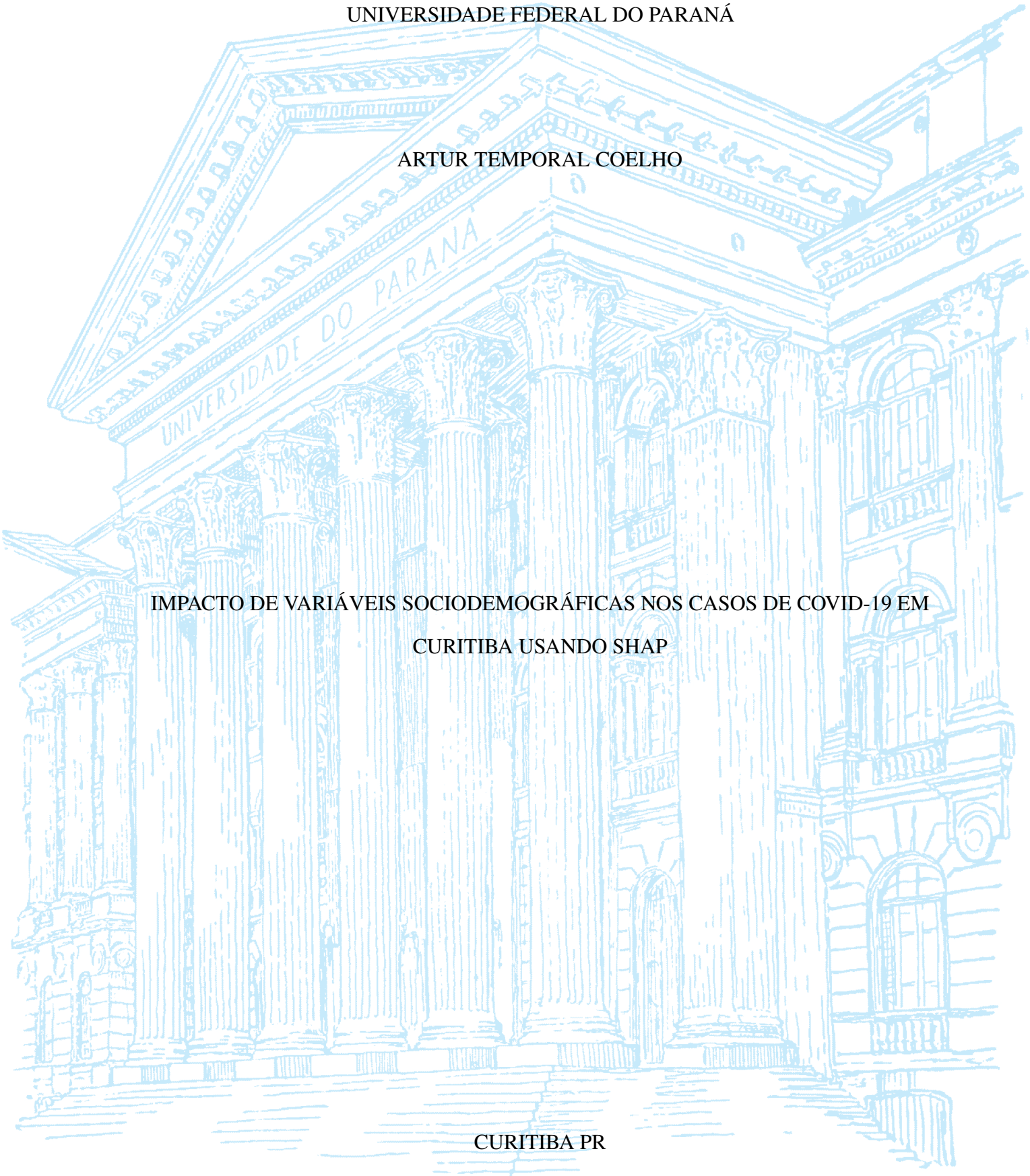
UNIVERSIDADE FEDERAL DO PARANÁ

ARTUR TEMPORAL COELHO



IMPACTO DE VARIÁVEIS SOCIODEMOGRÁFICAS NOS CASOS DE COVID-19 EM CURITIBA USANDO SHAP

CURITIBA PR

2024

ARTUR TEMPORAL COELHO

IMPACTO DE VARIÁVEIS SOCIODEMOGRÁFICAS NOS CASOS DE COVID-19 EM CURITIBA USANDO SHAP

Trabalho apresentado como requisito parcial à conclusão do curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: Eduardo Todt.

Coorientador: Marcos Sfair Sunye.

CURITIBA PR

2024

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho:

Impacto de Variáveis Sociodemográficas nos casos de COVID-19 em Curitiba Usando SHAP

Autor(es): ARTUR TEMPORAL COELHO

GRR 20190471 Nome:_____

GRR_____ Nome:_____

GRR_____ Nome:_____

Apresentação: Data: 13/12/2024 Hora: 10:30 Local: Dinf sala 102

Orientador: Eduardo Todt _____

Membro 1: Carlos Eduardo Ribeiro _____

Membro 2: Cleiton Almeida dos Santos _____

(nome) (assinatura)

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)				
Referência Bibliográfica	(00-10)				
Formato	(00-05)				
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)				
Desenvolvimento do Assunto	(00-05)				
Técnica de Apresentação	(00-03)				
Uso do Tempo	(00-02)				
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)		**********	**********	
NOTA FINAL		**********	**********	**********	90

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramitar o processo para CCOMP (Coordenação Ciência da Computação).

A minha parceira, que me apoiou imensamente por toda essa trajetória. Aos meus amigos, que fizeram essa jornada muito mais prazerosa. E aos meus pais, que me trouxeram até aqui.

AGRADECIMENTOS

Este trabalho não poderia ser realizado sem o apoio de diversos amigos, colegas e professores. Agradeço especialmente ao meu professor orientador, que fez parte deste percurso desde o início. Agradeço também aos meus colegas, que fizeram parte deste caminho. Sem eles, este trajeto não seria possível.

RESUMO

A pandemia de COVID-19 impactou profundamente a sociedade, expondo desigualdades sociais e desafiando as estruturas de saúde pública. Este estudo analisa como variáveis sociodemográficas influenciam a disseminação da doença nos bairros de Curitiba, explorando a relação entre fatores socioeconômicos e a taxa de casos confirmados. Para isso, foram utilizados dados do Censo Brasileiro de 2010 e da Secretaria Municipal de Saúde de Curitiba, consolidados em um dataset dos 75 bairros e 7 variáveis. Com o objetivo de modelar a taxa de atendimentos, um modelo de machine learning foi desenvolvido. O modelo selecionado, Extra Trees, apresentou desempenho robusto quando comparado com modelos de referência, sendo explicado com a ferramenta SHAP para a extração dos valores de *features*. Este método permitiu identificar as contribuições de variáveis e evidenciar como desigualdades sociais se refletem nos impactos da pandemia entre os bairros, proporcionando mais informações para subsidiar políticas públicas mais justas e eficazes.

Palavras-chave: SHAP. Explicação de Features. COVID-19.

ABSTRACT

The COVID-19 pandemic profoundly impacted our society, exposing social inequalities and challenging public health systems. This study examines how sociodemographic variables influence the spread of the disease across neighborhoods in Curitiba, exploring the relationship between socioeconomic factors and confirmed case rates. To this end, data from the 2010 Brazilian Census and the Curitiba Municipal Health Secretariat were consolidated into a dataset comprising 75 neighborhoods and seven variables. A machine learning model was developed to estimate the case rates, with the Extra Trees model demonstrating robust performance compared to reference models. The model's predictions were explained using the SHAP tool to extract feature importance values. This method identified the contributions of various variables and highlighted how social inequalities are reflected in the pandemic's impacts across neighborhoods. The findings provide valuable insights to support the development of more equitable and effective public policies.

Keywords: SHAP. Feature Explaining. COVID-19.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

SHAP	*Shapley Additive Explanations*
COVID-19	Doença por Coronavírus 2019
IBGE	Instituto Brasileiro de Pesquisa e Estatística
UPA	Unidade de Pronto Atendimento
SOTA	*State-Of-The-Art*
MAPE	Erro Percentual Absoluto Médio
MAE	Erro Médio Absoluto
RMSE	Raiz do Erro Quadrático Médio

SUMÁRIO

1 INTRODUÇÃO

Durante o evento da pandemia de COVID-19, a Secretaria Municipal de Saúde de Curitiba promoveu uma reorganização do sistema de saúde público da cidade, transformando algumas UPAs(Unidades de Pronto Atendimento) em um sistema híbrido, funcionando como centros de internamento. Enquanto isso, as unidades de saúde passaram a assumir o papel anteriormente desempenhado pelas UPAs, suspendendo a realização de exames de rotina.

Com essas mudanças, porém, unidades como as maternidades Victor Ferreira do Amaral e Centro Comunitário Bairro Novo, consideradas referência no atendimento de gestantes de risco habitual e parto humanizado, deixaram de realizar seus atendimentos normais para atender casos da COVID-19. Isso resultou, já no segundo ano da pandemia, em um aumento na razão de mortalidade materna. (A. C. R. Trovão, 2021)

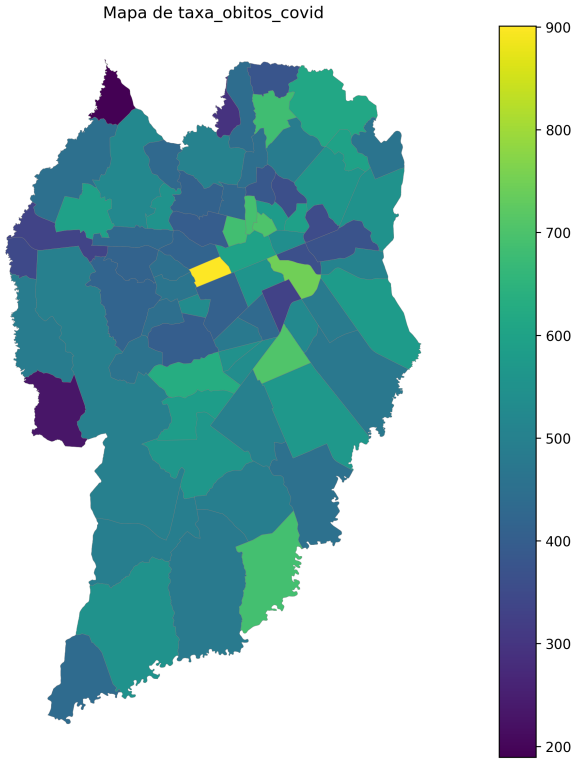


Figura 1.1: Mapa do número de mortes por pessoa registradas como decorrentes de COVID-19 nos bairros de Curitiba

A previsão da necessidade de atendimentos médicos é desafiadora devido à complexidade e variabilidade dos fatores que a influenciam, como desigualdades regionais, condições econômicas, acesso à saúde e políticas públicas. Nesse contexto, modelos de machine learning são ferramentas cada vez mais comuns na tomada de decisões públicas, embora eficazes em capturar padrões nos dados, muitas vezes são tratados como "caixas-pretas", dificultando a compreensão de quais variáveis têm maior impacto nas previsões. Por isso, há a necessidade de explicações por meio de métodos interpretáveis, que permitam identificar e quantificar a importância de cada variável, tornando o modelo mais transparente e útil para informar decisões.

Estes modelos são ferramentas valiosas para a sociedade, pois ajudam a identificar padrões e fatores que influenciam a disseminação da doença. Além disso, permitem projeções

mais precisas, que podem orientar políticas públicas, alocar recursos de saúde de forma eficiente e antecipar cenários críticos. Levando em conta a importância de variáveis sociodemográficas, como densidade populacional e condições econômicas, torna-se possível desenvolver estratégias específicas para mitigar o impacto de pandemias, promovendo ações mais justas e eficazes para diferentes comunidades.

Diante disso, o objetivo do trabalho é propor uma explicação para a influência dos fatores sociodemográficos na taxa de atendimentos de COVID-19 nos bairros de Curitiba, utilizando um modelo de regressão machine learning, assim como dados sociodemográficos e de saúde disponíveis para a cidade, e realizar sua explicação por meio da ferramenta SHAP.

Este trabalho está organizado em 6 capítulos, o primeiro, de introdução, apresenta uma visão geral sobre o problema. O segundo capítulo dá o contexto das tecnologias utilizadas. O terceiro traz trabalhos relacionados ao tema trabalhado. O quarto capítulo apresenta a proposta de solução do problema apresentado, os dados e os métodos utilizados para sua elaboração. O quinto capítulo trata da aplicação dos resultados obtidos. O sexto e último capítulo conclui o trabalho com a discussão desses resultados.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo vamos apresentar uma visão sobre os fundamentos e ferramentas utilizados na análise e explicação dos dados, desde a base matemática dos valores de Shapley, com um exemplo de jogo cooperativo justo, até uma breve explicação sobre o funcionamento do *framework* SHAP, destacando suas características internas que permitem explicações eficientes e aproximações dos valores SHAP.

A teoria dos jogos cooperativos é uma área da teoria dos jogos que estuda como grupos de jogadores colaboram para alcançar resultados vantajosos para todos os participantes. Os jogos cooperativos envolvem a formação de coalizões, onde os jogadores negociam e compartilham os benefícios de uma colaboração. O objetivo principal é entender como distribuir de forma justa os ganhos obtidos por essas coalizões, levando em consideração a contribuição individual de cada membro.

2.1 VALORES SHAPLEY

Em teoria dos jogos, Valor de Shapley (Shapley, 1951) é um **conceito de solução** que atribui um valor de recompensa a cada jogador de forma proporcional à sua contribuição marginal esperada em todas as possíveis coalizões de um jogo cooperativo.

Formalmente, dado um jogo de coalizão: começamos com um conjunto N (de n jogadores) e uma função $v : 2^N \to \mathbb{R}$, que associa valores a subconjuntos de jogadores, chamada de "função característica".

A função v é definida por: se S é uma coalizão de jogadores, então $v(S)$, chamado de "valor da coalizão S", descreve a soma total das recompensas aos membros de S que podem ser obtidas por meio dessa cooperação.

O valor de Shapley é uma forma de distribuir os ganhos totais entre os jogadores, supondo que todos colaborem formando uma coalizão. De acordo com o valor de Shapley, a quantidade que o jogador i recebe em um jogo de coalizão (v, N) é dada por:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left(v(S \cup \{i\}) - v(S)\right)$$

Onde n é o número total de jogadores e a soma se estende sobre $S \subseteq N \setminus \{i\}$. A fórmula pode ser interpretada da seguinte maneira: uma coalizão é formada um jogador por vez, com cada jogador obtendo sua "contribuição marginal"$v(S \cup \{i\}) - v(S)$ como compensação justa. Em seguida, para cada jogador que entra na coalizão, calcula-se a média dessa contribuição ao longo de todas as possíveis permutações em que a coalizão pode ser formada.

É considerado uma distribuição "justa"pois satisfaz os seguintes axiomas:

1. Eficiência: O valor total de um jogo deve ser igual a soma da contribuição de cada jogador.
2. Simetria: Dois jogadores são intercambiáveis se contribuem igualmente para a coalizão.
3. Propriedade do jogador nulo: Se um jogador não contribui para a coalizão, não recebe nenhum valor.
4. Aditividade: Combinando dois jogos, a contribuição de um jogador nele é a soma da contribuição do jogador para cada um dos jogos individuais.

Em suma, o valor de Shapley para cada jogador é dado pelo somatório de: O valor da função característica de sua coalizão, subtraído do valor da função para a coalizão com a remoção do jogador, multiplicado pelo peso, que é dado pelo número de coalizões nas quais o jogador aparece, dividido pelo número total de coalizões.

2.1.1 Valores de Shapley para 2 jogadores

Como exemplo, vamos imaginar um jogo fictício representando uma maratona de programação, e uma equipe (coalizão) formada por 2 jogadores e uma premiação recebida de R$100. Contudo, os jogadores, não contentes com a divisão proposta de 50% para cada, buscam uma alternativa, com a informação de que, participando da mesma competição individualmente, o jogador 1 receberia R$75, e o jogador 2 receberia R$50 de premiação. Eles decidem usar os valores de Shapley para realizar a distribuição.

Coalizão	Prêmio
C_{12}	100
C_1	75
C_2	50
C_0	0

Tabela 2.1: Tabela de coalizões e seus respectivos prêmios.

Assim, para obter o valor de Shapley para C_1:
Dado por

$$\phi_1(v) = \sum_{\substack{S \subseteq N \\ i \in S}} \frac{1}{n} \cdot v(S) - v(S \setminus \{1\})$$

Ou seja, para cada subconjunto em que 1 pertence, calcula-se a diferença do valor da função característica do subconjunto, para o valor da função característica no mesmo subconjunto retirando o elemento 1.

$$C_{12} - C_2 = 100 - 50 = 50$$

$$C_1 - C_0 = 75 - 0 = 75$$

$$\phi_1(v) = \frac{1}{2} \times (50 + 75) = 62.5$$

Obtemos então o valor de Shapley para o jogador 1 de 62.5

E o valor de Shapley para C_2:

$$C_{12} - C_1 = 100 - 75 = 25$$

$$C_2 - C_0 = 50 - 0 = 50$$

$$\phi_2(v) = \frac{1}{2} \times (25 + 50) = 37.5$$

Os jogadores então, com seus valores de Shapley calculados, os jogadores 1 e 2 recebem R$62.5 e R$37.5, respectivamente. Esta distribuição justa leva em consideração a contribuição esperada para cada jogador.

2.2 **SH**APLEY **A**DDITIVE EX**P**LANATIONS

A habilidade de interpretar corretamente a previsão de um modelo é de extrema importância, pois proporciona confiança ao usuário e serve como ferramenta para a melhoria do modelo, apoiando o entendimento do processo. Diante do crescente conflito entre acurácia e explicabilidade em modelos de machine learning, diversos métodos explicativos foram criados. Este trabalho utiliza o *framework* SHAP para explicar a importância das *features*.

O *framework* SHAP (Scott et al., 2017) é um método unificado de explicabilidade para modelos de machine learning que utiliza a teoria dos valores de Shapley para medir a contribuição de cada variável nas predições de modelos de forma agnóstica ao modelo, podendo ser aplicada a qualquer tipo de modelo. Os valores SHAP indicam a importância de cada variável, além de sua influência positiva ou negativa.

Os valores SHAP diferem dos valores de Shapley em não garantir o cumprimento dos axiomas que definem os valores de Shapley de acordo com Shapley (1951). O pacote unifica os diversos métodos de aproximação e explicação dos valores por meio de aproximações, inferências e métodos exatos. Neste trabalho, as características reais são denominadas "variáveis", enquanto "*features*"refere-se às suas representações numéricas ou categóricas nos modelos.

O algoritmo genérico (Scott et al., 2017), capaz de aproximar valores SHAP para qualquer modelo, consiste no cálculo da perturbação de *features*, executando o modelo para um subconjunto das possíveis permutações das *features*, assim como a execução para médias. Estes valores são então estimados por meio de regressão linear. São calculados os pesos baseados na probabilidade de aparição da *feature* e, por fim, uma média ponderada calcula o ganho médio esperado para cada *feature* no modelo.

Outro algoritmo implementado no *framework* é o TreeSHAP (Lundberg et al., 2018a), que alavanca a estrutura de cada árvore no modelo de ensamble, podendo calcular os valores

para modelos baseados em árvores em complexidade polinomial (Lundberg et al., 2020) [1]. Este algoritmo difere dos restantes em realizar os cálculos dos valores de acordo com as regras de inferência causal dispostas em Janzing et al. (2020).

O algoritmo consiste em, dado uma árvore do modelo:

- Calculam-se os pesos memoizados para cada nó, dado por 1 / coeficiente binomial do nó
- Para cada amostra, calcula-se a inferência no modelo, com cada *feature* faltante, adiciona o resultado em uma lista de resultados
- Finalmente, calcula-se a média ponderada para cada valor SHAP, com os pesos calculados previamente.
- Obtém-se por fim o ganho médio esperado, exato do modelo

[1] Existem implementações em complexidade linear (Bifet et al., 2022)

3 TRABALHOS RELACIONADOS

Este capítulo busca apresentar uma análise crítica de estudos semelhantes ao presente projeto, evidenciando as lacunas existentes e como a abordagem proposta contribui para avançar no entendimento da relação entre fatores socioeconômicos e a previsão de casos de COVID-19.

A pandemia de COVID-19 evidenciou disparidades sociais significativas, sendo mais prevalente entre minorias raciais e étnicas, bem como em populações de classes socioeconômicas mais baixas (Levy et al., 2022). Enquanto estes fatores são amplamente estudados a níveis municipais, estaduais e federais, os bairros têm grande influência na desigualdade de saúde dentro de cada cidade (Kawachi, 2003).

Khanijahani et al. (2021) Apresenta uma revisão de trabalhos sobre disparidades étnicas e socioeconômicas na pandemia, concluindo que mostram que minorias étnicas e pessoas que fazem parte de regiões desfavorecidas estão mais vulneráveis à pandemia.

Não encontramos trabalhos que discutam diretamente a explicação dos fatores socioeconômicos entre bairros, na pandemia de COVID-19, enquanto os trabalhos citados aqui discutem esses fatores nesse âmbito em parte, realizam estes estudos por meio de modelos que não incluem os fatores, analisando-os posteriormente de maneira anedótica, ou deixam de realizar explicações profundas sobre as *features*.

Ao fazer a previsão de casos diários de COVID-19 na Indonésia, Hasanah et al. (2023) utilizou 17 parâmetros de dados socioeconômicos e demográficos entre três centros populacionais, com diferentes distribuições populacionais. No qual foram realizadas previsões para cada dia utilizando o modelo Prophet (Developers, 2017).

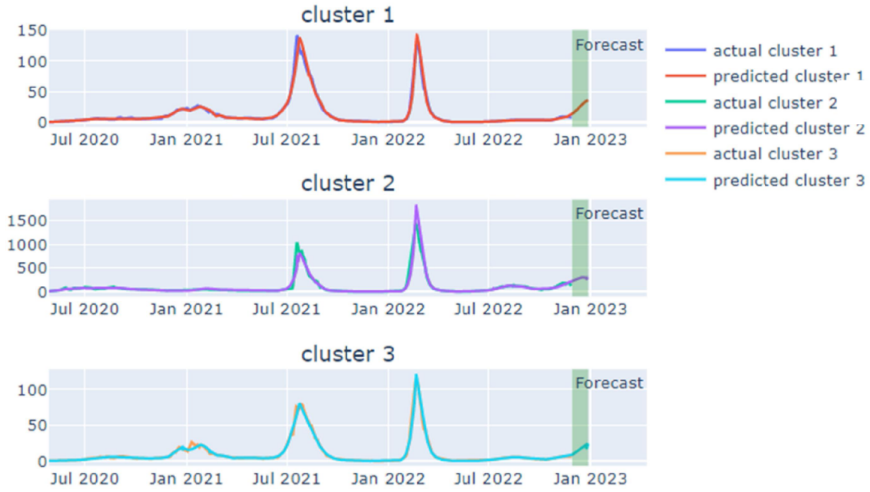


Figura 3.1: Número de casos por tempo, na legenda, os valores reais e previstos para cada *cluster*. Fonte: (Hasanah et al., 2023)

Separando os casos registrados em três regiões do país, foram realizadas previsões com séries temporais utilizando a média dos últimos sete dias, obtendo resultados próximos dos valores reais.

A análise final conclui que houve mais casos entre os centros com maior renda per capita, mostrando mais casos da doença, enquanto centros rurais, com população por área mais baixa, apresentaram menos casos. Porém, essas conclusões foram dadas por meio de uma análise

geral dos dados, não como um resultado do modelo, sendo inferidas pela diferença nos *clusters*, do número de casos para as características de cada.

Enquanto Levy et al. (2022) apresentou uma abordagem mais completa, realizando uma predição baseada na desigualdade socioeconômica nos bairros de São Francisco e em Wisconsin, utilizando a mobilidade urbana como parâmetro principal, buscando encontrar possíveis caminhos de transmissão da patologia.

O estudo apresenta um modelo inovador que compila um índice de disparidade chamado 'desigualdade baseada em mobilidade', encontrando uma melhoria significativa ao introduzi-la no modelo de previsão de casos. Enquanto o uso da mesma apenas não explica a qualidade dos modelos de Poisson utilizados. Em conclusão, apontam que houve um aumento de até 18,5% em casos de COVID-19 em bairros com população majoritariamente negra. E um aumento de risco de aproximadamente 60% em regiões de São Francisco pode ser explicado pelo modelo de disparidade social utilizado.

Enquanto o estudo de Morales et al. (2017) buscou compreender os fatores de risco a nível global para a pandemia de Influenza A(H1N1) de 2009. Por meio de modelos de regressão e simulações, o estudo buscou encontrar os principais fatores explicativos das disparidades nos números de óbitos da pandemia em questão.

Encontrando, em conclusão, que dentre os principais fatores, idade é mais proeminente, em conjunto com outras comorbidades de saúde pré-existentes. Esta conclusão, porém, se dá apenas pelas métricas retiradas dos modelos, sem uma explicação rigorosa das *features*.

Em Debjit et al. (2022), aplica modelos SOTA e calibragem de hiperparâmetros para classificar a presença de COVID-19 dado um conjunto de *features* representativas de sintomas, apresentando bons resultados na classificação e, posteriormente, explicando a importância das *features* por valores SHAP. O que garante uma maior explicação do funcionamento do modelo e, por consequência, um maior entendimento do problema.

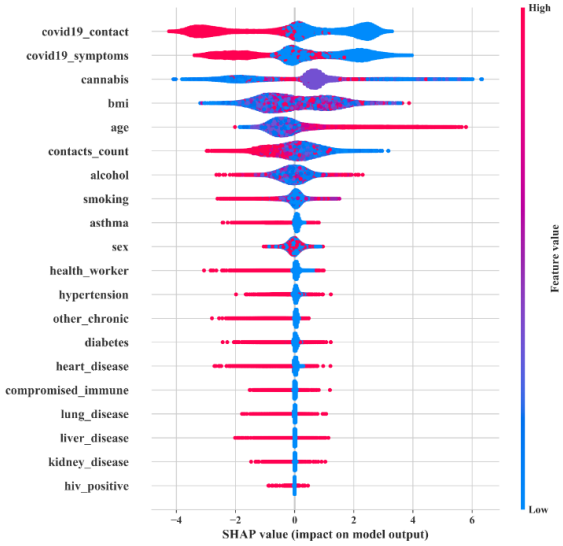


Figura 3.2: Valores SHAP para as *features* do dataset clinico utilizado para o trabalho. Fonte: (Debjit et al., 2022)

Com estas explicações, o trabalho espera auxiliar os profissionais de saúde na identificação da doença de maneira mais ágil, evidenciando a possibilidade do uso de ferramentas de machine learning explicativas em espaços de saúde com maior eficiência e confiabilidade.

Dentre os trabalhos apresentados, percebemos uma ausência de trabalhos que explicam a disparidade de casos a nível de granularidade de bairros. Sendo apresentados modelos de previsão gerais de casos, e trabalhos que realizam visões gerais sobre as disparidades.

4 PROPOSTA

Este capítulo descreve os procedimentos para a execução do trabalho, desde a organização do dataset, o treinamento do modelo de machine learning e a extração das importâncias das variáveis.

4.1 DATASET

O dataset utilizado foi construído a partir de dados do Censo demográfico do Brasil de 2010, somado a dados da Secretaria Municipal de Saúde de Curitiba, escolhendo as variáveis relevantes em diversas categorias. Nos quesitos socioeconômicos: população, renda e cor, também relevantes à saúde, como cobertura verde, que é a porcentagem da área do bairro disposta de vegetação. Os dados oriundos do censo do IBGE estão agregados por bairro, enquanto para os dados da Secretaria Municipal de Saúde de Curitiba realizamos a agregação na mesma granularidade.

Ao analisar o dataset, percebeu-se que, devido à grande disparidade de populações entre os bairros, há uma grande incidência de correlações triviais.

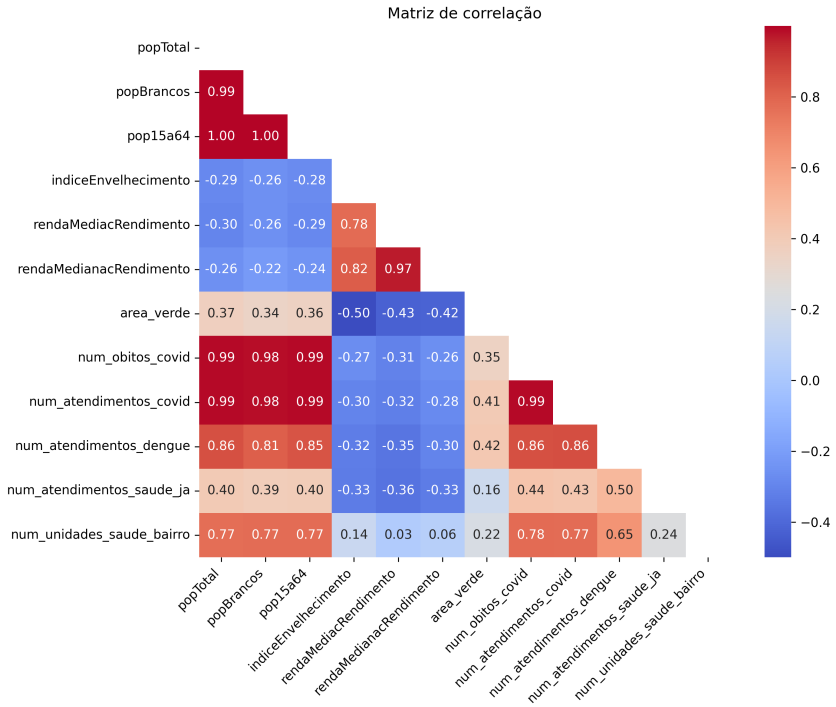


Figura 4.1: Matriz de correlação de dados originais do dataset, pode-se observar que ha correlações triviais em todos os dados relacionados a população, devido a natureza da separação dos dados em bairros

Portanto precisamos realizar a padronização dos dados por taxas. Os dados foram normalizados por população ou área (porcentagem de cobertura verde). Desta maneira, os dados são invariáveis à área ou população total de um bairro.

Com os dados normalizados, podemos observar correlações mais plausíveis, sem nenhuma correlação trivial. Desta maneira, o modelo construído terá uma base mais sólida e poderá mostrar resultados mais consistentes entre bairros, independente da população ou área.

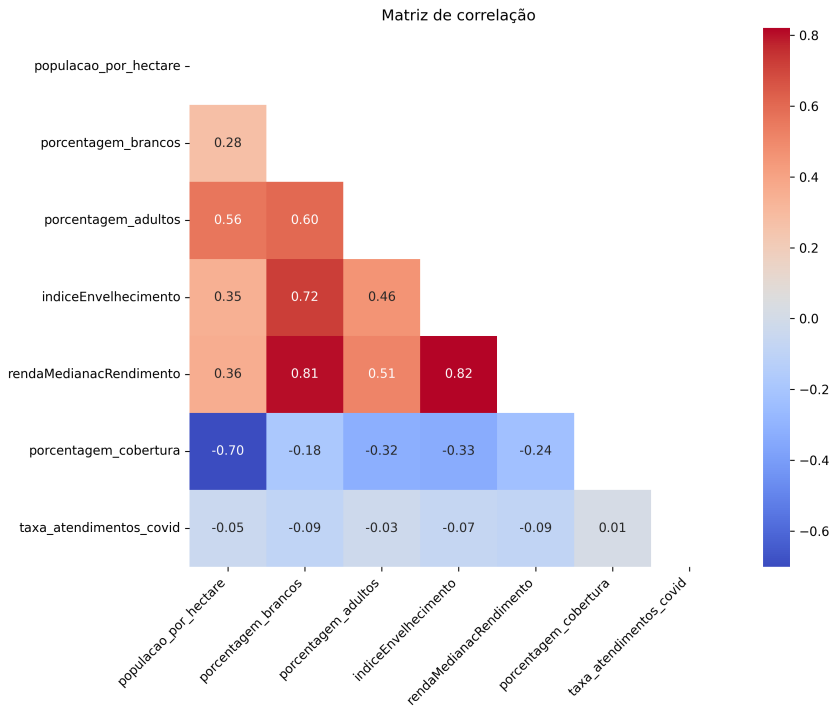


Figura 4.2: Matriz de correlação das variáveis escolhidas

Os mapas com os dados apresentados geograficamente estão disponíveis no Apêndice A.

Foram escolhidas 6 variáveis como *input*, para prever a 7ª variável como *output*. As variáveis levadas em consideração no modelo foram selecionadas devido à sua relevância e independência linear entre si, são as seguintes:

	Nome	**Descrição**
1	populacao_por_hectare	Descreve a densidade populacional.
2	porcentagem_brancos	Porcentagem de pessoas autodeclaradas brancas.
3	porcentagem_adultos	Porcentagem de indivíduos entre 15 e 64 anos.
4	indiceEnvelhecimento	Numero de idosos para cada 100 crianças de 0 a 14 anos.
5	rendaMedianacRendimento	Mediana da renda familiar por indivíduo.
6	porcentagem_cobertura	Porcentagem de cobertura verde na área total.
7	**taxa_atendimentos_covid**	Número de casos confirmados de COVID-19 para cada 100 mil habitantes.

Tabela 4.1: Tabela com variáveis e suas descrições

Sendo que iremos prever especificamente a **Taxa de atendimentos por COVID-19**

4.2 MODELO MACHINE LEARNING

Com os dados obtidos, vamos produzir um modelo de machine learning que seja capaz de inferir os dados desejados, e logo, explicá-los em relação aos valores SHAP das *features*. Para isso, temos que levar em consideração o perfil dos dados e do problema.

4.2.1 Limitações

Dado o dataset compilado, analisado e normalizado, com 75 linhas e 7 colunas, representando os bairros de Curitiba, e as variáveis escolhidas. Reconhecemos que o tamanho do dataset é considerado pequeno para o estudo, assim como a diferença de tempo entre os dados do censo e os dados da pandemia. Esses fatores podem ser limitantes para a performance de qualquer modelo criado, mesmo com as mitigações feitas para garantir a qualidade dos dados.

Precisamos, portanto, de um modelo que seja robusto a *outliers*, que são os valores numericamente distantes dos restantes, e que também seja resistente a *overfitting*, que é a adaptação excessiva do modelo aos dados, e que possa trabalhar com um tamanho reduzido de dados.

4.2.2 Seleção do modelo

Iniciamos por testar a execução de modelos mais comuns encontrados em trabalhos referência na área de explicação de variáveis (Liu et al., 2022) (Lundberg et al., 2018b) (Parsa et al., 2020). Vamos testar os principais modelos de regressão SOTA: Random Forest, Light Gradient Boosting Machine (LGBM) e Extreme Gradient Boosting (XGBoost). Também testaremos outros modelos disponíveis no pacote Scikit-learn (Pedregosa et al., 2011), com diversos modelos comuns disponíveis.

Modelos baseados em árvore, característica de todos os citados acima, possuem ainda a vantagem de serem apoiados pela própria ferramenta com o Tree SHAP (Lundberg et al., 2020). Desta maneira, a ferramenta é capaz de realizar a extração dos valores diretamente, em complexidade polinomial, sem a necessidade de métodos de aproximação.

Com a realização de testes empíricos, apresentados no próximo capítulo, percebemos que os modelos SOTA, apesar de obterem bons resultados em suas respectivas aplicações, foram mais sensíveis ao pequeno dataset do problema e obtiveram um ajuste inferior a modelos mais simples, como Random Forest; porém, o modelo mais vantajoso possui ainda mais características desejadas.

O modelo Extra Trees (Geurts et al., 2006) obteve os melhores resultados nas métricas medidas na aplicação do problema do trabalho e dataset usado. O modelo de ensemble é baseado em Random Forest, com a diferença de realizar os cortes dos nós em pontos aleatórios, gerando árvores completamente aleatórias. Além de ser um bom modelo para realizar a inferência dos dados, possui a característica desejável de facilitar a explicabilidade dos dados, fornecendo uma visão mais real da importância das *features* por meio dos valores SHAP (Genuer et al., 2010) (Lundberg et al., 2020).

4.2.3 Melhoria de parâmetros

Mesmo com modelos baseados em árvores obtendo bons resultados no geral com os parâmetros padrões, é interessante realizar a busca dos melhores parâmetros para o problema (Probst et al., 2019). Para melhorar a eficiência do modelo final, realizamos um processo de *grid search*, para buscar os parâmetros ideais para o modelo. O processo consiste em testar diversos parâmetros do modelo de maneira independente e abrangente, para obter as melhores métricas possíveis.

Além de, no nosso caso, encontrarmos melhores resultados em métricas de erro percentual, o número de árvores nos auxilia no processo posterior de extração de importâncias (Genuer et al., 2010), fornecendo resultados mais estáveis e reprodutíveis.

Parâmetro	Descrição	Val. Novo	Val. Original
`max_depth`	Profundidade máxima da árvore	Infinito	Infinito
`min_samples_leaf`	Número mínimo de amostras para ser considerado folha.	10	1
`min_samples_split`	Número mínimo de amostras para dividir um nó.	2	2
`n_estimators`	Número de árvores na floresta.	500	100
`random_state`	Controle de aleatoriedade, para a reprodutibilidade. (Sendo i o número do teste)	`i*7`	Aleatório

Tabela 4.2: Configuração de hiperparâmetros com Descrições e Valores Padrão

4.2.4 Treinamento

Devido às limitações citadas, iremos realizar, de maneira independente, 15 processos completos de treinamento e validação. Este processo utiliza um valor calculado do *random_state*, uma ferramenta que garante, tanto na separação dos dados entre treino e teste, quanto no treinamento, um nível de aleatoriedade e reprodutibilidade.

O processo é realizado em etapas, considerando o dataset já preparado:

1. Separação de dados em treino e teste, com uma distribuição de 80/20, respectivamente.
2. Ajuste do modelo, utilizando os parâmetros ótimos obtidos.
3. Inferência dos dados, utilizando os dados de teste.
4. Validação dos dados de teste, de acordo com as métricas escolhidas.
5. Armazenamento dos modelos e das métricas para análise futura.

Ao fim deste processo, obtemos os modelos ajustados de acordo com o dataset, assim como as métricas de cada um deles, realizando a validação com os dados de teste separados previamente. Assim, poderemos analisar sua eficácia na representação da variável de output em relação à real e, portanto, sua fidelidade ao problema real.

4.3 EXTRAÇÃO DOS VALORES SHAP

Dado os modelos, ajustados para o dataset gerados, vamos realizar a extração das importâncias das variáveis do modelo por meio da ferramenta SHAP. Dado o elevado número de árvores, temos uma maior estabilidade na importância das variáveis (Genuer et al., 2010). Portanto, para esta análise, selecionamos um modelo com métricas próximas à média de todos os modelos realizados.

O *framework* SHAP contém diversos métodos de explicação, sendo possível realizar explicações aproximadas para diversos tipos de modelos. Em Janzing et al. (2020), são expostos argumentos contrários para a explicação de modelos utilizando este método, por isso, a própria ferramenta inclui o método utilizando a abordagem "interventional", separando as dependências entre as *features* de acordo com as regras de inferência causal ditadas por Janzing et al. (2020). [1]

[1] A mudança no algoritmo Tree SHAP foi realizada após as críticas em Janzing et al. (2020)

Portanto, garantimos a utilização do algoritmo TreeSHAP (Lundberg et al., 2018a) para obtermos as vantagens expostas na seção 2.2. O módulo SHAP para Python (Lundberg, 2018) possui uma interface simples para a extração de valores SHAP, com um *wrapper* para uma implementação em C do algoritmo TreeSHAP (Lundberg et al., 2018a). Com isso, podemos garantir que as explicações obtidas possuem causalidade no modelo, entre as *features* e sua inferência.

5 RESULTADOS

Neste capítulo, são apresentados os resultados do modelo, os valores obtidos para avaliar o modelo de acordo com as métricas selecionadas e os valores SHAP obtidos para o dataset, que definem a importância das variáveis.

5.1 MÉTRICAS

As métricas que iremos calcular para nosso modelo de regressão são: Erro Médio Absoluto (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o Erro Percentual Absoluto Médio (MAPE). Para essas métricas, valores mais baixos indicam melhor desempenho. As métricas foram calculadas com os dados de teste separados anteriormente, de maneira a prover um valor que reflita o ajuste do modelo em dados reais.

5.2 TESTE DOS MODELOS REFERÊNCIA

Com os testes realizados, extraímos as métricas, assim podendo comparar os diversos modelos disponíveis. Notamos que entre os modelos testados, os modelos baseados em árvores se destacaram, com outros modelos comuns obtendo resultados bastante inferiores.

Modelo	**MAPE**
`ExtraTreesRegressor`	**0.154**
`RandomForestRegressor`	0.185
`LGBMRegressor`	0.185
`XGBRegressor`	0.175
`GradientBoostingRegressor`	0.211
`GaussianProcessRegressor`	0.228
`LinearSVR`	0.998
`MLPRegressor`	0.999

Tabela 5.1: Comparaçao dos modelos testados por MAPE

Ao fim, selecionamos o modelo Extra Trees por conta de sua maior eficácia nos testes de validação cruzada, superando os outros modelos de referência e por sua compatibilidade com as ferramentas de extração de explicações utilizadas.

5.3 RESULTADOS DO MODELO

Dado o modelo escolhido, os hiperparâmetros otimizados e o ajuste com o dataset, foram realizados diversos testes de validação cruzada, para maximizar a qualidade de representação do problema real.

Utilizamos principalmente a métrica MAPE, usualmente utilizada no contexto de regressões (Moreno et al., 2013), devido ao seu entendimento intuitivo e independente da magnitude dos dados. Percebemos então um valor de MAPE bom (Meade, 1983). Com o contexto de um dataset pequeno e trabalhos similares, que também trabalham com dados socioeconômicos, atingindo resultados semelhantes a Hasanah et al. (2023). Enquanto trabalhos

Tabela 5.2: Métricas obtidas na execução de 15 treinamentos independentes, para os dados de validação

	Mínimo	Média	Máximo	Desvio Padrão
RMSE	4563.82	7278.48	9580.11	1569.07
MAE	3387.61	5400.51	7410.59	1202.70
MAPE	0.0881	**0.1546**	0.2335	0.0403

que trabalham com séries temporais em contextos pandêmicos podem gerar resultados ainda melhores, como Soares et al. (2020).

Com a inferência dos dados de treino + teste, o MAPE é de 0.147. Esta baixa diferença nos indica que o modelo está bem generalizado e ajustado, ou seja, sem grande *overfitting* ou viés significativo, o que nos ajuda a garantir que nossas explicações são próximas da realidade dos dados.

5.4 ANÁLISE DOS ERROS

Podemos analisar os erros do nosso modelo de maneira explícita por meio de gráficos, para obter uma compreensão melhor da distribuição. Assim, podemos detectar qualquer tipo de desvio ou tendência, o que poderia indicar que o modelo obtido possui algum tipo de *overfitting*, necessitando de mais ajustes para representar melhor a realidade dos dados.

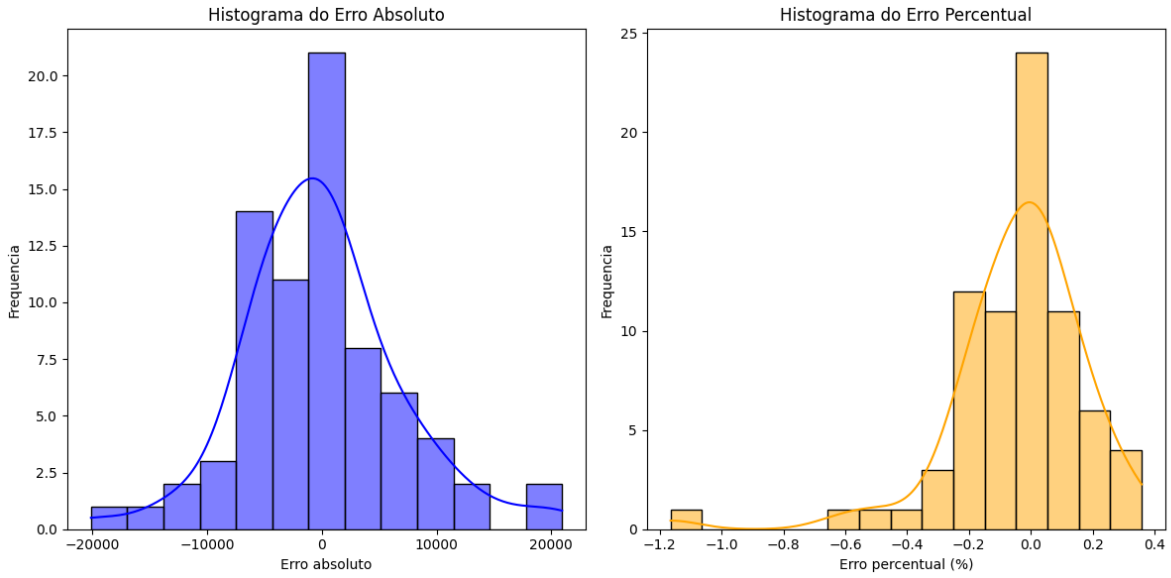


Figura 5.1: Gráfico da distribuição dos erros absolutos e percentuais do modelo em relação aos dados reais

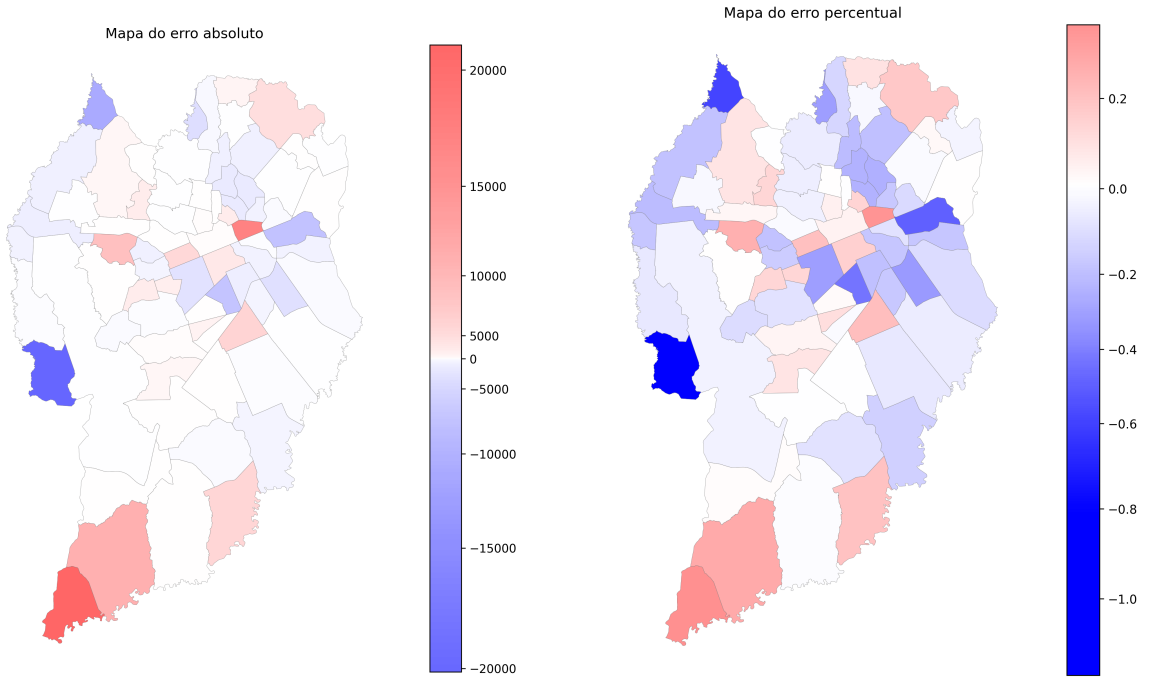
Podemos perceber que a distribuição absoluta se encontra uniformemente distribuída cerca do zero, mais uma confirmação do bom ajuste do modelo, contendo poucos *outliers*, mesmo que significativos. O gráfico de erro percentual mostra uma tendência de subestimação dos dados, o que deve ser levado em consideração, apesar de ser um fator comum em modelagens.

Uma parte significativa do erro pode ser atribuída a alguns *outliers* com altos erros, que coincidentemente são também *outliers* na variável objetivo.

Estes erros podem ser explicados quando observamos diretamente os *outliers* nos erros, como o bairro de São Miguel, um dos menos povoados da cidade, que pode ter uma grande evasão de casos, uma hipótese levantada devido à alta disponibilidade de unidades de saúde no

bairro vizinho, o CIC, ou reportagens baixas de casos devido à falta de unidades de saúde no local.

Observando também o bairro Caximba, notamos a maior incidência de casos da cidade, o que poderia ser dado por meio de um acréscimo de casos provenientes da cidade fronteiriça de Fazenda Rio Grande, e ao afastamento do centro da cidade a qual pertence.



(a) Mapa da diferença absoluta entre a inferência do modelo e o valor real.

(b) Diferença percentual entre a inferência do modelo e o valor real.

Figura 5.2: Comparação entre o erro absoluto e percentual das inferências do modelo.

A figura mostra a disposição dos erros geograficamente, podemos notar, na figura 5.2 (a), a distribuição dos erros cerca de zero, com poucos *outliers*, enquanto na figura 5.2 (b), o skew para a direita, representando uma subestimação nos valores.

Quando observamos os erros em comparação com os dados reais, notamos que o modelo é tendencioso para o centro das observações, sendo resistente a *outliers*.

5.5 EXPLICAÇÃO DAS *FEATURES*

Ao final, com o modelo ajustado ao dataset, extraímos as explicações de todas as observações. Os valores SHAP são apresentados por observação, ou seja, para cada observação e para cada *feature*, há um valor correspondente que informa o quanto a *feature* foi responsável pela inferência final.

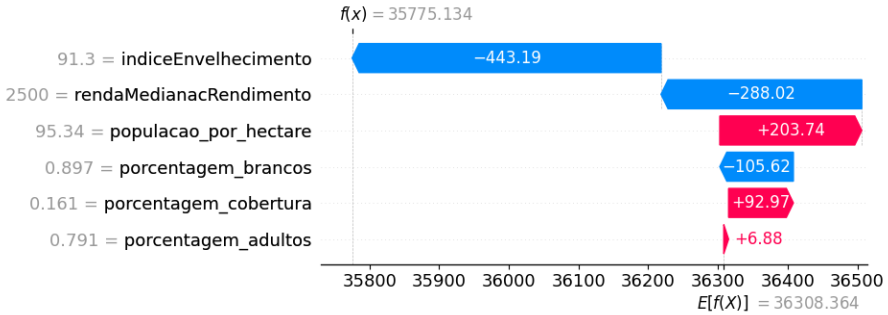


Figura 5.3: A influencia das *features* na observação do bairro Cristo Rei

Analisando a observação em 5.3 temos E[f(x)] = 36308 como o valor médio das previsões realizadas pelo modelo, em casos por 100 mil habitantes para o bairro. Logo f(x) = 35774 como o valor previsto (sendo 35688 o número real). [1]

Notamos, nesta observação, a atuação do modelo em levar em consideração um índice de envelhecimento e renda mais altos, assim como uma alta densidade populacional e população branca. E a direção na qual cada fator afeta a previsão como um todo.

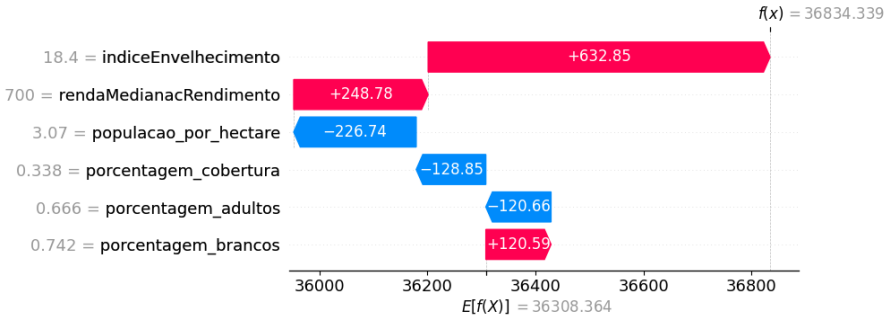


Figura 5.4: A influencia das *features* na observação do bairro Caximba

Observamos que na figura 5.4 o índice de envelhecimento e renda mais baixos tem um comportamento contrário, aumentando a previsão, enquanto uma densidade populacional mais baixa leva a previsão para valores menores.

[1] Devemos levar em conta que a população para cada bairro em média, é de aproximadamente 23 mil habitantes, ou seja, o número de casos por 100 mil habitantes adotado, nos da uma inflação aparente no número de casos.

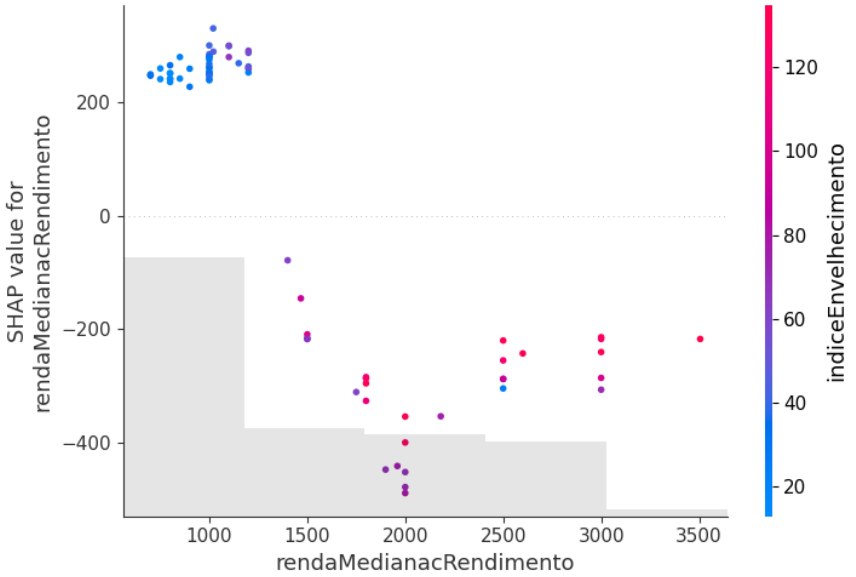


Figura 5.5: Valores SHAP do índice de envelhecimento por índice de envelhecimento

Porém, estas observações escondem um fator importante, que são as interações entre *features*. Em 5.5 observamos o comportamento da renda mediana em conjunto com o índice de envelhecimento. No eixo X, o valor da *feature* na observação; em Y, o valor SHAP para a mesma na inferência; em cinza, a distribuição em frequência da *feature*; e a coloração mostra o valor da *feature* de principal influência encontrado pelo modelo, o índice de envelhecimento.

Podemos observar que os valores baixos para a *feature* causam um valor alto de previsão, acompanhado de um índice de envelhecimento baixo.

Os gráficos de todas as variáveis utilizadas estão apresentados no Apêndice B.

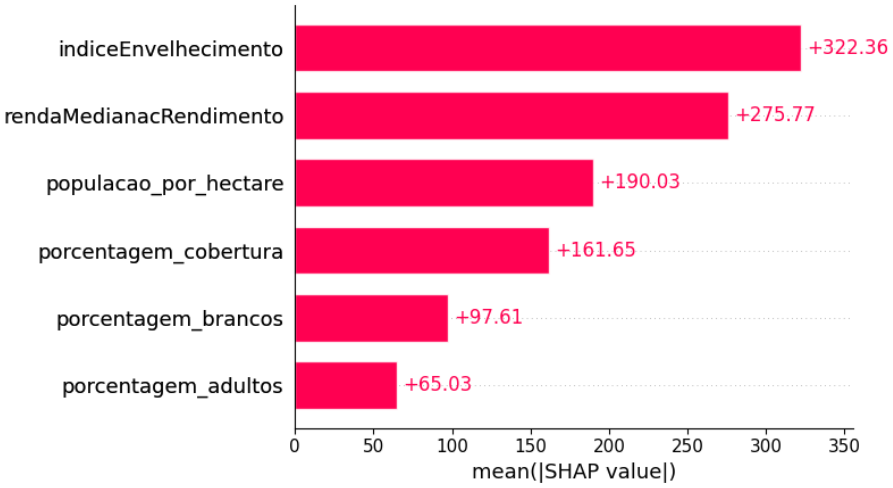


Figura 5.6: Média dos valores SHAP, representando a importância

A figura 5.6 apresenta diretamente a média de cada variável, com a média do valor SHAP das *features* em cada observação.

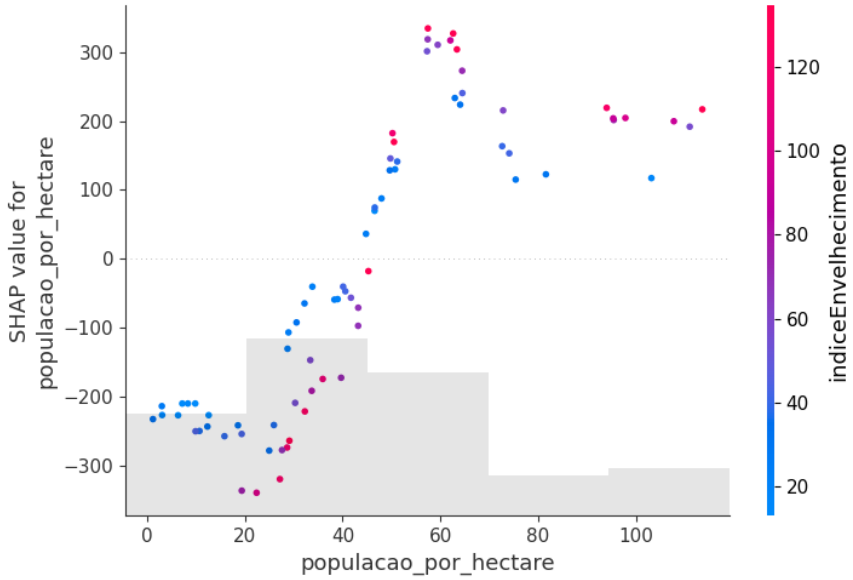


Figura 5.7: Valores SHAP da população por hectare pelo índice de envelhecimento

Para a *feature* de densidade percebemos que os valores SHAP acompanham o valor, com bairros menos povoados recebendo valores mais baixos, enquanto bairros mais povoados recebem um valor de influência mais alto; porém, esta *feature* já não apresenta uma grande separação visível em relação à variável secundária.

Podemos concluir que os valores SHAP seguem uma tendência intuitiva do entendimento do funcionamento da doença, expondo como os fatores sociodemográficos realmente afetam a propagação da mesma em um nível local.

Os dados também são corroborados com estudos sobre o assunto, em Aleta e Moreno (2020) são apresentadas correlações em diversos países nos quais populações jovens apresentam maiores incidências de COVID-19, devido também a fatores socioeconômicos.

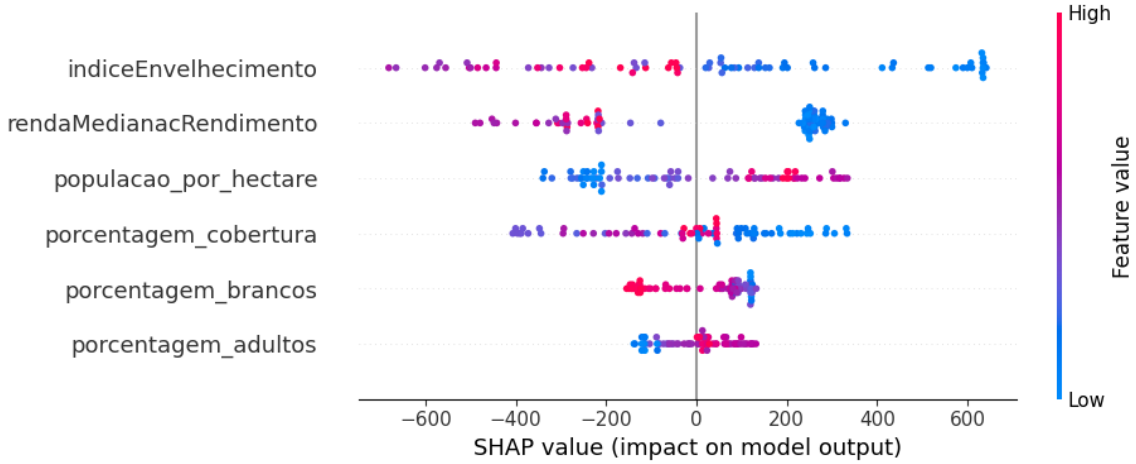


Figura 5.8: Distribuição dos valores SHAP

Por fim, na figura 5.8 podemos ver todos os pontos dos valores SHAP em cada observação como um todo, sendo a distância no eixo X o valor SHAP medido para a observação, e na coloração, a distribuição do valor da *feature*. Notamos uma boa separação dos valores entre as *features*, assim como sua influência em média. Notamos uma grande separação no índice de

envelhecimento e renda, com uma correlação inversa, já a densidade populacional apresenta uma correlação direta com a doença.

Podemos concluir que as explicações de *features* geradas para nosso modelo seguem uma explicação intuitiva, assim como o entendimento técnico para a pandemia (Carozzi et al., 2020) (Aleta e Moreno, 2020), e dado o algoritmo TreeSHAP (Lundberg et al., 2018a) implementado pelas regras de inferência causal (Janzing et al., 2020) podemos assegurar que há de fato uma causalidade entre o valor da *feature* e o valor SHAP da mesma no modelo. E isso, dado os valores de erro obtidos pelo modelo, podemos concluir que as explicações geradas são razoavelmente explicativas, na medida dos dados, ao problema real da pandemia.

6 CONCLUSÃO

Este capítulo conclui o estudo resumindo o que foi realizado. Também apresenta possíveis aplicações para os resultados obtidos por meio do modelo e das explicações. E, por último, apresenta oportunidades de pesquisa futura nas explicações de modelos de machine learning e a explicação dos fatores sociodemográficos em uma pandemia na cidade de Curitiba.

Modelos de previsão para casos de COVID-19 utilizando machine learning já são amplamente disponíveis desde os primórdios da pandemia, no entanto, este estudo dispõe de um modelo compreensível e explicável, utilizando somente os dados socioeconômicos públicos e apresentando explicações para os resultados de forma completa. Por meio do uso do modelo Extra Trees, obtemos resultados superiores em relação a outros modelos de referência na área da explicação de *features*, como LightGBM e XGBoost, no mesmo dataset. Os resultados e métricas obtidas nos habilitaram assim a validar a maior acurácia das explicações quanto ao problema real.

Estes resultados podem ser utilizados, dada a sua acurácia e erros aceitáveis, na tomada de decisões públicas em quanto a informar os responsáveis dos fatores sociodemográficos mais importantes, e como eles influenciam no número de casos, podendo então contribuir à sociedade com decisões melhor informadas.

Existem diversos trabalhos a serem realizados na área de explicações de modelos, tanto na área da saúde como em diversas outras. Podemos considerar a ampliação do escopo do trabalho com mais variáveis de interesse à políticas públicas. O trabalho pode ser expandido também no sentido da utilização de outros modelos, a fim de aproveitar melhor os dados geográficos, como modelos geoespaciais trabalhados em Nuijten et al. (2024). Com a utilização deste tipo de modelo, a interação dos bairros entre si poderia ser melhor aproveitada.

Em outras áreas de aplicação de machine learning, as aplicações são extensas, a explicação de modelos auxilia os usuários e desenvolvedores não apenas como ferramenta, auxiliando o desenvolvedor a compreender o problema e o objeto de trabalho de maneira profunda, mas também os usuários que, ao fim, podem realizar tomadas de decisão melhor embasadas.

REFERÊNCIAS

A. C. R. Trovão, E. C. da Costa, L. M. d. S. (2021). Impactos da covid-19 e da estratégia de administração da pandemia em curitiba nas metas ods 3. https://www.observatoriodasmetropoles.net.br/impactos-da-covid-19-e-da-estrategia-de-administracao-da-pandemia-em-curitiba-nas-metas-ods-3/. Acessado em 20/11/2024.

Aleta, A. e Moreno, Y. (2020). Age differential analysis of covid-19 second wave in europe reveals highest incidence among young adults. *MedRxiv*, páginas 2020–11.

Bifet, A., Read, J., Xu, C. et al. (2022). Linear tree shap. *Advances in Neural Information Processing Systems*, 35:25818–25828.

Carozzi, F., Provenzano, S. e Roth, S. (2020). Urban density and covid-19.

Debjit, K., Islam, M. S., Rahman, M. A., Pinki, F. T., Nath, R. D., Al-Ahmadi, S., Hossain, M. S., Mumenin, K. M. e Awal, M. A. (2022). An improved machine-learning approach for covid-19 prediction using harris hawks optimization and feature analysis using shap. *Diagnostics*, 12(5):1023.

Developers, F. (2017). Prophet diagnostics. https://facebook.github.io/prophet/docs/diagnostics.html. Accessed: 2024-12-06.

Genuer, R., Poggi, J.-M. e Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14):2225–2236.

Geurts, P., Ernst, D. e Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

Hasanah, S. N., Herdiyeni, Y. e Hardhienata, M. K. D. (2023). The impact of socioeconomic and demographic factors on covid-19 forecasting model. *Journal of Information Systems Engineering & Business Intelligence*, 9(1).

Janzing, D., Minorics, L. e Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. Em *International Conference on artificial intelligence and statistics*, páginas 2907–2916. PMLR.

Kawachi, I. (2003). *Neighborhoods and health*. Oxford University Press.

Khanijahani, A., Iezadi, S., Gholipour, K., Azami-Aghdash, S. e Naghibi, D. (2021). A systematic review of racial/ethnic and socioeconomic disparities in covid-19. *International journal for equity in health*, 20:1–30.

Levy, B. L., Vachuska, K., Subramanian, S. e Sampson, R. J. (2022). Neighborhood socioeconomic inequality based on everyday mobility predicts covid-19 infection in san francisco, seattle, and wisconsin. *Science advances*, 8(7):eabl3825.

Liu, Y., Liu, Z., Luo, X. e Zhao, H. (2022). Diagnosis of parkinson's disease based on shap value feature selection. *Biocybernetics and Biomedical Engineering*, 42(3):856–869.

Lundberg, S. (2018). Shap module. `https://shap.readthedocs.io/en/latest/`. Accessed: 2024-12-06.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N. e Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Lundberg, S. M., Erion, G. G. e Lee, S.-I. (2018a). Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J. et al. (2018b). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.

Meade, N. (1983). Industrial and business forecasting methods, lewis, cd, borough green, sevenoaks, kent: Butterworth, 1982. price:£ 9.25. pages: 144.

Morales, K. F., Paget, J. e Spreeuwenberg, P. (2017). Possible explanations for why some countries were harder hit by the pandemic influenza virus in 2009–a global mortality impact modeling study. *BMC infectious diseases*, 17:1–12.

Moreno, J. J. M., Pol, A. P., Abad, A. S. e Blasco, B. C. (2013). Using the r-mape index as a resistant measure of forecast accuracy. *Psicothema*, 25(4):500–506.

Nuijten, R. J., Coops, N. C., Theberge, D. e Prescott, C. E. (2024). Estimation of fine-scale vegetation distribution information from rpas-generated imagery and structure to aid restoration monitoring. *Science of Remote Sensing*, 9:100114.

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S. e Mohammadian, A. K. (2020). Toward safer highways, application of xgboost and shap for real-time accident detection and feature analysis. *Accident Analysis & Prevention*, 136:105405.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

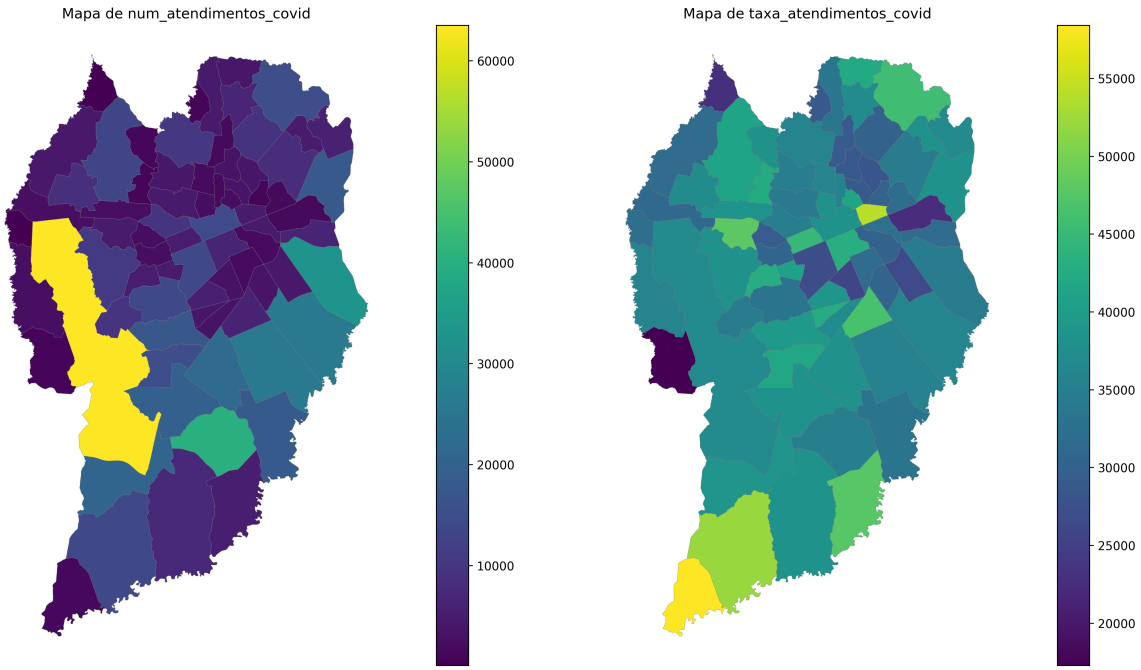
Probst, P., Wright, M. N. e Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301.

Scott, M., Su-In, L. et al. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774.

Shapley, L. S. (1951). Notes on the n-person game—ii: The value of an n-person game. *Lloyd S Shapley*.

Soares, L. D., Kunh, P. D., Corrêa, J. M. e dos Santos, J. A. A. (2020). Previsão de Óbitos por covid-19 no brasil utilizando redes neurais artificiais. *XL ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO*.

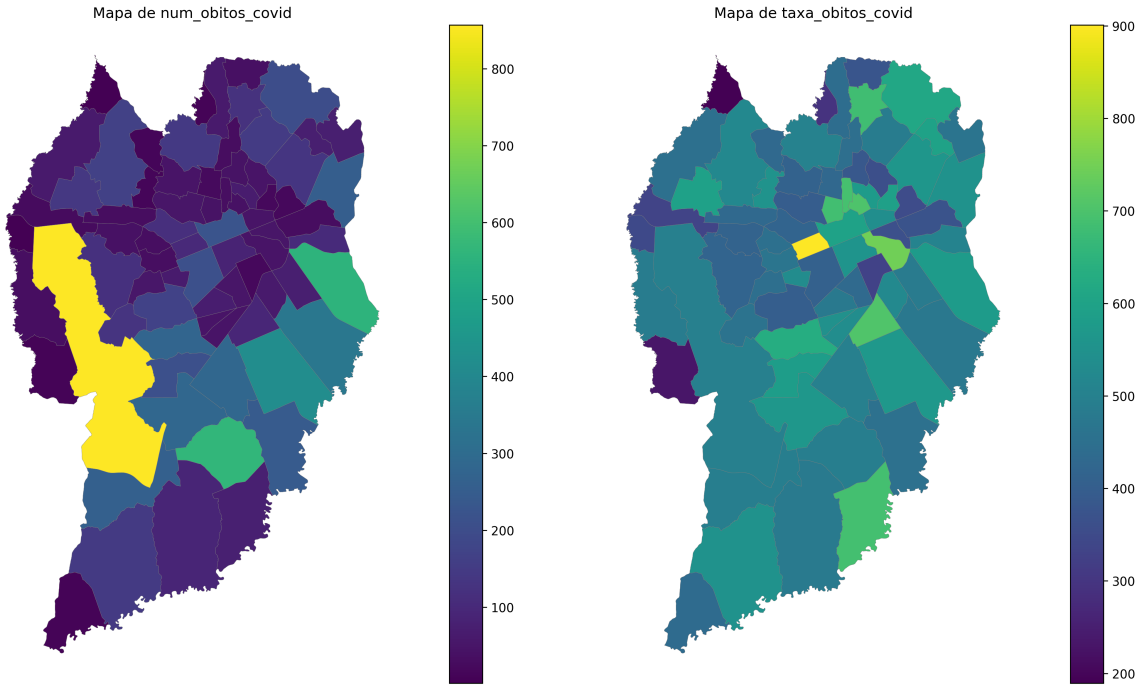
APÊNDICE A – MAPAS COM DADOS GEOGRÁFICOS



(a) Número total de casos registrados de COVID-19.

(b) Numero de casos por 100 mil habitantes.

Figura A.1: Comparação entre o numero absoluto e a taxa normalizado por população.

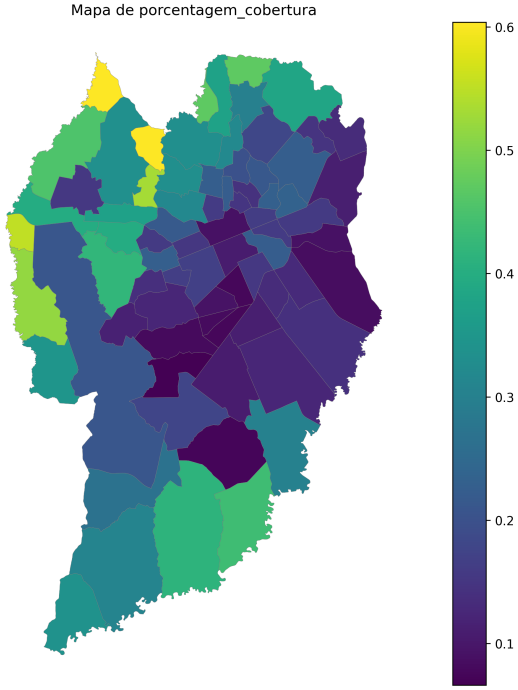


(a) Número total de óbitos registrados de COVID-19.

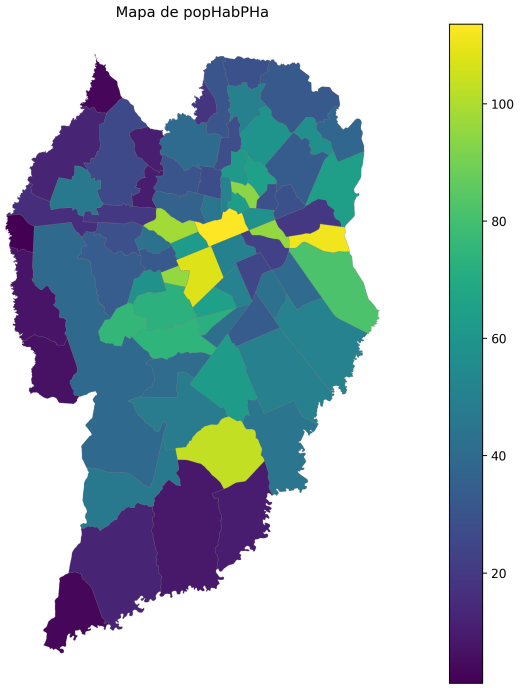
(b) Numero de óbitos por 100 mil habitantes.

Figura A.2: Comparação entre o número absoluto e a taxa normalizado por população.

Enquanto no número total, os dados são altamente tendenciosos quanto à população, no mapa com as taxas é possível visualizar as diferenças.

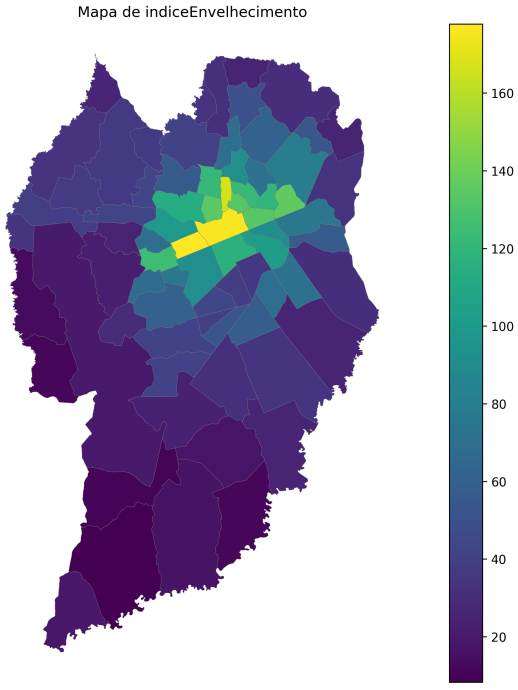


(a) Porcentagem da área coberta por vegetação detectável.

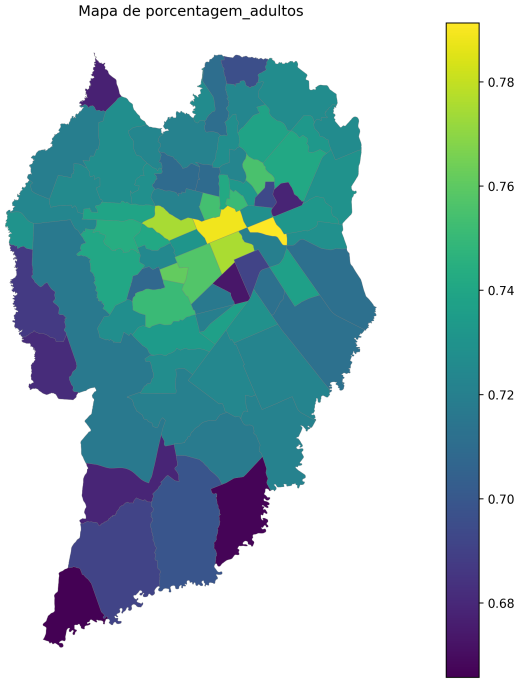


(b) Densidade populacional, em numero de habitantes por hectare.

Figura A.3: Notamos que ambos são inversamente proporcionais, no geral.

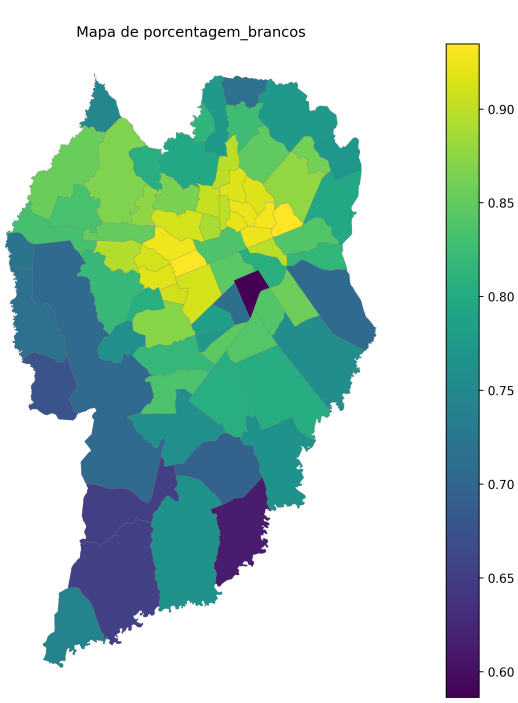


(a) Índice de envelhecimento, que é o número de idosos para cada 100 crianças de 0 a 14 anos.

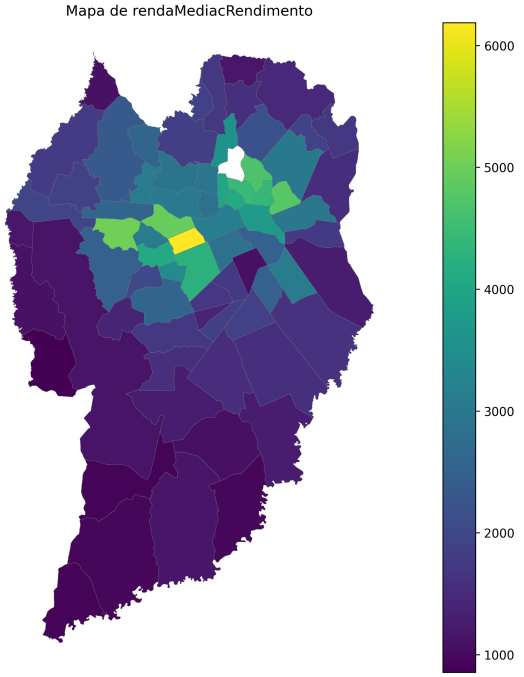


(b) porcentagem de adultos na população total.

Figura A.4: Ainda que centrados de maneira similar, nota-se uma maior concentração de idosos em bairros centrais, enquanto a população de adultos é distribuída mais uniformemente pela cidade.



(a) Porcentagem da população autodeclarada branca.



(b) Renda pessoal mediana, incluindo rendimentos somados a salários.

Figura A.5: Os mapas mostram a correlação, no caso da cidade de Curitiba ha uma correlação direta entre cor, e renda.

APÊNDICE B – DISTRIBUIÇÃO DE VALORES SHAP POR VARIÁVEL

Mostramos a relação de comportamento entre os valores das variáveis do modelo e sua razão nos valores SHAP das mesmas. Em destaque de cor, a segunda variável que mais interage e influencia a variável em foco.

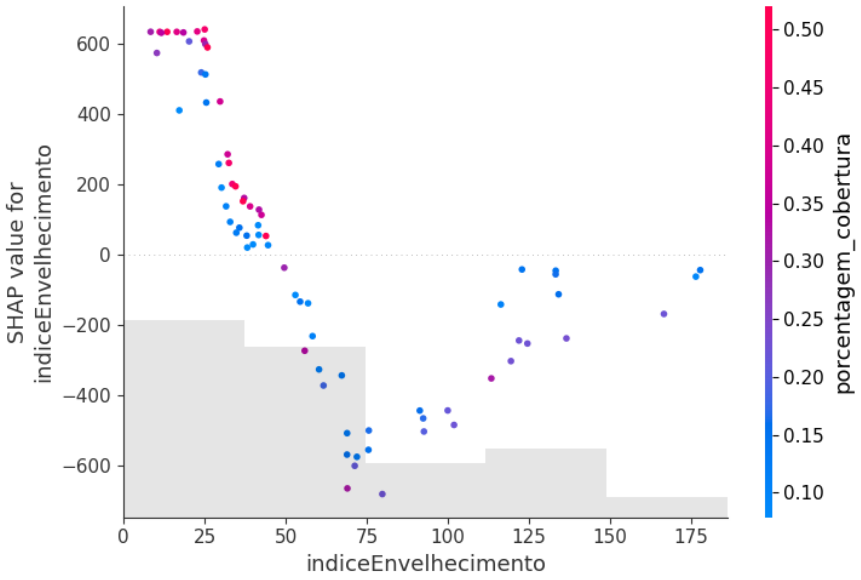


Figura B.1: Índice de envelhecimento por seus respectivos valores SHAP, com a porcentagem de cobertura verde

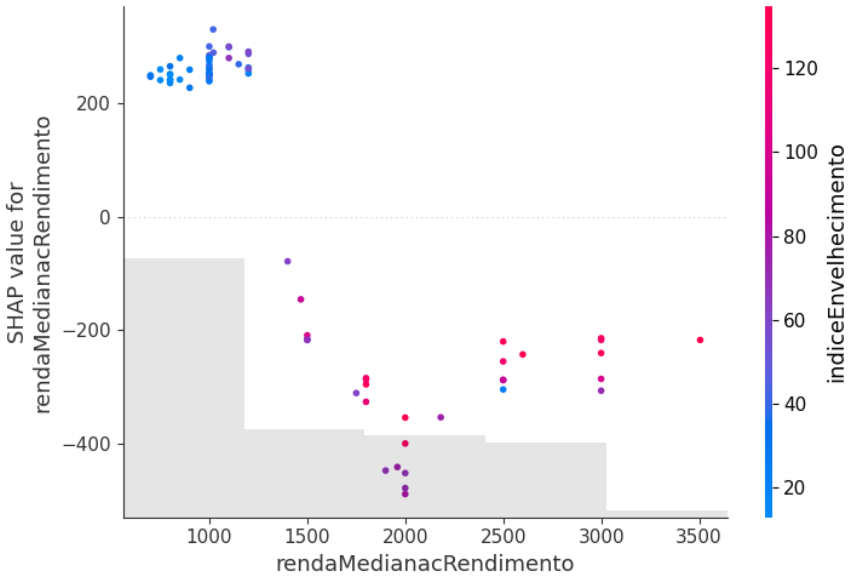


Figura B.2: Mediana do rendimento por seus respectivos valores SHAP, com o índice de envelhecimento

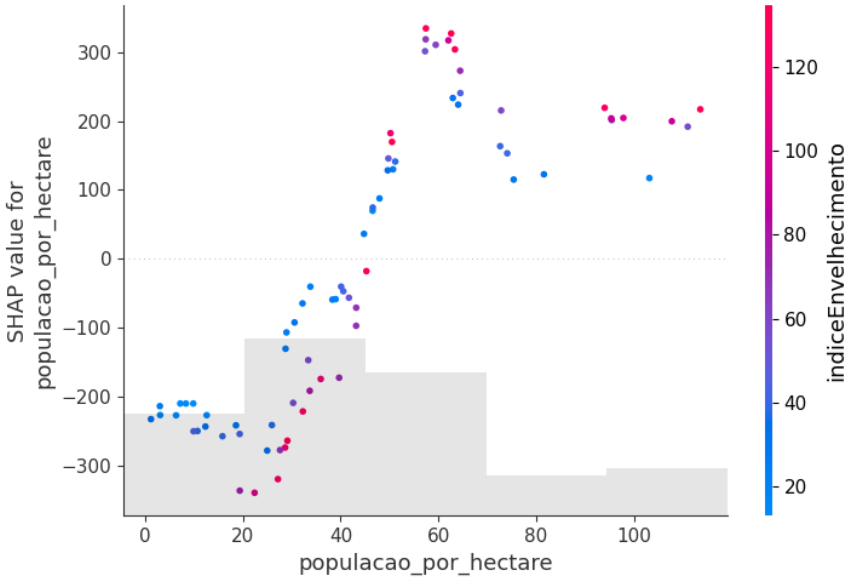


Figura B.3: Densidade por seus respectivos valores SHAP, com o índice de envelhecimento

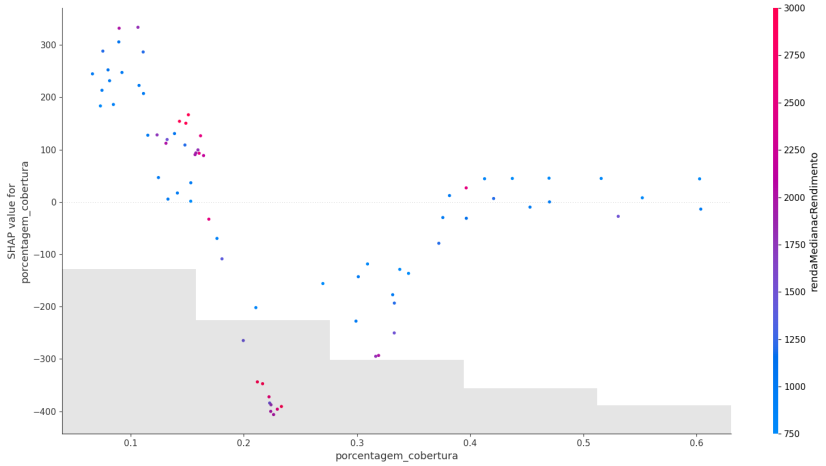


Figura B.4: Porcentagem de cobertura por seus respectivos valores SHAP, com a mediana do rendimento

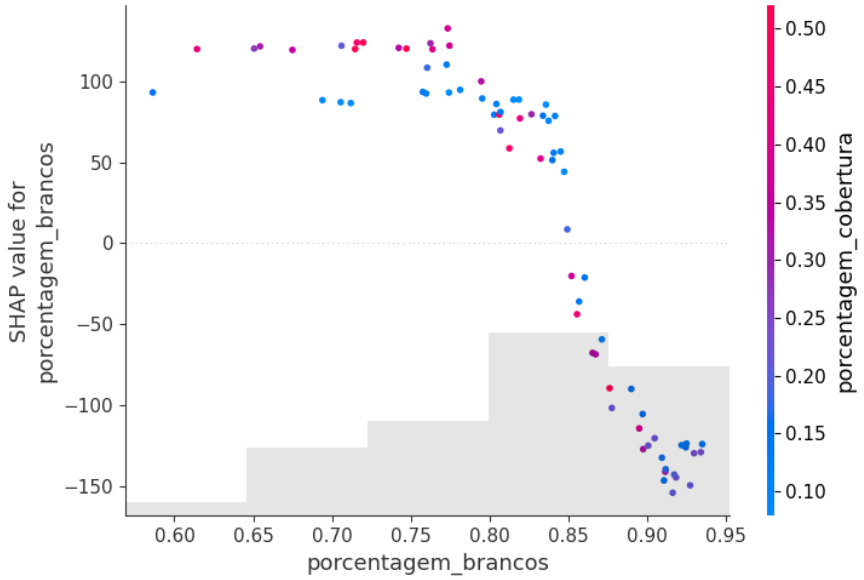


Figura B.5: Porcentagem de população autodeclarada branca por seus respectivos valores SHAP, com a porcentagem de cobertura verde

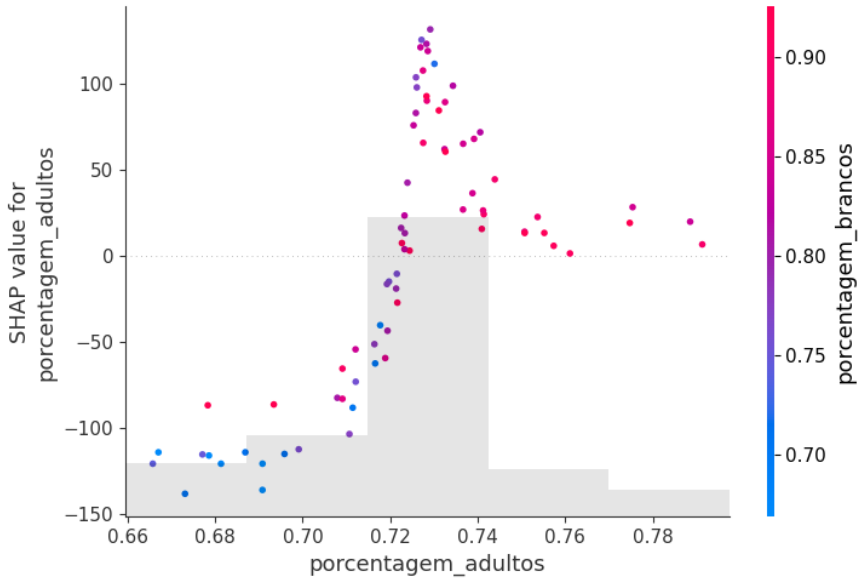


Figura B.6: Porcentagem de população adulta por seus respectivos valores SHAP, com a porcentagem de pessoas brancas