MODELOS COMPUTACIONAIS PARA PREDIÇÃO DA QUALIDADE SENSORIAL DE VINHOS A PARTIR DE CARACTERÍSTICAS QUÍMICAS

Allan Sampaio Pires[1], Gustavo Trentin[1,2], Caroline Costa Moraes[3], Sandro da Silva Camargo[1]

[1]Programa de Pós-Graduação em Computação Aplicada – Universidade Federal do Pampa (UNIPAMPA)
Caixa Postal 242 – 96.413-170 – Bagé – RS – Brasil

[2]Embrapa Pecuária Sul – Bagé – RS – Brasil.

[3]Campus Bagé – Universidade Federal do Pampa (UNIPAMPA)

{allanpires,gustavo.trentin,caroline.moraes,sandro.camargo}@unipampa.edu.br

Abstract. *The production of fine wines in the region of the Rio Grande do Sul Campaign has intensified in recent years, increasing its relevance in the local economy. In order to enhance this relevance, one of the alternatives is the production of wines with higher quality and greater added value. The quality of the wine is attributed by specialists, through a process of tasting called sensorial analysis, which is influenced by the physico-chemical properties. The present work sought to investigate and quantify the relationship between the chemical attributes and the quality of the wines produced, in order to potentiate the production of higher quality wines. For this, were created of predictive models of sensorial quality through techniques of statistical modeling and classification. The models created here show that it is possible to achieve a precision of more than 90% for the prediction of sensory quality.*

Resumo. *A produção de vinhos finos na região da Campanha do Rio grande do Sul tem se intensificado nos últimos anos, incrementado sua relevância na economia local. A fim de potencializar esta relevância, uma das alternativas é a produção de vinhos com maior qualidade e maior valor agregado. A qualidade do vinho é atribuída por especialistas, através de um processo de degustação chamado de análise sensorial, a qual é influenciada pelas propriedades físico-químicas. O presente trabalho buscou investigar e quantificar a relação dos atributos químicos com a qualidade dos vinhos produzidos, a fim de potencializar a produção de vinhos de maior qualidade. Para isso, foram criados de modelos preditivos de qualidade sensorial através de técnicas de modelagem estatística e de classificação. Os modelos aqui criados mostram que é possível atingir uma precisão superior a 90% para a predição da qualidade sensorial.*

1. Introdução

A vitivinicultura é uma ocupação agrícola que apresenta custo bastante elevado, porém sua lucratividade mostra-se superior, em grande parte das vezes, ao cultivo de grãos e

criação de gado. Como consequência, ao longo das últimas décadas, solos sob campo natural do Bioma Pampa, localizados na região da Campanha Gaúcha do Rio Grande do Sul, que historicamente eram utilizados para a criação de gado de corte, passaram também ao sistema de produção de frutas, incluindo o cultivo de uvas viníferas. Os vinhedos estão localizados tanto em propriedades familiares quanto em propriedades de grupos empresariais que utilizam grandes quantidades de terra [Brunetto, 2016]. Com isso, nos últimos anos, a produção de vinhos finos tem se intensificado na região da Campanha do Rio Grande do Sul, e já corresponde por 25% da produção nacional, ficando atrás apenas da Serra Gaúcha, que possui tradição nesta cadeia produtiva [Velloso, 2014]. Um dos aspectos que possuem influência decisiva no preço do vinho é sua qualidade sensorial. Assim, metodologias que possam contribuir para uma melhor compreensão dos fatores que influenciam a qualidade sensorial tem uma forte justificativa econômica. Neste contexto, o presente trabalho visa aplicar uma metodologia para predição da qualidade sensorial de vinho, que geralmente é atribuída através da degustação do produto por especialistas, a partir de seus atributos químicos. A abordagem utilizada envolveu a aplicação de técnicas de mineração de dados para identificar os relacionamentos entre os atributos químicos e a qualidade sensorial.

Outros autores possuem trabalhos correlatos onde buscam predizer o atributo referente a qualidade sensorial do vinho, utilizando técnicas de regressão, máquinas de vetor de suporte e redes neurais, como observado em por Cortez et al (2009). Existem trabalhos que também utilizam técnicas de árvores de decisão com o uso do algoritmo C4.5 [Lee et al, 2015]. Por outro lado, na literatura encontra-se também a aplicação de outras técnicas com a finalidade de predizer a qualidade sensorial dos vinhos, como análise de variância (ANOVA), análise de cluster Q e o método de análise otimizado de componentes principais [Fengjiao et al, 2015].

O restante deste trabalho está dividido da seguinte forma: A seção Material e Métodos apresenta a base de dados utilizada além dos métodos de regressão e classificação utilizados. A seção Resultados exibe os resultados obtidos após a aplicação e validação dos métodos sobre a base de dados. A seção Conclusões conclui o artigo fazendo uma reflexão sobre os resultados obtidos.

2. Material e Métodos

Dada a falta de disponibilidade de dados relativos à região da Campanha, foi utilizada uma base de dados pública, obtida no repositório da Universidade da Califórnia - Irvine (UCI), disponível em http://archive.ics.uci.edu/ml/, com dados de vinhos do tipo vinho verde, da região de Minho em Portugal, que conta com um total de 1599 amostras. Nos experimentos realizados, foi utilizado o programa *R*, versão 3.3.3, em conjunto com os pacotes *rpart*, *randomForest* e *caret*. A base de dados era composta por doze atributos com onze medições de características químicas das amostras e a respectiva qualidade sensorial, que foi obtida mediante a avaliação de três especialistas, que poderiam atribuir valores entre 0 (muito ruim) e 10 (excelente), tendo sido considerada apenas a mediana destas avaliações. Assim, os atributos disponíveis são os seguintes: fixed acidity (acidez fixa), volatile acidity (acidez volátil), citric acid (ácido cítrico), residual sugar (açúcar residual), chlorides (cloretos), free sulfur dioxide (dióxido de enxofre livre), free sulfur dioxide (dióxido de enxofre livre), total sulfur dioxide (dióxido de enxofre total), pH, sulphates (sulfatos), alcohol (álcool) e quality (qualidade). A tabela 1 mostra um resumo da estatística descritiva da base de dados.

O primeiro passo foi carregar a base de dados e fazer uma análise dos valores da variável qualidade, que pode verificar-se através da Fig. 1 a distribuição dos valores de qualidade dos vinhos na base de dados analisada, onde o eixo X representa a qualidade e o eixo Y a distribuição destes valores.

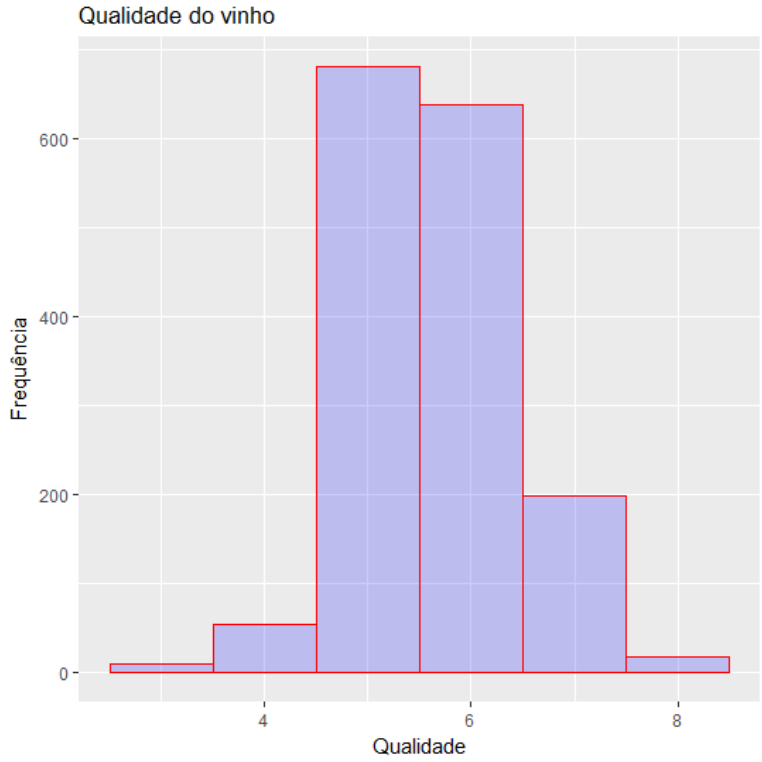

Figura 1. Histograma com a distribuição dos valores de qualidade dos vinhos

Verificou-se os valores mínimo e máximo, que foram 3 e 8 respectivamente, além de outras características estatísticas, que podem ser observadas na Tab. 1. Com base na observação destes valores, optou-se por uma nova classificação da qualidade dos vinhos, sendo esses classificados como “ruim”, “médio” e “bom”. Para isso utilizou-se faixas de valores da variável qualidade. Foi necessário fazer uma varredura para substituição dos valores, sendo que amostras que possuíam valores de qualidade menores ou igual a três receberam a classificação “ruim”, sendo classificados como “bom” os vinhos com valores de qualidade compreendidos entre quatro e seis, e acima de seis foram classificados como “bom”. Para treinamento dos modelos utilizados foram utilizados 70% dos dados disponíveis na base. Após isto foram aplicadas as técnicas de regressão linear, para a identificação dos atributos mais relevantes, aplicação de técnicas de aprendizagem supervisionada, no caso foram os algoritmos de classificação Random Forest e Rpart, para a obtenção de modelos preditivos e verificado seus medidores de eficiência através do método de validação cruzada k-fold. Para criação dos modelos preditivos, foram utilizados algoritmos de árvores de decisão (decision tree) e florestas aleatórias (Random Forest). As árvores de decisão são ferramentas poderosas e populares para classificação e previsão. As árvores de decisão representam regras, que podem ser compreendidas pelos humanos e usadas em sistema de conhecimento, como

banco de dados. Estes modelos são classificadores na forma de uma estrutura de árvore que consiste em nó de decisão que especifica um teste em um único atributo[8]. As florestas aleatórias são utilizadas para regressão e classificação. Esta metodologia inclui a construção de múltiplas árvores de decisão com os dados de treinamento fornecidos. Para verificação dos resultados foi utilizada a técnica de validação cruzada, que visa dividir o total das amostras em K segmentos, e com isso o modelo é testado K vezes, cada vez com uma versão do conjunto de treinamento em que um dos segmentos é omitido[9].

Tabela 1: Estatísticas descritivas dos atributos físico-químicos dos vinhos analisados

Atributo	Unidades	Valor mínimo	Valor máximo	Mediana	Média	Moda	Desvio padrão
1 - acidez fixa	g(ácido tartárico)/dm³	4.60	15.90	7.90	8.320	7.20	1.741096
2 - acidez volátil	g(ácido acético)/dm³	0.12	1.58	0.52	0.5278	0.60	0.1790597
3 - ácido cítrico	g/dm³	0.00	1.00	0.26	0.271	0.00	0.1948011
4 - açúcar residual	g/dm³	0.90	15.50	2.20	2.539	2.00	1.409928
5 - cloretos	g(cloreto de sódio)/dm³	0.012	0.611	0.079	0.08747	0.08	0.0470653
6 - dióxido de enxofre livre	mg/dm³	1.00	72.00	14.00	15.870	6.00	10.46016
7 - dióxido de enxofre total	mg/dm³	6.00	289.00	38.00	46.470	28.00	32.89532
8 - densidade	g/dm³	0.9901	1.004	0.9968	0.9967	0.9972	0.0018873
9 - pH	-	2.74	4.01	3.31	3.311	3.30	0.1543865
10 - sulfatos	g(sulfato de potássio)/dm³	0.33	2.00	0.62	0.6581	0.60	0.169507
11 - álcool	% Vol.	8.40	14.90	10.20	10.420	9.50	1.065668
12 - qualidade	-	3	8	6	5.636	5	0.8075694

3. Resultados

Neste trabalho foram utilizadas as técnicas de regressão linear simples, para detecção dos atributos de maior relevância, assim como o algoritmo Random Forest, que além de detectar os atributos mais importantes, também cria múltiplos modelos preditivos, e também o algoritmo Rpart, que cria um modelo preditivo baseado em árvores binárias. Os resultados obtidos são descritos a seguir.

3.1. Regressão Linear

A análise de regressão é um método matemático que aborda as relações entre múltiplas

variáveis. O método de análise de regressão mais simples e mais utilizado é a regressão linear simples, que mostra-se conveniente para a análise de um modelo com múltiplos atributos [Li et al, 2017]. Com a aplicação deste modelo, é possível perceber qual atributo da base de dados tem maior influência na predição da qualidade dos vinhos analisados.

Para a utilização da técnica citada, empregou-se a base de dados de treino, e concluiu-se que para predição da qualidade das amostras, primeiramente, a variável álcool se mostra mais importante, seguida do atributo sulfato. Os resultados obtidos podem ser observados na Tab. 2, onde percebe-se que os atributos álcool e sulfato obtiveram valores mais altos após a aplicação do método, concluindo-se então que estes são as características mais influentes na qualidade sensorial dos vinhos.

Tabela 2: Resultados obtidos com a aplicação da regressão linear

	Estimate	Std. Error	Value	Pr (> | t |)
(Intercept)	2.761e+01	2.518e+01	1.097	0.272991
1 - acidez fixa	2.553e-02	3.141e-02	0.813	0.416472
2 - acidez volátil	-1.028e+00	1.443e-01	-7.125	1.88e-12
3 - ácido cítrico	-4.524e-02	1.754e-01	-0.258	0.796500
4 - açúcar residual	1.540e-02	1.709e-02	0.901	0.367578
5 - cloretos	-1.795e+00	5.231e-01	-3.432	0.000622
6 - dióxido de enxofre livre	8.306e-03	2.603e-03	3.191	0.001457
7 - dióxido de enxofre total	-4.376e-03	8.694e-04	-5.034	5.61e-07
8 - densidade	-2.335e+01	2.573e+01	-0.907	0.364415
9 - pH	-4.480e-01	2.355e-01	-1.902	0.057371
10 - sulfatos	9.536e-01	1.311e-01	7.273	6.64e-13
11 - álcool	2.575e-01	3.153e-02	8.167	8.56e-16

3.2. Algoritmo Rpart

O algoritmo Rpart (Recursive PARTitioning) é utilizado para a construção de árvores binárias. A árvore é construída pelo seguinte processo: primeiro é encontrada a única variável que divide melhor os dados em dois grupos. Os dados são separados e, em seguida, esse processo é aplicado separadamente a cada subgrupo e assim recursivamente até que os subgrupos alcancem um mínimo tamanho ou até que nenhuma melhoria possa ser feita [Chitra e Alias Balamurugan, 2013]. Foram construídos dois modelos de árvores, um completo, que pode ser observado na Fig. 2, e outro mais enxuto que utiliza os conceitos de poda do algoritmo. A árvore podada gerada, que é mostrada na Fig. 3, aponta que o atributo químico mais importante para a qualidade sensorial do produto final é a concentração de álcool. Com concentração inferior a 11.55, o produto final tende a ter uma qualidade média. Com concentração maior que este limiar, o sulfato deve ser considerado a seguir, sendo que se sua concentração for superior a 0.685, deveremos considerar o dióxido de enxofre livre,

sendo este superior a 18.5, o vinho será considerado bom, e caso contrário, o vinho terá qualidade média. Caso o valor da concentração de sulfato seja inferior a explicitada anteriormente, a variável a ser considerada será o dióxido de sulfato total, sendo que na hipótese de seu valor ser inferior a 15.5, o sulfato será novamente analisado, sendo que se sua concentração for superior a 0.585, o vinho será classificado como bom, caso contrário, será qualificado como médio. E finalmente, caso o dióxido de sulfato total seja superior a 15.5, teremos o atributo químico dióxido de sulfato livre como determinante, sendo que, os produtos com concentração acima de 31.5 deste elemento serão classificados como bom, e os demais como médio.

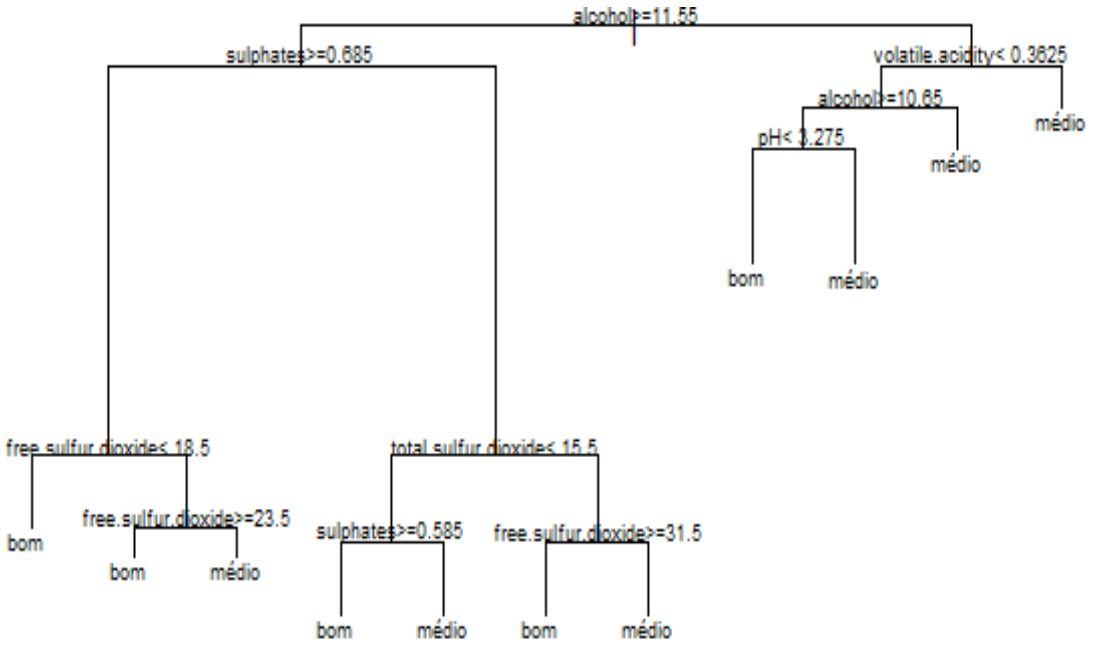

Figura 2. Árvore Rpart

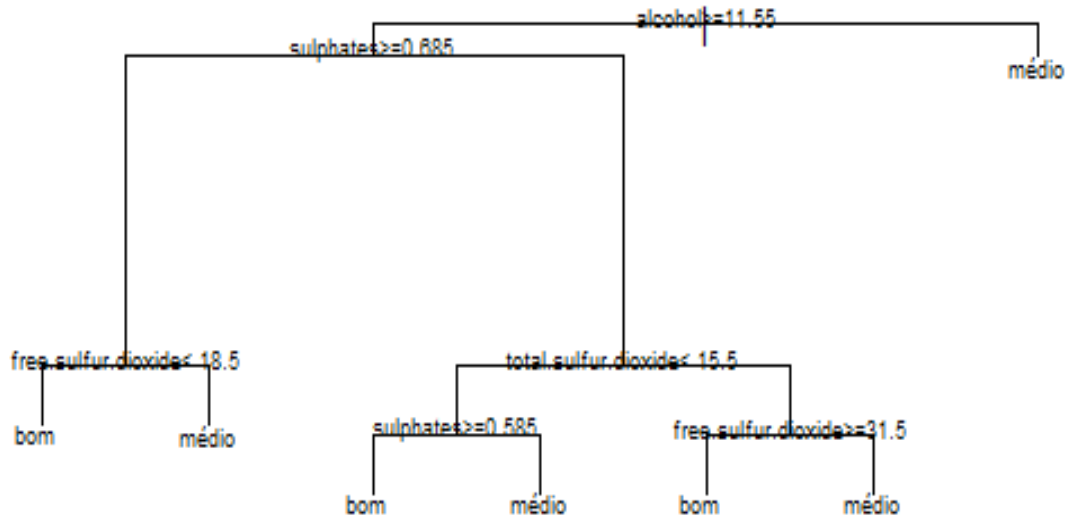

Figura 3. Árvore Rpart após a poda

Após a criação do modelo de árvore podada, foi gerado um gráfico, mostrado na Fig. 4

que possibilita observar os valores de qualidade reais e preditos, levando em consideração os atributos álcool e sulfatos, julgados fundamentais pelo algoritmo. Para isso, foi utilizado o conjunto de dados separados para teste.

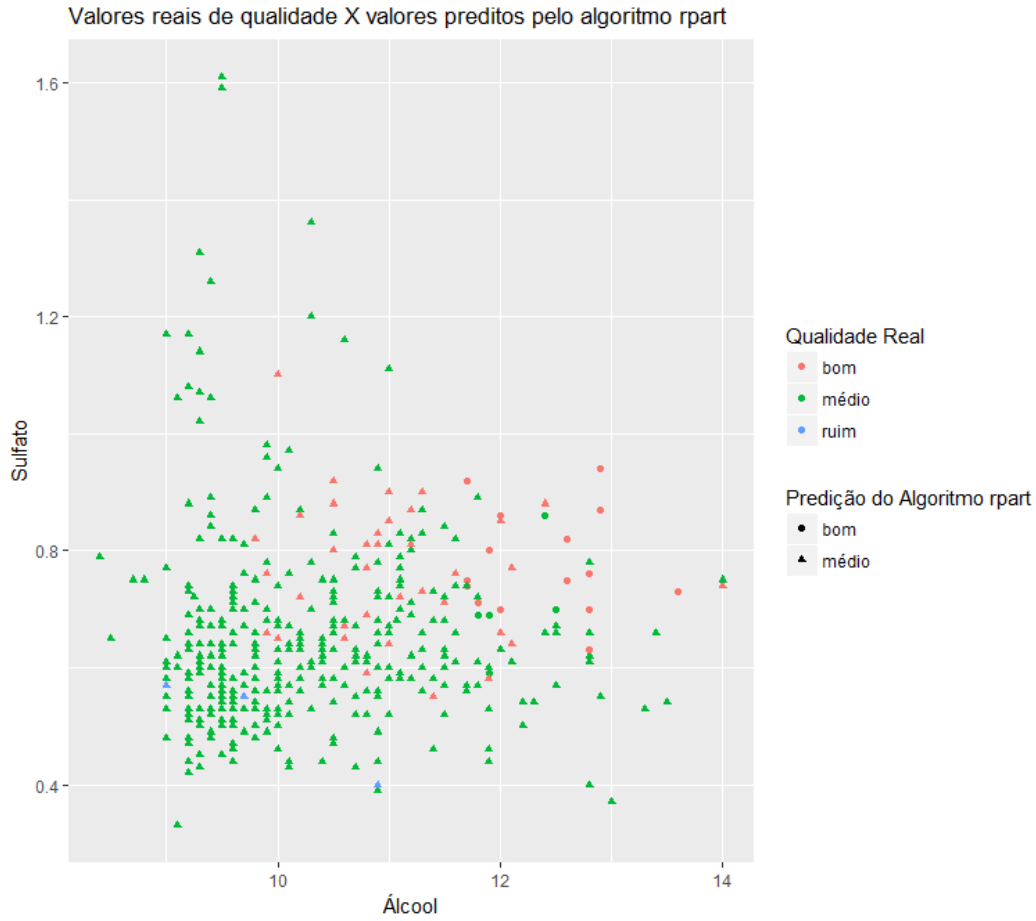

Figura 4. Valores reais e valores preditos pelo algoritmo

3.3. Algoritmo Random Forest

O algoritmo Random Forest é um tipo de aprendizado de máquina que consiste na criação de múltiplas árvores de decisão, utilizados para classificação e regressão, a partir dos dados de treinamento [Nakahara et al, 2017]. Neste caso a aplicação do algoritmo construiu quinhentas árvores de decisão e permitiu verificar quais variáveis são mais importantes para a classificação das amostras. A Figura 4, gerada a partir do algoritmo mostra a importância das variáveis, que se dá através do cálculo do índice de Gini [Gini, 1997].

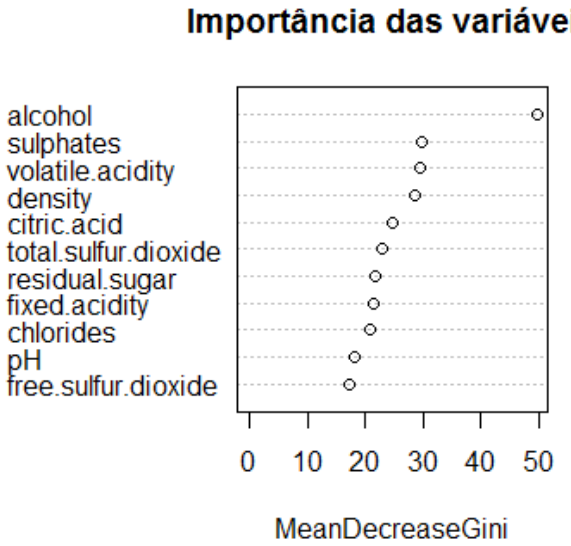

Figura 4: Importância das variáveis

3.4. Discussão dos Resultados

Para teste dos modelos foi utilizado o método de validação cruzada denominado k-fold, onde os dados são divididos em grupos e testados em múltiplas iterações. A precisão obtida em cada iteração é calculada para então obter a acurácia do modelo [Yadav e Shukla, 2016]. Os resultados destes testes podem ser vistos em tabela, denominada matriz de confusão, onde pode-se calcular a acurácia do modelo, somando-se os valores da diagonal principal da matriz e dividindo-se este resultado pelo total de elementos da tabela. Estes resultados são observados em Tab. 3 e Tab. 4. Com o método Rpart a precisão chegou a 88,5% enquanto o método Random Forest atingiu 90,04%. Vale ressaltar que os resultados obtidos com ambos os algoritmos foram coerentes, pois os atributos considerados relevantes por um método também foram julgados importantes pelo outro.

Tabela 3. Matriz de Confusão Rpart

Predição	bom	médio	ruim
bom	20	11	0
médio	38	380	3
ruim	0	0	0

Tabela 4. Matriz de Confusão Random Forest

Predição	bom	médio	ruim
bom	25	9	0
médio	33	382	3
ruim	0	0	0

4. Conclusões

As diferentes abordagens utilizadas apontaram respostas muito similares em relação a importância dos atributos, o que reforça a validade dos resultados. Os métodos de classificação utilizados mostraram-se eficientes nos testes realizados, tendo uma alta capacidade preditiva para a predição da qualidade sensorial a partir de características químicas. Desta forma, o trabalho aqui realizado apresenta um potencial de aplicação para a classificação dos vinhos da região da Campanha, agregando conhecimento no processo de produção e tornando-se uma alternativa para os produtores locais buscarem a melhoria da qualidade de seus produtos.

5. Referências

Brunetto, G., 2016, "Adubação e calagem em videiras cultivadas em solos arenosos no bioma pampa." em *Embrapa Uva e Vinho-Capítulo em livro técnico (INFOTECA-E)*.

Chitra, P. K. A., Alias Balamurugan, S. A., 2013, “Benchmark evaluation of classification methods for single label learning with R,” em *IEEE International Conference On Emerging Trends in Computing, Communication and Nanotechnology (ICECCN)*, Tirunelveli, Índia, p.746-752.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. e Reis, J., 2009, “Modeling wine preferences by data mining from physicochemical properties, Decision Support Systems,” Elsevier, Vol. 47, Issue 4, p. 547-553.

Fengjiao F., Jianping L., Guoming G., Chenxi M., 2015, "Mathematical model application based on statistics in the evaluation analysis of grape wine quality," *12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, Chengdu, China, p. 107-110.

Gini, C. “Concentration and dependency ratios,” Rivista di Politica Economica, vol. 87, pp. 769–789, 1997

Lee, S., Park, J., Kang, K., 2015, "Assessing wine quality using a decision tree," *2015 IEEE International Symposium on Systems Engineering (ISSE)*, Roma, Itália, p. 176-178.

Li, H., Pi, D., Wu Y., e Chen C., 2017, "Integrative Method Based on Linear Regression for the Prediction of Zinc-Binding Sites in Proteins," em *IEEE Access*, vol. 5, no. , p. 14647-14656.

Nakahara, H., Jinguji, A., Sato, S. e Sasao T., 2017, “A Random Forest Using a Multi-valued Decision Diagram on an FPGA,” em *IEEE 47th International Symposium on Multiple-Valued Logic (ISMVL)*, Novi Sad, Sérvia, p. 266-271

Prajwala, T. R., "A comparative study on decision tree and random forest using R tool," *International journal of advanced research in computer and communication engineering*, Vol. 4, Issue 1, p. 196-1.

Shevade, Shirish Krishnaj, and S. Sathiya Keerthi. "A simple and efficient algorithm for gene selection using sparse logistic regression." Bioinformatics19.17 (2003): 2246-2253.

Velloso, G. Os vinhos feitos no Pampa. Disponível em: <http://paladar.estadao.com.br/noticias/bebida,os-vinhos-feitos-nos-

pampas,10000008201>. Acesso em: 07 de julho de 2017.

Yadav S., Shukla, S., 2016, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” em *IEEE 6th International Conference on Advanced Computing (IACC)*, Bhimavaram, Índia, p. 78-83.