Ciência de dados aplicada à produção de maçãs: integração de dados de produção e indicadores meteorológicos para análise e previsão de produtividade

Yuri Padilha Alves[1], Henrique Dias Pereira[2]

Resumo

O clima é um fator fundamental para a agricultura e entender a influência das variáveis meteorológicas na produção é essencial para esse setor. O presente artigo apresenta o desenvolvimento de um projeto de ciência de dados que tem como objetivo gerar conhecimentos referentes à produção de maçãs através do relacionamento de dados históricos de produção e de variáveis meteorológicas e dessa forma, prever números de produção de safras futuras. A partir dos resultados gerados, espera-se que os mesmos se façam importantes para que as ações de tomada de decisão se tornem mais assertivas. Os resultados provenientes da análise de dados refletem algumas conclusões apresentadas na literatura, mas também indicam novas situações que merecem atenção. Com relação à predição de produção para safras futuras, os modelos de aprendizado de máquina se mostraram satisfatórios, apresentando números previstos próximos ao real e bons índices nas métricas de avaliação.

Palavras-chave

Análise de Dados, Aprendizado de Máquina, Fatores Climáticos, Previsão de Safra, Produção de Maçãs.

Data science applied to apple production: integration of production data and meteorological indicators for productivity analysis and forecasting

Abstract

The climate is a fundamental factor for agriculture and understanding the influence of weather variables on production is essential for this sector. This paper presents the development of a data science project that aims to generate knowledge about the production of apples through the relationship of historical data of production and weather variables and thus predict production numbers of future crops. From the results generated, it is expected that they will be important for the decision-making actions to become more assertive. The results from the data analysis reflect some conclusions presented in the literature, but also indicate new situations that deserve attention. Regarding the production prediction for future harvests, the machine learning models proved to be satisfactory, presenting predicted numbers close to the real ones and good indexes in the evaluation metrics.

Keywords

Data Analysis, Machine Learning, Climatic Factors, Crop Forecast, Apple Production.

I. Introdução

A prática agrícola é imprescindível para a sobrevivência da população humana e são evidentes os efeitos do clima sobre ela. Na fruticultura, as condições climáticas tendem a impactar em situações como o volume de produção dos frutos bem como sua cor, sabor e tamanho, por exemplo.

De acordo com Blain (2011), espécies frutíferas de clima temperado apresentam um período de repouso invernal fortemente condicionado pelas condições climáticas vigentes. Para cada novo ciclo vegetativo ser iniciado, é necessário que determinada espécie criófila sofra a ação de baixas temperaturas.

Na cultura da maçã algumas variáveis meteorológicas tendem a afetar o desenvolvimento do fruto, causando impacto no seu tamanho. Leite et al. (2021) destacam que a falta de frio hibernal, o excesso de chuvas durante a floração, temperaturas altas durante o ciclo vegetativo e estresse hídrico, são os eventos mais comuns que levam à redução da produção.

Dessa maneira, o projeto proposto visa a análise do histórico meteorológico confrontando com dados de produção, permitindo a validação dos comportamentos descritos em estudos realizados, podendo auxiliar na identificação de padrões e na previsão de produção de safras futuras.

O artigo está organizado da seguinte forma: a seção II busca promover a compreensão de conceitos relacionados ao projeto proposto, apresentando definições, funcionalidades,

[1]Aluno do curso de especialização em ciência de dados, Universidade de Caxias do Sul, Caxias do Sul, RS, Brasil; [2] Professor mestre do curso de especialização em ciência de dados, Universidade de Caxias do Sul, Caxias do Sul, RS, Brasil.

tecnologias, ferramentas e métodos utilizados no processo de desenvolvimento. Na seção III são expostos e detalhados alguns conhecimentos obtidos por meio do estudo. A seção IV traz as conclusões acerca da pesquisa bem como objetivos futuros e ao final a seção V expõe a bibliografia utilizada como embasamento para o artigo.

II. Materiais e Métodos

A. *Fontes de Dados*

O desenvolvimento do projeto fez uso de dados meteorológicos públicos coletados em uma estação meteorológica instalada na cidade de Vacaria, RS, Brasil e disponibilizados no formato .CSV no banco de dados meteorológicos do INMET (Instituto Nacional de Meteorologia) e de dados de produção de maçã, armazenados em um banco de dados Oracle, de cinco pomares de uma empresa também localizada na cidade de Vacaria, RS, Brasil e que possui uma área plantada total de 1.079 hectares.

Foram coletados dados relacionados às safras de 2010 até 2022, sendo que o período de safra é definido através do ciclo produtivo da maçã que é dividido em etapas que iniciam no mês de maio de um ano e terminam no final do mês de abril do próximo ano, conforme ilustrado na figura 1.

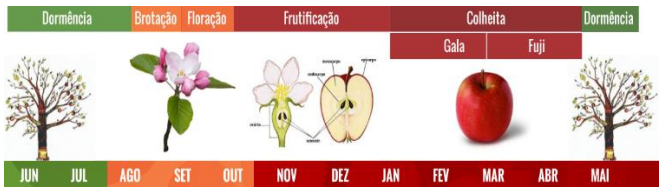

Figura 1: Ciclo de Produção da Maçã.
Fonte: Pomi Frutas.

Os dados de produção foram separados em grupos de calibre e tamanho da fruta para geração das análises. Segundo a FAEP (Federação da Agricultura do Estado do Paraná) (2022), o calibre pode ser classificado de acordo com o peso dos frutos (limites inferior e superior médios) expresso em gramas, sendo enquadrado em uma das classes estabelecidas na figura 2.

Classes ou Calibres	Peso Limite Inferior (em g)	Peso Limite Superior (em g)
60	279	-
70	241	278
80	213	240
90	190	212
100	172	189
110	157	171
120	142	158
135	127	141
150	115	126
165	105	114
180	96	104
198	87	95
220	78	86
250	67	77
300	50	66

Figura 2: Classes ou calibres da maçã, com base no número de frutos contidos em uma caixa com capacidade para conter 18 kg do produto.
Fonte: FAEP.

Os grupos de tamanho da fruta são estabelecidos de acordo com o calibre pela empresa que disponibilizou os dados para o estudo, conforme exposto na figura 3.

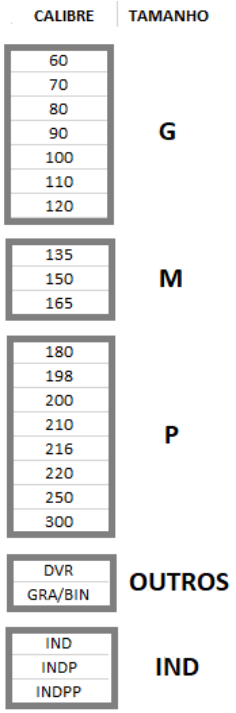

Figura 3: Grupos de tamanho da maçã.
Fonte: Próprio autor.

Os grupos G, M e P representam, respectivamente, os tamanhos grande, médio e pequeno, o grupo OUTROS consiste em calibres diversos, que não foram previamente classificados, e frutas a granel, já no grupo IND são consideradas frutas que serão comercializadas para a produção de produtos industrializados, como suco e vinagre, por exemplo.

B. *Extração, Transformação e Carga - ETL*

Para extrair, transformar e disponibilizar os dados para utilização foram desenvolvidos códigos na linguagem de programação Python que de acordo com Agarwal e Agarwal (2005) é uma linguagem de programação de propósito geral, que combina os paradigmas procedural, funcional e orientado a objetos e Sousa (2020) indica que sua utilização para desenvolver ETL em código é extremamente útil, trazendo versatilidade, versionamento e clareza no desenvolvimento de todo o processo.

Os códigos foram escritos e executados através do Jupyter Notebook que Pimentel et al. (2021) definem como uma ferramenta que através de sua interatividade permite análises de dados em tempo real, com o processo sendo documentado durante o desenvolvimento, resultados sendo exibidos de forma instantânea e discutidos imediatamente em linguagem natural.

Os resultados do processo de ETL foram exportados para arquivos .CSV que de forma simples e objetiva, segundo Oliveira (2015), podemos definir como sendo um arquivo de dados estruturados separados por vírgula e com a adição de algumas restrições.

C. *Mineração de Dados*

A etapa de mineração de dados foi desenvolvida utilizando a linguagem de programação Python e a ferramenta Jupyter Notebook, assim como na etapa de ETL.

Para analisar a correlação entre as variáveis meteorológicas e a quantidade de produção da maçã, foi utilizada a função Corr() da biblioteca Pandas. Coelho (2017), define Pandas como uma biblioteca licenciada com código aberto que oferece estruturas de dados de alto desempenho e de fácil utilização para a análise de dados, já Miranda et al. (2020), informa que a função Corr() tem como resultado uma matriz que informa a correlação entre cada par de colunas de um conjunto de dados, utilizando o método de Pearson, que é um cálculo desenvolvido por Karl Pearson que mede a correlação linear entre duas variáveis, tendo como resultado um valor entre -1 e 1.

A regressão linear múltipla foi o método escolhido para o modelo de predição de produção de safras futuras. Segundo Guimarães (2012), a regressão linear múltipla possui o objetivo de obter uma equação que explique satisfatoriamente a relação entre uma variável resposta Y (dependente) e duas ou mais variáveis explicativas X (independentes), possibilitando fazer a predição de valores de interesse.

Para obter as melhores *features* para o modelo de predição foi aplicado o método SelectKBest, da biblioteca Scikit-Learn, que Coppini (2019) define como um método que funciona de maneira que, dado um número inteiro k e um conjunto de dados X, retorna as k melhores *features* de X.

As métricas de avaliação utilizadas para o modelo de predição foram o Erro Percentual Absoluto Médio (MAPE), que de acordo com Kristiansen (2018) pode ser definido como a diferença absoluta média entre o valor real e o valor predito dividido pelo valor real e a Raiz do Erro Quadrático Médio (RMSE), que é calculada através da diferença entre a predição e o valor real, considerando um peso maior para desvios grandes pois são elevados ao quadrado, segundo dos Santos (2022).

Com o objetivo de gravar o modelo de predição em disco, para utilizá-lo após realizar os processos de treino e teste sem ter a necessidade de refazer estes dois processos, foi utilizada a função Dump() da biblioteca Joblib, que a comunidade Joblib (2022) apresenta como uma função que permite persistir um objeto Python arbitrário em um arquivo.

D. *Apresentação dos Resultados*

A ferramenta de *business intelligence* Microsoft Power BI foi utilizada para o agrupamento e apresentação dos resultados do projeto. A comunidade Microsoft (2022) informa que o Power BI é uma coleção de serviços de software, aplicações e conectores que funcionam em conjunto para transformar as origens de dados não relacionadas em informações coerentes, visualmente envolventes e interativas.

III. RESULTADOS

Para apresentação dos resultados foi desenvolvido um relatório no Microsoft Power BI que agrupa todos os dados provenientes da etapa de mineração de dados e está dividido em quatro páginas que são: Histórico Meteorológico, Histórico de Produção, Análise de Correlação e Análise Preditiva. O projeto pode ser acessado na íntegra no GitHub por meio do link a seguir: https://github.com/ypalves/MeteorologiaXProducao-Maca.git

A. *Histórico Meteorológico*

Nessa página são apresentados os dados históricos coletados no banco de dados meteorológicos do INMET no período de 01/05/2009 até 30/04/2021, que correspondem às safras de 2010 até 2021.

Na tela inicial são expostas em cartões as informações gerais de todas as variáveis presentes no conjunto de dados, sendo possível filtrar por safra ou visualizar os totais, conforme exibido na figura 4.

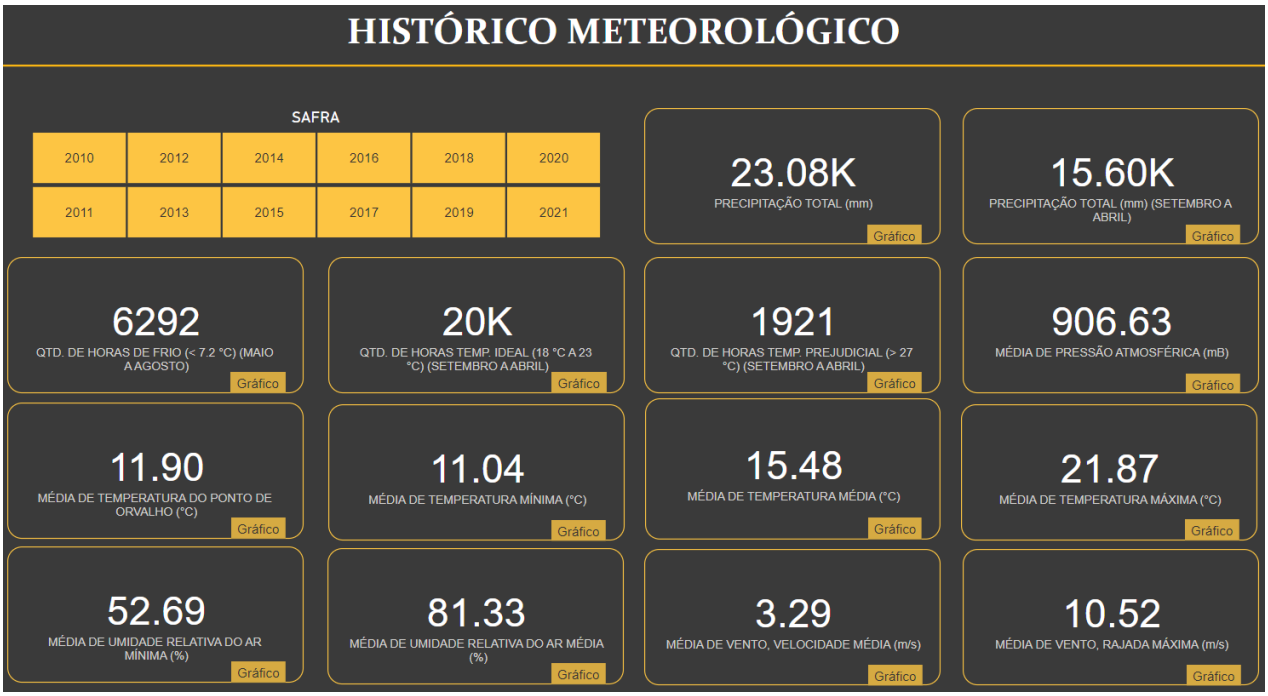

Figura 4: Histórico meteorológico – Informações gerais.
Fonte: Próprio autor.

A partir da tela de informações gerais é possível acessar uma visualização gráfica disponível em cada cartão que demonstra e compara o comportamento das variáveis. A figura 5 apresenta o gráfico do histórico da média de temperatura máxima por safra, medida em graus Celsius, onde se observa uma tendência de aumento, quando comparadas as safras iniciais e finais.

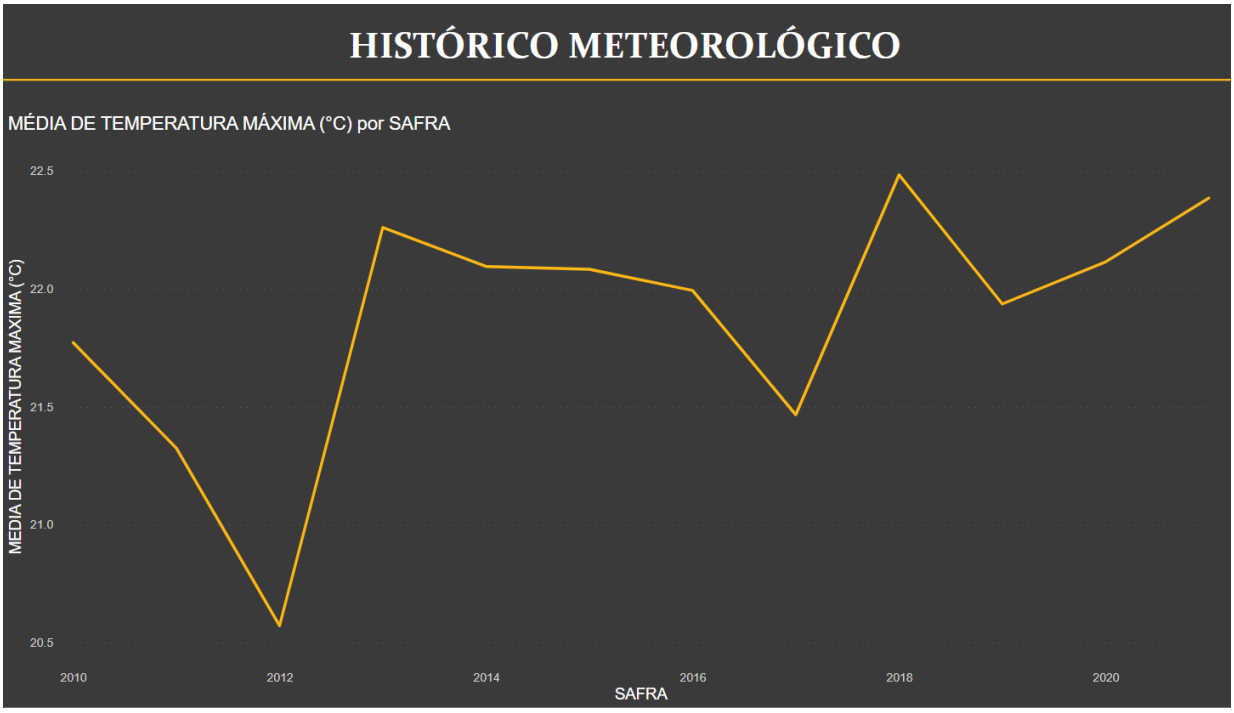

Figura 5: Histórico meteorológico – Gráfico da média de temperatura máxima por safra.
Fonte: Próprio autor.

B. Histórico de Produção

A página de histórico de produção apresenta os dados coletados no banco de dados da empresa referentes ao período de 01/01/2010 até 31/12/2021, que correspondem às safras de 2010 até 2021. Na figura 6 está demonstrada a tela inicial onde são expostos indicadores da quantidade total produzida em quilos, dos 5 calibres mais produzidos, da quantidade produzida em quilos por pomar e da quantidade produzida em quilos por tamanho da fruta, sendo possível filtrar por safra. Até a safra de 2018 os registros de produção não eram carregados para o banco de dados separados por pomar, dessa forma, para essas safras a quantidade produzida de cada pomar está totalizada no que foi denominado de POMAR GERAL.

Figura 6: Histórico de produção – Informações gerais.
Fonte: Próprio autor.

Cada indicador presente na tela de informações gerais possui um visual onde pode ser observada graficamente a evolução dos números durante o período selecionado. Através do gráfico *treemap* de quantidade em quilos por safra e tamanho do produto é possível constatar que nas safras finais se concentram as maiores quantidades de produção e o tamanho M se destaca como o mais produzido em relação aos demais, conforme representado na figura 7.

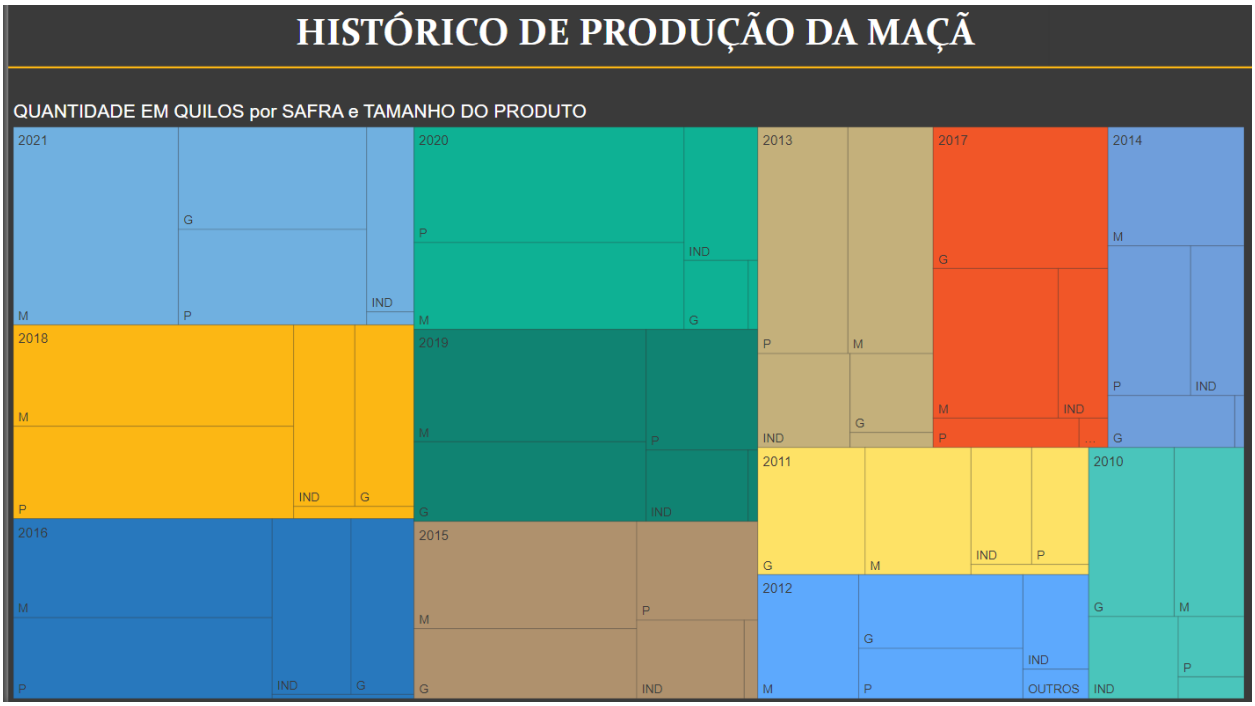

Figura 7: Histórico de produção – Gráfico de quantidade em quilos por safra e tamanho do produto.
Fonte: Próprio autor.

C. *Análise de Correlação*

Essa página exibe os resultados gerados a partir da análise de correlação realizada entre todas as variáveis do conjunto de dados meteorológicos e a quantidade em quilos produzida da maçã, extraída do conjunto de dados de produção. O agrupamento de tamanho da fruta foi o escolhido para essa análise e a figura 8 apresenta a primeira das três divisões da página.

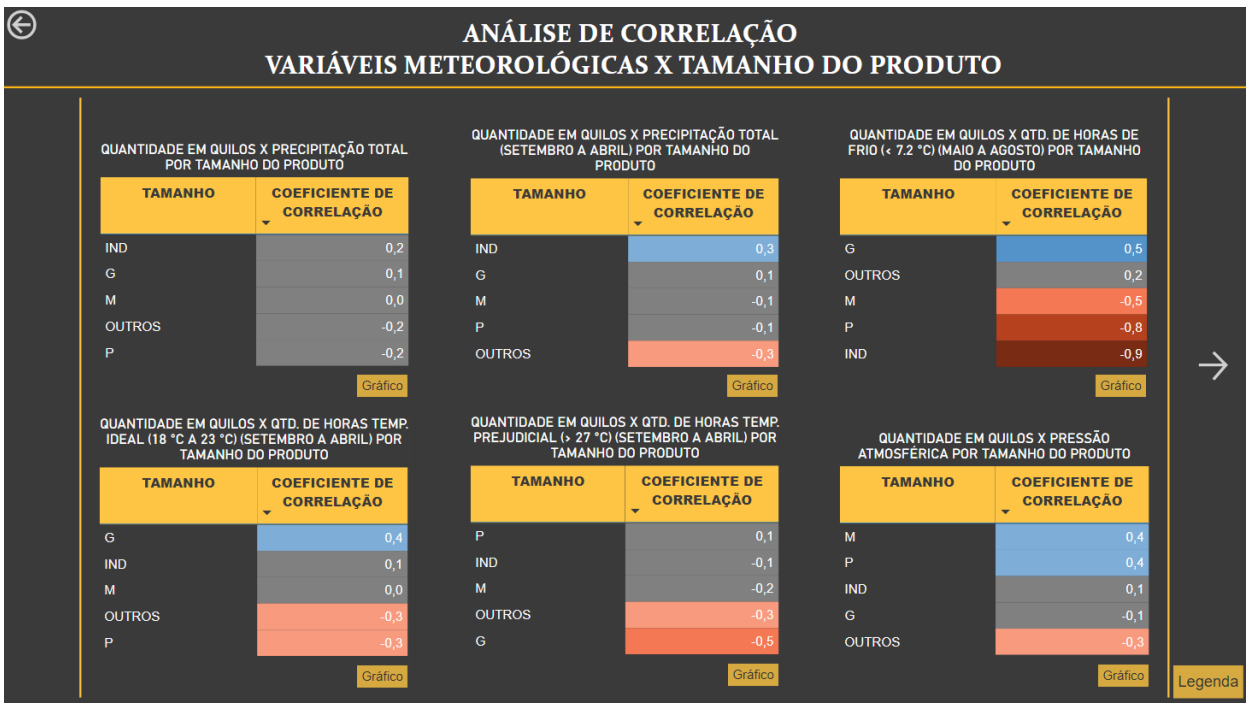

Figura 8: Análise de correlação – Informações gerais – 1ª Tela.
Fonte: Próprio autor.

O resultado da correlação entre as variáveis está representado através do coeficiente de Pearson. Para tornar a visualização mais clara, os níveis de correlação foram separados em cores, sendo assim, a figura 9 expõe a tela de legenda, que pode ser acessada a partir da tela inicial da página da análise de correção e é onde ocorre o detalhamento dos níveis e de qual cor os representa.

COR	NÍVEL DE CORRELAÇÃO
	CORRELAÇÃO NEGATIVA MUITO FORTE (-1 a -0,9)
	CORRELAÇÃO NEGATIVA FORTE (-0,8 a -0,7)
	CORRELAÇÃO NEGATIVA MODERADA (-0,6 a -0,5)
	CORRELAÇÃO NEGATIVA FRACA (-0,4 a -0,3)
	NÃO POSSUI CORRELAÇÃO (-0,2 a 0,2)
	CORRELAÇÃO POSITIVA FRACA (0,3 a 0,4)
	CORRELAÇÃO POSITIVA MODERADA (0,5 a 0,6)
	CORRELAÇÃO POSITIVA FORTE (0,7 a 0,8)
	CORRELAÇÃO POSITIVA MUITO FORTE (0,9 a 1)

Figura 9: Análise de correlação – Legenda.
Fonte: Próprio autor.

Com o objetivo de proporcionar uma visualização gráfica do comportamento da correlação entre as variáveis, foram disponibilizados gráficos que podem ser acessados a partir de cada bloco das telas de informações gerais.

A variável meteorológica que se mostrou mais significativa em relação a quantidade produzida em diversos tamanhos da fruta foi a da quantidade de horas de frio registradas entre os meses de maio e agosto, período de dormência da macieira. Hawerroth e Nachtigall (2016) informam que para que na primavera a macieira apresente brotação e floração uniforme é necessário que as plantas sejam expostas a regimes de baixas temperaturas durante o período de dormência e segundo Leite et al. (2018) no inverno, acúmulos de frio acima de 550 horas com temperatura menor que 7,2 °C são considerados satisfatórios para esses cultivares.

Os gráficos das figuras 10 e 11 permitem a comprovação da importância das horas de frio para a cultura da maçã, onde quando o tamanho da fruta é G a tendência é o aumento da produção se a quantidade de horas de frio for alta, entretanto quando o tamanho da fruta é P acontece o inverso, quanto maior a quantidade de horas de frio, menor a produção.

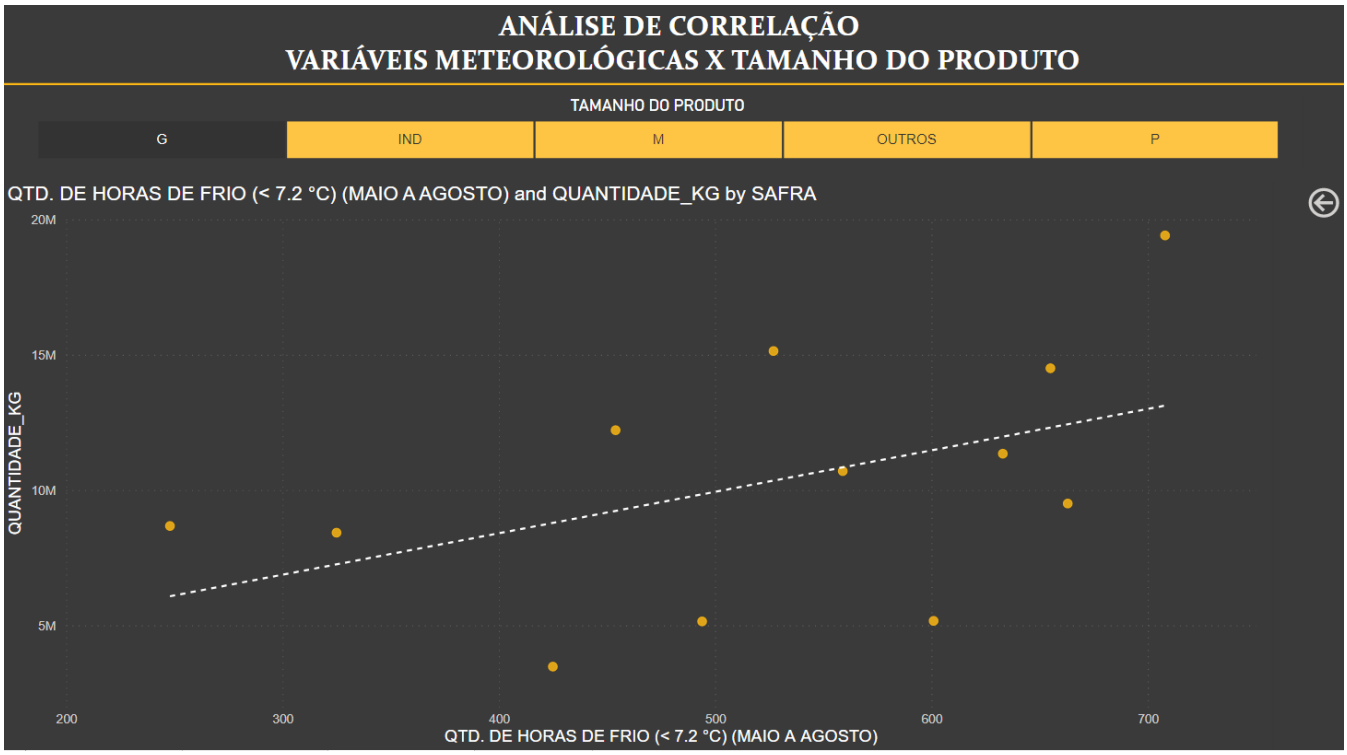

Figura 10: Análise de correlação – Gráfico de quantidade de horas de frio e quantidade em quilos por safra – Maçã de tamanho G.
Fonte: Próprio autor.

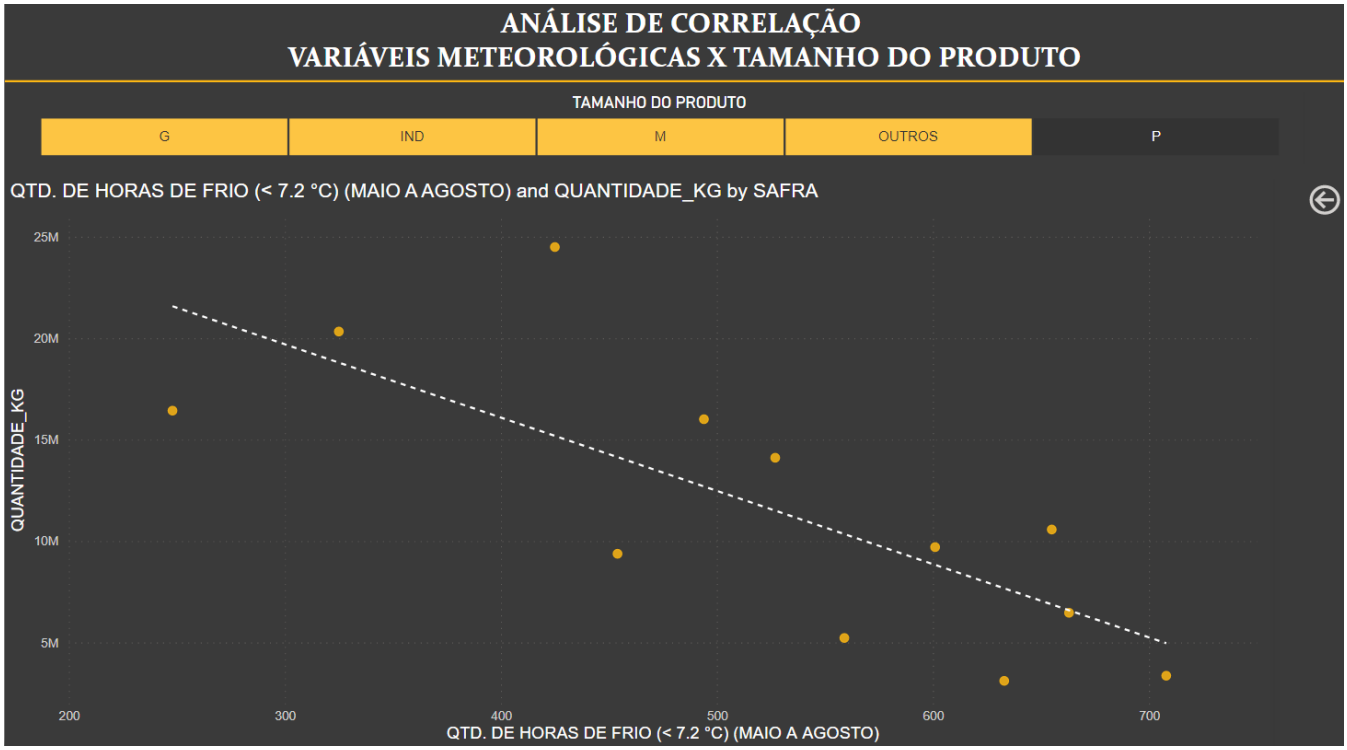

Figura 11: Análise de correlação – Gráfico de quantidade de horas de frio e quantidade em quilos por safra – Maçã de tamanho P.
Fonte: Próprio autor.

D. Análise Preditiva

Para o processo de análise preditiva os dados foram separados por grupo de tamanho da fruta com o intuito de desenvolver um modelo de predição individual para cada tamanho.

Os dados referentes às safras de 2010 até 2020 das *features* selecionadas pelo método SelectKBest foram escolhidos para o treino dos modelos e os dados referentes à safra de 2021 foram utilizados para o teste.

Bons resultados foram retornados nas métricas de avaliação após o desenvolvimento e aplicação dos modelos nos dados de teste. A figura 12 demonstra os valores das métricas MAPE e RMSE dos modelos de predição.

G
MAPE (%): 4,61
RMSE (KG): 697.805,48

M
MAPE (%): 7,91
RMSE (KG): 2.027.413,39

P
MAPE (%): 3,34
RMSE (KG): 471.848,97

OUTROS
MAPE (%): 4,00
RMSE (KG): 19.365,59

IND
MAPE (%): 1,15
RMSE (KG): 78.976,74

Figura 12: Análise preditiva – Métricas de avaliação.
Fonte: Próprio autor.

Na página de análise preditiva do relatório são apresentados blocos para cada tamanho da fruta onde, para a última safra fechada, os números da quantidade em quilos prevista pelo modelo, da quantidade real produzida em quilos e o percentual de diferença entre esses dois números são mostrados, já para a safra em andamento, que ainda não possui a contabilização da quantidade colhida, é apresentada a quantidade prevista pelo modelo. A figura 13 expõe a tela inicial da página de análise preditiva.

Figura 13: Análise preditiva – Informações gerais.
Fonte: Próprio autor.

Cada bloco da tela de informações gerais permite o acesso a uma visualização que compara graficamente os números de quantidade prevista em quilos e quantidade real em quilos da safra fechada e o número de quantidade prevista em quilos da safra em andamento. A visualização gráfica do totalizador das quantidades dos tamanhos da fruta é exposta na figura 14.

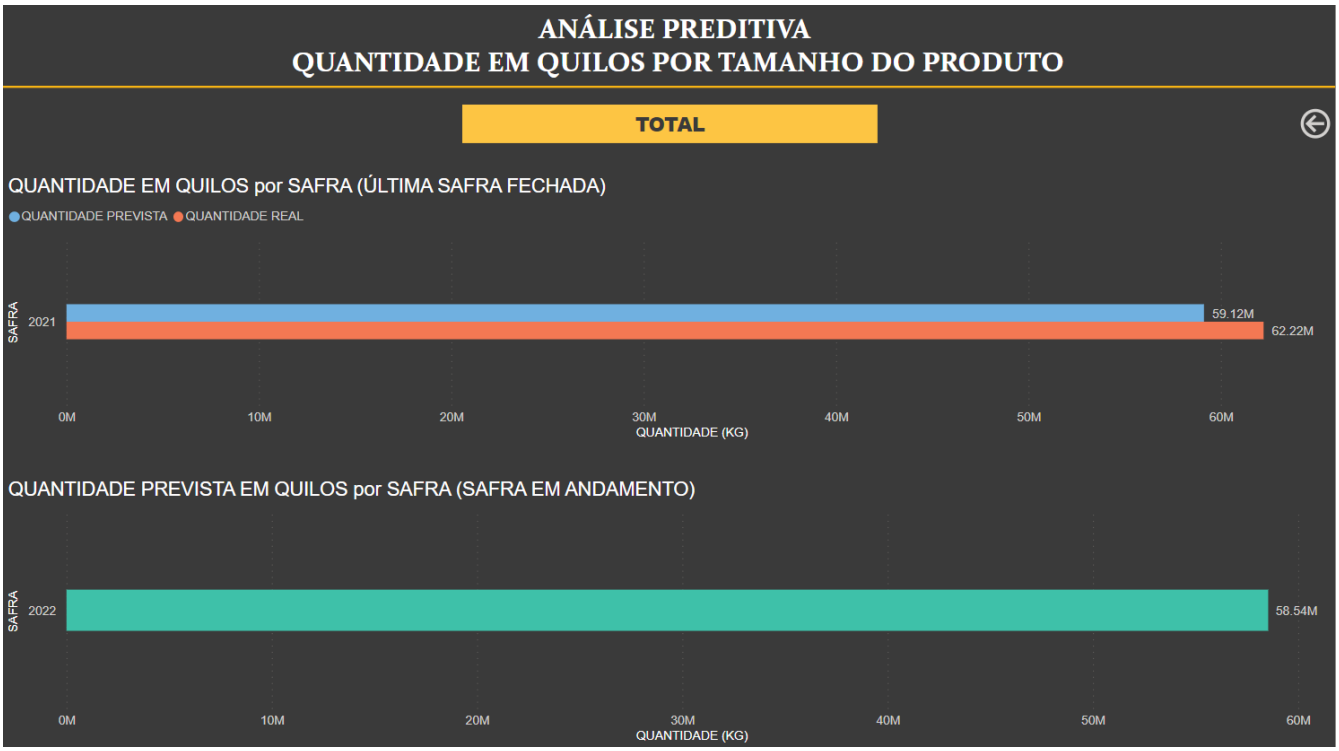

Figura 14: Análise preditiva – Gráfico de comparação entre quantidades previstas e reais em quilos por safra.
Fonte: Próprio autor.

IV. Conclusões

Mediante o estudo realizado pode-se constatar que as variáveis meteorológicas têm um impacto expressivo no processo de produção de maçãs. As soluções propostas buscam usufruir do potencial de tecnologias disponíveis no mercado para que juntas sejam um complemento eficaz para o aprimoramento da relação entre a meteorologia e a agricultura, tema que está em ênfase crescente.

Como atividades futuras do estudo, em alternativa ao algoritmo de regressão linear múltipla, serão realizados testes com outros algoritmos de aprendizado de máquina, também serão testadas alternativas de métricas de avaliação e será mensurada a possibilidade de adição de novas *features* para treinamento dos modelos de predição.

Os resultados apresentados mostram-se satisfatórios, possibilitando a análise de informações históricas relevantes para o negócio e também um olhar para o futuro através dos modelos preditivos desenvolvidos que, de acordo com os bons números nas métricas de avaliação, são uma fonte confiável de conhecimento.

V. Bibliografia

[1] Blain, G. C. "Mudanças climáticas e a fruticultura." Revista Brasileira de Fruticultura [online]. v. 33, n. spe1, pp. 7-12, 2011. https://doi.org/10.1590/S0100-29452011000500003

[2] Leite, G. B., Ricce, W., Massignan, A. M. "Efeitos do clima na Safra 2019/2020 da Maçã em Santa Catarina." Agropecuária Catarinense, 34(2), 60-64, 2021. https://doi.org/10.52945/rac.v34i2.1060

[3] Pomi Frutas "Processo de Produção" Disponível em: http://www.pomifrutas.com.br/processo-de-producao/ Acesso em: 15 de abril de 2022.

[4] Agarwal, K. K., Agarwal, A. "Python for CS1, CS2 and beyond." Journal of Computing Sciences in Colleges, 20(4):262–270, 2005.

[5] Sousa, M. V. C. "Data Warehouse e ETL para Dados da Saúde Pública." Instituto de Matemática e Estatística, Universidade de São Paulo, 2020.

[6] Pimentel, J. F., Oliveira, G. P., Silva, M. O., Seufitelli, D. B., Moro, M. M. "Ciência de dados com reprodutibilidade usando jupyter." Sociedade Brasileira de Computação, 2021.

[7] Oliveira, H. S. "CSVValidation: uma ferramenta para validação de arquivos CSV a partir de metadados." Master's thesis, Universidade Federal de Pernambuco, 2015.

[8] Coelho, A. S. "Introdução a análise de dados com python e pandas." Anais Eletrônicos ENUCOMP, pages 862–876, 2017.

[9] Miranda, A. E. B., de Souza, G. F., da Silva, J. L. R., dos Santos, J. P. Q., Cardoso, P. H., de Azevedo, M. D. "Correlação entre a notificação de sífilis, disponibilidade de penicilina e teste rápido: Uma análise a partir do sistema Retratos da Atenção Primária à Saúde." Revista Brasileira de Inovação Tecnológica em Saúde-ISSN: 2236-1103, 10(2), 11-11, 2020.

[10] Guimarães, P. R. B. "Métodos Quantitativos Estatísticos. (1ª ed.)" IESDE BRASIL S.A., 256 p. ISBN: 978-85-387-3028-6, 2012.

[11] Coppini, J. A. "Usando aprendizagem de máquina na criação de modelos para predizer resultados da Liga Nacional de Futsal do Brasil." Universidade Federal da Fronteira Sul, 2019.

[12] Kristiansen, T. "Forecasting Nord Pool day-ahead prices with Python." The Python Papers, 12(1), 2018.

[13] dos Santos, J. A. A. "Previsão do preço do café arábica: uma aplicação de redes neurais" CNN-BLSTM. Research, Society and Development, 11(3), e3511326101-e3511326101, 2022.

[14] Joblib. "joblib.dump". Disponível em: https://joblib.readthedocs.io/en/latest/generated/joblib.dump.html Acesso em: 17 de abril de 2022.

[15] Microsoft. "O que é Power BI?" Disponível em: https://docs.microsoft.com/pt-pt/power-bi/fundamentals/power-bi-overview Acesso em: 17 de abril de 2022.

[16] FAEP (Federação da Agricultura do Estado do Paraná). "Maçã: Classes ou Calibres" Disponível em: http://www.faep.com.br/comissoes/frutas/cartilhas/frutas/maca.htm Acesso em: 19 de abril de 2022.

[17] Hawerroth, F. J., Nachtigall, G. R. "Condições meteorológicas de outono e inverno e suas influências na safra de maçã 2016/17 na região de Vacaria, RS." Comunicado Técnico 190, ISSN 1808-6802, Embrapa, 2016.

[18] Leite, G. B., Petri, J. L.; Couto, M. "Dormência das Fruteiras de Clima Temperado." In: PIO, R. Cultivo de fruteiras de clima temperado em regiões subtropicais e tropicais. Lavras: Ed. UFL, p. 50-73, 2018.