Aprendizado de máquina para predizer a qualidade do leite

Gabriel Correia Granja
Universidade Federal de São Paulo
São José dos Campos, Brasil
correia.gabriel07@unifesp.br

***Abstract*—Observando cada vez mais o avanço da Inteligência Artificial, principalmente na área de Aprendizado de Máquina, seus recursos recorrentemente vão sendo expandidos e aplicados dentro das indústrias para a melhoria e inovação da qualidade dos produtos produzidos. Tendo conhecimento dos conceitos dessa área tecnológica, este artigo busca aplicar os conceitos de Aprendizado de Máquina relacionados com a indústria de laticíneos, demonstrando como a Inteligência Artificial pode auxiliar no processo da fabricação do leite.**

I. Introdução

A indústria de laticíneos é formada por um conjunto de fábricas na qual trabalham na produção de leite e seus derivados, como por exemplo: iogurtes e queijos. A cadeia de leite possui a sua grande importância dentro da indústria de alimentos, sendo responsável por 12% do total do valor gerado na área de alimentos. Além disso, pelo fato do leite e seus derivados serem altamente perecíveis, necessita-se de cuidados especiais acerca da questão de produção, armazenamento, processamento e distribuição para que apresentem a qualidade ideal para o devido consumo [2].

Visto que dentro do cenário competitivo atual, elaborar e gerenciar conhecimento que auxilie especialistas no controle de processos industriais pode ser produtivo dentro de uma determinada indústria. Nesse caso, a Inteligência Artificial é uma área de conhecimento que fornece diversos modelos de apoio para a decisão e controle de acordo com fatos e conhecimentos empíricos e teóricos [8].

Em função disso, o objetivo desse artigo é demonstrar por meio de testes e análise, uma aplicação de algoritmos de aprendizado de máquina supervisionados sobre um conjunto de dados que indicam a qualidade do leite, que nos possibilite pontecial crescimento de produção e melhoria de qualidade em relação ao produto, sendo capaz de nos ajudar a predizer a rotulação do leite, com melhor eficiência e menor tempo de execução

Os algoritmos de aprendizado de máquina supervisionado utilizados para o propósito desse trabalho são: Naive Bayes [11], Regressão Linear [6] e *Multi-Layer Perceptron* (MLP) [7]. A base de dados análisada foi retirada da plataforma de ciência de dados *Kaggle* [9].

Logo, tendo o contexto e objetivo desse artigo já definidos, o restante dele está dividido em seções e organizado da seguinte forma: Na Seção 2 apresenta-se brevemente os trabalhos e pesquisas que já integraram Inteligência Artificial dentro da indústria de determinado setor e serviram de inspiração e auxílio para esse projeto. Na Seção 3 é explicado qual a metodologia que foi seguida para o estudo de cada algoritmo aplicado ao *dataset*. A descrição da base de dados, juntamente com a configuração do ambiente e discussão dos resultados são feitas na Seção 4. Por último, tem-se a conclusão com base em tudo o que foi observado.

II. Trabalhos Relacionados

Como o seguinte tema desse trabalho propõem a incorporação de algoritmos de Inteligência Artificial para a avaliação do leite, auxiliando em seu processo fabricação e manipulação dentro da indústria, projetos relacionados com Inteligência Artifical aplicada para inspecionar e analisar a qualidade de um produto foram pesquisados como fonte de inspiração e auxílio dentro da temática abordada neste artigo.

No artigo de Inspeção Visual Automática de Peças Cerâmicas via Inteligência Artificial [1], os autores desenvolveram um sistema visual automático, baseado em uma técnica de Inteligência Artificial definida como lógica difusa, sendo o sistema utilizado no inspecionamento de peça cerâmicas. Tal proposta tem como fundamentação a demanda por desenvolvimento novos métodos e sistema para inspeção final de qualidade de revestimentos cerâmicos.

No projeto de Classificação de defeito em lotes numa indústria farmacêutica: Uma abordagem prática com aprendizado de máquina em processos de qualidade [5] foi proposto a integração de técnicas de aprendizado de máquina para o controle de qualidade, concentrando-se na solução de questões relacionadas com perdas referentes aos lotes de produtos. Dentro desse estudo, os algoritmos de aprendizado de máquina utilizados para a exploração do conjunto de dados coletados foram: *Support Vector Machines* (SVM), Florestas Aleatórias e *Gradient Boosting*.

Para o artigo sobre Aprendizado de Máquina Aplicado a Avaliação da Qualidadede Frutos [3], oa autores buscam a exploração de técnicas de visão computacional e aprendizado de máquina para a classificação de frutas, com ênfase na qualidade e controle pós-colheita. A classificação das frutas é realizada por implementação de redes neurais artificiais, como o *Convolutional Neural Network* e *MultiLayer Perceptron*, sendo ela obtida atráves da análise sobre as características físicas da fruta.

Por fim, na dissertação de pós-graduação de Analisando Cerveja Artesanal por meio de 3 Modelos de Classificação e Aprendizado de Máquina [10] tem como objetivo a aplicação de métodos de aprendizado de máquina visando a correta classificação dos dados disponíveis nas receitas, possibilitando realizar previsões mais acertivas de produto final. Os três modelos utilizados para a categorização entre os estilos de cerveja foram: o *Random Forest Classifier*, a *Logistic Regression* e *XGBoost Classifier*.

III. Metodologia

Os passos para esse estudo seguem o seguinte o fluxograma abaixo na figura 1:

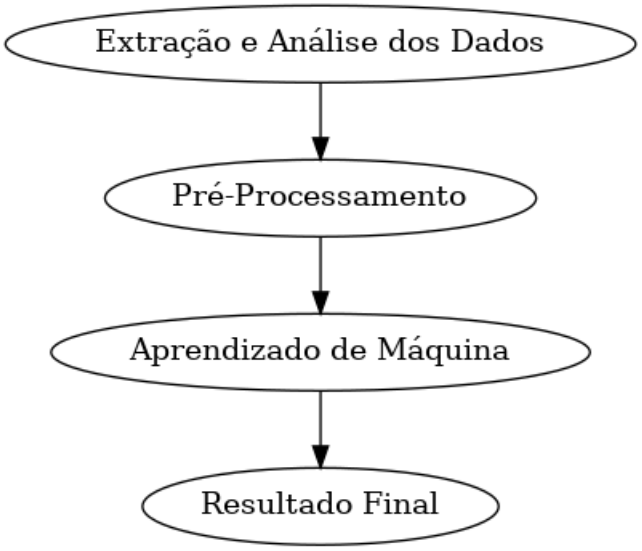

Fig. 1. Fluxograma da metodologia empregada.

A. Extração e Análise dos Dados

A extração de dados foi feita a partir da leitura do arquivo "*milknews*.csv", proveninetes do site "*Kaggle*". A base de dados possui 1059 linhas e 8 colunas. Seus detalhes mais específicos e estatísticos serão analisados na seção IV de Análise Experimental.

B. Pré-processamento

Após a extração e análise, executou-se o pré-processamento dos dados sendo essencial para a construção modelos preditivos robustos. Inicializamos definindo atributos e a classe a ser prevista. Em seguida, dividimos o conjunto de dados em treinamento e teste para avaliação. A normalização é aplicada para garantir consistência na escala dos atributos, evitando impactos desproporcionais. Essas etapas são cruciais para desenvolver modelos eficazes e avaliá-los adequadamente.

C. Aprendizado de Máquina

Feito o pré-processamento, aplicou-se 3 técnicas de aprendizado de máquina supervisionado no qual objetivo de cada algoritmo é a construção de um classificador que defina corretamente classe de novos dados que ainda não foram classificados [4].

A primeira técnica implementada corresponde ao Algoritmo de Naive Bayes que utiliza a Regra de Bayes juntamente com uma forte suposição de que os atributos são incondicionalmente independentes dada a classe. Apesas de que suposição de independência é frequentemente violada na prática, o Naive Bayes, no entanto, muitas vezes oferece precisão de classificação competitiva. [11].

A segunda técnica refere-se ao algoritmo de Regressão Linear, em que se baseia em um modelo matemático capaz de descrever a relação entre duas ou mais variáveis do tipo quantitativo. Caso o modelo for a equação de uma reta e incidir unicamente sobre as duas variáveis, então define-se por regressão linear simples [6].

A última técnica é sobre o algoritmo de *Multi-Layer Perceptron* (MLP) que consiste em múltiplas camadas de elementos de processamento simples, de dois estados, do tipo sigmoide, ou neurônios, que interagem por meio de conexões ponderadas. Após uma camada inferior de entrada, geralmente existem várias camadas intermediárias, ou ocultas, seguidas por uma camada de saída no topo [7].

D. Resultado Final

Ao final de tudo, examinamos os resultados referentes à acurácia e tempo de execução de cada algoritmo para propor a determinada conclusão desse projeto.

IV. Análise Experimental

A. Conjunto de dados

Como já foi dito, o *dataset* é composto por 1059 linhas de dados divididos entre 8 colunas, que representam os tipos de dados apurados em relação aos leites averiguados, sendo eles o seu pH, a sua temperatura, o seu sabor, o seu odor, a sua gordura, a sua turbidez, a sua cor e por último a sua qualificação.

Os parâmetros de sabor, gordura e turbidez atribuem o valor 1 para condições ideais do leite e valor 0 caso contrário. Enquanto que o ph, a temperatura e a cor atribuem seus valores reais no conjunto de dados. Para a coluna de qualificação os valores atribuidos são *low* (ruim), *medium* (moderado) e *high* (bom).

Sabendo dessas informações realizamos uma descrição estatística sobre os dados, como é mostrado na figura 2:

	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
count	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000	1059.000000
mean	6.630123	44.226629	0.546742	0.432483	0.671388	0.491029	251.840415
std	1.399679	10.098364	0.498046	0.495655	0.469930	0.500156	4.307424
min	3.000000	34.000000	0.000000	0.000000	0.000000	0.000000	240.000000
25%	6.500000	38.000000	0.000000	0.000000	0.000000	0.000000	250.000000
50%	6.700000	41.000000	1.000000	0.000000	1.000000	0.000000	255.000000
75%	6.800000	45.000000	1.000000	1.000000	1.000000	1.000000	255.000000
max	9.500000	90.000000	1.000000	1.000000	1.000000	1.000000	255.000000

Fig. 2. Descrição estatística do *dataset*.

Com base nessas descrições, temos os seguintes boxplots e gráficos de barra das variáveis sobre a qualificação do leite, demonstrados nas figuras 3 e 4:

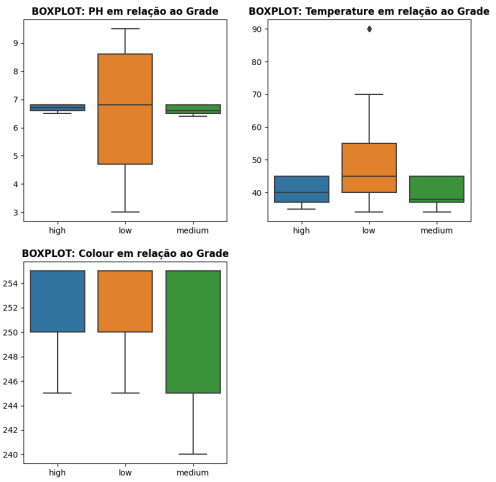

Fig. 3. *Boxplots* referentes às variáveis Ph, Temperatura e Cor.

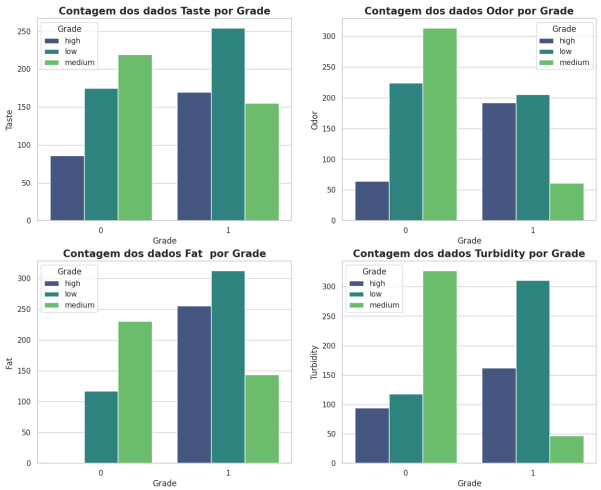

Fig. 4. Gráficos de barra referente às variáveis Sabor, Odor, Gordura e Turbidez.

B. Configuração do ambiente e dos algoritmos

O estudo desse projeto foi desenvolvido dentro do *Google Colab*, havendo um número de 2 CPUs disponíveis, 107,7 *gigabytes* de armazenamneto e 12,7 *gigabytes* de memória RAM.

O código executado foi escrito em *Python* na versão *Python3* e as módulos empregados para o pré-processamento dos dados e aplicação dos 3 algoritmos encontram-se inseridos no biblioteca *scikit-Learn*, também conhecida como *sklearn*, responsável por implementar modelos de aprendizado de máquina e modelagem estatística, possibilitando realizar diversos modelos de aprendizado de máquina para regressão, classificação, clustering e ferramentas estatísticas para analisá-los.

Para o pré-processamento dos dados, como já foi visto na seção III, o conjunto será separado em 2 porções sendo uma para treinamento e outra para teste. Nesse caso, foram separados 70% dos dados para treninamento e 30% para testagem.

C. Resultados e discussão

Passado os dados como parâmetros para ambos os algoritmos de aprendizado superviosnado, após o pré-processamento, testou-se 30 vezes a execução de cada algoritmo para verificar sua acurácia e tempo de execução para predizer a qualidade do leite acerca dos dados testados. Para facilitar o balanço, organizou-se os resultados em 2 gráficos de linha, sendo o primeiro a respeito da comparação de acurácia representado na figura 5 e o segundo sobre a comparação do tempo de execução na figura 6

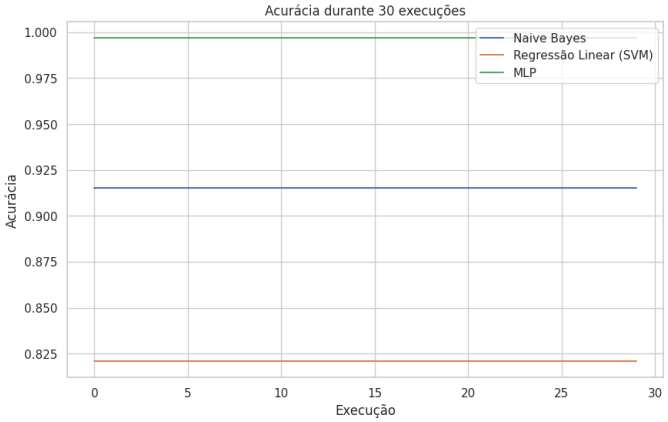

Fig. 5. Acurácia dos 3 algoritmos.

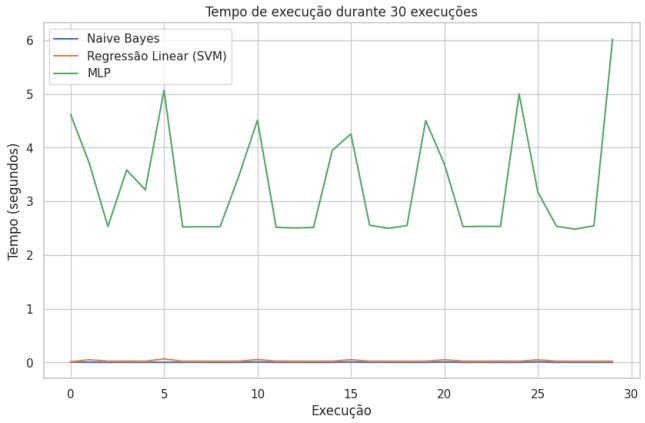

Fig. 6. Tempo de execução para os 3 algoritmos.

O valor de acurácia permaneceu constante para ambos, apontando o algoritmo MLP com o melhor resultado e o algoritmo de Regressão com o pior. No entanto, tratando-se sobre o tempo de execução, apesar dos valores apresentarem oscilações e não serem constantes, o algoritmo MLP foi o que obteu o maior tempo de execução disparadamente, enquanto

que os algoritmos de Naive Bayes e Regressão Linear apresentaram números equivalentes ou quase em determinadas execuções.

Conclusão

Portanto, diante dos resultados obtidos a respeito da predição da qualidade do leite, a escolha entre os algoritmos dependerá do compromisso desejado entre acurácia e tempo de execução, sendo o MLP preferível em cenários em que a precisão é prioritária, apesar de maior demanda de tempo. Em contrapartida, os algoritmos de Naive Bayes e Regressão Linear oferecem um equilíbrio aceitável entre acurácia e eficiência temporal, tornando-os alternativas viáveis em determinados contextos.

References

[1] Bueno, M. L., Stemmer, M. R., and Borges, P. Inspeção visual automatica de peças cerâmicas via inteligencia artificial. *Cerâmica Industrial 5*, 5 (2000), 29–37.

[2] da Silva, R. M. Processo de industrialização do leite.

[3] Junior, A. B. A., and da Silva, W. G. Aprendizado de máquina aplicado a avaliação da qualidade de frutos. *Academic Journal on Computing, Engineering and Applied Mathematics 4*, 2 (2023), 83–86.

[4] Ludermir, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados 35* (2021), 85–94.

[5] Malakin, L. A., Mielke, L. V., Abreu, L. R. d., Barrofaldi, R. d. C. Z., Miranda Neto, M., Louzada, F., and Rezende, S. O. Classificação de defeito em lotes numa indústria farmacêutica: uma abordagem prática com aprendizado de máquina em processos de qualidade. *Anais* (2021).

[6] Martins, M. E. G. Regressão linear simples. *Revista de Ciência Elementar 7*, 3 (2019).

[7] Pal, S. K., and Mitra, S. Multilayer perceptron, fuzzy sets, classifiaction.

[8] Sellitto, M. A. Inteligência artificial: uma aplicação em uma indústria de processo contínuo. *Gestão & Produção 9* (2002), 363–376.

[9] Shrijayan Rajendran. Milk quality prediction. Disponível em: https://www.kaggle.com/datasets/cpluzshrijayan/milkquality. Acesso em 5 de dezembro 2023, 2022.

[10] VIEIRA, W. D. C. Analisando cerveja artesanal por meio de 3 modelos de classificação e aprendizado de máquina.

[11] Webb, G. I., Keogh, E., and Miikkulainen, R. Naïve bayes. *Encyclopedia of machine learning 15*, 1 (2010), 713–714.