UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JOSÉ AUGUSTO BEZERRA NETO

PREDIÇÃO DO PREÇO DO LEITE UTILIZANDO DADOS DE COOPERATIVAS

CAMPINA GRANDE - PB

2022

JOSÉ AUGUSTO BEZERRA NETO

PREDIÇÃO DO PREÇO DO LEITE UTILIZANDO DADOS DE COOPERATIVAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Eanes Torres Pereira

CAMPINA GRANDE - PB

2022

JOSÉ AUGUSTO BEZERRA NETO

PREDIÇÃO DO PREÇO DO LEITE UTILIZANDO DADOS DE COOPERATIVAS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Eanes Torres Pereira

Orientador – UASC/CEEI/UFCG

Professor Dr. Herman Martins Gomes

Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni

Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 07 de abril de 2022.

CAMPINA GRANDE - PB

ABSTRACT

Currently, the volume of data generated is increasing more and more, making it possible to use techniques that detect patterns, generating relevant information for those who need it. One such technique is supervised learning, which is one of the machine learning techniques that deals with pattern recognition. This technique explores the construction of algorithms capable of learning and making predictions from a data set. Thus, a model trained with milk data would be able to suggest prices to be paid to the milk producer. Generally, the producer does not know for sure how much he will receive per liter of milk produced, because the milk is delivered during a month and he will only know the value for sure in the next month, causing uncertainty. To alleviate this problem, this work trained a regression model capable of predicting the price of milk, obtaining metrics $R^2 = 0.98$ and RMSE = 0.014.

Predição do Preço do Leite Utilizando Dados de Cooperativas

José Augusto Bezerra Neto
jose.augusto.neto@ccc.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

Eanes Torres Pereira
eanes@computacao.ufcg.edu.br
Universidade Federal de Campina Grande
Campina Grande, Paraíba, Brasil

Resumo

Na atualidade, o volume de dados gerados vem aumentando cada vez mais, tornando possível a utilização de técnicas que detectam padrões, gerando informação relevante para aquele que necessita. Uma dessas técnicas é a aprendizagem supervisionada, que é uma das técnicas de aprendizado de máquina que lida com o reconhecimento de padrões. Essa técnica explora a construção de algoritmos, capazes de aprender e realizar predições a partir de um conjunto de dados. Assim, um modelo treinado com dados de leite seria capaz de sugerir preços a serem pagos ao produtor de leite. Geralmente, o produtor não sabe ao certo o quanto receberá por litro de leite produzido, pois, a entrega do leite é feita durante um mês e ele só saberá do valor ao certo no próximo mês, ocasionando uma incerteza. Para amenizar esse problema, este trabalho treinou um modelo de regressão capaz de predizer o preço do leite, obtendo métricas $R^2 = 0,98$ e RMSE = 0,014.

Palavras-Chave

Aprendizagem supervisionada, modelo de regressão, predição do preço do leite, leite.

1. Introdução

A produção de leite se estende por todo o território brasileiro, onde as regiões Sudeste, Sul e Centro-Oeste são as mais produtivas, com Minas Gerais sendo o maior produtor de leite atualmente. O leite é um produto de grande importância do setor agropecuário, em termos econômicos e alimentares. Segundo o IBGE, em 2020 foram adquiridos mais de 20 bilhões[1] de litros de leite cru pelos estabelecimentos que atuam sob algum tipo de inspeção sanitária [5].

O mercado de leite sofre muitas variações nos preços, principalmente em períodos de entressafra, ocasionando uma produção menor do leite e o aumento do custo de produção, com adição de suplementos e mão de obra, dificultando o lado do produtor. Um fator que impacta o produtor é a existência de mais de uma política para precificar o leite, em que algumas empresas levam em conta só a quantidade de leite produzido, só a qualidade ou ambas. Assim, gerando uma discriminação de preços por parte das empresas que pagam preços diferentes pelo mesmo produto. Outro fator que preocupa o produtor é a incerteza do quanto ele vai receber por litro de leite produzido. O produtor entrega o leite diariamente e só saberá o quanto vai receber após um mês de entrega ou mais.

Uma forma de ajudar na organização do produtor para períodos de entressafra e uma forma de amenizar a preocupação é a predição de preços futuros. Essa predição se dá pela coleta de preços pagos anteriormente aplicados a métodos estatísticos de série temporal para prever os preços dos meses subsequentes. Outra técnica de predição a partir de dados anteriores é a aprendizagem supervisionada, que lida com o reconhecimento de padrões. Um bônus que a aprendizagem supervisionada possui em comparação com a mencionada anteriormente é a possibilidade de adicionar um conjunto de variáveis maior para gerar um resultado mais preciso, tornando a solução mais robusta. Dessa forma, o presente trabalho tem como objetivo treinar um modelo de regressão para predizer o preço do leite, utilizando dados de cooperativas referentes ao preço pago anteriormente, a qualidade do leite e os preços dos derivados de leite.

Este Trabalho de Conclusão de Curso (TCC) foi realizado em cooperação com um representante de cooperativas de leite. Como há o objetivo de desenvolver um software comercial para predição do preço do leite para as cooperativas de leite do Brasil, alguns detalhes técnicos foram intencionalmente omitidos deste relatório visando preservar propriedade intelectual estratégica para o software que será desenvolvido. Dentre os detalhes que foram intencionalmente omitidos estão: tipo de algoritmo de regressão específico, parâmetros de treinamento e teste do algoritmo, variáveis utilizadas para regressão e fontes dos dados obtidos. Considera-se que a omissão desses detalhes não reduzem o mérito deste TCC.

Inicialmente, a Seção 2 apresenta a fundamentação teórica, descrevendo alguns algoritmos de aprendizagem. A Seção 3 lista alguns trabalhos relacionados com a pesquisa realizada. A Seção 4 apresenta a metodologia utilizada. A Seção 5 relata os resultados e discussões. Por fim, na Seção 6 é descrita as considerações finais do trabalho realizado.

2. Fundamentação teórica

Existem vários algoritmos de aprendizagem supervisionada para o reconhecimento de padrões. Esta seção, introduz os conceitos

[1] Soma dos quatro trimestres de 2020.

básicos sobre alguns dos principais algoritmos utilizados na aprendizagem supervisionada.

2.1 Regressão linear

Segundo James et al. [3] na regressão linear simples, temos um algoritmo que descreve o relacionamento entre duas variáveis através de uma equação matemática, uma variável dependente e a outra independente, estimando valores para a variável dependente, com base nos valores conhecidos da variável independente.

Ele também descreve que a regressão múltipla, é uma extensão da simples, utiliza duas ou mais variáveis independentes para a predição. Na prática é um algoritmo mais aplicável por permitir que mais de um preditor possa ser utilizado, como por exemplo em dados de vendas e publicidade. Como verificar o impacto nas vendas com publicidade em TV, jornal e rádio, nesse caso temos três preditores para a venda, com a regressão simples ficaria difícil realizar essa análise.

2.2 Regressão logística

Segundo James et al. [3] a regressão logística é um modelo utilizado quando a variável dependente é categórica, geralmente o resultado é binário, 0 ou 1. É um modelo com resultado que mostra a probabilidade daquele evento acontecer. Diferente da regressão linear, a regressão logística utiliza uma função logística no modelo, de modo a garantir que o valor da probabilidade esteja entre 0 e 1, no qual a regressão linear não garante isso.

2.3 Redes neurais

As redes neurais são uma categoria de algoritmos que simulam o comportamento de neurônios, dispostos em um conjunto de camadas que podem ser ativados ou não, a depender das funções de ativação em cada nó.

A Figura 1 ilustra uma representação das camadas de uma rede neural. A camada de entrada são os dados que serão utilizados para a rede aprender, cada uma das ligações com a camada intermediária possui um peso associado que ao se conectarem com o nó na camada intermediária é passado para uma função de ativação do nó que irá determinar se aquele nó será ativado ou não e por fim os nós intermediários são conectados a camada de saída que refere-se ao resultado final. Em redes mais complexas existe mais de uma camada intermediária, com o mesmo processo de pesos nas conexões e passados pela função de ativação. Segundo Bishop [2], o processo de aprendizagem ocorre por meio do algoritmo chamado *backpropagation,* algoritmo esse que propaga o erro entre as camadas, para frente e para trás, dessa forma é possível determinar o gradiente da função de erro. Assim [2] descreve que a partir do gradiente da função de erro são feitos cálculos dos ajustes nos pesos, permitindo assim que a rede neural aprenda.

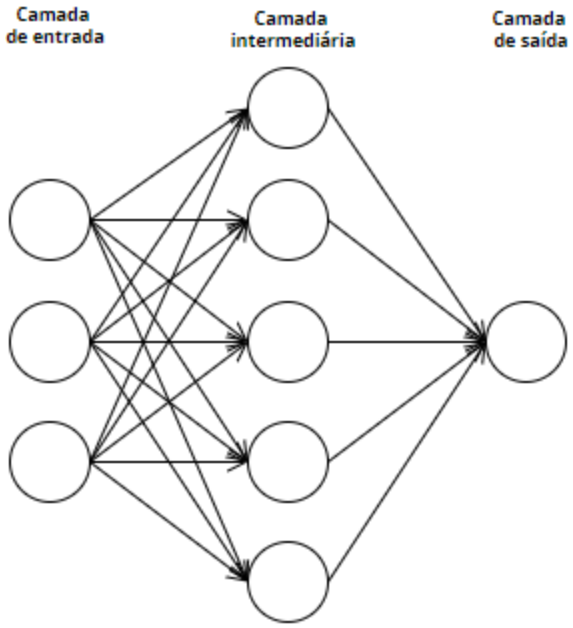

Figura 1: Representação de uma rede neural

3. Trabalhos relacionados

Alves et al. [1] aplicaram o método Box-Jenkins para previsão de preços de leite e para identificar variações sazonais nos preços recebidos pelos produtores de leite. Foi utilizada uma série de preços médios pagos ao produtor de janeiro de 2000 a julho de 2014. Com o modelo X-12 ARIMA foi possível identificar sazonalidade nos preços referentes ao leite, e dessa forma pôde-se ajustar o modelo para realizar a previsão de preços, e foi identificado que no período da previsão o valor a ser recebido pelos produtores seria menor.

Sarmento, Beduschi e Zen [4] também aplicaram o método Box-Jenkins para previsão de preço do leite, com um diferencial de que se utilizou uma função de transferência para tornar mais preciso o modelo desenvolvido. Foram realizados testes em cinco estados no Brasil de janeiro de 1986 a outubro de 2006. Para a função de transferência foi considerada a taxa de câmbio como variável explicativa para o preço do leite. Como resultado, a inclusão da função de transferência foi positiva para três estados analisados, necessitando de novos estudos para os outros dois restantes.

4. Metodologia

Para a obtenção dos resultados deste trabalho foram realizadas as seguintes etapas: (i) coleta de dados; (ii) validação dos resultados; (iii) implementação de algoritmo de regressão e avaliação.

4.1 Coleta de dados

Os dados utilizados neste trabalho foram disponibilizados por algumas cooperativas de leite no Brasil, em um período de 18

meses, totalizando 590 amostras dos produtores de leite associados às cooperativas, com informações sobre a qualidade do leite, preço pago ao produtor e os preços dos derivados de leite comercializados pela indústria de laticínios.

O passo seguinte foi realizar a limpeza nesses dados, visto que algumas amostras não possuíam todos os dados referentes ao leite, o que resultou num total de 409 amostras finais para serem utilizadas.

Em seguida foi realizada a normalização dos dados para garantir que todas as variáveis possuíssem o mesmo grau de importância no treinamento.

4.2 Validação dos resultados

Para validar o modelo produzido é necessário dividir o conjunto de dados em treinamento e teste, de forma que essa separação seja mutuamente exclusiva. Essa separação pode ser realizada de forma aleatória, e pode utilizar uma técnica chamada de validação cruzada, que particiona o conjunto em vários subconjuntos mutuamente exclusivos, e a partir desses subconjuntos é realizado o treinamento do modelo, validando com os subconjuntos de teste correspondentes. Na Figura 2, é apresentada uma representação visual dessa técnica.

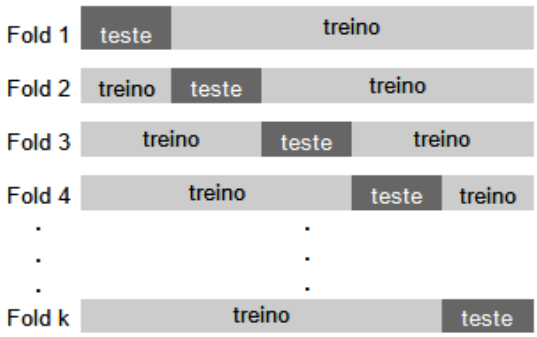

Figura 2: Representação da validação cruzada.

As métricas utilizadas para avaliar o desempenho do modelo foram: R-quadrado (R^2) e raiz do erro quadrático médio (RMSE) com os dados de treinamento segmentados em 10 partes (k-fold = 10).

4.2.1 R-quadrado (R^2)

R-quadrado ou coeficiente de determinação é uma medida estatística que mostra o quanto o modelo está ajustado aos dados. Segundo James et al. [3] a função se dá da seguinte forma: $R^2 = 1 - \frac{RSS}{TSS}$.

O RSS é a soma residual dos quadrados, que mede a quantidade de variabilidade que o modelo não explica após a regressão, em que n é a quantidade de observações, y_i é o valor observado e $\widehat{y_i}$ é o valor estimado (previsão).

$$RSS = \sum_{i=1}^{n} \left(y_i - \widehat{y_i}\right)^2$$

O TSS é a soma total dos quadrados, que mede a variância total entre a variável dependente e sua média geral, em que y_i é o valor observado e $\overline{y}$ é a média das observações.

$$TSS = \Sigma \left(y_i - \overline{y}\right)^2$$

James et al. [3] descreve que se o R^2 está próximo de 1 indica que a regressão explicou uma grande proporção da variabilidade, de forma oposta se o valor estiver próximo a 0 indica que a regressão não explicou bem a variabilidade. Assim quanto mais próximo de 1 o valor de R^2 estiver, melhor o modelo de regressão analisado.

4.2.2 Raiz do erro quadrático médio (RMSE)

Raiz do erro quadrático médio é uma métrica que calcula a raiz da média da diferença entre o valor real e o predito ao quadrado. A equação do RMSE se dá da seguinte forma: $RMSE = \sqrt{MSE}$. No qual o MSE é o erro quadrático médio. A diferença entre o RMSE e o MSE está no valor quadrático que uma tem e outra não. Logo o RMSE é melhor para análise e pode ser representado em porcentagem.

Segundo James et al. [3] a função do MSE é a seguinte: $MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{f}(x_i)\right)^2$, onde $\widehat{f}\left(x_i\right)$ é a previsão para a i-ésima observação. A variação do valor do MSE é de 0 a infinito, e quanto menor o valor do MSE melhor o modelo analisado.

4.3 Algoritmo de regressão

Para este trabalho, foi escolhido um modelo de regressão para predizer os preços do leite pagos aos produtores. Um modelo de regressão é um algoritmo de aprendizagem supervisionada em que um algoritmo é treinado para prever uma saída a partir de um conjunto de valores contínuos.

Esta etapa consistiu em executar o algoritmo de regressão e ajustar os hiperparâmetros de forma a maximizar as métricas R^2 e RMSE como mostrado na Figura 3.

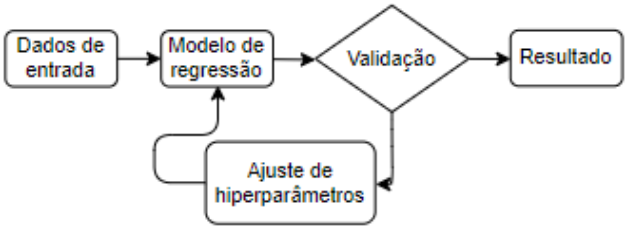

Figura 3: Etapas do treinamento do modelo de regressão

5. Resultados

Nesta seção, são apresentados os resultados dos treinamentos realizados.

Após ajustar os hiperparâmetros do modelo de regressão utilizado, realizamos a validação do modelo utilizando a validação cruzada com 10 *folds*. Na Tabela 1, podemos

visualizar o resultado das métricas de cada *fold*, com a média e o desvio padrão no treino e no teste.

Podemos observar que em cada subconjunto dos dados na validação cruzada o modelo se ajustou muito bem aos dados, com R^2 por volta de 98% e com um erro médio de $\pm\ 1,4\%$. O desvio padrão também nos mostra que o conjunto de resultados é bem homogêneo, com os valores próximos da média.

Outra forma de visualizar os resultados obtidos está presente na Tabela 2, que possui as diferenças de preços do valor real com o valor predito acumulados. Nessa tabela, podemos observar que a diferença entre o valor real pago ao produtor e o valor predito pelo modelo em sua grande parte está entre R$ 0,01 e R$ 0,005. A diferença de um centavo é aceitável para a precificação do leite. Na Figura 4, temos outra visualização do subconjunto de validação dos preços a serem pagos aos produtores de leite, mais especificamente o k-fold = 2.

Logo, podemos dizer que o modelo de regressão implementado neste trabalho é capaz de predizer o preço a ser pago ao produtor de leite com alta precisão, podendo ser utilizado pelas cooperativas com a finalidade de auxiliar nas tomadas de decisão, sobre como precificar de forma mais justa os produtores associados.

k-fold	**R^2**	**RMSE treinamento**	**RMSE teste**
1	0,980182	0,012178	0,013243
2	0,986679	0,010989	0,013232
3	0,98151	0,013628	0,018836
4	0,983107	0,015451	0,015746
5	0,973356	0,011787	0,01617
6	0,985996	0,014436	0,012959
7	0,986069	0,013955	0,014263
8	0,984342	0,01341	0,013097
9	0,972887	0,011811	0,012542
10	0,979165	0,009301	0,011706
Média	**0,9813293**	**0,0126946**	**0,0141794**
Desv. Pad.	**0,005010853**	**0,001822660**	**0,002143253**

Tabela 1: R^2 e RMSE da validação cruzada do modelo de regressão

k-fold	$\geq\ 0,3$	$\geq\ 0,2$	$\geq\ 0,1$	$\geq\ 0,05$	$\geq\ 0,01$	$\geq\ 0,005$	$\geq\ 0,001$	$\geq\ 0,0001$	$< 0,0001$
1	0	0	0	0	18	33	40	41	0
2	0	0	0	0	18	36	40	41	0
3	0	0	0	1	21	28	39	41	0
4	0	0	0	0	20	30	37	40	1
5	0	0	0	0	24	32	38	41	0
6	0	0	0	0	21	30	40	40	1
7	0	0	0	0	21	32	40	41	0
8	0	0	0	0	19	28	41	41	0
9	0	0	0	0	17	27	37	41	0
10	0	0	0	0	14	20	37	40	0

Tabela 2: Diferença acumulada em reais, entre o valor real do leite pago ao produtor e o valor predito pelo modelo

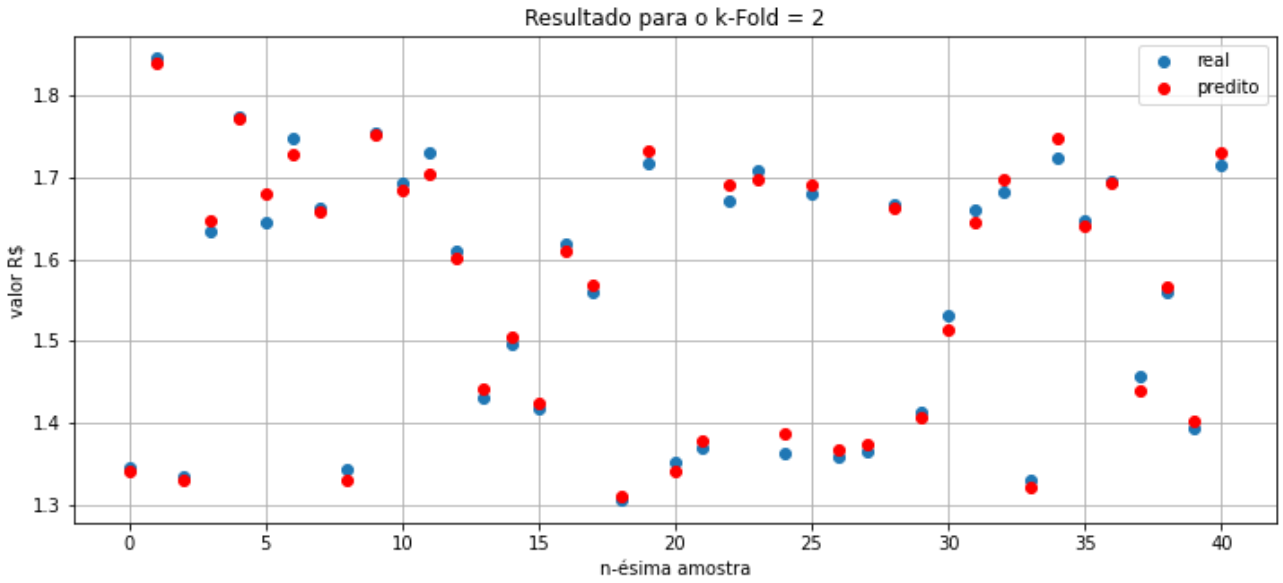

Figura 4: Preços reais do leite pago ao produtor e preços preditos pelo modelo do subconjunto da validação cruzada

6. Considerações finais

Neste trabalho, foram aplicadas técnicas de aprendizagem supervisionada, para predizer o preço a ser pago ao produtor de leite. Através dos resultados obtidos, pode-se concluir que o modelo de regressão treinado é capaz de realizar a predição dos preços de leite, com uma boa confiança, dado as métricas obtidas, o tornando útil para o uso nas cooperativas, a fim de auxiliar nas tomadas de decisão.

Por outro lado, é importante mencionar que esta abordagem pode ter suas limitações, visto que a amostra de dados foi pequena. Dessa forma, sugere-se, que em trabalhos futuros, a utilização de um histórico maior de dados, possa melhorar esse aspecto . Com um conjunto de dados maior é possível melhorar ainda mais a precisão e confiança do modelo produzido.

7. Agradecimentos

Gostaria de agradecer a minha família, por me apoiarem em todos os momentos da minha vida, principalmente nos mais difíceis, ao meu professor orientador Eanes, por todo o direcionamento e conhecimento passado, aos meus colegas de classe e demais professores que fizeram parte dessa minha jornada.

8. Referências

[1] ALVES, F. F.; SOUSA, L. V. de C.; ERVILHA, G. T. Planejamento e Previsão do Preço do Leite em Minas Gerais: Análise Empírica com Base no Modelo X12- ARIMA. **Revista de Economia e Agronegócio**, [S. l.], v. 12, n. 1,2,3, 2015.

[2] BISHOP, C. M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

[3] JAMES, Gareth et al. **An introduction to statistical learning**. New York: Springer, 2013.

[4] SARMENTO, P. H. L.; BEDUSCHI, G.; ZEN, S. **Função de Transferência Para Previsão de Preço do Leite nos Principais Estados Produtores**. *In:* XLV Congresso da Sociedade Brasileira de Economia, Administração e Sociologia Rural, jul. 2007, Londrina, PR.

[5] IBGE. **Indicadores IBGE: estatística da produção pecuária**. Rio de Janeiro: IBGE, 2006. Disponível em: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=72380 Acesso em: 10 mar. 2022.