UNIVERSIDADE DE ARARAQUARA
MESTRADO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

Germano de Melo Matos Trevisan

O USO DA MINERAÇÃO DE DADOS NA DESCOBERTA DE CONHECIMENTO EM EMPRESA DO SETOR AGRÍCOLA

Dissertação apresentada ao Programa de Mestrado Profissional em Engenharia de Produção da Universidade de Araraquara – UNIARA – como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção, Área de Concentração: Gestão Estratégica e Operacional da Produção.

Prof. Dr. Fábio Ferraz Júnior
Orientador

Araraquara, SP – Brasil
2017

FICHA CATALOGRÁFICA

T739u Trevisan, Germano de Melo Matos
O uso da mineração de dados na descoberta de conhecimento em empresa do setor agrícola/Germano de Melo Matos Trevisan. – Araraquara: Universidade de Araraquara, 2017.
157f.

Dissertação (Mestrado)- Mestrado Profissional em Engenharia de Produção – Universidade de Araraquara-UNIARA

Orientador: Prof. Dr. Fábio Ferraz Júnior

1. Cana de açúcar. 2. Produtividade. 3. Mineração de dados. I.Título.

CDU 62-1

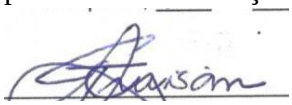
REFERÊNCIA BIBLIOGRÁFICA

TREVISAN, G. M. M. **O uso da mineração de dados na descoberta de conhecimento em empresa do setor agrícola**. 2017. 157f. Dissertação de Mestrado em Engenharia de Produção – Universidade de Araraquara, Araraquara-SP.

ATESTADO DE AUTORIA E CESSÃO DE DIREITOS

NOME DO AUTOR: Germano de Melo Matos Trevisan
TÍTULO DO TRABALHO: O uso da mineração de dados na descoberta de conhecimento em empresa do setor agrícola
TIPO DO TRABALHO/ANO: Dissertação / 2017

Conforme LEI Nº 9.610, DE 19 DE FEVEREIRO DE 1998, o autor declara ser integralmente responsável pelo conteúdo desta dissertação e concede a Universidade de Araraquara permissão para reproduzi-la, bem como emprestá-la ou ainda vender cópias somente para propósitos acadêmicos e científicos. O autor reserva outros direitos de publicação e nenhuma parte desta dissertação pode ser reproduzida sem a sua autorização.



Germano de Melo Matos Trevisan
Universidade de Araraquara – UNIARA
Rua Carlos Gomes, 1217, Centro. CEP: 14801-340, Araraquara-SP
Email (do autor): gm.mt@hotmail.com

UNIVERSIDADE DE ARARAQUARA - UNIARA
MESTRADO PROFISSIONAL EM ENGENHARIA DE PRODUÇÃO

FOLHA DE APROVAÇÃO

Dissertação apresentada ao Programa de Mestrado Profissional em Engenharia de Produção da Universidade de Araraquara – UNIARA – para obtenção do título de Mestre em Engenharia de Produção.

Área de Concentração: Gestão Estratégica e Operacional da Produção.

NOME DO AUTOR: **GERMANO DE MELO MATOS TREVISAN**

TÍTULO DO TRABALHO:

" O USO DA MINERAÇÃO DE DADOS NA DESCOBERTA DE CONHECIMENTO EM EMPRESA DO SETOR AGRÍCOLA."

Assinatura do(a) Examinador(a)	**Conceito**
Fábio Ferraz Júnior **Prof(a). Dr(a). Fábio Ferraz Júnior (Orientador(a))** **Universidade de Araraquara - UNIARA**	(X)Aprovado ( ) Reprovado
Prof(a). Dr(a). Jorge Alberto Achcar **Universidade de Araraquara - UNIARA**	(X)Aprovado ( ) Reprovado
Prof(a). Dr(a). Edson Walmir Cazarini **Universidade de São Paulo – USP**	(X)Aprovado ( ) Reprovado

Versão definitiva revisada pelo(a) Orientador(a) em: 21/10/17

Fábio Ferraz Júnior
Prof(a). Dr(a). Fábio Ferraz Júnior (orientador(a))

Dedico este trabalho a todos que acreditaram em mim, no meu esforço e na minha dedicação para ser um profissional e um ser humano melhor.

AGRADECIMENTOS

Agradeço a Deus, por me permitir cursar o Mestrado, mesmo diante da correria do cotidiano, das dificuldades e do tempo escasso. Agradeço por chegar ao fim de mais essa jornada, com bom humor e sabedoria para aproveitar os ensinamentos adquiridos.

Agradeço à minha família, por entender o tempo dedicado aos estudos e por me apoiar incondicionalmente.

Agradeço aos amigos, pela motivação e pelo companheirismo.

A todos o meu muito obrigado.

RESUMO

A cana-de-açúcar é um dos principais produtos da economia brasileira. Dessa forma, são indispensáveis estratégias para reduzir os custos e otimizar os seus processos produtivos para se obter melhor produtividade. Vários são os fatores que influenciam a produtividade e a qualidade da cana-de-açúcar, sendo o objetivo principal da agroindústria sucroalcooleira a recuperação máxima da sacarose da cana-de-açúcar ao menor custo possível. A eficiente gestão das condições ambientais, como clima, solo, manejo empregado durante o plantio, variedade, tipo de muda, época de corte e estágio de desenvolvimento da cultura, resulta em maior produtividade. Com isso, justifica-se o estudo dos fatores que podem influenciar a produtividade dessa cultura por meio da aplicação de técnicas de apoio ao planejamento e/ou tomada de decisão, como, por exemplo, a técnica de mineração de dados. Com ela, há a possibilidade de extrair informações ocultas e gerar dados que favoreçam a descoberta de conhecimento, explorando padrões existentes no banco de dados. Diante do exposto, apresenta-se como objetivo da pesquisa identificar e classificar os fatores que impactam na produtividade agrícola da cana-de-açúcar, utilizando a técnica de mineração de dados para auxiliar os gestores na descoberta de informação e conhecimento. O método utilizado foi a pesquisa aplicada, ou seja, a partir da introdução das técnicas de mineração de dados em empresa do setor agrícola. Os resultados da pesquisa foram obtidos a partir da aplicação dessas técnicas, compostas por análises da árvore de decisão (especificamente árvore de classificação) e floresta aleatória, nas quais as variáveis contínuas, Tonelada de Cana por Hectare (TCH), Açúcar Total Recuperável (ATR) e TCH*ATR, foram transformadas em classes de acordo com os quartis da distribuição normal. As principais conclusões demonstraram que as técnicas de mineração de dados podem ser empregadas com sucesso no agronegócio da cana-de-açúcar, utilizando dados climáticos (meteorológicos) e de manejo para realizar uma análise exploratória das variáveis mais importantes na definição de classes de TCH, ATR e TCH*ATR.

Palavras-chave: Cana-de-açúcar. Produtividade. Mineração de dados. KDD.

ABSTRACT

*Sugarcane is one of the main products of the Brazilian economy. In this way, strategies are indispensable to reduce costs and optimize your production processes for better productivity. Several factors influence the productivity and quality of sugarcane, with sugarcane sucrose being the main objective of sugarcane agroindustry, at the lowest possible cost. The efficient management of environmental conditions, such as climate, soil, as well as the management used at the time of planting, variety, type of seedlings, cutting season and stage of development of the crop, results in greater productivity. Therefore, the study of the factors that can influence the productivity of this crop is justified by the application of techniques to support planning and / or decision making, such as the data mining technique. Using the Data Mining application tasks makes it possible to extract hidden information and generate data that favors the discovery of knowledge by exploiting existing patterns in the database. In view of the above, the objective of the research is to identify and classify the factors that impact sugarcane agricultural productivity, using the data mining technique to assist managers in the discovery of information and knowledge. The applied method is the applied research, with application of the techniques of data mining in company of the agricultural sector. The results of the research were obtained through the application of data mining techniques composed of analyzes of the decision tree (specifically classification tree) and random forest, where the continuous variables TCH, ATR and TCH were transformed into classes according to the Quartiles of the normal distribution. The main conclusions demonstrated that data mining techniques can be used successfully in the agribusiness of sugarcane using climatic (meteorological) and management data to perform an exploratory analysis of the most important variables in the definition of classes of HCT, ATR And TCH * ATR.*

Key-words*: Sugarcane. Productivity. Data Mining. KDD.*

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Abreviaturas e Siglas

AGE	Assessoria de Gestão Estratégica
APLA	Arranjo Produtivo Local do Álcool
ATR	Açúcar Total Recuperável
BNDES	Banco Nacional de Desenvolvimento Econômico e Social
CART	*Classification and Regression Trees*
CEISE	Centro Nacional das Indústrias do Setor Sucroenergético e Biocombustíveis
CONAB	Companhia Nacional de Abastecimento
COP	Conferência das partes
CRISP-DM	*Cross-Industry Standard Process of Data Mining*
DBMS	*Database management system*
DW	*Data Warehouse*
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
EPR	Erro padrão residual
EQM	Erro quadrático médio
ERP	*Enterprise Resource Planning*
ESALQ	Escola Superior de Agricultura Luiz de Queiroz
HSD	*Honestly significant difference*
IA	Inteligência Artificial
IAA	Instituto do Açúcar e do Álcool
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	*Knowledge Discovery in Databases*
KNN	*K-nearest neighbor*
MAPA	Ministério da Agricultura, Pecuária e Abastecimento
MDIC	Ministério da Indústria, Comércio Exterior e Serviços
OLAP	*On Line Analytical Processing*
OOB	*Out-of-bag*
PIB	Produto Interno Bruto
PIMS	*Process Information Management Systems*
SECEX	Secretaria de Comércio Exterior
SEMMA	*Sample, Explore, Modify, Model, Assesment*
SGE	Secretaria de Gestão Estratégica
TCH	Tonelada de Cana por Hectare
TI	Tecnologia da Informação
TIC	Tecnologia da Informação e Comunicação
UFPR	Universidade Federal do Paraná
ÚNICA	União da Indústria de Cana-de-açúcar
USDA	*United States Department of Agriculture*
USP	Universidade de São Paulo
WSD	W*holly significant difference*

SUMÁRIO

1 INTRODUÇÃO

De acordo com o Ministério da Agricultura brasileiro – MAPA (BRASIL, 2016), a partir dos dados da Secretaria de Comércio Exterior – SECEX/Ministério da Indústria, Comércio Exterior e Serviços – MDIC, as empresas do setor do agronegócio exportaram no ano de 2015 produtos que totalizaram US$ 88,2 bilhões, com saldo da balança comercial positivo em US$ 75,15 bilhões. Nesse montante, o complexo sucroalcooleiro contribuiu com US$ 8,53 bilhões, sendo o quarto maior setor em valor exportado.

Dados apresentados pela Companhia Nacional de Abastecimento – CONAB (2016) demonstram que os resultados supracitados obtidos pelo setor foram possíveis através do processamento de 665,6 milhões de toneladas de cana-de-açúcar (*saccharum*), representando um crescimento de 4,9% em relação à safra anterior, quando foram cultivadas em 8.654,2 mil hectares, apresentando uma redução de 3,9%, se comparada com a safra 2014/15.

Esses dados evidenciam a importância da cana-de-açúcar para a economia brasileira. O seu planejamento e a sua gestão em todas as fases de sua produção, na medida em que otimizam o retorno econômico dessa cultura, sobretudo diante da forte crise que vive o setor, são indispensáveis para que o Brasil se mantenha como maior produtor sucroalcooleiro do mundo. Assim, reduzir os custos e otimizar os processos produtivos da cana-de-açúcar é imprescindível para obter melhor produtividade (TOMAZELA; CAMPOS; DANIEL, 2015).

Marchiori (2004) e Tomazela, Campos e Daniel (2015) destacam que, devido à importância do Brasil no cenário internacional da cana-de-açúcar e de seus derivados, conforme demonstrado pelos dados do MAPA e, considerando-se que a cana-de-açúcar tem uma época durante o ano em que ocorre maior concentração de sacarose nos colmos, tornando-se mais propícia para a colheita, é importante conhecer todas as informações relevantes durante o processo de desenvolvimento e crescimento da cana, pois o segmento sucroalcooleiro tem como principal finalidade a recuperação máxima da sacarose da cana-de-açúcar ao menor custo possível. Com isso, justifica-se o estudo dos fatores que podem influenciar na produtividade dessa cultura, por meio da aplicação de técnicas de apoio ao planejamento e/ou tomada de decisão, como por exemplo, a técnica de mineração de dados.

A descoberta do conhecimento oculto nas grandes bases de dados de empresas de diversos setores, de maneira automática ou semiautomática, é o objetivo da mineração de dados, uma técnica que permite maior agilidade no processo de tomada de decisão pelos gestores (PASTA, 2011).

Nota-se que o armazenamento dos dados gerados na atividade produtiva possibilita extrair conhecimento em base de dados, selecionando e processando dados com a finalidade de identificar novos padrões, dar maior precisão em padrões conhecidos e modelar fenômenos do mundo real, com o intuito de descobrir padrões de comportamento implícito na base de dados, bem como suas relações de causa e efeito. Desse modo, as informações contidas nessas bases de dados, processadas e analisadas de forma correta, tornam-se requisitos primordiais na tomada de decisões, e em muito podem contribuir para a otimização da produtividade da cana-de-açúcar (TOMAZELA; CAMPOS; DANIEL, 2015; PASTA, 2011).

Dentre essas bases de dados, cita-se os ambientes de gestão das operações agrícolas de cana-de-açúcar, um setor que, nos últimos anos, aumentou o volume de dados recolhidos no campo a cada safra, os quais, quando avaliados em conjunto, fornecem informações sobre como aumentar a produtividade e a eficiência no campo.

Para Janzen (2015), essa geração de dados é composta por:

- Dados agronômicos: derivados das atividades e condições das áreas cultivadas, como análise do solo, informações nutricionais, população de plantas, entre outros;
- Dados dos equipamentos: associados com o funcionamento da máquina, incluindo o consumo de combustível e indicadores de performance;
- Dados climatológicos: compostos por informações sobre precipitação, vento, temperatura e outras condições climáticas.

Considera-se uma necessidade fundamental para as empresas agrícolas a missão de gerir informações, haja vista a existência de diversos fatores que impactam na produtividade da lavoura e a existência da utilização das tecnologias da informação no campo. A informação tem um papel fundamental no desenvolvimento estratégico da empresa, e cabe ao gestor trabalhá-la de forma que ela sirva como um elemento chave para a tomada de decisões, desde que essa informação seja precisa, segura e esteja à disposição. Processar e analisar as informações geradas pelas enormes bases de dados atuais de maneira correta estão entre os requisitos essenciais para uma boa tomada de decisão (CUNHA, 2011; PASTA, 2011).

Desse modo, busca-se o uso estratégico da informação, sendo que a utilização da técnica de mineração de dados possibilita a extração de informações implícitas existentes nos Bancos de Dados, contribuindo com esse process para otimização da produtividade da lavoura de cana-de-açúcar e, consequentemente, ocasionar maior lucratividade ou competitividade para os produtores de cana-de-açúcar (PASTA, 2011; TOMAZELA; CAMPOS; DANIEL, 2015).

1.1 Problema da pesquisa

Marin et al. (2016) analisaram a trajetória de rendimentos do setor canavieiro nas duas últimas décadas, a fim de determinar em que níveis eles devem ser acelerados para que se obtenha, em 2024, uma maior produção de cana sem a expansão da área de produção. Considerando essa série histórica de ganho de rendimento, o estudo avalia que o ritmo atual de crescimento não será suficiente para atender a demanda projetada sem que haja uma expansão de área de 5% a 45% para cenários de baixa e alta demanda, respectivamente.

No entanto, conforme dados apresentados pela CONAB (2016), a partir de um levantamento realizado em 11 safras (período analisado entre as safras 2000/01 e 2013/14), demonstra-se um crescimento médio na produtividade de 1% ao ano, o que torna esse cenário de expansão um desafio.

Para que a produtividade da cana-de-açúcar aumente, diversos atributos devem ser considerados, como o ambiente de produção que, de acordo com Prado et al. (2008), é definido em função das condições físicas, hídricas, morfológicas, químicas e mineralógicas dos solos, sob manejo adequado da camada arável, associadas às condições da subsuperfície do solo e ao clima, além de variáveis meteorológicas e de manejo.

Considerando esse cenário, torna-se relevante a análise de dados históricos para extração de informações que podem ser utilizadas pelos gestores, objetivando a obtenção de conhecimento que gere vantagem competitiva sustentável (PASTA, 2011).

Desse modo, em um ambiente extremamente mutável, como naquele das empresas agrícolas, torna-se necessária a aplicação de técnicas e ferramentas automáticas que agilizem o processo de extração de informações relevantes de grandes volumes de dados, contemplando os atributos descritos por Prado et al. (2008).

Para isso, a utilização de ferramentas que auxiliem na busca, seleção e extração de informações relevantes em grandes bases de dados tem sido considerada pelas empresas, com o principal objetivo de minimizar o trabalho manual e disponibilizar informações corretas aos tomadores de decisões (PASTA, 2011).

Uma metodologia emergente, que tenta solucionar esse problema da análise de grandes quantidades de dados, é a Descoberta de Conhecimento em Banco de Dados - *Knowledge Discovery Databases* (KDD), constituída por um processo que envolve a preparação dos dados, análise estatísticas, mineração de dados, evolução e interpretação dos dados para descobrimento do conhecimento (QUIROZ-GIL; VALENCIA, 2012).

O foco do problema desta dissertação encontra-se no uso de informações ocultas na base de dados com o intuito de gerar conhecimentos aos gestores, como apoio para o planejamento estratégico e operacional, devido a sua importação nas tomadas de decisões.

A aplicação da técnica de mineração de dados se dará numa empresa agrícola, haja vista a variedade de dados armazenados e números de fatores que impactam na produtividade da lavoura e necessidade de manutenção e acréscimo da produtividade do canavial, a fim de se manter competitiva. Assim, o problema consiste em averiguar se a utilização da técnica de mineração de dados, em face da enorme disponibilidade de dados armazenados, pode contribuir como uma metodologia adequada que resulte em informações úteis aos gestores.

Nota-se de acordo com algumas pesquisas em diversos setores, que utilização das tarefas de aplicação de mineração de dados possibilita extrair informações ocultas e gerar dados que favorecem a descoberta de conhecimento, explorando padrões existentes no banco de dados. Diante do exposto, a seguinte questão norteadora da pesquisa foi: de que forma as técnicas de extração de informações podem auxiliar os gestores de empresas agrícolas produtoras de cana-de-açúcar? Mais pragmaticamente, como a gestão da informação obtida pelo uso de técnicas de extração de informação pode ajudar os profissionais da empresa a auxiliarem na tomada de decisões estratégicas para o gerenciamento das suas atividades agrícolas?

Para responder a essas questões, a Descoberta de Conhecimento em Base de Dados, por meio da técnica de mineração denominada árvore de decisão (*decision tree*) e floresta aleatória (*random forest*) foram considerados os métodos de análises mais apropriados, pois o problema de pesquisa envolve a variável produtividade do canavial (– Tonelada de Cana por Hectare - TCH e Açúcar Total Recuperável - ATR) em três classes (alta, média e baixa), que são variáveis categóricas.

A escolha do uso de mineração de dados para auxiliar na tomada de decisão, através da aplicação de tal técnica, deve-se a algumas vantagens que a mineração de dados proporciona, dentre elas, leva-se em consideração o fato de serem de fácil compreensão e das variáveis envolvidas poderem ser usadas sem necessidade de normalização.

Nesse caso, a técnica da árvore de decisão é utilizada para descobrir regras de classificação para um atributo a partir da subdivisão dos dados. Trata-se de uma técnica simples, e mais popular na abordagem de representação de classificadores, sendo utilizada por pesquisadores de várias disciplinas, tais como estatística e aprendizado de máquina. Não necessita de parâmetros de configuração, e apresenta um bom grau de assertividade, no

entanto, demanda uma análise detalhada dos dados que serão usados para garantir bons resultados (CAMILO; SILVA, 2009; MAIMON; LIOR, 2010).

Já a floresta aleatória (*random forest*), é um algoritmo classificador que faz uso do método de árvore de decisão, mas se diferencia um pouco dos algoritmos de árvores de decisão, pois consiste:

> [...] em uma técnica de agregação de classificadores do tipo árvore de decisão, construídos de forma que sua estrutura seja composta de maneira aleatória. Para determinar a classe de uma instância, o método combina o resultado de várias árvores de decisão, por meio de um mecanismo de votação (LORENZETTI; TELÖCKEN, 2016, p.1).

É um algoritmo poderoso do que comparado somente a uma árvore de decisão, boa taxa de acerto quando testado em diferentes conjuntos de dados, técnica exata, evitam ajustes, menos sensíveis à ruídos e proporcionam classificação aleatória das árvores sem intervenção humana (LORENZETTI; TELÖCKEN, 2016).

Além disso, o fato dos modelos obtidos com as técnicas de árvores de decisão e florestas aleatórias serem de fácil compreensão possibilita às pessoas sem conhecimentos estatísticos interpretar tais modelos, corroborando com a intenção de auxiliar os gestores na tomada de decisões.

1.2 Objetivos

1.2.1 Objetivo geral

O presente trabalho teve como objetivo desenvolver um modelo utilizando árvores de decisão e floresta aleatória que determina qual cenário de ambiente de produção, clima e manejo produz maiores magnitudes para os valores de desempenho de produtividade para o canavial (ATR e TCH).

1.2.2 Objetivos específicos

Constituem objetivos específicos da presente pesquisa para o ambiente em estudo:

- Aplicar e comparar as técnicas de árvores de decisão (*decision tree*) e floresta aleatória (*random forest*), aplicadas a problemas de mineração de dados;
- Verificar se existe um relacionamento significativo entre os fatores de manejo, clima e ambiente de produção e a produtividade do canavial;

- Definir quais os fatores dentro do ambiente de produção, clima e manejo que mais influenciam ATR e TCH;
- Identificar e classificar os fatores que impactam na produtividade agrícola da cana-de-açúcar, utilizando a técnica de mineração de dados para auxiliar os gestores na descoberta de informação e conhecimento.

1.3 Justificativas

Para justificar o desenvolvimento desta pesquisa apresenta-se dois aspectos: um social e outro científico.

Sobre a relevância social se destaca a importância do setor agrícola produtor de cana-de-açúcar:

- Na geração de emprego no setor agrícola – de acordo com Neves e Trombin (2014): 283.647 pessoas no cultivo da cana-de-açúcar, durante a safra 2013/2014.
- Moraes et al. (2015), em seu estudo mostrou que os trabalhadores envolvidos com a cana-de açúcar recebem salários maiores, são mais escolarizados e têm uma proporção maior de emprego formal quando comparados com a média desses indicadores para as outras culturas analisadas. Foi possível ainda verificar que os descendentes dos empregados da lavoura canavieira apresentam indicadores socioeconômicos melhores, além de terem uma mobilidade maior para outros setores fora do agrícola.
- Projeções realizadas com base em dados da Assessoria de Gestão Estratégica - AGE/MAPA e da Secretaria de Gestão Estratégica - SGE/Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) indicam um aumento de 59% na área nos próximos 10 anos, totalizando 14,339 milhões de hectares.
- Marin et al. (2016), o desafio é aumentar a produtividade da cana existente, dadas as preocupações sobre a conversão de novas áreas e a crescente demanda mundial por açúcar e etanol vindo da cana. O rendimento médio nacional da produção nessas condições é de 62% do potencial – ou seja, há oportunidades de incremento de 38%. Para os pesquisadores, caso o cenário se torne mais favorável, com o segmento atingindo uma produtividade média de 80% da sua capacidade, a demanda de cana no curto prazo seria atendida integralmente mesmo diante da possibilidade de redução de 18% na área de cana em um cenário de baixa demanda ou uma expansão de 13% para um cenário de alta demanda.

- Nova Cana (2016b), pelo compromisso feito na Conferência do Clima na Conferência das Partes - COP 21, o governo estabelecia que o etanol manteria sua participação na matriz energética brasileira em 2030. Projeções feitas pelo Nova Cana mostravam que o Brasil teria que produzir cerca de 50 bilhões de litros de etanol para que essa meta fosse cumprida. O problema é que o governo não disse como essa meta seria atingida. A situação difícil que o setor sucroenergético vem vivendo na última meia década tem estrangulado os investimentos em novas usinas e no aumento do cultivo de cana-de-açúcar no Brasil. Por esse motivo muitos desconfiam da capacidade do país de cumprir essa meta.
- De acordo com Nyko et al. (2013), além das adversidades climáticas, a produtividade agrícola do canavial foi afetada também pelo processo de mecanização do plantio e colheita, quadro varietal, com baixos investimentos em pesquisa.
- Ainda de acordo com os autores, o desenvolvimento de uma variedade leva em média 10 anos para ser comercializada e mais 5 anos para ocorrer a consolidação nas áreas cultivadas, com investimentos previstos em torno de R$ 150 milhões para o desenvolvimento completo.
- Pode-se depreender que variedades mais antigas vêm ocupando espaço cada vez maior nas lavouras de cana, ou seja, o ritmo de substituição de variedades antigas por novas variedades está se reduzindo constantemente na última década. Isso ocorre, pois para grandes empresas de genética agrícola, cujo mercado é global, a cultura da cana-de-açúcar é menos importante. A apropriabilidade dos resultados no mercado de cana é bastante inferior quando comparada com a de outras culturas, o que pode gerar menores incentivos para a inovação e desenvolvimento de novas variedades (NYKO et al., 2013).
- Os avanços gerados pelas iniciativas públicas de pesquisa básica não vêm sendo suficientes para engendrar estímulos e ciclos virtuosos de inovação setorial. Ademais, não foram estabelecidos elos eficientes de transferência de conhecimento e tecnologia da academia para o setor privado. Pode-se identificar até mesmo escassez de mão-de-obra qualificada para ser incorporada pelo setor privado (NYKO et al., 2013).
- Ao analisar o sistema de produção usado atualmente na cultura da cana-de-açúcar, nota-se que existe um grande volume de informações associadas à planta, ao solo, ao clima e ao meio físico de produção. Esses fatores influenciam a produtividade, os custos operacionais, os investimentos e os impactos ambientais. Os efeitos e as

interações entre essas variáveis são complexos e demandam o auxílio da tecnologia da informação para viabilizar o armazenamento, a análise e o diagnóstico na gestão agrícola (NOVA CANA, 2016b).

Sob esse olhar fica evidente a relevância social desta pesquisa, pois o seu resultado pode contribuir com a melhoria na produção agrícola brasileira de cana-de-açúcar, obtendo assim vantagem competitiva.

Já a relevância científica desta pesquisa se justifica pela carência de pesquisas científicas que:

a) Evidenciam o uso da tecnologia e pesquisa no setor agrícola com o objetivo de armazenar e processar informações sobre o solo, clima, características das variedades agrícolas, custos e desempenhos operacionais;

b) Demonstrem como as organizações brasileiras do setor agrícola produtor de cana-de-açúcar estão estruturando as suas ferramentas de gestão de informação aliada à análise de dados, interpretação e gestão do conhecimento para apoiarem os seus processos estratégicos;

c) Apontem o uso de novas tecnologias para o desenvolvimento de novas variedades de cana e práticas de manejo que suportem o aumento de produtividade e contribua para o desafio de crescimento do setor;

Neste item, é feita uma revisão bibliográfica, com um breve relato de outros trabalhos desenvolvidos com a utilização de diferentes técnicas de mineração de dados. Além disso, diversas pesquisas atestam a eficácia da técnica de mineração de dados, aplicada ao setor agrícola:

- Garcia e Camolesi (2015), em seu trabalho com aplicação do KDD, observaram como as técnicas de mineração de dados podem prover subsídios valiosos para a tomada de decisão no que diz respeito à gestão das lavouras de cana-de-açúcar, com intuito de melhorar a produtividade dos canaviais. O processo KDD utiliza a técnica de mineração denominada árvore de decisão, que mostra os fatores com maior influência na produtividade agrícola com análise dos dados sobre as culturas de cinco anos de duas unidades agrícolas na região de Assis/SP. Nesse estudo, que contemplou 23 variáveis, identificou-se que a variedade de cana-de-açúcar cultivada foi a que representou maior influência nas unidades pesquisadas.
- Tomazela, Campos e Daniel (2015) publicaram um trabalho intitulado “Mineração de Dados aplicada à produtividade de cana-de-açúcar”, que tem como objetivo identificar e caracterizar grupos de acordo com a produtividade da cana, utilizando

mineração de dados para propiciar maior precisão no gerenciamento desses grupos. O método de pesquisa é caracterizado como Modelagem, porque envolve tratamento estatístico dos dados coletados e aplicação da técnica de mineração de dados, denominada clusterização, e algoritimo *k-means*, pela sua complexidade linear e utilização da ferramenta *Weka*. Como resultado, foi possível identificar 3 *clusters*, caracterizados em produtividade baixa, média e alta. O solo, a fertilidade e a textura foram característicos em cada um dos clusters. Já os níveis de adubação foram semelhantes nos 3 clusters, além de ter sido identificado um grupo de talhões com fertilidade muito baixa.

- Di Girolamo Neto, Rodrigues e Meira (2014), através de modelos de predição da ferrugem do cafeeiro por técnica de mineração, constataram que modelos desenvolvidos fornecem subsídios para o monitoramento de doenças em anos de alta carga pendente de frutos do que outros modelos existentes, além de prover uma possibilidade de monitoramento em anos de baixa pendente de frutos. Além disso, os autores ressaltam que os melhores modelos foram gerados pelas técnicas de máquinas de vetores suporte e florestas aleatórias.
- Nonato e Oliveira (2013) verificaram a aderência de técnicas de mineração de dados voltadas para problemas de classificação de dados na identificação automatizada de áreas cultivadas com cana-de-açúcar, em imagens do satélite Landsat 5/TM. Os resultados reforçam o forte potencial das árvores de decisão no processo de classificação e identificação de áreas cultivadas com cana-de-açúcar, em diferentes cidades produtoras no Estado de São Paulo.
- Antunes, Oliveira e Rodrigues (2011), cujo trabalho tem como objetivo aplicar técnicas de mineração de dados para classificação das fases fenológicas da cultura da cana-de-açúcar no estado de São Paulo, utiliza dados do sensor MODIS e de precipitação que auxiliam na caracterização do ciclo de desenvolvimento da cultura. A descoberta do conhecimento pode ser feita através de regras de decisão relevantes para especialistas, revelando a aderência de técnica de mineração de dados em problemas de classificação de imagens de satélite. A técnica de classificação utilizada foi a árvore de decisão com o algoritmo J48.
- Souza et al. (2010) mapearam a produtividade da cana-de-açúcar em uma área de 23ha, para analisar os atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística, classificados em três níveis de produção para

indução de árvore de decisão no programa SAS *Enterprise Miner*, o que permitiu verificar que a altitude é a variável com maior potencial para interpretar os mapas de produtividade de cana-de-açúcar, auxiliando na agricultura de precisão e se mostrando uma ferramenta adequada para definição de zonas de manejo em área cultivada com essa cultura.

1.4 Contribuição esperada

Sobre a relevância científica, verifica-se que a evolução da tecnologia no campo ocasionou um aumento da quantidade e da heterogeneidade dos dados armazenados nas bases de dados organizacionais, causando uma grande complexidade na obtenção de conhecimento adequado e útil como auxílio ao gestor. Para esses casos, a técnica de mineração de dados surge com o objetivo de buscar e extrair o conhecimento dessas bases de dados, auxiliando o gestor no processo de tomada de decisão operacional e estratégica.

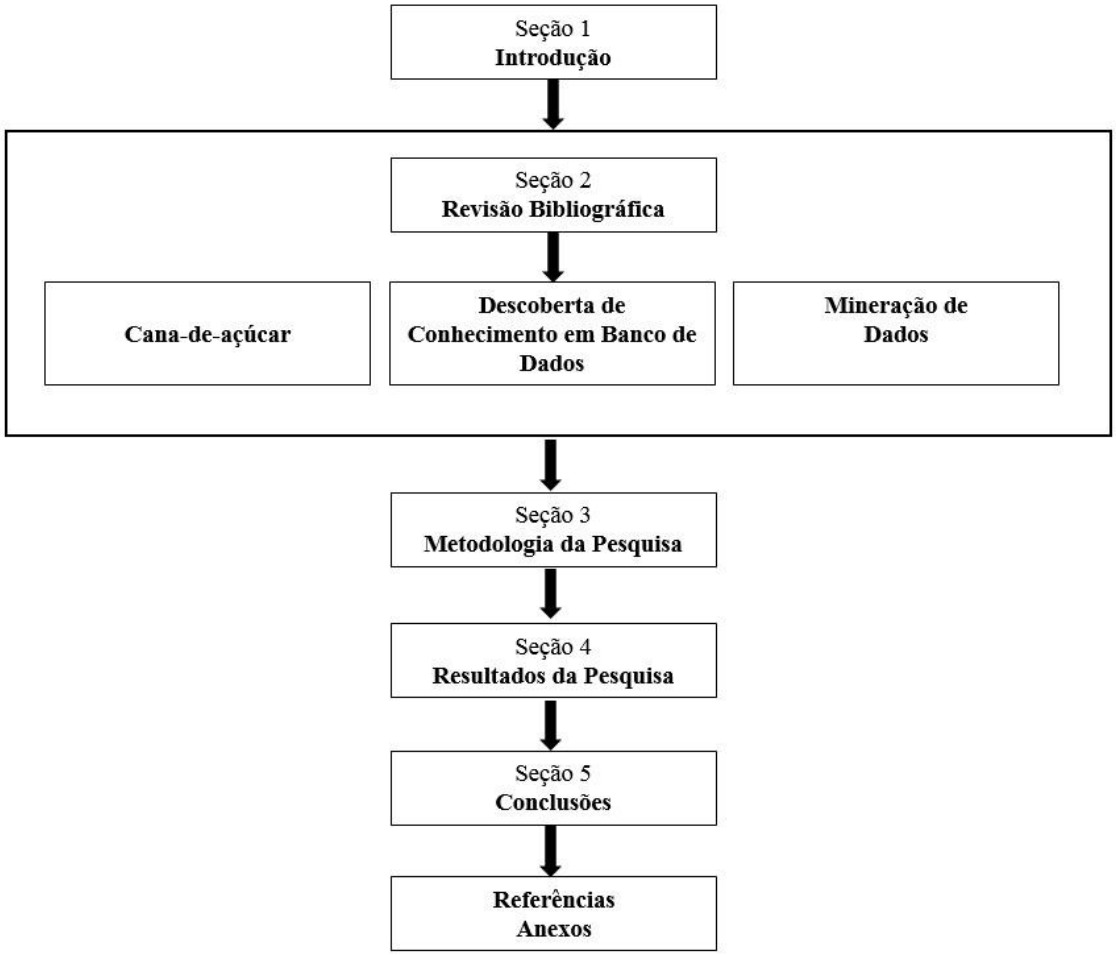
Este trabalho propõe, assim, a aplicação de Descoberta de Conhecimento em Base de Dados, por meio de um estudo realizado sobre os dados operacionais de uma empresa operadora agrícola, utilizando as técnicas de mineração de dados denominadas árvores de decisão (*decision tree*) e florestas aleatórias (*random forest*), com objetivo de identificar e classificar os fatores que impactam a produtividade do canavial.

Espera-se que esta pesquisa venha contribuir para o aprimoramento do sistema de gestão das operações agrícolas e que seja um instrumento útil para a área acadêmica, caracterizando-se como um trabalho com visão acadêmico-empresarial.

1.5 Estrutura do trabalho

A dissertação é composta por cinco seções e um conjunto de anexos considerados de interesse do trabalho desenvolvido. Uma visão geral da estrutura pode ser observada pela Figura 1:

Figura 1 – Estrutura da Dissertação.



Fonte: Elaborada pelo próprio autor (2016).

Nesta seção 1, apresentam-se uma contextualização breve do ambiente da pesquisa, os objetivos, a justificativa e os resultados esperados com a pesquisa.

Na seção 2, por meio de uma revisão bibliográfica, é abordada uma visão setorial da produção de cana-de-açúcar no Brasil. São apresentados os objetivos e conceitos, tipos de abordagem e metodologia relacionados ao processo de Descoberta de Conhecimento em Base de Dados. Além disso, uma abordagem sobre a Mineração de Dados mostra as técnicas mais utilizadas para validação dos modelos a serem utilizados no decorrer deste estudo, destacando as tarefas de árvores de decisão (*decision tree*) e florestas aleatórias (*random forest*).

A seção 3 tem como objetivo demonstrar a Metodologia de Pesquisa empregada e descrever o conjunto de atividades realizadas, que suportam esta pesquisa, de modo sistematizado, bem como o desenvolvimento das etapas baseadas na Descoberta de Conhecimento em Base de Dados.

Os resultados das práticas descritas na seção 3 são apresentados na seção 4 por intermédio da sua aplicação em banco de dados de empresa do setor agrícola produtora de cana-de-açúcar, por meio de uma pesquisa-ação, com demonstração real e discussão dos resultados obtidos baseados em técnicas de mineração de dados.

E, por último, na seção 5, são apresentadas as considerações finais do trabalho desenvolvido, identificando-se as principais contribuições para a área de Gestão Operacional e Estratégica no setor agrícola, sendo lançadas linhas orientadoras para trabalhos a serem desenvolvidos no futuro pelo setor empresarial e acadêmico.

2 FUNDAMENTAÇÃO TEÓRICA

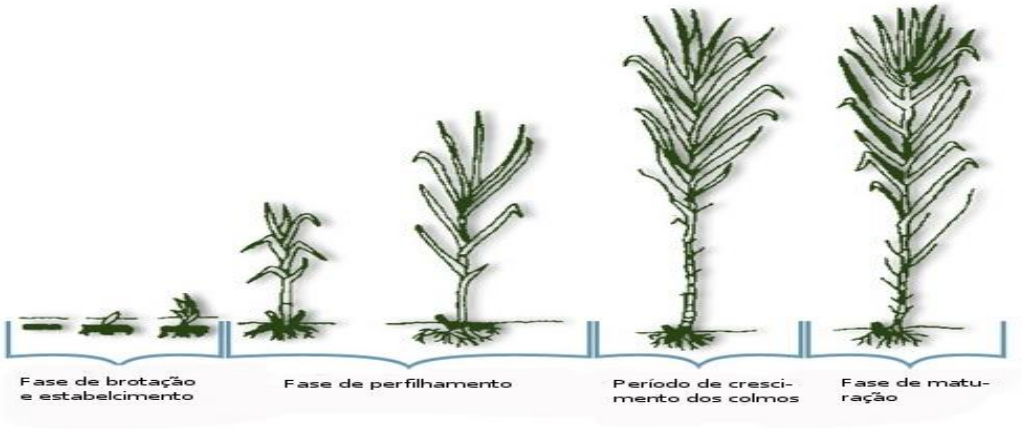
2.1 Fatores que impactam na gestão da cana-de-açúcar

O que torna o Brasil um dos mais tradicionais produtores de cana-de-açúcar é o seu cultivo em diferentes regiões brasileiras. Conforme os dados da CONAB (2016), a área colhida de cana-de-açúcar destinada à atividade sucroalcooleira na safra 2015/16 foi de 8.654,2 mil hectares, sendo São Paulo o maior produtor, com 52% (4.498,3 mil hectares), seguido por Goiás com 10,4% (885,8 mil hectares), Minas Gerais com 10,1% (866,5 mil hectares), Mato Grosso do Sul com 7% (596,8 mil hectares), Paraná com 6% (515,7 mil hectares), Alagoas com 3,7% (323,6 mil hectares), Pernambuco com 3% (254,2 mil hectares) e Mato Grosso com 2,7% (232,8 mil hectares); os outros 14 estados produtores possuem áreas menores que 1% da área total do país, compreendendo, juntos, 5,1% da área total do país. Sendo assim, a cana é cultivada em vários tipos de solos que estão sob influência de diferentes climas, resultando em diversos tipos de ambientes para a sua produção (DIAS, 1997).

São vários os fatores que podem influenciar na produtividade e na qualidade da cana-de-açúcar que, ao final, representa a integração das diferentes condições a que fica submetida a cultura (MARCHIORI, 2004). Os principais são: a interação edafoclimática, o manejo da cultura, a variedade cultivada, a queda do ritmo de ganhos de rendimentos agrícolas, a deterioração de indicadores de difusão tecnológica, além de problemas identificados que precisam de solução a partir de políticas públicas, como a dificuldade por parte de produtores e usinas independentes em acessar linhas de financiamento para ampliar o cultivo de cana e investir em máquinas e equipamentos, a insuficiência de recursos direcionados à Pesquisa e Desenvolvimento no âmbito do setor, tais como manejo de canaviais, etanol de segunda geração, motores mais eficientes na queima do etanol, desenvolvimento de outros usos e produtos derivados da cana-de-açúcar (CESAR et al., 1987; NYKO et al., 2013; NEVES; TROMBIN, 2014).

O estudo da cultura no seu ambiente de desenvolvimento pode gerar informações para adequar o melhor manejo e variedade cultivada para os específicos ambientes de produção (solo e clima). Assim, é possível extrair do ambiente de produção os melhores resultados através do melhor rendimento da cultura, o que pode proporcionar maior lucratividade ou competitividade para os produtos de cana-de-açúcar (MAULE, 2001). O processo de desenvolvimento da cana pode ser representado pela Figura 2 e pelo Quadro 1:

Figura 2 – Fases do desenvolvimento da cana.



Fonte: EMBRAPA (2016a).

Quadro 1 – Fases do desenvolvimento da cana-de-açúcar.

Brotação e emergência	O broto rompe as folhas da gema e se desenvolve em direção à superfície do solo. Ao mesmo tempo surgem as raízes do tolete. A emergência do broto ocorre de 20 a 30 dias após o plantio. O broto é um caule em miniatura que surge acima da superfície do solo (chamado de colmo primário). Esta fase depende da qualidade da muda, ambiente, época e manejo do plantio. Neste estágio ocorre, ainda, o enraizamento inicial (duas a três semanas após a emergência) e o aparecimento das primeiras folhas.
Perfilhamento **Início do perfilhamento e formação da touceira** **Auge do perfilhamento**	Perfilhamento é o processo de emissão de colmos por uma mesma planta, os quais recebem a denominação de perfilhos. O processo de perfilhamento é regulado por hormônios e resulta no crescimento de brotos que vão em direção à superfície do solo. Esses brotos aparecem de 20 a 30 dias após a emergência do colmo primário. Por meio desse processo, ocorre a formação da touceira da cana-de-açúcar e a população de colmos que será colhida. É importante destacar que a formação do sistema radicular da touceira é resultado do desenvolvimento das raízes de cada perfilho. O auge do perfilhamento ocorre quando há a total cobertura do solo pela folhagem das plantas. Cada touceira possui o máximo de perfilhos.
Crescimento dos colmos **Crescimento radicular vigoroso:** **Definição da população final de colmos:**	A partir do auge do perfilhamento, os colmos sobreviventes continuam o crescimento e desenvolvimento, ganhando altura e iniciando o acúmulo de açúcar na base. O crescimento é estimulado por luz, umidade e calor. Durante esta fase, as folhas mais velhas começam a ficar amareladas e secam. O crescimento do sistema radicular torna-se mais intenso, tanto nas laterais quanto em profundidade. A maior parte das raízes estão nos primeiros 40 centímetros de profundidade, sendo esta a zona principal no que concerne à absorção de água e nutrientes por parte da cultura. O canavial pode atingir altura acima de três metros, com a população final de colmos, variando em função das condições de clima e solo.
Maturação dos colmos **Maturação inicial:** **Maturação do terço médio:** **Maturação final:** **Momento de colheita:**	A maturação inicia-se junto com o crescimento intenso dos colmos sobreviventes do perfilhamento das touceiras. É válido mencionar, novamente, que o excesso de açúcar permanece armazenado na base de cada colmo. Quando as touceiras atingem altura igual ou superior a dois metros, nota-se o amarelecimento e a consequente seca das folhas que se encontram na altura mediana da planta, indicando que já está sendo depositado açúcar nessa região. No período entre o outono e o inverno, com a presença de chuvas variáveis e temperaturas mais baixas, existe maior atividade de maturação e menor atividade de crescimento, sendo que há intenso armazenamento de açúcar. É definido em função da variedade, época de plantio e consequente duração do ciclo, manejo da maturação e condições climáticas no ambiente.

Fonte: EMBRAPA (2016a).

“Com todas as informações disponíveis sobre o ciclo da planta, é possível identificar as relações e a influência dos fatores envolvidos no processo de produção, favorecendo a previsão de problemas, o manejo e a tomada de decisão” (EMBRAPA, 2016a, p. 1).

A análise de crescimento e cultura da cana-de-açúcar, segundo a EMBRAPA (2012), é realizada por meio da estimativa de índices morfofisiológicos, que necessitam de avaliações sequenciais em intervalos regulares, do acúmulo de biomassa seca e da área foliar. Também pode-se avaliar o crescimento da cana através de análises dinâmicas de perfilhamento e desenvolvimento das folhas, sendo que esses fatores limitam a quantidade de irradiação solar fotossinteticamente ativa que a cultura da cana pode absorver, impactando na eficiência do uso dessa cana.

> A dinâmica de crescimento dos colmos é a variável que apresenta maior correlação positiva com a produtividade, contudo, a mesma pode variar com a variedade e/ou as condições do ambiente de cultivo. A dinâmica foliar é também crucial na determinação da produtividade da cana-de-açúcar, uma vez que o baixo desenvolvimento foliar pode limitar expressivamente o rendimento da cultura, devido à redução na interceptação da irradiação solar incidente e, consequentemente, no menor acúmulo de biomassa (EMBRAPA, 2012, p. 5).

A quantidade de sacarose acumulada no colmo da cana, segundo a mesma fonte, é determinada pela anatomia dos colmos, ou seja, pela sua capacidade de armazenamento, além do seu metabolismo e transporte de açúcar nas folhas e nos drenos. Assim sendo, as características morfológicas da cana estão intimamente relacionadas com o acúmulo de sacarose, sendo determinantes da sua capacidade de armazenamento nas células do parênquima do colmo (EMBRAPA, 2012).

É imprescindível, então, conhecer a dinâmica de crescimento da cana, pois somente dessa forma pode-se aprimorar as práticas culturais, assim como é possível melhorar o aproveitamento das cultivares altamente produtivas em ambientes de produção diferentes.

Segundo Marchiori (2004), as informações da cana em relação ao seu comportamento futuro de maturação (acúmulo de sacarose) podem definir o manejo de variedades e, até mesmo, a época ideal da colheita.

De acordo com Rosseto (2016a), define-se o processo de maturação da cana-de-açúcar como um processo fisiológico constituído pela formação de açúcares nas folhas e seu deslocamento e armazenagem no colmo. Segundo a autora da Embrapa, essa maturação pode ocorrer sob três aspectos:

- Botânico: a cana só é considerada madura após a emissão de flores (Figura 2) e a formação de sementes. Na reprodução por toletes, a maturação é considerada quando as gemas estão em condições de dar origem a novas plantas;
- Fisiológico: a maturação ocorre quando o colmo atinge seu máximo armazenamento de açúcar (sacarose);
- Econômico: quando a cana atinge o teor mínimo de sacarose de 13% do peso do colmo, necessário para que possa ser viável industrialmente.

Esse armazenamento e transporte de água do açúcar se processa de forma lenta, iniciando já nos primeiros meses do crescimento da cana e permanecendo até o completo desenvolvimento dos colmos da planta. O acúmulo de sacarose atinge o seu máximo no momento em que ocorrem condições que venham a limitar ou restringir o seu crescimento (deficiência hídrica, insuficiência de nutrientes e condições desfavoráveis do clima). Esses fatores obrigam a planta a encerrar seu processo de crescimento e a amadurecer (ROSSETO, 2016a).

Por isso, conhecer previamente os fatores que podem restringir o crescimento da cana, alcançando o nível máximo de sacarose e forçando seu amadurecimento, é imprescindível para uma boa gestão da produtividade da cana-de-açúcar.

Segundo a Embrapa, usar fertilizantes em excesso pode favorecer de forma intensa o crescimento da cana, atrasando sua maturação. Além disso, "a farta quantidade de nitrogênio existente na época da colheita leva ao baixo conteúdo de sacarose da planta. Da mesma forma, a água em abundância durante todo o ciclo da cana prejudica sua maturação" (ROSSETO, 2016a, p. 1). É imprescindível planejar e gerir essas informações:

> É comum o emprego de práticas culturais de manejo de adubação e de irrigação com o objetivo de favorecer o amadurecimento da cana-de-açúcar com elevados teores de sacarose. Na medida em que as plantas apresentarem colmos bem desenvolvidos, a adubação e a irrigação devem ser limitados. Além destas práticas, o uso de maturadores tem sido amplamente utilizado (ROSSETO, 2016b, p. 1).

Os maturadores são considerados produtos elaborados quimicamente, que induzem ao amadurecimento da cana, o que causa a translocação e o armazenamento dos açúcares na planta. Os maturadores representam importante insumo utilizado de forma ampla para antecipar e otimizar o planejamento da colheita.

Segundo Rosseto (2016a, p. 1), para determinar se a cana-de-açúcar se encontra no ponto de maturação, utiliza-se o refratômetro de campo, aparelho que fornece a porcentagem

de sólidos solúveis do caldo (chamado de brix), que está ligado ao teor de sacarose da cana-de-açúcar. Após essa medição, é feita uma análise laboratorial.

Dentro dessas fases de desenvolvimento da cana, Marchiori (2004) lista alguns fatores que devem ser analisados para a otimização do acúmulo de sacarose nos colmos:

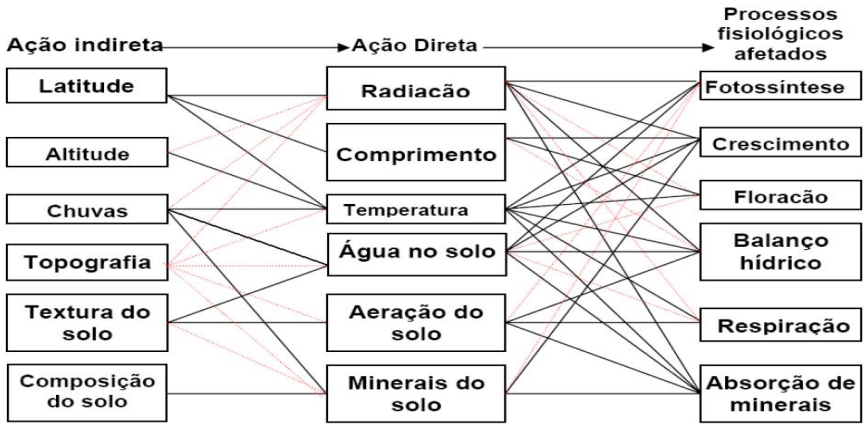
- Umidade do solo: Marchiori (2004) citou em seu trabalho que a umidade do solo é o fator mais influente para a absorção dos nutrientes do solo, o que se reflete, na cultura da cana, diretamente no desenvolvimento da planta e na produção de açúcar. Já em 1939, segundo estudos de Hartt, demonstrou-se que as plantas de cana providas de água se desenvolviam melhor que as outras sem tanta água (MARCHIORI, 2004). Na pesquisa, comprovou-se que as folhas das plantas cultivadas em solo úmido continham até dez vezes mais sacarose do que as outras, privadas de água. Concluiu o estudo que grandes provisões de água para o desenvolvimento da cana-de-açúcar são essenciais para que a planta forme grande quantidade de sacarose nas folhas, para seu transporte para o talo e seu reflexo no caldo.
- Temperatura: vários são os autores que entendem ser este o fator de maior relevância para o desenvolvimento da cana-de-açúcar. Marchiori (2004) incluiu em seus estudos as pesquisas de Humbert (1968), Alexandrer (1973), Ometo (1979, 1981) e Barbieri (1993), nas quais todos os resultados chegaram à importância da temperatura para a maturação fisiológica da cana-de-açúcar, pois afeta diretamente a absorção de água e nutrientes pela planta, por meio de fluxo transpiratório.
- Luminosidade: ligada ao processo de fotossíntese da planta, a luz solar está diretamente ligada ao armazenamento de açúcares e à acumulação de amido nas folhas. Luminosidade em abundância, conjugada a outros fatores importantes, representa a eficiência das plantas de cana (MARCHIORI, 2004).
- Nutrientes minerais: estes nutrientes influenciam diretamente na maturação da cana-de-açúcar conforme a época em que estão disponíveis para as plantas. Se forem tardias, as ofertas de nitrogênio no solo não atendem a sua finalidade e não favorecem o acúmulo de sacarose desejado. O nitrogênio deve ser oferecido de forma suficiente para a planta no período inicial, ou seja, na formação da cultura e na formação dos perfilhos, pois, no período posterior, de crescimento acelerado (desde que haja uma situação de umidade e temperatura) e, caso a oferta de

nitrogênio tenha sido insuficiente no período de formação, é baixo o número de perfilhos que participará desta etapa de crescimento (MARCHIORI, 2004).

- Floração: a floração é uma resposta aos estímulos ambientais que a cana recebe. A transformação da gema apical de vegetativa para reprodutiva é uma consequência desses estímulos e passa a impedir de forma parcial o crescimento da planta, interrompe a formação de novos entrenós e cria uma situação favorável para a acumulação de sacarose (MARCHIORI, 2004).

Dessa forma, conforme se extrai da literatura, a temperatura, a umidade e a luminosidade são as variáveis determinantes do clima que se refletem diretamente no desenvolvimento da cana-de-açúcar. Marchiori (2004), que expôs em sua Tese diversos outros estudos, concluiu que o clima é o fator que mais influencia na produtividade da cana. Estudou, contudo, a interação do clima com outras variáveis que podem afetar na produtividade. Na Figura 3, as linhas cheias correspondem a uma ação mais efetiva do que as linhas interrompidas.

Figura 3 – Classificação dos fatores de produção vegetal que afetam direta e indiretamente os processos fisiológicos das plantas



Fonte: Marchiori (2004).

Segundo Segato et al. (2006 apud ANDRETTA, 2012), são diversas as variáveis ambientais, de manejo e genéticas nas fases do ciclo fenológico da cana-de-açúcar. Considerando somente as ambientais, a brotação e o enraizamento são influenciados pelas variáveis temperatura, umidade relativa do ar, doenças, pragas, plantas daninhas, textura e estrutura do solo. O perfilhamento é influenciado pelas variáveis: radiação solar, temperatura, umidade, aeração, vento, doenças, pragas e plantas daninhas. E a maturação é influenciada pelas variáveis: temperatura, umidade, vento, pragas, doenças e características do solo.

2.2 Conceito e etapas do processo de KDD

Begoli e Horey (2012) referem-se ao KDD como um conjunto de atividades destinadas a extrair novos conhecimentos a partir de conjuntos de dados complexos. O processo de KDD é muitas vezes interdisciplinar e abrange a ciência da computação, a estatística, a visualização e a experiência de domínio. Em anos recentes, grandes quantidades de dados se tornaram cada vez mais disponíveis em volumes significativos. Esses dados têm muitas fontes, incluindo atividades *online* (redes e mídia sociais), telecomunicativas (computação móvel, estatísticas de chamadas), científicas (simulações, experiências sensoras ambientais) e agrupamento das fontes tradicionais (formulários, inquéritos).

Para Han e Kamber (2006), como consequência, o KDD tornou-se estrategicamente importante para as grandes empresas comerciais, organizações governamentais e instituições de pesquisa, pois ele possibilita a produção de conhecimento a partir de grandes conjuntos de dados, e isso é muito útil especialmente para organizações de grandes empresas compostas por múltiplas sub-organizações.

Um KDD requer a adoção de práticas organizacionais e tecnológicas eficazes. Assim, os processos de descoberta de conhecimento são compostos de: coleta de dados, armazenagem e práticas de organização; compreensão e aplicação efetiva da organização dos dados; adoção de métodos analíticos de dados (incluindo ferramentas); entendimento do domínio do problema, da natureza e da estrutura dos dados subjacentes (HAN; KAMBER, 2006).

O KDD é como o processo global de descoberta de conhecimento útil a partir de dados. A mineração de dados é um passo particular nesse processo de aplicação de algorítmos específicos para extração de padrões (modelos) a partir de dados. Os passos adicionais no processo KDD englobam a preparação de dados, a seleção de dados, os dados de limpeza, a incorporação do conhecimento prévio apropriado e uma interpretação correta dos resultados de mineração que devem disponibilizar um conhecimento útil derivado a partir dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

KDD evoluiu e continua evoluindo, a partir da interseção da pesquisa em áreas como bancos de dados, aprendizado de máquina, reconhecimento de padrões, estatísticas, inteligência artificial, raciocínio com a incerteza, aquisição de conhecimento para sistemas especialistas, visualização de dados, descoberta de máquina, descoberta científica, recuperação de informação e computação de alto desempenho. Os sistemas de software KDD incorporam teorias, algoritmos e métodos de todos esses campos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

Sass (2013) reflete que o processo do conhecimento engloba a transformação dos dados em informação e conhecimento. Nesse contexto, os dados podem ser compreendidos como a matéria bruta para esse processo e guardam os aspectos dos fenômenos que estão sendo estudados. Assim, a informação pode ser entendida como o resultado de um processamento executado nesses dados e o conhecimento é um conjunto de argumentos e explicações que interpretam o conjunto de informações.

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996b), as etapas do processo do KDD englobam:

1. Aprendizado do domínio da aplicação: inclui o conhecimento prévio relevante e os objetivos da aplicação;

2. Criação de um conjunto de dados de destino: inclui a seleção de um conjunto de dados concentrando-se em um subconjunto de variáveis ou amostras de dados em que a descoberta deve ser executada;

3. Limpeza de dados e pré-processamento: incluem as operações básicas, tais como remoção de ruído, recolha da informação necessária para modelar, decidindo sobre as estratégias para lidar com a falta de campos de dados e representando informações da sequência de tempo e as mudanças conhecidas, bem como decidir questões de *database management system (*DBMS), como tipos de dados, esquema e mapeamento e valores desconhecidos;

4. Redução de dados e projeção: inclui encontrar recursos úteis para representar os dados, dependendo do objetivo da tarefa e usando métodos de redução de dimensionalidade ou de transformação para diminuir o número efetivo de variáveis sob consideração ou para encontrar representações invariáveis para os dados;

5. Escolha da função de exploração de dados: inclui decidir a finalidade do modelo derivado pelo algoritmo de extração de dados (por exemplo, compactação, classificação, regressão, e aglomeração);

6. Escolha do algoritmo(s) de mineração de dados: inclui a seleção do método(s) a ser utilizado(s) para a busca de padrões nos dados, tais como decidir quais os modelos e parâmetros podem ser apropriados (por exemplo, modelos para dados categóricos são diferentes dos modelos de vetores mais reais), combinando um método de mineração de dados em particular com os critérios gerais do processo de KDD (por exemplo, o usuário pode ser mais interessado em entender o modelo do que suas capacidades de previsão);

7. Mineração de dados: inclui a busca de padrões de interesse em uma forma de representação particular ou um conjunto de tais representações, incluindo as regras de

classificação ou árvores, regressão, *clustering*, sequência de modelagem, dependência e análise linear;

8. Interpretação: inclui a interpretação dos padrões descobertos, provavelmente retornando para qualquer uma das etapas anteriores, bem como a possível visualização dos padrões extraídos, removendo padrões redundantes ou irrelevantes e traduzindo as informações úteis em termos compreensíveis pelos utilizadores;

9. Uso do conhecimento emergente, inclui a incorporação desse conhecimento para o sistema de desempenho, tomando medidas com base no conhecimento, ou simplesmente o documentando e o denunciando às partes interessadas, bem como realizando a verificação de resolver possíveis conflitos com prévio conhecimento.

Neste trabalho será dado destaque à etapa de mineração de dados, apresentada com outro subtítulo, a seguir, por ser o foco neste estudo.

2.3 Mineração de dados

Para Castanheira (2008), a mineração de dados pode ser compreendida como uma ferramenta utilizada para a descoberta de novos padrões, correlações e tendências entre as informações de uma empresa, por intermédio da análise de grandes quantidades de dados armazenados em *Data Warehouse (DW)*, utilizando técnicas de reconhecimento de matemática, estatística e padrões.

A mineração de dados envolve modelos que determinam os padrões de dados observados. Os modelos ajustados desempenham o papel de conhecimento inferido. Decidir em utilizar ou não os modelos mostra como a mineração de dados é conhecimento útil e faz parte de um processo de KDD interativo global para que o julgamento humano subjetivo seja geralmente necessário. Uma grande variedade e número de algoritmos de mineração de dados são descritos na literatura, a partir dos campos de estatísticas, reconhecimento de padrões, aprendizagem de máquina e bases de dados. Assim, uma discussão da visão geral pode, muitas vezes, consistir em longas listas de algoritmos aparentemente não relacionadas e altamente específicas. Aqui vamos dar um ponto de vista um tanto reducionista (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

A maior parte dos algoritmos de mineração de dados podem ser vistos como composições de algumas técnicas e princípios básicos. Em particular, algoritmos de mineração de dados consistem em grande parte de alguma mistura específica de três componentes (FAYYAD, PIATETSKY-SHAPIRO; SMYTH, 1996b):

- O modelo: existem dois fatores relevantes, a função do modelo (como exemplo, classificação e agrupamento) e a forma de representação do modelo (por exemplo, uma função linear de variáveis múltiplas e a Função Densidade de Probabilidade Gaussiana). Um modelo contém parâmetros que devem ser determinados a partir dos dados.
- O critério de preferência: uma base para a preferência de um modelo ou um conjunto de parâmetros sobre a outra, dependendo dos dados apresentados. O critério é normalmente algum tipo de função de bondade de ajuste do modelo aos dados, talvez temperado por uma suavização, para evitar o excesso de montagem, ou gerar um modelo com muitos graus de liberdade a ser condicionada pelos dados fornecidos.
- O algoritmo de busca: a especificação de um algoritmo para encontrar determinados modelos e parâmetros, dados fornecidos, um modelo (ou família de modelos) e um critério de preferência.

Um algoritmo específico de mineração de dados é criação do modelo/preferência componentes de pesquisa (por exemplo, um modelo de classificação baseado em uma árvore de decisão). Além disso, os algoritmos muitas vezes diferem em termos da representação do modelo (por exemplo, linear e hierárquica), e também modelo de preferência ou métodos de busca são frequentemente similares em diferentes algoritmos. A literatura sobre algoritmos de aprendizagem frequentemente não indica claramente o modelo de representação, ou método de pesquisa utilizado, estes são muitas vezes misturados na descrição de um determinado algoritmo (FAYYAD, PIATETSKY-SHAPIRO; SMYTH, 1996b).

Two crowns corporation (2005) menciona que a mineração de dados aproveita os avanços nos campos da inteligência artificial (IA) e estatística. Ambas as disciplinas têm trabalhado em problemas de reconhecimento de padrões e classificação e têm feito grandes contribuições para a compreensão e a aplicação de redes neurais e árvores de decisão. A mineração de dados não substitui as técnicas estatísticas tradicionais. Pelo contrário, é uma extensão de métodos estatísticos que é, em parte, o resultado de uma grande mudança na comunidade estatística. O desenvolvimento da maioria das técnicas estatísticas era, até recentemente, com base na teoria elegante e métodos analíticos que funcionaram muito bem nas quantidades modestas de dados que estão sendo analisados. O aumento da potência dos computadores e seu baixo custo, juntamente com a necessidade de analisar enormes conjuntos de dados com milhões de linhas, têm permitido o desenvolvimento de novas técnicas baseadas em uma exploração de força bruta de soluções possíveis.

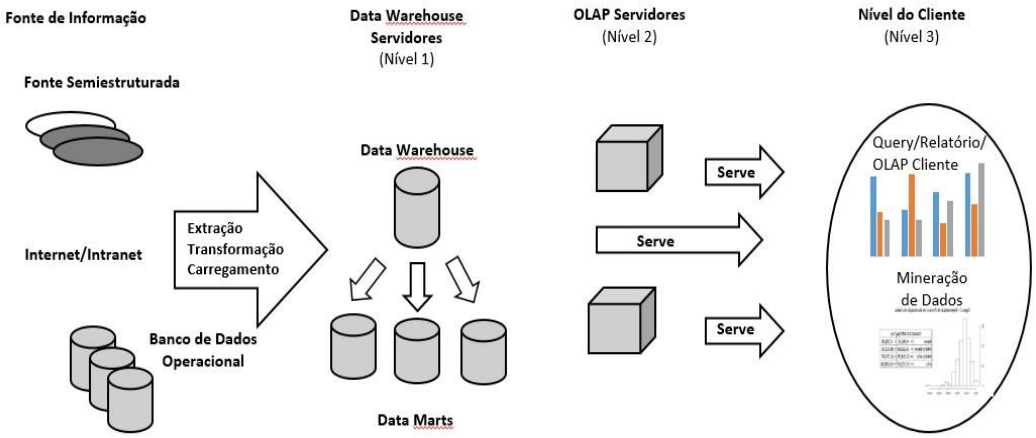
Maimon e Lior (2010), afirmam que muito pode ser obtido em mineração de dados, explorando o conhecimento sobre a estatísticas, embora as ferramentas de visualização das estatísticas geralmente não estejam calibradas para os conjuntos de dados comumente tratados na mineração de dados. Por exemplo, o Boxplots, usado para comparações visuais de lotes de dados. A amostragem também é uma área muito bem desenvolvida de estatísticas, mas é usualmente usado em mineração de dados no nível muito básico. Segundo os autores a a estatística tornou-se uma disciplina fechada, com jargão científico e objetivos acadêmicos que favorecem provas analíticas ao invés de práticos métodos de aprendizagem a partir de dados. Ainda, se faz necessário distinguir entre a matemática teórica estatística e seu uso como uma ferramenta em muitos estudos experimentais de investigação científica, e destacam que a metodologia de computação e muitas outras questões ser incorporadas às estatísticas tradicionais.

Novas técnicas incluem algoritmos relativamente recentes, como redes neurais e árvores de decisão e novas abordagens para algoritmos mais antigos, como análise discriminante. Em virtude de trazer o aumento do poder do armazenamento de dados sobre os enormes volumes de dados disponíveis, essas técnicas podem aproximar quase qualquer forma funcional ou interação por conta própria. Técnicas estatísticas tradicionais contam com um modelador para especificar a forma funcional e interações. O ponto-chave é que a mineração de dados é a aplicação dessas e de outras técnicas de IA. A mineração de dados é uma ferramenta para aumentar a produtividade das pessoas tentando construir modelos preditivos (TWO CROWNS CORPORATION, 2005).

Nesse aspecto, Maimon e Lior (2010) ressaltam que a produção constante de informação ultrapassou a capacidade analítica do homem, o que torna o processo mineração de dados importante para as empresas no que tange ao processo de geração de conhecimento através da abundância dos dados.

Ainda segundo os autores, a abundância de dados faz com a Mineração de Dados se torne parte integrante da Tecnologia da Informação (TI). Conforme ilustrado na Figura 4, são três níveis do aspecto de apoio à decisão de TI, que parte das fontes de dados (tais como bancos de dados operacionais, dados semiestruturados, não estruturados e relatórios, sites da Internet, etc.), a primeira camada é o armazém de dados, seguida por servidores *On Line Analytical Processing* (OLAP) e termina com ferramentas de análise, onde as ferramentas de mineração de dados *mining* são as mais avançadas.

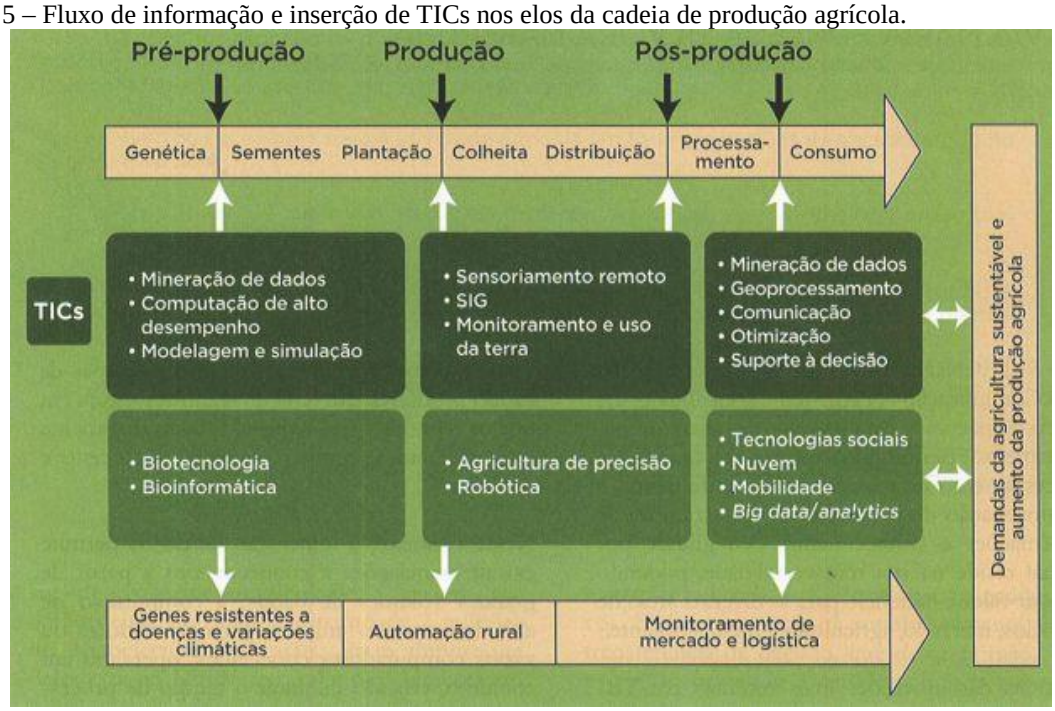
Figura 4 - Os níveis de Suporte para Decisão de TI.



Fonte: Adaptado de Maimon e Lior (2010, p. 7).

Esse conceito apresentado é reforçado no trabalho de Massruhá e Leite (2016), que analisa o uso da Tecnologia da Informação e Comunicação (TIC), num setor específico: a agricultura, e ressalta que a tecnologia tem contribuído para diversas áreas de conhecimento, auxiliando no armazenamento e processamento de grande volumes de dados, automatização de processos e intercâmbio de informações e conhecimento, além de agregar valor e benefício para as diversas áreas de negócio, mercado, agricultura e meio ambiente.

Figura 5 – Fluxo de informação e inserção de TICs nos elos da cadeia de produção agrícola.



Fonte: Massruhá (2016).

Nesse contexto, a mineração de dados permite extrair associações e conhecimento a partir de grandes volumes de dados. A computação de alto desempenho utiliza supercomputadores ou vários computadores conectados, operando em conjunto, visando diminuir o tempo de processamento de sistemas complexos informatizado. No entanto, o risco da abordagem integrada de TI vem do fato de que as técnicas de DM são mais complexas do que OLAP, por exemplo, pois os usuários precisam ser treinados adequadamente para condução da melhor técnica de mineração de dados para extração de conhecimentos e definição das tecnologias que irão propiciar o monitoramento do uso da terra, o desenvolvimento e a disseminação de sistemas automatizados de suporte à decisão (MAIMON; LIOR, 2010; MASSRUHÁ; LEITE, 2016).

Vale destacar que a escolha de quais técnicas de mineração de dados aplicar, envolve uma análise preliminar de qual será a tarefa de mineração a ser realizada. As exigências das tarefas de mineração e as suas características influenciam a viabilidade entre os métodos de mineração e os problemas do negócio (PASTA, 2011).

2.3.1 Metodologia da mineração de dados

Atualmente existem diversas metodologias desenvolvidas com o objetivo de definir e padronizar as fases da mineração de dados. Dentre as quais se destacam: a *Sample, Explore, Modify, Model, Assesment* (SEMMA), desenvolvida pela SAS, e a *Cross-Industry Standard Process of Data Mining* (CRISP-DM), que, de acordo com diversos autores, é a mais utilizada. Segundo Maimon e Lior (2010), a metodologia CRISP-DM surgiu em uma tentativa de padronizar o processo de mineração de dados. Quando comparada ao SEMMA, tem a vantagem de ser neutra no que tange à adoção de ferramentas de mineração de dados.

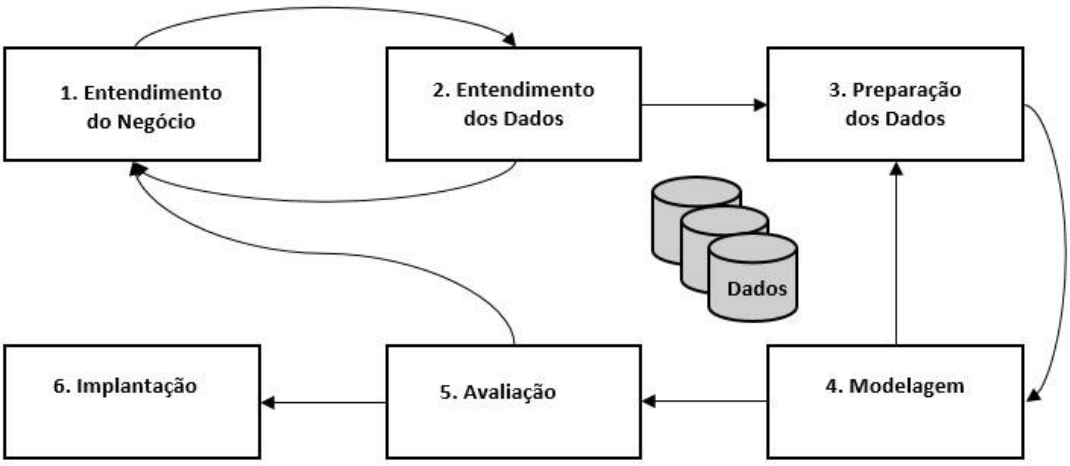
Neste trabalho, é adotada a metodologia CRISP-DM como modelo. Sua escolha é fundamentada pela disponibilidade de literatura disponível e por ser considerada o padrão de maior aceitação, conforme citado no trabalho de Camilo e Silva (2009).

A Figura 6 representa o clico de vida da metodologia CRISP-DM, que consiste em seis fases, segundo Maimon e Lior (2010):

- Entendimento do Negócio: o foco nesta etapa é entender qual o objetivo que se deseja atingir com o projeto de mineração de dados e quais requisitos necessários para alcançá-lo, com base na perspectiva de negócio. Esse processo é necessário para definição do problema e criação do plano para execução das próximas fases.

- Entendimento dos dados: refere-se à fase de recolhimento inicial dos dados e contempla atividades que possibilitam ao pesquisador familiaridade com as fontes dos dados.
- Preparação dos dados: considerando que os dados são oriundos de diversas fontes é comum que eles não estejam preparados para a sua mineração e, nesta fase, tabelas, registros e atributos são transformados e limpos. O processo de limpeza de dados geralmente envolve filtrar, combinar e preencher valores vazios.
- Modelagem: nesta fase, aplicam-se as técnicas (algoritmos) de mineração. A definição da(s) técnica(s) depende dos objetivos desejados. Os parâmetros são calibrados de forma a atingir o volume ótimo.
- Avaliação: considerada uma fase crítica, pois os modelos são avaliados para verificar se possibilitam cumprir os objetivos propostos no processo inicial. A verificação é realizada por testes e validações, visando obter confiabilidade nos modelos.
- Implantação: busca-se nesta fase a extração de conhecimentos dos modelos utilizados com os dados e posterior apresentação dos resultados aos envolvidos no processo.

Figura 6 – O ciclo CRISP – DM.



Fonte: Mainom e Lior (2010, p. 1033).

A aplicação desta metodologia em projetos de mineração de dado garante uma maior confiabilidade, redução de custos de execução, maior segurança, exequibilidade e viabilidade, visto que, trata-se de uma metodologia completa e documentada, uma vez que suas fases estão devidamente organizadas, estruturadas e definidas, o que leva à um projeto que pode ser

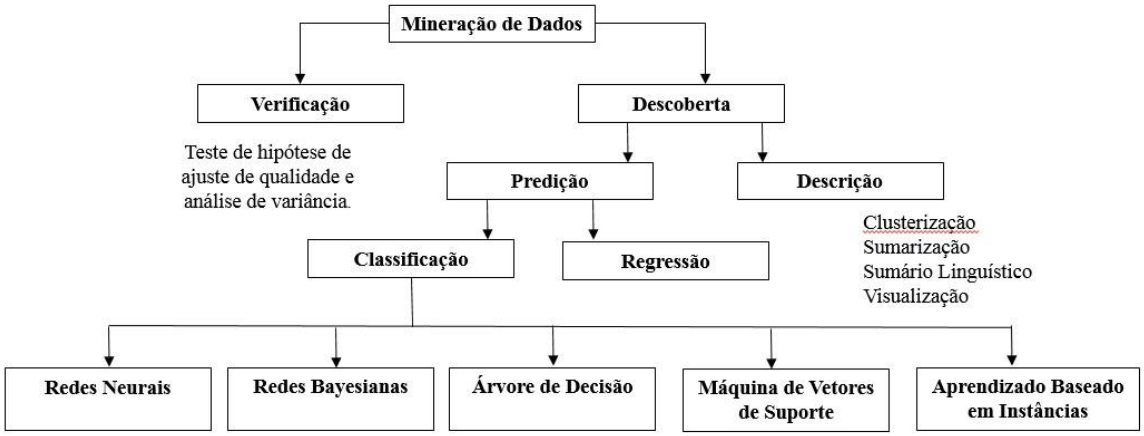
facilmente compreendido ou revisto. E como resultado da sua aplicação obtém-se um conjunto de documentação sobre o processo em vários relatórios - estudo do negócio, relatório inicial de dados, relatório de descrição e qualidade dos dados, relatório da modelagem relatório da avaliação, plano de implementação, manutenção e relatório final (MATTOZO, 2007).

2.3.2 Principais tarefas da mineração de dados

A mineração de dados é comumente classificada pela sua capacidade em realizar determinadas tarefas. Na literatura, é encontrada uma quantidade diferenciada de tarefas para mineração de dados.

De acordo com Maimon e Lior (2010), essas tarefas podem ser unidas em dois grupos, como se pode ver na Figura 7:

Figura 7 – Tarefas de mineração de dados.



Fonte: Miamon e Lior (2010, p. 06).

Como a literatura descreve grande número de tarefas de mineração de dados, este trabalho apresenta as que foram aplicadas na pesquisa.

2.3.2.1 Classificação: mapeamento dos dados de entrada

Uma classificação mapeia dados dentro de várias classes pré-definidas (FAYYAD, PIATETSKY-SHAPIRO; SMYTH, 1996b). Sass (2013) ressalta que a categorização de dados em classes tem o intuito de descobrir relacionamentos entre um atributo meta e um conjunto de atributos de previsão.

A tarefa de classificação pode ser entendida como uma função de aprendizado que realiza o mapeamento dos dados de entrada, ou dos conjuntos de dados de entrada, dentro de um número finito de classes. Nesse contexto, o objetivo de um algoritmo de classificação

seria encontrar correlações entre os atributos de uma classe, de forma que o processo de classificação possa ser usado para predizer uma determinada classe. Tem por princípio a descoberta de algum tipo de relacionamento entre os atributos preditivos e o atributo meta, com o objetivo de descobrir um novo conhecimento, o qual possa ser aplicado na previsão de uma nova classe, ainda desconhecida (CASTANHEIRA, 2008; PASTA, 2011).

Pasta (2011), através de dados de uma instituição de ensino, exemplifica que um algoritmo de classificação pode analisar os dados do quadro 2, a fim de determinar quais valores dos atributos preditivos devem ser relacionados, com cada um dos atributos objetivos. Com base nesta descoberta, o conhecimento gerado pode contribuir para a previsão de futuras evasões por parte dos acadêmicos da respectiva instituição.

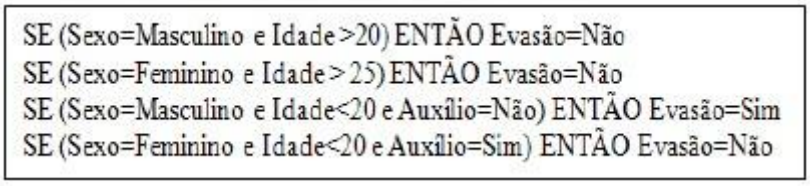
Quadro 2 – Entrada de dados para a tarefa de classificação.

Sexo	Idade	Auxílio	Evasão
Masculino	26	Sim	Não
Feminino	19	Não	Sim
Masculino	19	Não	Sim
Masculino	30	Não	Não
Feminino	20	Sim	Não
Feminino	29	Não	Não
Masculino	18	Não	Sim

Fonte: Pasta (2011, p. 72).

Pasta (2011) descreve que a representação do conhecimento é descoberta na forma de regras do tipo SE-ENTÃO, e a interpretação é, "SE os atributos preditivos satisfazem a uma condição no antecedente da regra, ENTÃO a classe é indicada na consequente da regra". Tal composição é apresentada na Figura 8 e mostra as regras extraídas de um algoritmo de classificação, tendo como atributos os do quadro acima.

Figura 8 – Regras de classificação



Fonte: Pasta (2011, p. 72).

Às vezes, um especialista classifica uma amostra do banco de dados, e este a classificação é então usado para criar o modelo que será aplicado a toda a base de dados, além de ser necessário fazer alguns experimentos com algoritmos disponíveis com o intuito de

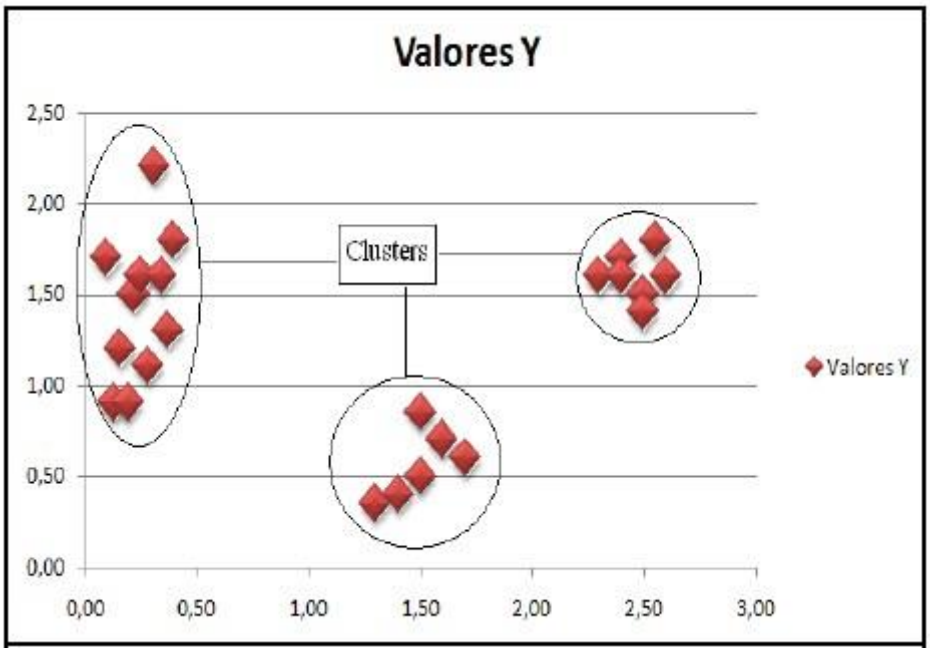
checar qual melhor se adequa a aplicação em questão, pois são diversos algoritmos aplicado nas tarefas de classificação, com destaque para as Redes Neurais, Back-Propagation, Classificadores Bayesianos e Algoritmos Genéticos (*TWO CROWNS CORPORATION*, 2005; PASTA, 2011).

2.3.2.2 Clusterização

A Clusterização mapeia um item de dados em uma das várias classes categóricas (ou *clusters*) em que as classes devem ser determinadas a partir da classificação de dados, ao contrário do que ocorre quando as classes são pré-definidas. Os *clusters* são definidos pela descoberta de agrupamentos naturais de itens de dados com base em métricas de similaridade ou de densidade de probabilidade (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

De acordo com Sass (2013), a clusterização separa grupos heterogêneos, a priori, em subgrupos mais homogêneos. Como exemplo, tem-se o fato de agrupar clientes por região de país e agrupar clientes com comportamento de compra similar. Pasta (2011), através da Figura 9 exemplifica o uso da tarefa de Clusterização, que é considerada como uma tarefa que identifica um conjunto finito de categorias com o objetivo de descrever os dados, e principalmente fazer a partição da base de dados em um número determinado de clusters, nos quais as instâncias destes clusters sejam similares, conforme demonstrado abaixo:

Figura 9 – Exemplo da visualização de clusters.



Fonte: Pasta (2011, p. 76).

Nota-se que os dados analisados podem ser agrupados em classes ou clusters de elementos similares. Nenhuma informação sobre a existência de determinadas classe é dada ao sistema, a descoberta é feita pelo próprio algoritmo, que agrupa os dados em classes com as mesmas características. Na clusterização não há classes pré-definidas, ao contrário da classificação.

2.4 Métodos de mineração de dados (técnicas)

Maimon e Lior (2010) relatam que a abundância de informação faz com que exista um conjunto de objetivos de mineração de dados, e a taxonomia da mineração de dados permite a compreensão da relação existente entre os objetivos e as técnicas de mineração, constituídos basicamente por dois tipos:

- Orientado à verificação: o sistema verifica as hipóteses do utilizador;
- Orientado à descoberta: o sistema descobre novas regras e padrões automaticamente.

Quanto à orientação para a descoberta, esta pode ser dividida em:

- Previsão: baseada na construção de um modelo de comportamento capacitado para previsão de valor de uma ou mais variáveis relacionadas com itens novos (para quais o modelo não foi treinado);
- Descrição: com objetivo de compreender como os dados subjacentes estão se relacionando com as suas partes.

A aprendizagem indutiva é o que determina a maior parte das técnicas de mineração de dados orientadas para a descoberta, sendo que os modelos são construídos de forma explícita ou implícita pela generalização de um número suficiente de exemplos de treino. Já os métodos de verificação possibilitam a avaliação de hipótese proposta por fonte externa. Ambos os métodos consideram as técnicas de estatísticas tradicionais, porém menos associados à área de mineração de dados, cujos problemas objetivam a descoberta de novas hipóteses, descartando-se aquelas já conhecidas (MAIMON; LIOR, 2010).

A aprendizagem de máquina, por sua vez, baseia-se no princípio indutivo, no qual as informações de um conjunto de exemplos fornecem informação para a generalização do todo. Consiste na capacidade da máquina em melhorar o desempenho em determinada tarefa através da experiência passada (MITCHELL, 1997).

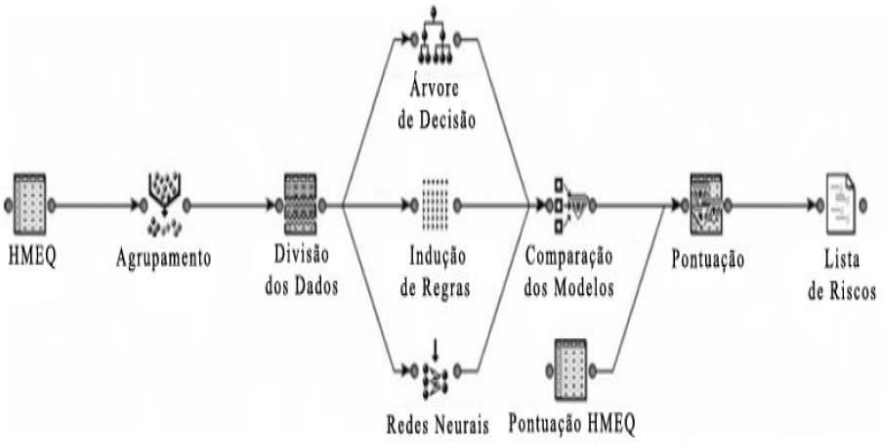
O processo de aprendizado indutivo pode ser, assim, dividido em supervisionado e não supervisionado por reforço. Tradicionalmente, os métodos de mineração de dados também são

separados em aprendizado supervisionado (preditivo) e não supervisionado, com uma divisão muito tênue entre os métodos.

Os autores Maimon e Lior (2010) destacam que o processo de aprendizado supervisionado se refere às técnicas que tentam descobrir a relação entre os atributos de entrada (variáveis independentes) e o atributo de saída (variável dependente). Como os exemplos fornecidos ao algoritmo contêm a variável resposta desejada, facilita o ajuste do algoritmo de aprendizado, de acordo com o viés apresentado. Dessa maneira, no processo de aprendizado supervisionado podem haver correspondentes a dois objetivos de mineração de dados: classificação (no caso de rótulo de dados discretos, a qual também pode ser utilizada pelo método não supervisionado) ou regressão (em caso de uma variável resposta contínua) (REZENDE, 2003; McCUE, 2007; FACELI et al., 2011). Já o aprendizado não supervisado não exige uma pré-categorização para os registros, ou seja, não há necessidade de um atribuo alvo (CAMILO; SILVA, 2009).

McCue (2007) recomenda que durante o processo de mineração de dados, diversas técnicas devem ser testadas e combinadas, pois este processo possibilita comparações com o objetivo de definir a melhor técnica (ou combinação de técnicas) a ser utilizada, conforme demonstrado na Figura 10.

Figura 10 – Processo de comparação com algumas técnicas.



Fonte: McCue (2007, p.126).

O processo proposto por McCue (2007), anteriormente relatatado, corrobora com a afirmação de que um algoritmo de mineração de dados em particular é, geralmente, uma instanciação do modelo / preferências / componentes de pesquisa, de preferência na busca de um modelo de métodos que muitas vezes são semelhantes em diferentes algoritmos. A literatura sobre algoritmos frequentemente não indica, com clareza, a representação do

modelo, critério de preferência ou método de pesquisa utilizado, os quais são, muitas vezes, misturados em uma descrição de um algoritmo específico (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996b).

No aprendizado supervisionado, os exemplos fornecidos ao algoritmo apresentam a variável resposta desejada, o que permite o ajuste do algoritmo de aprendizado conforme o viés mostrado, sendo a relação descoberta representada em uma estrutura denominada modelo. Esses modelos geralmente descrevem e explicam fenômenos que estão escondidos no conjunto de dados e que podem ser usados para prever a resposta. Nesse tipo de aprendizado, tem-se dois principais modelos: classificação (modelos classificadores), no caso de rótulo de dados discreto e regressão, no caso uma variável resposta contínua (REZENDE, 2003; MAIMON; LIOR 2010; FACELI et al., 2011).

No aprendizado não supervisionado, o algoritmo usado tem como função analisar os exemplos disponíveis quanto às possíveis relações, agrupamento ou padrões que os dados apresentam. Os autores Maimon e Lior (2010) comentam que o objetivo da tarefa de agrupamento (*clustering)* é descritiva. Ao buscar descobrir um novo conjunto de categorias, os novos grupos são de interesses em si mesmos e sua avaliação é intrínseca.

2.4.1 Métodos/técnicas de mineração de dados

Conforme Harrison (1998), não se pode falar em técnica ou técnicas próprias para resolver todos os problemas de mineração de dados. Ao contrário, devem ser testados diferentes métodos que se adaptam a diferentes propósitos, pois cada um oferece suas vantagens e desvantagens, necessitando ser avaliado em relação ao problema em questão. Contudo, é importante destacar que o usuário precisa ter relativa familiaridade com as técnicas disponíveis, pois somente assim poderá optar por uma delas, tornando a escolha mais simples.

Chen, Han e Yu (1996), afirmam que diferentes esquemas de classificação podem ser aplicados para categorização de métodos de mineração de dados sobre os tipos de base de dados a serem estudados, os tipos de conhecimentos a serem descoberto e os tipos de técnicas a serem utilizadas, conforme descrito a seguir:

- Com que tipo de bases de dados se deve trabalhar: a classificação pode ser feita de acordo com os tipos de dados sobre as quais as técnicas de mineração serão aplicadas, como: base de dados relacionais, bases de dados de transação, orientadas a objetos, dedutivos, espaciais, temporais, informações da internet, bases textuais;

- Qual o tipo de conhecimento deve ser explorado: vários tipos de conhecimento podem ser descobertos por meio da extração de dados, incluindo regras de associação, regras características, regras de classificação, regras discriminantes, grupamentos;
- Qual tipo de técnica deve ser utilizado: a extração pode ser categorizada de acordo com as técnicas de mineração de dados subordinadas. Como exemplo, extração dirigida a dados ou a questionamento, extração de dados interativa, ou ainda, pode ser categorizada, ou seja, subordinada, tal como: extração de dados baseadas em generalização, padrões, em teorias estatísticas ou matemáticas, entre outras.

Para McCue (2007), "durante o processo de mineração, diversas técnicas devem ser testadas e combinadas a fim de que comparações possam ser feitas e então a melhor técnica (ou combinação de técnicas) seja utilizada".

O extenso número de problemas de mineração de dados faz com que não haja uma técnica que possa ser aplicada para resolução de todos eles, pois cada problema possui suas peculiaridades, o que exige a aplicação de técnicas distintas para a resolução de problemas com propósitos diferentes. As diferentes técnicas para extração do conhecimento em base de dados consistem na aplicação de um mais algoritmo, implementados em ferramentas acadêmicas ou comerciais, para descobrir o conhecimento a partir da base de dados a ser explorada (PASTA, 2011).

Ainda, de acordo com o autor, as técnicas de mineração de dados aplicadas atualmente são extensões naturais ou generalização de métodos analíticos já conhecidos. O que se altera é intenção de aplicação destas técnicas com o objetivo de proporcionar a descoberta de conhecimentos que auxiliem os gestores no processo de tomada de decisão, devido ao crescente armazenamento de dados e à redução nos custos de processamento.

A literatura descreve muitos métodos de mineração de dados para aprendizado supervisionado ou não supervisionado. Este trabalho apresenta, de modo sucinto, os algoritmos de mineração de dados comumente encontrados nos métodos apresentados por Camilo e Silva (2009) e Maimon e Lior (2010): Redes Neurais, Árvores de Decisão, Máquinas de Vetores de Suporte, *Random Forest, Naives Bayes, K-Nearest Neighbor, Análise* de Regressão e Regressão Logística. Nela, os métodos são classificados de acordo com as tarefas que realizam.

Nesta pesquisa tem-se a tarefa de classificação através das técnicas de Árvores de Decisão e Random Forest (Floresta Aleatória).

As seguir são descritas sucintamente as principais características destas técnicas de mineração de dados, anteriormente citadas.

2.4.1.1 Árvores de Decisão

Para Camilo e Silva (2009), Árvores de Decisão (*Decision Trees*) é um método de classificação que funciona como um fluxograma em forma de árvore, onde cada nó (não folha) indica um teste feito sobre um valor (por exemplo, idade > 20):

> As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Após a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore (mediante os testes nos nós não-folhas) começando no nó raiz até chegar a uma folha. Pela estrutura que formam, as árvores de decisões podem ser convertidas em Regras de Classificação. O sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade. Apesar de ser uma técnica extremamente poderosa, é necessária uma análise detalhada dos dados que serão usados para garantir bons resultados (CAMILO; SILVA, 2009, p. 5).

Portanto, essa é uma técnica simples, segundo a literatura, pois não se exige parâmetros de configuração, apresentando um nível satisfatório de assertividade. Contudo, conforme se extrai da citação, uma avaliação detalhada dos dados a serem usados é o que garante resultados positivos para essa técnica de mineração de dados. Para Harrison (1998), uma das suas principais vantagens é o fato de que o modelo é bem explicável, uma vez que tem a forma de regras explicitas.

A árvore da decisão, segundo Rabelo (2007, p. 29), é uma "técnica que utiliza a recursividade para particionamento da base de dados na construção de uma árvore de decisão. Cada nó não terminal desta árvore representa um teste ou decisão sobre o item de dado".

Em geral, a técnica de árvore de decisão é apropriada às tarefas de classificação e regressão, e alguns dos algoritmos de árvore de decisão são: CART, CHAID, C5.0, ID-3 (CHEN; HAN; YU, 1996), dentre outros (DIAS, 2001).

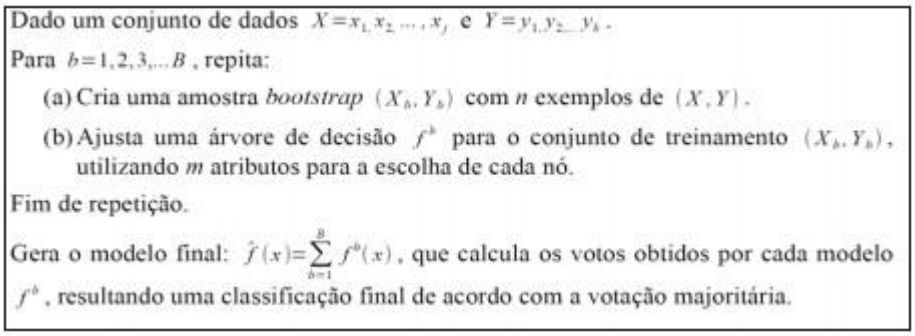
2.4.1.2 Random Forest (Florestas Aleatórias)

Segundo Rodrigues et. al (2012 apud DI GIROLAMO NETO; RODRIGUES; MEIRA, 2014), *Random forest* (florestas aleatórias) é uma técnica de que vem crescendo nos últimos anos, sobretudo nos estudos relacionados ao uso da terra. Para Caruana, Karampatziakis e Yessenalina (2008), é uma das técnicas mais precisas que existem em mineração de dados quando comparadas a outras técnicas também utilizadas, como redes

neurais artificiais ou SVM. É uma técnica computacionalmente muito efetiva e, ainda, conforme Breiman (2001), evita sub reajustes e se mostra pouco sensível a ruídos. Dessa forma, as florestas aleatórias vêm se mostrando muito importante no sentido de gerar informações úteis como suporte para mapear áreas agrícolas por meio de métodos computacionais.

Trata-se de uma técnica de classificação e regressão desenvolvida por Breiman (2001), que consiste num conjunto de árvores de decisão combinada para solucionar problemas de classificação. Cada árvore de decisão é construída utilizando uma amostra aleatória inicial dos dados e, a cada divisão desses dados, um subconjunto aleatório de m atributos é utilizado para a escolha dos atributos mais informativos. No final, a técnica de Random Forest gera uma lista dos atributos mais importantes no desenvolvimento da floresta, que são determinados pela importância acumulada do atributo nas divisões dos nós de cada árvore da floresta (JAMES; HASTIE; TIBSHIRANI, 2013). Os principais passos do algoritmo Random Forest podem ser vistos na Figura 11.

Figura 11 – Algoritmo básico da técnica Random Forest.



Fonte: Breiman (2001).

2.4.2 Ferramentas de mineração de dados

Cruz (2007) tem demonstrado em seus trabalhos a ampla gama de ferramentas de mineração de dados existentes. Optou-se, então, por citar aquelas mais utilizadas, conforme mostrado no Quadro 3.

Quadro 3 – Principais fermentas de mineração de dados.

Ferramenta	Versão	Licença	Disponibilidade	Uso	Arquitetura
Alyuda Neuro Inteligence	F	C	S	C	S
BrainMaker	F	C	N	A/C	S
BSVM	F	F	S	A	S
Clementine	F	C	N	C	S/C S
DTREG	F	C	S	A/C	S
EQUBITS Foresight (tm)	F	C	S	A/C	S
EWA Systems	F	C	N	A/C	S/C S
GhostMiner	F	C	N	A/C	S
Gist	F	F	S	A	S
Gornik	F	C	N	C	S/C S
Insightful Miner	F	C	S	A/C	S/C S
Kernel Machines	F	F	S	A	S
Knowledge Miner	F	C	S	A/C	S
KXEN	F	C	N	C	S/C S
LIBSVM	F	F	S	A	S
MATLAB NN Toolbox	F	C	S	A	S
MCubiX from Diagnos	F	C	N	C	S
MemBrain	F	F	S	A	S
NeuralWorks Predict	F	C	S	C	S
NeuroSolutions	F	C	S	A/C	S/C S
NeuroXL	F	C	N	C	S
IPNNL Software	B	F	S	A	S
Oracle Data Mining	F	C	S	C	S,CS,PP
Orange	F	F	S	A	S
PcSVM	B	P	S	A	S
R	F	P	S	A	S
SAS Enterprise Miner	F	C	S	A/C	CS
StarProbe	F	C	S	A/C	S/C S
STATISTICA NN	F	C	S	A	S/C S
SvmFu 3	B	P	S	A	S
SVM-light	F	F	S	A	S
TANAGRA	F	F	S	A	S
HhinkAnalitics	F	C	N	C	CS
Tiberius	F	C	S	A/C	S/C S
Weka	F	P	S	A	S
XLMiner	F	C	S	A/C	S

Fonte: Cruz (2007, p.45).

O autor classificou as ferramentas de mineração de dados em razão de alguns critérios, conforme se pode observar no Quadro 3 acima (CRUZ, 2007, p. 45):

a) versão - final (F) ou beta (B);

b) licença - comercial (C), freeware e shareware (F) ou pública (P);

c) disponibilidade – se é ou não disponibilizada uma versão de demonstração (Demo) ou a ferramenta é totalmente operacional para download (Download);

d) aplicação de uso - acadêmica (A) ou comercial (C);

e) a arquitetura - Stand alone (S), Cliente/Servidor (C/S) ou Processamento Paralelo (PP).

Dessa forma, em razão da grande disponibilidade de ferramentas de mineração de dados disponíveis (o autor identificou 159, reunindo apenas as mais importantes no Quadro 3), é necessário que o usuário faça uma seleção a fim de otimizar seu trabalho em relação aos interesses pretendidos com a ferramenta escolhida. Segundo Cruz (2007), para a escolha das ferramentas selecionadas no Quadro 3 foi levado em conta: a aplicabilidade da tarefa de Descoberta de Regras de Associação e a utilização da técnica de Associação, aliada ao fato da ferramenta ser de licença Livre.

3 METODOLOGIA: APLICAÇÃO DAS TÉCNICAS DE MINERAÇÃO DE DADOS EM EMPRESA DO SETOR AGRÍCOLA

A presente pesquisa classifica-se quanto a abordagem como sendo uma pesquisa quali-quantitativa, objetivando análises de profundidade e as inferências tem como referência a própria teoria, além do uso de recursos e de técnicas de mineração de dados. De acordo com Turrioni e Melo (2012), este tipo de abordagem também pode ser denominado pesquisa combinada, pois o pesquisador pode combinar aspectos das pesquisas.

De natureza aplicada, que de acordo com Turrioni e Melo (2012), se fortalece por seu interesse na prática, com aplicação dos resultados na solução de problemas reais, além de que estes resultados podem despertar o interesse comercial através do desenvolvimento de novas soluções aplicadas à processos ou produtos demandados pelo mercado.

De caráter exploratório, pois tem como objetivo proporcionar maior aproximação com o problema por meio de estabelecimento de critério, método e técnicas para elaboração da pesquisa, e visa oferecer informações sobre o problema com vistas a torná-lo explícito ou a construir hipóteses (GIL, 2002).

De acordo com Holanda e Riccio (2010), o cenário atual exige que os estudos relacionados a temas de ambiente organizacional sejam conduzidos de forma sistêmica, uma vez que as empresas se encontram inseridas num cenário de mudanças. A emergência da informação e do conhecimento são aspectos fundamentais e determinantes de nova sociedade, o que leva a um novo entendimento das organizações e suas dinâmicas de funcionamento.

Segundo Audy, Andrade e Cidral (2005, p. 82):

> O objetivo ao explorar diferentes maneiras de pensar a organização é poder desenvolver um novo enfoque, transformador da realidade atual. O administrador que pretende trabalhar dentro dessa nova visão terá que admitir que as organizações são "complexas, ambíguas e paradoxais" e sua tarefa será aprender a gerenciar a complexidade e a incerteza. Para ele, as soluções simplistas e apressadas podem causar grande prejuízo à sobrevivência organizacional e é preciso ler, compreender e identificar o significado de cada problema para então agir de forma eficaz.

Corroborando com a necessidade de alinhar as expectativas organizacionais às mudanças conjunturais se percebe que os problemas estão muitas vezes correlacionados à aplicação do conhecimento, muitas vezes oriundos de pesquisas básicas. No entanto, a aplicação do conhecimento não se consistiu como uma sequência lógica da pesquisa básica, faz-se necessário uma ciência própria, a ciência ação (HOLANDA; RICCIO, 2010).

Ainda segundo Ackoff, (1999 apud HOLANDA; RICCIO, 2010), o raciocínio analítico predispõe a ações de separar e tratar as coisas de maneira distintas. No entanto, uma análise simplista não permite a compreensão de um sistema, cujas essências de suas propriedades não são compartilhadas por suas partes.

Tais percepções e entendimento direcionam a utilização do enfoque sistêmico no desenvolvimento de soluções de sistemas de informação. Holanda e Riccio (2010) afirmam que diversos teóricos concordam com a natureza de colaboração da pesquisa-ação, mesmo que o processo de validação da participação do indivíduo seja duvidoso, muitos reconhecem na pesquisa-ação uma possibilidade de melhorar a ação social. Desta maneira, a busca pela solução de um problema prático através da geração de conhecimento se torna a principal estratégia da pesquisa-ação (TURRIONI; MELLO, 2012).

Para Tripp (2005), a ação do indivíduo isolado tem reflexo em toda a organização, o que corrobora com a ideia de que a pesquisa-ação funciona melhor com o envolvimento e comprometimento de diversos indivíduos no processo.

McKay e Marshall (2001) destacam que a essência da pesquisa-ação está no próprio nome, uma justaposição de pesquisa + ação, ou seja, prática + teoria. Assim como uma abordagem de pesquisa comprometida com a produção de conhecimento por meio da busca de soluções de problemas ou melhorias em situações práticas. Esse conhecimento baseado em situações práticas é importante para incorporação nos conteúdos acadêmicos de diversas disciplinas, a pesquisa-ação deveria produzir uma transição da teoria para prática e também da prática para a teoria (TRIPP, 2005).

Ainda Tripp (2005) destaca que a participação do indivíduo não é o único fator determinante para caracterização do tipo da pesquisa-ação que será executado. De acordo com Turrioni e Mello (2012) esta caracterização depende dos seus objetivos e do contexto no qual a pesquisa será aplicada.

Desta maneira é possível identificar algumas modalidades diferentes da pesquisa-ação. Uma pesquisa-ação prática, com abordagem técnica com ação do pesquisador na implantação de uma prática já existente em sua esfera de atuação. Neste caso o pesquisador define ou projeta as mudanças que serão implantadas para suportar as melhorias desejadas. E por último a pesquisa-ação política, cuja ação está no processo de alteração da cultura institucional e/ou de suas limitações (TRIPP, 2005).

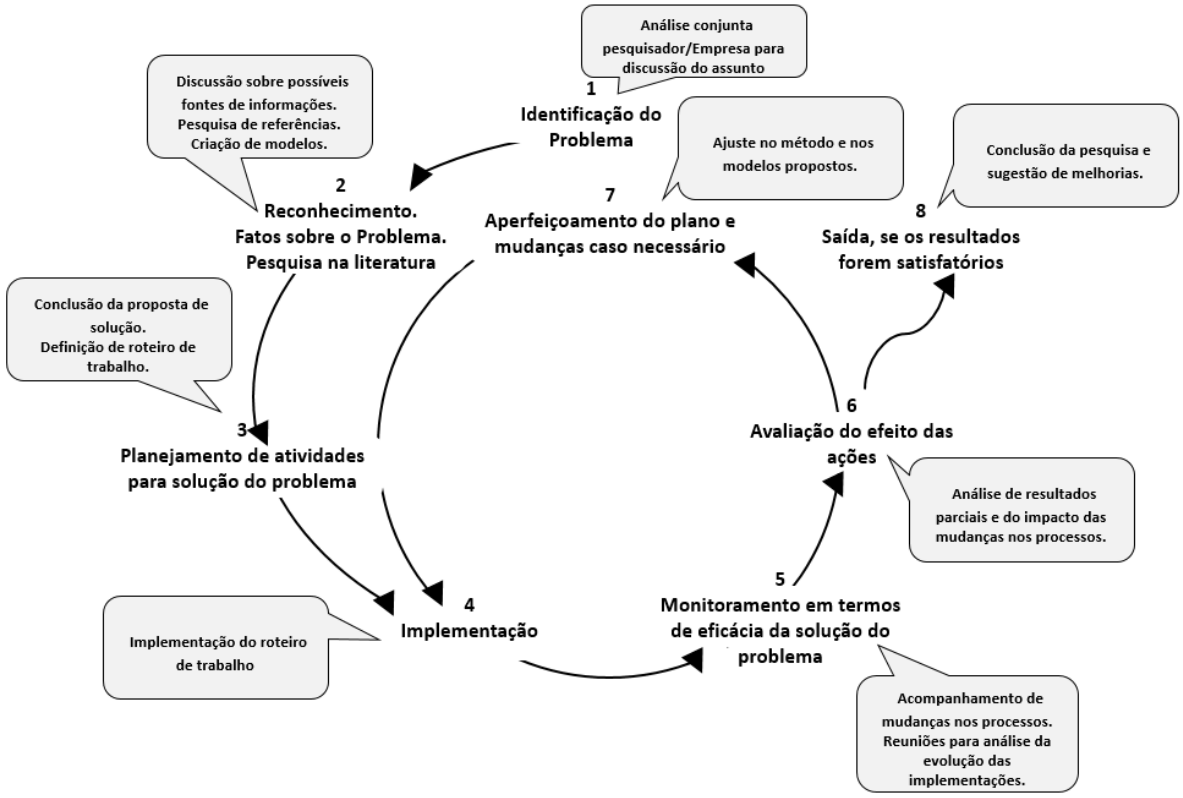
Assim, a pesquisa-ação é apropriada para esta pesquisa, visto que a questão da pesquisa é descrever ao longo do projeto o desdobramento das ações do grupo e da empresa envolvidos nos processos de mudanças ou melhorias. Ao utilizar a observação participante, o

pesquisador pode interferir no objetivo de estudo, cooperando com os demais participantes da ação com o intuito de solucionar um problema e contribuindo para a geração de conhecimento TURRIONI; MELLO, 2012, p. 154).

3.1 Aplicação do Método de Pesquisa-ação

O desenvolvimento desta pesquisa é baseado no roteiro apresentado por McKay e Marshall (2001), ilustrado na Figura 12, e adaptado por Costa et al. (2013), e que está em consonância com o modelo apresentado por Turroni e Mello (2012), apresentando com clareza o planejamento da pesquisa-ação, com detalhamento das fases, etapas e atividades da estrutura que suporta pesquisa-ação.

Figura 12 – Roteiro de Trabalho.



Fonte: Mckay e Marshall (2001, apud COSTA et al., 2013).

De acordo com Costa et al. (2013), o diagrama possibilita a ilustração estática do ciclo da pesquisa-ação.

Cada ciclo do processo de pesquisa-ação acontece em cinco fases: planejar; coletar dados; analisar dados e planejar ações; implementar ações; avaliar resultados e gerar relatórios.

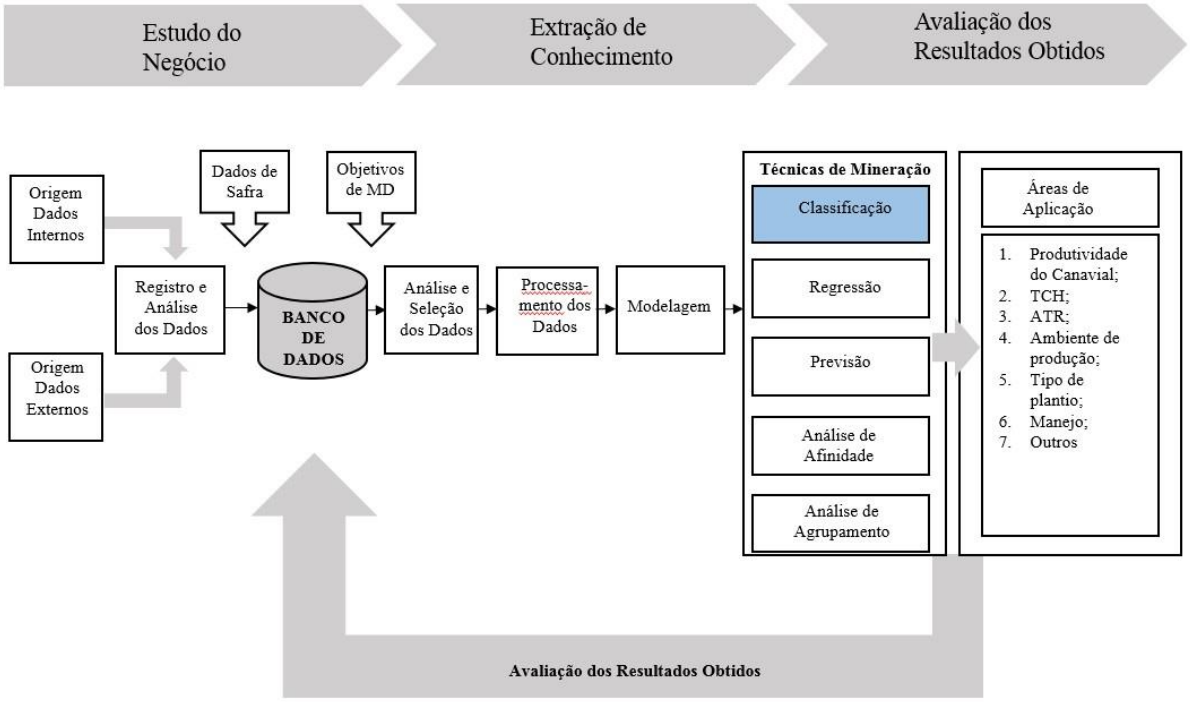
Conforme demonstrado no roteiro ilustrado na Figura 12, de início, foi realizada uma pesquisa bibliográfica dos assuntos abordados nas seções anteriores, com o objetivo de adquirir embasamento teórico no que se refere à importância da cultura da cana-de-açúcar e aos fatores que impactam na sua produtividade, bem como conhecer o processo da descoberta de conhecimento em banco de dados e o processo de *Data Mining* – mineração de dados (processo, modelos, técnicas utilizadas, entre outros).

Além disso, buscou-se conhecer os problemas enfrentados pela administração, como também especialistas da área de planejamento agronômico para identificação dos fatores que impactam na produtividade da cana-de-açúcar, suas prioridades e dificuldades no acesso ao conhecimento. O número de variedades cultivadas e volume de dados armazenados proporcionam diversas análises e podem indicar quais variedades apresentam melhor desempenho na região estudada e quais são as melhores práticas de manejo com objetivo de alcançar maiores índices produtivos. É importante, para o processo de tomada de decisões nas operações agrícolas, conhecer estas particularidades e, assim, realizar com maior precisão as recomendações agronômicas, o que demanda ferramentas de apoio que auxiliem neste processo.

Com base nas seções de trabalho com a equipe da organização responsável pela gestão das operações agrícolas para obtenção de experiências sobre os fatores (variáveis) de influência na produtividade agrícola, foram apontados diversos fatores (variáveis) com diferentes graus de impactos e facilidade de manejo (atuação). Logo, com as variáveis selecionadas, iniciou-se a análise de quais variáveis estavam contempladas no sistema de informação utilizado pela empresa.

Após as etapas acima, foi utilizada a metodologia CRISP-DM para aplicação das técnicas de mineração de dados, a qual, segundo Maimon e Lior (2010), é dividida em seis etapas principais: Estudo do Negócio; Estudo de Dados; Preparação dos Dados; Modelagem; Avaliação e Implementação. Mattozo (2007) afirma que, ao integrar a sistematização desses componentes numa perspectiva organizacional, torna-se possível o desenvolvimento de projetos. A metodologia sugerida contempla 3 etapas: Estudo do Negócio, Extração de Conhecimento e Avaliação dos Resultados Obtidos, conforme Figura 13 a seguir:

Figura 13 - Etapas da metodologia CRISP-DM.



Fonte: Adaptado de Matozzo (2007, p.135).

Na Figura 13, nota-se que, de início, os dados são obtidos a partir de diferentes fontes. Após o registro da análise do dado, é criado um Banco de Dados, que tem como objetivo dar suporte à próxima etapa do processo, correspondente à extração de conhecimentos. Por último, a avaliação dos resultados alcançados é concretizada na etapa de aplicação à atividade de produção de cana-de-açúcar.

De acordo com Mattozo (2007), a metodologia apresenta três questões fundamentais para o desenvolvimento do trabalho:

- O levantamento de informações sobre a empresa para criação de um modelo com a finalidade de definir as necessidades, os fluxos de informações e dados necessários para o atendimento dos objetivos de negócio;
- A definição do banco de dados, cuja escolha depende do tamanho e da quantidade de variáveis presentes. As ferramentas de mineração de dados podem ser aplicadas diretamente em bancos de dados relacionais ou também em um *data warehouse* ou *data mart* que consigam lidar com dados armazenados, complicados e complexos.
- A escolha das técnicas e ferramentas de mineração de dados depende dos tipos de dados disponíveis e de qual o tipo de correlação a ser pesquisado entre as variáveis do problema para deduções sobre as relações existentes.

Descreve-se, a seguir, em detalhes, a metodologia proposta com base na revisão teórica realizada na seção anterior que proporcionou a identificação dos requisitos necessários para desenvolvimento e alcance dos objetivos propostos nesta pesquisa. Assim, esta seção apresenta a aplicação da metodologia e encontra-se estruturada com as seguintes etapas, conforme proposto por Mattozo (2007) e Maimon e Lior (2010):

A) Estudo do Negócio

A1) Descrição da empresa, processos operacionais e gestão do negócio;

A2) Obtenção e Exploração dos Dados:

Avalição dos Dados Internos Disponíveis;

Importação dos Dados e Constituição do Banco de Dados Inicial;

Filtragem e Limpeza dos Dados;

A3) Constituição da Base de Dados com dados de safras e fatores que impactam na produtividade do canavial:

Seleção das Informações Elegíveis.

B) Extração de Conhecimentos

B1) Análise e Seleção dos Dados;

B2) Pré-processamento e Transformação dos Dados;

B3) Modelagem;

B4) Avaliação dos Resultados.

C) Avaliação dos Resultados Obtidos

3.2 Descrição da área experimental e práticas agrícolas utilizadas

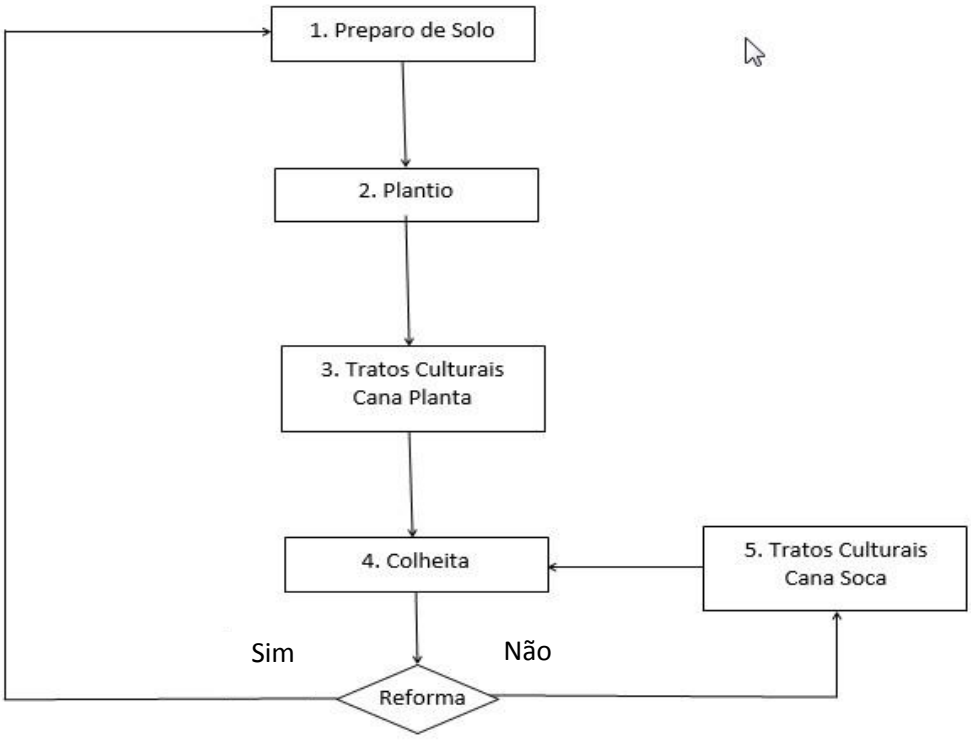
O caso da pesquisa aqui descrita foi desenvolvido em empresa agrícola produtora de cana-de-açúcar com experiência de 72 anos no setor, responsável pela administração de 4,6 milhões de toneladas de cana-de-açúcar cultivadas em 51 mil hectares, nas seguintes coordenadas: 22g 26m latitude e 50g 20m longitude, com temperatura média de 23,1°C, precipitação média anual de 1430,4 mm e altitude média de 440 metros.

Principais Números:

1. Volume de canavial administrado: 4,6 milhões de toneladas
2. Volume de área explorada: 51 mil hectares
3. Número de funcionários: 1.260
4. Participação na produção de cana nos municípios de atuação: 36%

No aspecto operacional, as principais atividades realizadas pela empresa no cultivo da cana-de-açúcar, conforme demonstrado na Figura 14, são:

Figura 14 - Fluxograma do processo produtivo da cana-de-açúcar.



Fonte: Elaborado pelo próprio autor (2016).

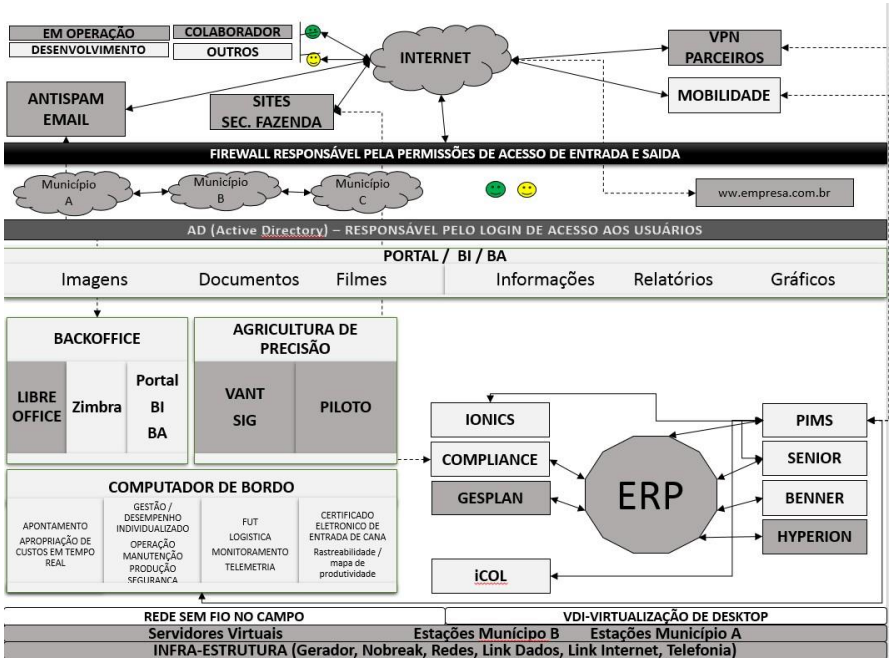
As etapas podem ser descritas como:

1. Preparo do solo: levantamento topográfico, calagem, descompactação e outras.
2. Plantio: Mecanizado, semimecanizado e manual.
3. Tratos Culturais da Cana Planta: combate de plantas daninhas, pragas, doenças, adubação, visando proporcionar a melhor condição para desenvolvimento da planta e maior produtividade.
4. Colheita: corte, transbordamento e transporte.
5. Tratos Culturais Cana Soca: adubação, combate de plantas daninhas, pragas, doenças, adubação, visando proporcionar a melhor condição para desenvolvimento da planta e maior produtividade.

Com relação à tecnologia, a empresa conta com um moderno sistema de informações – *Enterprise Resource Planning* (ERP) Oracle – integrado ao sistema de controle de produção – *Process Information Management Systems* (PIMS) – em ambiente *web* para operação de seus controles operacionais, contabilidade fiscal-tributária, gestão de custo e orçamento, suprimento de matérias e insumos, planejamento financeiro e agrícola de suas atividades, processo que gera segurança, qualidade e velocidade às informações gerenciais, contribuindo

para o processo decisório e gestão de risco. As informações oriundas de equipamentos do processo para softwares são dispostas em nível informação, para que as informações geradas atinjam níveis altos dentro da indústria, cuja arquitetura da informação pode ser definida, conforme a Figura 15, que representa o mapa tecnológico da empresa.

Figura 15 - Mapa tecnológico.



Fonte: Elaborado pelo autor (2016).

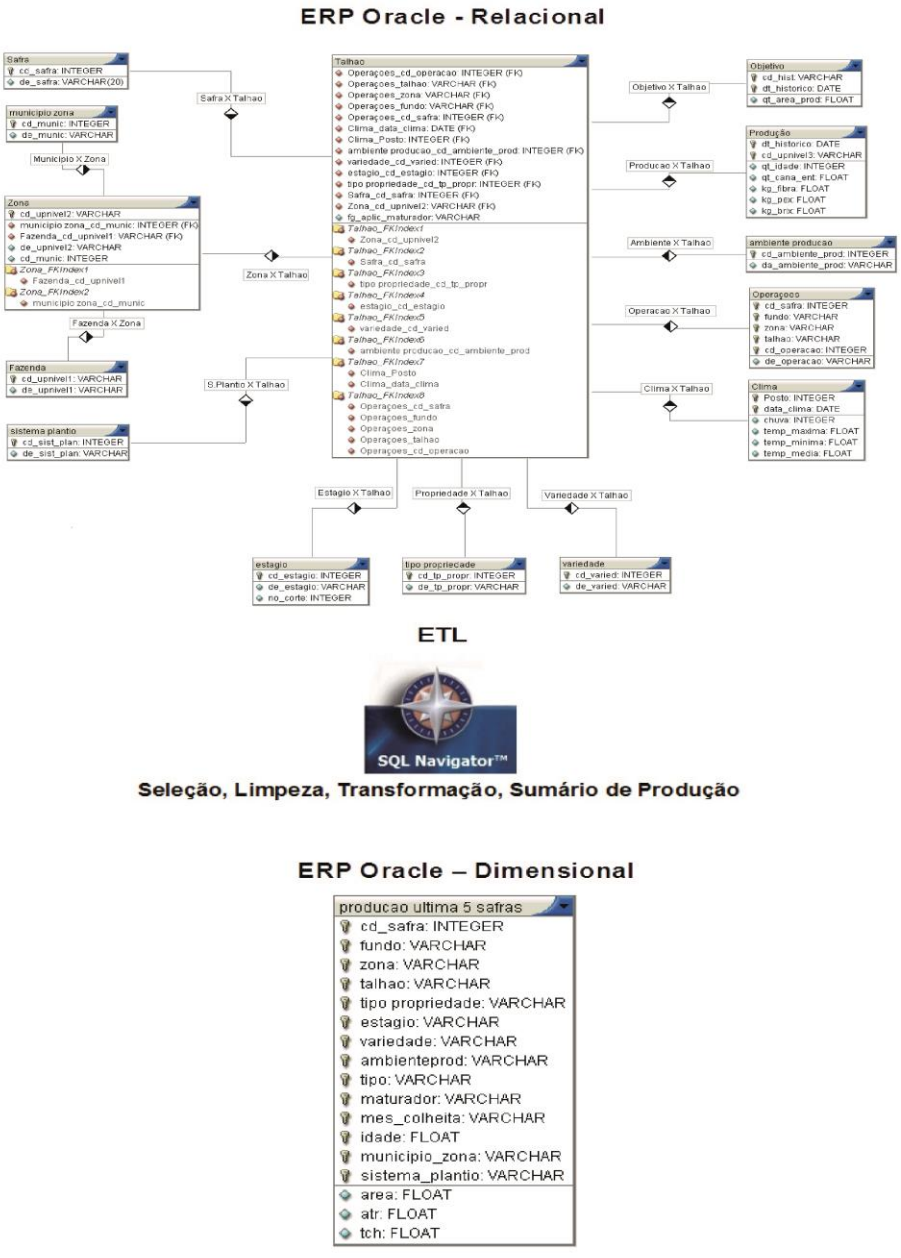
O suporte tecnológico ao processo de descoberta de conhecimento em banco de dados é fundamental para expansão e utilização da técnica de mineração de dados. O mapa de tecnologia adotado pela empresa propicia o monitoramento do uso da terra e o desenvolvimento e a disseminação de sistemas automatizados de suporte à decisão. E neste contexto, se destaca a mineração de dados com a solução para extração de associações e conhecimentos a partir de grandes volumes de dados gerados através da modelagem e simulação para geração de cenários e fenômenos com objetivo de otimizar modelos que levem à solução ótima de um problema (MAIMON; LIORM, 2010; MASSRUHÁ; LEITE, 2016).

3.3 Obtenção e exploração dos dados

Para o presente trabalho foram utilizados dados primários obtidos do sistema interno da empresa e externo da estação meteorológica na sede. Para constituição do banco de dados inicial dos parâmetros de fatores operacionais e climatológicos que impactam na produtividade do canavial, foi desenvolvida uma rotina em *Structured Query Language*

(SQL) para extração de dados, que tem como objetivo sumarizar as informações de uma base relacional *Oracle* dentro de um ERP Agrícola, conforme a seleção de dados, para um determinado agrupamento com as informações demonstradas na Figura 16 e formação de um banco de dados dimensional:

Figura 16 - Banco de dados Oracle dimensional.



Fonte: Elaborado pelo autor (2016).

A obtenção de dados para comprovar a aplicabilidade da metodologia, mediante identificação do problema e solução, baseando-se na metodologia CRISP-DM como ferramenta para identificação e classificação dos fatores que impactam na produtividade da cana-de-açúcar, foi alcançada através de dados históricos armazenados na base de dados da empresa, sendo as principais fontes dos dados os sistemas legados utilizados pela empresa, armazenados em Banco de Dados *Oracle*.

Nota-se que os dados empregados nesta pesquisa são extraídos do sistema PIMS, que oferece suporte a todas as etapas do processo produtivo, do preparo do solo à colheita de cana-de-açúcar, do controle do processo na indústria ao gerenciamento de ativos (TOTVS, 2016). Esse sistema alimenta uma base de dados *Oracle* onde foi aplicada a mineração de dados.

Na sequência das etapas descritas acima, foi aplicado um extrator para acesso aos dados necessários para aplicação das ferramentas. Após a definição da estrutura de aquisição de dados do processo de produção de cana-de-açúcar, com base no fluxo do processo, foram desenvolvidos os módulos para a continuidade do processo de obtenção e exploração do banco de dados e geração de conhecimento.

3.4 Variáveis avaliadas

As variáveis classificatórias ou independentes avaliadas foram:

- Safra: identifica o ano da safra em que o processo de colheita foi realizado. Neste caso, o período analisado corresponde aos meses entre abril e março, totalizando 12 meses, sendo a operação realizada entre o período da colheita e da entressafra. Os dados utilizados neste estudo são referentes às últimas 5 safras: 2011/2012, 2012/2013, 2013/2014, 2014/2015 e 2015/2016.
- Fundo: identifica o nome da propriedade onde o canavial encontra-se cultivado.
- Zona: tem por finalidade identificar as divisões de áreas para cultivo dentro de um fundo, ou seja, na propriedade.
- Talhão: representa uma subdivisão dentro da zona para atividades de manejo operacional.
- Tipo de propriedade: identifica se o canavial está implantado em área própria da empresa, arrendada, parceria ou fornecedor parceiro, conforme tipo de contratação realizada pela empresa.
- Município (Zona): representa o município onde o canavial está implantado.

- Estágio: o estágio de corte é representado por um número que determina a idade do canavial e o número de cortes, sendo classificados em: primeiro corte 12M, primeiro corte 18M, primeiro corte inverno, segundo corte, terceiro corte, quarto corte, quinto corte, sexto corte e sétimo corte. Vale ressaltar que a cultura procedente da cana planta, por via de regra, vem seguida de duas a quatro culturas de rebrota (cana soca), podendo em certos casos ultrapassar os oito ciclos (MARCHIORI, 2004).
- Variedade: identifica o tipo de variedade analisada e diz respeito a cultivar da cana-de-açúcar. O banco de dados contempla 51 variedades. Nesta pesquisa, foram consideradas as principais variedades cultivadas pela empresa: CTC 4 (ciclo médio) e RB 966928 (ciclo precoce), que correspondem a 46% dos hectares cultivados pela empresa. A CTC 4 é caracterizada como Média/tardia, indicada para cultivo em ambientes A a C, com boa aceitação na região de Goiás e na região Oeste paulista. Variedade com excelente brotação de soqueiras, inclusive em cana crua. Período de colheita entre os meses de junho e outubro. Já a variedade RB 966928 é caracterizada pela maturação precoce e média. Foi desenvolvida pela Universidade Federal do Paraná (UFPR), com boa aceitação na região Sul e Sudeste do Brasil. Apresenta elevado teor de sacarose. Indicada para cultivo em ambientes de médio a alto potencial. Variedade com excelente germinação em cana-planta, brotação em soqueiras muito boa. Colheita recomendada para o período de abril a maio (RIDESA, 2010).
- Ambiente: o atributo ambiente de produção contém o código referente à classificação do tipo do solo de acordo com a classificação brasileira de tipos de solo. Traz informação do solo em vários níveis:

  a) O primeiro nível diz respeito à classe do solo, de acordo com a morfologia (latossolo, argissolo, etc);

  b) O segundo nível considera as cores no horizonte B (horizontes são camadas mais ou menos paralelas à superfície do terreno, diferenciadas pela cor, textura e estrutura);

  c) O terceiro nível considera as condições químicas do horizonte subsuperficial (eutrófico, distrófico, etc). Detalhes dessa tipificação podem ser encontrados em Prado et al. (2008).

A base de dados estudada tem 5 tipos de ambientes de produção representada por cinco diferentes códigos: A – Alta; B – Média Alta; C – Média; D – Média baixa e E – Baixa. A textura refere-se à proporção de argila, silte e areia do solo; para isso, são utilizados os seguintes códigos: 1 – solo argiloso; 2 – solo arenoso e 3- solo argiloso/arenoso.

- Tipo: neste campo é representado o tipo de canavial colhido, classificado em cana planta, ou seja, o canavial que está no primeiro corte, ou cana soca, o canavial que será colhido a partir do segundo corte até a sua exaustão; quanto maior a idade do canavial, menor a sua produtividade em TCH.
- Maturador: informa se houve aplicação de produto que acelera o clico de maturação da cana, com o objetivo de concentrar mais ATR. O processo de maturação da cana-de-açúcar pode ser definido como o processo fisiológico que envolve a formação de açúcares nas folhas e seu deslocamento e armazenamento no colmo.
- Mês da Colheita: corresponde ao atributo que identifica o mês em que a operação de colheita foi realizada.
- Idade (meses): identifica os meses entre o plantio e a realização do primeiro corte, ou a idade entre cortes, caso o canavial seja cana soca. Tal procedimento de manejo pode permitir reais ganhos de produtividade da lavoura canavieira.
- Trimestre safra: indica o trimestre em que a cana foi colhida, por exemplo, trimestre (T1): abril, maio e junho; trimestre (T2): julho, agosto e setembro e trimestre (T3): outubro, novembro e dezembro.
- Área: consta a quantidade de área cultivada por zona representada por hectares cultivados.
- Operação: neste campo é identificado o tipo de manejo para tratos culturais referentes à aplicação de fertilizantes no período pós-colheita, sendo estes tratos representados por aplicação de resíduos do processo industrial, como a vinhaça ou aplicação de fertilizantes minerais.
- Tipo de plantio: apresenta o tipo de operação de plantio realizado, classificado em: plantio manual convencional, manual com torta, semimecanizado e mecanizado.

As variáveis agronômicas avaliadas foram:

- TCH: O atributo TCH – Tonelada de Cana por Hectare – representa o volume de cana produzida em um hectare cultivado, sendo representado em tonelada.

- ATR: O atributo ATR- Açúcar Total Recuperado – informa a quantidade de açúcar produzido em kg de açúcar por hectare; indicador que representa a quantidade total de açúcares da cana (sacarose, glicose e frutose).
- TCH*ATR: um índice representado pelo produto da multiplicação entre TCH e ATR.

As variáveis meteorológicas avaliadas foram:

- Precipitação acumulada até o corte: identifica o volume de chuva nos respectivos meses de safra, representado em milímetros.
- Temperatura média: em graus Celsius, avaliada durante o período experimental.
- Evapotranspiração média: em milímetros, avaliada durante o período experimental.
- Fotoperíodo: horas de luz média durante o período do levantamento dos dados.

3.5 Processamento de dados e análises estatísticas (extração de conhecimento)

3.5.1 Limpeza e composição do banco de dados para análises

Na fase de composição do banco de dados, para a realização das análises estatísticas, foram considerados os seguintes critérios:

1. Dados de 5 anos de produção de cana-de-açúcar de uma operadora agrícola (safras de 2010 a 2016); dados das características dos locais de produção de uma operadora agrícola (safras de 2011 a 2016).
2. Duas variedades foram consideradas para a composição do banco de dados, a CTC 4 (ciclo médio) e RB 966928 (ciclo precoce), que correspondem a 46% dos hectares cultivados pela empresa.
3. Produtividade de cana por hectare (TCH) válida (acima de 30t/ha e abaixo de 250t/ha) e valores de ATR (kg/t) válidos entre 80 e 200.
4. Dados próximos a sede da estação meteorológica para a composição da planilha do balanço hídrico e fonte de dados externos relacionados aos fatores climáticos: precipitação evapotranspiração, luminosidade e temperatura média.

3.5.2 Análises estatísticas

3.5.2.1 Árvore de decisão

A árvore de decisão é um método não-paramétrico supervisionado de aprendizagem utilizado para classificação e regressão. Uma árvore é cultivada por um particionamento recursivo binário utilizando a variável resposta na fórmula especificada e escolhendo

partições. A árvore de decisão funciona de forma hierárquica a fim de descrever a partição de *x* possíveis observações em sub-regiões correspondentes às folhas. Uma árvore de decisão é desenhada da esquerda para a direita ou começando a partir da raiz. O primeiro nó é uma raiz e os nós subsequentes são as folhas (RIPLEY, 2008).

A principal diferença entre um método paramétrico para um não-paramétrico é que o método paramétrico é composto de duas etapas, 1ª etapa: é feita uma pressuposição do modelo sobre a forma funcional de f (ex: em uma regressão múltipla $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$,assume-se que f é linear de X), 2ª etapa: Após o modelo selecionado, é utilizado um procedimento que usa os dados do treinamento para *ajuste* ou *treinar* o modelo. Nesse caso, do modelo linear, é necessário estimar os parâmetros $\beta_0, \beta_1, \ldots, \beta_p$. Encontrando valores que: $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$. A abordagem mais comum nesse caso é a dos quadrados mínimos ordinários. No caso de um método não paramétrico, não se procura explicitamente pressuposições sobre a forma funciona de f. Ao invés disso, o método procura uma estimativa de f que chega o mais perto possível dos pontos dos dados sem ser muito exigente (JAMES; HASTIE; TIBSHIRANI, 2013).

A árvore de decisão funciona da seguinte forma: primeiro é encontrado um ponto de particionamento de uma simples variável. Os dados são separados e, então, esse processo é aplicado separadamente para cada subgrupo, o processo é feito recursivamente até que os subgrupos alcancem um tamanho mínimo ou até que não haja melhorias no particionamento (THERNEAU; ATKINSON; FOUNDATION, 2015).

O critério de particionamento da árvore de decisão, supondo um nó "A", funciona da seguinte forma, se particionarmos um nó A em dois filhos, A_L e A_R (esquerda e direita):

$$P(A_L)r(A_L) + P(A_R)r(A_R) < P(A)r(A) \qquad (1)$$

Usando a fórmula 1, uma forma possível de construir uma árvore é escolher uma partição que maximiza Δr e diminui o risco.

Relacionando o ajuste do modelo da regressão linear com a árvore de decisão em um variável resposta do tipo contínua, em modelos lineares uma forma de verificar o ajuste é através do R^2 e da erro padrão residual (EPR) que são dados por:

$$R^2 = \frac{SQT - SQR}{SQT} = 1 - \frac{SQR}{SQT} \ , \qquad (2)$$

Em que SQT é a soma de quadrados totais definida como:

$$SQT = \sum (y_i - \bar{y})^2 \ , \qquad (3)$$

Portanto, o R^2 mensura a proporção da variabilidade de Y que pode ser explicada em X (variável independente).

Além disso, o erro padrão residual:

$$\text{(EPR)} = \sqrt{\frac{1}{n-2}SQR} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \quad (4)$$

Em que SQR é a soma dos quadrados residuais, que é dada por:

$$\text{SQR: } \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = e_1^2 + e_2^2 + \cdots + e_n^2, \quad (5)$$

Em que $e_i = y_i - \hat{y}_i$ representa o i-ésimo resíduo. Isto é, a i-ésima resposta observada e o i-ésimo valor que é predito pelo modelo linear.

Nos métodos de mineração de dados, para variáveis classificatórias como variável resposta, como em árvore de decisão, uma forma alternativa ao erro padrão residual (EPR) é a mensuração da taxa de erro de classificação (E), que é definida como uma fração das observações do banco de dados de treinamento que não pertencem a classe mais comum para a classe em questão:

$$E = 1 - \max_k(\hat{p}_{mk}) \quad (6)$$

Aqui $\hat{p}_{mk}$ representa a proporção das observações de treinamento na m-ésima região que são da classe k-ésima.

Quando variável resposta é contínua, a árvore de regressão e uma regressão linear continuam sendo diferentes em sua essência, enquanto que uma regressão linear assume o modelo:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \quad (7)$$

As árvores de regressão assumem o modelo da forma:

$$f(X) = \sum_{m=1}^{M} c_m . 1(X \in R_m) \quad (8)$$

Onde $R_1,...,R_M$ representam a partição da variável. A escolha de usar regressão linear ou árvore de regressão é em função do problema em específico, ou seja, se a relação entre as variáveis independentes com a variável dependente não é puramente linear, a ávore de regressão pode ser uma boa alternativa. Quando se trata de visualização de modelos, os modelos de árvores também são indicados (JAMES; HASTIE; TIBSHIRANI, 2013).

No presente trabalho, a árvore de decisão foi utilizada para definir a partir de que ponto variáveis de ambiente de produção, clima e manejo influenciam altas ou baixas magnitudes de ATR e TCH.

3.5.2.2 Floresta aleatória

Em árvores de decisão padrões, cada nó é dividido usando a melhor separação entre todas as variáveis. Em uma floresta aleatória, cada nó é dividido usando o melhor entre um subconjunto de indicadores escolhidos aleatoriamente naquele nó. Esta estratégia é um pouco contraintuitiva e funciona bem quando comparada a muitos outros classificadores, incluindo a análise discriminante, suporte de vetor de máquina e redes neurais, além de ser uma técnica robusta contra o *overfitting* (BREIMAN, 2001). Os modelos de floresta aleatória incluem muitos modelos de árvores de classificação (padrões) e não apenas um. Assim, a floresta aleatória tende a ter uma maior resistência às mudanças de dados e ao ruído (que são variáveis de pouca influência sobre a variável objetivo).

No processo de construção das árvores, cada vez que uma partição é considerada, uma amostra aleatória de *m* preditores é escolhida como candidata para particionar o conjunto completo de *p* preditores. Uma amostra dos *m* preditores é selecionada em cada partição e, tipicamente, é preferível escolher $m \approx \sqrt{p}$, que é o número de preditores considerado em cada partição. Geralmente, usa-se a raiz quadrada do número total de preditores. De forma mais sucinta, a floresta aleatória não utiliza todas as variáveis em cada partição, contornando também a presença de colinearidade entre as variáveis (JAMES; HASTIE; TIBSHIRANI, 2013).

A estatística gerada denominada de "*Out-of-bag*" (OOB) é relacionada com os dados que não entraram na amostragem bootstrapping. Na análise de floresta aleatória de classificação, o erro OOB é a taxa de observações classificadas como erradas. A acurácia é denominada como $Acurácia(\%) = 100 - OOB_{erro}$.

O método de floresta aleatória é baseado numa técnica denominada de *bagging* (*Bootstrap aggregation*) que é um procedimento de propósito geral com o objetivo de reduzir a variância do método de aprendizagem estatística.

Supondo um conjunto de dados com n observações independentes $Z_1,...,Z_n$, cada uma com variância σ^2, portanto, a variância média é dada por σ^2/n. Em outras palavras, fazendo a média de observações reduz a variância. Uma forma natural de reduzir a variância e aumentando a predição da acurácia do modelo é pegar vários conjuntos de dados da

população, construir um modelo de predição separado usando cada conjunto de treinamento e fazendo a média das predições resultantes. Assim, calculando:

$$\hat{f}^1(x), \hat{f}^2(x), \ldots, \hat{f}^B(x) \quad (9)$$

Usando B conjunto de dados de treinamento e fazendo a média deles para obter um modelo de aprendizagem estatístico de baixa variância simples que é dados por:

$$\hat{f}_{avg(x)} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^b(x) \quad (10)$$

Como na prática não se tem múltiplos conjuntos de dados, é feito o bootstrap, pegando amostras repetidas de um (simples) conjunto de dados de treinamento. O método de treinamento é feito no b-ésimo conjunto de dados de treinamento para obter $\hat{f}^{*b}(x)$, e finalmente fazer a média de todas as predições, para chegar em na bagging que pode ser descrita como:

$$\hat{f}_{bag} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \quad (11)$$

Considerando uma variável resposta contínua, aplicando bagging num modelo de árvore de regressão, simplesmente é construída B árvores de regressão usando B treinamento via bootstrapping, e feita a média das predições resultantes. Essas árvores são crescidas profundamente e não são podadas. Portanto, cada árvore individualmente tem alta variância mas baixo viés (JAMES; HASTIE; TIBSHIRANI, 2013).

Considerando uma variável composta de classes, no banco de dados teste é gravado a classe predita por cada árvore B, e no final a classe mais votada é a de ocorrência mais provável entre as B árvores preditoras.

Sobre a importância das variáveis e ajuste do modelo de floresta aleatória, em um estudo com 100 variáveis das quais 6 eram preditores relevantes, as variáveis principais foram selecionadas 50 % das vezes em cada partição. Portanto, não é necessário se preocupar com seleção de variáveis em floresta aleatória (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

No presente trabalho, a floresta aleatória foi utilizada para definir a partir de que ponto variáveis de ambiente de produção, clima e manejo influenciam altas ou baixas magnitudes de ATR e TCH.

3.5.2.3 Teste de Tukey

Os testes de comparações múltiplas, ou testes de comparações de médias, servem como um complemento do teste F, para detectar diferenças entre os tratamentos. Utilizado para testar todo e qualquer contraste entre duas médias, neste contexto, o teste de Tukey é

uma ferramenta recomendada para comparar dois ou mais grupos. O Teste proposto por Tukey é também conhecido como teste de Tukey da diferença honestamente significativa (*honestly significant difference*) (HSD) e teste de Tukey da diferença totalmente significativa (*wholly significant difference) (*WSD).

O teste de Tukey pode ser descrito como:

$$\Delta = q\frac{\sqrt{QMR}}{\sqrt{J}} \tag{12}$$

Onde:

q : amplitude total estudentizada, é função (I, graus de liberdade do resíduo da análise de variância e α).

QMR: é o desvio padrão residual do ensaio, ou seja, a raiz quadrada do quadrado médio do resíduo da análise de variância.

J: é o número de repetições das médias confrontadas no contraste (FERREIRA, 2000).

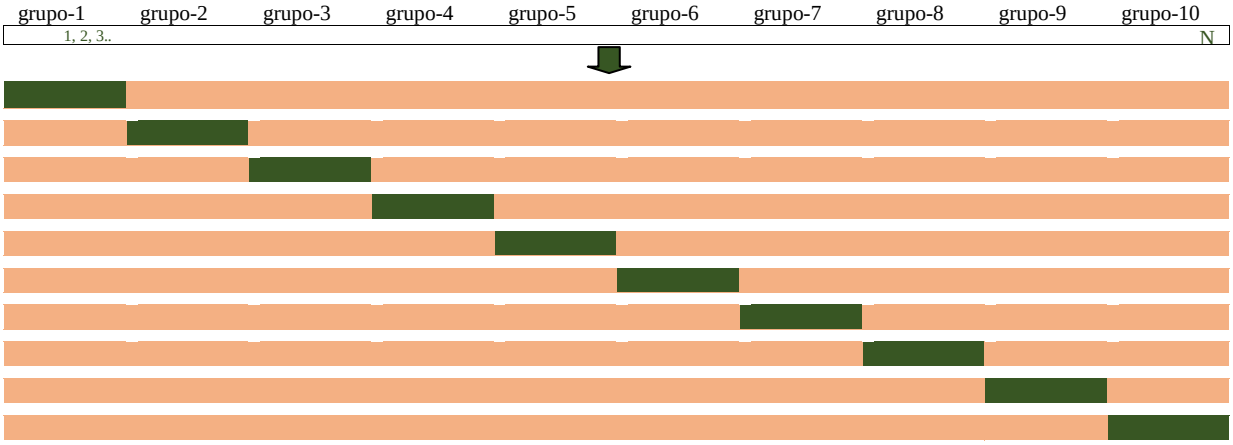
3.5.2.4 Validação cruzada

Na ausência de um banco de dados muito grande designado para o teste que pode ser usado para estimar diretamente a taxa de erro, várias técnicas podem ser utilizadas para estimar esse valor usando o banco de dados de treinamento disponível, dentre elas a validação cruzada.

No presente trabalho foi utilizado o método de validação cruzada k-grupos[1], mais especificamente o 10-fold. Neste método, o banco de dados em questão é repartido aleatoriamente em 10 partes iguais, sem sobreposição. Nove partes do banco de dados são utilizadas para treinamento (cor laranja) e $1/10$ do banco de dados é utilizado para validação - cor verde (Figura 17) (JAMES; HASTIE; TIBSHIRANI, 2013). Contudo, também é possível destinar 15% do banco de dados principal exclusivamente para validação, como no presente trabalho.

[1] Método k-fold cross-validation (CV).

Figura 17 – Esquema do funcionamento do método 10-fold cross-validation no banco de dados.



Fonte: Elaborado pelo autor (2017).

O primeiro grupo é tratado como validação, e o método é ajustado considerando o k-1 grupos (ou folds). Para variáveis respostas contínuas, o erro quadrático médio (EQM) é então computado, EQM_1. Esse procedimento é repetido k vezes, cada vez um grupo diferente é tratado como validação. O processo resulta em k estimativas do erro, no presente caso, EQM_1, EQM_2, ..., EQM_{10}. As estimativas do erro do k-grupos via validação cruzada (VC) são computadas fazendo-se a média dos valores computados:

$$VC_{(K)} = \frac{1}{k}\sum_{i=1}^{k} EQM_i \qquad (13)$$

Na prática adotando-se k=5 ou k=10 tem relação com vantagens computacionais, contudo, estudo demonstram que k ≥ 10 produzem melhores estimativas para o banco de dados em questão (WILLIAMS, 2011). Para variáveis respostas categóricas na mineração de dados, ao invés de quantificar o erro com o EQM, é utilizado o número de observações classificadas erroneamente (Err):

$$VC_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \qquad (14)$$

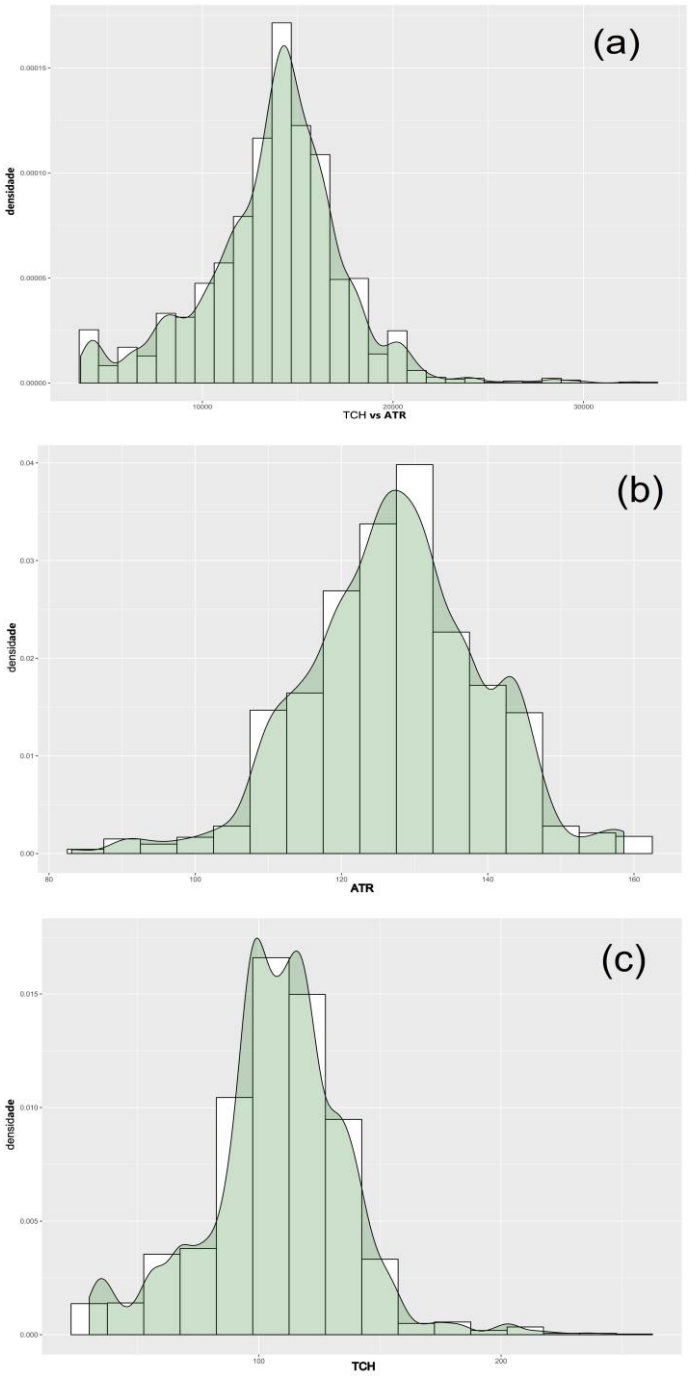
Onde, $Err_i = I(y_i \neq \hat{y}_i)$. A taxa de erro do 10-fold CV e o erro do banco de dados de validação são definidos de forma análoga. Todas as análises foram realizadas no software e ambiente R, IDE *RStudio* (R CORE TEAM, 2016) e R *Data Mining* (*ratlle*) (WILLIAMS, 2011). As análises de árvores de decisão foram realizadas no pacote *rpart* (THERNEAU; ATKINSON; RIPLEY, 2015). As análises de floresta aleatória foram realizadas no pacote *random Forest* (LIAW; WIENER, 2002). O teste de Tukey foi feito no pacote *agricolae* do R (DE MENDIBURU, 2014). Os scripts R principais do presente trabalho, estatística descritiva, árvores de decisão e florestas aleatórias, pode ser encontrado no Apêndice B.

4 RESULTADOS E DISCUSSÃO

4.1 Transformação das variáveis TCH, ATR e TCH*ATR em classes

Para verificar a distribuição dos dados organizados em classe para as variáveis avaliadas, foram plotados histogramas juntamente com a densidade (Figura 18).

Figura 18 – Histograma com densidade sobreposta para as variáveis TCH*ATR (a), TCH (b) e ATR (c).



Fonte: Elaborado pelo autor (2017).

Além dos histogramas, para verificar a normalidade e/ou necessidade de transformação de dados, os valores da estatística W de Shapiro-Wilk foram de TCH= 0,96938, ATR= 0,99117 e TCH*ATR= 0,96348, todos com p-valores <0,001, indicando desvio de normalidade. Portanto, foi realizado o teste da melhor potência de Box-Cox (1964), que produziu valores de lambda (λ) próximo da unidade para as três variáveis avaliadas, indicando a não necessidade da transformação de dados. Uma análise detalhada da estatística descritiva, boxplots e gráficos de dispersão das variáveis TCH e ATR com relação as variáveis independentes, pode ser encontrada no Apêndice A.

Para realizar as análises da árvore de decisão (especificamente árvore de classificação) e floresta aleatória, as variáveis contínuas TCH, ATR e TCH*ATR foram transformadas em classes de acordo com os quartis da distribuição normal. As classes definidas foram: "baixa" (25% das observações para baixo), "baixa-média" (50% das observações para baixo), "média-alta" (75% das observações para baixo) e "alta" (100% das observações para baixo).

Tabela 1 - Número de observações em cada classe de quartil, média, desvio padrão, mínimo, máximo e coeficiente de assimetria para as variáveis TCH, ATR e TCH*ATR.

Classe dos quartis	N*	Média	Desvio padrão	Mínimo	Máximo	Assimetria
TCH						
Baixa	537	71,44	19,32	30,02	94,24	0,0064
Baixa-Média	536	101,16	4,15	94,26	108,70	
Média-Alta	536	116,37	4,18	108,71	124,39	
Alta	536	142,15	18,83	124,41	247,90	
ATR						
Baixa	576	112,05	7,14	83,09	119,71	-0,2257
Baixa-Média	564	123,93	2,24	119,72	127,54	
Média-Alta	575	130,97	2,17	127,55	135,52	
Alta	560	142,45	5,54	135,55	158,61	
TCH*ATR						
Baixa	537	8851,73	2393,87	3611,84	11817,84	-0,0497
Baixa-Média	536	13157,99	686,22	11827,50	14201,31	
Média-Alta	536	14870,47	503,22	14201,57	15801,22	
Alta	536	18001,51	2473,92	15807,46	32940,83	

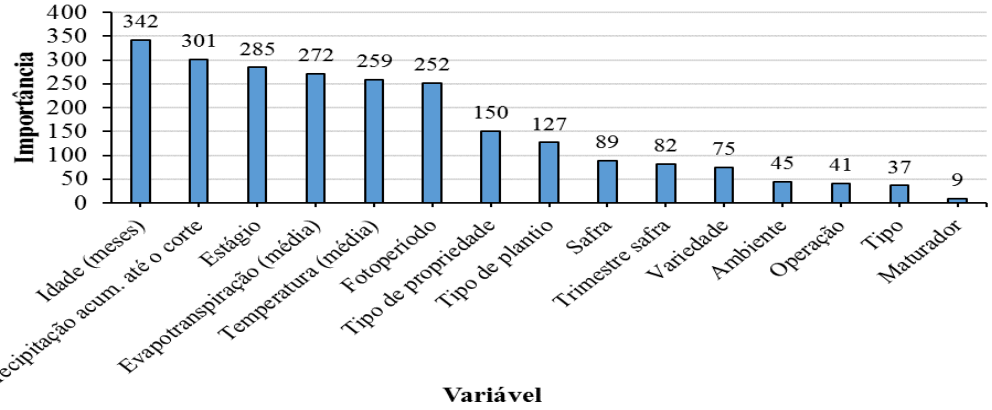
N: Número de observações em cada classe.
Fonte: Elaborado pelo autor (2016).

As variáveis TCH, ATR e TCH*ATR apresentaram número de observações similares em cada classe. Isso se deve ao coeficiente de assimetria que apresentaram valores de baixa magnitude. A assimetria refere-se à forma da curva de uma distribuição de frequência, mais especificamente do polígono de frequência ou do histograma. Denomina-se assimetria o grau de afastamento de uma distribuição da unidade de simetria (CORREA, 2003). Considerando a distribuição normal, a variável TCH pode ser considerada como simétrica (assimetria ≈ zero), ao passo que ATR e TCH*ATR são consideradas como assimétricas à esquerda e negativas. Para as três variáveis, os desvios padrão foram maiores para as classes "baixa" e "alta" que são compostas de valores extremos para a variável avaliada (Tabela 1).

4.2 Árvores de decisão (considerando a variável "safra")

A importância das variáveis que mais discriminam a variável resposta em cada partição na árvore de decisão (particionamento recursivo) é mensurada a partir da influência atribuída a cada variável na redução da função de perda[2]. Esse método não fornece qual classe tem maior peso quando a variável independente é um fator, ou seja, composta de vários níveis (KUHN, 2016). As cinco variáveis independentes que apresentaram maior importância para classificar as classes de TCH (baixa, baixa-média, média-alta e alta) foram: idade (meses), precipitação acumulada até o corte, estágio, evapotranspiração média e temperatura média. Portanto, três variáveis climáticas, dentre as cinco principais, influenciam os valores de TCH. As variáveis que menos influenciam TCH são: maturador, tipo, operação, ambiente e variedade (Figura 19).

Figura 19 - Importância das variáveis da árvore de decisão para a variável resposta TCH (considerando a variável "safra").



Fonte: Elaborado pelo autor (2016).

[2] **Função de perda**: Erro quadrático médio (EQM) em variáveis respostas contínuas e número de observações classificadas erroneamente (Err) em dados discretos.

As variáveis com menores efeitos na árvore de decisão também são importantes para definir os valores de TCH. Contudo, seus efeitos ficam suprimidos pelos efeitos de variáveis meteorológicas, que apresentam resultado mais expressivo na variável resposta. Além disso, o efeito do maturador no valor do TCH não foi expressivo, embora conclusões não devam ser tomadas sobre esse aspecto, porque na análise da árvore de decisão não está sendo mensurada se o maturador tem efeito de precocidade na cana, e sim o seu efeito sobre o valor final da TCH.

No banco de dados utilizado para a árvore de decisão, a estatística mais importante é a acurácia de predição dos dados de validação cruzada, que para a variável TCH apresentou magnitude de 72,82. Portanto, 72,82% das classes de TCH foram classificadas corretamente quando utilizadas as variáveis independentes (Tabela 2). A acurácia é a medida mais conhecida e a mais usada para medir o desempenho de classificação. Para um classificador dentro de uma matriz de confusão, a acurácia é definida com o número de itens categorizados corretamente, dividido pelo número total de itens. O ideal é que a acurácia alcance valores próximos de 100%, ou seja, que o erro de classificação que é *100-acurácia*, seja minimizado (ZUMEL; MOUNT, 2014).

Na validação cruzada, o conjunto total de dados foi dividido em 10 subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto foi utilizado para teste e os 10-1 restantes foram utilizados para estimação dos parâmetros e calculada a acurácia do modelo. Este processo foi realizado por 10 vezes alternando de forma circular o subconjunto de teste. Uma vez que a taxa de classificação é definida como 100 – acurácia, logo, o valor do erro de classificação para a validação foi de 27,18% (Tabela 3).

Na árvore de decisão, considerando a TCH, a acurácia para os dados de treinamento, no geral, é maior que os dos dados da validação, uma vez que no treinamento o banco de dados utilizado é maior (Tabela 2).

Tabela 2 - Estatísticas de classificação da árvore de decisão para a variável resposta TCH (considerando a variável "safra").

Descrição	Acurácia de predição (%)	Taxa do erro de classificação (%)
Predição para os dados do treinamento [(1)]	79,21	20,79
Predição para os dados da validação cruzada [(2)]	72,82	27,18

[(1)]: Erro do nó da raiz * erro relativo * 100%, [(2)]: Erro do nó da raiz * erro da validação cruzada (10-fold CV) * 100%.
Fonte: Elaborado pelo autor (2016).

A matriz de confusão exibe a distribuição das observações nos termos de suas classes reais e suas classes preditas. Na análise da árvore de decisão, para a classe TCH "baixa", 453 das observações foram classificadas corretamente; contudo, essa classe apresentou classificação errada de 16,93%, porque 57 + 27 + 31 = 116 observações não foram classificadas com acerto. Além disso, a classe TCH "baixa" apresenta acerto de 100 – 16,93% = 83,07% das observações (Tabela 4). A classe TCH "baixa-média" apresentou 389 das observações classificadas corretamente e erro de 16,98%. A classe TCH "média-alta" apresentou 412 das observações classificadas corretamente e erro de 18,69%. A classe TCH "alta" apresentou 445 das observações classificadas corretamente e erro de 16,24%, uma vez que 18 + 25 + 64 = 107 observações não foram classificadas corretamente. Portanto, para a classe TCH "alta", 83,76% dos indivíduos foram classificados corretamente (Tabela 3).

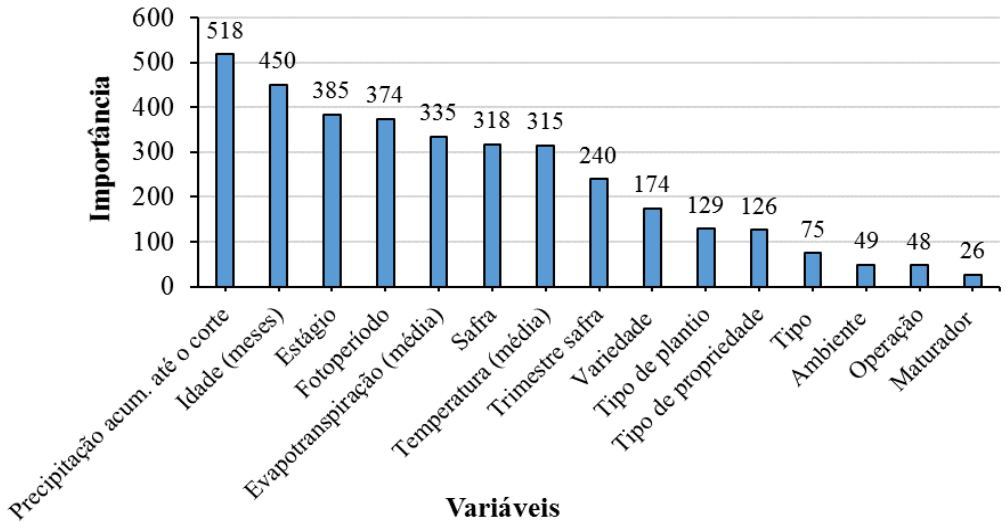
Tabela 3 - Matrix de confusão para os dados do treinamento da árvore de decisão para a variável resposta TCH (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	453	57	27	32	0,1693
Baixa-Média	42	389	33	25	0,1698
Média-Alta	24	65	412	34	0,1869
Alta	18	25	64	445	0,1624

Tamanho da amostra, N = 2145.
Fonte: Elaborado pelo autor (2016).

Na análise da árvore de decisão, as cinco variáveis independentes que apresentaram maior importância para classificar as classes de ATR ("baixa", "baixa-média", "média-alta" e "alta") foram: precipitação acumulada até o corte, idade (meses), estágio, fotoperíodo e evapotranspiração média. Portanto, três variáveis climáticas, dentre as cinco principais, influenciam os valores de ATR (Figura 20). As variáveis que menos influenciam ATR são: maturador, operação, ambiente, tipo e tipo de propriedade (Figura 20).

Figura 20 - Importância das variáveis da árvore de decisão para variável resposta ATR (considerando a variável "safra").



Fonte: Elaborado pelo autor (2016).

Considerando ATR ("baixa", "baixa-média", "média-alta" e "alta") como variável resposta na árvore de decisão, 89,93% dos dados são classificados corretamente no banco de dados do treinamento, ao passo que 83,74% dos dados de ATR são classificados de forma correta utilizando a validação cruzada com 10 subconjuntos de dados (Tabela 4). Na prática, os valores de treinamento e validação são apenas um indicativo de modelagem, uma vez que os dados do modelo final devem ser avaliados com o banco de dados do teste (WILLIAMS, 2011).

Tabela 4 - Estatísticas de classificação para a árvore de decisão da variável resposta ATR (considerando a variável "safra").

Descrição	Acurácia de predição (%)	Taxa do erro de classificação (%)
Predição para os dados do treinamento [(1)]	89,93	10,07
Predição para os dados da validação cruzada [(2)]	83,74	16,26

[(1)]: Erro do nó da raiz * erro relativo * 100%, [(2)]: Erro do nó da raiz * erro da validação cruzada (10-fold CV) * 100%.
Fonte: Elaborado pelo autor (2016).

Para a classe ATR "baixa", 515 das observações foram classificadas corretamente na análise da árvore de decisão. Todavia, essa classe apresentou classificação errada de 6,13%, porque 20 + 5 + 11 = 36 observações não foram classificadas corretamente. A classe ATR

"baixa" apresenta acerto de 100 – 6,13% = 93,87% das observações (Tabela 6). A classe ATR "baixa-média" apresentou 509 das observações classificadas corretamente e erro de 11,03%. A classe ATR "média-alta" indicou 489 das observações classificadas corretamente e erro de 10,50%. A classe ATR "alta" apresentou 528 das observações classificadas corretamente e erro de 9,38%, uma vez que 7 + 5 + 49 = 61 observações não foram classificadas com acerto. Portanto, para a classe ATR "alta", 90,62% dos indivíduos foram classificados corretamente (Tabela 5).

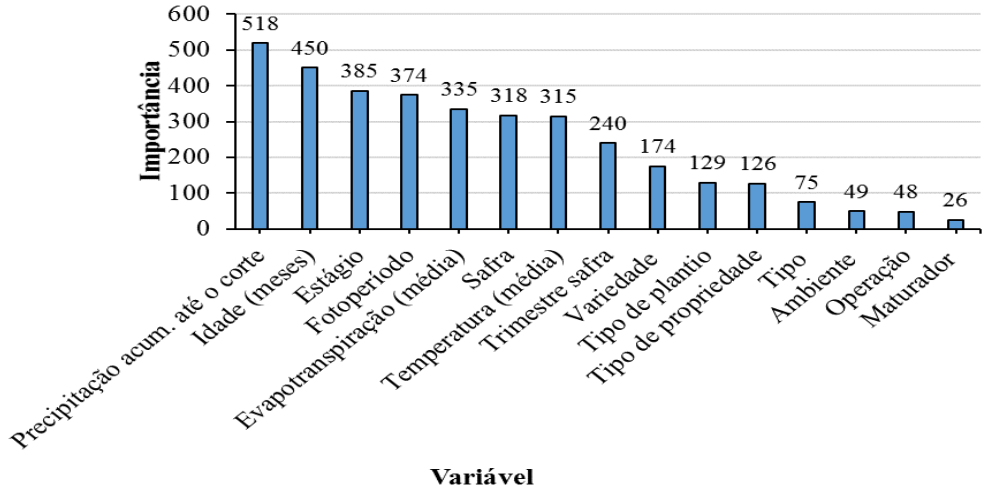
Tabela 5 - Matrix de confusão para os dados do treinamento da árvore de decisão para a variável resposta ATR (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	515	20	5	11	0,0613
Baixa-Média	37	509	32	3	0,1103
Média-Alta	17	30	489	18	0,1050
Alta	7	5	49	528	0,0938

Tamanho da amostra, N = 2275
Fonte: Elaborado pelo autor (2016).

Na análise da árvore de decisão, as cinco variáveis independentes que apresentaram maior importância para classificar as classes de TCH*ATR (baixa, baixa-média, média-alta e alta) foram: precipitação acumulada até o corte, idade (meses), estágio, fotoperíodo e evapotranspiração média. Portanto, três variáveis climáticas, dentre as cinco principais, influenciam os valores de TCH*ATR (Figura 21). As variáveis que menos influenciam ATR são: ambiente, maturador, operação, tipo e tipo de propriedade (Figura 21).

Figura 21 - Importância das variáveis da árvore de decisão para a variável resposta TCH*ATR (considerando a variável "safra").



Fonte: Elaborado pelo autor (2016).

Considerando TCH*ATR (baixa, baixa-média, média-alta e alta) como variável resposta na análise da árvore de decisão, 80,79% dos dados são classificados corretamente no banco de dados do treinamento, ao passo que 74,45% dos dados de ATR são classificados de forma correta utilizando validação cruzada com 10 subconjuntos de dados (Tabela 6).

Tabela 6 - Estatísticas de classificação da árvore de decisão para a variável resposta TCH*ATR (considerando a variável "safra").

Descrição	Acurácia de predição (%)	Taxa do erro de classificação (%)
Predição para os dados do treinamento [(1)]	80,79	19,21
Predição para os dados da validação cruzada [(2)]	74,45	25,55

[(1)]: Erro do nó da raiz * erro relativo * 100%, [(2)]: Erro do nó da raiz * erro da validação cruzada (10-fold CV) * 100%.
Fonte: Elaborado pelo autor (2016).

Na análise da árvore de decisão, para a classe TCH*ATR "baixa", 455 das observações foram classificadas corretamente. Além disso, essa classe apresentou classificação errada de 13,60%, porque 39 + 20 + 26 = 85 observações não foram classificadas corretamente. Ainda na classe TCH*ATR "baixa", esta mostra acerto de 100 – 13,60% = 86,40% das observações (Tabela 5). A classe TCH*ATR "baixa-média" apresentou 402 das observações classificadas corretamente e erro de 16,05%. A classe TCH*ATR "média-alta" apresentou 439 das observações classificadas corretamente e erro de 20,46%. A classe TCH*ATR "alta" apresentou 436 das observações classificadas corretamente e erro de 13,55%, uma vez que 23 + 24 + 34 = 81 observações não foram classificadas corretamente. Portanto, para a classe TCH*ATR "alta", 86,45% dos indivíduos foram classificados corretamente (Tabela 7).

Tabela 7 - Matrix de confusão para os dados do treinamento da árvore de decisão para a variável resposta TCH*ATR (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	455	39	20	26	0,1360
Baixa-Média	33	402	43	19	0,1605
Média-Alta	26	71	439	55	0,2046
Alta	23	24	34	436	0,1355

Tamanho da amostra, N=2145.
Fonte: Elaborado pelo autor (2016).

4.3 Floresta aleatória (considerando a variável “safra”)

Na análise da floresta aleatória para TCH foram simuladas 500 árvores, três variáveis testadas em cada nó obtendo, assim, taxa da estimativa do erro OOB de 16,80% e acurácia de 83,20% para os dados do treinamento (Tabela 8).

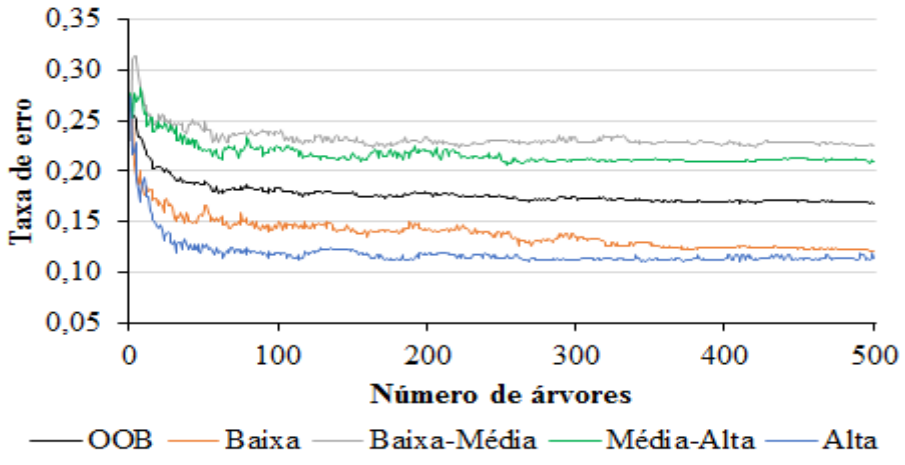
Tabela 8 - Estatísticas de classificação da floresta aleatória para a variável resposta TCH (considerando a variável “safra”).

Descrição	Valor
Número de árvores	500
Número de variáveis testadas em cada partição	3
Acurácia de predição (%)	83,20
Taxa de estimativa do erro OOB (Taxa do erro de classificação (%))	16,80

Fonte: Elaborado pelo autor (2016).

Considerando o número de árvores × taxa do erro, é perceptível que, com o aumento do número de arvores no modelo da floresta aleatória, acontece diminuição do erro OOB e, portanto, ocorre aumento da acurácia. A análise da floresta aleatória apresenta ganho de acurácia de 83,20 – 79,21 ≈ 4% quando comparada à árvore de decisão. Sendo assim, a análise da árvore de decisão apresenta maior acurácia na classificação de TCH do que apenas uma árvore. As classes TCH “baixa” e TCH “alta” foram aquelas de menor taxa no que diz respeito ao erro de classificação, provavelmente, porque são valores extremos dentro da distribuição normal e se distinguem das outras classes (Figura 22).

Figura 22 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta TCH (considerando a variável “safra”).



Fonte: Elaborado pelo autor (2016).

A Tabela 9 mostra as variáveis mais importantes ordenadas pela diminuição na acurácia e no índice Gini, caso a variável seja excluída. Quanto maior os valores da tabela, mais a variável é importante. As cinco variáveis mais importantes para a predição das classes de TCH na análise floresta aleatória foram: estágio, idade (meses), temperatura (média), safra e evapotranspiração (média). As variáveis menos relacionadas com as classes de TCH foram: maturador, variedade, tipo, ambiente e operação (Tabela 9).

O índice Gini usado pelo algoritmo CART (árvore de classificação e árvore de regressão) é uma medida de quantas vezes um elemento escolhido aleatoriamente do conjunto seria classificado de modo incorreto se fosse feito aleatoriamente, de acordo com a distribuição de classificação no subconjunto. Quanto mais a precisão da floresta aleatória diminui, devido à exclusão (ou permutação) de uma única variável, mais importante é considerada a variável. A diminuição média no coeficiente de Gini é uma medida de como cada variável contribui para a homogeneidade dos nós e das folhas na floresta aleatória resultante. Contudo, em um banco de dados com variáveis mistas (numérica e categóricas) o índice Gini pode ser viesado e produzir maiores valores para variáveis com maior número de categorias ou de acordo com escala utilizada (STROBL et al., 2007).

Tabela 9 - Coeficientes padronizados da importância das variáveis dentro de cada classe de TCH, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta TCH (considerando a variável "safra").

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Estágio	60,22	50,75	59,80	58,86	72,32	143,21
Idade (meses)	50,79	44,50	51,74	48,48	67,88	169,72
Temperatura (média)	43,10	37,92	47,64	44,20	60,70	109,19
Safra	48,70	41,59	45,14	48,10	53,90	75,86
Evapotranspiração (média)	42,56	36,72	43,29	45,72	53,52	123,04
Fotoperíodo	39,84	33,89	41,76	39,89	53,28	99,05
Precipitação acum. até o corte	39,19	34,15	43,93	43,04	52,28	113,15
Tipo de plantio	34,82	29,45	37,18	37,24	46,93	52,49
Trimestre safra	31,37	27,19	35,88	35,85	39,45	43,67
Tipo de propriedade	24,67	33,33	31,99	24,86	37,86	42,70
Operação	25,18	21,67	24,69	13,88	35,14	24,93
Ambiente	28,16	19,00	18,92	28,56	34,04	34,41
Tipo	18,82	18,47	20,34	22,08	28,21	18,92
Variedade	21,79	21,28	21,09	25,69	25,18	26,13
Maturador	5,49	-1,37	15,67	8,05	17,16	6,48

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
(2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
Fonte: Elaborado pelo autor (2016).

Já na análise da floresta aleatória referente à classe TCH "baixa", 403 das observações foram classificadas corretamente. No entanto, essa classe apresentou classificação errada de 12,01%, porque 29 + 12 + 13 = 54 observações não foram classificadas de modo correto. Além disso, a classe TCH "baixa" aponta acerto de 100 – 12,01% = 87,99% das observações (Tabela 10). A classe TCH "baixa-média" apresentou 376 das observações classificadas corretamente e erro de 22,47%. A classe TCH "média-alta" apresentou 393 das observações classificadas corretamente e erro de 20,93%. A classe TCH "alta" apresentou 433 das observações classificadas corretamente e erro de 11,45%, uma vez que 3 + 21 + 30 = 54 observações não foram classificadas de maneira correta. Portanto, para a classe TCH "alta", 88,55% dos indivíduos foram classificados com acerto (Tabela 10).

Tabela 10 - Matrix de confusão para os dados do treinamento da floresta aleatória para a variável resposta TCH (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	403	29	12	14	0,1201
Baixa-Média	42	376	48	19	0,2247
Média-Alta	3	42	393	59	0,2093
Alta	5	21	30	433	0,1145

Tamanho da amostra, N = 1929
Fonte: Elaborado pelo autor (2016).

Na análise da floresta aleatória para ATR foram simuladas 500 árvores, três variáveis testadas em cada nó obtendo, dessa forma, taxa da estimativa do erro OOB de 5,17% e acurácia de 98,43% para os dados do treinamento (Tabela 11).

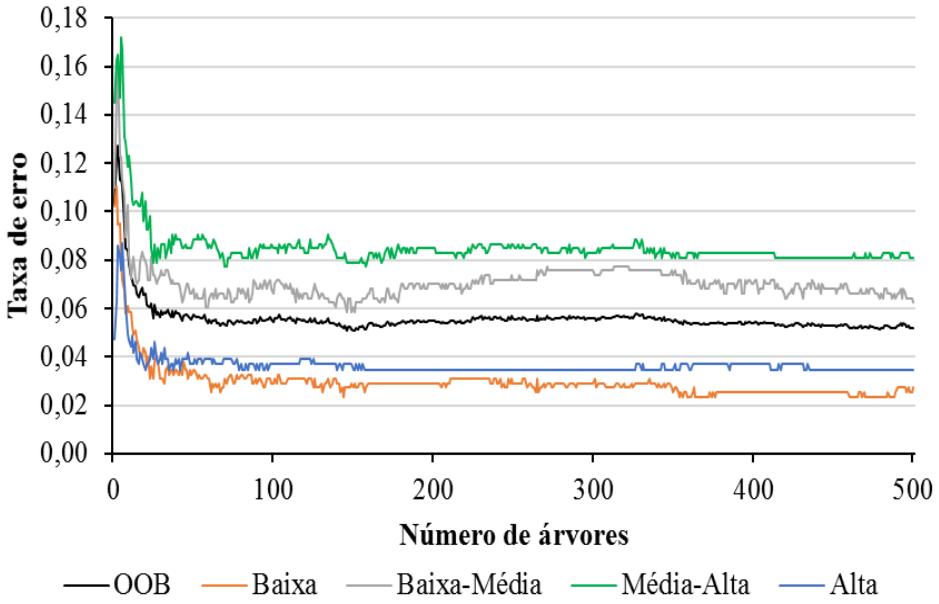
Tabela 11 - Estatísticas de classificação da floresta aleatória para a variável resposta ATR (considerando a variável "safra").

Descrição	Valor
Número de árvores	500
Número de variáveis testadas em cada partição	3
Acurácia de predição (%)	98,43%
Taxa de estimativa do erro OOB (Taxa do erro de classificação (%)	5,17%.

Fonte: Elaborado pelo autor (2016).

Considerando o número de árvores × taxa do erro, pode-se perceber que, com o aumento do número de árvores no modelo da floresta aleatória, ocorre diminuição do erro OOB e, como consequência, há aumento da acurácia. A análise da floresta aleatória mostra um ganho de acurácia de 98,43 – 89,93 ≈ 8,5% quando comparada à árvore de decisão para ATR. Portanto, a análise da árvore de decisão apresenta maior acurácia na classificação de ATR do que apenas uma árvore. As classes ATR "baixa" e ATR "alta" foram as que apresentaram menor taxa de erro de classificação, provavelmente, porque são valores extremos dentro da distribuição normal e se distinguem das outras classes (Figura 23).

Figura 23 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta ATR (considerando a variável "safra").



Fonte: Elaborado pelo autor (2016).

Quanto maior os valores na Tabela 9, mais as variáveis contribuem para discriminar o ATR. As cinco variáveis mais importantes para a predição das classes de ATR, de acordo com a acurácia na análise floresta aleatória, foram: idade (meses), estágio, precipitação acumulada até o corte, temperatura (média) e evapotranspiração (média). As variáveis menos relacionadas com as classes de TCH foram: maturador, variedade, tipo, ambiente e operação (Tabela 12).

Tabela 12 - Coeficientes padronizados da importância das variáveis dentro de cada classe de ATR, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta ATR (considerando a variável "safra").

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Idade (meses)	45,83	55,30	60,08	45,43	68,42	187,00
Estágio	53,86	56,18	52,17	40,68	63,87	139,57
Precipitação acum. até o corte	47,61	48,77	52,18	41,43	59,42	158,43
Temperatura (média)	37,83	53,37	48,01	38,69	57,32	151,68
Evapotranspiração (média)	44,38	48,11	45,88	34,24	56,95	144,67
Trimestre safra	49,27	46,10	50,35	44,77	54,04	111,67
Fotoperíodo	46,62	48,78	45,95	40,31	52,44	155,70
Tipo de propriedade	27,16	29,36	34,50	30,62	40,99	46,41
Safra	37,06	35,47	35,52	31,94	40,88	76,52
Tipo de plantio	37,03	36,69	33,72	23,05	39,55	61,67
Operação	18,65	25,42	28,95	19,97	36,86	27,57
Ambiente	20,64	21,88	25,45	23,82	35,25	27,93
Tipo	21,60	23,82	27,13	20,03	32,01	26,54
Variedade	23,52	26,14	29,45	29,41	29,64	44,52
Maturador	12,04	10,77	12,54	8,97	19,22	7,55

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
(2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
Fonte: Elaborado pelo autor (2016).

Quanto à análise da floresta aleatória para a classe ATR "baixa", 536 das observações foram classificadas de forma correta. Porém, essa classe mostrou classificação errada de 2,72%, porque 12 + 2 + 1 = 15 observações não foram classificadas corretamente. A mesma classe ATR "baixa", ainda, indica acerto de 100 – 2,72% = 97,28% das observações (Tabela 13). A classe ATR "baixa-média" apresentou 496 das observações classificadas corretamente e erro de 6,24%. A classe ATR "média-alta" apontou 477 das observações classificadas corretamente e erro de 8,09%. Já a classe ATR "alta" apresentou 418 das observações classificadas corretamente e erro de 3,46%, pois 2 + 13 = 15 observações não foram classificadas corretamente. Portanto, para a classe ATR "alta", 96,54% dos indivíduos foram classificados de modo correto. Todavia, os dados da Tabela 14 devem ser verificados com cautela, uma vez que o banco de dados do treinamento tende a superestimar os valores da acurácia do modelo, sendo que esses valores decrescem no banco de dados durante o teste (JAMES; HASTIE; TIBSHIRANI, 2013).

Tabela 13 - Matrix de confusão para os dados do treinamento da floresta aleatória para a variável resposta ATR (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	536	12	2	1	0,0272
Baixa-Média	10	496	21	2	0,0624
Média-Alta	2	22	477	18	0,0809
Alta	0	2	13	418	0,0346

Tamanho da amostra, N=2118
Fonte: Elaborado pelo autor (2016).

No que se refere à análise da floresta aleatória para TCH*ATR, foram simuladas 500 árvores, três variáveis testadas em cada nó obtendo, dessa maneira, taxa da estimativa do erro OOB de 16,28% e acurácia de 83,72% para os dados do treinamento (Tabela 14).

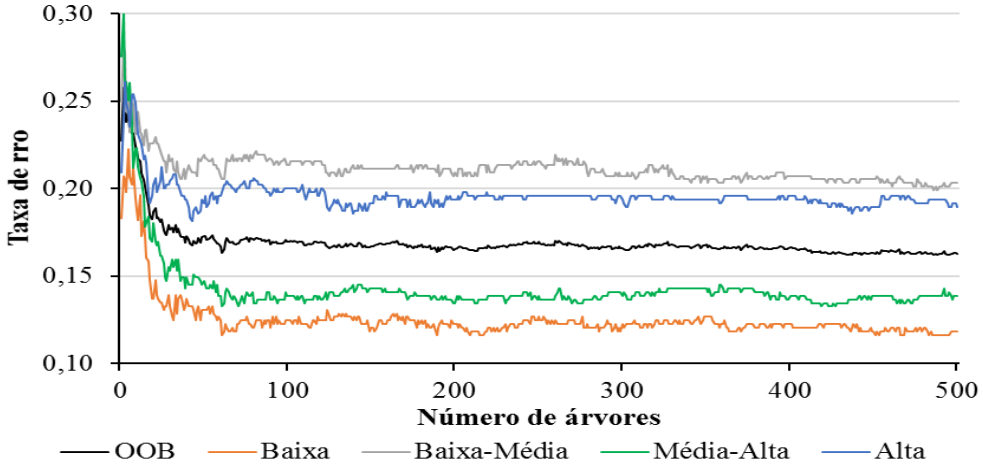
Tabela 14 - Estatísticas de classificação da floresta aleatória para a variável resposta TCH*ATR (considerando a variável "safra").

Descrição	Valor
Número de árvores	500
Número de variáveis testadas em cada partição	3
Acurácia de predição (%)	83,72
Taxa de estimativa do erro OOB (Taxa do erro de classificação (%))	16,28

Fonte: Elaborado pelo autor (2016).

Ao considerar o número de árvores × taxa do erro, observa-se que, com o aumento do número de árvores no modelo da floresta aleatória, acontece diminuição do erro OOB, ocorrendo, como consequência, aumento da acurácia. A análise da floresta aleatória indica ganho de acurácia de 83,72 – 80,79 ≈ 3% quando comparada à árvore de decisão. Portanto, a análise da árvore de decisão aponta maior acurácia na classificação de TCH*ATR do que apenas uma árvore. As classes TCH*ATR "baixa" e TCH*ATR "alta" foram aquelas que apresentaram menor taxa de erro de classificação, provavelmente, porque são valores extremos dentro da distribuição normal e se distinguem das outras classes (Figura 24).

Figura 24 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta TCH*ATR (considerando a variável “safra”).



Fonte: Elaborado pelo autor (2016).

Quanto maior os valores na Tabela 15, mais as variáveis contribuem para discriminar o TCH*ATR. As cinco variáveis mais importantes para a predição das classes de TCH*ATR, considerando a acurácia na análise da floresta aleatória, foram: estágio, idade (meses), evapotranspiração (média), temperatura (média) e precipitação acumulada até o corte. As variáveis menos relacionadas com as classes de TCH*ATR foram: maturador, tipo, variedade, ambiente e operação (Tabela 15).

Tabela 15 - Coeficientes padronizados da importância das variáveis dentro de cada classe de TCH*ATR, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta TCH*ATR (considerando a variável “safra”).

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Estágio	66,53	59,70	51,30	60,91	79,57	132,65
Idade (meses)	55,23	49,78	51,74	52,16	65,87	184,07
Evapotranspiração (média)	44,54	38,45	40,97	40,24	55,32	115,92
Temperatura (média)	39,40	40,28	45,22	44,53	52,63	120,68
Precipitação acum. até o corte	38,12	39,01	39,46	38,65	51,12	104,12
Safra	44,99	41,81	37,42	41,20	48,26	77,04
Fotoperíodo	36,31	33,18	39,69	37,77	47,09	91,98
Tipo de plantio	34,10	33,82	34,43	38,17	45,05	57,07
Trimestre safra	37,54	37,69	37,13	40,95	43,58	56,45
Tipo de propriedade	23,15	28,56	28,14	24,78	38,69	33,79
Operação	25,69	17,69	26,80	16,83	34,92	26,90
Ambiente	29,42	18,94	12,83	26,62	33,09	31,29

Variedade	21,95	20,56	21,97	29,89	27,03	30,19
Tipo	19,13	20,20	19,28	20,14	24,84	18,16
Maturador	11,09	5,00	19,29	7,75	21,67	7,89

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória. (2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
Fonte: Elaborado pelo autor (2016).

Em relação à análise da floresta aleatória para a classe TCH*ATR "baixa", 418 das observações foram classificadas corretamente. No entanto, essa classe apresentou classificação errada de 11,81%, porque 29 + 9 + 18 = 56 observações não foram classificadas de forma correta. Além disso, a classe TCH*ATR "baixa" apresenta acerto de 88,19% das observações (Tabela 16). A classe TCH*ATR "baixa-média" apontou 392 das observações classificadas corretamente e erro de 20,33%. A classe TCH*ATR "média-alta" indicou 416 das observações classificadas corretamente e erro de 13,87%. Já a classe TCH*ATR "alta" apresentou 389 das observações classificadas de modo correto e erro de 18,96%, uma vez que 13 + 18 + 60 = 91 observações não foram classificadas corretamente. Portanto, para a classe TCH*ATR "alta", 81,04% dos indivíduos foram classificados corretamente (Tabela 16).

Tabela 16 - Matrix de confusão para os dados do treinamento da floresta aleatória para a variável resposta TCH*ATR (considerando a variável "safra").

Preditos / Reais	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	418	29	9	18	0,1181
Baixa-Média	23	392	65	12	0,2033
Média-Alta	2	26	416	39	0,1387
Alta	13	18	60	389	0,1896

Tamanho da amostra, N=1996
Fonte: Elaborado pelo autor (2016).

Ao levar em conta as análises de árvore de decisão e floresta aleatória, o efeito da safra é importante para determinar os valores de TCH e principalmente de ATR. Na sequência da variável estágio (número de cortes na cana) e dos fatores climáticos, o efeito da safra aparece como o mais importante, pois esta é uma combinação dos efeitos meteorológicos, de manejo e do ambiente testado (solo). Contudo, para criar um modelo de predição futura dos valores de TCH e ATR, a variável safra deve ser tirada do modelo. Essas análises encontram-se presentes no tópico seguinte.

4.4 Árvore de decisão e floresta aleatória (desconsiderando a variável “safra”)

Para a variável TCH, o método floresta aleatória apresentou menor taxa de erro ao alocar TCH nas classes corretas: "baixa", "baixa-média", "média-alta" e "alta", e consequentemente, a acurácia de predição foi maior. A acurácia da árvore de decisão e floresta aleatória decresce nas etapas do treinamento, validação e teste dos dados; contudo, o método da floresta aleatória apresenta menor redução da acurácia. A taxa de erro da árvore de decisão foi em média duas vezes maior que o método da floresta aleatória (Tabela 17).

O modelo é construído utilizando o banco de dados de treinamento. O banco de dados da validação é usado para verificar a performance do modelo e o banco de dados do teste consiste em observações selecionadas aleatoriamente do banco de dados geral, as quais não foram utilizadas nas etapas do treinamento e validação dos modelos. Isso assegura a obtenção de estimativas não viesadas do desempenho do modelo em novas observações (WILLIAMS, 2011).

Tabela 17 - Acurácia de predição e taxa do erro de classificação para os bancos de dados do treinamento, da validação cruzada e teste comparando árvore de decisão e floresta aleatória pata TCH.

Predição*	Acurácia de predição (%)		Taxa do erro de classificação (%)	
	Árvore de decisão	Floresta aleatória	Árvore de decisão (1)	Floresta aleatória (2)
Banco de dados do treinamento	71	88	29	12
Banco de dados da validação cruzada	65	82	35	18
Banco de dados do teste	63	80	37	20

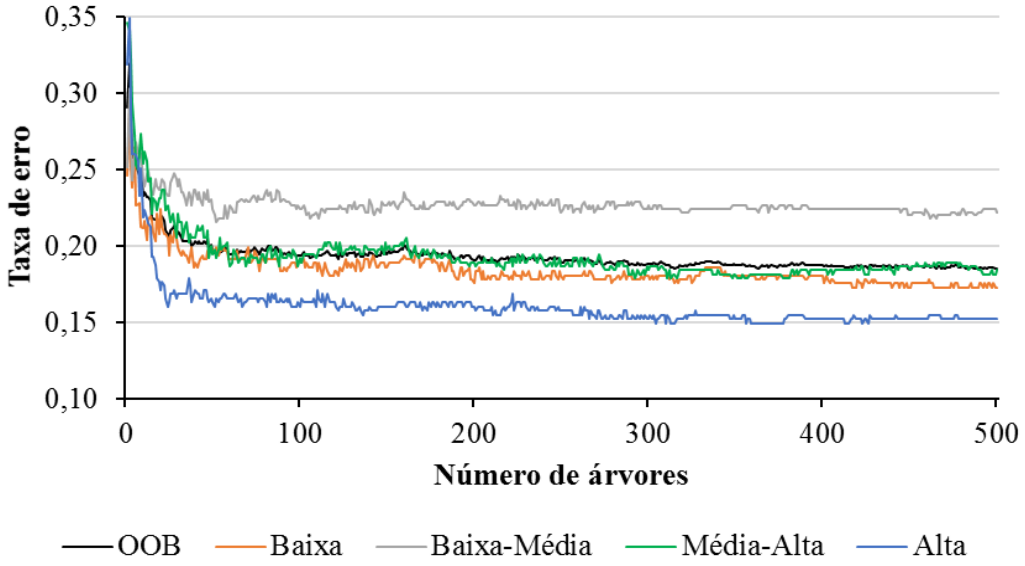
* O banco de dados geral foi repartido em: 70% treinamento, 15% validação e 15% teste. (1): Complexidade da árvore de classificação = 0,0027; (2): Erro geral do modelo de árvore de decisão com 500 árvores, 3 variáveis testadas em cada partição e taxa de estimativa do erro OOB = 18,53%.
Fonte: Elaborado pelo autor (2016).

Considerando o número de árvores × taxa do erro, é perceptível que, com o aumento do número de arvores no modelo da floresta aleatória, ocorre diminuição do erro OOB e, portanto, ocorre aumento da acurácia.

A análise da floresta aleatória apresenta ganho de acurácia de $80 - 63 \approx 17\%$ quando comparada à árvore de decisão no banco de dados do teste. Sendo assim, a análise da árvore de decisão apresenta maior acurácia na classificação de TCH do que apenas uma árvore. As classes TCH “baixa” e TCH “alta” foram aquelas que apresentaram menor taxa de erro de

classificação, provavelmente, porque são valores extremos dentro da distribuição normal e se distinguem das outras classes (Figura 25). Apenas uma árvore de decisão produz um modelo muito simples ou muito específico. Dessa forma, a análise de floresta aleatória produz várias árvores e as combina em um modelo que pode ser utilizado para tomada de decisão ou de classificação (WILLIAMS, 2011).

Figura 25 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta TCH.



Fonte: Elaborado pelo autor (2016).

A taxa de erro de classificação no banco de dados do teste representa o erro médio, caso um novo banco de dados seja gerado na usina e testado com o modelo previamente treinado e validado, sendo, portanto, o mais importante entre os três (treinamento, validação e teste).

A maior taxa de erro na análise da floresta aleatória foi obtida na classe TCH “média-alta”, com valor de 29,17%. Para a análise da árvore de decisão, a maior taxa de erro foi obtida na classe TCH “baixa”, com valor de 45,59%. Na floresta aleatória as classes TCH “baixa” e “alta”, tiveram taxas de erro de classificação menor que as classes intermediárias, provavelmente porque são valores extremos (Tabela 18).

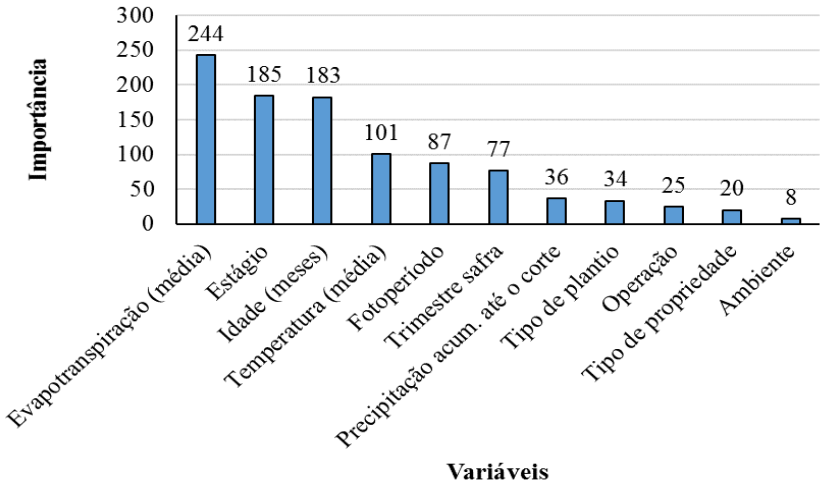
Tabela 18 - Comparação entre a árvore de decisão × floresta aleatória considerando o banco de dados teste para a variável TCH.

Árvore de decisão					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	37	14	10	7	0,4559
Baixa-Média	6	47	12	9	0,3649
Média-Alta	9	10	52	6	0,3247
Alta	10	6	18	63	0,3505
Floresta aleatória					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	49	10	1	0	0,1833
Baixa-Média	0	57	4	6	0,1493
Média-Alta	1	10	51	10	0,2917
Alta	0	7	7	72	0,1628

Fonte: Elaborado pelo autor (2016).

As variáveis que mais discriminam as classes de TCH na análise da árvore de decisão são: evapotranspiração (média), estágio, idade (meses), temperatura (média) e fotoperíodo. As variáveis que menos discriminam as classes de TCH são: ambiente, tipo de propriedade, operação, tipo de plantio e precipitação acumulada até o corte (Figura 26). Para TCH, a árvore foi podada considerando (complexidade) Cp = 0,0027; portanto, as variáveis variedade, tipo e maturador não entraram na análise da árvore de decisão.

Figura 26 - Importância das variáveis da árvore de decisão para a variável resposta TCH.



Fonte: Elaborado pelo autor (2016).

Na Tabela 19, quanto maior os valores mais as variáveis são importantes. As cinco variáveis mais importantes para a predição das classes de TCH na análise floresta aleatória, de acordo com a acurácia, foram: estágio, idade (meses), evapotranspiração (média), temperatura (média) e precipitação acumulada até o corte. As variáveis menos relacionadas com as classes de TCH foram: maturador, operação, tipo, ambiente e variedade (Tabela 19). Para TCH os resultados da importância das variáveis foram similares aos da árvore de decisão.

Tabela 19 - Coeficientes padronizados da importância das variáveis dentro de cada classe de TCH, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta TCH.

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Estágio	61,31	58,4	60,25	63,57	76,92	65,24
Idade (meses)	51,17	60,4	54,47	51,94	73,09	100
Evapotranspiração (média)	56,27	45,47	53,58	57,18	66,5	78,78
Temperatura (média)	47,4	46,44	50,47	51,00	65,42	70,08
Precipitação acum. até o corte	44,66	43,98	45,3	39,62	57,7	69,11
Fotoperíodo	41,31	38,99	40,89	41,89	54,69	58,56
Tipo de plantio	34,68	32,02	35,68	38,98	47,25	29,97
Trimestre safra	32,11	28	33,78	33,91	37,29	22,28
Tipo de propriedade	31,98	35,53	33,59	31,33	41,95	26,94
Variedade	21,66	20,03	22,25	26,55	25,82	13,34
Ambiente	27,79	23,93	20,67	25,87	35,44	19,15
Tipo	20,83	20,4	20,4	22,66	29,6	10,06
Operação	23,3	25,52	20,22	13,06	31,93	12,61
Maturador	6,99	-3,69	13,73	7,22	14,55	3,24

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
(2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
Fonte: Elaborado pelo autor (2016).

O maturador não apresentou muito efeito sobre TCH, ou seja, o valor final do TCH em si não é modificado com a aplicação do maturador. A variável tipo (cana planta ou cana soca), mesmo que apresente pouco efeito sobre a classificação de TCH, é bem relacionada com a variável estágio, que apresenta como classe o número de cortes a que a cana foi submetida. As variáveis apresentam produtividade similar, pois elas não discriminaram as classes de TCH e são, portanto, semelhantes, e a escolha das variedades deve ser baseada em outros parâmetros, como precocidade e adequação a determinado ambiente. A operação (aplicação de vinhaça ou fertilizante) também não mostrou grande efeito sobre os valores finais de TCH. Isso

demonstra que ambos os manejos conduzem a valores similares de TCH, embora mais análises devam ser feitas para explorar essa variável (Tabela 19).

Para a variável ATR, o método floresta aleatória apresentou menor taxa de erro ao alocar ATR nas classes corretas: "baixa", "baixa-média", "média-alta" e "alta" do que a árvore de decisão e, consequentemente, a acurácia de predição foi maior. A acurácia da floresta aleatória foi menor na etapa da validação, ao passo que para a árvore de decisão não houve diferença do erro entre os bancos de dados da validação e do teste. A taxa de erro da árvore de decisão foi, em média, nove vezes maior que o método da floresta aleatória no banco de dados do treinamento, três vezes maior na validação cruzada e quatro vezes maior no banco de dados do teste (Tabela 20). O algoritmo da floresta aleatória tende a produzir modelos acurados, porque os grupos de árvores gerados reduzem a instabilidade observada quando se constrói uma única árvore (WILLIAMS, 2011).

Tabela 20 - Acurácia de predição e taxa do erro de classificação para os bancos de dados do treinamento, da validação cruzada e do teste comparando árvore de decisão e floresta aleatória para ATR.

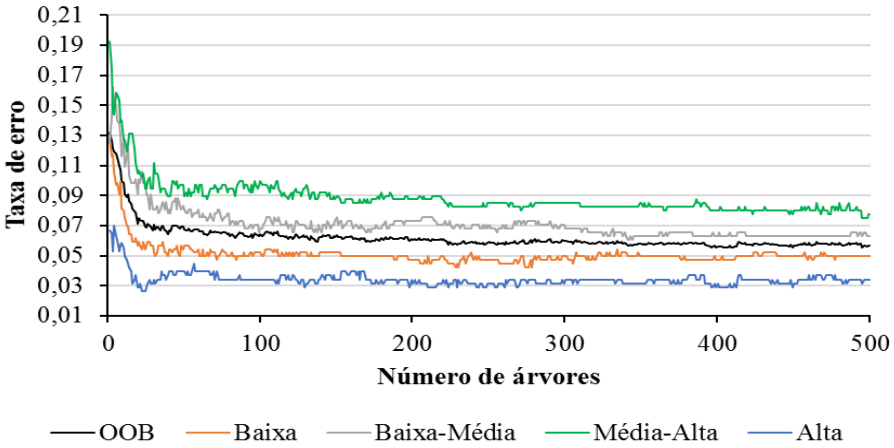
Predição*	Acurácia de predição (%)		Taxa do erro de classificação (%)	
	Árvore de decisão	Floresta aleatória	Árvore de decisão (1)	Floresta aleatória (2)
Banco de dados do treinamento	82	98	18	2
Banco de dados da validação cruzada	77	92	23	8
Banco de dados do teste	77	94	23	6

*: O banco de dados geral foi repartido em: 70% treinamento, 15% validação e 15% teste. (1): Complexidade da árvore de classificação=0,0043; (2): Erro geral do modelo de árvore de decisão com 500 árvores, 3 variáveis testadas em cada partição e taxa de estimativa do erro OOB= 5,65%.
Fonte: Elaborado pelo autor (2016).

Quando se considera o número de árvores × taxa do erro, nota-se que, com o aumento do número de arvores no modelo da floresta aleatória, ocorre diminuição do erro OOB e, consequentemente, aumento da acurácia. A análise da floresta aleatória apresenta ganho de acurácia de $94 - 77 \approx 17\%$ em relação à árvore de decisão no banco de dados do teste. Portanto, a análise da árvore de decisão apresenta maior acurácia na classificação de ATR do que apenas uma árvore. As classes ATR “baixa” e ATR “alta” foram aquelas que indicaram

menor taxa de erro de classificação, provavelmente, porque são valores extremos dentro da distribuição normal e se distinguem das outras classes (Figura 27).

Figura 27 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta ATR.



Fonte: Elaborado pelo autor (2016).

A maior taxa de erro na análise da floresta aleatória foi obtida na classe ATR "média-alta", com valor de 12,16%. Para a análise da árvore de decisão, a maior taxa de erro ocorreu na classe ATR "média-alta", com valor de 31,65%. Na floresta aleatória e árvore de decisão, as classes ATR "baixa" e "alta" tiveram taxas de erro de classificação menor que as classes intermediárias, provavelmente porque são valores extremos (Tabela 21).

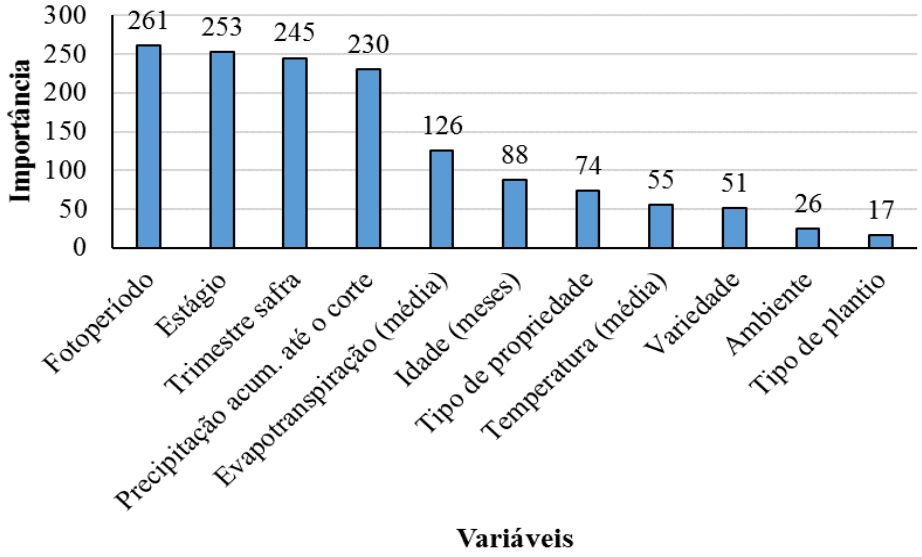
Tabela 21 - Comparação entre a árvore de decisão × floresta aleatória considerando o banco de dados do teste para a variável ATR.

Árvore de decisão					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	68	6	7	3	0,1905
Baixa-Média	11	59	13	3	0,3140
Média-Alta	6	5	54	14	0,3165
Alta	2	2	8	81	0,1290
Floresta aleatória					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	81	1	0	0	0,0122
Baixa-Média	3	73	3	1	0,0875
Média-Alta	0	3	65	6	0,1216
Alta	0	0	3	70	0,0411

Fonte: Elaborado pelo autor (2016).

As variáveis que mais discriminam as classes de ATR na análise da árvore de decisão são: fotoperíodo, estágio, trimestre safra, precipitação acumulada até o corte e evapotranspiração. As variáveis que mais discriminam as classes de ATR são: tipo de plantio, ambiente, variedade, temperatura (média) e tipo de propriedade (Figura 28).

Figura 28- Importância das variáveis da árvore de decisão para a variável resposta ATR.



Fonte: Elaborado pelo autor (2016).

As cinco variáveis mais importantes para a predição das classes de ATR na análise da floresta aleatória, de acordo com a acurácia, foram: idade (meses), temperatura (média), precipitação acumulada até o corte, estágio e fotoperíodo. As variáveis menos relacionadas com as classes de ATR foram: maturador, ambiente, operação, tipo e variedade (Tabela 22).

Tabela 22 - Coeficientes padronizados da importância das variáveis dentro de cada classe de ATR, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta ATR.

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Idade (meses)	46,41	56,58	58,55	45,08	69,3	97,53
Temperatura (média)	41,68	52,99	54,45	48,46	63,55	88,06
Precipitação acum. até o corte	46,17	50,99	54,39	46,27	63,14	84,33
Estágio	49,67	55,49	53,21	40,07	65,91	65,23
Fotoperíodo	52	53,73	49,77	51,59	61,68	85,43
Trimestre safra	48,97	48,78	49,39	46,9	54,52	59,25
Evapotranspiração (média)	48,99	53,37	48,42	41,75	66,67	83,56

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia (1)	Gini (2)
Tipo de propriedade	29,58	31,84	34,41	34,05	38,6	29,64
Tipo de plantio	39,46	38,54	34,15	26,7	47,31	31,34
Variedade	24,65	27,6	33,87	30,88	32,37	22,27
Tipo	19,32	26,27	25,75	20,93	31,59	11,88
Operação	15,47	23,87	25,64	19,24	33,25	13,17
Ambiente	19,66	19,54	21,54	21,84	31,55	13,19
Maturador	13,28	11,17	11,76	8,2	18,14	4,17

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
(2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória.
Fonte: Elaborado pelo autor (2016).

Assim como para TCH, os valores do ATR dependem da idade com que a cana é colhida, bem como de fatores meteorológicos, como temperatura, precipitação e fotoperíodo (Tabela 22). Considerando a média das duas variedades, a cana colhida com ou sem maturador não apresentou diferenças no valor final ATR. Portanto, além de ser uma questão de manejo, a utilização de maturadores não produz perdas de ATR, apesar de antecipar a colheita.

Para a variável TCH*ATR, o método floresta aleatória apontou menor taxa de erro ao alocar TCH*ATR nas classes corretas: "baixa", "baixa-média", "média-alta" e "alta" e, consequentemente, a acurácia de predição foi maior. A acurácia da floresta aleatória foi menor na etapa da validação, ao passo que para a árvore de decisão a taxa do erro de classificação foi menor na etapa do teste. A taxa de erro da árvore de decisão foi, em média, duas vezes maior que o método da floresta aleatória no banco de dados do treinamento, uma vez e meia maior na validação cruzada e, em média, duas vezes maior no banco de dados do teste (Tabela 23).

Tabela 23 - Acurácia de predição e taxa do erro de classificação para os bancos de dados do treinamento, da validação cruzada e do teste comparando árvore de decisão e floresta aleatória para TCH*ATR.

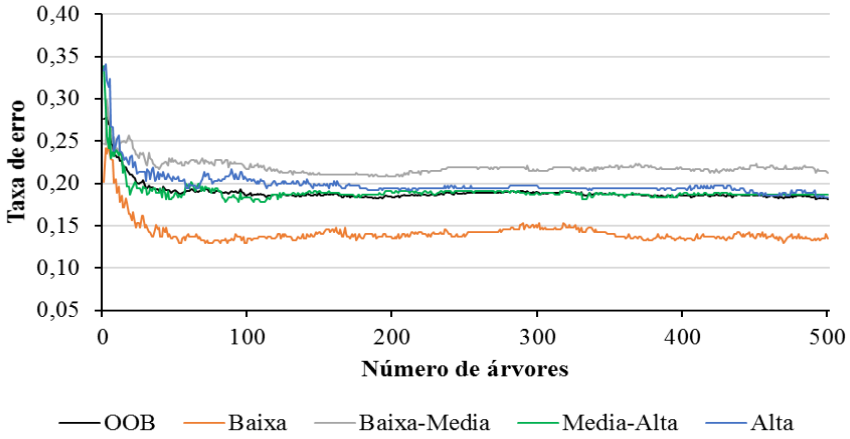
Predição*	Acurácia de predição (%)		Taxa do erro de classificação (%)	
	Árvore de decisão	Floresta aleatória	Árvore de decisão (1)	Floresta aleatória (2)
Banco de dados do treinamento	74	89	26	11

Banco de dados da validação cruzada	63	78	37	22
Banco de dados do teste	61	80	39	20

*: O banco de dados geral foi repartido em: 70% treinamento, 15% validação e 15% teste. (1): Complexidade da árvore de classificação = 0,0032; (2): Erro geral do modelo de árvore de decisão com 500 árvores, 3 variáveis testadas em cada partição e taxa de estimativa do erro OOB = 17,71%.
Fonte: Elaborado pelo autor (2016).

Considerando o número de árvores × taxa do erro para TCH*ATR, é perceptível que, com o aumento do número de árvores no modelo da floresta aleatória, ocorre diminuição do erro OOB, havendo, como consequência, aumento da acurácia. A análise da floresta aleatória apresenta ganho de acurácia de 80 – 61 ≈ 19% se comparada à árvore de decisão no banco de dados do teste. Portanto, a análise da árvore de decisão apresenta maior acurácia na classificação de TCH*ATR do que apenas uma árvore. As classes TCH*ATR "baixa" e TCH*ATR "média-alta" foram aquelas que indicaram menor taxa de erro de classificação, ao passo que as classes "alta" e "baixa-média" apresentaram maior erro de classificação (Figura 29).

Figura 29 - Número de árvores × erro para as classes da árvore de decisão e erro OOB para a variável resposta TCH*ATR.



Fonte: Elaborado pelo autor (2016).

A maior taxa de erro na análise da floresta aleatória foi obtida na classe TCH*ATR "média-alta", com valor de 22,97%. Para a análise da árvore de decisão, a maior taxa de erro se deu na classe TCH*ATR "alta", com valor de 43,00%. Na floresta aleatória, as classes TCH*ATR "baixa" e "alta" tiveram taxas de erro de classificação menor que as classes intermediárias, provavelmente, porque são valores extremos (Tabela 24).

As taxas de erro de classificação para essa variável foram mais próximas dos erros de TCH do que aqueles de ATR. Isso pode ser explicado porque TCH apresenta maior variância e magnitude, e quando esses valores são multiplicados por ATR, a análise de floresta aleatória de TCH*ATR fica mais próxima de TCH do que de ATR.

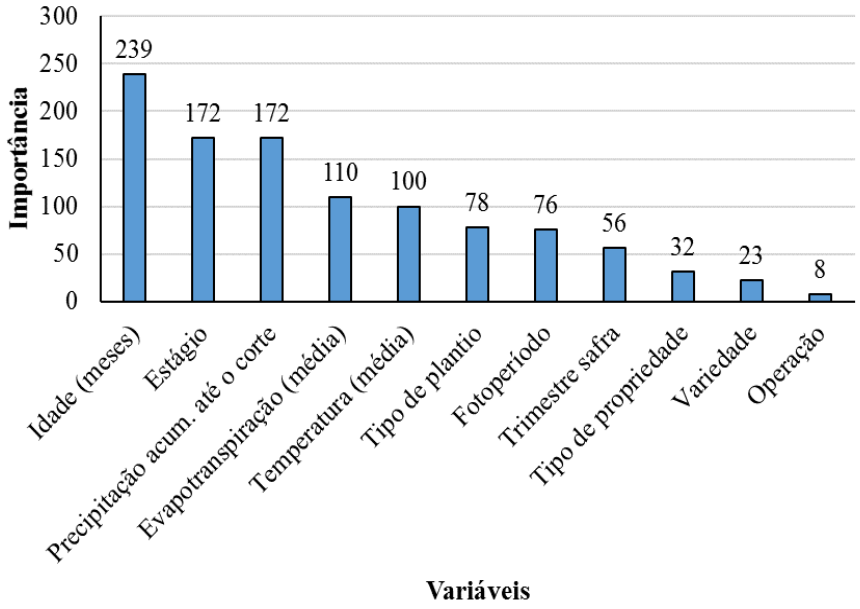
Tabela 24 - Comparação entre a árvore de decisão × floresta aleatória considerando o banco de dados do teste para a variável TCH*ATR.

Árvore de decisão					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	46	11	5	7	0,3333
Baixa-Média	8	38	14	5	0,4154
Média-Alta	6	13	53	10	0,3537
Alta	15	5	23	57	0,4300
Floresta aleatória					
Reais\Preditos	Baixa	Baixa-Média	Média-Alta	Alta	Erro de classificação
Baixa	51	9	2	1	0,1905
Baixa-Média	4	48	8	0	0,2000
Média-Alta	1	8	57	8	0,2297
Alta	4	1	11	72	0,1818

Fonte: Elaborado pelo autor (2016).

As variáveis que mais discriminam as classes de TCH*ATR para a árvore de decisão são: idade (meses), estágio, precipitação até o corte, evapotranspiração (média) e temperatura média. As variáveis que menos discriminam as classes de TCH*ATR são: operação, variedade, tipo de propriedade, trimestre safra e fotoperíodo. Uma vez que a árvore foi podada com complexidade, Cp = 0,0032. Logo, as variáveis tipo (cana planta ou soca) e maturador não entraram na análise de árvore de decisão (Figura 30).

Figura 30 - Importância das variáveis da árvore de decisão para a variável resposta TCH*ATR.



Fonte: Elaborado pelo autor (2016).

As cinco variáveis mais importantes para a predição das classes de TCH*ATR na análise da floresta aleatória, de acordo com a acurácia, foram: idade (meses), estágio, temperatura (média), precipitação acumulada até o corte e evapotranspiração (média). As variáveis menos relacionadas com as classes de TCH*ATR foram: tipo, maturador, ambiente, variedade e tipo de propriedade (Tabela 25). Para TCH*ATR os resultados da importância das variáveis na floresta aleatória foram similares aos da árvore de decisão.

Tabela 25 - Coeficientes padronizados da importância das variáveis dentro de cada classe de TCH*ATR, acurácia média decrescente de todas as classes, média do índice Gini de impureza dos nós da floresta aleatória para a variável resposta TCH*ATR.

Classes / Variável	Baixa	Baixa-Média	Média-Alta	Alta	Acurácia [1]	Gini [2]
Idade (meses)	57,01	55,91	61,56	53,65	72,36	104,35
Estágio	68,93	60,59	58,48	60,35	82,97	70,82
Temperatura (média)	48,89	45,47	52,00	50,32	61,81	75,37
Precipitação acum. até o corte	45,58	47,88	45,54	42,85	60,55	65,49
Evapotranspiração (média)	49,58	44,32	44,91	46,28	59,51	71,52
Trimestre safra	38,07	36,86	40,55	44,36	46,28	26,7
Tipo de plantio	36,37	30,83	37,39	35,19	46,29	29,85
Fotoperíodo	36,62	33,04	37,39	36,87	44,52	49,28
Operação	25,40	15,02	28,94	16,66	33,98	13,55

Tipo de propriedade	27,21	27,70	28,36	29,19	37,24	22,84
Variedade	24,71	24,14	23,45	31,98	29,51	17,8
Ambiente	28,26	17,56	21,47	24,17	32,56	17,88
Maturador	10,84	4,77	20,38	5,77	21,96	4,58
Tipo	22,46	20,2	18,36	19,21	26,97	9,64

(1): Diminuição média na acurácia, caso a variável da i-ésima linha seja deletada da análise floresta aleatória. (2): Diminuição média no índice Gini, caso a variável da i-ésima linha seja deletada da análise floresta aleatória. Fonte: Elaborado pelo autor (2016).

Considerando as análises de árvore de decisão e floresta aleatória, e desconsiderando a variável safra do modelo, os fatores climáticos são os mais importantes para definir os valores de TCH e ATR, principalmente as variáveis: evapotranspiração (média), temperatura (média), precipitação acumulada até o corte e fotoperíodo. Os tipos de manejos que podem ser feitos e impactar TCH e ATR são relacionados ao estágio (número de cortes na cana), idade (meses) em que a cana é colhida, tipo de plantio (mecanizado, semimecanizado, manual convencional e manual com torta) e planejamento do trimestre em que a cana será planta e colhida.

A análise de árvore de decisão pode ser empregada no agronegócio da cana-de-açúcar, utilizando dados climáticos (meteorológicos) e de manejo para realizar uma análise exploratória das variáveis mais importantes na definição de classes de TCH, ATR e TCH*ATR (“baixa”, “baixa-média”, “média-alta” e “alta”). As análises das florestas aleatórias podem ser usadas no agronegócio da cana-de-açúcar, utilizando dados climáticos (meteorológicos) e de manejo para realizar uma análise exploratória das variáveis mais importantes na definição de classes de TCH, ATR e TCH*ATR (“baixa”, “baixa-média”, “média-alta” e “alta”), com erro de classificação provável de 20% para TCH, 6% para ATR e também 20% para TCH*ATR.

4.5 Teste de média para os manejos: estágio, idade e tipo de plantio em relação a TCH e ATR

Foram realizadas análises de variância (ANAVA) para as fontes de variação “estágio” e “variedade”, onde foi constatado pelo teste F que existem diferenças estatísticas para TCH e ATR de acordo com a idade do canavial (Estágio) para a cultivar CTC4. Com relação a variedade RB966928 o teste F detectou diferenças apenas para TCH (Dados não apresentados).

A variedade CTC4 apresentou ATR máximo no primeiro corte com 18 meses e primeiro corte no inverno e valores mínimos no primeiro corte com a cana bis com 12 meses e no sétimo corte. Para TCH a CTC4 apresentou maior valor no primeiro corte da cana bis com

12 meses e menores valores no quinto e sexto corte. A variedade RB966928 não apresentou diferenças no valor do ATR para os cortes pelo teste de Tukey; contudo, para o TCH os maiores valores foram obtidos do primeiro corte no inverno ao quarto corte. No quinto e no sexto corte, houve decréscimo do valor de TCH para RB966928 (Tabela 26).

Tabela 26 - Teste de Tukey comparando o efeito do estágio do corte sobre as cultivares CTC4 e RB966928 para as variáveis TCH e ATR.

CTC4					
Estágio	ATR		Estágio	TCH	
Primeiro corte 12m	121,84	C	Primeiro corte 12m	100,39	bcd
Primeiro corte 18m	136,07	A	Primeiro corte 18m	105,46	b
Primeiro corte Inverno	138,62	A	Primeiro corte Inverno	98,71	bcd
Pri corte bis 12m	86,42	D	Pri corte bis 12m	151,82	a
Segundo corte	130,97	Ab	Segundo corte	110,06	b
Terceiro corte	131,33	Ab	Terceiro corte	110,91	b
Quarto corte	126,97	Ab	Quarto corte	103,37	bc
Quinto corte	124,77	bc	Quinto corte	87,67	d
Sexto corte	128,99	Ab	Sexto corte	90,99	cd
Sétimo corte	123,10	bc	Sétimo corte	104,87	bc
Oitavo corte	131,65	ab	Oitavo corte	92,77	bcd
Nono corte	125,21	abc	Nono corte	92,85	bcd
RB966928					
Estágio	ATR		Estágio	TCH	
Primeiro corte 12m	124,57	a	Primeiro corte 12m	104,40	ab
Primeiro corte Inverno	126,54	a	Primeiro corte Inverno	109,68	a
Primeiro corte 18m	121,22	a	Primeiro corte 18	125,30	a
Segundo corte	124,21	a	Segundo corte	112,19	a
Terceiro corte	124,68	a	Terceiro corte	113,63	a
Quarto corte	123,55	a	Quarto corte	110,75	a
Quinto corte	121,98	a	Quinto corte	91,95	b
Sexto corte	121,00	a	Sexto corte	102,11	ab

* médias seguidas pela mesma letra não diferem entre si pelo teste de Tukey (10%).
Fonte: Elaborado pelo autor (2016).

A variável idade do canavial corresponde aos meses entre o plantio e realização do primeiro corte, ou a idade entre cortes, caso o canavial seja cana soca. Essa variável é altamente relacionada com os valores de ATR e TCH a serem obtidos em cada safra. Cada

genótipo ou variedade apresenta uma particularidade quando se trata da idade do canavial e da produtividade.

A variedade CTC4 apresentou os maiores valores de TCH quando colhida com 17 meses de idade (117,81 $t.ha^{-1}$) e o segundo maior valor com 13 meses de idade (107,24 $t.ha^{-1}$), ao passo que a variedade RB966928 mostrou maior produtividade quando colhida com 16 meses de idade (193,03 $t.ha^{-1}$), ou com 14 e 15 meses de idade (126,79 $t.ha^{-1}$). Em termos práticos, considerando todo o banco de dados, a variedade CTC4 foi colhida 240 vezes com 13 meses e a RB966928 foi colhida 348 vezes com 12 meses (Tabela 25). Além disso, a variedade RB966928 apresentou maior TCH (114,28 $t.ha^{-1}$) que a variedade CTC4 (103,74$t.ha^{-1}$). Esses valores diferem estatisticamente com $p < 0,0001$ (Tabela 27).

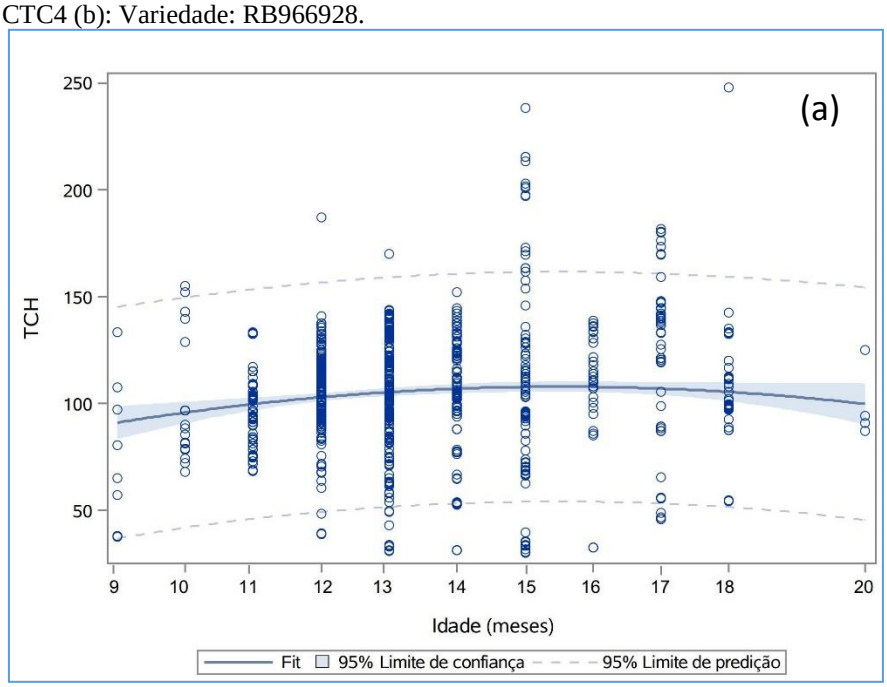
Tabela 27 - Estatística descritiva para TCH em relação à idade em que a cana é colhida, comparação entre as duas variedades.

Variedade	Idade (meses)	N	TCH Média	Desvio Padrão	Mínimo	Máximo
CTC4	9	9	72,67	34,45	37,52	133,36
	10	24	98,09	25,74	68,08	154,82
	11	115	98,87	15,74	68,52	133,28
	12	325	104,49	15,52	38,83	187,01
	13	240	107,24	26,62	30,80	169,96
	14	147	102,94	27,64	31,17	152,03
	15	108	102,37	49,68	30,02	238,27
	16	31	106,84	29,18	32,52	138,76
	17	68	117,81	39,46	46,01	181,54
	18	93	102,68	21,33	54,44	247,90
	20	4	99,33	17,28	87,09	124,87
Média	-	-	103,74	-	-	-
RB966928	9	28	103,92	23,72	35,77	146,19
	10	35	108,40	26,73	33,31	153,16
	11	237	101,73	29,77	32,97	196,06
	12	348	111,29	28,09	31,96	208,27
	13	163	112,31	27,31	57,27	185,63
	14	90	126,79	33,18	67,17	233,65
	15	66	126,79	29,56	31,15	203,71
	16	6	193,03	28,08	135,72	204,58
	17	4	115,36	28,40	73,11	134,62
	18	1	86,42	-	86,42	86,42
Média	-	-	114,28	-	-	-
p-valor			<0,0001			

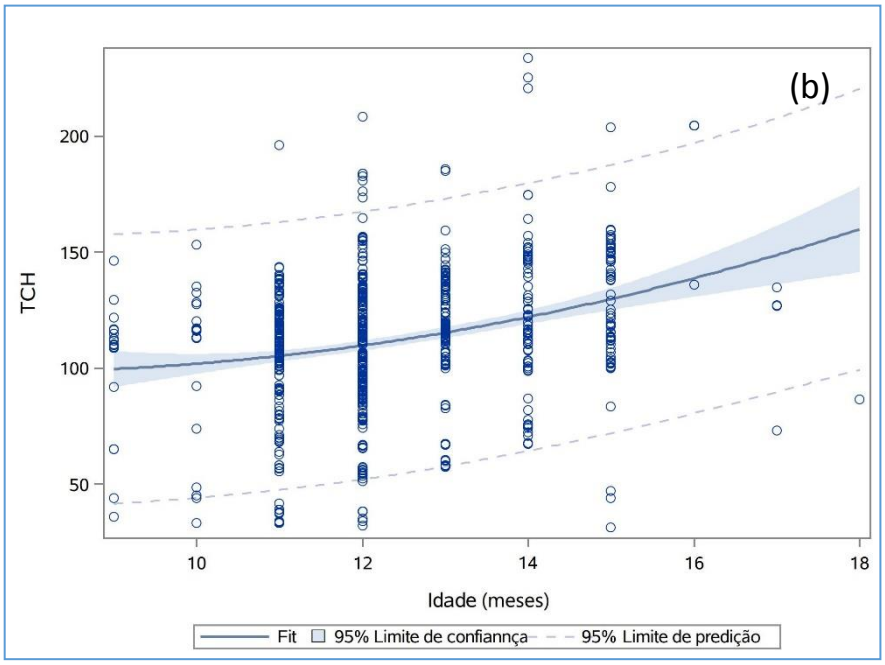
Fonte: Elaborado pelo autor (2016).

A Figura 31 ilustra que, com o aumento da idade em que a cana é colhida, a variedade CTC4 apresenta um pico de produtividade e começa a decrescer a partir do 18º mês (Tabela 28, Figura 31). A variedade RB966928 aponta um aumento progressivo de TCH com decréscimo quando colhida a partir do 17º mês (Tabela 27, Figura 31). Os limites de confiança ou de predição[3] expresso nos gráficos são utilizados para definir limites aos quais a verdadeira média da variável avaliada se encontra inclusa, sendo o valor 95% o mais usado porque resulta em um bom equilíbrio entre precisão (que é refletido na largura do intervalo de confiança) e confiabilidade (conforme expresso pelo nível de confiança) (FERREIRA, 2015). A particularidade é que o limite de predição será sempre mais amplo que o limite de confiança. Mesmo com o aumento do número de observações o limite de confiança apresenta pouca mudança de lugar, ao passo que o limite de confiança vai em direção a média geral dentro de cada mês estudado.

Figura 31 - Gráfico de dispersão entre a idade em que a cana é colhida com relação aos valores de TCH obtidos. (a): Variedade CTC4 (b): Variedade: RB966928.



[3] Limites de confiança e de predição: São também denominados de intervalos de confiança e de predição.



Fonte: Elaborado pelo autor (2016).

A variedade CTC4 apresentou os maiores valores de ATR quando colhida com 18 meses de idade (138,67) e o segundo maior valor com 11 meses de idade (134,83), ao passo que a variedade RB966928 obteve maior ATR quando colhida com 16 meses de idade (133,68) ou com 13 meses de idade (128,17). Em termos práticos, ao observar todo o banco de dados, a variedade CTC4 foi colhida 352 vezes com 12 meses e a RB966928 foi colhida 381 vezes com 12 meses (Tabela 29). Ademais, a variedade RB966928 apresentou menor TCH (122,71) que a variedade CTC4 (128,35). Esses valores diferem estatisticamente com $p < 0,0001$ (Tabela 28).

Tabela 28 - Estatística descritiva para ATR em relação à idade em que a cana é colhida, comparação entre as duas variedades.

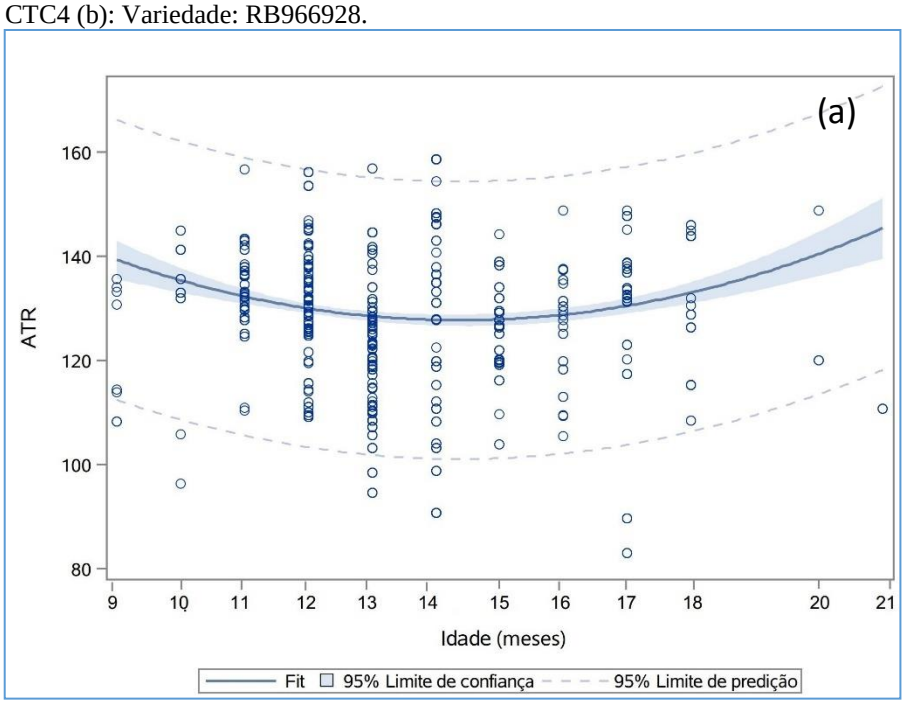
Variedade	Idade (meses)	N	ATR Médio	Desvio Padrão	Minimo	Maximo
CTC4	9	9	120,75	12,28	108,31	135,61
	10	26	134,04	10,69	96,27	144,92
	11	120	134,83	6,30	110,35	156,69
	12	352	133,47	10,50	109,10	156,20
	13	254	120,91	13,05	94,53	156,83
	14	150	131,31	20,12	90,74	158,61
	15	110	126,14	7,04	103,94	144,32
	16	34	127,85	10,20	105,46	148,75
	17	72	130,07	17,86	83,09	148,75

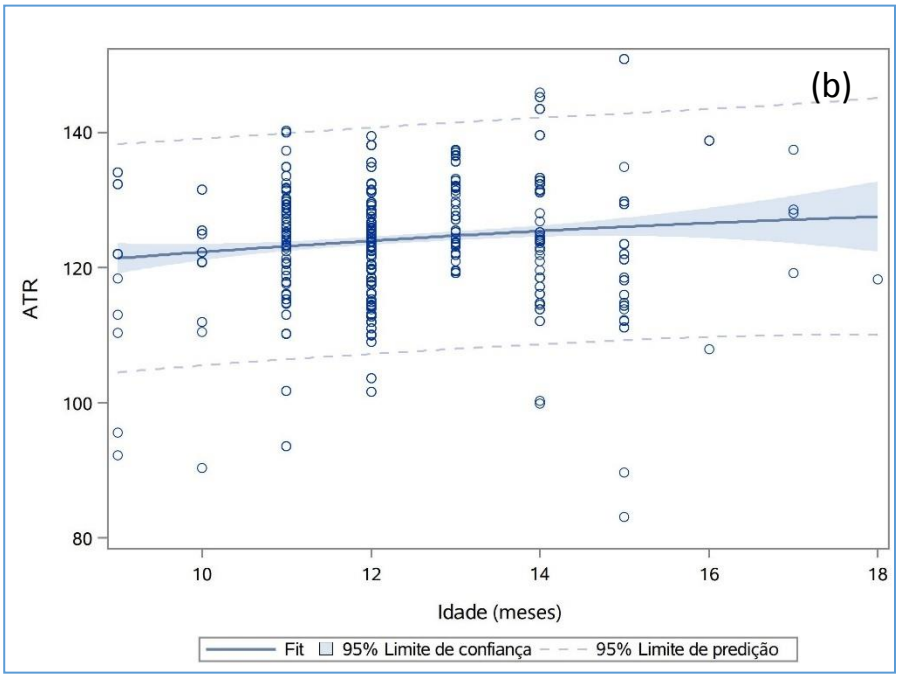
Variedade	Idade (meses)	N	ATR Médio	Desvio Padrão	Minimo	Maximo
	18	93	138,67	10,13	108,51	145,97
	20	4	127,19	14,38	120,00	148,75
	21	4	110,74	0,00	110,74	110,74
Média	-	-	128,35	-	-	-
RB966928	9	29	123,81	10,69	92,19	134,10
	10	39	121,50	6,90	90,29	131,50
	11	248	123,72	8,26	93,51	140,27
	12	381	122,10	7,65	101,59	139,41
	13	163	128,17	5,75	119,22	137,48
	14	96	127,72	10,40	99,94	145,85
	15	77	122,33	11,15	83,09	150,87
	16	6	133,68	12,62	107,91	138,83
	17	4	128,35	7,46	119,22	137,48
	18	1	118,27	-	118,27	118,27
Média	-	-	122,71	-	-	-
p-valor			< 0,0001			

Fonte: Elaborado pelo autor (2016).

A Figura 32 ilustra que a variedade CTC4 apresenta os maiores valores de ATR quando colhida com 10 a 12 meses ou com 17 e 18 meses (Tabela 29, Figura 35). A variedade RB966928 apresenta um aumento progressivo de ATR com relação à idade em que a cana é colhida (Tabela 28, Figura 32).

Figura 32 - Gráfico de dispersão entre a idade em que a cana é colhida com relação aos valores de ATR obtidos. (a): Variedade CTC4 (b): Variedade: RB966928.





Fonte: Elaborado pelo autor (2016).

Foram realizadas análises de variância (ANAVA) para as fontes de variação “tipo de plantio” e “trimestre safra”, onde foi constatado pelo teste F que existem diferenças estatísticas para TCH e ATR de acordo com o manejo para as cultivares CTC4 e RB966928 (Dados não apresentados).

Os diferentes tipos de plantio não influenciaram as médias de ATR para a variedade CTC4, contudo, para os valores de TCH o tipo de plantio mecânico apresentou maior média e o manual convencional menor média. Os tipos semimecanizados e manual com torta apresentaram valores intermediários. A variedade RB966928 apresentou maior média de ATR e TCH nos tipos semimecanizados e menores médias nos outros três tipos restantes (Manual com Torta, Manual Convencional e Mecânico) (Tabela 29).

Tabela 29 - Teste de Tukey comparando o efeito do tipo de plantio sobre as cultivares CTC4 e RB966928 para as variáveis TCH e ATR.

CTC4					
Tipo de plantio	ATR		Tipo de plantio	TCH	
Semimecanizado	135,04	A	Mecânico	115,59	a
Manual com Torta	131,26	A	Semimecanizado	111,89	ab
Manual Convencional	129,67	A	Manual com Torta	107,97	ab
Mecânico	120,99	B	Manual Convencional	99,078	b
RB966928					

Tipo de plantio	ATR		Tipo de plantio	TCH	
Semimecanizado	125,92	A	Semimecanizado	129,07	a
Manual com Torta	125,43	B	Mecânico	119,65	b
Manual Convencional	124,72	B	Manual com Torta	111,34	b
Mecânico	119,76	B	Manual Convencional	109,67	b

* médias seguidas pela mesma letra não diferem entre si pelo teste de Tukey (10%).
Fonte: Elaborado pelo autor (2016).

Tanto a variedade CTC4 como a RB966928 apresentaram maiores valores de ATR quando colhidas no trimestre T2, julho, agosto e setembro, e maiores valores de TCH quando colhidas no trimestre T1 abril, maio e junho (Tabela 30). Cada trimestre de safra diferiu estatisticamente entre si pelo teste de Tukey. Logo, existem diferenças no valor final do TCH e ATR quando o trimestre da colheita é alterado (Tabela 30).

Tabela 30 - Teste de Tukey comparando o efeito trimestre da safra sobre as cultivares CTC4 e RB966928 para as variáveis TCH e ATR.

CTC4					
Trimestre safra	ATR		Trimestre safra	TCH	
T2	138,87	a	T1	109,87	a
T3	126,70	b	T2	106,06	ab
T1	118,70	c	T3	102,65	b
RB966928					
Trimestre safra	ATR		Trimestre safra	TCH	
T2	129,56	a	T1	115,95	a
T3	125,26	b	T2	109,90	b
T1	120,74	c	T3	84,04	c

* médias seguidas pela mesma letra não diferem entre si pelo teste de Tukey (10%). * trimestre (T1): abril, maio e junho; trimestre (T2): julho, agosto e setembro e trimestre (T3): outubro, novembro e dezembro.
Fonte: Elaborado pelo autor (2016).

5 CONCLUSÃO

A aplicação de técnicas de mineração de dados com o objetivo de extrair novos conhecimentos ocultos em base de dados permite gerar informações úteis aos gestores subsidiando-os no processo decisório.

O objetivo desta pesquisa aplicada, empregando as técnicas de mineração de dados em empresa agrícola produtora de cana-de-açúcar, foi desenvolver um modelo utilizando árvores de decisão e floresta aleatória que determina qual cenário de ambiente de produção, clima e manejo produz maiores magnitudes para os valores de ATR e TCH. Além de contribuir como um complemento às técnicas de gestão já utilizadas pela empresa no processo de apoio ao negócio e assim proporcionar melhores resultados produtivos, operacionais e financeiros.

O modelo desenvolvido permite que árvores de decisão e principalmente florestas aleatórias possam ser utilizadas no agronegócio da cana-de-açúcar empregando dados climáticos (meteorológicos) e de manejo para prever classes de TCH, ATR e TCH*ATR com erro de classificação provável de 20% para TCH, 6% para ATR e 20% para TCH*ATR.

Com o desenvolvimento do modelo e aplicação das técnicas de mineração foi possível verificar que existe um relacionamento significativo entre os fatores de manejo, clima e ambiente de produção e a produtividade do canavial. Neste caso, observa-se até o momento que os tipos de manejo identificados e que ser realizados podem impactar TCH e ATR são relacionados ao: estágio (número de cortes na cana), idade em meses em que a cana é colhida, tipo de plantio (mecanizado, semimecanizado, manual convencional e manual com torta) e planejamento do trimestre em que a cana será plantada e colhida.

O TCH pode ser maximizado quando a variedade CTC4 é colhida com 13 meses de idade ao passo que a variedade RB966928 apresenta maior TCH se colhida dos 14 aos 16 meses de idade.

Tanto a variedade CTC4 como a RB966928 apresentaram maiores valores de ATR quando colhidas no trimestre T2 (julho, agosto e setembro) e maiores valores de TCH quando colhidas no trimestre T1 (abril, maio e junho).

A utilização das técnicas de mineração de dados mostrou-se útil para o descobrimento do conhecimento que se encontrava escondido na base de dados da empresa. A consistência das tarefas de classificação dos fatores que impactam na produtividade do canavial geradas pela ferramente R, foram avaliadas e comprovadas pela equipe de planejamento agrícola da empresa, a qual irá incorpar as técnicas de mineração para extração de novos conhecimentos úteis para a tomada de decisão.

5.1 Contribuições

A extração de conhecimento em grandes bases de dados utilizando mineração de dados objetiva buscar informações que é o resultado do processamento executado nesses dados, e gerar conhecimento, que é um conjunto de argumentos e explicações interpretando as informações processadas. Desta forma, as principais contribuições são:

O uso da metodologia CRISP-DM possibilita a resolução de problemas de extração de informações de uma forma organizada e progressiva, tendo como início uma análise de alto nível, a qual busca a compreensão das regras do negócio, direcionando-se para a definição e implantação de modelos que permitem a obtenção efetiva dos objetivos da mineração.

A utlização da metodologia no ambiente proposto, permitiu a viabilidade e a utilidade prática da metodologia em um estudo de caso real, sendo que os resultados poderão auxiliar os gestores a elucidar características relevantes em relação a diversas situações observadas neste estudo. As conclusões permitiram mostrar a relevência da metodologia CRISP-DM na obtenção dos resultados da mineração de dados.

O resultado da utilização desta metodologia tende a proporcionar uma melhor interpretação das atividades inerentes ao uso das técnicas de mineração de dados pelos gestores da empresa, haja vista que os mesmos não estão familiarizados com tais técnicas e terão mais um recurso a sua disposição para auxiliar nas tomadas de decisões.

A análise de dados feita pelo uso de técnicas de mineração de dados é ainda um pouco difundida na empresa, assim sendo este estudo e as sugestões para trabalhos futuros visam contibuir para que o uso das técnicas e metodologias de mineração de dados seja utilizados como um diferencial competitivo também no setor agrícola.

Nesta pesquisa foi demonstrada a relevância do processo de mineração de dados a obtenção de informações no que se refere a análise das informações contidas no banco de dados da empresa. Assim, teve-se o objetivo de desenvolver um modelo que permitisse identificar e classificar os fatores que impactam na produtividade da cana-de-açúcar por meio da aplicação das técnicas de árvore de decisão e floresta aleatória.

Entretanto, vale salientar que para cada objetivo desejado devem-se aplicar tarefas e técnicas específicas para se conseguir qualidade nos resultados esperados.

Além disso, a presença do profissional da área da atividade avaliada também não pode ser descartada. Ele participa desde o início como conhecedor do domínio do problema até o final na análise de viabilidade dos resultados.

Outro ponto essencial é a qualidade de dados utilizados na mineração. Para que o resultado possa ser utilizado por profissionais no processo de tomada de decisão, é necessário

que os dados sejam coletados e armazenados de maneira precisa. Dados imprecisos podem não descrever corretamente uma condição de manejo ou climatológico.

No setor agrícola onde os resultados obtidos sofrem influência de fatores climáticos, de manejo e mercadológicos, a utilização das técnicas de mineração de dados estão se tornando obrigatórias.

Alguns desafios e dificuldades foram encontrados durante o desenvolver desta pesquisa, entre eles:

a) A dificuldade de definição dos atributos, por parte dos gestores, que compuseram a base de dados para a mineração;

b) Ausência de estações metereológicas em todas as áreas produtivas, o que limitou o número de amostras observadas;

5.2 Sugestões para trabalhos futuros

Após o estudo abordado nesta dissertação, estabelecem-se algumas recomendações para pesquisas de mesmo cunho. Os principais são:

a) Aplicação de novas tarefas e técnicas de mineração de dados não contempladas neste estudo para dados de empresas agrícolas produtoras de cana-de-açúcar;

b) Implementação de algoritmos de Mineração de Dados junto a ferramenta de gestão agrícola oportunizando ao próprio gestor elaborar sua mineração;

c) Desenvolvimento de uma pesquisa contemplando outras variáveis climáticas e de manejo que possam impactar na produtividade do canavial possibilitando novos resultados para decisões mais elaboradas.

d) Cruzamento do histórico de dados de operação com o histórico de dados espaciais, de modo a produzir uma análise geográfica georreferenciada na linha de tempo da sequência de operações e de seus respectivos resultados.

e) Aplicação de técnicas estatística tradicional no banco de dados estudado para comparação de resultados apurados com as técnicas de mineração de dados aplicadas.

REFERÊNCIAS

AGROINFORMÁTICA, n. 8, Bento Gonçalves, 2011.

ANDRADE, E. T.; CARVALHO, S. R. G.; SOUZA, L. F. **Programa do proálcool e o etanol no Brasil**. Disponível em: <http://www.uff.br/enzimo/arquivos/arqix001.pdf>. Acesso em: 15 set. 2016.

ANDRETTA, R. L. **Influência de variáveis climáticas, da água disponível no solo, e dos eventos el niño e la niña na produtividade da cana-de-açúcar no estado do paraná**. 2012. 149 f. Dissertação de Mestrado em Angronomia, Universidade Federal do Paraná, Curitiba, 2012.

ANTUNES, J. F. G.; OLIVEIRA, S. R. M.; RODRIGUES, L. H. A. Mineração de dados para classificação das fases fenológicas da cultura da cana-de-açúcar utilizando dados do sensor modis e de precipitação. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 8., 2011, Bento Gonçalves. **Anais**... Florianópolis: UFSC; Pelotas: UFPel, 2011.

AUDY, J. L. N.; ANDRADE, G. K.; CIDRAL, A. **Fundamentos de sistemas de informação.** Porto Alegre: Artmed, 2005.

BALANÇA COMERCIAL DO AGRONEGÓCIO. **Síntese dos resultados do mês, do acumulado no ano e doze meses**. Disponível em: <http://www.agricultura.gov.br/internacional/indicadores-e-estatisticas/balanca-comercial>. Acesso em: 02 jul. 2016.

BANCO DO NORDESTE. **A evolução da produção de etanol no Brasil, no período de 1975 a 2009**. Disponível em: <http://www.bnb.gov.br/projwebren/Exec/artigoRenPDF.aspx?cd_artigo_ren=1342>. Acesso em: 15 set. 2016.

BANCO NACIONAL DO DESENVOLVIMENTO ECONÔMICO-BNDES. **Perspectivas do investimento 2007/2010**. TORRES FILHO, E. T.; PUGA, F. P. (Orgs.). Rio de Janeiro: BNDES, 2007.

BEGOLI, E.; HOREY, J. **Design Principles for Effective Knowledge Discovery from Big Data.** Computational Sciences & Engineering Division Oak Ridge National Laboratory. Tennessee, USA: Oak Ridge, 2012. Disponível em: <http://cda.ornl.gov/publications_2012/Publication_36116.pdf>. Acesso em: 15 ago. 2016.

BEUREN, I. M. (Org.). **Como elaborar trabalhos monográficos em contabilidade**: teoria e prática. 3. ed. São Paulo: Atlas, 2003.

BIOSEV. **Setor sucroalcooleiro**. Disponível em: <http://ri.biosev.com/biosev/web/conteudo_pt.asp?idioma=0&conta=28&tipo=30884>. Acesso em: 28 set. 2016.

BONOMA, T.V. Case research in marketing: opportunities, problems, and process. **Journal of Marketing Research**, v.22, p. 209, maio 1985.

BORDA, J. C. B.; GOMES, C.; REZENDE, F. Setor sucroalcooleiro enfrenta uma das maiores crises da história. **Jornal da Globo,** Rio de Janeiro, set. 2014. Disponível em:

<http://g1.globo.com/jornal-da-globo/noticia/2014/07/setor-sucroalcooleiro-enfrenta-uma-das-maiores-crises-da-historia.html>. Acesso em: 05 set. 2016.

BOX, G. E. P.; COX, D. R. An analysis of transformations (with discussion). **Journal of the Royal Statistical Society B**, v. 26, p.211-252, 1964.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento-MAPA. **Cana-de-açúcar**. Disponível em: <http://www.agricultura.gov.br/vegetal/culturas/cana-de-acucar>. Acesso em: 16 set. 2016.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BRESSAN, F. O método do estudo de caso. **Administração On Line**, São Paulo, v.1, n.1, 2000. Disponível em: <http://www.fecap.br/adm_online>. Acesso em: 25 ago. 2016.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas. Goiás: UFG, 2009. Relatório técnico. Disponível em: <www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf>. Acesso em: 10 jul. 2016.

CARUANA, R.; KARAMPATZIAKIS, N.; YESSENALINA, A. An empirical evaluation of supervised learning in high dimensions. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 25, HELSINKI, PROCEEDINGS. **Paper**s… Helsinki: ACM, p.96-103, 2008.

CASTANHEIRA, L.G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões**. 2008. Dissertação de Mestrado em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

CESAR, M. A. A.; DELGADO, A. A.; CAMARGO, A. P.; BISSOLI, B. M. A.; SILVA, F. C. Capacidade de fosfatos naturais e artificiais em elevar o teor de fósforo no caldo de cana-de-açúcar (cana-planta), visando o processo industrial. **STAB:** Açúcar, Álcool e Subprodutos, v.6, p.32-38, 1987.

CHEN, M. S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and data Engineering**, v. 8, n. 6, p. 866-883, 1996.

COMPANHIA NACIONAL DE ABASTECIMENTO-CONAB. **Séries históricas de área plantada, produtividade e produção, relativas às safras 1976/77 a 2014/15 de grãos, 2001 a 2014 de café, 2005/06 a 2014/15 de cana-de-açúcar**. Disponível em: <http://www.conab.gov.br/conteudos>. Acesso em: 02 jul. 2016.

CORREA, S. M. B. B. **Probabilidade e estatística**. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003.

COSTA, J. F. A. **Um ambiente gráfico para facilitar tarefas de data mining via ferramenta R**. 2011. Dissertação de Mestrado em Tecnologias e Sistemas de Informação, Engenharia e Gestão de Sistemas de Informação, Universidade do Minho, Lisboa, 2011.

COSTA, E. P.; POLITANO, P. R.; PEREIRA, N. A. Exemplo de aplicação do método de pesquisa-ação para a solução de um problema de sistema de informação em uma empresa

produtora de cana-de-açúcar. **Gest. Prod. São Carlos**, v.21, n. 4, oct/dec. 2014. Disponível em: <http://www.scielo.br> Acesso em: 02 jun. 2015.

CTC. **Variegates CTC (1 a 15)**. Disponível em: <www.coplana.com/gxpfiles/ws001/design/.../VariedadesCana/Variedade_CTC_115.pd>. Acesso em: 5 out. 2016.

CUNHA, M. J. **Descoberta do conhecimento de base de dados como ferramenta aplicada em processos sucroalcooleiros.** 2011.Tese de Doutorado em Engenharia Mecânica, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2011.

CRUZ, A. J. R. **Data Mining via redes neuronais artificiais e máquinas de vectores de suporte**. 2007. 123 f. Dissertação de Mestrado em Sistemas de Informação, Universidade do Minho, Lisboa, 2007.

DE MENDIBURU, F. Agricolae: statistical procedures for agricultural research. **R package version**, v. 1, p. 1-6, 2014. Disponível em: <https://cran.rproject.org/web/packages/agricolae/vignettes/tutorial.pdf >. Acesso em: 29 set. 2016.

DIAS, F. L. F. **Relação entre a produtividade, clima, solos e variedades de cana-de-açúcar, na Região Noroeste do Estado de São Paulo**. 1997. 64p. Dissertação de Mestrado em Engenharia de Sistemas Agrícolas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 1997.

DIAS, M. M. **Um Modelo de Formalização do Processo de Desenvolvimento de Sistemas de Descoberta de Conhecimento em Banco de Dados.** 2001. 212 f. Tese de Doutorado em Engenharia da Produção, Universidade Federal de Santa Catarina – UFSC, Florianópolis, 2001.

DIGIOVANI, M. S. **Em meio a forte crise, o setor sucroalcooleiro apresenta as projeções para a safra 2015/2016**. Disponível em: <http://www.sistemafaep.org.br/wp-content/uploads/2015/06/EmMeioForteCrise4.pdf>. Acesso em: 18 set. 2016.

DI GIROLAMO NETO, C. D.; RODRIGUES, L. H. A.; MEIRA, C. A. A. Modelos de predição da ferrugem do cafeeiro (Hemileia vastatrix Berkeley & Broome) por técnicas de mineração de dados. **Coffee Science**, v. 9, n. 3, p. 408-418, jul./set. 2014.

EMBRAPA. Agência Embrapa de Informação Tecnológica. **Cana-de-açúcar.** Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_73_22122006154841.html>. Acesso em: 2 set. 2016a.

_____. _____. **Fenologia**. Disponível em:<http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_68_22122006154840.html>. Acesso em: 15 ago. 2016b.

_____. **Análise quantitativa de crescimento em cana-de-açúcar:** uma introdução ao procedimento prático. Aracaju, SE: Embrapa Tabuleiros Costeiros, 2012.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011. 394 p.

FACHIN, O. **Fundamentos de metodologia**. 5. ed. São Paulo: Saraiva, 2003.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From *DM* to Knowledge Discovery in Databases. **AI Magazine**, v. 17, n. 3, 1996a.

_____. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **ACM**, v. 39,n.11, 1996b.

FERREIRA, P. V. **Estatística experimental aplicada à agronomia**. 3. ed. Maceió, AL: Edufal, 2000.

FERREIRA, V. A. **Estatística aplicada**. Campo Grande: Universidade Estácio de Sá, 2015. 136 p.

GARCIA, E.; CAMOLESI JUNIOR., L. Classificação de fatores que mais impactam a produtividade da cana-de-açúcar usando mineração de dados. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA (SBIAgro), 5., 2015. **Anais**...Universidade Estadual de Ponta Grossa, Ponta Grossa, 2015.

GIL, A. C. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GIOLO, S. R. **Análise de Regressão e Correlação**. Curitiba: Universidade Federal do Paraná, 2005.

HAN, J.; KAMBER, M. **Data Mining:** Concepts & Techniques. University of Illinois at Urbana-Champaign: Elsevier, 2006.

HARRISON, T. H. **Intranet data warehouse**. São Paulo: Berkeley Brasil, 1998.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**. 2.ed. New York: Springer, 2009.

HOLANDA, V. B.; RICCIO, E. L. **A utilização da pesquisa ação para perceber e implementar sistemas de informações empresariais**. 2010. Disponível em: <http://www.tecsi.fea.usp.br> Acesso em: 02 jun. 2015.

IDEA ONLINE. **Fatores para o sucesso da terceirização das operações de plantio e colheita de cana-de-açúcar**. Disponível em: <http://www.ideaonline.com.br/artigo/fatores-para-o-sucesso-da-terceirizacao-das-operacoes-de-plantio-e-colheita-de-cana-de-acucar.html>. Acesso em: 05 set. 2016.

JAMES, G.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning:** with applications in R. London: Springer, 2013. 429 p.

JANZEN, T. **Defining "Big Data" in Agriculture**. 2015. Disponível em: <http://janzenlaw.blogspot.com.br/2015/02/defining-big-data-in-agriculture.html>. Acesso em: 10 set. 2016.

KOCHE, J.C. **Fundamentos de metodologia científica**: teoria da ciência e prática de pesquisa. 14.ed. Petrópolis: Vozes, 1997.

KUHN, M. **A Short Introduction to the caret Package**. 2016. Disponível em: <https://cran.r-project.org/web/packages/caret/vignettes/caret.pdf>. Acesso em: jun 2017.

LIAW, A.; WIENER, M. Classification and regression by random Forest. **R news**, v. 2, n. 3, p. 18-22, 2002.

LOPES, M. A.; CONTINI, E. Agricultura, sustentabilidade e tecnologia. **Revista Agroanalysis**, fev. 2012.

LORENZZET, C. D. C.; TELÖCHEN, A. V. Estudo comparativo entre os algoritmos de Mineração de Dados Random Forest e J48 na tomada de decisão. **Revista eletrônica Unicruz**. Disponível em: <revistaeletronica.unicruz.edu.br/index.php/computacao/article/download/4023/737>. Acesso em: 25 set. 2016.

MAIMON, O.; LIOR, R. **Data Mining and Knowledge Discovery Handbook**. New York: Springer, 2010.

MARCHIORI, L. F. **Influência da época de plantio e corte na produtivade de cana-de-açúcar**. 2004. 277 f. Tese de Doutorado em Engenharia de Sistemas Agrícolas, Escola Superior de Agricultura "Luiz de Queiroz", Piracicaba, 2004.

MARIN, F. R.; GERALDO, B. M., CASSMAN JUNIOR, K. G.; GRASSINI, P. Prospects for Increasing Sugarcane and Bioethanol Production on Existing Crop Area in Brazil. **Revista Bioscience.** Disponível em: <http://bioscience.oxfordjournals.org/content/early/2016/02/12/biosci.biw009.abstract>. Acesso em: 15 out.2016.

MASSRUHÁ, S. M. F. S.; LEITE, M. A. A. Agricultura Digital. **RECoDAF – Revista Eletrônica Competências Digitais para Agricultura Familiar**, Tupã, v. 2, n. 1, p. 72-88, jan./jun. 2016.

MATTOZO, T.C. **Análise de desempenho de vendas em telecomunicações utilizando técnicas de mineração de dados.** 2007. Dissertação de Mestrado em Engenharia de Produção, Universidade Federal do Rio Grande do Norte, Natal, 2007.

MAULE, R. F.; MAZZA, J. A.; MARTHA JUNIOR, G. B. Produtividade agrícola de cultivares de cana-de-açúcar em diferentes solos e épocas de colheita. **Sci. agric.** [online], v.58, n.2, pp.295-301, 2001. Disponível em: <http://dx.doi.org/10.1590/S0103-90162001000200012>. Acesso em: 15 out. 2016.

MCKAY, J.; MARSHALL, P. **The dual imperatives of action research.** Churchlands, Australia: Edith Cowan University, 2001. Disponível em: <http://www.rbsv.eu/courses/rmtw/mtrl?AR?>. Acesso em: 02 jun. 2015.

McCUE, C. **Data Mining and Predictive Analysis** - Intelligence Gathering and Crime Analysis. Elsevier, 2007.

MIGUEL, P. A. C. Estudo de caso na engenharia de produção: estruturação e recomendações para sua condução. **Produção**, v. 17, n. 1, jan./abr. 2007, p.216-229.

MORAES, M. A. F. D.; OLIVEIRA, F. C. B.; DIAZ-CHAVEZ, R. A. Socio-economic impacts of Brazilian sugarcane industry. **Environmental Development**, v.16, p.31-43. dez, 2015. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2211464515000627>. Acesso em: 25 jul. 2016.

NEVES, M. F.; TROMBIN, V. G. **A dimensão do setor sucroenergético**: mapeamento e quantificação da safra 2013/14. Ribeirão Preto: Markestrat Consultoria, 2014.

NAKANO, B. Métodos de pesquisa adotados na engenharia de produção e gestão de operações. In: MIGUEL, P. C. (Org.). **Metodologia da pesquisa em engenharia de produção e gestão de operações**. Rio de Janeiro: Elsevier, 2010.

NYKO, D.; VALENTE, M. S.; MILANEZ, A. Y.; TANAKA, A. K. R.; RODRIGUES, A. V. P. A evolução das tecnologias agrícolas do setor sucroenergético: estagnação passageira ou crise estrutural? **BNDES Setorial**, Rio de Janeiro, n. 37, p. 399-442, 2013.

NONATO, R. T.; OLIVEIRA, S. R. 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.

NOVA CANA. **Atolada em dívida, indústria do açúcar vive o dilema de outros emergentes**. Disponível em: <http://www.novacana.com/n/industria/usinas/atolada-divida-industria-acucar-vive-dilema-emergentes-280915/> Acesso em: 28 set. 2015a.

_____. **Como é feito o transporte da cana-de-açúcar no Brasil**. Disponível em: <http://www.novacana.com/cana/transporte-da-cana-brasil/> Acesso em: 05 set. 2015b.

_____. Gerenciamento agrícola e tecnologia da informação nos canaviais. Disponível em:<https://www.novacana.com/cana/gerenciamentoagricolatecnologiainformacao/?tmpl=component&print=1 2/8>. Acesso em: jun. 2016a.

_____. Governo diz que Brasil produzirá 54 bilhões de litros de etanol em 2030. Disponível em: <http://www.novacana.com>. Acesso em: dez. 2016b.

PASTA, A. **Aplicação da técnica de Data Mining na base de dados do ambiente de gestão educacional**: um estudo de caso de uma instituição de ensino superior de Blumenau-SC. 2011. Dissertação de Mestrado em Computação Aplicada, Universidade do Vale do Itajaí, São José, 2011. Disponível em: <http://www.uniedu.sed.sc.gov.br/wpcontent/uploads/2013/10/Arquelau-Pasta.pdf >. Acesso em: 17 ago. 2016.

PRADO, H.; PADUA JUNIOR, A. L.; GARCIA, J. C.; MORAES, J. F. L.; CARVALHO, J. P.; DONZELI, P. L. Solos e ambientes de produção. In: DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M.; LANDELL, M. G. A. **Cana-de-açúcar**. Campinas: Instituto Agronômico, 2008. p. 179-204.

QUIROZ GIL, N. L.; VALENCIA, C. A. Aplicación del proceso de KDD en el contexto de bibliomining: El caso Elogim. **Revista interamericana de bibliotecología**, v. 35, n. 1, p. 97-108, 2012.

RABELO, E. **Avaliação de técnicas de visualização para mineração de dados**. 2007. 103 f. Dissertação (Mestrado em Ciência da Computação). Universidade Estadual de Maringá, Maringá, 2007.

R CORE TEAM. **R**: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria, 2016. Disponível em: <http://www.R-project.org/>. Acesso em: 17 out. 2016.

REZENDE, S. O. **Sistemas Inteligentes**: fundamentos e aplicações. São Paulo: Barueri, 2003.

RIDESA. **Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro**. Catálogo nacional de variedades "RB" de cana-de-acúcar / Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro. Curitiba, 2010.

RIPLEY, B. D. Tree-structured classifiers. In: Idem. **Pattern Recognition and Neural Networks**. Cambridge: Cambridge University Press, 2008, p. 416.

ROSSETTO, R. Agência Embrapa de Informação Tecnológica. **Cana-de-açúcar.** Disponível em: <http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/arvore/CONTAG01_73_22122006154841.html> Acesso em: 2 set. 2016a.

_____. _____. **Maturação**. Disponível em: <http://w ww.agencia.cnptia.embrapa.br/gestor/cana-de acucar/arvore/CONTAG01_90_22122006154841.html>. Acesso em: 18 set. 2016b.

SASS, S. R. R. **Abordagens de descoberta de conhecimento em bases de dados aplicadas ao cadastro territorial multifinalitário**. Dissertação de Mestrado em Ciências Cartográficas, Universidade Estadual Paulista, Presidente Prudente, 2013.

SILVA, J. E. A. **Desenvolvimento de um modelo de simulação para auxiliar o gerenciamento de corte, carregamento e transporte de cana-de-açúcar**. Dissertação de Mestrado em Engenharia de Produção, Universidade Federal de São Carlos - UFSCAR, 2006.

SOUZA, E. F. M.; PETERNELLI, L. A.; MELLO, M. P. **Software Livre R:** aplicação estatística. Disponível em: <http://www.de.ufpb.br/~tarciana/MPIE/ApostilaR.pdf>. Acesso em: 15 out. 2016.

SOUZA, Z. M. CERRI, D. G. P. COLET, M. J.; RODRIGUES, L. H. A.; MAGALHÃES, P. S. G.; MANDONI, R. J. A. Análise dos atributos do solo e da produtividade da cultura de cana-de-açúcar com o uso da geoestatística e árvore de decisão. **Ciência Rural**, v.40, p. 840- 847, 2010.

STROBL, C.; BOULESTEIX, A. L.; ZEILEIS, A.; HOTHORN, T. Bias in random forest variable imp ortance measures: illustrations, sources and a solution. **BMC Bioinformatics**, v. 8, n. 25, p. 8-25, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1796903/>. Acesso em: 30 nov. 2016.

THERNEAU, T. M.; ATKINSON, E. J.; FOUNDATION, M. **An introduction to recursive partitioning using the RPART routines**. 2015. Disponível em: < https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf >. Acesso em: 10 nov.2015.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. D. **RPART**: recursive partitioning and regression trees. R package version 4.1-10. 2015. Disponível em: <http://CRAN.R-project.org/package=rpart >. Acesso em: 17 out. 2015.

TOLEDO, M. Crise no setor canavieiro provoca fechamento de usinas e demissões. **Folha de S. Paulo** (*online*). São Paulo, 13 jul. 2014. Disponível em: <http://www1.folha.uol.com.br/mercado/2015/07/1655141-crise-no-setor-canavieiro-provoca-fechamento-de-usinas-e-demissoes.shtml>. Acesso em: 05 set. 2016.

TOMAZELA, M. G.; CAMPOS, F. C.; DANIEL L. A.; L. **Mineração de dados aplicada à produtividade de cana-de-açúcar.** In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUCAO, 35. Fortaleza, CE, Brasil, 13 a 16 out. 2015. Disponível em: <http://www.abepro.org.br/biblioteca/TN_STO_206_226_27367.pdf>. Acesso em: 22 set. 2016.

TOTVS. Disponível em: <https://www.totvs.com >. Acesso em: 24 jul. 2016.

TRIPP, D. Pesquisa-ação: uma introdução metodológica. **Educação e Pesquisa**, São Paulo, v. 31, n. 3, p. 443-466, set./dez. 2005. Disponível em: <http://www.scielo.br> Acesso em: 02 jun. 2015.

TSAI, H. H. Global data mining: an empirical study of current trends, future forecasts and technology diffusions. **Expert Systems with Applications**, v. 39, n. 9, p. 8172–8181, 2012. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0957417412001704>. Acesso em: 29 jul. 2016.

TURTELLI, C. Crise deixa dez usinas paradas na atual safra de cana-de-açúcar. **Folha de S. Paulo** (*online*). São Paulo, 24 abr. 2014. Disponível em: <http://www1.folha.uol.com.br/cotidiano/ribeiraopreto/2014/04/1444575-crise-deixa-dez-usinas-paradas-na-atual-safra-de-cana-de-acucar.shtml>. Acesso em 05 set. 2016.

TURRIONI, J. B.; MELLO, C. H. P. **Metodologia de pesquisa em engenharia de produção**: estratégias, métodos e técnicas para condução de pesquisas quantitativas e qualitativas. 2012. Programa de Pós-graduação em Engenharia de Produção, Universidade Federal de Itajubá, Itajubá, 2012.

_____. Pesquisa-ação na engenharia de produção. In: TURRIONE, J. B.; MELLO, C. H. P. **Metodologia de pesquisa em engenharia de produção e gestão de operações.** Rio de Janeiro: Elsevier, 2012.

TWO CROWNS CORPORATIONS. **Introduction to Data Mining and Knowledge Discovery**. 3 ed. Potomac: Two Crowns, 2005.

UNIÃO DA INDÚSTRIA DE CANA-DE-AÇÚCAR-UNICA. **Linha do Tempo**. Dis ponível em: <https://www.unica.com.br/linhadotempo/index.html>. Acesso em: 13 mar. 2016.

WILLIAMS, G. **Data mining with rattle and R.** The art of escavating data for knowledge discovery. New York: Springer, 2011.

YIN, R. K. **Estudo de caso**: planejamento e métodos. 3. ed. Porto Alegre: Bookman, 2005.

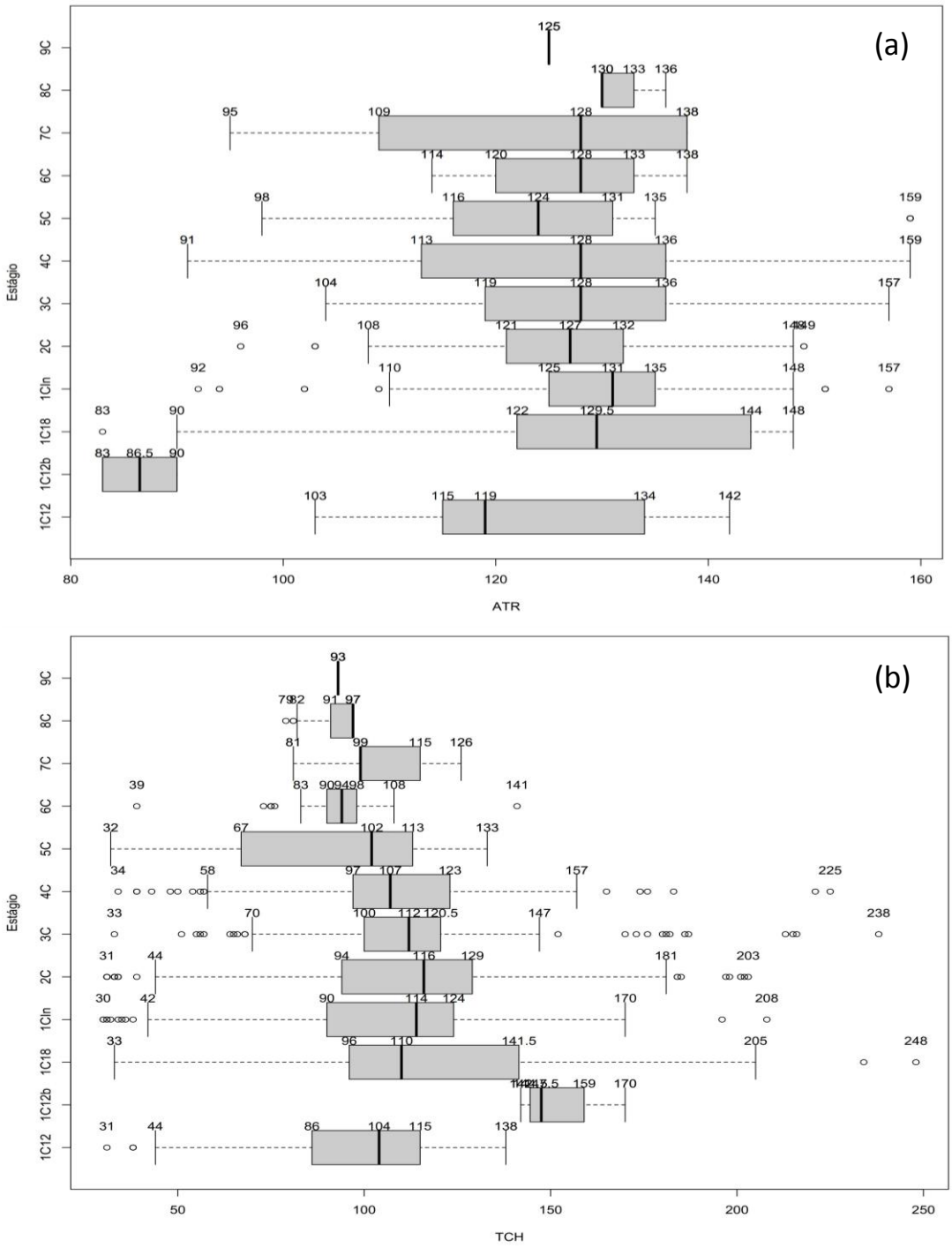
ZHANG, H. The optimality of naive bayes. In: **FLAIRS CONFERENCE – AAAI**, University of New Brunswick Fredericton, New Brunswick, Canada, 2004. Disponível em: <http://www.cs.unb.ca/~hzhang/publications/FLAIRS04ZhangH.pdf>. Acesso em: jun. 2016.

ZUMEL, N.; MOUNT, J. Practical Data Science with R. **Manning Publications Co**., 2014. Disponível em: <http://dl.acm.org/citation-cfm?id=2614429>. Acesso em: jun. 2016.

APÊNDICES

Apêndice A – Gráficos descritivos das variáveis respostas TCH e ATR vs variáveis independentes utilizadas nas análises de árvore de decisão e floresta aleatória.

Figura 33 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Estágio.



*****Boxplots para variáveis compostas de classes e gráfico de dispersão para variáveis independentes numéricas.
*Nos boxplots estão marcadas as estatísticas: mínimo, limite inferior, quartil 1, mediana, quartil 3, limite superior e valor máximo. Nos gráficos de dispersão está marcada a correlação de Pearson entre as duas variáveis em questão.

Figura 34 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Maturador.

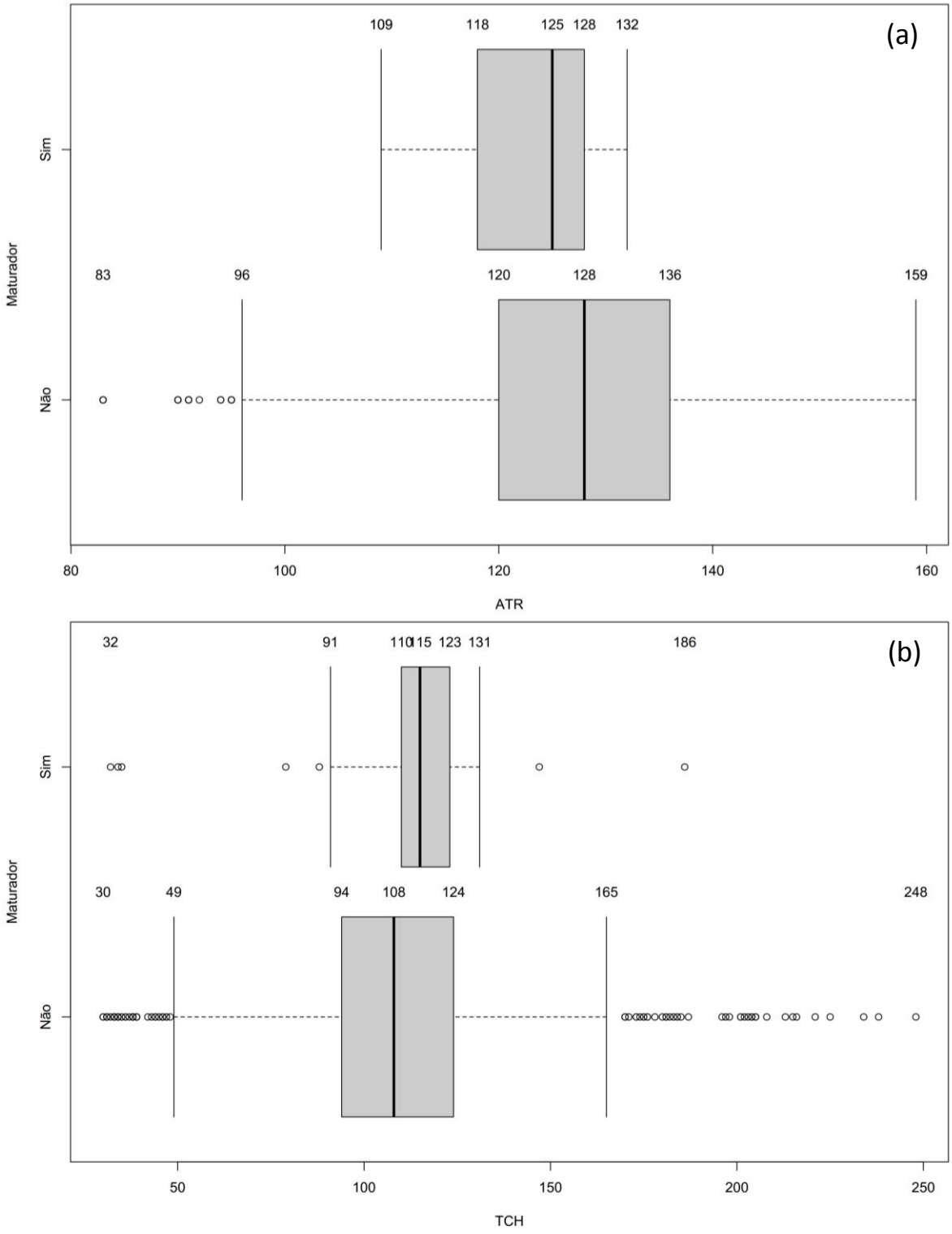
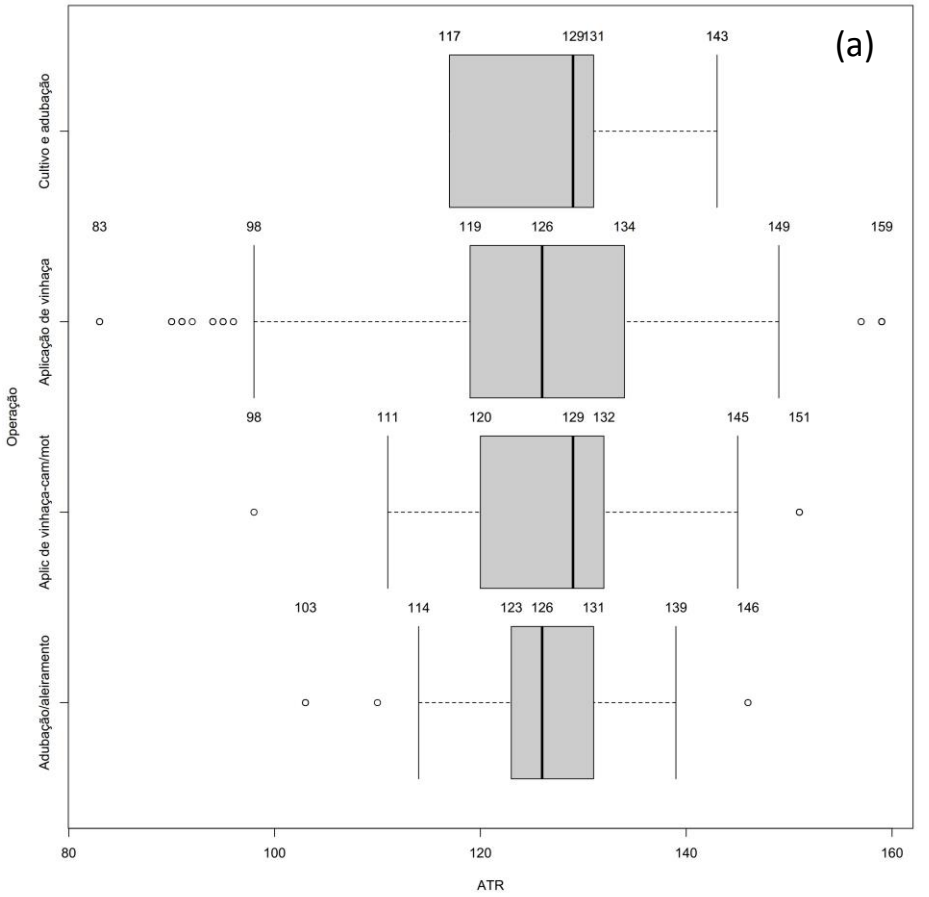


Figura 35 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Operação.



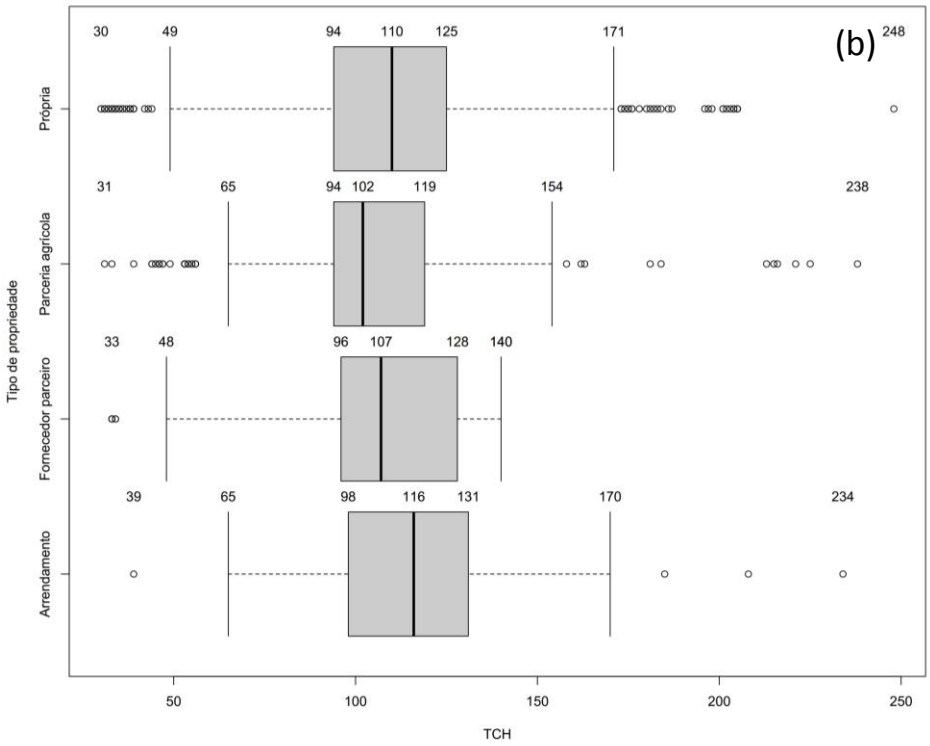


Figura 36 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Tipo.

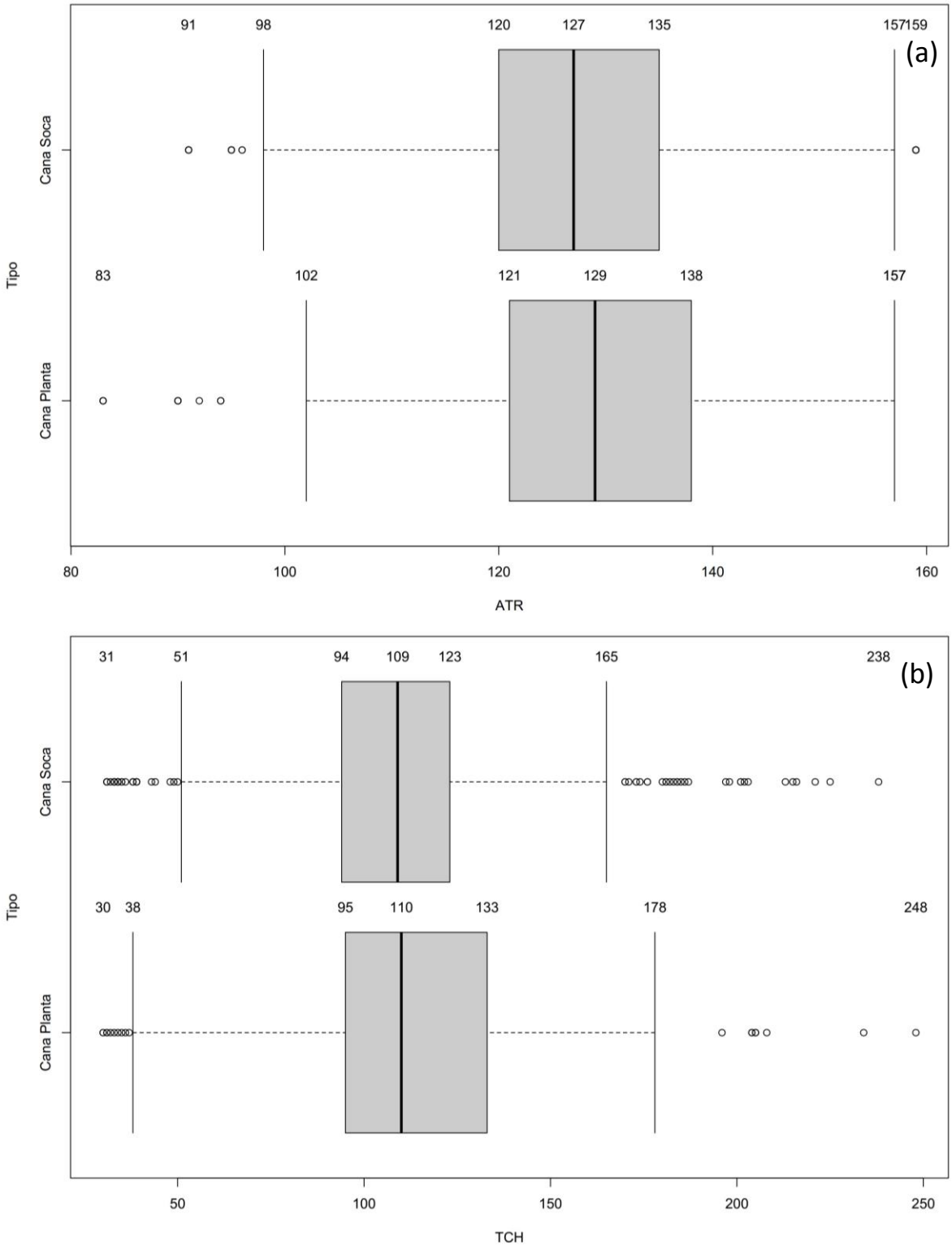


Figura 37 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Trimestre safra.

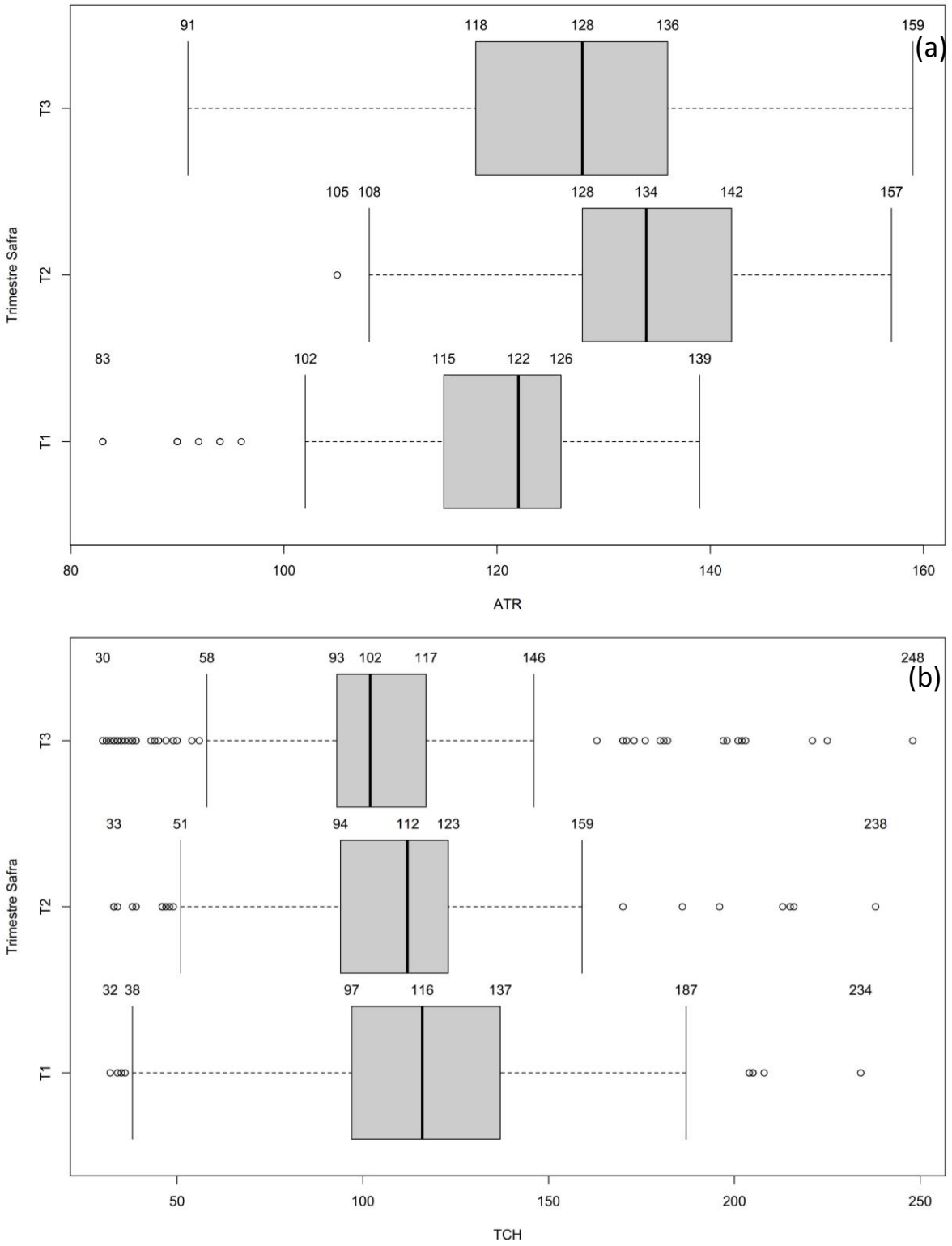
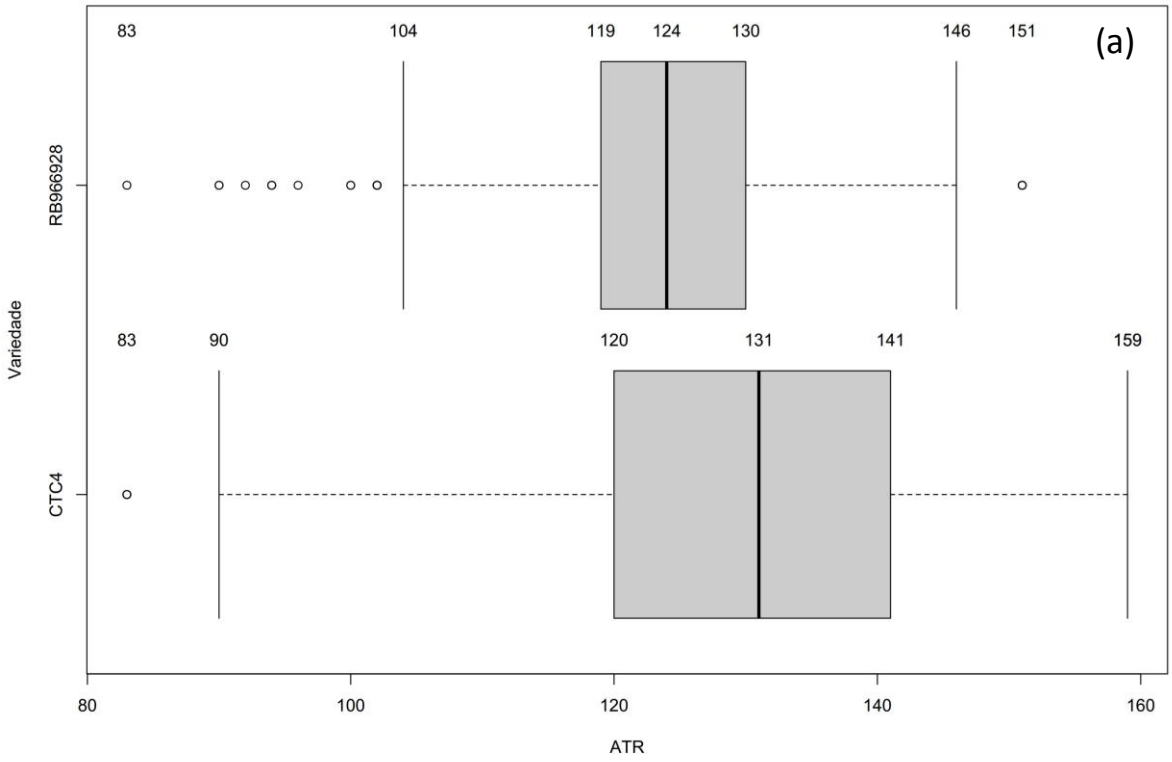


Figura 38 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Variedade.



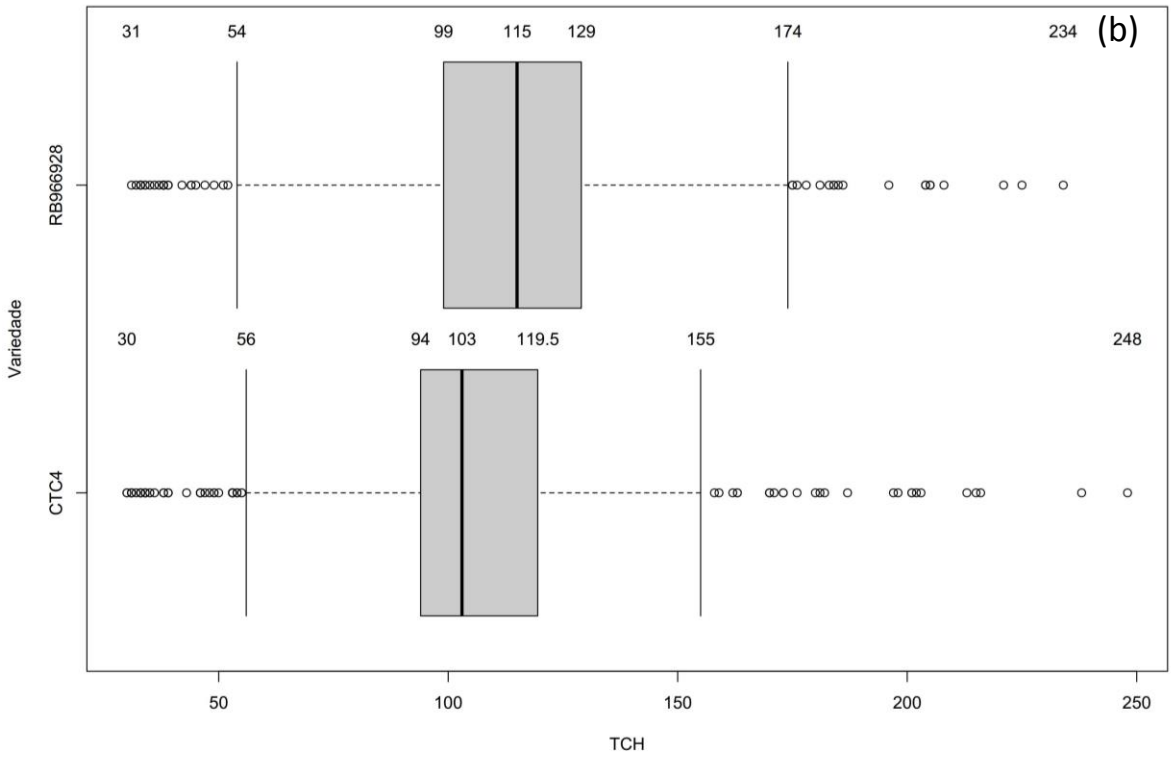


Figura 39 - Boxplot entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Ambiente.

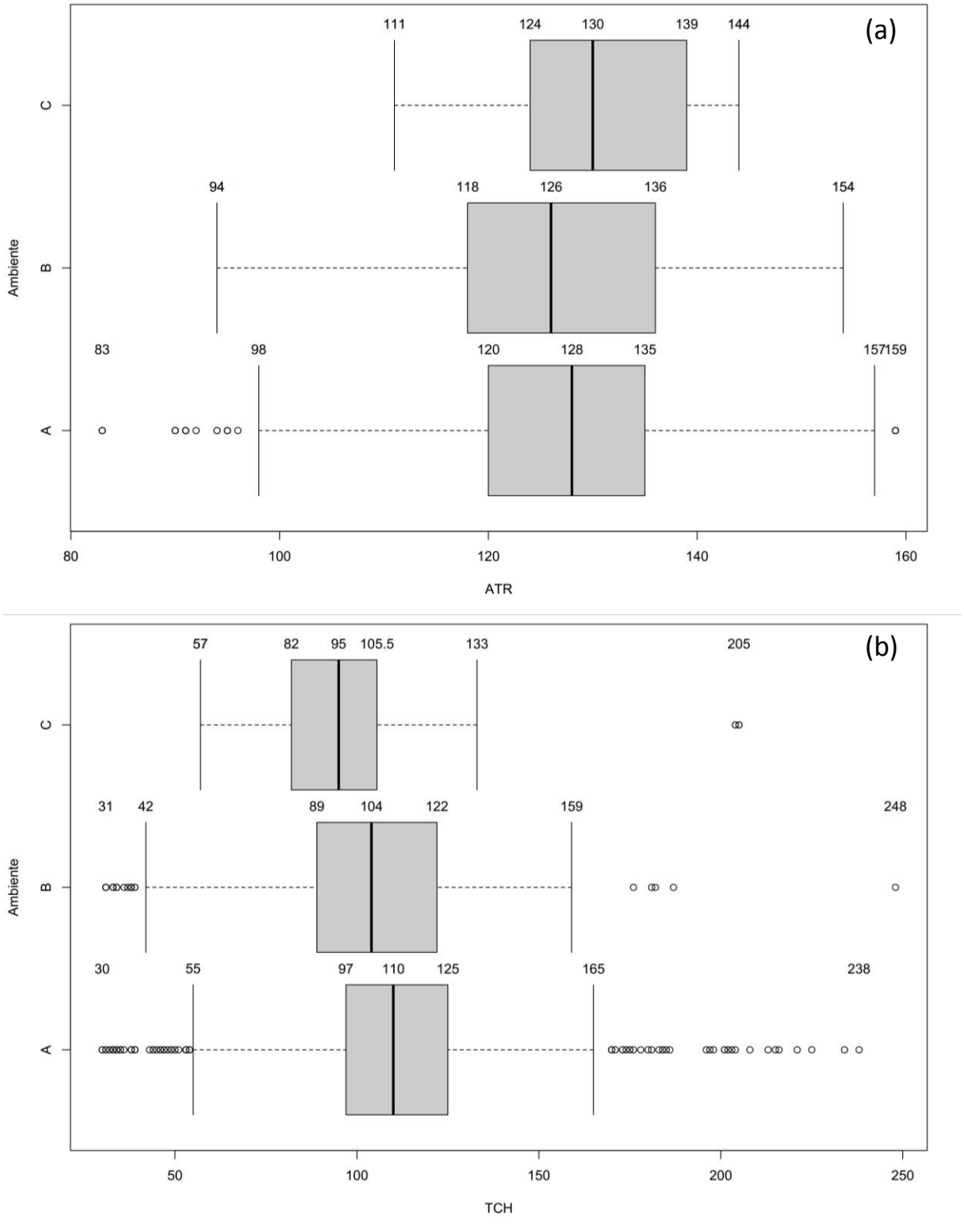


Figura 40 - Gráfico de dispersão entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Fotoperíodo (horas).

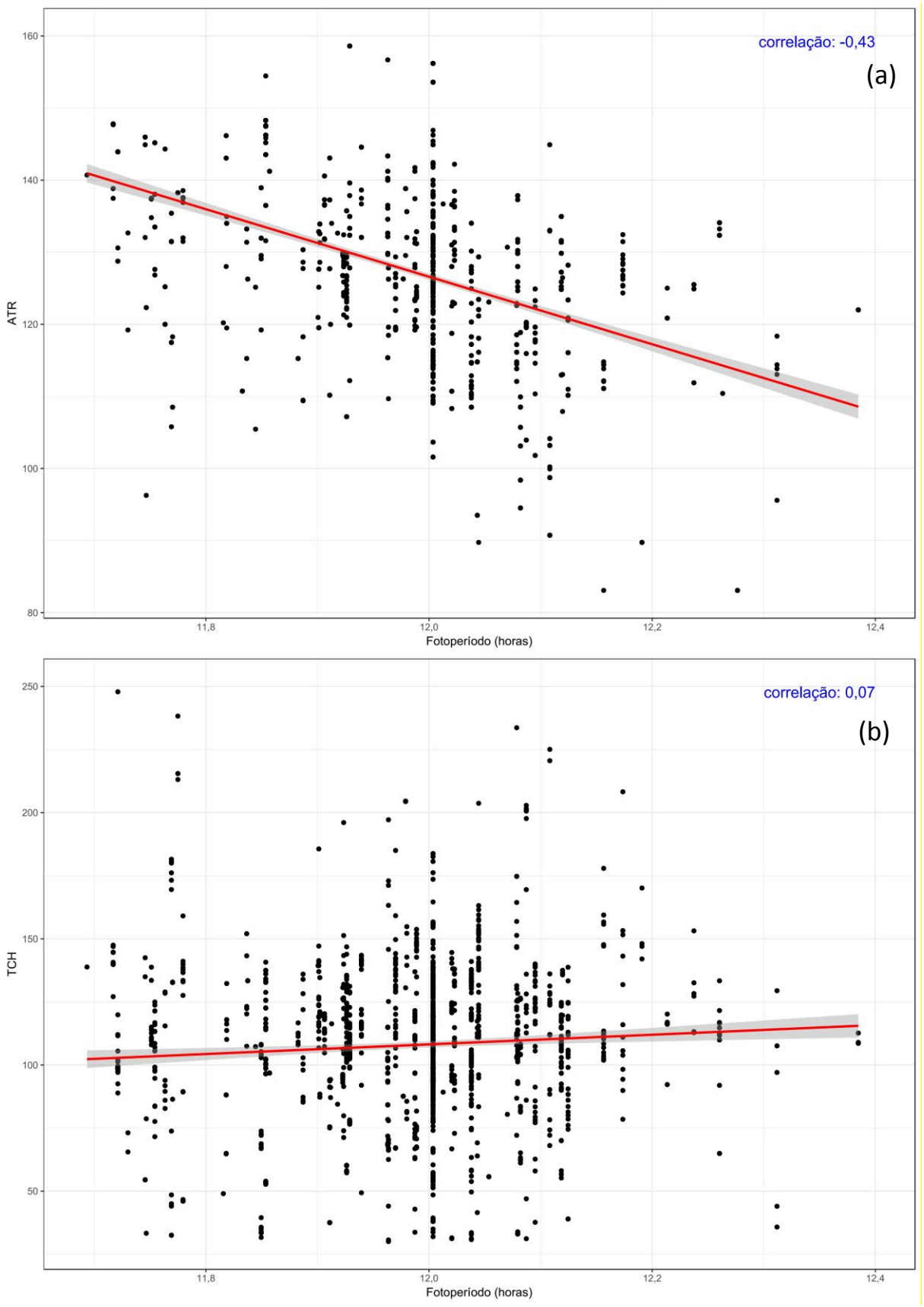
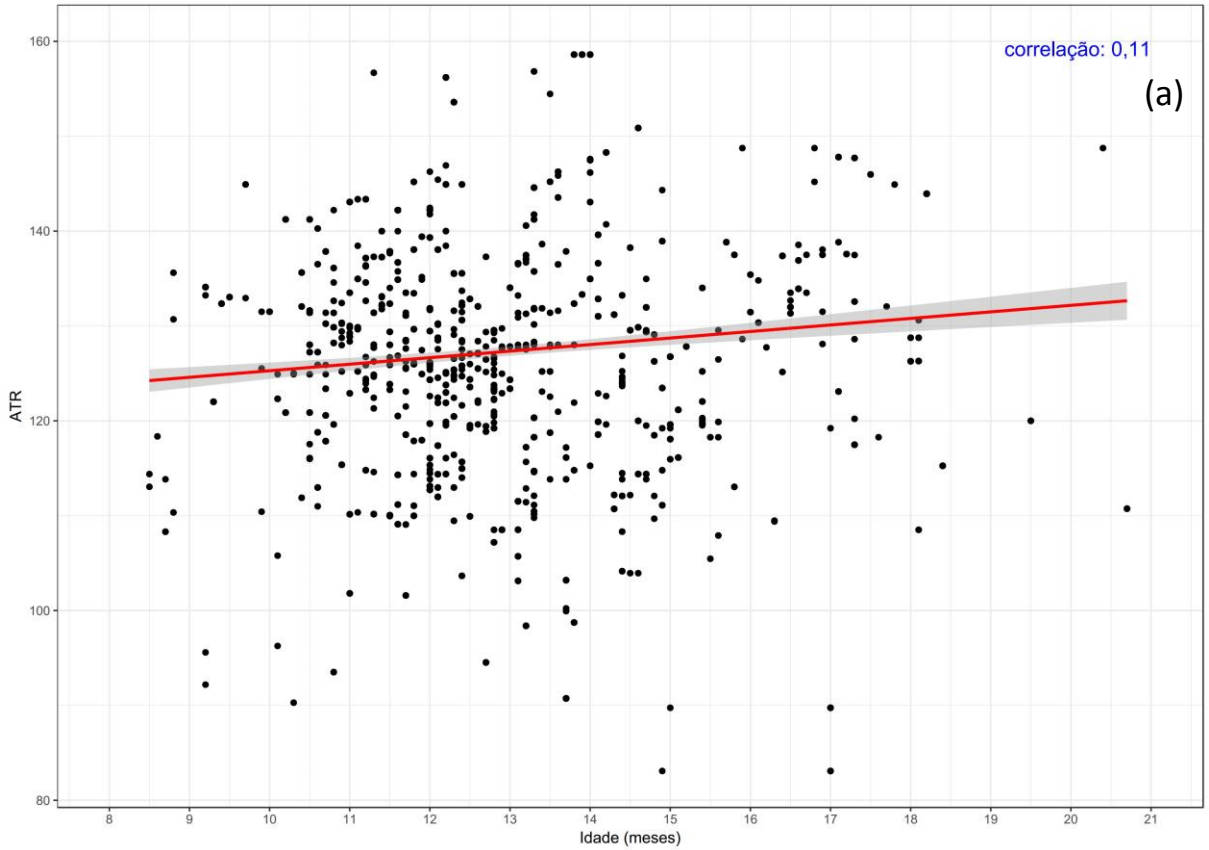


Figura 41 - Gráfico de dispersão entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Idade (meses).



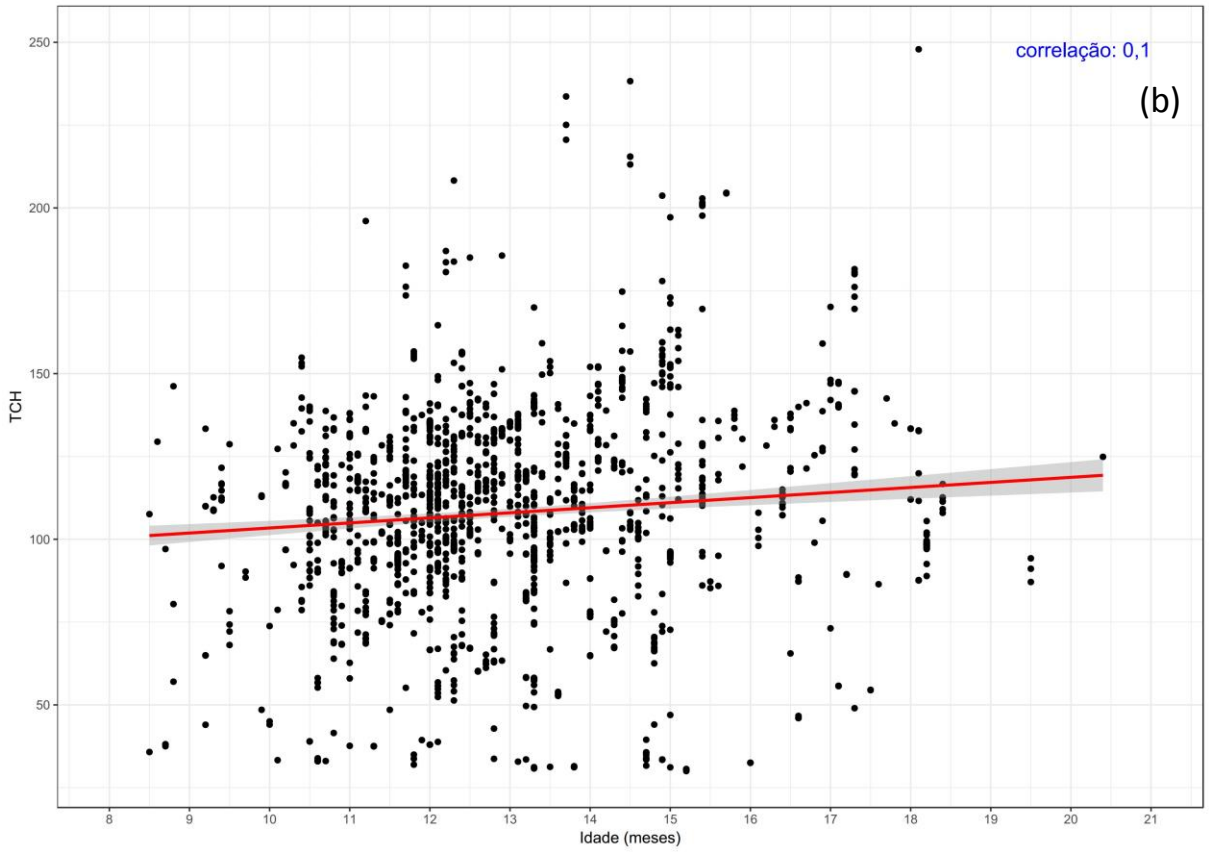
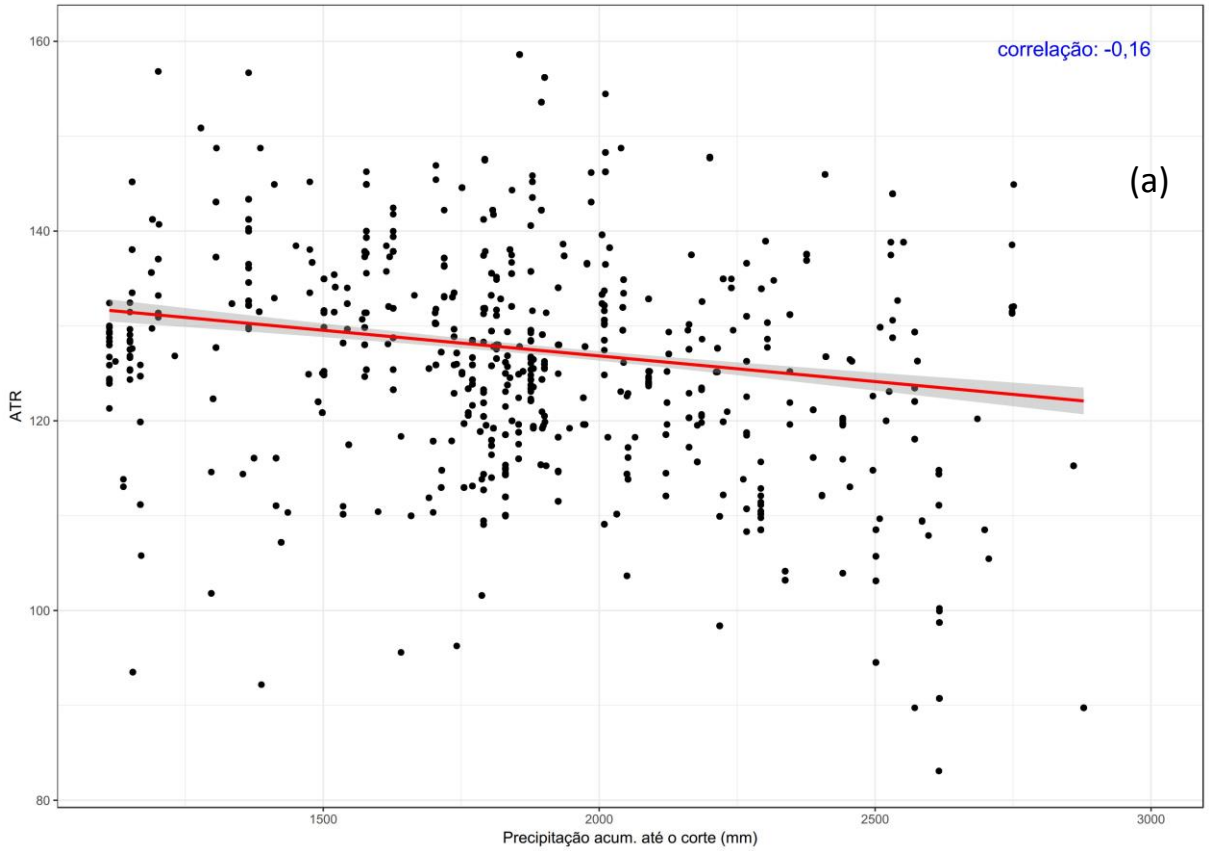


Figura 42 - Gráfico de dispersão entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Precipitação acum. até o corte (mm).



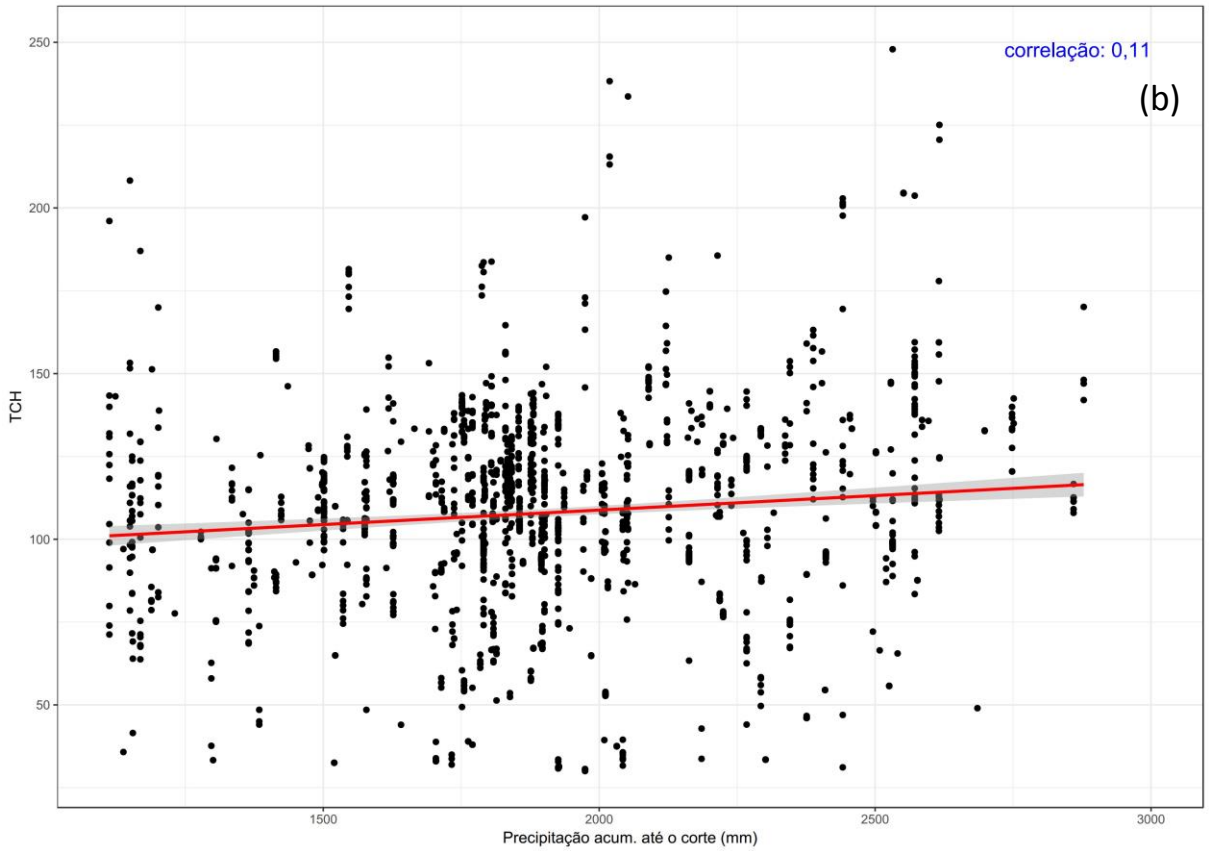
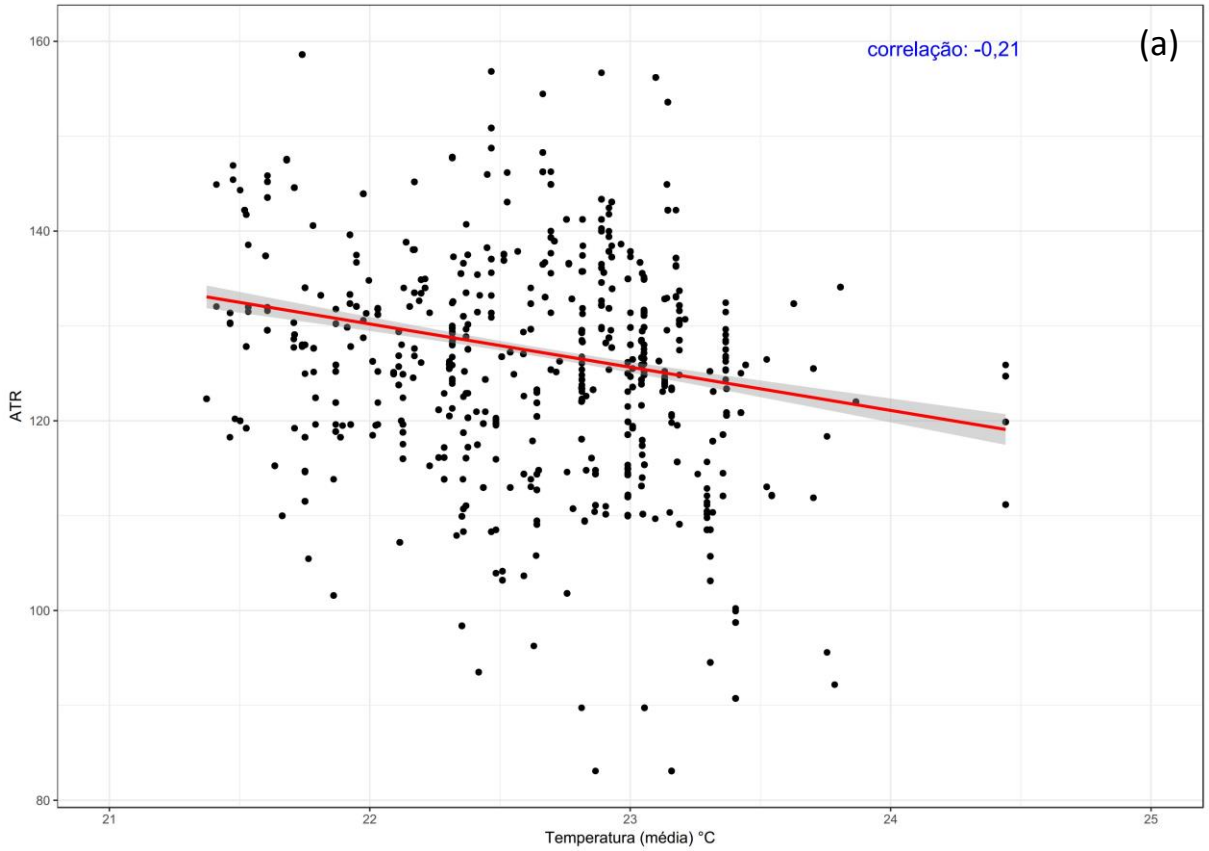


Figura 43 - Gráfico de dispersão entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Precipitação acum. até o corte (mm).



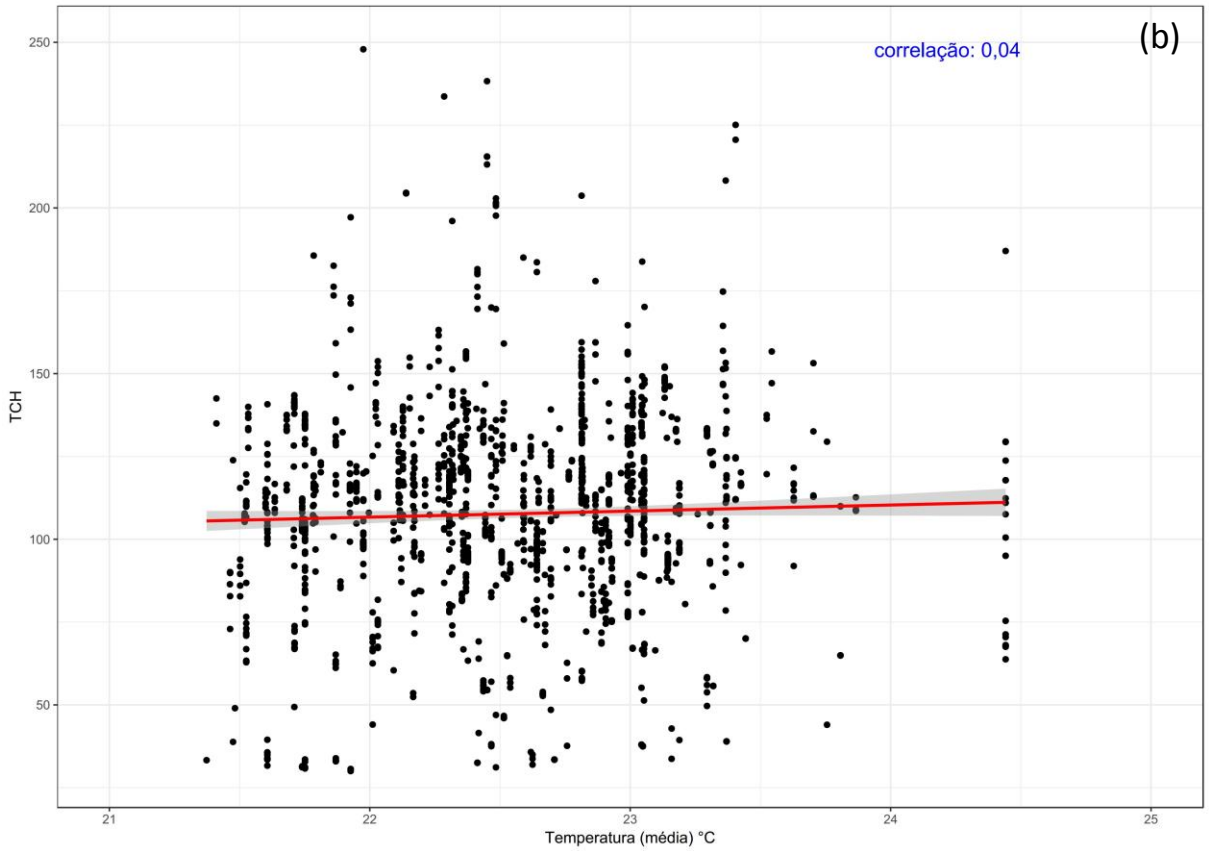
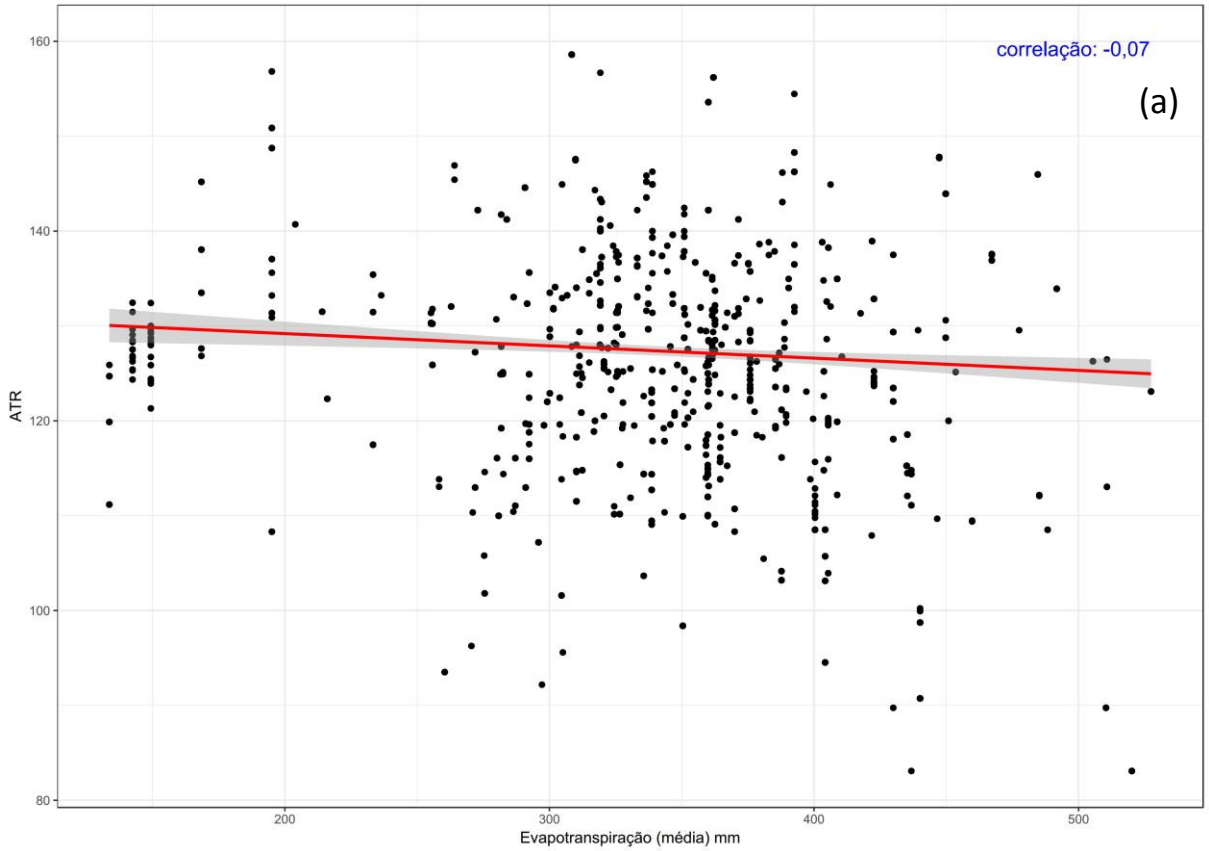
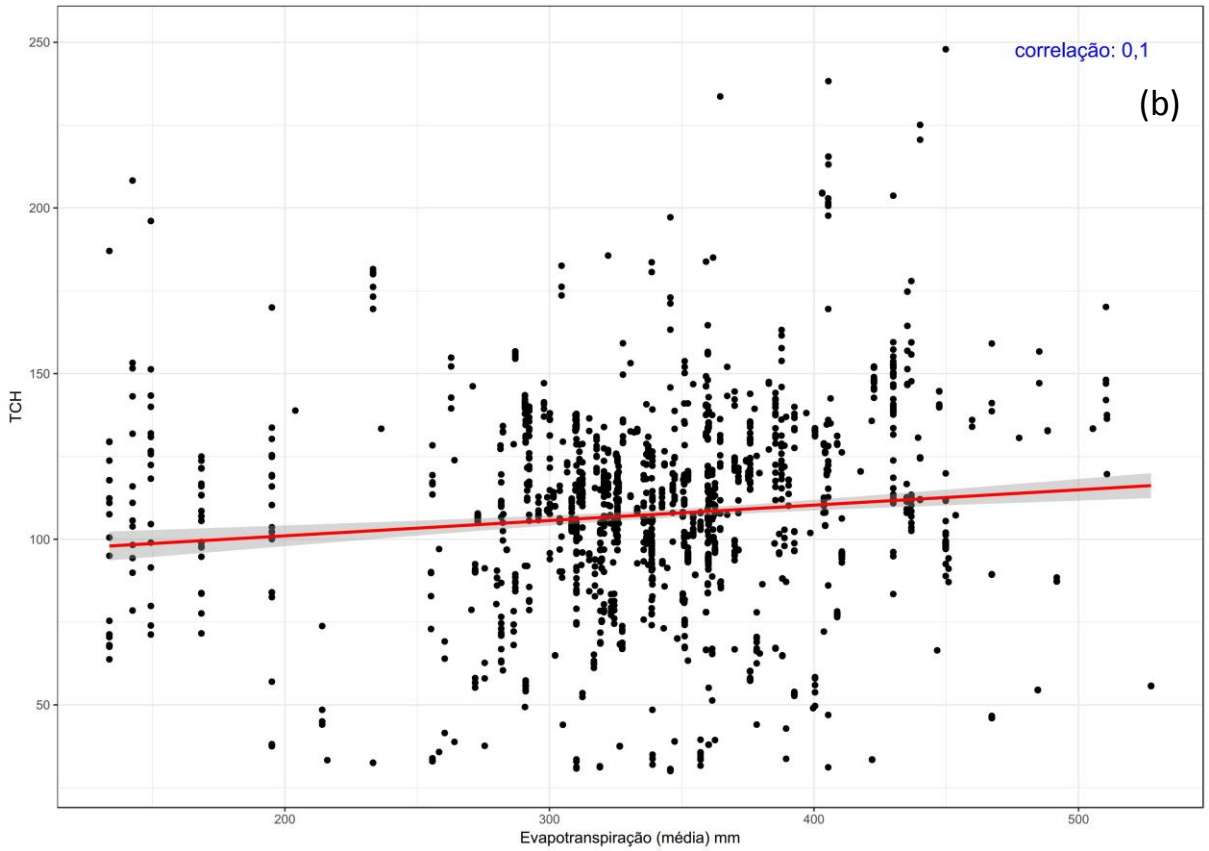


Figura 44 - Gráfico de dispersão entre as variáveis dependentes ATR (a) e TCH (b) com a variável independente Evapotranspiração média (mm).





Apêndice B – Script R utilizados para gerar a estatística descritiva, árvores de decisão e floresta aleatória para as variáveis respostas TCH, ATR e TCH*ATR.

```
#####< Instalando pacotes necessários para as análises >#######
#install.packages("rattle", repos="http://rattle.togaware.com")
#devtools::install_github("JackStat/ModelMetrics")
#install.packages("rpart")
#install.packages("rpart.plot")
#install.packages("party")
#install.packages("caret")
#install.packages("randomForest")
#install.packages("RColorBrewer")
#install.packages("Rcpp")
#install.packages("partykit")
#install.packages("magrittr")
#install.packages("data.table")
#install.packages("agricolae")
#install.packages("readxl")


#########################################################################
#########
#) Definindo diretório e carregando pacotes
#########################################################################
#########

getwd()
setwd("C:/Germano_Analises_Estatisticas")

library(rattle)         #Função: rattle()
library(rpart)          #Função: rpart()
library(rpart.plot)
library(party)          #Função ctree(), cforest()
library(caret)          #Função: cforestStats()
library(randomForest)#Funções: randomForest()
library(RColorBrewer)
library(Rcpp)
library(partykit)
library(readxl)   #Função: read_excel()
library(data.table) #Função: setDT
library(agricolae)

#Lendo Banco de dados
DATA1<-read_excel("DATA1.xls")

#Redefinindo e ordenando fatores
DATA1$TCH_classe<-factor(DATA1$TCH_classe)
levels(DATA1$TCH_classe)
DATA1$TCH_classe<- factor(DATA1$TCH_classe,ordered = T,levels= c("Baixa",
"Baixa-Média", "Média-Alta", "Alta"))
#
DATA1$ATR_classe<-factor(DATA1$ATR_classe)
levels(DATA1$ATR_classe)
DATA1$ATR_classe<- factor(DATA1$ATR_classe,ordered = T,
                          levels= c("Baixa", "Baixa-Média", "Média-Alta",
"Alta"))
#
DATA1$TCH_x_ATR_classe<-factor(DATA1$TCH_x_ATR_classe)
levels(DATA1$TCH_x_ATR_classe)
DATA1$TCH_x_ATR_classe<- factor(DATA1$TCH_x_ATR_classe,ordered = T,
```

```
                                    levels= c("Baixa", "Baixa-Média", "Média-
Alta", "Alta"))

#Eliminando valores maiores que 25 meses de colheita.
DATA1$Idade_Meses[DATA1$Idade_Meses>25]<-NA

##############################################################################
#########
#) Histogramas e teste de normalidade (Shapiro-Wilk)
##############################################################################
#########
# Histograma para TCH
attach(DATA1)
ggplot(DATA1, aes(x=TCH)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead
of count on y-axis
                 binwidth=15,
                 colour="black", fill="white") +
  scale_x_continuous(30:250) +
  geom_density(alpha=.2, fill="darkgreen")  # Overlay with transparent
density plot
summary(TCH)
detach(DATA1)

# Histograma para ATR
attach(DATA1)
ggplot(DATA1, aes(x=ATR)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead
of count on y-axis
                 binwidth=5,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="darkgreen")  # Overlay with transparent
density plot
summary(ATR)
detach(DATA1)

#Histograma para TCH vs ATR
attach(DATA1)
ggplot(DATA1, aes(x=TCH_x_ATR)) +
  geom_histogram(aes(y=..density..),      # Histogram with density instead
of count on y-axis
                 #binwidth=20,
                 colour="black", fill="white") +
  geom_density(alpha=.2, fill="darkgreen")  # Overlay with transparent
density plot
summary(TCH_x_ATR)
detach(DATA1)

#Teste de normalidade para os e resíduos de TCH, ATR e TCH*ATR
shapiro.test(DATA1$TCH)
shapiro.test(DATA1$ATR)
shapiro.test(DATA1$TCH_x_ATR)

#Criando ANOVA para teste de normalidade dos resíduos: TCH
TEST<-aov(TCH ~ as.factor(T_Propriedade) + as.factor(Estagio) +
            as.factor(Variedade) + as.factor(Operacao) +
            as.factor(Maturador) +  as.factor(Tipo_Plantio) +
            as.factor(Trimestre_Safra) + Fotoperiodo + Et_Media +
            Temperatura_Media + Precipitacao_acum_corte  +
            Idade_Meses, data=DATA1)
summary(TEST)
```

```
plot(resid(TEST))
shapiro.test(residuals(TEST))
hist(resid(TEST))
plot(TEST)
# Cálculo de Box-Cox para verificar se a variável resposta necessita ser
transformada
BoxCoxTrans(DATA1$TCH, na.rm = T)

#Criando ANOVA para teste de normalidade dos resíduos: ATR
TEST<-aov(ATR ~ as.factor(T_Propriedade) + as.factor(Estagio) +
            as.factor(Variedade) + as.factor(Operacao) +
            as.factor(Maturador) +  as.factor(Tipo_Plantio) +
            as.factor(Trimestre_Safra) + Fotoperiodo + Et_Media +
            Temperatura_Media + Precipitacao_acum_corte  +
            Idade_Meses,  data=DATA1 )
summary(TEST)
plot(resid(TEST))
shapiro.test(residuals(TEST))
hist(resid(TEST))
plot(TEST)
# Cálculo de Box-Cox para verificar se a variável resposta necessita ser
transformada
BoxCoxTrans(DATA1$ATR, na.rm = T)

#Criando ANOVA para teste de normalidade dos resíduos: ATR
TEST<-aov(TCH_x_ATR ~ as.factor(T_Propriedade) + as.factor(Estagio) +
            as.factor(Variedade) + as.factor(Operacao) +
            as.factor(Maturador) +  as.factor(Tipo_Plantio) +
            as.factor(Trimestre_Safra) + Fotoperiodo + Et_Media +
            Temperatura_Media + Precipitacao_acum_corte  +
            Idade_Meses,  data=DATA1 )
summary(TEST)
plot(resid(TEST))
shapiro.test(residuals(TEST))
hist(resid(TEST))
plot(TEST)
# Cálculo de Box-Cox para verificar se a variável resposta necessita ser
transformada
BoxCoxTrans(DATA1$TCH_x_ATR, na.rm = T)

#########################################################################
#########
#) Criando Boxplots para variáveis indep. compostas de classes vs (TCH e
ATR)
#########################################################################
#########
### TCH
attach(DATA1)
TCH<-round(as.numeric(TCH), 0)
#Criando boxplot TCH vs Tipo de propriedade
BOX<-boxplot(TCH ~ T_Propriedade, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
```

```
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[T_Propriedade=="Arrendamento"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(TCH[T_Propriedade=="Fornecedor parceiro"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(TCH[T_Propriedade=="Parceria agrícola"]),INF.BOX3,
SUP.BOX3)
sevennum4<-c(fivenum(TCH[T_Propriedade=="Própria"]),INF.BOX4, SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "TCH", ylab="Tipo de propriedade", font.main=2)

DATA1<-as.data.table(DATA1)
class(Estagio)
DATA1<-setDT(DATA1)[Estagio=='Decimo corte', Estagio := '10C']
DATA1<-setDT(DATA1)[Estagio=='Decimo corte bis', Estagio := '10Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo primeiro cort', Estagio := '11C']
DATA1<-setDT(DATA1)[Estagio=='Decimo Pri corte bis', Estagio := '11Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo segundo corte', Estagio := '12C']
DATA1<-setDT(DATA1)[Estagio=='Decimo Seg corte bis', Estagio := '12Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo terceiro cort', Estagio := '13C']
DATA1<-setDT(DATA1)[Estagio=='Decimo quarto corte', Estagio := '14C']
DATA1<-setDT(DATA1)[Estagio=='Decimo quinto corte', Estagio := '15C']
DATA1<-setDT(DATA1)[Estagio=='Primeiro corte 12M', Estagio := '1C12']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis 12M', Estagio := '1C12b']
DATA1<-setDT(DATA1)[Estagio=='Primeiro corte 18M', Estagio := '1C18']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis 18M', Estagio := '1C18b']
DATA1<-setDT(DATA1)[Estagio=='Primeiro Corte Inver', Estagio := '1CIn']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis Inver', Estagio := '1CInb']
DATA1<-setDT(DATA1)[Estagio=='Segundo corte', Estagio := '2C']
DATA1<-setDT(DATA1)[Estagio=='Segundo corte bis', Estagio := '2Cb']
DATA1<-setDT(DATA1)[Estagio=='Terceiro corte', Estagio := '3C']
DATA1<-setDT(DATA1)[Estagio=='Terceiro corte bis', Estagio := '3Cb']
DATA1<-setDT(DATA1)[Estagio=='Quarto corte', Estagio := '4C']
DATA1<-setDT(DATA1)[Estagio=='Quarto corte bis', Estagio := '4Cb']
DATA1<-setDT(DATA1)[Estagio=='Quinto corte', Estagio := '5C']
DATA1<-setDT(DATA1)[Estagio=='Quinto corte bis', Estagio := '5Cb']
DATA1<-setDT(DATA1)[Estagio=='Sexto corte', Estagio := '6C']
DATA1<-setDT(DATA1)[Estagio=='Sexto corte bis', Estagio := '6Cb']
DATA1<-setDT(DATA1)[Estagio=='Setimo corte', Estagio := '7C']
DATA1<-setDT(DATA1)[Estagio=='Setimo corte bis', Estagio := '7Cb']
DATA1<-setDT(DATA1)[Estagio=='Oitavo corte', Estagio := '8C']
DATA1<-setDT(DATA1)[Estagio=='Oitavo corte bis', Estagio := '8Cb']
DATA1<-setDT(DATA1)[Estagio=='Nono corte', Estagio := '9C']
DATA1<-setDT(DATA1)[Estagio=='Nono corte bis', Estagio := '9Cb']

Estagio<-as.factor(as.character(DATA1$Estagio))
#Criando boxplot TCH vs Estagio
BOX<-boxplot(TCH ~ Estagio, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
```

```
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
INF.BOX5<-BOX$stats[1,5] #lower.box
SUP.BOX5<-BOX$stats[5,5] #upper.box
INF.BOX6<-BOX$stats[1,6] #lower.box
SUP.BOX6<-BOX$stats[5,6] #upper.box
INF.BOX7<-BOX$stats[1,7] #lower.box
SUP.BOX7<-BOX$stats[5,7] #upper.box
INF.BOX8<-BOX$stats[1,8] #lower.box
SUP.BOX8<-BOX$stats[5,8] #upper.box
INF.BOX9<-BOX$stats[1,9] #lower.box
SUP.BOX9<-BOX$stats[5,9] #upper.box
INF.BOX10<-BOX$stats[1,10] #lower.box
SUP.BOX10<-BOX$stats[5,10] #upper.box
INF.BOX11<-BOX$stats[1,11] #lower.box
SUP.BOX11<-BOX$stats[5,11] #upper.box
INF.BOX12<-BOX$stats[1,12] #lower.box
SUP.BOX12<-BOX$stats[5,12] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Estagio=="1C12"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Estagio=="1C12b"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(TCH[Estagio=="1C18"]),INF.BOX3, SUP.BOX3)
sevennum4<-c(fivenum(TCH[Estagio=="1CIn"]),INF.BOX4, SUP.BOX4)
sevennum5<-c(fivenum(TCH[Estagio=="2C"]),INF.BOX5, SUP.BOX5)
sevennum6<-c(fivenum(TCH[Estagio=="3C"]),INF.BOX6, SUP.BOX6)
sevennum7<-c(fivenum(TCH[Estagio=="4C"]),INF.BOX7, SUP.BOX7)
sevennum8<-c(fivenum(TCH[Estagio=="5C"]),INF.BOX8, SUP.BOX8)
sevennum9<-c(fivenum(TCH[Estagio=="6C"]),INF.BOX9, SUP.BOX9)
sevennum10<-c(fivenum(TCH[Estagio=="7C"]),INF.BOX10, SUP.BOX10)
sevennum11<-c(fivenum(TCH[Estagio=="8C"]),INF.BOX11, SUP.BOX11)
sevennum12<-c(fivenum(TCH[Estagio=="9C"]),INF.BOX12, SUP.BOX12)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
text(x=sevennum5,labels = sevennum5, y=5.5)
text(x=sevennum6,labels = sevennum6, y=6.5)
text(x=sevennum7,labels = sevennum7, y=7.5)
text(x=sevennum8,labels = sevennum8, y=8.5)
text(x=sevennum9,labels = sevennum9, y=9.5)
text(x=sevennum10,labels = sevennum10, y=10.5)
text(x=sevennum11,labels = sevennum11, y=11.5)
text(x=sevennum12,labels = sevennum12, y=12.5)
title(xlab = "TCH", ylab="Estágio", font.main=2)

levels(Variedade)
#Criando boxplot TCH vs variedades
BOX<-boxplot(TCH ~ Variedade, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Variedade=="CTC4"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Variedade=="RB966928"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
```

```
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "TCH", ylab="Variedade", font.main=2)

#Criando boxplot TCH vs Ambiente de produção
BOX<-boxplot(TCH ~ Ambiente, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Ambiente=="A"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Ambiente=="B"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(TCH[Ambiente=="C"]),INF.BOX3, SUP.BOX3)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
title(xlab = "TCH", ylab="Ambiente", font.main=2)

#Criando boxplot TCH vs Tipo de cana
BOX<-boxplot(TCH ~ Tipo, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Tipo=="Cana Planta"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Tipo=="Cana Soca"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "TCH", ylab="Tipo", font.main=2)

#Criando boxplot TCH vs Operação
BOX<-boxplot(TCH ~ Operacao, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Operacao=="Adubação/aleiramento"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(TCH[Operacao=="Aplic de vinhaça-cam/mot"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(TCH[Operacao=="Aplicação de vinhaça"]),INF.BOX3,
```

```
SUP.BOX3)
sevennum4<-c(fivenum(TCH[Operacao=="Cultivo e adubação"]),INF.BOX4,
SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "TCH", ylab="Operação", font.main=2)

#Criando boxplot TCH vs Trimestre Safra
BOX<-boxplot(TCH ~ Trimestre_Safra, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Trimestre_Safra=="T1"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Trimestre_Safra=="T2"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(TCH[Trimestre_Safra=="T3"]),INF.BOX3, SUP.BOX3)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
title(xlab = "TCH", ylab="Trimestre Safra", font.main=2)

#Criando boxplot TCH vs Tipo de plantio
BOX<-boxplot(TCH ~ Tipo_Plantio, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Tipo_Plantio=="Manual com Torta"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(TCH[Tipo_Plantio=="Manual Convencional"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(TCH[Tipo_Plantio=="Mecânico"]),INF.BOX3, SUP.BOX3)
sevennum4<-c(fivenum(TCH[Tipo_Plantio=="Semimecanizado"]),INF.BOX4,
SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "TCH", ylab="Tipo de plantio", font.main=2)

#Criando boxplot TCH vs Maturador
BOX<-boxplot(TCH ~ Maturador, horizontal = T, axes = T, staplewex =
```

```
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(TCH[Maturador=="Não"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(TCH[Maturador=="Sim"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "TCH", ylab="Maturador", font.main=2)

### ATR
attach(DATA1)
ATR<-round(as.numeric(ATR), 0)
#Criando boxplot ATR vs Tipo de propriedade
BOX<-boxplot(ATR ~ T_Propriedade, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[T_Propriedade=="Arrendamento"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(ATR[T_Propriedade=="Fornecedor parceiro"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(ATR[T_Propriedade=="Parceria agrícola"]),INF.BOX3,
SUP.BOX3)
sevennum4<-c(fivenum(ATR[T_Propriedade=="Própria"]),INF.BOX4, SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "ATR", ylab="Tipo de propriedade", font.main=2)

DATA1<-as.data.table(DATA1)
class(Estagio)
DATA1<-setDT(DATA1)[Estagio=='Decimo corte', Estagio := '10C']
DATA1<-setDT(DATA1)[Estagio=='Decimo corte bis', Estagio := '10Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo primeiro cort', Estagio := '11C']
DATA1<-setDT(DATA1)[Estagio=='Decimo Pri corte bis', Estagio := '11Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo segundo corte', Estagio := '12C']
DATA1<-setDT(DATA1)[Estagio=='Decimo Seg corte bis', Estagio := '12Cb']
DATA1<-setDT(DATA1)[Estagio=='Decimo terceiro cort', Estagio := '13C']
DATA1<-setDT(DATA1)[Estagio=='Decimo quarto corte', Estagio := '14C']
DATA1<-setDT(DATA1)[Estagio=='Decimo quinto corte', Estagio := '15C']
DATA1<-setDT(DATA1)[Estagio=='Primeiro corte 12M', Estagio := '1C12']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis 12M', Estagio := '1C12b']
DATA1<-setDT(DATA1)[Estagio=='Primeiro corte 18M', Estagio := '1C18']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis 18M', Estagio := '1C18b']
```

```
DATA1<-setDT(DATA1)[Estagio=='Primeiro Corte Inver', Estagio := '1CIn']
DATA1<-setDT(DATA1)[Estagio=='Pri corte bis Inver', Estagio := '1CInb']
DATA1<-setDT(DATA1)[Estagio=='Segundo corte', Estagio := '2C']
DATA1<-setDT(DATA1)[Estagio=='Segundo corte bis', Estagio := '2Cb']
DATA1<-setDT(DATA1)[Estagio=='Terceiro corte', Estagio := '3C']
DATA1<-setDT(DATA1)[Estagio=='Terceiro corte bis', Estagio := '3Cb']
DATA1<-setDT(DATA1)[Estagio=='Quarto corte', Estagio := '4C']
DATA1<-setDT(DATA1)[Estagio=='Quarto corte bis', Estagio := '4Cb']
DATA1<-setDT(DATA1)[Estagio=='Quinto corte', Estagio := '5C']
DATA1<-setDT(DATA1)[Estagio=='Quinto corte bis', Estagio := '5Cb']
DATA1<-setDT(DATA1)[Estagio=='Sexto corte', Estagio := '6C']
DATA1<-setDT(DATA1)[Estagio=='Sexto corte bis', Estagio := '6Cb']
DATA1<-setDT(DATA1)[Estagio=='Setimo corte', Estagio := '7C']
DATA1<-setDT(DATA1)[Estagio=='Setimo corte bis', Estagio := '7Cb']
DATA1<-setDT(DATA1)[Estagio=='Oitavo corte', Estagio := '8C']
DATA1<-setDT(DATA1)[Estagio=='Oitavo corte bis', Estagio := '8Cb']
DATA1<-setDT(DATA1)[Estagio=='Nono corte', Estagio := '9C']
DATA1<-setDT(DATA1)[Estagio=='Nono corte bis', Estagio := '9Cb']

attach(DATA1)
Estagio<-as.factor(as.character(DATA1$Estagio))
levels(Estagio)
#Criando boxplot ATR vs Estagio
BOX<-boxplot(ATR ~ Estagio, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
INF.BOX5<-BOX$stats[1,5] #lower.box
SUP.BOX5<-BOX$stats[5,5] #upper.box
INF.BOX6<-BOX$stats[1,6] #lower.box
SUP.BOX6<-BOX$stats[5,6] #upper.box
INF.BOX7<-BOX$stats[1,7] #lower.box
SUP.BOX7<-BOX$stats[5,7] #upper.box
INF.BOX8<-BOX$stats[1,8] #lower.box
SUP.BOX8<-BOX$stats[5,8] #upper.box
INF.BOX9<-BOX$stats[1,9] #lower.box
SUP.BOX9<-BOX$stats[5,9] #upper.box
INF.BOX10<-BOX$stats[1,10] #lower.box
SUP.BOX10<-BOX$stats[5,10] #upper.box
INF.BOX11<-BOX$stats[1,11] #lower.box
SUP.BOX11<-BOX$stats[5,11] #upper.box
INF.BOX12<-BOX$stats[1,12] #lower.box
SUP.BOX12<-BOX$stats[5,12] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Estagio=="1C12"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Estagio=="1C12b"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(ATR[Estagio=="1C18"]),INF.BOX3, SUP.BOX3)
sevennum4<-c(fivenum(ATR[Estagio=="1CIn"]),INF.BOX4, SUP.BOX4)
sevennum5<-c(fivenum(ATR[Estagio=="2C"]),INF.BOX5, SUP.BOX5)
sevennum6<-c(fivenum(ATR[Estagio=="3C"]),INF.BOX6, SUP.BOX6)
sevennum7<-c(fivenum(ATR[Estagio=="4C"]),INF.BOX7, SUP.BOX7)
sevennum8<-c(fivenum(ATR[Estagio=="5C"]),INF.BOX8, SUP.BOX8)
sevennum9<-c(fivenum(ATR[Estagio=="6C"]),INF.BOX9, SUP.BOX9)
```

```
sevennum10<-c(fivenum(ATR[Estagio=="7C"]),INF.BOX10, SUP.BOX10)
sevennum11<-c(fivenum(ATR[Estagio=="8C"]),INF.BOX11, SUP.BOX11)
sevennum12<-c(fivenum(ATR[Estagio=="9C"]),INF.BOX12, SUP.BOX12)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
text(x=sevennum5,labels = sevennum5, y=5.5)
text(x=sevennum6,labels = sevennum6, y=6.5)
text(x=sevennum7,labels = sevennum7, y=7.5)
text(x=sevennum8,labels = sevennum8, y=8.5)
text(x=sevennum9,labels = sevennum9, y=9.5)
text(x=sevennum10,labels = sevennum10, y=10.5)
text(x=sevennum11,labels = sevennum11, y=11.5)
text(x=sevennum12,labels = sevennum12, y=12.5)
title(xlab = "ATR", ylab="Estágio", font.main=2)

levels(Variedade)
#Criando boxplot ATR vs variedades
BOX<-boxplot(ATR ~ Variedade, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Variedade=="CTC4"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Variedade=="RB966928"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "ATR", ylab="Variedade", font.main=2)

#Criando boxplot ATR vs Ambiente de produção
BOX<-boxplot(ATR ~ Ambiente, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Ambiente=="A"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Ambiente=="B"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(ATR[Ambiente=="C"]),INF.BOX3, SUP.BOX3)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
title(xlab = "ATR", ylab="Ambiente", font.main=2)

#Criando boxplot ATR vs Tipo de cana
BOX<-boxplot(ATR ~ Tipo, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
```

```
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Tipo=="Cana Planta"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Tipo=="Cana Soca"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "ATR", ylab="Tipo", font.main=2)

#Criando boxplot ATR vs Operação
BOX<-boxplot(ATR ~ Operacao, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Operacao=="Adubação/aleiramento"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(ATR[Operacao=="Aplic de vinhaça-cam/mot"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(ATR[Operacao=="Aplicação de vinhaça"]),INF.BOX3,
SUP.BOX3)
sevennum4<-c(fivenum(ATR[Operacao=="Cultivo e adubação"]),INF.BOX4,
SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "ATR", ylab="Operação", font.main=2)

#Criando boxplot ATR vs Trimestre Safra
BOX<-boxplot(ATR ~ Trimestre_Safra, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Trimestre_Safra=="T1"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Trimestre_Safra=="T2"]),INF.BOX2, SUP.BOX2)
sevennum3<-c(fivenum(ATR[Trimestre_Safra=="T3"]),INF.BOX3, SUP.BOX3)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
```

```
title(xlab = "ATR", ylab="Trimestre Safra", font.main=2)

#Criando boxplot ATR vs Tipo de plantio
BOX<-boxplot(ATR ~ Tipo_Plantio, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
INF.BOX3<-BOX$stats[1,3] #lower.box
SUP.BOX3<-BOX$stats[5,3] #upper.box
INF.BOX4<-BOX$stats[1,4] #lower.box
SUP.BOX4<-BOX$stats[5,4] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Tipo_Plantio=="Manual com Torta"]),INF.BOX1,
SUP.BOX1)
sevennum2<-c(fivenum(ATR[Tipo_Plantio=="Manual Convencional"]),INF.BOX2,
SUP.BOX2)
sevennum3<-c(fivenum(ATR[Tipo_Plantio=="Mecânico"]),INF.BOX3, SUP.BOX3)
sevennum4<-c(fivenum(ATR[Tipo_Plantio=="Semimecanizado"]),INF.BOX4,
SUP.BOX4)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
text(x=sevennum3,labels = sevennum3, y=3.5)
text(x=sevennum4,labels = sevennum4, y=4.5)
title(xlab = "ATR", ylab="Tipo de plantio", font.main=2)

#Criando boxplot ATR vs Maturador
BOX<-boxplot(ATR ~ Maturador, horizontal = T, axes = T, staplewex =
1,col="grey80")
#coletando valores de limites inferiores e superiores do boxplot
INF.BOX1<-BOX$stats[1,1] #lower.box
SUP.BOX1<-BOX$stats[5,1] #upper.box
INF.BOX2<-BOX$stats[1,2] #lower.box
SUP.BOX2<-BOX$stats[5,2] #upper.box
#criando vetor com as setes estatisticas principais (min, limite inf.,
Quartil1, mediana, Quartil3, limite sup., max)
sevennum1<-c(fivenum(ATR[Maturador=="Não"]),INF.BOX1, SUP.BOX1)
sevennum2<-c(fivenum(ATR[Maturador=="Sim"]),INF.BOX2, SUP.BOX2)
#plotando as setes estatisticas no boxplot
text(x=sevennum1,labels = sevennum1, y=1.5)
text(x=sevennum2,labels = sevennum2, y=2.5)
title(xlab = "ATR", ylab="Maturador", font.main=2)

#########################################################################
#########
#) Criando Gráficos de dispersão para variáveis independentes numéricas vs
TCH e ATR
#########################################################################
#########
options(OutDec = ",")
#TCH
#Gráfico de dispersão e correlação: Idade_Meses vs TCH
DATA1<-as.data.frame(DATA1)
CORR<-correlation(DATA1[,c("Idade_Meses", "TCH")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Idade_Meses, y = TCH)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
```

```
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 21, y = 250), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Idade (meses)", limits = c(8,21), breaks =
c(8:21))+theme_bw()

#Gráfico de dispersão e correlação:Precipitação acum. até o corte (mm) vs
TCH
CORR<-correlation(DATA1[,c("Precipitacao_acum_corte", "TCH")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Precipitacao_acum_corte, y = TCH)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 3000, y = 250), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Precipitação acum. até o corte (mm)")+theme_bw()

#Gráfico de dispersão e correlação: Temperatura (média) °C vs TCH
CORR<-correlation(DATA1[,c("Temperatura_Media", "TCH")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Temperatura_Media, y = TCH)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 24.5, y = 250), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Temperatura (média) °C", limits = c(21.0,25), breaks =
c(21:25)) +theme_bw()
summary(DATA1$Temperatura_Media)

#Gráfico de dispersão e correlação: Evapotranspiração (média) mm vs TCH
CORR<-correlation(DATA1[,c("Et_Media", "TCH")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Et_Media, y = TCH)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 527, y = 250), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Evapotranspiração (média) mm")+theme_bw()
summary(DATA1$Et_Media)

#Gráfico de dispersão e correlação: Idade_Meses vs TCH
CORR<-correlation(DATA1[,c("Fotoperíodo (horas)", "TCH")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Fotoperiodo, y = TCH)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 12.4, y = 250), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Fotoperíodo (horas)", limits =
c(11.69,12.40))+theme_bw()
summary(DATA1$Fotoperiodo)

##ATR
#Gráfico de dispersão e correlação: Idade (meses) vs ATR
DATA1<-as.data.frame(DATA1)
CORR<-correlation(DATA1[,c("Idade_Meses", "ATR")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Idade_Meses, y = ATR)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 21, y = 160), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Idade (meses)", limits = c(8,21), breaks =
c(8:21))+theme_bw()

#Gráfico de dispersão e correlação: Precipitação acum. até o corte (mm) vs
ATR
```

```
CORR<-correlation(DATA1[,c("Precipitacao_acum_corte", "ATR")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Precipitacao_acum_corte, y = ATR)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 3000, y = 160), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Precipitação acum. até o corte (mm)")+theme_bw()

#Gráfico de dispersão e correlação: Temperatura (média) °C vs ATR
CORR<-correlation(DATA1[,c("Temperatura_Media", "ATR")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Temperatura_Media, y = ATR)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 24.5, y = 160), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Temperatura (média) °C", limits = c(21.0,25), breaks =
c(21:25))+theme_bw()
summary(DATA1$Temperatura_Media)

#Gráfico de dispersão e correlação: Evapotranspiração (média) mm vs ATR
CORR<-correlation(DATA1[,c("Et_Media", "ATR")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Et_Media, y = ATR)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 527, y = 160), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Evapotranspiração (média) mm")+theme_bw()
summary(DATA1$Et_Media)

#Gráfico de dispersão e correlação: Fotoperíodo (horas) vs ATR
CORR<-correlation(DATA1[,c("Fotoperiodo", "ATR")])
caption <- paste("correlação:",CORR$correlation[1,2])
ggplot(DATA1, aes(x = Fotoperiodo, y = ATR)) +
  geom_point() +  stat_smooth(method = "lm", col = "red") +
geom_text(aes(x, y, label = caption),col="blue",
data = data.frame(x = 12.4, y = 160), hjust = 1, vjust = 1, size = 5) +
scale_x_continuous("Fotoperíodo (horas)", limits = c(11.69,12.40))
+theme_bw()
summary(DATA1$Fotoperiodo)

#########################################################################
#########
#) Análises: Árvores de decisão e florestas aleatórias
#)Fazendo estudo de TCH como variável compostas de classes (fator)
# (Desconsiderando a variável SAFRA)
#########################################################################
#########
setwd("C:/Germano_Analises_Estatisticas")
setwd("E:/Análises Estatísticas/Germano")
crs<-NA
crs$dataset<-read.csv("DATA1.csv", sep=",", na.strings=c(".", "NA", "",
"?"), strip.white=TRUE, encoding="ISO-8859-1")
sapply(crs$dataset, levels)

#TCH categórica
#Construindo bancos de dados, treinamento, validação e teste
crv$seed<-123;set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset) # 2275 observações
crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7*crs$nobs) # 1592
observações
crs$validate <- sample(setdiff(seq_len(nrow(crs$dataset)), crs$train),
```

```
0.15*crs$nobs) # 341 observações
crs$test <- setdiff(setdiff(seq_len(nrow(crs$dataset)), crs$train),
crs$validate) # 342 observações

# Selecionando variáveis
crs$input <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Idade_Meses", "Trimestre_Safra", "Operacao", "Tipo_Plantio",
     "Maturador", "Precipitacao_acum_corte", "Temperatura_Media",
"Et_Media",
     "Fotoperiodo")

crs$numeric <- c("Idade_Meses", "Precipitacao_acum_corte",
"Temperatura_Media", "Et_Media",
     "Fotoperiodo")

crs$categoric <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Trimestre_Safra", "Operacao", "Tipo_Plantio", "Maturador")

crs$target  <- "TCH_classe"
crs$risk    <- NULL
crs$ident   <- NULL
crs$ignore  <- c("TCH", "ATR", "TCH_x_ATR", "ATR_classe",
"TCH_x_ATR_classe", "Safra")
crs$weights <- NULL

#TCH categórica: Árvore de decisão
library(rpart, quietly=TRUE) #carregando pacote da análise de árvores de
decisão

crv$seed<-123;set.seed(crv$seed) #definindo semente: 123
# Build the modelo de árvore de decisão.
crs$rpart <- rpart(TCH_classe ~ .,
    data=crs$dataset[crs$train, c(crs$input, crs$target)],
    method="class",
    parms=list(split="information"),
      control=rpart.control(cp=0.0027,
        usesurrogate=0,
        maxsurrogate=0))

#Gerando um texto visual da árvore
print(crs$rpart)
printcp(crs$rpart)
TREE_SUMMARY<-summary(crs$rpart)
barchart(TREE_SUMMARY$variable.importance) #Importancia das variáveis na
árvore

#Construindo modelo de floresta aleatória usando abordagem tradicional
crv$seed<-9857365;set.seed(crv$seed)
sapply(crs$dataset, class)
crs$rf <- randomForest::randomForest(TCH_classe ~ .,
  data=crs$dataset[crs$sample, c(crs$input, crs$target)],
  ntree=500,
  mtry=3,
  importance=TRUE,
  na.action=randomForest::na.roughfix,
  replace=FALSE)

#Gerando saida textual do modelo de floresta aleatória.
crs$rf

# Mostrando árvore de n° 500.
```

```
printRandomForests(crs$rf, 500)

#Listando a importancia das variáveis
rn <- round(randomForest::importance(crs$rf), 2)
rn[order(rn[,3], decreasing=TRUE),]

#Plotando a importancia relativa das variaveis.
p <- ggVarImp(crs$rf,main="TCH:Importancia relativa das variaveis (Floresta
aleatória) ")
p

#########################################################################
#########
#) Predição com os modelos de árvores de decisão e floresta aleatória
# (Desconsiderando a variável SAFRA)
#########################################################################
#########

#Predição em TCH: Árvore de decisão
#Avaliando o desempenho do modelo no banco de dados de treinamento
#Gerando matriz de erro para o modelo de árvore de decisão
#Obtendo a resposta com o modelo de árvore de decisão
crs$pr <- predict(
  crs$rpart, newdata=crs$dataset[crs$sample, c(crs$input, crs$target)],
type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$sample, c(crs$input, crs$target)]$TCH_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$sample, c(crs$input,
crs$target)]$TCH_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

#Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#Predição em TCH: Floresta aleatória
#####################################
# Gerando matriz de erro para o modelo de floresta aleatória
# Obtendo a resposta com o modelo de floresta aleatória
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$sample,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$TCH_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$TCH_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
```

```
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

#Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#Avaliando o desempenho do modelo no banco de dados de validação
# Gerando um matriz de erro para árvore de decisão.
# Obtendo a resposta com o modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$validate, c(crs$input,
crs$target)], type="class")

# Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$validate, c(crs$input, crs$target)]$TCH_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

# Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$validate, c(crs$input,
crs$target)]$TCH_classe, crs$pr)
round(per, 2)

# Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando um matriz de erro para o modelo de floresta aleatória.
# Obtendo a resposta com o modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$validate,
c(crs$input, crs$target)]))

# Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$TCH_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

# Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$TCH_classe, crs$pr)
round(per, 2)

# Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#=========================================================
# Avaliando o desempenho do modelo de teste.
# Gerando um matriz de erro para modelo de árvore de decisão.
# Obtendo a resposta com o modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$test, c(crs$input,
crs$target)], type="class")

# Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$test, c(crs$input, crs$target)]$TCH_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))
```

```
# Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$test, c(crs$input,
crs$target)]$TCH_classe, crs$pr)
round(per, 2)

# Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando um matriz de erro para o modelo de floresta aleatória.
# Obtendo a resposta com o modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$test,
c(crs$input, crs$target)]))

# Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$test, c(crs$input, crs$target)])$TCH_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predicta"))

# Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$test, c(crs$input,
crs$target)])$TCH_classe, crs$pr)
round(per, 2)

# Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Salvando o projeto (variável crs) no HD do computador.
save(crs, file="DATA1.TCH.rattle", compress=TRUE)
########################################################################
#########
#) Análises: Árvores de decisão e florestas aleatórias
#)Fazendo estudo de ATR como variável compostas de classes (fator)
# (Desconsiderando a variável SAFRA)
########################################################################
#########

# construindo os bancos de dados de treinamento, validação e teste
crv$seed<-42;set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset) # 2275 observations
crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7*crs$nobs) # 1592
observations
crs$validate <- sample(setdiff(seq_len(nrow(crs$dataset)), crs$train),
0.15*crs$nobs) # 341 observations
crs$test <- setdiff(setdiff(seq_len(nrow(crs$dataset)), crs$train),
crs$validate) # 342 observations

#Construindo modelo de árvore aleatória usando abordagem tradicional
crs$input <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Idade_Meses", "Trimestre_Safra", "Operacao",
     "Tipo_Plantio", "Maturador", "Precipitacao_acum_corte",
"Temperatura_Media",
     "Et_Media", "Fotoperiodo")
```

```
crs$numeric <- c("Idade_Meses", "Precipitacao_acum_corte",
"Temperatura_Media", "Et_Media",
     "Fotoperiodo")

crs$categoric <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Trimestre_Safra", "Operacao", "Tipo_Plantio",
     "Maturador")

crs$target  <- "ATR_classe"
crs$risk    <- NULL
crs$ident   <- NULL
crs$ignore  <- c("Safra", "TCH", "ATR", "TCH_x_ATR", "TCH_classe",
"TCH_x_ATR_classe")
crs$weights <- NULL

# Árvore de decisão
library(rpart, quietly=TRUE)

# Definindo uma semente.
crv$seed<-123;set.seed(crv$seed)
# Build the modelo de árvore de decisão.
crs$rpart <- rpart(ATR_classe ~ .,
    data=crs$dataset[crs$train, c(crs$input, crs$target)],
    method="class",
    parms=list(split="information"),
      control=rpart.control(cp=0.004300,
        usesurrogate=0,
        maxsurrogate=0))

# Generate a textual view of the modelo de árvore de decisão.
print(crs$rpart)
printcp(crs$rpart)
cat("\n")

#Construindo modelo de árvore aleatória usando abordagem tradicional.
crv$seed<-123;
set.seed(crv$seed)
crs$rf <- randomForest::randomForest(ATR_classe ~ .,
  data=crs$dataset[crs$sample, c(crs$input, crs$target)],
  ntree=500,
  mtry=3,
  importance=TRUE,
  na.action=randomForest::na.roughfix,
  replace=FALSE)

#Gerando saída textual do modelo de floresta aleatória
crs$rf

#Listando a importância das variáveis
rn <- round(randomForest::importance(crs$rf), 2)
rn[order(rn[,3], decreasing=TRUE),]

# Mostrando a árvore de nº 500.
printRandomForests(crs$rf, 500)
#============================================================
# Avaliando desempenho geral do banco de dados de treinamento.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$sample, c(crs$input,
crs$target)], type="class")
```

```
#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$sample, c(crs$input, crs$target)]$ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$sample, c(crs$input,
crs$target)]$ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$sample,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#============================================================
# Avaliando desempenho geral do banco de dados de validação.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$validate, c(crs$input,
crs$target)], type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$validate, c(crs$input, crs$target)]$ATR_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$validate, c(crs$input,
crs$target)]$ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))
```

```
# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$validate,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$ATR_classe, crs$pr)
round(per, 2)
#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#=============================================================
# Avaliando desempenho geral do banco de dados teste.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$test, c(crs$input,
crs$target)], type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$test, c(crs$input, crs$target)]$ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$test, c(crs$input,
crs$target)]$ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$test,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$test, c(crs$input, crs$target)])$ATR_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$test, c(crs$input,
crs$target)])$ATR_classe, crs$pr)
round(per, 2)
```

```
#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Salvando o projeto (variável crs) no HD do computador.
save(crs, file="DATA1.ATR.rattle", compress=TRUE)

############################################################################
#########
#) Análises: Árvores de decisão e florestas aleatórias
#)Fazendo estudo de TCH*ATR como variável compostas de classes (fator)
# (Desconsiderando a variável SAFRA)
############################################################################
#########
# construindo os bancos de dados de treinamento, validação e teste
crv$seed<-123
set.seed(crv$seed)
crs$nobs <- nrow(crs$dataset) # 2275 observations
crs$sample <- crs$train <- sample(nrow(crs$dataset), 0.7*crs$nobs) # 1592
observations
crs$validate <- sample(setdiff(seq_len(nrow(crs$dataset)), crs$train),
0.15*crs$nobs) # 341 observations
crs$test <- setdiff(setdiff(seq_len(nrow(crs$dataset)), crs$train),
crs$validate) # 342 observations

#Construindo modelo de árvore aleatória usando abordagem tradicional
crs$input <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Idade_Meses", "Trimestre_Safra", "Operacao",
     "Tipo_Plantio", "Maturador", "Precipitacao_acum_corte",
"Temperatura_Media",
     "Et_Media", "Fotoperiodo")

crs$numeric <- c("Idade_Meses", "Precipitacao_acum_corte",
"Temperatura_Media", "Et_Media",
     "Fotoperiodo")

crs$categoric <- c("T_Propriedade", "Estagio", "Variedade", "Ambiente",
     "Trimestre_Safra", "Operacao", "Tipo_Plantio",
     "Maturador")

crs$target  <- "TCH_x_ATR_classe"
crs$risk    <- NULL
crs$ident   <- NULL
crs$ignore  <- c("Safra", "TCH", "ATR", "TCH_x_ATR", "TCH_classe",
                 "ATR_classe")
crs$weights <- NULL

#============================================================
# Árvore de decisão
library(rpart, quietly=TRUE)

# Definindo uma semente.
crv$seed<-123
set.seed(crv$seed)

# Build the modelo de árvore de decisão.
crs$rpart <- rpart(TCH_x_ATR_classe ~ .,
    data=crs$dataset[crs$train, c(crs$input, crs$target)],
    method="class",
```

```
    parms=list(split="information"),
      control=rpart.control(cp=0.003200,
        usesurrogate=0,
        maxsurrogate=0))
# Generate a textual view of the modelo de árvore de decisão.
print(crs$rpart)
printcp(crs$rpart)
cat("\n")

#=============================================================
#Construindo modelo de árvore aleatória usando abordagem tradicional.
crv$seed<-123
set.seed(crv$seed)
crs$rf <- randomForest::randomForest(TCH_x_ATR_classe ~ .,
  data=crs$dataset[crs$sample, c(crs$input, crs$target)],
  ntree=500,
  mtry=3,
  importance=TRUE,
  na.action=randomForest::na.roughfix,
  replace=FALSE)

#Gerando saída textual do modelo de floresta aleatória
crs$rf

#Listando a importância das variáveis
rn <- round(randomForest::importance(crs$rf), 2)
rn[order(rn[,3], decreasing=TRUE),]

# Mostrando a árvore de n° 500.
printRandomForests(crs$rf, 500)

#=============================================================
# Avaliando desempenho geral do banco de dados de treinamento.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$sample, c(crs$input,
crs$target)], type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$sample, c(crs$input, crs$target)]$TCH_x_ATR_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$sample, c(crs$input,
crs$target)]$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$sample,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
```

```
table(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$sample, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

#============================================================
# Avaliando desempenho geral do banco de dados de validação.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$validate, c(crs$input,
crs$target)], type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$validate, c(crs$input, crs$target)]$TCH_x_ATR_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$validate, c(crs$input,
crs$target)]$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$validate,
c(crs$input, crs$target)]))
#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$validate, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))
```

```
#=============================================================
# Avaliando desempenho geral do banco de dados teste.
# Gerando uma matriz de erro do modelo de árvore de decisão.
# Obtendo a resposta do modelo de árvore de decisão.
crs$pr <- predict(crs$rpart, newdata=crs$dataset[crs$test, c(crs$input,
crs$target)], type="class")

#Gerando matriz de confusão mostrando contagens
table(crs$dataset[crs$test, c(crs$input, crs$target)]$TCH_x_ATR_classe,
crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))
?table
#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(crs$dataset[crs$test, c(crs$input,
crs$target)]$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Gerando uma matriz de erro do Modelo de floresta aleatória.
# Obtendo a resposta do Modelo de floresta aleatória.
crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[crs$test,
c(crs$input, crs$target)]))

#Gerando matriz de confusão mostrando contagens
table(na.omit(crs$dataset[crs$test, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr,
        useNA="ifany",
        dnn=c("Real", "Predita"))

#Gerando matriz de confusão mostrando proporções
per <- rattle::errorMatrix(na.omit(crs$dataset[crs$test, c(crs$input,
crs$target)])$TCH_x_ATR_classe, crs$pr)
round(per, 2)

#Calculando porcentagem de erro geral
cat(100*round(1-sum(diag(per), na.rm=TRUE), 2))

# Calculando porcentagem de erro geral média das classes
cat(100*round(mean(per[,"Error"], na.rm=TRUE), 2))

# Salvando o projeto (variável crs) no HD do computador.
save(crs, file="DATA1.TCHvsATR.rattle", compress=TRUE)


############<< FINAL DO SCRIPT R >>##############################
```