V SIMPÓSIO DA CIÊNCIA DO AGRONEGÓCIO

Aplicação de Técnicas de Mineração em Dados de Propriedades Leiteiras do Município de Derrubadas - RS

Luciano Moraes da Luz Brum[1], Vinicius do Nascimento Lampert[2], Sandro da Silva Camargo[3], Fábio André Eickhoff[4]

[1]*Mestrando no Programa de Pós-graduação em Computação Aplicada (PPGCAP), lucianobrum@unipampa.edu.br*
[2]*Pesquisador da Empresa Brasileira de Pesquisa Agropecuária e Docente no Programa de Pós-Graduação em Computação Aplicada (PPGCAP)*
[3]*Docente no Programa de Pós-Graduação em Computação Aplicada (PPGCAP) e no Curso de Engenharia de Computação da Universidade Federal do Pampa (Unipampa)*
[4]*Engenheiro agrônomo e extensionista agropecuário da Associação Riograndense de Empreendimentos de Assistência Técnica e Extensão Rural – EMATER/RS*

Resumo. *Os avanços da tecnologia e dos dispositivos de armazenamento de dados proverão um grande volume de dados. Os desafios surgem quando é necessário extrair informações destes dados brutos que sejam relevantes e uteis na tomada de ações e em processos decisórios. Neste presente trabalho é abordado o uso de técnicas de mineração de dados para a análise de indicadores de dimensões econômica, social, ambiental e produtiva de propriedades rurais de um município da região noroeste do Rio Grande do Sul. 54 propriedades rurais do município de Derrubadas foram escolhidas para este estudo de caso por serem consideradas como prioritárias para intervenção com políticas públicas. O propósito é construir um modelo preditivo que determine o índice de desenvolvimento das unidades de produção que um produtor de leite se enquadra com base nos valores dos indicadores e determinar quais destes são os mais relevantes nesta categorização. Os resultados deste trabalho podem apontar as razões das altas ou baixas médias de produção de leite deste município, direcionando ações e políticas públicas que auxiliem os produtores de leite.*
Palavras-chave: Propriedades, mineração, indicadores.

Application of Data Mining Techniques in Data of Dairy Farms of the city of Derrubadas - RS

Abstract. *Advances in technology and data storage devices have provided a great volume of data. Challenges arise when it is necessary to extract information from these raw data that is relevant and useful in taking actions and in decision-making processes. In this work, the use of data mining techniques for the analysis of indicators of economic, social, environmental and productive dimensions of rural properties of a city in the northwestern region of Rio Grande do Sul is approached. 54 rural properties in the city of Derrubadas were chosen for this case study because they are considered as priorities for intervention with public policies. The purpose is to construct a predictive model that determines the development index of the production units that a milk producer fits based on the values of the indicators and determine which of these are the most relevant in this categorization. The results of this work can point out the reasons of the high or low average of milk production of this municipality, directing actions and public policies that assist the milk producers.*
Keywords: *Properties, mining, indicators.*

1 Introdução

No Brasil, a atividade leiteira é uma atividade econômica de fundamental importância. Esta atividade possui um grande potencial de crescimento para os próximos anos, onde as taxas de crescimento anuais podem ficar entre 2,6% e 3,4%, correspondendo a 44,7 bilhões de litros de leite produzidos até o fim das projeções (BRASIL, 2014). A figura 1 apresenta a produção média de leite no Brasil entre 2013 e 2015. De acordo com o Atlas Socioeconômico do Rio Grande do Sul (2017), o estado de Minas de Gerais é o maior produtor com 27% do total e o Rio Grande do Sul (RS) é o segundo produtor nacional com cerca de 13% da produção ou 4,6 bilhões de litros em média no triênio 2013-2015.

A produção de leite no RS tem uma importância notória. Existem pelo menos 173.706 propriedades rurais que produzem alguma quantidade de leite. A produção destinada à indústria

equivale a R$ 4,2 bilhões por ano, o que representa R$ 8,5 milhões por ano para cada um dos 497 municípios gaúchos. O volume total de leite é de 4,5 milhões de litros de leite. Cada propriedade no RS gera em média uma renda estimada em R$ 2.210,48 por mês (EMATER, 2017). A produção de leite é a condição de atividade estratégica para a agricultura familiar, por colaborar para o desenvolvimento de muitas regiões do RS, pelo importante papel na composição da renda dos agricultores/as no processo de desenvolvimento econômico e social do país (EMATER, 2013).

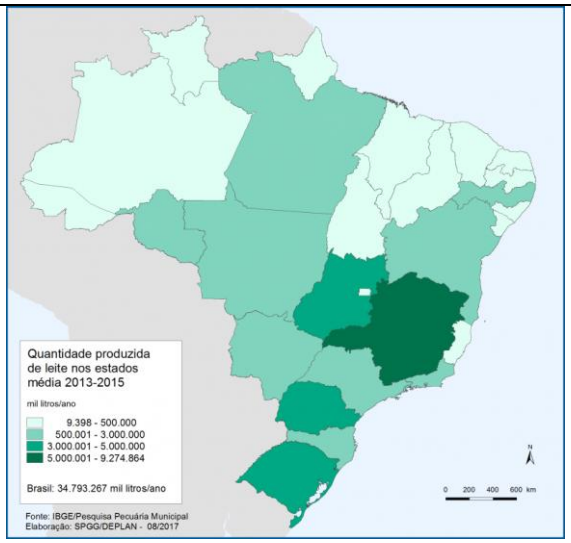
A pecuária de leite, além de sua grande importância econômica, apresenta aspectos sociais relevantes, em função principalmente de proporcionar condições de vida e trabalho para uma grande quantidade de famílias no meio rural. Na Região Noroeste há uma preocupação com relação ao êxodo rural e ao futuro dos agricultores familiares, já que na maioria dos casos a sucessão familiar está muito ameaçada (SILVA et al., 2016). Esta região é a principal produtora de leite do estado (Figura 2).

Nesta região observa-se que os agricultores que estão localizados no Corede Celeiro (Figura 3) estão entre os mais pobres do estado. Em torno de 11% das famílias de baixa renda ou de extrema pobreza estão no meio rural. O que reforça o entendimento do alto grau de pobreza deste Corede é a posição que os municípios ocupam no ranking do PIB estadual. Neste Corede estão os municípios com os piores índices do estado.

Neste sentido, a Emater/RS realizou um diagnóstico de propriedades com agricultura familiar atendendo a chamada pública SAF/ATER nº 07/2013, Lote 23 da secretaria da agricultura familiar vinculada ao ministério do desenvolvimento agrário com vistas ao desenvolvimento sustentável das Unidades de Produção Familiar com atividade leiteira identificadas como prioritárias. O trabalho contou com uma equipe de extensionistas vinculados aos 24 escritórios municipais (Figura 4) - correspondendo aos municípios integrantes deste Lote - bem como técnicos do escritório regional de Ijuí, além do apoio de técnicos do Escritório Central. Os fatores econômicos não são os únicos que podem garantir a permanência do produtor na atividade. Apesar da sua importância, segundo Ferraz (2003), na avaliação da sustentabilidade outras dimensões devem ser consideradas a fim de que por meio de indicadores possam ser analisados os fatores intrínsecos da atividade em cada uma dessas dimensões e também de suas inter-relações. A carência de indicadores que auxiliem no planejamento dos agricultores e entre os técnicos constitui um dos grandes desafios a serem superados pela pesquisa (MARCO REFERENCIAL EM AGROECOLOGIA, 2006).

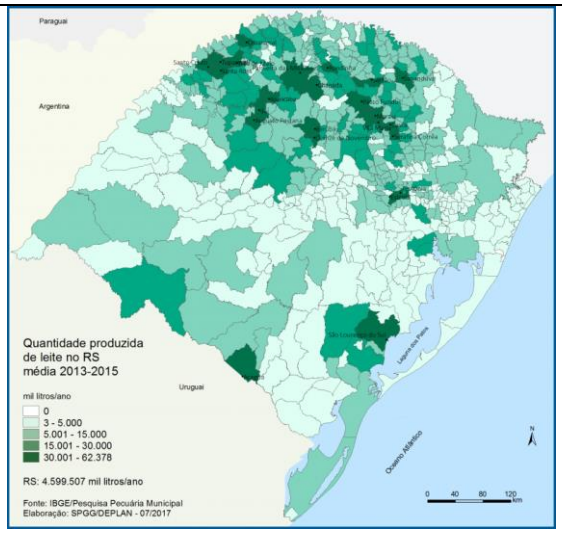
Neste sentido, para realização deste estudo foi utilizada a versão beta do método para análise sistêmica e evolução temporal de indicadores que está em fase de desenvolvimento e validação pela Embrapa e Emater/RS. O método tem como objetivo permitir identificar os gargalos e principais aspectos que estão estagnados e não avançam; aumentar a eficiência na coleta, registro e processamento das informações e melhorar o conhecimento sobre territórios. Na sua concepção serão criados parâmetros específicos que permitem verificar o comportamento instantâneo e a dinâmica de variação dos indicadores ao longo do tempo. Espera-se também que as informações geradas pelo instrumento constituam um subsídio importante para a ação da assistência técnica e extensão rural (ATER), e para trabalhos de pesquisa que buscam sistemas de produção mais sustentáveis.

Figura 1: Produção de leite no Brasil 2013-2015



Fonte: Atlas Socioeconômico do Rio Grande do Sul, 2017.

Figura 2: Produção de leite no RS 2013-2015

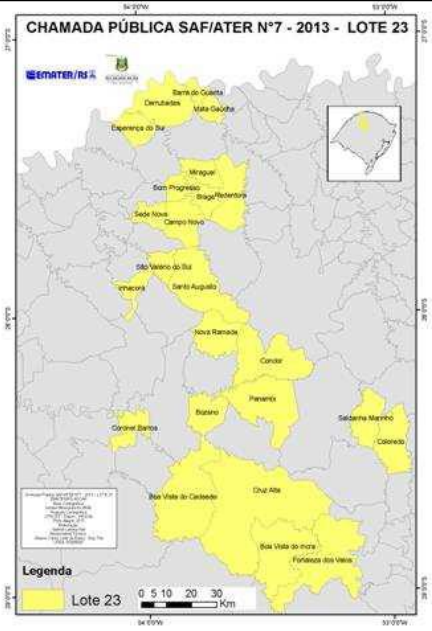


Fonte: Atlas Socioeconômico do Rio Grande do Sul, 2017.

Figura 3 - Mapa do Corede Celeiro

Fonte: Emater (2013).

Figura 4 - Localização geográfica dos municípios pertencentes ao Lote 23 da Chamada Pública SAF/ATER nº 07/2013.



Fonte: Emater (2013).

Este método surgiu com uma demanda da Emater/RS a partir da necessidade de interpretar as centenas de questionários dos diagnósticos obtidos com a chamada leite. Eram muitos dados e era complexo a análise comparativa entre propriedades ao longo de três anos.

Na presente pesquisa, foi escolhido o município de Derrubadas-RS para essa análise por ser um dos pioneiros na coleta de dados usando esta metodologia e por pertencer às regiões prioritárias quanto à intervenção e estabelecimento de políticas públicas, apresentado um desafio em elaborar estratégias de inclusão social e produtiva desses agricultores familiares nestes municípios onde o êxodo rural e população vem diminuindo mais rapidamente nos últimos anos.

A partir do Índice de Desenvolvimento das Unidades de Produção Familiar (ID-UPF) proposto no método, o propósito deste trabalho foi identificar quais os aspectos dentre as dimensões social, econômica, produtiva e ambiental tiveram maior influência para que algumas propriedades tivessem melhores ou piores índices. Para isso, serão executadas as etapas do processo de Descoberta de Conhecimento em Banco de Dados (DCBD) (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996). Os resultados podem ser importantes para auxiliar na definição de estratégias de intervenção nas propriedades com atividade leiteira carentes no município de Derrubadas-RS.

Este trabalho está estruturado em 4 seções, incluindo esta introdução apresentando um panorama histórico da produção de leite na região. A seção 2 aborda a metodologia adotada no trabalho. A seção 3 revela os resultados obtidos após a aplicação das técnicas de mineração de dados e uma breve discussão. A seção 4 apresenta as considerações finais deste trabalho e, logo após, as referências bibliográficas que embasaram este estudo.

2 Material e Métodos

O município de Derrubadas está localizado na região noroeste do estado do Rio Grande do Sul, possui população estimada em 3.190 habitantes de acordo com o último censo e uma área territorial de 361 km² (IBGE, 2010). Derrubadas caracteriza-se por uma agricultura familiar composta por 676 famílias, das quais 285 produzem leite comercialmente, através de um rebanho de 3.226 vacas ordenhadas e uma produção de 10 milhões de litros anuais (EMATER, 2017).

A escala de produção dos produtores é baixa, sendo que 36 produtores produzem até 50 litros/dia, 63 produzem de 51 a 100 litros/dia, 42 entre 101 e 150 litros/dia, 47 entre 151 a 200 litros/dia, 27 entre 201 e 300 litros/dia, 13 entre 301 a 500 litros/dia e apenas 07 acima de 501 litros/dia. A adoção de tecnologias é baixa, sendo que apenas 05 produtores fazem controle leiteiro, 05 produtores fornecem a ração conforme a produção dos animais e 180 usam inseminação artificial.

Quanto a infraestrutura, 200 produtores contam com local adequado para ordenha, 67 destas salas possuem fosso ou rampa, 67 possuem estrutura para alimentação individualizada e separada da ordenha. No que ser refere ao sistema de ordenha, 205 usam o balde ao pé, 20 usam transferidor de leite, 8 usam ordenhadeira canalizada e apenas 2 ordenham manualmente. Quanto a armazenagem do leite, 230 usam resfriadores de expansão direta e 05 resfriadores de tarro. Para limpeza dos equipamentos 135 usam água aquecida.

As principais dificuldades relatas são a falta ou a deficiência de mão-de-obra, dificuldade de acesso ao crédito, baixa escala de produção, condução da atividade realizada de forma pouco profissional; processos de gerenciamento e planejamento pouco aplicados; baixa qualidade do leite; baixa produtividade e rentabilidade; pouco uso de tecnologias e geração de renda mensal insatisfatória.

Portanto, para atingir o objetivo deste trabalho, foram coletados dados de produtores de leite, englobando 36 indicadores, sendo na dimensão social: políticas públicas, satisfação, energia elétrica, estradas, independência de insumos e facilidade do trabalho; na dimensão econômica: renda familiar per capita, gestão econômica, potencial de produção, estoque de semoventes, diversidade de atividades, participação do leite na renda e independência financeira; na dimensão ambiental: manejo do esgoto, manejo do lixo orgânico, manejo do lixo inorgânico, manejo de embalagens, manejo de dejetos de animais, manejo de resíduos agrícolas, manejo do solo, conservação do solo, adequação da reserva legal, adequação app e manejo da pastagem e na dimensão produtiva: disponibilidade hídrica, potencial produtivo, diversificação de forrageiras, grau de intensificação, melhoramento genético, precocidade de novilhas, intervalo entre partos, eficiência na reconcepção, sanidade do rebanho, produção diária de leite,

produção de leite por vaca e controle leiteiro. Esses indicadores foram uma amostra dos presentes no questionário que foram selecionadas pela Emater a fim de abastecer o método em desenvolvimento. Essas informações foram armazenadas e processadas por uma planilha em Excel.

A aquisição dos dados foi feita com o objetivo de extrair características ou reconhecer padrões sobre estes produtores. Para a extração de padrões, foram utilizados classificadores. Para a construção dos classificadores, foram executadas as etapas do processo de DCBD através da *IDE* (ambiente de desenvolvimento integrado) *RStudio* e da linguagem R (FAYYAD, PIATETSKY-SHAPIRO, SMYTH, 1996). O *RStudio* é uma ferramenta *open source* que funciona nos sistemas operacionais Linux, Windows e Mac OS (RStudio, 2017). Ela possui ferramentas e bibliotecas para plotagem de gráficos, cálculos estatísticos, além de diversos algoritmos de mineração de dados.

O pré-processamento de dados constitui as etapas iniciais do processo de DCBD. Apenas foram executadas as etapas 1, 3 e 4. A etapa 2, integração dos dados, não foi executada, pois tem-se como única fonte de dados uma planilha eletrônica (.xls). O público-alvo da pesquisa foram 60 produtores de leite do município de Derrubadas no ano de 2014 e 54 nos anos de 2015 e 2016. Neste trabalho, foram analisados apenas os dados do ano de 2016.

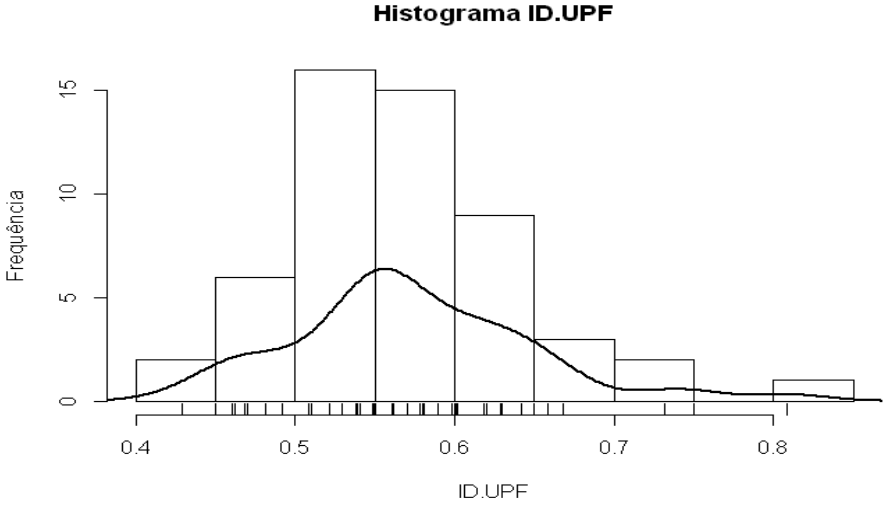
Inicialmente, os indicadores coletados foram normalizados em valores de 0 a 1 através de um processo de interpolação com valores de referência da região obtidos com especialistas. O objetivo é padronizar os dados para facilitar a análise e visualização dos dados e, ainda, permite a aplicação de algoritmos específicos de mineração de dados.

A etapa 1 consistiu da limpeza dos dados. Havia uma coluna que não possuía dados válidos para a análise, portanto, foi eliminada e a planilha foi salva em uma nova versão.

A etapa 3 consistiu da seleção dos dados para análise. Foi utilizado o indicador "ID.UPF" como variável dependente e o restante dos indicadores como variáveis independentes. O indicador "ID.UPF" é dependente, pois este é a média ponderada dos outros indicadores.

A etapa 4 consistiu da transformação dos dados em padrões apropriados para análise. Para efetuar essa etapa, foi utilizado o *RStudio* e a linguagem de programação R. O indicador "ID.UPF" foi transformado em dado categórico dividido em três faixas. Para determinar as faixas de valores, foi analisado o histograma deste indicador no ano de 2016, conforme mostra a figura 5. As categorias "ALTO", "MÉDIO" e "BAIXO" foram atribuídas, respectivamente, às faixas de valores [0.00, 0.55], [0.55, 0.625] e [0.625, 1.00].

Figura 5: Histograma do indicador ID.UPF.



Fonte: Autores, 2017.

Etapa 5: mineração de dados. Foram utilizados algoritmos de indução de árvores de decisão para determinar a influência dos diferentes indicadores no "ID.UPF". Conforme Han e Kamber (2006), indução de árvores de decisão é o aprendizado das árvores de decisão sobre atributos categóricos das tuplas de treinamento. Árvores de decisão são utilizadas na tarefa de classificação. Na literatura, existem diferentes algoritmos para esta tarefa, como o *CART*, *C4.5*, *C5.0*, *ID3*, todos adotando estratégias gulosas. Estas foram utilizadas porque não requerem conhecimento do domínio, são muito utilizados em análises exploratórias de conhecimento, funcionam bem com dados multidimensionais, a representação final em forma de árvore de decisão é intuitiva, treinamento e classificação são rápidos e possuem uma boa acurácia (HAN, KAMBER, 2006).

A interpretação destes resultados pode dar um norteamento na construção de políticas públicas de auxílio aos produtores de leite com dificuldades neste município, através da coleta, processamento e interpretação destes dados, auxiliando assim, em processos decisórios.

A acurácia destas técnicas foram determinadas através do *10-fold cross-validation (CV)*. O CV produz uma boa estimativa de erro por utilizar todas as informações como treinamento e teste e tende, em geral, a ter um viés e variância relativamente baixas (HAN, KAMBER, 2006).

4 Resultados e Discussões

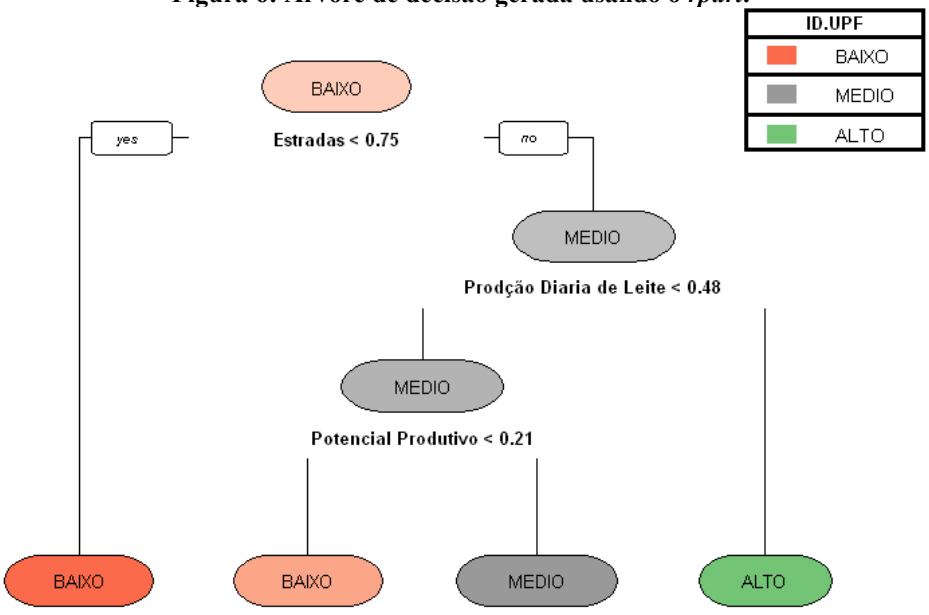
Nesta seção, são apresentados os resultados gerados pelos algoritmos de árvores de decisão, a precisão de cada modelo e uma discussão dos resultados. As figuras 6 e 7 ilustram as árvores geradas pelos algoritmos *rpart* e *C5.0*, respectivamente.

Na figura 6, em cada nó, temos as categorias do indicador ID.UPF indicadas pela cor. A cor verde indica a maioria das amostras com ID.UPF alto, cinza indica maioria com ID.UPF médio e vermelho indica maioria com ID.UPF baixo. Quanto maior a intensidade da cor, maior é a porcentagem de casos naquela categoria. Na figura 7, temos nos nós folhas, o número de casos naquele nó e a porcentagem de casos nas categorias de ID.UPF. A cor preta indica o nível baixo, o cinza indica nível médio e o branco indica nível alto deste indicador.

Por ordem de impacto no ID.UPF, o *rpart* e o *C5.0* retornaram os seguintes indicadores:

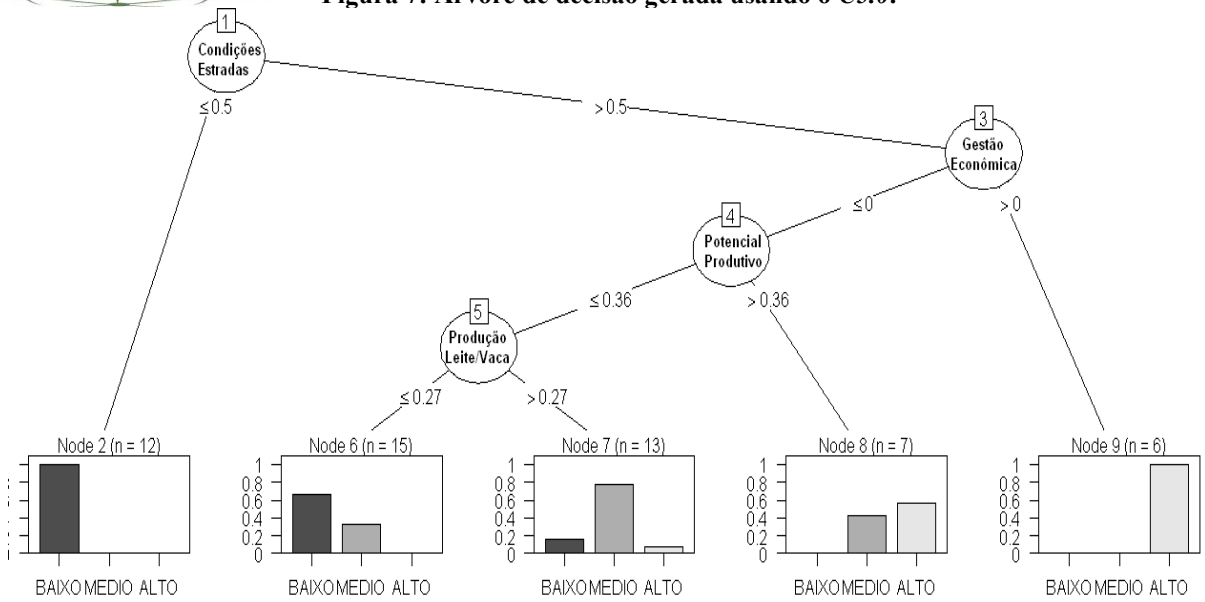
- *Rpart*: Presença de estradas, produção diária de leite e potencial produtivo.
- *C5.0:* Presença de estradas, gestão econômica, potencial produtivo e produção de leite por vaca.

Figura 6: Árvore de decisão gerada usando o *rpart*.



Fonte: Autores, 2017.

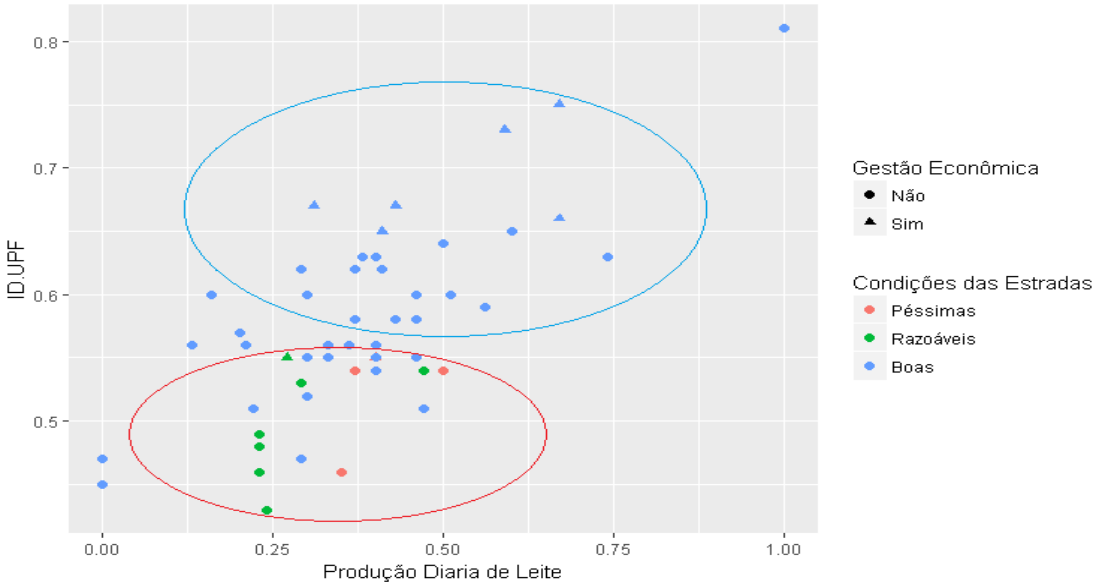
Figura 7: Árvore de decisão gerada usando o *C5.0*.



Fonte: Autores, 2017.

Ambos algoritmos apontam que a dificuldade de acesso à propriedades define todos os produtores com a média de indicadores no nível baixo. Portanto, a melhoria na acessibilidade destas propriedades pode influenciar na melhoria geral de todos os outros indicadores, proporcionando uma maior produtividade e sucesso de produtores de leite. Por outro lado, seguindo o *C5.0*, verifica-se que a presença de estradas e produtores que realizam gestão econômica determina produtores com a média de indicadores mais altas. No caso do *rpart*, além da presença de estradas, produtores com altos níveis no indicador produção diária de leite também possuem altos níveis nos demais indicadores. A produção diária de leite por si só depende de inúmeros fatores que fogem do escopo deste trabalho. Portanto, nestes resultados, destacam-se a importância da gestão econômica, produção diária de leite e a acessibilidade em propriedades produtoras de leite de Derrubadas. A figura 8 apresenta um gráfico com os indicadores apontados como relevantes pelos algoritmos, confirmando visualmente os resultados gerados.

Figura 8: Relação entre Produção diária de leite, condições das estradas, gestão e ID.UPF.



Fonte: Autores, 2017.

Foram obtidas as seguintes acurácias utilizando o método *10-fold cross-validation*:

- **Algoritmo *rpart*: Precisão de 63,33%.**
- **Algoritmo *C50*: Precisão: de 65,33%.**

O algoritmo *C5.0* produz um resultado com uma melhor acurácia em comparação com o *rpart*. Portanto, a confiabilidade dos resultados é maior no *C5.0*.

5 Considerações Finais

Neste trabalho foi analisada a influência dos indicadores no ID.UPF. Pôde-se concluir que produtores que tem péssimas e razoáveis condições de estradas tendem a ter níveis mais reduzidos nos outros indicadores. São necessárias políticas públicas que visem auxiliar tais produtores de leite neste sentido. Também foi apontado que efetuar a gestão econômica e possuir alta produção diária de leite, além de boas condições de estradas, determina produtores com altos níveis, em média, em todos indicadores. Como trabalhos futuros, sugere-se:

- Uma análise mais aprofundada utilizando algoritmos genéticos para efetuar novas descobertas de relacionamentos entre indicadores.
- Obter mais amostras e dados para otimizar as estimativas dos modelos preditivos.

Referências Bibliográficas

ATLAS SOCIOECONÔMICO DO RIO GRANDE DO SUL. Secretaria de Planejamento, Governança e Gestão. Disponível em: <http://www.atlassocioeconomico.rs.gov.br/leite>. Acesso em: 16 de setembro de 2017.

BRASIL. Ministério da Agricultura, Pecuária e Abastecimento. Projeções do Agronegócio: Brasil 2013/2014 a 2023/2024, Projeções de Longo Prazo. 5. ed. Brasília: MAPA/ACS, 2014. 100p.

EMATER, ASCAR-RS, Relatório socioeconômico da cadeia produtiva do leite no RS. POA, RS, 2017. 64p.

EMATER, ASCAR-RS, projeto de assistência técnica e extensão rural para promoção da agricultura familiar sustentável na cadeia produtiva do leite. Porto Alegre, RS, 2013. 108p.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. Advances in Knowledge Discovery and Data Mining. 1996. *AI Magazine*, v. 17 n° 3, 1996.

FERRAZ, J.M.G. Proposta Metodológica para a Escolha de Indicadores de Sustentabilidade. In: MARQUES, J.F.; SKORUPA, L.A.; FERRAZ, J.M.G. Indicadores de sustentabilidade em agroecossistemas. Jaguariúna, SP: Embrapa Meio Ambiente: 2003. p.59-72.

HAN, J.; KAMBER, M. Data Mining: Concepts and Techniques. 2° ed. Morgan Kauf. Publishers, p. 5–7, 2006.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. Disponível em: <http://cidades.ibge.gov.br/xtras/perfil.php?codmun=430632>. Acesso em: 21 de maio de 2017.

MARCO REFERENCIAL EM AGROECOLOGIA. Empresa Brasileira de Pesquisa Agropecuária, Brasília, DF: Embrapa Informação Tecnológica, 2006. 70p.

RSTUDIO. RStudio (2017). Disponível em: < https://www.rstudio.com/products/RStudio/> Acesso em: 31 de agosto de 2017.

SILVA, G. M.; LAMPERT, V. N.; WEILLER, O. H. et al. Indicadores de Sustentabilidade na Visão de Agricultores Familiares como Instrumento para Gestão de Unidades de Produção com Pecuária de Leite.In: CONGRESSO DA SOCIEDADE BRASILEIRA DE SISTEMA DE PRODUÇÃO, 11., 2016, Pelotas. Abordagem sistêmica e sustentabilidade: produção agropecuária, consumo e saúde: anais. Pelotas: SBSP, 2016.