamento remoto e metodologias de aplicação**. [s.l.] : Viçosa, MG: Editora UFV, 2011., 2011.

NONATO, R. T.; OLIVEIRA, S. R. de M. Técnicas de mineração de dados para identificação de áreas com cana-de-açúcar em imagens Landsat 5. **Engenharia Agrícola**, Jaboticabal, v.33, n.6, p.1268-1280, nov./dez. 2013.

NOVO, E. M. L. de M. **Sensoriamento remoto: princípios e aplicações.** São Paulo: Edgard Blücher, 2010 308p

PEREIRA, L. C. **Aptidão agrícola das terras e sensibilidade ambiental: proposta metodológica**. 122p. Tese (Doutorado em Planejamento e Desenvolvimento Rural Sustentável) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, 2002.

PEREIRA,L.C.; SILVEIRA, M.A.da; LOMBARDINETO,F. Agroecologia e aptidão agrícola das terras: as bases científicas para a agricultura sustentável. In: CONGRESSO BRASILEIRO DE AGROECOLOGIA, 1.; SEMINÁRIOINTERNACIONAL SOBRE AGROECOLOGIA, 4.; SEMINÁRIO ESTADUALSOBRE AGROECOLOGIA, 5., 2003, Porto Alegre. **Anais...** Porto Alegre: EMATER/RS:ASCAR, 2003. 4 p

PEREIRA, L. C.; LOMBARDI NETO, F. **Avaliação da aptidão agrícola das terras: proposta metodológica**: Documentos. Jaguariuna: Embrapa Meio Ambiente, 2004. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/CNPMA/5805/1/documentos_43.pdf>. Acesso em 11 de jan 2018

PLANO Agrícola e Pecuário, 2014/2015. Brasília, DF,2015. Disponível em: <http://www.agricultura.gov.br/arq_editor/PAP%202014-2015.pdf>. Acesso em 24 jan 2018

PLATT, J. **Sequentional minimal optimization: A fast algorithm for training support vector machines**. Technical Report Technical Report 98-14, Microsoft Research, 1998.

PRICE, M. H. **Mastering ArcGIS**. [S.l.]: New York, NY : McGraw-Hill Education, c2016., 2016.

QGIS. QGIS Documentation**. Disponível em <https://docs.qgis.org/2.14/pt_BR/docs/gentle_gis_introduction/spatial_analysis_interpolation.html> Acesso em 22 out 2018**

QUARTAROLI, C. F.; MIRANDA, E. E. de; VALLADARES, G. S.; HOTT, M. C.; CRISCUOLO, C.; GUIMARÃES, M. **Avaliação da adequação do uso das terras agrícolas**

no Nordeste do Estado de São Paulo em 1988 e 2003. Campinas: Embrapa Monitoramento por Satélite, 2006. 36 p. (Embrapa Monitoramento por Satélite. Documentos, 57).
QUARTAROLI, C. F.; VICENTE, L. E.; ARAUJO, L. S. de Sensoriamento remoto. In: TÔSTO, S. G.; RODRIGUES, C. A. G.; BOLFE, E. L.; BATISTELLA, M. (Ed.). **Geotecnologias e geoinformação**. Brasília, DF: Embrapa, 2014. p. 61-79. (Coleção 500 Perguntas, 500 Respostas)

QUINLAN, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

RAMALHO FILHO, A.; BEEK, K. J. **Sistema de avaliação da aptidão agrícola das terras**. 3.ed. Rio de Janeiro: EMBRAPA-CNPS, 1995. 65p.

REZENDE, S. O.; PUGLIESI, J. B., MELANDA, E. A., PAULA, M. F. Mineração de Dados. In: **Sistemas Inteligentes. Fundamentos e aplicações**. Organização: Solange Oliveira Rezende. Barueri, SP: Manole, 2005.

RIGAUX, P.; SCHOLL, M.; VOISARD, A. Spatial databases with application to GIS. San Francisco: Academic Press, 2000. 408 p. Apud ADAMI. Estimativa de áreas agrícolas por meio de técnicas de sensoriamento remoto, geoprocessamento e amostragem. 183 f. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2004

RODRIGUES, A. C.M. **Mapeamento Multitemporal do uso e cobertura do solo do município de São Sebastião-SP, utilizando técnicas de segmentação e classificação de imagens TM Landsat e HRV-SPOT.** São José dos Campo:INPE, 94p. 2000. (INPE-7510-PUD/39).

RODRIGUES, A. C.M. **Mapeamento Multitemporal do uso e cobertura do solo do município de São Sebastião-SP, utilizando técnicas de segmentação e classificação de imagens TM Landsat e HRV-SPOT.** São José dos Campo:INPE, 94p. 2000. (INPE-7510-PUD/39). Apud DAINESE, R. C**.; Sensoriamento remoto e geoprocessamento aplicado ao estudo temporal do uso da terra e na comparação entre classificação não-supervisionada e análise visual**. 2001. 185 F. Dissertação (Mestrado em Agronomia - Área de Concentração em Energia na Agricultura) – Universidade Estadual Paulista "Júlio Mesquita Filho", Faculdade de Ciências Agronômicas, Campus de Botucatu, Botucatu, 2001

RUDORFF, B. F. T.; MOREIRA, M. A.; **Sensoriamento remoto aplicado à agricultura**. São José dos Campo: INPE, 19p. 2002.

SANO, E.E.; DOS SANTOS, C.C.M.; DA SILVA, E.M.; CHAVES, J.M. **Fronteira Agrícola do Oeste Baiano: Considerações sobre os Aspectos Temporais e Ambientais**. Revista Geociências, v. 30, n. 3 p. 479-489, 2011. Disponível em <www.revistageociencias.com.br/geociencias-arquivos/30_3/Art%2013_Sano.pdf> Acesso 24 nov 2018

SANTOS, W. G. DOS; MARTINS, J. I. F. **O Zoneamento Agrícola de Risco Climático e sua contribuição à agricultura brasileira**. Revista de Política Agrícola, v. 25, n. 3, p. 73–94, 29 dez. 2016.

TAN, P-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining**. Rio de Janeiro: Editora Ciência Moderna Ltda, 2009. 896 p.

THOME, R.; **Interoperabilidade em geoprocessamento: conversão entre modelos conceituais de sistemas de informação geográfica e comparação com o padrão OPEN GIS**. São José dos Campos. 196p. (INPE-7266-TDI/708). Dissertação (Mestrado em Computação Aplicada) - INPE, 1998.

USGS. **United States Geological Survey**. Disponível em <https://www.usgs.gov/> Acesso 30 maio 2018

VALLADARES, G. S.; QUARTAROLI, C. F.; HOTT, M. C.; MIRANDA, E. E. de; NUNES, R. da S.; KLEPKER, D.; LIMA, G. P. **Mapeamento da aptidão agrícola das terras do Estado do Maranhão**. Campinas: Embrapa Monitoramento por Satélite, 2007. 25 p.

VAPNIK, V.N. **The nature of statistical learning theory**. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

Weka. **Waikato Environment for Knowledge Analysis.** Disponivel em < www.cs.waikato.ac.nz/ml/weka/ > Acesso 15 março 2019

WITTEN, I. H.; FRANK, E; HALL, M. A.. **Data mining : practical machine learning tools and techniques.** 3rd ed. Amsterdam; Boston, MA: Morgan Kaufman, 2011.