UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS SOCIAIS APLICADAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS CONTÁBEIS
CURSO DE DOUTORADO EM CIÊNCIAS CONTÁBEIS

INTELIGÊNCIA ARTIFICIAL, PREDIÇÃO INFORMACIONAL E O RISCO DE SOLVÊNCIA NA SAÚDE SUPLEMENTAR BRASILEIRA

LINHA DE PESQUISA: INFORMAÇÃO CONTÁBIL PARA USUÁRIOS INTERNOS

JOÃO PESSOA
2024

MARÍLIA AUGUSTA RAULINO JÁCOME

INTELIGÊNCIA ARTIFICIAL, PREDIÇÃO INFORMACIONAL E O RISCO DE SOLVÊNCIA NA SAÚDE SUPLEMENTAR BRASILEIRA

Tese apresentada ao Programa de Pós-Graduação em Ciências Contábeis da Universidade Federal da Paraíba, do Centro de Ciências Sociais Aplicadas, como requisito para obtenção do grau de Doutora em Ciências Contábeis.

Linha de pesquisa: Informação contábil para usuários internos

Orientador: Prof. Dr. Wenner Glaucio Lopes Lucena.

JOÃO PESSOA

2024

MARÍLIA AUGUSTA RAULINO JÁCOME

INTELIGÊNCIA ARTIFICIAL, PREDIÇÃO INFORMACIONAL E O RISCO DE SOLVÊNCIA NA SAÚDE SUPLEMENTAR BRASILEIRA

Esta tese foi julgada adequada para obtenção do grau de Doutora em Ciências Contábeis, em linha de pesquisa de Informação contábil para Usuários Internos, e aprovada em forma final pelo Programa de Pós-Graduação em Ciências Contábeis da Universidade Federal da Paraíba.

BANCA EXAMINADORA:

gov.br Documento assinado digitalmente
WENNER GLAUCIO LOPES LUCENA
Data: 19/03/2024 17:51:57-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Wenner Glaucio Lopes Lucena
Universidade Federal da Paraíba

gov.br Documento assinado digitalmente
ROBERIO DANTAS DE FRANCA
Data: 19/03/2024 09:28:44-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Robério Dantas de França
Membro Interno – PPGCC/UFPB

gov.br Documento assinado digitalmente
CLAUDIMAR PEREIRA DA VEIGA
Data: 06/03/2024 09:58:35-0300
Verifique em https://validar.iti.gov.br

Profª. Drª Claudimar Pereira da Veiga
Membro Externo – UFPR

gov.br Documento assinado digitalmente
ANAILSON MARCIO GOMES
Data: 18/03/2024 17:50:27-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Anailson Márcio Gomes
Membro Externo – UFRN

gov.br Documento assinado digitalmente
FLAVIO LUIZ DE MORAES BARBOZA
Data: 18/03/2024 18:09:34-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Flávio Luiz de Moraes Barboza
Membro Externo – UFU

Catalogação na publicação
Seção de Catalogação e Classificação

J17i Jácome, Marilia Augusta Raulino.
Inteligência artificial, predição informacional e o risco de solvência na saúde suplementar brasileira / Marilia Augusta Raulino Jácome. - João Pessoa, 2024.
124 f. : il.

Orientação: Wenner Glaucio Lopes Lucena.
Tese (Doutorado) - UFPB/CCSA.

1. Inteligência artificial. 2. Machine learning. 3. Predição informacional. 4. Risco de solvência. 5. Saúde suplementar. I. Lucena, Wenner Glaucio Lopes. II. Título.

UFPB/BC CDU 004.8(043)

Elaborado por Walqueline da Silva Araújo - CRB-15/514

Aos meus pais, Jônatas e Edileuza, meu irmão, Jônatas Jr, e Geovana, minha sobrinha, vocês são o que tenho de mais bonito nessa vida! Amo vocês!

A história da ciência é também a história de mulheres corajosas que desafiaram o status quo, iluminando o caminho para descobertas extraordinárias. Que esta tese seja um tributo a todas as mentes femininas que com muito talento exploraram os limites do conhecimento e abriram novos horizontes no vasto universo do saber.

AGRADECIMENTOS

Agradecer, neste momento, é um ato de consciência das minhas limitações. É reconhecer que por trás desta tese concluída, há uma equipe de pessoas reais que contribuíram, aconselharam, incentivaram e acreditaram que eu seria capaz, sobre as quais cito nominalmente:

Deus, eu o vi na inspiração da ideia desta tese, o vi na companhia diária nesses 4 anos e o vi na generosidade de todas as mãos que me ajudaram. Eu o vi em tudo. Você foi e é o meu maior orientador. Eu o amo. Obrigada.

Vovó Lilia, Vovó Penha e Vovô Jácome (todos *in memoriam*), vocês me mostraram que o amor verdadeiro está para além dessa vida e que é possível agradecer pela herança imaterial tão presente em cada ato ou gesto meu. Amo vocês, para além dessa vida. obrigada.

Jônatas (pai), Edileuza (mãe), Jônatas Jr (irmão) e Geovana (sobrinha), vocês são o que tenho de mais belo e valioso nessa vida. É por vocês sempre, amo vocês. Obrigada.

Família e amigos(as), vocês aqueceram o meu coração e não pouparam incentivos. Vocês são a melhor rede de apoio que eu poderia ter tido. Amo vocês, obrigada.

Colegas de turma (Caritsa, Geisa, Gilson, Jaqueline, Thamirys, Risolene e em especial Ingrid), conhecer e caminhar academicamente com vocês foi um presente. Obrigada pelo apoio, online ou presencial.

UFPB e PPGCC (em especial à Coordenação, Wilma e Cecília), pela estrutura, corpo docente e suporte administrativo ao longo desses 4 anos. Vocês foram impecáveis. Obrigada.

Prof. Wenner Lucena, pela paciência, sugestões e orientações ao longo das disciplinas, trabalhos e tese. Obrigada.

Aos professores da banca examinadora, pela participação e contribuições, especialmente a Profª Rossana Guerra, que além das valiosas sugestões para a tese, sempre me trouxe incentivo durante a construção do trabalho. Profª Rossana, a senhora foi e é uma inspiração. Obrigada.

Igor, Márcia e Gabi, cada dica, ajuste, leitura e releitura fizeram com que esta tese fosse concluída. A vocês, um muito obrigada bem especial.

Aos pesquisadores(as) na área de saúde suplementar e inteligência artificial e à Agência Nacional de Saúde Suplementar (ANS), vocês me antecederam e abriram o caminho para a colaboração e avanço das pesquisas acadêmicas nesse segmento. Obrigada.

Com o coração aliviado e sensação de dever cumprido, a mais sincera gratidão por ter concluído esse ciclo. Obrigada.

RESUMO

JÁCOME, Marília Augusta Raulino. **Inteligência artificial, predição informacional e o risco de solvência na saúde suplementar brasileira. 2024.** 126 f. Tese (Doutorado em Ciências Contábeis) – Programa de Pós-Graduação em Ciências Contábeis da Universidade Federal da Paraíba, 2024.

Esta tese investiga o impacto da inteligência artificial (IA) na predição de despesas assistenciais e no risco de solvência das Operadoras de Planos de Saúde (OPS) brasileiras. Utilizando machine learning, a análise de cluster com modelos *Fuzzy C-Means* e *K-means* segmentou dados reais em grupos de risco. Para prever despesas assistenciais, foram empregados algoritmos *K-nearest neighbors* (KNN), *Random Forest* e *Extreme Gradient Boosting (XGBoost)*. Os resultados mostraram acurácia de 99,06% a 99,26% na predição das despesas das OPS Alfa, Beta e Gama, indicando que a IA pode impactar positivamente o risco de solvência das OPS e otimizar a alocação de recursos. Diretrizes foram propostas para a gestão de riscos de solvência das OPS, promovendo a integração da IA nas práticas operacionais e contribuindo para um sistema de saúde mais resiliente e sustentável. As contribuições abrangem aspectos econômicos, sociais e de governança corporativa, oferecendo aplicações práticas que podem influenciar positivamente a economia, sociedade e meio ambiente na saúde suplementar brasileira.

Palavras-chave: Inteligência artificial; *machine learning*; predição informacional; risco de solvência; saúde suplementar.

ABSTRACT

JÁCOME, Marília Augusta Raulino. **Artificial intelligence, informational prediction and the risk of solvency in Brazilian supplementary healthcare. 2024.** 126 f. Tese (Doutorado em Ciências Contábeis) – Programa de Pós-Graduação em Ciências Contábeis da Universidade Federal da Paraíba, 2024.

This thesis investigates the impact of Artificial Intelligence (AI) on predicting healthcare expenses and solvency risk in Brazilian Health Insurance Companies (HICs). Using cluster analysis with Fuzzy C-Means and K-means, real-world databases integrating accounting, healthcare, and sociodemographic data were segmented into risk groups. For predicting healthcare expenses (claims), machine learning algorithms such as K-nearest neighbors (KNN), Random Forest, and Extreme Gradient Boosting (XGBoost) were used. Results show AI models achieving 99.06% to 99.26% accuracy in predicting expenses for Alpha, Beta, and Gamma HICs, indicating AI's potential to positively impact solvency risk and optimize financial cycles and resource allocation. Based on these findings, guidelines for solvency risk management were proposed, marking a significant step towards integrating AI into HIC practices, enhancing system resilience and sustainability. Contributions span economic, social, and corporate governance aspects, demonstrating both theoretical advancement and practical applications with potential positive impacts on the Brazilian supplementary health sector.

Key words: Artificial intelligence; *machine learning*; informational prediction; solvency risk; supplementary health.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SIGLAS

ANA - Agência Nacional de Águas

ANAC - Agência Nacional de Aviação Civil

ANATEL - Agência Nacional de Telecomunicações

ANCINE - Agência Nacional de Cinema

ANEEL - Agência Nacional de Energia Elétrica

ANM - Agência Nacional de Mineração

ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis

ANPD - Agência Nacional de Proteção de Dados

ANS – Agência Nacional de Saúde Suplementar

ANSAU - Agência Nacional de Saúde

ANTT - Agência Nacional de Transportes Terrestres

ANVISA - Agência Nacional de Vigilância Sanitária

CF – Constituição Federal

DIOPS - Documento de Informações periódicas das Operadoras de Planos de Saúde

IA – Inteligência Artificial

IDSS - Índice de Desempenho da Saúde Suplementar

IESS – Instituto de Estudos de Saúde Suplementar

KMPG - Audit, Tax e Advisory

OPS – Operadora de Planos de Saúde Suplementar

PEONA - Provisão para Eventos Ocorridos e Não Avisados

PIB – Produto Interno Bruto

PPA - Procedimentos Previamente Acordados

RN – Resolução Normativa

ROE - Retorno sobre o Patrimônio Líquido

SUS – Sistema Único de Saúde

SUMÁRIO

1 INTRODUÇÃO

1.1 **Contextualização do tema**

A demanda por serviços de saúde é um elemento inerente à vida humana e à manutenção desta, sendo considerado direito universal, reconhecido a partir da Constituição Federal do Brasil de 1988, delegando ao Estado o dever de garanti-la a todos os cidadãos. Para a materialização deste direito, o Estado estruturou o Sistema Único de Saúde (SUS), consolidando então a sua obrigatoriedade de promover o bem-estar físico, psíquico e social aos cidadãos brasileiros.

Acerca deste Sistema de Saúde Brasileiro, Malta e Merhy (2001) afirmam que a política de saúde no Brasil foi marcada por uma trajetória de constantes reduções de investimentos, repercutindo nos aspectos acerca da qualidade dos serviços demandados. Assim, ao se considerar as constantes e crescentes demandas pelos serviços de saúde e em contraponto, às reduções de recursos direcionados ao SUS e consequente não suprimento das demandas da população, surge a oportunidade do oferecimento de serviços privados de assistência à saúde, originando a saúde suplementar no cenário brasileiro.

Este segmento surgiu sob a ausência da regulação, contudo, em 1998 foi instituído o primeiro marco regulatório, a Lei nº 9.656 intitulada como Lei dos planos de saúde (marco regulatório dos planos de saúde) e posteriormente a Lei nº 9.961 de 2000, que criou a Agência Nacional de Saúde Suplementar [ANS], estabelecendo os mecanismos de atuação desta agência.

A saúde suplementar possui características próprias, desde a concepção de atores, quanto à relação entre estes. Por atores, são definidos: as Operadoras de Planos de Saúde (OPS), que comercializam os serviços de assistência à saúde; os beneficiários, que contratam os planos de saúde e usufruem da assistência; os prestadores, que são contratados pelas OPS para prestarem os serviços de assistência à saúde aos beneficiários e a ANS, figurando como agência reguladora das relações existentes nesse segmento de mercado.

Esta função regulatória possui relação direta com a existência das múltiplas imperfeições do setor, inclusive por considerar que estas contribuem para o desequilíbrio econômico e informacional entre os atores do segmento da saúde suplementar (Andrade; Maia, 2009). Sobre estas imperfeições, se destacam os conflitos decorrentes da ausência de política de reajuste, reembolso de contratos, prazos de carência, aspectos de vigência, renovação

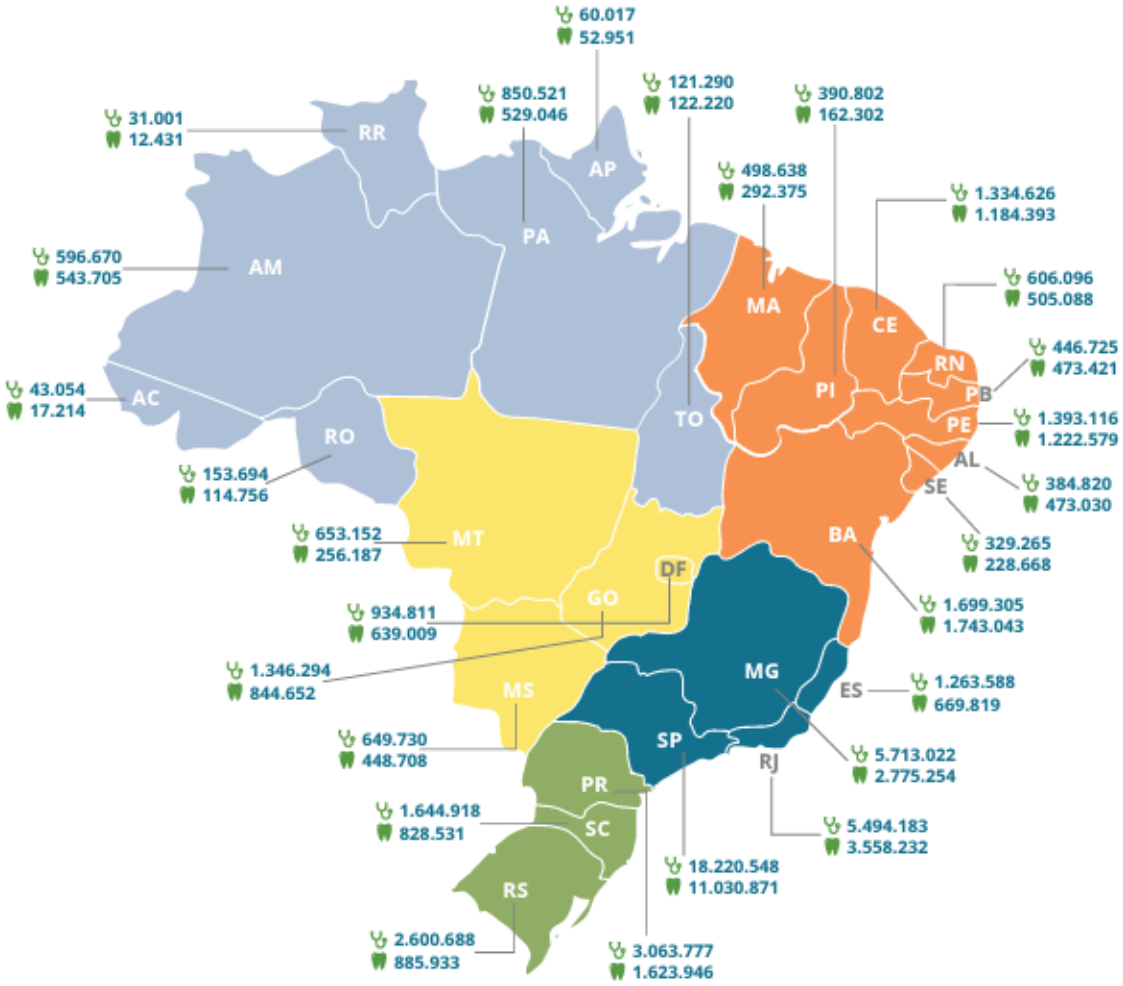
contratual e ainda sobre a insolvência das OPS quando se caracteriza a descontinuidade na prestação de serviços aos beneficiários.

Especificamente sobre as OPS, a regulação (sob o prisma econômico e financeiro) impõe estruturas e obrigações a estas, a fim de monitorar o desempenho econômico e os riscos de continuidade da operação, cumprindo sua finalidade de fomentar o equilíbrio e sustentabilidade ao mercado de saúde suplementar. Para tanto, a regulação estabelece que a contabilidade é responsável pela produção das informações remetidas à ANS (em um primeiro momento) acerca dos aspectos econômicos e financeiros das OPS, e posteriormente remetidas à sociedade em forma de indicadores, promovendo a sinalização acerca das condições econômicas e financeiras das OPS.

Diante da necessidade informacional e acompanhamento das OPS, Salvatori e Ventura (2012) destacam que a operação no segmento da saúde suplementar possui muitos riscos vinculados e sobre estes, Guimarães e Alves (2009) esclarecem que estão atrelados à não previsão, pelas OPS, do gasto futuro com seus beneficiários, incluindo novos procedimentos no rol de cobertura dos planos de saúde, o aumento do custo da assistência (decorrente do avanço tecnológico) e o risco inerente ao negócio, que juntos contribuem para a insolvência deste mercado, podendo comprometer a assistência aos beneficiários.

Especificamente, acerca do impacto social causado pela insolvência, destaca-se que a saúde suplementar, até dezembro de 2023, alcançou a cobertura assistencial médica de cerca de 51 milhões de brasileiros, o que representa 25% da população total do Brasil (ANS, 2023), conforme demonstra a Figura 1 abaixo. Importante destacar que os 26 Estados do Brasil possuem a cobertura da saúde suplementar por meio da existência de Operadoras de Planos de Saúde e respectivamente de seus beneficiários, indicando que o impacto social não se daria apenas localmente, em uma região específica, mas sim com potencial de maior impacto, considerando a amplitude de cobertura e atuação da saúde suplementar.

Figura 1 - Distribuição de beneficiários pelas Unidades da Federação (médico hospitalar e exclusivamente odontológico)



Fonte: SIB/ANS/05-2023.

Portanto, a insolvência deste setor, ensejaria em uma sobrecarga do SUS, impossibilitando a prestação da assistência à saúde, considerando a limitação de investimentos, estrutura física, de pessoal, materiais e medicamentos.

Sob o prisma econômico, este segmento movimentou até dezembro de 2023, R$ 440 bilhões de reais, entre receitas e despesas, com expectativa anual que corresponde a aproximadamente 14% do Produto Interno Bruto (PIB) do Brasil, o que notadamente infere relevância econômica (ANS, 2023). Até o final do terceiro trimestre de 2023, a cadeia produtiva da saúde suplementar registrou o total de 4,8 milhões de vínculos empregatícios registrados formalmente, fato que confere a relevância econômica e social, indicando as possíveis consequências e impactos sociais e econômicos que a insolvência pode ocasionar.

No setor da saúde suplementar, em que a descontinuidade da operação de um agente pode afetar significativamente outros agentes (consumidores e prestadores de serviços, por exemplo), o Estado, por meio da ANS, exerce o controle da entrada e saída de ofertantes,

estabelecendo regras que impeçam comportamentos oportunistas/imprudentes, monitorando a situação econômico-financeira e impondo medidas preventivas para situações que indiquem o risco de insolvência.

Desse modo, frente ao risco de insolvência das OPS, a ANS estabeleceu mecanismos de garantias financeiras e assistenciais, destacando algumas regras prudenciais que dispõem sobre as seguintes medidas: Provisão para Eventos Ocorridos e Não Avisados - PEONA, Provisão de Eventos a Liquidar, Ativos garantidores, Capital Baseado em risco e Margem de Solvência, sendo esta última conceituada como "a capacidade da OPS honrar com todos os custos assistenciais e compromissos financeiros assumidos, mesmo nas situações mais adversas" (Oliveira *et al*., 2015, p.1).

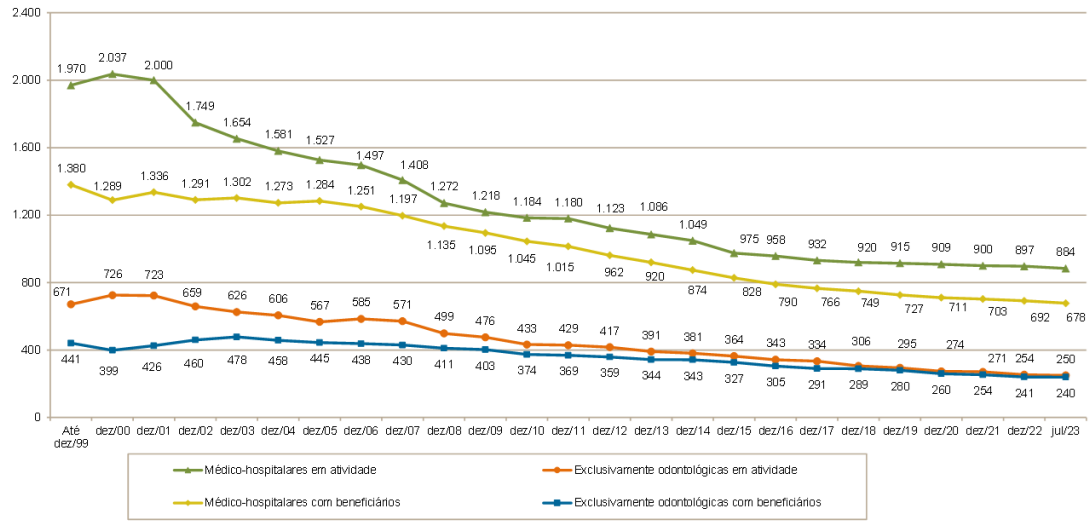
Essencialmente a ANS realiza o acompanhamento econômico e financeiro das OPS por meio das informações contábeis remetidas periodicamente, transformando estas informações contábeis em indicadores econômico-financeiros para cumprir a função de monitoramento e promoção do equilíbrio do segmento da saúde suplementar. Desse modo, trimestralmente as OPS remetem à ANS, o DIOPS (Documento de Informações periódicas das Operadoras de Planos de Saúde), contendo o balancete contábil e fluxo de caixa, ambos trimestrais, para cálculo da margem de solvência e a sua suficiência e patrimônio mínimo ajustado. De maneira complementar, para fechamento de exercício, há a obrigatoriedade de envio do PPA – Procedimentos Previamente Acordados e o Relatório Circunstanciado emitido pela Auditoria independente, para as OPS de médio e grande porte (que possuem acima de 20 mil beneficiários).

Complementar aos já citados, utilizando as informações contábeis remetidas pelas OPS, a ANS editou duas Resoluções Normativas que dispõem sobre a adoção de práticas mínimas de governança corporativa, com ênfase em controles internos e gestão de riscos, para fins de solvência das operadoras de planos de assistência à saúde. A esse respeito, a RN 443/2019 e RN 518/2022, estabeleceram mais 12 indicadores de monitoramento: Margem de lucro líquida; Retorno sobre o Patrimônio Líquido (ROE); Sinistralidade; Percentual de Despesas Administrativas em relação às Receitas de Contraprestação; Percentual de Despesa Comercial em relação à Receita de Contraprestações; Percentual de Despesas Operacionais em relação às Receitas Operacionais; Índice de Resultado Financeiro; Liquidez Corrente; Capital de terceiros sobre capital próprio; Prazo Médio de Contraprestações a receber; Prazo médio de pagamento de eventos; Variação de custos.

Vê-se então que a contabilidade serve ao segmento da saúde suplementar, sobretudo para sinalizar ao mercado e ao órgão regulador acerca da solvência das OPS e por consequência, contribuir para o equilíbrio econômico, financeiro e informacional da saúde suplementar, incluindo os beneficiários dos planos de saúde, que podem se utilizar dessas informações para a escolha na tomada de decisão de contratualização com as OPS e otimização na alocação de recursos.

Sob esta perspectiva informacional da contabilidade, a solvência sinaliza acerca das possibilidades relacionadas à continuidade das operações das OPS, em contrapartida, o estado de insolvência da OPS infere no encerramento das atividades, e sobre este aspecto, a Figura 2 demonstra a sucessiva redução no quantitativo de OPS no exercício das suas atividades. Dentre os motivos atrelados à insolvência das OPS, a Comissão de Inquéritos de Insolvência da ANS (2018), apurou que de 2012 a 2018, as OPS que se tornaram insolventes possuíam falhas de controles internos, gestão de risco e má gestão.

Figura 2 – Operadoras de planos privados de saúde em atividade no Brasil (dezembro/1999 a jul/2023)



Fonte: Sistema de informações de beneficiários/ANS/MS (2023) e CADOP/ANS/MS (2023).

Observa-se a redução média de 55% no quantitativo de OPS em atividade, desde a criação da ANS e a instituição da regulação prudencial setorial até o ano de 2023. Este fato aponta para a necessidade constante e latente da instituição e permanente desenvolvimento da gestão de riscos de solvência nas OPS, iniciando pela gestão de riscos dos sinistros [gastos com a utilização de serviços pelos beneficiários], que em 2023 atingiu o patamar de 88,2% (ANS, 2023).

Outrossim, Araújo e Silva (2018) definem a sinistralidade como um índice (em percentual) calculado ao se relacionar os sinistros ocorridos (custos assistenciais) e o prêmio recebido (receitas da assistência). Este indicador permite aos beneficiários, prestadores de serviços, órgão regulador e a própria OPS, a identificação quanto à sustentabilidade e solvência das OPS, que se relaciona diretamente com a garantia de prestação dos serviços contratados.

Sob a ótica do mercado, o constante encerramento das atividades de OPS pode ensejar na concentração de mercado em grandes *players*, que em posições mais favorecidas, acabam por absorver OPS singulares e autônomas para compor grandes conglomerados. Importante destacar que, conforme a ANS (2023), as duas maiores operadoras concentram 17,7% do total de beneficiários de planos médico-hospitalares do país e as 5 maiores concentram 37,4%, fato que reforça os possíveis impactos sociais para os beneficiários, resultado da falta de concorrência, ocasionando "perda de poder de escolha" ou até mesmo aumento nos valores cobrados, devido à alta concentração do mercado.

Visualizando os impactos sociais e econômicos decorrentes do fenômeno da insolvência na saúde suplementar, a produção informacional pode potencializar a acurácia das decisões das OPS neste cenário complexo e permeado de riscos. Neste percurso de potencializar a acurácia, a tecnologia, e mais especificamente a inteligência artificial, se apresenta como ferramenta para ampliar as possibilidades de tratamento de dados que compõem os cenários decisórios das organizações. Chen, Chiang e Storey (2012) afirmam que a aplicação da inteligência artificial aos negócios é relevante porque as organizações dispõem, cada vez mais, de grande volume e variações de dados e por isto necessitam cada vez mais de ferramentas analíticas de dados para maior acurácia em seus processos decisórios.

Assim, esta tese propõe um encontro uníssono entre a inteligência artificial, dados contábeis, assistenciais, sociodemográficos e o segmento da saúde suplementar, a fim de proporcionar possibilidades preditivas acerca do risco de solvência das OPS, isto por compreender que em um cenário complexo e permeado de riscos, manter a operação ativa e sustentável é um exercício desafiador para as OPS, inclusive para fins de sinalização à ANS e aos demais usuários das informações.

1.2 INDICAÇÃO DO PROBLEMA DE PESQUISA

A partir da contextualização acima, emerge o potencial da discussão acerca do risco de solvência na saúde suplementar brasileira, incluindo as possibilidades do uso da tecnologia para análises preditivas informacionais. Nesse sentido, esta tese busca discorrer sobre a seguinte questão norteadora:

Como a aplicação de inteligência artificial na predição informacional das despesas assistenciais pode impactar o risco de solvência das operadoras de planos de saúde da saúde suplementar brasileira?

Desse modo, a abordagem proposta **investigou o impacto da inteligência artificial na predição informacional das despesas assistenciais na saúde suplementar brasileira e sua relação com o risco de solvência das Operadoras de Planos de saúde**.

A partir de informações e registros contábeis agregados a dados assistenciais e sociodemográficos, defende-se a tese que: **a utilização de modelagens preditivas informacionais da previsibilidade de gastos [despesas assistenciais], impacta positivamente o risco de solvência das operadoras de planos de saúde brasileiras**.

1.3 OBJETIVOS

1.3.1 OBJETIVO GERAL

Investigar o impacto da inteligência artificial na predição informacional das despesas assistenciais na saúde suplementar brasileira e sua relação com o risco de solvência das Operadoras de Planos de saúde.

1.3.2 OBJETIVOS ESPECÍFICOS

a) Calcular a sinistralidade das OPS brasileiras, segregando por região geográfica, porte e modalidade;

b) Aplicar modelos de inteligência artificial para a predição das despesas assistenciais em operadoras de planos de saúde, considerando informações contábeis, assistenciais e sociodemográficas;

c) Analisar o nível de acurácia das modelagens preditivas informacionais na previsibilidade das despesas assistenciais e seu impacto no risco de solvência;

d) Propor diretrizes para análises preditivas informacionais, a partir da inteligência artificial, visando aperfeiçoar a gestão de risco de solvência das OPS na saúde suplementar brasileira.

1.4 JUSTIFICATIVA

1.4.1 Ineditismo/ originalidade

O ineditismo dessa tese se deve à lacuna de pesquisas publicadas que apresentem soluções que envolvam inteligência artificial ou qualquer outro artefato tecnológico que se apliquem aos dados contábeis e assistenciais a fim de predizer informações que possam impactar a solvência das OPS brasileiras. Assim, esta tese traz características que se direcionam ao ineditismo sob a ótica metodológica e ainda por trazer a investigação no ambiente brasileiro.

No que se refere aos aspectos de originalidade, a tese incorpora a aplicação de artefatos tecnológicos como meios para predição informacional, a partir dos dados contábeis e assistenciais de despesas, a fim de promover cenários de decisões internas antecipadas em relação ao risco de solvência das OPS brasileiras.

Alguns estudos apresentados na literatura possuem aspectos que tangenciam com o problema principal levantado nesta tese, a exemplo, Areias e Carvalho (2021) verificaram como o resseguro impactaria a solvência das operadoras em um cenário hipotético de longo prazo. Coelho de Sá *et al* (2017) aferiram a probabilidade de ruína de uma operadora de plano de saúde no horizonte finito de dez anos e os resultados encontrados reforçaram a importância de os gestores medirem de forma antecipada o impacto de ações estratégicas para a manutenção da solvência da operadora.

Em um cenário de dados alarmantes sobre o aumento da sinistralidade e seu potencial risco, Araújo e Silva (2018) buscaram compreender as mudanças ocorridas no setor de saúde suplementar nos últimos anos, através da análise temporal de séries históricas relacionadas ao setor, encontrando resultados que apontam que, mesmo com o aumento da demanda, houve uma diminuição do número de operadoras em atividade no país. Os autores ainda concluem que o aumento da sinistralidade oferece riscos à sobrevida e à abertura de novas operadoras. Esta

visão acaba por corroborar que a diminuição do número de operadoras pode conduzir o país a uma oligopolização do setor com uma demanda crescente do número de beneficiários.

Em uma outra pesquisa, Bragança *et al* (2019) analisam que em um contexto de crescimento de usuários, contudo, a quantidade de operadoras de planos de saúde (OPS) vem reduzindo desde a criação da ANS, e por isso se dedicaram a analisar a influência da regulação e das intervenções da ANS na continuidade das OPS, apontando como resultado uma influência significativa para a previsão de insolvência das OPS.

Assim, para além da discussão tangencial já produzida por trabalhos anteriores, esta tese se propõe a trazer como seu objetivo a busca por apresentar a combinação de artifícios tecnológicos, dados contábeis e assistências como possibilidades para predição informacional, produzindo impacto no risco de solvência na saúde suplementar, uma vez que em nenhum estudo anterior esta abordagem foi proposta.

1.4.2 Relevância

Destaca-se a relevância desta tese sob três perspectivas: econômica, social e informacional, estando ambas intrinsecamente relacionadas ao tamanho deste segmento e a representatividade para a economia e sociedade brasileira.

A saúde suplementar é um segmento de mercado caracterizado por sua complexidade e particularidade no que diz respeito ao objeto de negócio ofertado, o serviço de assistência à saúde. Sob o ponto de vista econômico, até o mês de dezembro do ano de 2023, 25% dos brasileiros eram beneficiários de planos privados assistência médica, ou seja, cerca de 51 milhões de cidadãos, responsáveis por movimentar (até novembro/23) o montante de R$ 205 bilhões de reais em receitas de contraprestações e 178 bilhões de reais em despesas assistenciais (ANS, 2023).

Dado relevante é que, além dos beneficiários citados acima, este segmento representa a fonte de renda e vínculo empregatício de aproximadamente 4,8 milhões de trabalhadores em sua cadeia produtiva (IESS, 2022).

Ao considerar os montantes circulantes nesta cadeia de valor, em 2022, por meio do processo de ressarcimento ao Sistema Único de Saúde, as OPS enviaram o total de R$ 600 milhões referente ao gasto por utilização dos seus respectivos beneficiários de qualquer serviço usufruído na rede pública de saúde, contribuindo então para a manutenção da assistência pública à saúde.

Sob a perspectiva social, a relevância se concentra na prestação de 1,8 bilhão de serviços de saúde, entre consultas, exames, terapias, cirurgias em 2022. Esse quantitativo reflete a utilização de 51 milhões de beneficiários, utilização esta que, caso houvesse colapso por insolvência neste segmento, acarretaria sobrecarga para o sistema público de saúde do Brasil e consequente déficit na prestação do serviço ao cidadão.

Em que pese o referido aspecto social, é necessário ressaltar que o acesso aos planos de assistência médica se dá em um mercado regulado pela ANS, o que garante aos cidadãos o acesso a um mercado descentralizado, cujos preços possuem limite máximo estabelecido pelo regulador e o cidadão possui livre direito de escolha. Contudo, em um cenário de concentração de OPS, em virtude do alto índice de insolvência, o processo de precificação pode se tornar um fator que inviabiliza o acesso do cidadão à saúde suplementar, demonstrando então ser este aspecto relevante para a condução desse estudo.

Acerca da perspectiva informacional, a regulação assegura que haja o equilíbrio informacional entre os atores desta cadeia de valor, para tanto, a contabilidade serve à sociedade como instrumento de prestação de contas, evidenciando aspectos relativos à situação econômica e financeira das OPS. No entanto, a partir de uma abordagem gerencial, a contabilidade também se mostra relevante para os usuários internos das informações, quais sejam: Gestores financeiros, Diretores, Conselhos de Administração, Conselhos Fiscais, Auditores internos, Gestores de risco e de *compliance*, pois atua como a linguagem que sinaliza acerca do cumprimento da regulação econômica das garantias financeiras, incluindo a margem de solvência das OPS.

A relevância informacional deste trabalho, se direciona para a proposição de produção informacional, majoritariamente para os usuários internos, utilizando dados contábeis e assistenciais como meios potencialmente capazes de impactar positivamente o risco de solvência das OPS, promovendo então novas possibilidades que viabilizam a tomada de decisão antecipada e estratégica em relação ao risco de solvência.

1.4.3 Contribuição

Os resultados propostos nesta tese, contribuem para a sustentabilidade do mercado da saúde suplementar, na medida em que apresenta aos usuários internos das informações contábeis, a predição das despesas assistenciais como uma alternativa para que estes possam

implementar mecanismos de monitoramento somados à atuação estratégica e preventiva face as despesas assistenciais que impactam diretamente o risco de solvência das OPS.

Importante destacar que a predição das despesas assistenciais não enseja na prática de seleção de risco como medida anterior ao ingresso dos beneficiários aos planos de saúde, fato que iria contrariar os princípios da regulação, mas, se refere à possibilidade de iniciar a trajetória do beneficiário na saúde suplementar de maneira planejada e acima de tudo, cercada pelo aspecto preventivo.

Lidar preventivamente com as despesas assistenciais, enseja em, mesmo que indiretamente, atuar para a redução das fraudes e gastos indevidos pelos beneficiários das OPS. Destaca-se que, de acordo com o estudo desenvolvido IESS (Instituto de Estudos na Saúde Suplementar), em 2017 o prejuízo decorrente destes, alcançou a cifra de R$ 28 bilhões, o que representou 19,1% do total das despesas assistenciais neste mesmo ano.

Por outro lado, a atuação preventiva das OPS, em relação às despesas assistenciais, promoverá, consequentemente, maiores cuidados com a saúde e incentivos de cuidados aos seus beneficiários, no tocante aos programas de prevenção e outros mecanismos de gestão da carteira e monitoramento constante dos grupos de risco, emergindo então, resultados positivos e concretos em favor da saúde dos cidadãos que possuem planos privados de assistência médica.

Pelo exposto, a tese é relevante por ter o propósito finalístico de contribuir para a sustentabilidade do mercado de saúde suplementar, o que por consequência implica na continuidade da cadeia de valor e prestação de serviços de assistência à saúde aos seus beneficiários.

Além disso, os resultados possibilitarão identificar em quais regiões e modalidades, os impactos no risco de solvência serão mais altos, sendo possível agregar novos conhecimentos acerca da importância da predição informacional e os seus benefícios no mercado de saúde suplementar, cujos estudos ainda são escassos na área acadêmica.

2 REVISÃO DA LITERATURA

2.1 TEORIA DA REGULAÇÃO ECONÔMICA

O estudo da teoria da regulação econômica remonta ao surgimento dos chamados "monopólios naturais", conceito que surgiu na revolução industrial. Baumol (1977) fornece a seguinte definição formal de monopólio natural: "[uma] indústria na qual a produção de várias empresas é mais cara do que a produção por monopólio", ou seja, é um monopólio em uma indústria cuja característica principal são altos custos de infraestrutura, o que se configura em

uma barreira à entrada de outros fornecedores, logo, este primeiro fornecedor em um mercado, possui uma vantagem acentuada sobre concorrentes potenciais.

Durante esse período de rápido crescimento industrial, se tornou evidente que alguns segmentos industriais, devido à natureza de seus produtos e custos fixos substanciais, eram naturalmente inclinados ao monopólio. Esse contexto fez emergir a preocupação com a posição de monopólio assumida por empresas, inclusive considerando que estas poderiam potencializar seu poder de mercado para prejudicar os consumidores.

Nesse contexto, as discussões acerca da teoria da regulação econômica nasceram a partir da análise de conflitos vivenciados no mercado de monopólio natural. Alfred Marshall, um dos economistas pioneiros da teoria do consumidor e da concorrência imperfeita, trouxe uma das primeiras contribuições para a teoria da regulação econômica em seu livro "Princípios de Economia" (1890). Nesta obra, Marshall discutiu a necessidade de regulamentação governamental em setores que a concorrência pura e perfeita não era possível, lançando luz sobre a necessidade de maior discussão e desenvolvimento desta teoria.

No entanto, foi a partir das décadas de 1930 e 1940 que a teoria da regulação econômica começou a se desenvolver mais sistematicamente. Economistas como George Stigler e Ronald Coase contribuíram para a compreensão dos incentivos que as empresas reguladas têm para influenciar reguladores e explorar sua posição de monopólio.

O artigo seminal de Ronald Coase, "*The Nature of the Firm*" (A Natureza da Empresa/firma), publicado em 1937, contribuiu significativamente para a teoria da regulação econômica. Embora o foco central tenha sido em questões de organização industrial e teoria das empresas, as ideias discutidas, trouxeram influência acerca da compreensão da regulação em mercados com estruturas de propriedade distintas.

Coase (1937) introduziu o conceito de custos de transação, abordando que a existência das empresas pode ser um caminho para a redução do custo de transação, à medida em que reduzem os custos de negociação, que normalmente ocorreriam caso as transações fossem conduzidas inteiramente no mercado. Essa ideia tem implicações importantes para a teoria da regulação econômica, pois sugere que a escolha entre a regulação governamental e a coordenação de mercado depende dos custos de transação envolvidos.

A partir das ideias trazidas por Coase (1937), a discussão acerca da Teoria da Regulação evoluiu com o trabalho seminal de Stigler (1971), "*The Theory of Economic Regulation*", em que o autor debate quais os atores serão beneficiados, a partir da regulação, ou até mesmo quais arcarão com as obrigações oriundas da instituição da regulação. Outro aspecto central desta

discussão se direcionou sobre qual a forma da política regulatória e os efeitos dela sobre a alocação dos recursos, o que na visão de Veljanovski (2010) se configura como a resposta ao questionamento: “por que temos a regulação que temos?”.

Além da relação que atribui a existência da regulação meramente como resposta às falhas de mercado, Stigler (1971, p. 3) assevera que “a regulação se institui primordialmente para a proteção ou benefício do público em geral ou uma grande – subclasse - de público”. A discussão de Stigler se direciona à teoria da captura regulatória e, mais especificamente, como as indústrias frequentemente influenciam reguladores em seu benefício. Nesse sentido, as empresas buscam capturar reguladores, pressionando por políticas que favoreçam seus interesses privados em detrimento do interesse público. Desse modo, Stigler (1971) ressalta que o processo regulatório não é imune à interferência de interesses privados e que a captura regulatória ocorre quando as indústrias conseguem direcionar as políticas regulatórias para atender aos seus próprios interesses.

Um argumento que o autor destaca é que as agências reguladoras são, em grande parte, compostas por indivíduos que têm ou tiveram ligações com as indústrias que regulam, criando um ambiente propício para a captura. O autor ressalta que a presença de atores econômicos com fortes incentivos para influenciar a regulamentação pode levar a resultados regulatórios que não servem ao interesse público, o que consequentemente contribui para minar a eficácia da regulação. Stigler (1971) traz a reflexão sobre a necessidade de se entender o equilíbrio de poder entre reguladores e indústrias reguladas como parte fundamental da análise regulatória.

Dessa forma, as contribuições de Stigler são amplamente reconhecidas como um marco importante, sendo considerada parte importante da análise econômica da regulação, fornecendo uma estrutura conceitual para avaliar as complexas interações entre governos e setores regulados. Essa importância, tornou o seu trabalho uma referência para as discussões posteriores.

Em 1985 o estudo seminal de Samuelson e Nordhaus, "*The Theory of Public Utility Pricing*" (A Teoria da Precificação de Utilidades Públicas), sobre a regulação de utilidades públicas discutiu sobre os monopólios naturais, fornecendo uma estrutura teórica para a determinação de preços regulatórios e ainda discutindo acerca do equilíbrio entre proteger os interesses dos consumidores e garantir que as empresas pudessem operar eficientemente. Este trabalho explorou a regulação de empresas de serviços públicos, como eletricidade, água e gás, e como os reguladores podem determinar preços justos para essas empresas, considerando seu *status* de monopólio natural.

Samuelson e Nordhaus (1985), introduziram o conceito de "preço regulatório", que é o preço pelo qual uma empresa regulada pode cobrar seus serviços, permitindo-lhe cobrir seus custos e obter um retorno justo sobre o capital investido, mas sem explorar o monopólio. Logo, os autores destacaram a importância de encontrar o equilíbrio entre garantir que as empresas de serviços públicos possam operar de forma viável, mas também evitar que explorem sua posição de monopólio em detrimento dos consumidores.

Os autores também trouxeram a discussão acerca do *Trade-off* entre eficiência econômica e equidade na regulação de utilidades públicas, de modo que a ênfase se direcionou ao acesso aos serviços públicos, com vistas a garantir que preços acessíveis estejam disponíveis para todos os consumidores.

As discussões permitiram a evolução do pensamento sobre a Teoria da Regulação Econômica, e mais especificamente, acerca do contexto brasileiro, estabelecer o *modus operandi* da regulação se configura em um paradigma para a Administração Pública, uma vez que esta atividade se direciona às formas relacionais entre o Estado e as empresas e público em geral.

No contexto brasileiro, embora a regra geral estabeleça o princípio da livre concorrência (CF, art. 170), a intervenção direta ou indireta do Estado, implica em atuar na regulação de mercados a fim de evitar excessos e abusos de poder econômico, por parte dos atores de cada segmento.

Para Salvatori e Ventura (2012), essa estrutura de regulação do Estado Brasileiro sofreu influência advinda dos modelos norte-americano e britânico, influência esta explicada como um formato que fomenta a garantia da transparência e participação das partes que recebem o efeito da regulação.

No Brasil, as Agências Reguladoras surgiram de uma proposta de Reforma do Estado (1995-2002) como uma medida de monitoramento e aprimoramento da regulação das principais atividades econômicas privatizadas. Nesse sentido, além do combate às falhas de mercado, tais como: seleção adversa, o risco moral e a assimetria de informação, Peltzman (2004) indica que a regulação mediaria a relação entre agentes econômicos, assegurando a competitividade de setores e promovendo a proteção dos interesses dos consumidores, parte mais frágil da relação.

Quanto a este aspecto de mediação, Salvatori e Ventura (2012) afirmam que à medida que esses conflitos forem mediados ativamente, por meio da regulação, é possível pressupor uma atuação em conformidade e transparente durante todo o processo, tornando a relação

entre os agentes equilibrada, inclusive sob o aspecto informacional. Logo, há o destaque para o aspecto técnico atrelado à regulação.

No contexto brasileiro, a estrutura regulatória é composta por agências reguladoras, independentes e autônomas para supervisionar e regulamentar setores específicos da economia. As principais agências reguladoras, suas funções e data de criação (conforme os respetivos sites oficiais) seguem:

- Agência Nacional de Energia Elétrica (ANEEL): Regula o setor de energia elétrica (criada em 1996);
- Agência Nacional de Telecomunicações (ANATEL): Regula o setor de telecomunicações (criada em 1997);
- Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP): Regula o setor de petróleo, gás natural e biocombustíveis (criada em 1997);
- Agência Nacional de Saúde Suplementar (ANS): Regula o mercado de planos de saúde e seguros de saúde (criada em 2000);
- Agência Nacional de Transportes Terrestres (ANTT): Regula o transporte rodoviário, ferroviário e aquaviário, bem como a concessão de rodovias e ferrovias (criada em 2001);
- Agência Nacional de Aviação Civil (ANAC): Regula a aviação civil, incluindo aeroportos, companhias aéreas e serviços de transporte aéreo (criada em 2005);
- Agência Nacional de Vigilância Sanitária (ANVISA): Regula a segurança e qualidade de produtos farmacêuticos, alimentos, cosméticos e produtos de saúde (criada em 1999);
- Agência Nacional de Águas (ANA): Regula o uso e a gestão dos recursos hídricos no Brasil, incluindo rios, lagos e aquíferos (criada em 2000);
- Agência Nacional de Cinema (ANCINE): Regula o setor audiovisual, incluindo produção, distribuição e exibição de filmes e conteúdo audiovisual (criada em 2001);
- Agência Nacional de Mineração (ANM): Regula a exploração mineral e a gestão de recursos minerais no país (criada em 2017);
- Agência Nacional de Saúde (ANSAU): Regula o setor de saúde, com ênfase na promoção da saúde e prevenção de doenças (criada em 2000);
- Agência Nacional de Proteção de Dados (ANPD): Regula a proteção de dados pessoais e a privacidade no ambiente digital (criada em 2018).

Todas estas agências reguladoras atuam em segmentos específicos, com autonomia técnica (prevista nas respectivas Leis de criação), buscando garantir que a regulação seja baseada em critérios técnicos e objetivos, reduzindo questões como captura regulatória e equilíbrio informacional. De maneira geral, vê-se que a criação das agências e consequentemente as discussões regulatórias são recentes no contexto brasileiro, o que confere espaço para estudos e maiores aprofundamentos conceituais e aplicações práticas.

2.2 TEORIA DA SINALIZAÇÃO

O estudo da teoria da sinalização teve origem na área da economia da informação, partindo de um cenário que prevê a interação entre agentes econômicos com informações assimétricas, sendo compradores e vendedores, em uma relação que ambos não dispõem de todas as informações, contudo, o trabalho seminal de Spence (1973) acabou por dar notabilidade a esta discussão (Boulding; Kirmani, 1993).

Stiglitz (2000), esclarece que a assimetria informacional se configura no momento em que em uma transação econômica (compra, venda ou contrato), uma das partes possui informações que a outra parte não possui, o que resulta em desequilíbrio acerca do conhecimento considerado importante para a transação. Em uma relação assimétrica, uma das partes pode explorar sua vantagem de informação, levando a decisões inadequadas ou alocação de recursos de forma ineficiente.

Desse modo, discutir sobre o aspecto informacional no mercado é compreender que a assimetria informacional existe entre os atores que produzem e que utilizam estas informações. Todavia, a teoria da sinalização se põe como uma proposta de busca por demonstrar como a assimetria pode ser reduzida a partir da sinalização de informações que emanam das empresas para o mercado.

Como base para esta teoria, Spence (1973) exemplificou o mercado de trabalho e a função de sinalização advinda da educação, no cenário em que os empregadores não possuem acesso às informações sobre a qualidade dos candidatos, contudo, os candidatos sinalizam por meio do seu grau educacional a sinalização acerca da sua qualidade. Nesse sentido, a educação é trazida como um sinal confiável de comunicação, uma vez que Spencer interpretou a educação como meio de comunicação (Connelly *et al*., 2010).

A sinalização é uma abordagem aplicável em mercados com assimetria de informações e a aplicação desta no mercado de seguros tem gerado *insights* relevantes para abordar os desafios relacionados à seleção adversa e ao risco moral, inerentes a esse setor. A pesquisa

seminal de Michael Rothschild e Joseph Stiglitz, "*Equilibrium in Competitive Insurance Markets*" (1976), abordou como seguradoras e segurados enfrentam o problema da seleção adversa. A pesquisa demonstrou que, em mercados de seguro, competitivos, as políticas de prêmios diferenciados podem servir como sinais, permitindo a segmentação dos riscos, assim, as seguradoras podem usar os prêmios mais altos como um sinal para atrair os segurados de maior risco, ao passo que os prêmios mais baixos sinalizam segurados de menor risco. Isso lança luz sobre como os agentes podem utilizar informações assimétricas para construir relações contratuais eficientes e promover o equilíbrio do mercado.

Araujo e Novinski (2009), na pesquisa "*Signaling and competition in the insurance market*", também aplicaram a teoria da sinalização ao mercado de seguros. Os autores exploraram como as seguradoras podem usar a sinalização para competir de maneira estratégica. A pesquisa se direcionou a observar como as ações e estratégias adotadas pelas seguradoras, que são observáveis pelos consumidores, podem funcionar como sinais de sua qualidade, transparência e confiabilidade. Desse modo, sinais como “processos de gestão de riscos transparentes” e “cumprimento rigoroso das normas regulatórias”, podem criar diferenciação e atrair segurados conscientes da qualidade.

Os autores apontaram que as seguradoras que oferecem sinais críveis de qualidade, têm maior probabilidade de atrair um *pool* de segurados mais favorável, reduzindo a probabilidade de sinistros graves e promovendo uma melhor alocação de riscos.

Ainda de maneira complementar, a pesquisa de Araujo e Novinski (2009), destaca que a sinalização não é apenas uma ferramenta de mercado, utilizada estrategicamente, mas também pode ter implicações regulatórias. A adoção de políticas que promovem a transparência e a sinalização pode ser benéfica para o setor de seguros, criando um ambiente onde as seguradoras competem não apenas com base no preço, mas também com base na qualidade dos serviços oferecidos.

Desse modo, concluem enfatizando a capacidade das seguradoras de usar a sinalização como estratégia competitiva para superar a assimetria de informações, atrair segurados preocupados com qualidade e eficiência, e contribuir para um mercado de seguros mais estável e eficiente. Isso destaca a relevância contínua da teoria da sinalização na análise do funcionamento e das dinâmicas dos mercados de seguros.

A aplicação da teoria da sinalização no mercado de saúde suplementar brasileiro envolve o uso de sinais que podem ajudar a reduzir a assimetria de informação entre as

operadoras de planos de saúde, prestadores de serviço e os beneficiários (consumidores dos planos de saúde).

Para tanto, a contabilidade das OPS serve como fonte de produção de informações requeridas pelo órgão regulador, para monitoramento da situação econômica e financeira das OPS, reportando trimestralmente o balancete contábil e a demonstração dos fluxos de caixa e, anualmente, o Relatório Circunstanciado da Auditoria Independente (para as OPS de médio e grande porte). A partir deste fornecimento, a ANS, seguindo padrões e métricas instituídas em seus normativos, elabora e divulga indicadores que buscam informar à sociedade, dentre outros aspectos, sobre o desempenho, assistência e situação econômica e financeira das OPS.

Estes sinais, representados pelos indicadores divulgados pela ANS, visam fornecer transparência, monitoramento do desempenho das operadoras e a possibilidade de auxílio aos consumidores no momento de decisão de alocação de recursos em planos de saúde. Os principais indicadores, são:

- **Índice de Desempenho da Saúde Suplementar (IDSS)**: indicador composto que avalia o desempenho das operadoras em quatro áreas-chave: Qualidade, Garantia de Acesso, Sustentabilidade no Mercado e Gestão de Processos de Saúde;
- **Taxas de Reclamações e Negativas de Cobertura**: informações sobre as reclamações recebidas de consumidores e as taxas de negativas de cobertura pelas operadoras de planos de saúde;
- **Índice de Satisfação do Beneficiário**: Avaliação do nível de satisfação dos beneficiários com seus planos de saúde;
- **Taxas de Reajuste de Mensalidades**: Publicação das taxas de reajuste autorizadas para os planos de saúde, permitindo que os consumidores saibam quanto suas mensalidades podem aumentar;
- **Qualificação de Prestadores**: Informações sobre hospitais e estabelecimentos qualificados pelas operadoras, incluindo informações sobre especialidades e localização;
- **Índice de Operadoras**: Classificação das operadoras com base em critérios como desempenho financeiro e qualidade dos serviços prestados;
- **Índice de Sinistralidade**: Índice de despesas assistenciais, ou despesas médicas, ou ainda, sinistralidade. Mostra a relação entre as despesas assistenciais e o total

das receitas com operação de planos de saúde da operadora, acrescido do valor absoluto das contraprestações de corresponsabilidade cedida;

- **Liquidez corrente**: Mostra a relação entre os ativos conversíveis em dinheiro no curto prazo e as dívidas de curto prazo;
- **Margem de Lucro Líquido**: Mostra a relação entre o resultado líquido e o total das receitas com operação de planos de saúde (contraprestações efetivas);
- **Prazo médio de pagamento de eventos**: Indica o tempo médio que a operadora leva para pagar os eventos assistenciais;
- **Prazo médio de recebimento**: Indica o tempo médio que a operadora leva para receber os eventos assistenciais;
- **Índice combinado**: Mostra a relação entre as despesas operacionais (administrativas, comercialização e assistenciais) e as receitas (contraprestações efetivas e receitas administrativas).

A obrigatoriedade de envio desses dados é prevista nas Resoluções Normativas da ANS e se aplica a todas as OPS registradas e ativas no cadastro da ANS, independente do porte e modalidade. Desse modo, a regulação serve à sociedade, produzindo sinalização para a redução da assimetria informacional existente no segmento da saúde suplementar, permitindo maior poder de escolha na alocação do capital e reduzindo os riscos contratuais.

2.3 RISCO DE SOLVÊNCIA NA SAÚDE SUPLEMENTAR

Se constituindo sob a égide de um segmento permeado por relações conflituosas, a saúde suplementar também é traspassada por riscos inerentes à própria atividade/negócio, principalmente no que se refere a situações em que o beneficiário utilizará ilimitadamente os serviços que as OPS comercializam. Essas situações se manifestam considerando a livre vontade do beneficiário em contratar serviços de assistência à saúde privada, gerando a obrigação, para este, de desembolso financeiro contínuo em favor das OPS.

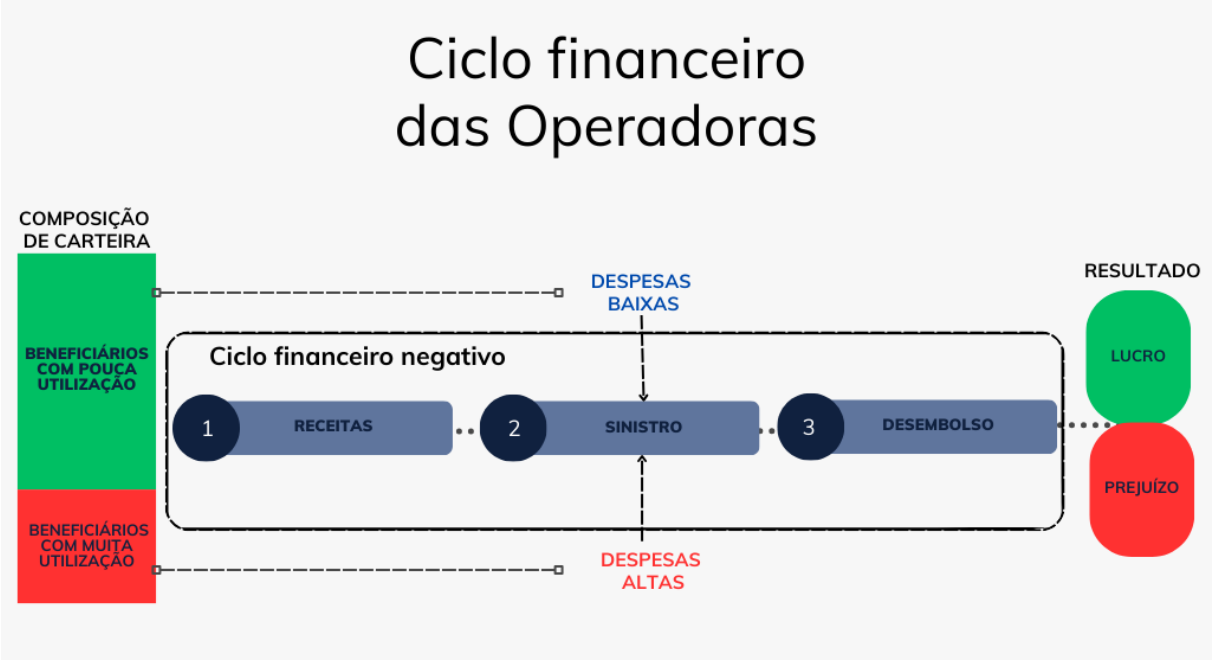
Nesse sentido, como mediação destes conflitos, a Lei nº 9.961, de 28 de janeiro de 2000 criou a ANS, designando-a a função regulatória total desse mercado. No exercício de suas funções, a ANS, por meio da edição da Resolução Normativa Nº 209, de 22 de dezembro de 2009, estabeleceu os critérios de manutenção de Recursos Próprios Mínimos e constituição de Provisões Técnicas a serem observados pelas operadoras de planos privados de assistência à

saúde. Ambos os critérios de manutenção se direcionam à manutenção da garantia da prestação dos serviços de saúde contratados pelos beneficiários.

Neste ponto, o esforço da regulação se direciona para a busca da correta gestão de ativos para a cobertura de passivos inerentes tanto na manutenção das OPS, como a cobertura dos sinistros. Destaca-se que, em um período anterior à regulação, aproximadamente de 1960 ao ano 2000, os incentivos negativos do ciclo financeiro invertido geravam descapitalização setorial, o que implica em não observância dos conceitos de liquidez e solvência, exceto para as OPS cuja modalidade de seguradoras já possuía regulação específica.

O conceito de ciclo financeiro no mercado de seguros e planos de saúde, especialmente no modelo de contratos a preço preestabelecido, se dá na medida em que os serviços contratados são entregues após o pagamento pelos beneficiários (Plantin; Rochet, 2007), conforme a Figura 3 retrata.

Figura 3 - Ciclo financeiro negativo das Operadoras de Planos de Saúde



Fonte: Elaborado e adaptado pela autora (2023).

Levi, Lee e Gibson (2009), explicam que "o *negative cash flow cycle*" ocorre quando as operadoras recebem os prêmios dos segurados antecipadamente, mas precisam arcar com os custos dos serviços médicos prestados aos beneficiários posteriormente. Essa lacuna temporal entre recebimentos e desembolsos pode promover um ambiente de pressões financeiras

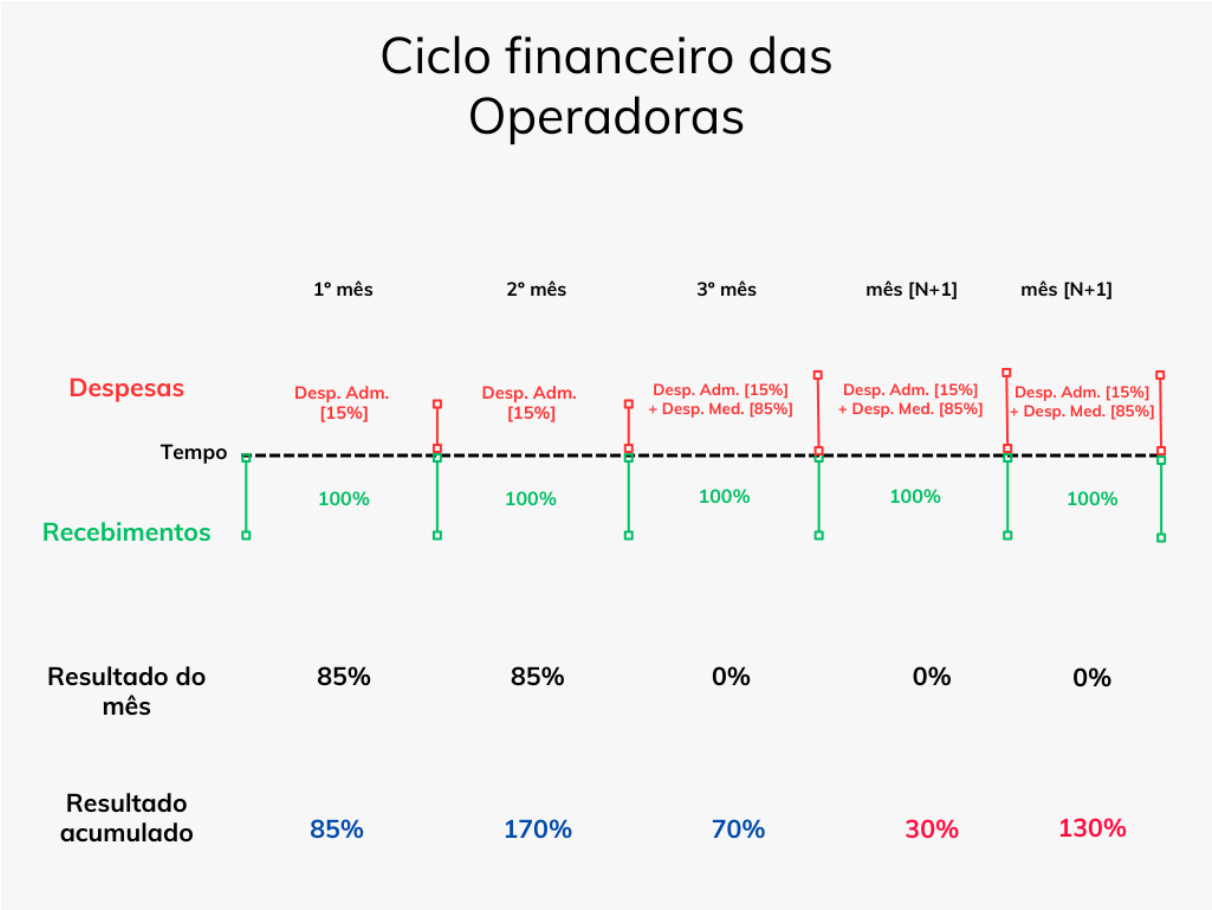
significativas, impactando a liquidez e a capacidade das operadoras de cumprir suas obrigações, ou seja, afetar a solvência.

Langabeer e Helton (2018, p. 32), reforçam que:

> "a gestão inadequada do ciclo financeiro pode levar a pressões significativas sobre o capital das operadoras, à medida que as receitas de mensalidades são recebidas antecipadamente, enquanto os custos decorrentes da prestação dos serviços de assistência à saúde incorrem de forma contínua e não obedecem ao ritmo das contraprestações recebidas".

Dessa maneira, o ciclo financeiro negativo das OPS é marcado pelo risco do sinistro, ou seja, o risco de utilização do plano de saúde, o que implica, para a operadora, a necessidade de utilização da contraprestação (receita) desembolsada pelo beneficiário em seu ciclo de contratação, sendo este risco o fator definidor do equilíbrio financeiro de qualquer OPS. De forma detalhada, a Figura 4, apresenta um exemplo hipotético, que expressa o descompasso presente no ciclo financeiro das OPS, revelando a necessidade da gestão de riscos do sinistro e, consequentemente, do ciclo financeiro.

Figura 4 - Ciclo financeiro das Operadoras de Planos de Saúde



Fonte: Elaborado e adaptado pela autora (2023).

A figura 4, formatada com o objetivo de clarificar um ciclo financeiro hipotético de uma OPS, esclarece que o ciclo financeiro favorável à OPS se encontra no mês 1 e mês 2, contudo, em uma simulação da ocorrência de sinistro nos meses seguintes, o resultado acumulado demonstra como o equilíbrio financeiro da OPS pode ser afetado pelo ciclo financeiro negativo, favorecendo então um cenário de insolvência.

Com esse olhar específico para o risco de solvência das OPS, após a criação da Lei dos Planos de Saúde e posteriormente da ANS, a regulação estabeleceu critérios de manutenção das OPS, tornando o critério de solvência amplamente almejado pelas OPS e conhecido na saúde suplementar. Mesmo em um contexto onde não há uma definição única sobre solvência, Silva *et al.* (2016) apresentam a insolvência como o não cumprimento das obrigações empresariais com suas dívidas.

Para Chung, Tan e Holdsworth (2008), o estado de insolvência caracteriza o momento em que a companhia sinaliza não conseguir cumprir com suas obrigações (passivos) até o prazo de vencimento, ou até mesmo o momento em que o passivo desta companhia passa a exceder o ativo total.

Sob uma perspectiva semelhante, Altman e Hotchkiss (2010) caracterizam a insolvência como uma condição temporária, em que se identifica a falta de liquidez ou a não capacidade de cumprir com as obrigações. Todavia, este estado, ainda que temporário, pode sinalizar indício de falência, se configurando então em fator crítico e não apenas temporário.

Especificamente na saúde suplementar, a regulação estabeleceu critérios de manutenção das operações das OPS, e de forma mais específica, a RN nº 209/2009 define a margem de solvência correspondente à suficiência do Patrimônio Líquido ajustado por efeitos econômicos para cobrir o maior montante entre os seguintes valores:

> I - 0,20 (zero vírgula vinte) vezes a soma dos últimos doze meses: de 100% (cem por cento) das contraprestações/prêmios na modalidade de preço preestabelecido, e de 50% (cinquenta por cento) das contraprestações/prêmios na modalidade de preço pós-estabelecido; ou
> II – 0,33 (zero vírgula trinta e três) vezes a média anual dos últimos trinta e seis meses da soma de: 100% (cem por cento) dos eventos/sinistros na modalidade de preço preestabelecido e de 50% (cinquenta por cento) dos eventos/sinistros na modalidade de preço pós-estabelecido.

Sobre a solvência das OPS como indicador crítico no segmento da saúde suplementar, Guimarães e Alves (2009) apontam que o estado oposto à solvência é a insolvência e esta última ocorre quando o patrimônio líquido da operadora de plano de saúde é igual ou inferior a zero, fato que vem de forma recorrente sendo observado na saúde suplementar do Brasil.

Além deste aspecto da solvência, no exercício de suas funções, a ANS, por meio da edição da Resolução Normativa Nº 205, de 09 de outubro de 2009, estabeleceu exigências informacionais às OPS, atribuindo a estas a obrigação de fornecê-las e divulgar aos seus beneficiários (e ao mercado de forma geral): informações assistenciais; informações acerca da criação de reservas e garantias; informações sobre o índice de sinistralidade; divulgação das demonstrações contábeis, ao final do exercício; além do estabelecimento da previsão de auditorias independentes nas OPS de grande porte (as que possuem mais de 100.000 mil beneficiários).

Dentre estas exigências, a ANS realiza o acompanhamento do índice de sinistralidade, definido por Araújo e Silva (2018) como a relação entre as despesas e a receita das operadoras de saúde suplementar, medida em percentual. Esse conceito está estritamente relacionado à representação entre as despesas assistenciais e o total das receitas com operação de planos de saúde (contraprestações efetivas), ou seja, a medida de risco das OPS.

Quanto ao parâmetro da sinistralidade ideal para as OPS, Araújo e Silva (2018) assegura que a média tida como aceitável pela maior parte das empresas do segmento é de 75%, portanto, é por meio desta medida que se torna possível acompanhar os limites de custeio dos serviços assistenciais e as despesas do negócio. Observa-se então que, quanto maior for a sinistralidade, menor será a margem de lucratividade das OPS, haja vista se tratar de duas grandezas inversamente proporcionais.

O Estudo da Capitolio sobre sinistralidade na saúde suplementar, com a janela temporal de 2004 a 2018, demonstra que em apenas dois anos (2004 e 2005), a média do índice de sinistralidade permaneceu abaixo o limite aceitável de 75% (Capitolio, 2019). No entanto, em que pese exceder este limite aceitável, a sinistralidade no ano de 2022 se apresentou em 85,8%, o que implica imputar às OPS 14,2% das receitas assistenciais para arcar com todas as outras obrigações não assistenciais (ANS, 2022). Isto, consequentemente, impacta os resultados financeiros, trazendo riscos de insolvência a médio e longo prazo.

Os fatores que podem exercer impacto sobre a sinistralidade estão relacionados ao cuidado preventivo e contínuo com a saúde do beneficiário, doenças e lesões pré-existentes à contratação do plano, aumento do custo tecnológico com a assistência à saúde, ampliação do rol de procedimentos e serviços oferecidos pelas OPS, além do risco inerente ao negócio da saúde suplementar, que inclui a imprevisibilidade dos gastos com a prestação do serviço.

Ancorada na discussão acerca da sinistralidade e a sua relação com a margem de solvência, esta tese busca trazer a contribuição em torno de como a tecnologia e as modelagens

preditivas podem produzir informações para os usuários internos, auxiliando na tomada de decisão acerca de fatores que impactam a solvência das OPS. Esta discussão continua no tópico abaixo que versa sobre o aparato tecnológico que pode potencializar a produção informacional preditiva.

2.4 INTELIGÊNCIA ARTIFICIAL E APRENDIZAGEM DE MÁQUINA

A inteligência artificial (IA), como disciplina, iniciada pelo cientista John McCarthy e formalmente divulgada na Conferência de Dartmouth (*Dartmouth Summer Research Project on Artificial Intelligence*) em um workshop de verão de 1956 que é amplamente considerado o evento fundador da inteligência artificial como um campo.

Makridakis (2017) afirma que a IA está presente nas atividades diárias das pessoas e das empresas, ao citar os recursos de reconhecimento de voz, de face, de digital e ainda as funções de sugestões de escrita dos smartphones.

Pan (2016) cita o sistema Watson, desenvolvido pela empresa IBM, utilizado operacionalmente nos hospitais, detectando alternativas de diagnóstico de câncer, buscando informações dentre milhões de registros de pacientes. Schwab (2016) acrescenta que a IA tem facilidade em realizar correspondências de padrões e automatizar processos, o que faz a adoção da tecnologia ser recomendável às organizações.

Adicionalmente, Igarashi *et al*. (2008) demonstraram a aplicabilidade de ferramentas de IA ao contexto científico, sendo possível observar a pluralidade no que se refere aos diversos campos de pesquisa. Os autores evidenciaram a utilização da IA para:

- Linguagem e sistema para o desenvolvimento de aplicações – na área de ciência da computação;
- Fornecer previsões climáticas – na área meteorológica;
- Modelos de processos de desenvolvimento de software para sistemas de agentes – na área de engenharia de *software*;
- Limite de oscilação cíclica em uma aeronave – na área de aviação, dentre outros.

Com tamanha amplitude e potencial escalável de aplicabilidade a muitos segmentos da sociedade, a IA é definida por Dolgui *et al*. (2018), como sendo a capacidade de uma máquina ou equipamento pensar, aprender e agir como seres humanos. Mishra *et al*. (2019), destacam que a inteligência artificial inclui todas as máquinas ou equipamentos que utilizam capacidades computacionais para trabalhar e funcionar como humanos ou para substituir humanos. Desse

modo, vê-se que a performance da figura humana é a principal fonte de inspiração para o desenvolvimento da inteligência artificial.

Russell e Norvig (2010) explicam que a IA visa criar sistemas que possam perceber o ambiente, aprender com dados, compilar informações e tomar decisões de forma autônoma. Nesse sentido, uma das áreas fundamentais da IA é o aprendizado de máquina ou *machine learning*, que envolve o desenvolvimento de algoritmos que permitem às máquinas aprender com dados e melhorar seu desempenho ao longo do tempo, ou seja, o aprendizado a partir de dados de mundo real e o resumo em uma descrição matemática assertiva para a tomada de decisões.

Como definição, Khalid *et al.* (2021) trazem que o aprendizado de máquina envolve o desenvolvimento de modelos, principalmente modelos estatísticos que podem ser construídos e fornecer resultados preditivos. Nesse sentido, na perspectiva destes autores, o aprendizado de máquina é basicamente o subcampo da inteligência artificial que envolve a detecção automática de padrões em dados, sendo originalmente descrito como um programa que aprende a realizar uma tarefa automaticamente a partir da experiência (ou seja, dados), em vez de ser programado explicitamente.

Logo, Khalid *et al.* (2021) defendem que os algoritmos de aprendizado de máquina têm a capacidade de evoluir por meio do treinamento e podem processar grandes volumes de dados e extrair informações significativas usando uma variedade de técnicas de programação. Dessa forma, eles podem aprender com os dados fornecidos e melhorar as interações anteriores.

O aprendizado de máquina é amplamente dividido em três categorias:

(1) Aprendizagem Supervisionada: Para treinamento de um modelo, é utilizado um conjunto de dados com características (variáveis) e rótulos (resultado ou classe de interesse). Assim, é gerada uma função que mapeia recursos para rótulos e então a utiliza para prever os rótulos de dados não rotulados.

    Os algoritmos de aprendizagem supervisionada mais importantes são *random forest* (floresta aleatória), *decision tree* (árvore de decisão), *K-nearest neighbords support* (K-vizinhos mais próximos), *support vecto machine* (máquina de vetores de suporte linear), *support vecto machine* (máquina de vetores de suporte RBF), Bayes ingênuo, regressão polinomial e redes neurais artificiais.

(2) Aprendizagem Não Supervisionada: Neste tipo de abordagem, os dados de treinamento não são rotulados, logo o sistema tenta aprender “sem a ajuda de um

professor". A aprendizagem não supervisionada revela padrões ocultos em dados não rotulados e formula conclusões a partir deles.

(3) Aprendizagem por Reforço: Combina aprendizagem supervisionada e não supervisionada. O algoritmo na aprendizagem por reforço maximiza a precisão por meio de tentativa e erro.

De maneira complementar, Dhamija e Bag (2020) argumentam que a observação precisa do mundo real são dados e os programas de computador aprendem sobre o meio ambiente a partir dos dados para gerar resultados inteligentes. Nessa perspectiva, várias ferramentas são usadas em inteligência artificial, desde ferramentas simples de modelagem de regressão até ferramentas avançadas, como aprendizado profundo (visão computacional avançada).

Destaca-se então que a capacidade de aprendizado, citada por Dhamija e Bag (2020) é essencial para muitas aplicações práticas da IA, tais como reconhecimento de voz, processamento de linguagem natural e condução autônoma em diversas indústrias, desde assistentes virtuais em dispositivos móveis até diagnósticos médicos avançados e automação industrial.

Na perspectiva de Kaufman e Santaella (2020), os sistemas de aprendizado de máquina, apesar da evolução, ainda carecem da capacidade de compreender o significado, não possuindo senso intuitivo ou até mesmo a capacidade de formar conceitos abstratos. Na visão das autoras, essa lacuna impacta na fragilidade de produzir analogias e generalizações, isto por não existir a capacidade de compreender o funcionamento do mundo a partir da observação ou até mesmo do ambiente.

Ainda é preciso destacar que os avanços da IA trouxeram benefícios únicos, originais e carregados de ineditismo para a humanidade, contudo, não se pode deixar de citar os desafios éticos e regulatórios, permeados de aspectos complexos. Como exemplo destas complexidades, Kaufman e Santaella (2020) citam o sistema Watson da IBM, que apresenta 90% de taxa de sucesso se comparado aos 50% dos médicos (humanos) nos diagnósticos de câncer de pulmão, sendo capaz de processar grandes volumes de dados, estabelecendo correlações entre sintomas e/ou imagens em uma dimensão de tamanha capacidade e velocidade, que atualmente se torna impossível de ser alcançada por um ser humano.

Arora *et al.* (2019) denominam este grande volume de dados de *big data* e, enfatizam que na área da saúde, estes dados surgiram como consequência da transformação de grandes volumes de prontuários para o formato eletrônico, despertando então o interesse em produzir previsões de risco por meio de algoritmos de aprendizagem de máquina.

Outro exemplo que demonstra as possibilidades a partir do uso de *big data*, é o Medline (banco de dados) da Biblioteca Nacional de Medicina dos EUA, que tem capacidade para indexar mais de 5.600 periódicos e milhões de registros médicos, históricos de saúde de pacientes e estudos de caso que tem o potencial de oferecer importantes insights com o uso de algoritmos de IA. O volume total de dados pode dobrar a cada cinco anos, conforme estimativas baseadas nos últimos cinco anos.

Santos *et al.* (2020) utilizaram a IA para desenvolver e comparar o desempenho preditivo de algoritmos de aprendizado de máquina para estimar a sobrevivência e qualidade de vida em pacientes críticos com diagnóstico de câncer. Os autores trabalharam com seis algoritmos de aprendizado de máquina e aplicaram em uma base de 777 pacientes internados em Unidade de Terapia Intensiva de dois hospitais públicos brasileiros, especializados em tratamento de câncer. O estudo apontou como conclusão que os algoritmos preditores discriminaram bem o risco de vida dos pacientes oncológicos, contribuindo para informar pacientes e médicos sobre o prognóstico (do paciente) e ajudar a tomar melhores decisões em relação às intervenções, com vista à maior permanência e longevidade de vida.

Kaufman (2020) defende que a contribuição da IA na área de saúde transcende até mesmo apenas o ato de diagnosticar doenças, mas ainda pode abranger de forma significativa a prevenção de epidemias e potenciais anomalias individuais. Dessa forma, vê-se que a aplicação da IA na área da saúde, pode configurar uma ruptura em como os Governos e instituições de saúde (públicas e privadas) estruturam e concebem políticas, ações e processos de prevenção para a população.

Para além da área de saúde, Streit e Borenstein (2009) analisam a regulamentação do setor financeiro combinando a lógica nebulosa e um modelo baseado em agentes, em que o comportamento destes foi simulado por meio de regras *fuzzy*, com base na análise do conteúdo de jornais e entrevistas com especialistas.

Para traduzir os complexos regulamentos do segmento agrícola em conjuntos de regras processáveis por máquinas executadas por módulos especializados de sistemas de informação de gestão agrícola (FMIS), Espejo-Garcia *et al*. (2018) utilizaram a inteligência artificial para formatar uma ferramenta que transforma regulamentações textuais de pesticidas em regras processáveis por máquinas, reduzindo a complexidade, o tempo despendido para a tradução manual destas regras, vinculando e otimizando a intervenção humana.

Nesse sentido, a utilização da IA têm sido cada vez mais comum e consequentemente mais abrangente em sua aplicação a diversos segmentos, o que na visão de Khalid *et al.* (2021)

pode guardar relação com o acesso ao desenvolvimento da IA, que inicialmente estava limitado às grandes corporações, dado os altos investimentos necessários, contudo, a pulverização das linguagens de programação tem possibilitado o acesso mais amplo e democrático aos benefícios da IA.

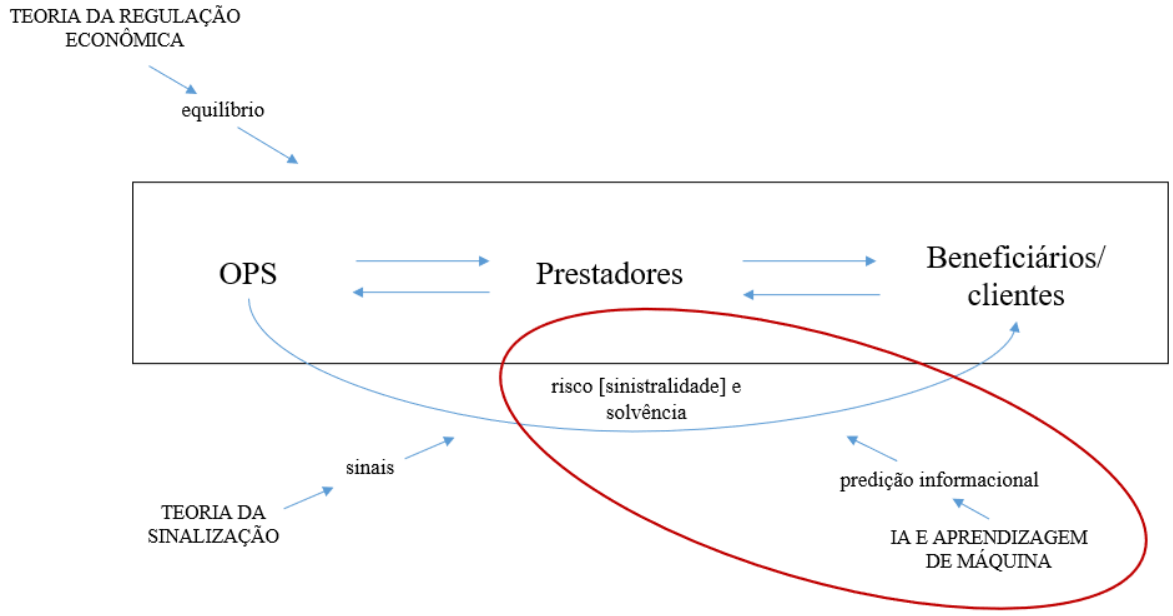
Desse modo, ao considerar a relevância do uso da IA e *machine learning* no contexto de mercado das organizações, esta tese propõe a aplicação dessa tecnologia para predições informacionais na saúde suplementar.

2.5. HIPÓTESES DE PESQUISA

Com a pretensão de verificar o impacto da inteligência artificial na predição informacional das despesas assistenciais na saúde suplementar brasileira e sua relação com o risco de solvência das Operadoras de Planos de saúde, foram formuladas duas hipóteses de pesquisa. Estas hipóteses nasceram da discussão teórica demonstrada na figura 5, em que a Teoria da Regulação Econômica explica a busca pelo equilíbrio econômico, informacional e financeiro introduzido pela regulação econômica no segmento da saúde suplementar, por meio da lei nº 9656 e posterior lei nº 9961 e seus desdobramentos.

De maneira complementar, a teoria da sinalização embasa a discussão da emissão de sinais para redução da assimetria informacional, podendo otimizar a decisão de alocação de recursos, tanto dos gestores das OPS quanto dos beneficiários ao escolher a OPS a ser contratada. Nessa perspectiva, a sinistralidade é o sinal que indica sobre a solvência da OPS, tanto ao órgão regulador, ANS, quanto à sociedade. Por tamanha relevância, a IA se apresenta como uma alternativa para potencializar a produção de informação preditiva acerca da sinistralidade das OPS.

Figura 5 – Relação entre a estrutura teórica e as hipóteses de pesquisa



Fonte: Elaborado e adaptado pela autora (2023).

2.5.1 Inteligência artificial, predição informacional e risco de solvência

A inteligência artificial tem se tornado uma importante ferramenta de aplicação no segmento da saúde, principalmente no que se refere a diagnóstico e prognóstico, melhorando o processo de tomada de decisão clínica entre profissionais da área. Algumas pesquisas científicas direcionadas a diagnóstico e previsibilidade de desenvolvimento de doenças estão sendo conduzidas de maneira experimental. Dado o cenário crescente de casos de câncer no Brasil e no mundo, por exemplo, é cada vez mais importante melhorar as decisões prognósticas para pacientes com neoplasia, por isso, GSF Silva *et al.* (2023), na pesquisa "Aprendizado de máquina para previsão longitudinal de risco de mortalidade em pacientes com neoplasia maligna em São Paulo" treinaram algoritmos para predição do risco de morte de pacientes com neoplasia, a fim de fornecer subsídios para seu manejo clínico.

Para predizer sobre o risco de morte por COVID-19, Wichmann *et al.* (2023) treinaram e avaliaram o desempenho de três algoritmos, com dados de pacientes de 18 hospitais localizados em todas as cinco regiões do Brasil, produzindo resultados que contribuíram para a replicação desta pesquisa e traçando possíveis horizontes de perfis de pacientes mais propensos a complicações decorrentes da COVID-19.

Observa-se então que a incorporação da inteligência artificial (IA) no âmbito da saúde não apenas aprimora a precisão diagnóstica, mas também desempenha um papel fundamental na personalização dos tratamentos, sugerindo terapias mais eficazes e personalizadas, contribuindo assim para uma abordagem mais direcionada e eficiente no combate ao câncer.

A partir do exposto, no campo de aplicação da saúde suplementar, espera-se que a inteligência artificial demonstre potencial de predição informacional das despesas assistenciais das operadoras de planos de saúde brasileiras, identificando padrões e tendências, permitindo que as operadoras ajustem suas estratégias e tomem decisões mais informadas.

H1: A inteligência artificial possui capacidade de predição informacional das despesas assistenciais das operadoras de planos de saúde brasileiras.

Considerando o potencial de predição informacional trazido pela inteligência artificial, acrescenta-se que, esta predição precisa e oportuna, oferecerá às operadoras maior capacidade de antecipar variações nas despesas assistenciais. Isso pode permitir uma alocação mais eficiente de recursos, a implementação de medidas preventivas e uma resposta ágil a mudanças nas condições de mercado, impactando positivamente a solvência das operadoras.

H2: A predição informacional das despesas assistenciais impacta positivamente o risco de solvência das operadoras de planos de saúde brasileiras.

3. ESTRATÉGIA METODOLÓGICA

3.1 DELINEAMENTO DA PESQUISA

Quanto à área de estudo da ciência, a pesquisa é aplicada porque apresenta como motivação a necessidade de produzir conhecimento para a aplicação de seus resultados no mercado brasileiro de saúde suplementar. Barros e Lehfeld (2000, p.78) salientam que a pesquisa aplicada tem como objetivo “contribuir para fins práticos, visando à solução mais ou menos imediata do problema encontrado na realidade”.

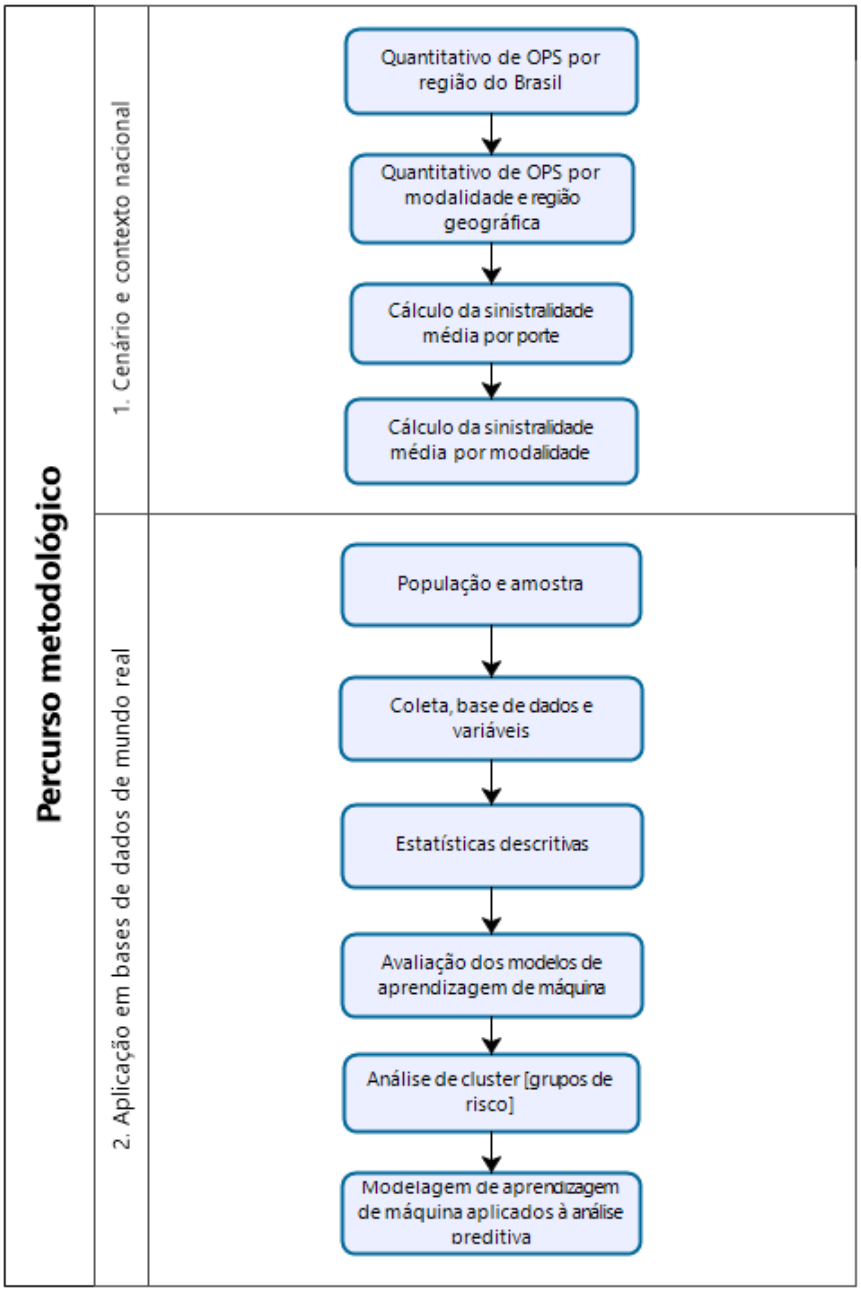
Quanto aos objetivos, se caracteriza como exploratória e descritiva, isto porque busca explorar aspectos que envolvem a utilização da inteligência artificial na predição informacional e sua relação com o risco de solvência na saúde suplementar brasileira.

A natureza da pesquisa é quantitativa se centrando na objetividade, considerando a necessidade de dados brutos para a compreensão da realidade (Fonseca, 2002). Acerca dos procedimentos para o alcance dos objetivos propostos, foram realizadas modelagens preditivas e estimações em um conjunto de variáveis que buscaram identificar padrões para predição informacional acerca das despesas assistenciais na saúde suplementar brasileira.

A primeira fase do trabalho se constitui em apresentar um cenário geral das Operadoras de planos de saúde brasileiras, agrupadas por região, modalidade, bem como calcular a sinistralidade média por porte e modalidade, uma vez que é a medida de risco de insolvência adotada como parâmetro de mercado. A segunda fase se constitui a partir da coleta das variáveis (contábeis, assistenciais e demográficas) para estimação de modelagens preditivas e algoritmos de aprendizagem de máquina que sinalizaram acerca do poder de predição informacional das despesas assistenciais das operadoras de planos de saúde.

Estas fases são sequenciais, ou seja, primeiro é realizada a fase 1 e posteriormente a fase 2, a figura 6 detalha as atividades realizadas em cada fase.

Figura 6 – Percurso metodológico



Fonte: Elaboração própria, 2023.

3.2 DADOS E MÉTODOS

3.2.1 Cenário e contexto nacional

O cenário nacional compreende 1.127 operadoras de planos de saúde médico e odontológico, registradas na base da ANS, em exercício e com beneficiários ativos no exercício de 2023. Contudo, para melhor retratar o contexto desta pesquisa, foram excluídas as modalidades de OPS: exclusivamente odontológicas (249 OPS), administradoras de benefícios

(176 OPS), filantrópicas (32 OPS) e seguradoras (8 OPS), totalizando 465 OPS excluídas. Além destas exclusões, 85 OPS foram retiradas por apresentarem dados incompletos de receitas e despesas, totalizando a amostra final em 577 OPS, conforme demonstra a tabela 1, OPS por região do Brasil.

Tabela 1 – Quantitativo de OPS por região do Brasil

Região	Quantitativo de OPS
Norte	19
Nordeste	59
Centro-oeste	49
Sudeste	342
Sul	108
Total	**577**

Fonte: ANS, 2023.

As 249 OPS exclusivamente odontológicas foram excluídas deste cenário, em virtude do baixo risco de insolvência apresentado por este segmento, principalmente ao se considerar que estas possuem historicamente o índice médio de sinistro de 43,1% (ANS, 2023), característica que difere do objeto de pesquisa desta tese, uma vez que a sinistralidade não representa o mesmo risco para OPS odontológicas e OPS de saúde médica.

Considerando aspectos peculiares e ainda a previsão de regulação [regras e normativos] específica, que diferem das demais modalidades, as administradoras de benefícios, filantropia e seguradoras foram excluídas deste cenário inicial. As administradoras de benefícios possuem um modelo de negócio de intermediação entre empresas contratantes e as OPS, por este motivo, a regulação prudencial quanto aos aspectos econômicos e financeiros não são aplicáveis, o que implica na desnecessidade de constituição de reservas e provisões que façam frente aos riscos da operação, uma vez que estes riscos são assumidos pelas OPS e não pelas administradoras.

As OPS, cuja modalidade é a filantropia, se diferenciam das demais modalidades por se beneficiarem de imunidade tributária, isenção de imposto de renda (IR), contribuição social (CSLL) e contribuições previdenciárias e ainda redução de alíquotas de impostos indiretos, o que torna impraticável o parâmetro da comparabilidade entre essa modalidade de OPS e as demais. As OPS seguradoras, pela atuação no mercado regulado pela SUSEP (Superintendência de Seguros Privados) possuem modelo de operação diferente das demais modalidades de OPS, uma vez que o consumidor do seguro tem um contrato (apólice) que define as condições e os limites de reembolso, bem como a cobertura e a abrangência geográfica dos serviços, assim, as seguradoras não estão expostas ao risco como as demais OPS.

Desse modo, a tabela 2 representa o quantitativo final das OPS que compuseram o cenário nacional aplicável a esta tese, demonstrando a modalidade da OPS e a respectiva quantidade existente por região geográfica do Brasil.

Tabela 2 – Quantitativo de OPS por modalidade e região geográfica

Modalidade	Quantitativo de OPS	REGIÃO GEOGRÁFICA				
		Norte	Nordeste	Centro Oeste	Sudeste	Sul
Autogestão	107	3	15	18	51	20
Cooperativa médica	262	11	25	22	146	58
Medicina de grupo	208	5	19	9	145	30
Total	**577**	**19**	**59**	**49**	**342**	**108**

Fonte: ANS, 2023.

Observa-se então que aproximadamente 59% do total das OPS se concentram na região sudeste do Brasil e, o Estado de São Paulo possui aproximadamente 54% (dos 59%) do total das OPS da região sudeste, sugerindo uma possível relação com a alta faixa populacional do Estado.

3.2.2 Coleta, base de dados e variáveis

3.2.2.1 – Fase 1: Cálculo da Sinistralidade

A coleta dos dados contábeis das OPS se deu exclusivamente por meio do acesso público a base de dados da Agência Nacional de Saúde Suplementar (ANS), disponível em seu sítio eletrônico. O período para coleta e posterior cálculo da sinistralidade compreendeu os trimestres dos anos de 2018 a 2023, considerando a disponibilidade dos dados (base da ANS).

Importante destacar que, por meio da Lei Nº 9.656 de 1998, há a obrigatoriedade que os dados contábeis disponibilizados pelas OPS de grande porte (acima de 100 mil beneficiários), sejam auditados por auditoria independente, contribuindo então para maior confiabilidade nos dados fornecidos.

Para o cálculo da sinistralidade as variáveis contábeis coletadas foram:

Tabela 3 - Sumário das variáveis para cálculo da sinistralidade

CATEGORIA	VARIÁVEL	UNIDADE TEMPORAL
DESPESAS/RECEITA	Despesa_Assistencial (sinistro)	Trimestre
	Receita_Operadora (prêmio)	Trimestre

Fonte: Elaboração própria, 2023.

Com as variáveis coletadas, o cálculo do índice de sinistralidade se deu a partir da Equação 01:

$$\text{Sinistralidade (k)} = [\text{Despesa assistencial (sinistro)}/\text{Receita (prêmio)}]x100 \quad (1)$$

Nesta etapa, a sinistralidade será observada a partir de agrupamentos por graus de semelhança entre as OPS, considerando o "porte", "região" e "modalidade", além da média geral por trimestre, para melhor indicar a realidade dos grupos de OPS.

3.2.2.2 – Fase 2: Testes das modelagens preditivas e algoritmos de aprendizagem de máquina (*machine learning*)

a) Aplicações em base de dados de mundo real - Operadoras de Planos de Saúde Brasileiras: Alfa, Beta e Gama.

Nesta fase 2, denominada "Testes das modelagens preditivas e algoritmos de aprendizagem", a população foi selecionada considerando o critério de índice de sinistralidade, entre o intervalo a partir da média nacional de 2023, qual seja, 88,20% e abaixo de 100%, pela compreensão de que as OPS que apresentam índices de sinistralidade acima de 100% possivelmente estão enfrentando algumas medidas tomadas pela ANS, como por exemplo: notificações sobre o indicador da sinistralidade; exigência de plano de ação e trabalho para reverter a situação (incluindo medidas para reduzir despesas, aumentar a eficiência operacional ou ajustar as mensalidades) e intervenção regulatória (por exemplo: proibição de venda de novos planos ou a alienação compulsória da carteira de beneficiários).

Desse modo, com esse parâmetro de intervalo do índice de sinistralidade, a população foi composta por 50 OPS, distribuídas conforme Tabela 4:

Tabela 4 – População da pesquisa [fase 2]

Porte	Região	Modalidade	Sinistralidade média
Pequeno [27]	Norte [1] Nordeste [3] Centro Oeste [2] Sudeste [17]	Autogestão [14] Cooperativa [5]	92,76%

	Sul [4]	Medicina de Grupo [8]	Intervalo [88,95% - 97,60%]
Médio [16]	Norte [-] Nordeste [2] Centro Oeste [1] Sudeste [8] Sul [5]	Autogestão [7] Cooperativa [6] Medicina de Grupo [3]	92,16% Intervalo [88,41% - 97,70%]
Grande [7]	Norte [-] Nordeste [3] Centro Oeste [-] Sudeste [4] Sul [-]	Autogestão [2] Cooperativa [5] Medicina de Grupo [-]	91,79% Intervalo [88,47% - 96,36%]

Fonte: Elaboração Própria, (2023).

Após contato e convite para participação e cessão da base de dados para esta pesquisa, a amostra final foi formada por 3 OPS brasileiras, de portes, regiões e modalidades distintas. Destaca-se, como reforço, que a amostra final foi composta por critério exclusivo de disponibilidade e acesso à base de dados.

A partir de então, estas OPS serão denominadas, respectivamente, OPS Alfa, OPS Beta e OPS Gama, a fim de garantir o sigilo e confidencialidade sobre as informações cedidas pelas mesmas. Para uma melhor visualização, a Tabela 5 caracteriza estas OPS por: porte; quantidade de beneficiários (não sendo exposta a quantidade exata por sigilo); região; modalidade e medida da sinistralidade da carteira.

Tabela 5 – Caracterização das OPS Alfa, Beta e Gama

Operadora	**Porte**	**Beneficiários**	**Região**	**Modalidade**	**Sinistralidade**
OPS Alfa	Pequeno	Acima de 18 mil e abaixo de 20 mil	Centro-oeste	Autogestão	88,95%
OPS Beta	Médio	Acima de 55mil e abaixo de 60 mil	Nordeste	Medicina de grupo	89,63%
OPS Gama	Grande	Acima de 110 mil e abaixo de 125 mil	Sudeste	Cooperativa médica	88,41%

Fonte: Elaboração Própria, (2023).

A OPS Alfa possui mais de duas décadas de atuação no mercado de saúde suplementar e por ser da modalidade autogestão, a mesma possui comercialização exclusiva de plano coletivo empresarial.

A OPS Beta possui mais de duas décadas de atuação no mercado de saúde suplementar e por ser da modalidade medicina de grupo, a mesma possui comercialização de plano individual ou familiar; coletivo por adesão e coletivo empresarial. Esta OPS se caracteriza por possuir uma carteira de aproximadamente 48% de beneficiários vinculados ao plano individual ou familiar, cujo dados de mercado apontam um índice de sinistralidade superior aos planos coletivos (ANS, 2023).

A OPS Gama possui mais de duas décadas de atuação no mercado de saúde suplementar e por ser da modalidade cooperativa médica, a mesma possui comercialização de plano individual ou familiar; coletivo por adesão e coletivo empresarial. Esta OPS se caracteriza por possuir uma carteira que demonstra equilíbrio, sendo aproximadamente 40% de beneficiários vinculados a planos individual ou familiar, 42% coletivo empresarial e 18% coletivo por adesão.

As bases fornecidas pelas OPS Alfa, OPS Beta e OPS Gama são compostas por 15 variáveis, segmentadas três categorias: perfil sociodemográfico; econômico e financeiro; indicadores de saúde, conforme Tabela 6. Para fins de unificação de análise e comparabilidade entre os pares do segmento da saúde suplementar, o período de coleta das variáveis compreendeu os anos 2018 a 2023, para todas as OPS participantes.

Tabela 6 - Sumário das variáveis

CATEGORIA	VARIÁVEL	ESTATÍSTICAS/VALORES
PERFIL SOCIODEMOGRÁFICOS	1. Sexo	1. Feminino 2. Masculino
	2. Idade	-
	3. Raça	1.Branco 2.Negro 3.Pardo 4.Amarelo 5.Indígena
	4. Região	1.Norte 2.Nordeste 3.Centro Oeste 4.Sudeste 5.Sul
	5. Estado civil	1.Divorciado 2.Casado 3.Solteiro 4.Viúvo
ECONÔMICO E FINANCEIRO	6. Despesa_Operadora [assistencial]	-
	7. Receita_Operadora	-
	8. Renda	-
	9. Ano	2018 < 2019 < 2020 < 2021 < 2022 < 2023
INDICADORES DE ESTADO DE SAÚDE	10. Câncer	1.Não 2.Sim
	11. Diabetes	1.Não 2.Sim
	12. Pressão_alta	1.Não 2.Sim
	13. Colesterol_alto	1.Não 2.Sim 3.Não diagnosticado
	14. Fumante	1.Não 2.Sim

	15. Rotina_Exames	1.Não 2.Sim

Fonte: Elaboração Própria, (2023).

As bases de dados são compostas por fatores, que são variáveis de classe (frequência) e numéricas (representado pelas estatísticas). Para a operacionalização destas bases de dados e desenvolvimento dos algoritmos de aprendizagem de máquina, foi utilizada a linguagem de programação *Phyton*.

A seleção das variáveis do estado de saúde, tomou por base o estudo de Wichmann *et al.* (2023) e GSF Silva *et al.* (2023), alinhados ao critério de disponibilidade na base de dados das OPS participantes. As variáveis da categoria econômica e financeira surgem como inovação desta tese, representando os dados que compõem o índice de sinistralidade das OPS. Assim, os modelos de aprendizagem de máquina foram treinados com estas variáveis e validados por cada modelo detalhado no item 3.2.3.

Importante destacar a compatibilidade existente acerca da padronização de dados no segmento da saúde suplementar, sendo responsabilidade da ANS a edição de padrões que permitem a uniformização entre dados de OPS e consequentemente a comparabilidade entre as OPS do setor.

Desse modo, para fundamentar esta afirmação, por meio da vivência da pesquisadora desta tese no segmento da saúde suplementar, 15 OPS (grande porte, médio e pequeno porte) de todas as regiões do Brasil foram consultadas (por meio digital) e confirmaram a compatibilidade dos dados, o que permite vislumbrar uma aplicação estendida e ampla para toda e qualquer OPS que disponibilizar sua base de dados.

3.2.3 Tratamento quantitativo dos dados

Para Ghosh e Dubey (2013), a tecnologia de mineração de dados e aprendizagem de máquina tem sido considerado como meio útil para identificar padrões e tendências de grande volume de dados. Assim, para esta etapa, o objetivo se direciona para a buscar a identificação de padrões e tendências nos dados da base das OPS Alfa, Beta e Gama. A abordagem combinada de aprendizagem de máquina não supervisionada (análise de *cluster*) e supervisionada, foi utilizada e validada por Teixeira *et al*. (2023), no estudo em que identificou numerosos grupos de mortalidade em várias regiões do Brasil.

Para tanto, essa estratégia se subdividiu em dois momentos, sucessivos e distintos: *análise de cluster – segmentação por grupos de risco* e *modelagem de aprendizagem de máquina aplicados à análise preditiva das despesas assistenciais*.

3.2.3.1 Análise de *Cluster* – Segmentação por grupos de risco

A análise de *Cluster* ou aprendizagem não-supervisionada tem por objetivo que os modelos de aprendizagem os dados criem grupos de segmentação, considerando as classes por proximidade das informações. Assim, dentro de cada grupo os elementos devem ser semelhantes entre si e diferentes dos elementos de outros grupos.

Especificamente nesta tese, a análise de *cluster* foi utilizada para segmentar a população em grupos de riscos, considerando o perfil e as variáveis na Tabela 5. Para tanto, com base em Ghosh e Dubey (2013), foi escolhido os algoritmos (modelo de aprendizagem de máquina): *K-Means* o *Fuzzy C-Means*.

a) *K-Means* e o *Fuzzy C-Means*

Para Ghosh e Dubey (2013), o agrupamento *K-Means* ou *Hard C-Means* é basicamente um método de particionamento aplicado para analisar dados e trata observações dos dados como objetos baseados em localizações e distância entre vários pontos de dados de entrada. Particionando os objetos em *cluster*s mutuamente exclusivos (K), de tal forma que os objetos dentro de cada *cluster* permaneçam tão próximos entre si e o mais longe possível de objetos em outros aglomerados.

O *K-means* é então um algoritmo não supervisionado que por meio da distância euclidiana, separa os dados em grupos com aqueles que obtiverem maior proximidade em relação ao centroide (Bezdek, 1981). O modelo é descrito pela equação a seguir:

$$j = \sum_{j=1}^{k} \sqrt{\sum_{i=1}^{n_j} (x_i - c_i)^2} \quad (2)$$

onde:
K = quantidade de centroides;
x = vetor de observações por variável
c = centroide.

Os algoritmos de *fuzzy c-means*, são baseados nos números nebulosos, onde não é por meio da matemática tradicional binária mas sim por probabilidades. Assim, aplicado ao modelo de agrupamento, o *fuzzy c-means* estima probabilidades, onde cada observação pode pertencer em uma classe ou outra, dada a distância dos pontos para o centroide.

3.2.3.2 Modelos de aprendizagem de máquina

a) Modelagem de aprendizagem de máquina aplicados à análise preditiva das despesas assistenciais

A partir das 15 variáveis destacadas na Tabela 5 – sumário de variáveis, foram selecionados três modelos de aprendizagem de máquina, supervisionados, para a previsão das despesas, sendo: K-Vizinhos Próximos (K-*nearest neighbors*), Florestas Aleatórias e *XGBoost*.

Durante a fase de pré-processamento dos dados foram implementadas técnicas de normalização e codificação. Variáveis categóricas foram transformadas utilizando codificação One-Hot, uma técnica que preserva a natureza não ordinal das categorias.

Para reduzir a dimensionalidade dos dados, foi aplicado o logaritmo natural na variável alvo (despesa) e a normalização nas demais variáveis numéricas. Utilizou-se o método MinMaxScaler para a normalização dos dados, uma técnica empregada para reescalar os valores das variáveis numéricas dentro de um intervalo entre 0 e 1. Esta abordagem é particularmente vantajosa quando se trabalha com algoritmos que são sensíveis à variação de escala entre os atributos, como os baseados em distâncias ou aqueles que implementam regularização.

A escolha por esta técnica deve-se ao fato de que a normalização MinMax não distorce as diferenças nos intervalos de valores das características e preserva as informações originais, mesmo após a aplicação da escala. Para as categorias, foram criadas variáveis *dummy*.

Para validação dos modelos, a amostra foi dividida em 70% para treino e 30% teste. Esta segregação ocorreu em virtude da margem de treino dos modelos, a fim de verificar os melhores parâmetros para a replicação na base de teses, conforme validação de Teixeira *et al.* (2023). Como estratégia metodológica, os 30% corresponde aos dados de 2022 e 2023, de forma que os algoritmos treinados foram utilizados para previsão dos dados dos exercícios de 2022 e 2023. Além disso, dentre os 70% da base de treino, foi aplicada a validação cruzada como parte do procedimento de otimização de hiper parâmetros utilizando a técnica de Grid Search.

A técnica de validação cruzada se dá por meio da geração de amostras aleatórias da base original, a fim de treinar o mesmo modelo com diferentes combinações de parâmetros para

otimizar a relação erro-parâmetros (Ramezan *et al.*, 2019), evitando o sobre ajuste do modelo treinado para com a variável alvo, isto é, dada as diversas combinações dos parâmetros, a validação cruzada busca minimizar o erro do modelo, sendo representado pelo RMSE (*Root mean squared error* - raiz quadrada do erro-médio).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=i}^{n} (x'_i - y_i)^2} RMSE = \sqrt{\frac{1}{n}\sum_{i=i}^{n} (x'_i - y_i)^2} \quad (3)$$

Onde:
x’ = observação prevista,
y = observação real,
n = tamanho da amostra.

A métrica adotada para validar os modelos foi o RMSE, que busca minimizar a distância entre o valor previsto e a variável alvo. Assim, quanto menor o valor para essa métrica, mais robusto o modelo pode ser considerado.

a.1) K-Vizinhos Próximos ou *K-nearest neighbors*

O modelo *KNN* ou K-Vizinhos Próximos é um modelo que agrupa dados com distância próxima e então realiza a predição por meio desta similaridade (Borde *et al.*, 2017). Em sua pesquisa, Borde *et al. (2017)* realizou um estudo comparativo de vários algoritmos de aprendizado de máquina com o objetivo examinar a viabilidade desses algoritmos e selecionar o mais viável, obtendo como resultado o modelo KNN, sendo esta a pesquisa base para a escolha deste modelo. Sua equação é dada por:

$$\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \qquad \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2} \quad (4)$$

Onde,
K = quantidade de vizinhos,
x = variáveis independentes,
y = variável dependente.

Por meio da validação cruzada é possível então buscar o valor de K que minimiza o erro, de outro modo, busca-se então a menor distância entre os fatores.

a.2) Florestas Aleatórias ou *Random Forest*

Os algoritmos de Florestas Aleatórias, florestas de árvores de decisão ou Random Forest são modelos *ensemble*, ou seja, combinam diversas árvores de decisão a fim de verificar o menor erro por cada árvore (Gong *et al.*, 2018). Assim, o parâmetro afim de maximizar essa relação é o *mtry*, que representa o número de variáveis em cada nódulo das árvores. É dado por:

$$mtry = \sqrt{n(x)} \qquad (5)$$

Onde *n* representa o número de variáveis da base de dados x.

a.3) *Extreme Gradient Boosting (XGBoost)*

O *XGBoost* (*extreme Gradient Boosting*) é um modelo baseado em árvores de decisão, diferenciando em seu algoritmo que funciona como uma técnica *ensemble*, conforme Chen e Guestrin (2016). Desse modo, são selecionadas árvores distintas com erros elevados e então o algoritmo treina novos modelos com base no resultado das árvores treinadas anteriormente.

Os parâmetros são:
max_depth = o limite máximo de profundidade de cada árvore, com função de mitigar o sobreajuste;
min_child_weight = o limite mínimo de somatório dos pesos, com função de mitigar o sobreajuste;
subsample = determina a porcentagem de cada observação para ser selecionada randomicamente na criação de novas árvores, quanto menor o valor, mais conservador o algoritmo tende a ser, porém, leva para subajuste;
colsample_bytree = determina porcentagem de cada variável selecionada aleatoriamente para treinar novas árvores;
eta = determina a taxa de aprendizagem do modelo, o nível de robustez varia com o conservadorismo do modelo, quanto menor, mais robusto tende a ser, porém, a intensidade computacional aumenta.

3.3 INTERAÇÃO ENTRE OBJETIVOS E ESTRATÉGIA METODOLÓGICA

Com o objetivo de demonstrar o alinhamento entre os objetivos da tese e o percurso metodológico a ser utilizado, o quadro resumo abaixo detalha sistemática e sequencialmente a

ordem dos objetivos e as respectivas estratégias metodológicas, contudo, as especificações acerca de cada estratégia estão dispostas no item 3.

Quadro 01 – Interação entre objetivos e metodologia

OBJETIVOS ESPECÍFICOS	ESTRATÉGIA METODOLÓGICA
a) Calcular a sinistralidade das OPS brasileiras, segregando por região geográfica, porte e modalidade.	- Metodologia própria da ANS. Coleta de variáveis contábeis [receita e despesa assistencial].
b) Aplicar modelos de inteligência artificial para a predição das despesas assistenciais em operadoras de planos de saúde, considerando informações contábeis, assistenciais e sociodemográficas.	- Base de dados de mundo real de 3 OPS brasileiras. - Modelo de aprendizagem não supervisionado - Análise de *cluster*s (modelos *K-Means* e *Fuzzy C-means)*. - Modelagem de aprendizagem de máquina aplicados à análise preditiva (K-Vizinhos Próximos, Florestas Aleatórias e *XGBoost*).
c) Analisar o nível de acurácia das modelagens preditivas informacionais na previsibilidade das despesas assistenciais e seu impacto no risco de solvência.	- Acurácia dos modelos por meio do RMSE. - Indicar o melhor modelo de aprendizagem de máquina para predição informacional das despesas assistenciais das OPS.
d) Propor diretrizes para análises preditivas informacionais, a partir da inteligência artificial, visando aperfeiçoar a gestão de risco de solvência das OPS na saúde suplementar brasileira.	- Elaboração de grupo de diretrizes que, a partir dos resultados identificados nas etapas anteriores, podem orientar os gestores de OPS sobre novas perspectivas de utilização da predição informacional acerca do risco de solvência das OPS.

Fonte: Elaboração própria, 2023.

4. RESULTADOS E DISCUSSÃO

Nesta seção, foram descritos, analisados e discutidos os resultados encontrados na pesquisa. Inicialmente, foi retratado o cenário e contexto nacional das OPS brasileiras, com um olhar aprofundado para a sinistralidade e, posteriormente, realizou-se testes das modelagens preditivas e algoritmos de aprendizagem em base de dados de mundo real de três OPS, com variações de porte, região e modalidade, a fim de verificar qual modelo traz mais acurácia na predição informacional das despesas assistenciais das OPS.

4.1 Cenário e contexto nacional

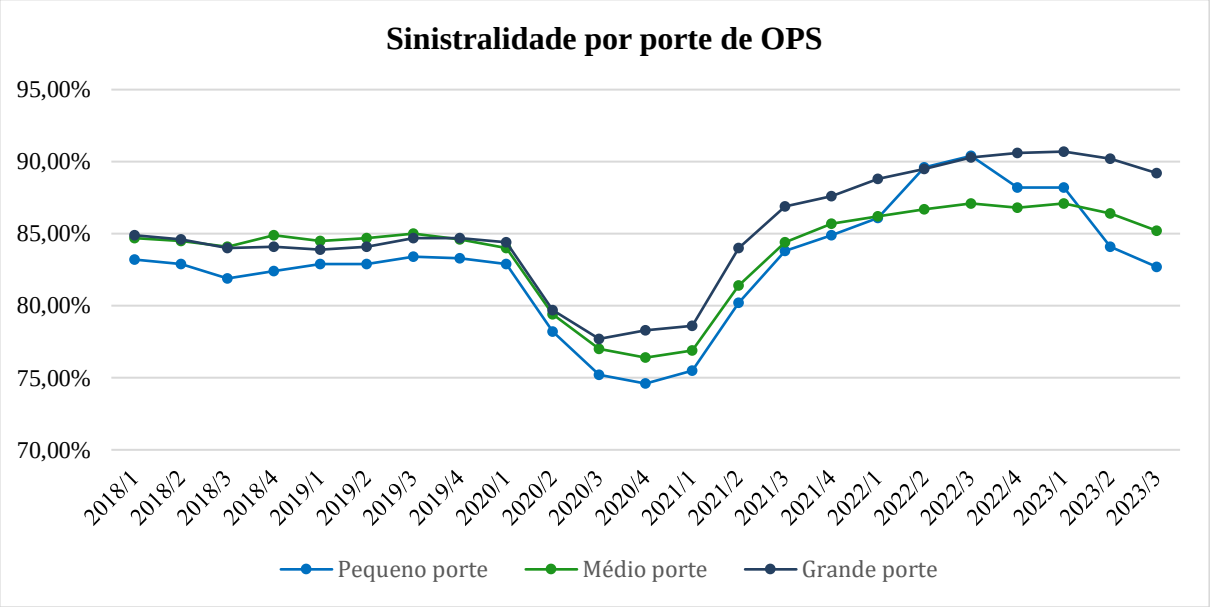
Para retratar o cenário e contexto nacional, inicialmente destaca-se que a média nacional de sinistralidade das OPS médico-hospitalar é de 88,20% (ANS, 2023). Desse modo, o gráfico

1 apresenta uma análise abrangente da sinistralidade das 577 OPS ao longo dos últimos seis anos, segmentado por porte (pequeno, médio e grande) e trimestre. Os dados revelam padrões e variações significativas, proporcionando uma visão detalhada das tendências no setor.

A análise da evolução trimestral permite observar flutuações sazonais e identificar possíveis correlações entre o porte da operadora e sua performance em termos de sinistralidade. De maneira mais específica, nota-se que o segundo, terceiro e quarto trimestres de 2020 foram fortemente afetados pelo efeito da pandemia da COVID-19, fenômeno que reduziu significativamente as despesas assistenciais das OPS.

Sobre este fenômeno, cabe ressaltar que as medidas de contenção da pandemia, como *lockdowns* e restrições à circulação, impactaram diretamente na busca e efetiva utilização dos serviços de saúde não relacionados à COVID-19. As OPS registraram adiamentos de procedimentos eletivos, a realização de consultas médicas remotamente e a procura por determinados tipos de cuidados de saúde diminuiu. A combinação desses fatores criou um ambiente desafiador, marcado pela pressão sobre as OPS e as conduzindo a reavaliação de suas estratégias para adaptação às novas condições e, ao mesmo tempo, garantia da continuidade da assistência aos beneficiários.

Gráfico 1 – Média da sinistralidade por porte de OPS



Fonte: Elaboração própria, 2023 – Dados coletados do painel econômico-financeiro da ANS.

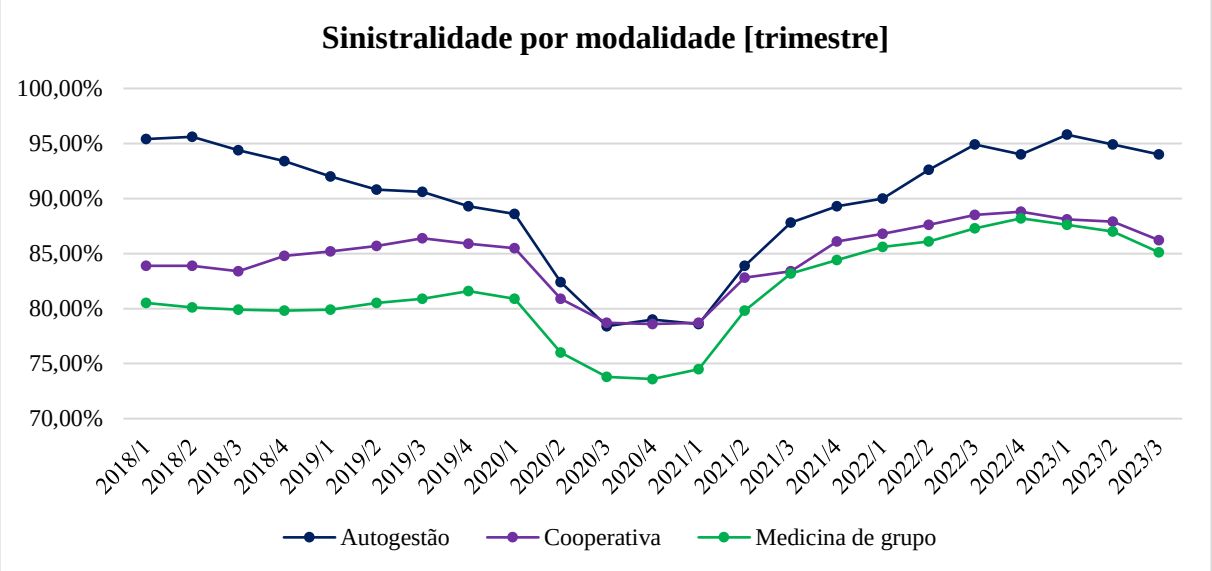
Adicionalmente, o Gráfico 1 evidencia nuances importantes da sinistralidade, destacando possíveis desafios enfrentados por operadoras de diferentes portes em momentos específicos, especificamente a partir do segundo trimestre de 2021, período marcado por altas

despesas com a COVID-19. A curva crescente da sinistralidade, após este trimestre de 2021, representa a demanda por saúde, reprimida nos últimos trimestres de 2020, mas ainda retrata o aumento exponencial no número de casos da doença, que exigiu uma resposta rápida e abrangente do setor de saúde. Logo, a necessidade de tratamento intensivo, uso de equipamentos médicos específicos e a contratação de profissionais de saúde especializados para lidar com a demanda extraordinária contribuíram para um aumento substancial das despesas assistenciais e consequentemente afetando a sinistralidade das OPS.

Observa-se ainda que há uma correspondência entre o tamanho (porte) da OPS e o índice da sinistralidade, isto porque o tamanho da carteira de beneficiários, pode influenciar o risco de sinistro para a OPS. Desse modo, as OPS de grande porte apresentam sinistralidade média acima das de médio e pequeno porte, exceto no exercício de 2019.

O Gráfico 2 exibe a média da sinistralidade trimestral das OPS entre os anos de 2018 e 2023, categorizando-as por modalidade, incluindo autogestão, medicina de grupo e cooperativa médica. Essa segmentação permite uma análise visual das diferenças no índice de sinistralidade entre as modalidades, oferecendo *insights* acerca da perspectiva da eficiência operacional e despesas assistenciais de cada grupo de OPS. A modalidade de autogestão, cujos beneficiários são, em sua maioria, os próprios funcionários ou membros de uma determinada empresa, associação ou grupo específico (ANS, 2023) apresentaram maior média de sinistralidade em todos os anos dispostos no gráfico 2, sugerindo que a faixa etária das carteiras das OPS nesta modalidade pode ser um fator de explicação direta do aumento da sinistralidade.

Gráfico 2 – Média da evolução trimestral da sinistralidade por modalidade de OPS

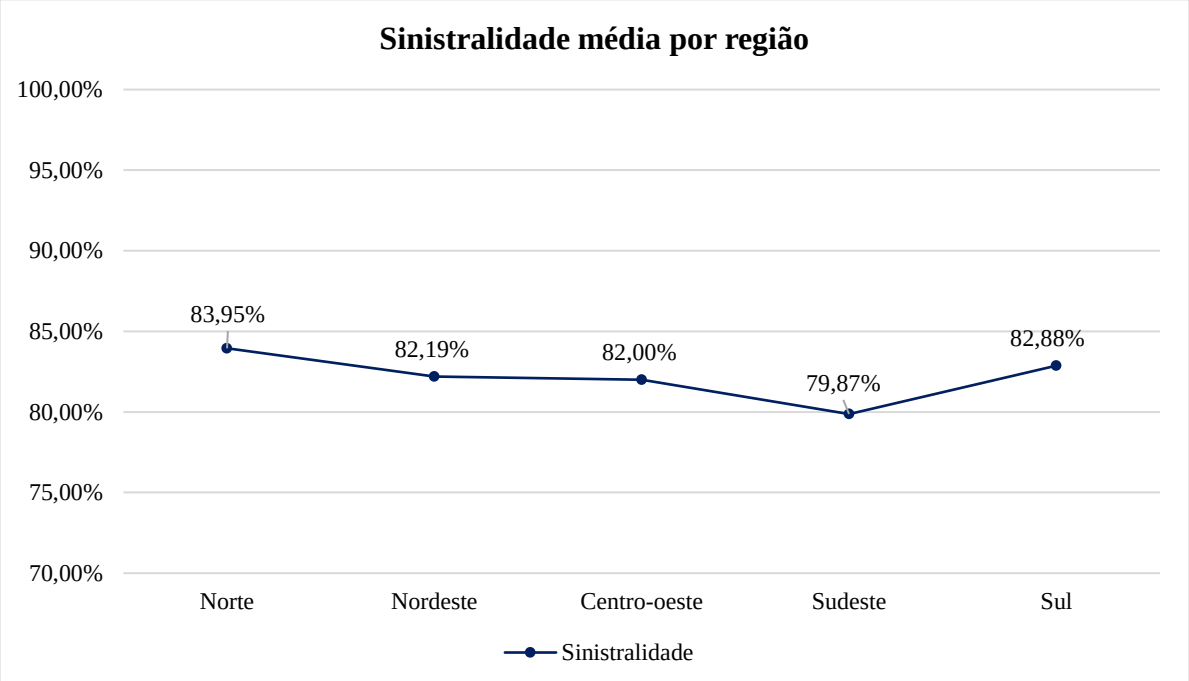


Fonte: Elaboração própria, 2023 – Dados coletados das demonstrações contábeis disponibilizadas na base da ANS.

Ao observar as tendências demonstradas trimestralmente no Gráfico 2, é possível identificar padrões distintos de sinistralidade em cada modalidade, demonstrando que as especificidades acerca de estrutura, faixa etária da carteira, composição de gênero, região geográfica podem exercer alguma influência sobre este índice. Desse modo, até mesmo por uma visão mais simples, reguladores, gestores e demais stakeholders podem compreender a dinâmica do mercado de saúde suplementar, e de maneira específica, os gestores das OPS podem se utilizar dessa visão para a tomada de decisões estratégicas informadas e mais assertivas acerca da alocação dos recursos.

A partir das observações do índice de sinistralidade por porte das OPS e modalidades, torna-se importante retratar este mesmo índice por região geográfica do Brasil. Desse modo, o Gráfico 3 retrata a sinistralidade média das OPS conforme região geográfica, em que a região Norte apresenta o maior índice de sinistralidade, sendo 83,95%, seguido da região Sul com 82,88%, região Nordeste com 82,19%, região Centro-oeste com 82% e por fim a região Sudeste com índice de 79,87%.

Gráfico 3 –Sinistralidade por região geográfica



Fonte: Elaboração própria, 2023 – Dados coletados das demonstrações contábeis disponibilizadas na base da ANS.

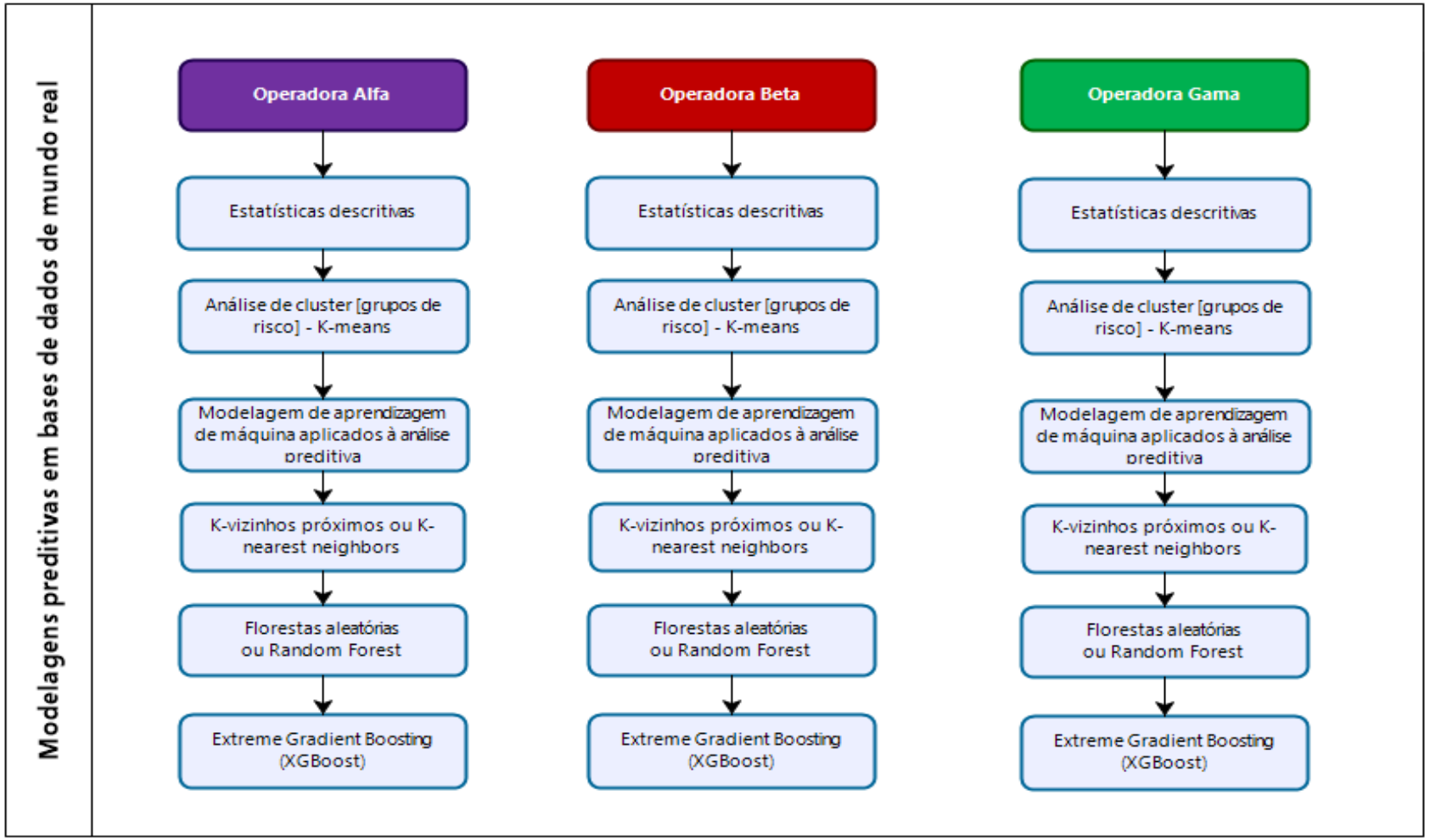
É certo destacar que a variação regional no Brasil introduz complexidades adicionais, entre as quais cito:

1. Perfil demográfico e epidemiológico: Áreas com uma população mais envelhecida podem ter uma maior demanda por serviços de saúde.
2. Infraestrutura de saúde: Regiões com uma infraestrutura mais robusta podem ter uma capacidade maior de atender às necessidades de saúde, influenciando a sinistralidade.
3. Economia local: A situação econômica pode influenciar a capacidade das pessoas de adquirir planos de saúde, assim, em áreas com maior estabilidade econômica, as pessoas podem ter mais recursos para investir em cuidados de saúde.
4. Padrões culturais e comportamentais: Diferenças nos padrões culturais e comportamentais das populações regionais podem impactar os hábitos de saúde, influenciando a frequência e a natureza das demandas por serviços assistenciais.
5. Oferta de profissionais de saúde: Disponibilidade e distribuição de profissionais de saúde em diferentes regiões podem afetar a capacidade de atendimento e influenciar a sinistralidade.

4.2 Modelagens de aprendizagem de máquina em bases de dados de mundo real

A fim de testar as modelagens de aprendizagem de máquina, a Figura 7 demonstra a sequência metodológica para as OPS Alfa, Beta e Gama. De maneira contínua, o resultado será apresentado de forma individual, ou seja, por OPS, iniciando pelas estatísticas descritivas; análise de *cluster* e os modelos de aprendizagem de máquina para predição das despesas assistenciais, sendo: K-vizinhos próximos ou *K-nearest neighbors*; Florestas aleatórias ou *Randon forest* e XGBoost ou *Extreme gradiente boosting*.

Figura 7 – Modelagens de aprendizagem de máquina em bases de dados de mundo real



Fonte: Elaboração própria, (2023).

4.2.1 Operadora Alfa

Os resultados da OPS Alfa estão segmentados em: a) estatísticas descritivas; b) análise de *cluster* – grupos de risco e c) algoritmos de aprendizagem de máquina – análise preditiva.

a) Estatísticas descritivas

A Tabela 7 apresenta uma análise detalhada das variáveis relacionadas aos perfis sociodemográficos, indicadores econômicos e financeiros e indicadores de estado de saúde dos beneficiários da OPS Alfa. Na categoria das variáveis de perfil sociodemográfico, observa-se uma distribuição de gênero equilibrada, com 54.4% de mulheres e 45.6% de homens. A idade média dos beneficiários é de 47.7 anos, variando de 18 a 85 anos. A maioria dos beneficiários é de raça branca (70.3%), e a distribuição geográfica indica uma representação total da região Centro Oeste (localização da OPS Alfa). Quanto ao estado civil, a maioria é casada (48.9%).

Sob o aspecto econômico e financeiro, destaca-se que a despesa assistencial média é de R$ 5.712, com uma ampla variação (mínimo: R$ 0, máximo: R$ 552.898). A receita apresenta uma média de R$ 12.352.1, variando de R$ 0 a R$ 1.798.079. A renda média dos beneficiários é de R$ 34.900.3, com uma grande dispersão (mínimo: R$ 0, máximo: R$ 301.284).

Os indicadores de estado de saúde revelam que a maioria dos beneficiários não possui câncer (90.7%), diabetes (88.6%), pressão alta (65.5%), colesterol alto (68.9%), não é fumante (85.2%), mas realiza rotineiramente exames (60.1%). Esses dados fornecem uma visão abrangente do perfil dos beneficiários e dos fatores que podem influenciar as despesas assistenciais desta carteira.

Tabela 7 - Sumário das variáveis OPS Alfa

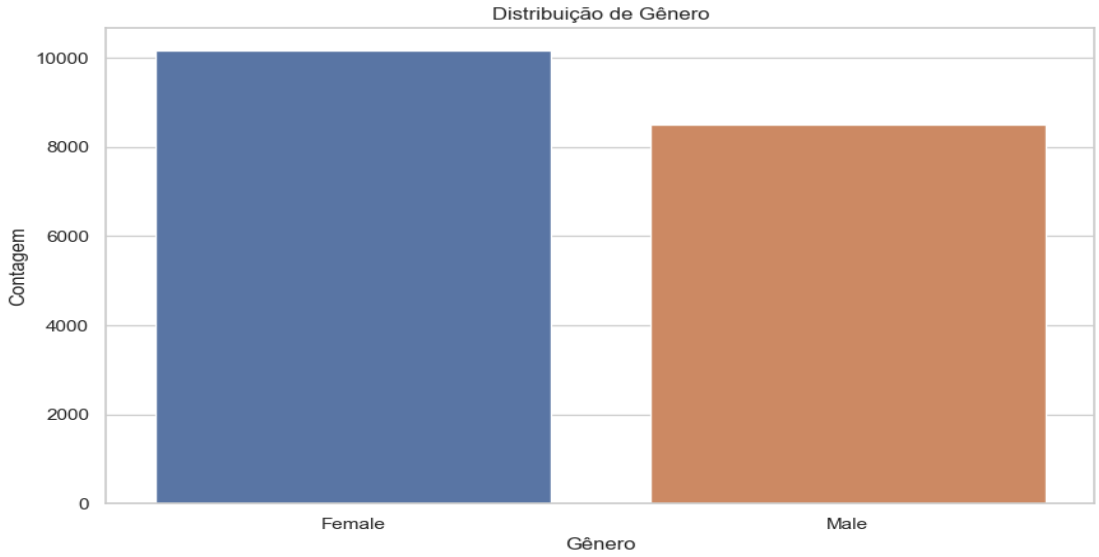
CATEGORIA	VARIÁVEL	ESTATÍSTICAS/VALORES	FREQUÊNCIA
PERFIL SOCIODEMOGRÁFICOS	1. Sexo	1. Feminino 2. Masculino	1.10156 (54.4%) 2. 8506 (45.6%)
	2. Idade	Média (sd) : 47.7 (17.8) min < med < max: 18 < 47 < 85	68 valores distintos
	3. Raça	1.Branco 2.Negro 3.Pardo 4.Amarelo 5.Indígena 6.Multiplas raças	13126 (70.3%) 1020 (5.5%) 3422 (18.3%) 339 (1.8%) 577 (3.1%) 182 (1.0%)
	4. Região	1.Norte 2.Nordeste 3.Centro Oeste 4.Sudeste 5.Sul	3. 100%
	5. Estado civil	1.Divorciado 2.Casado	2804 (15.1%) 9132 (48.9%)

		3.Solteiro 4.Viúvo	5462 (29.3%) 1264 (6.8%)
ECONÔMICO E FINANCEIRO	6. Despesa_Operadora [assistencial]	Média (sd): 5712 (15196.6) min < med < max: 0 < 1187 < 552898	7866 valores distintos
	7. Receita_Operadora	Média (sd): 12352.1 (47400.9) min < med < max: 0 < 1399 < 1798079 IQR (CV): 6612.5 (3.8)	8679 valores distintos
	8. Renda	Média (sd) : 34900.3 (37839.3) min < med < max: 0 < 24000 < 301284	7806 valores distintos
	9. Ano	Média (sd): 2020.5 (1.7) min < med < max: 2018 < 2021 < 2023 IQR (CV): 3 (0)	2018: 3109 (16.7%) 2019: 3109 (16.7%) 2020: 3110 (16.7%) 2021: 3109 (16.7%) 2022: 3109 (16.7%) 2023: 3116 (16.7%)
INDICADORES DE ESTADO DE SAÚDE	10. Câncer	1.Não 2.Sim	16934 (90.7%) 1728 (9.3%)
	11. Diabetes	1.Não 2.Sim	16541 (88.6%) 2121 (11.4%)
	12. Pressão_alta	1.Não 2.Sim	12222 (65.5%) 6440 (34.5%)
	13. Colesterol_alto	1.Não 2.Sim 3.Não diagnosticado	12864 (68.9%) 5791 (31.0%) 7 (0.0%)
	14. Fumante	1.Não 2.Sim	15891 (85.2%) 2771 (14.8%)
	15. Rotina_Exames	1.Não 2.Sim	7453 (39.9%) 11209 (60.1%)

Fonte: Elaboração Própria, (2023).

O Gráfico 4 apresenta uma distribuição dos beneficiários por gênero, evidenciando uma ligeira predominância do público feminino, que representa 54.4% do total, em comparação aos beneficiários masculinos, que correspondem a 45.6%. A diferença percentual não representa um desequilíbrio entre ambos os gêneros, contudo, essa distribuição de gênero pode influenciar diferentes padrões de utilização dos serviços assistenciais, sendo um aspecto relevante a ser considerado na compreensão dos fatores que impactam as despesas assistenciais da OPS Alfa.

Gráfico 4 –Distribuição de gênero

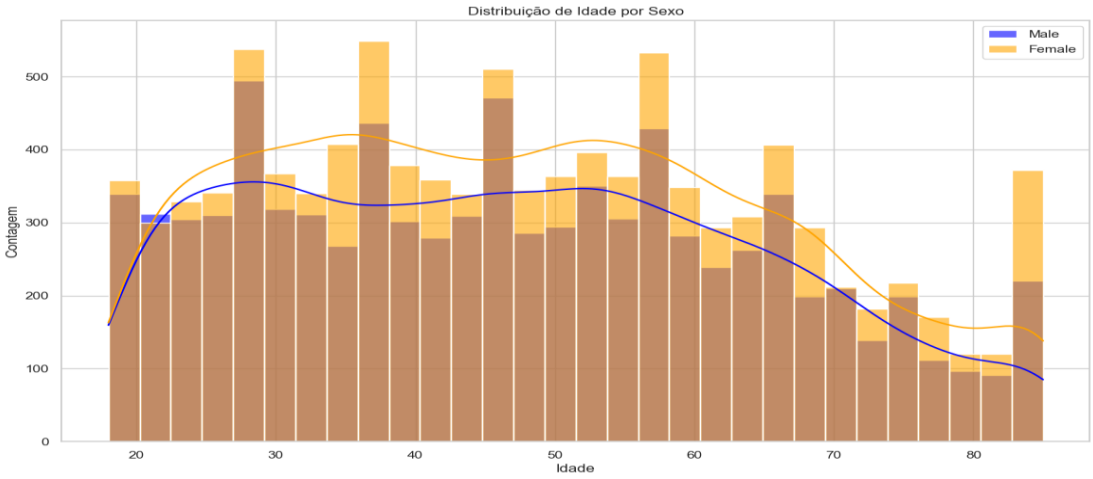


Fonte: Elaboração própria (2023).

A análise da distribuição da idade por gênero, disposta no Gráfico 5, revela uma variação que abrange o intervalo de 18 a 85 anos. Nota-se uma concentração de beneficiários nas faixas etárias de homens e mulheres entre 30 e 58 anos. Esse fato sugere uma representação significativa de adultos de meia-idade fato que merece atenção pois pode influenciar estratégias e políticas internas da OPS Alfa que visem atender às necessidades específicas desse grupo etário, acerca de aspectos de promoção à saúde.

A partir do Gráfico 5, é possível indicar que as concentrações na faixa etária entre 30 e 58 anos podem sugerir padrões comportamentais, desafios de saúde específicos ou mesmo influências socioculturais que podem ser exploradas em estudos mais aprofundados.

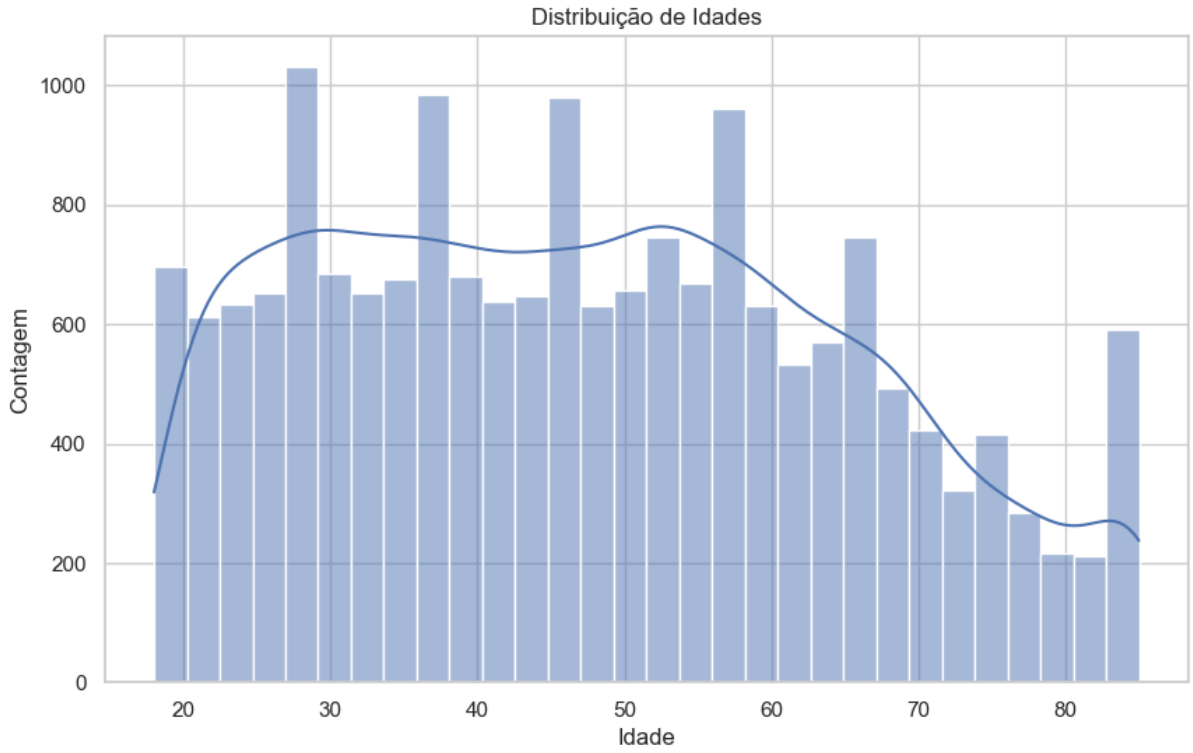
Gráfico 5 – Distribuição da idade por gênero



Fonte: Elaboração própria (2023).

O Gráfico 6 demonstra a distribuição por idade, sugerindo que a OPS Alfa atende predominantemente a adultos em diferentes estágios da vida, com uma representatividade de beneficiários nas faixas etárias consideradas maduras. Essa concentração de beneficiários na faixa etária entre 30 e 58 anos pode influenciar estratégias de oferta de serviços de saúde preventiva e especializada, considerando as demandas típicas dessa fase da vida.

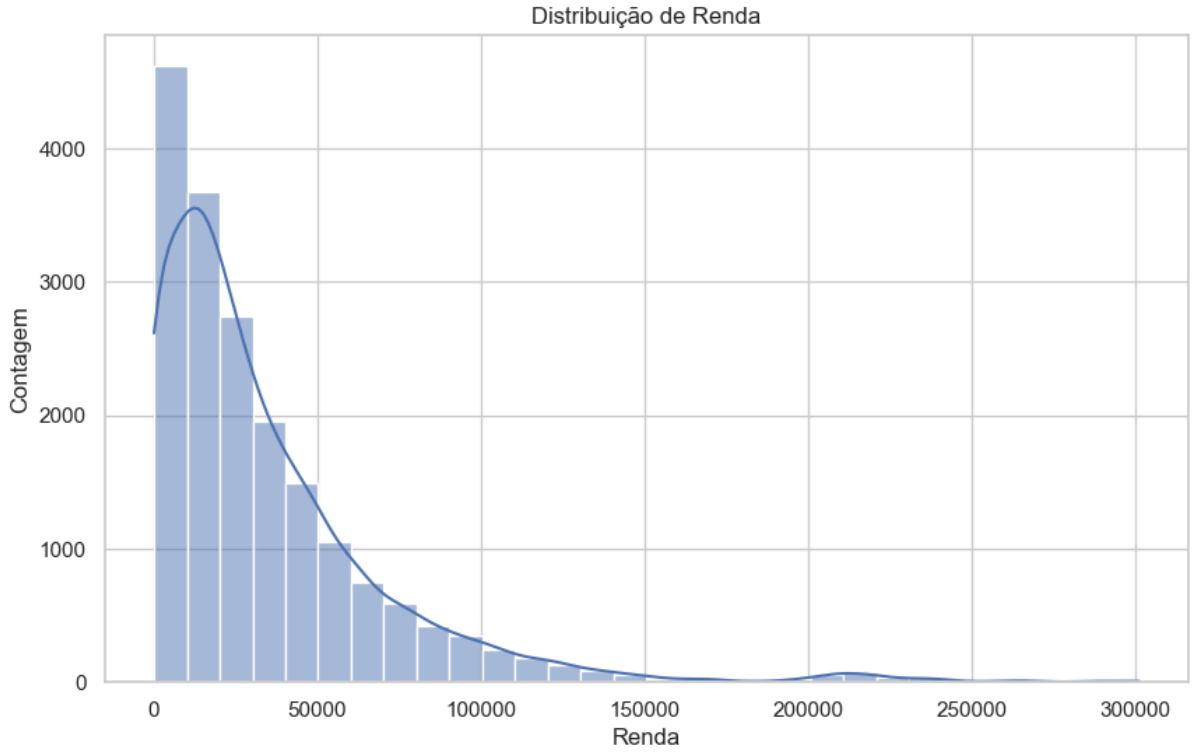
Gráfico 6 –Distribuição por idade



Fonte: Elaboração própria (2023).

O Gráfico 7 demonstra a representação visual da probabilidade de diferentes faixas de renda entre os beneficiários da OPS Alfa. No eixo horizontal, os intervalos de renda são dispostos, enquanto o eixo vertical exibe a probabilidade associada a cada faixa, sendo observado que a primeira faixa de renda concentra a maior probabilidade de representação dos beneficiários da OPS Alfa.

Gráfico 7 – Distribuição de renda



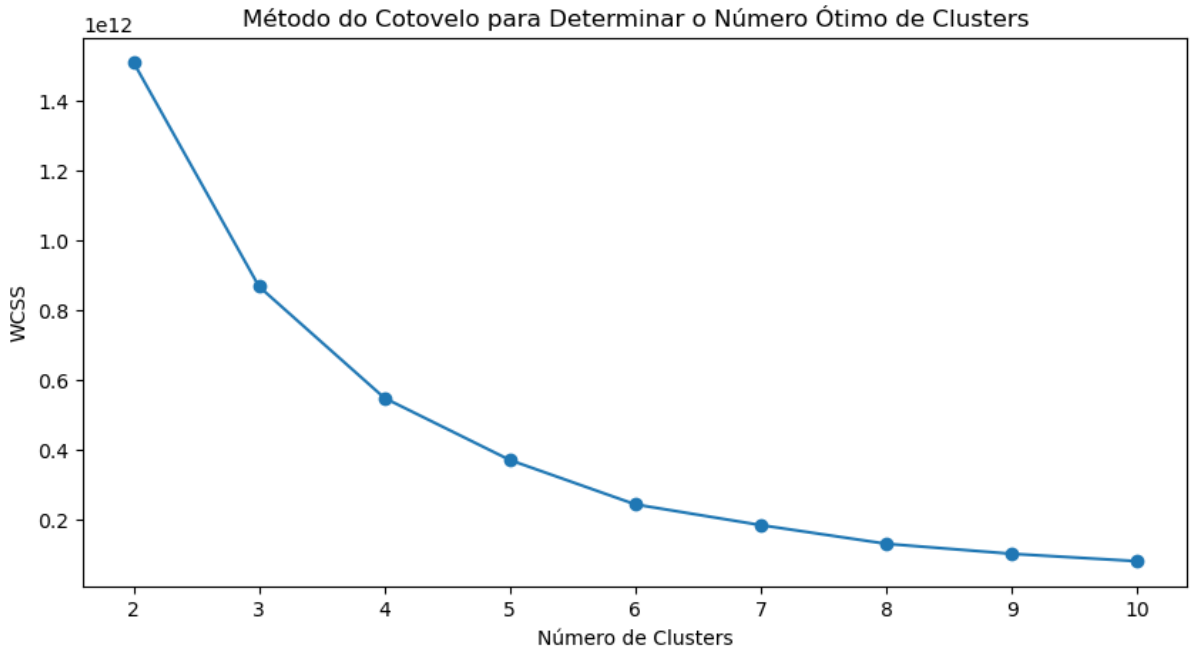
Fonte: Elaboração própria (2023).

b) Análise de *cluster* – grupos de risco

De maneira antecedente à análise de *clusters*, utilizou-se o método do cotovelo para definição no número ótimo de *clusters*, método que avalia a variância explicada em relação ao número de *clusters*. Assim, o Gráfico 8 exibe o eixo horizontal, representando o número de *clusters*, variando de 2 a 10, enquanto o eixo vertical mostra a soma dos quadrados intra*cluster*s (WCSS), uma métrica que reflete a dispersão dos pontos dentro de cada *cluster*. A curva no Gráfico 10 representa a variação do WCSS em relação ao número de *cluster*s e ao analisá-la, observamos que a curva apresenta um padrão descendente acentuado, iniciando com um valor alto no eixo vertical (1.4) e diminuindo progressivamente à medida que o número de *clusters* aumenta. O ponto onde essa curva apresenta uma mudança na taxa de declínio, formando um "cotovelo", indica o número ótimo de *clusters*, especificamente 3.

Portanto, foi considerado que três *clusters* corresponde a quantidade mais apropriada para segmentar os dados, proporcionando uma estrutura de agrupamento eficaz e representativa da variabilidade nos dados analisados.

Gráfico 8 – Método do cotovelo para determinar o número ótimo de *clusters*

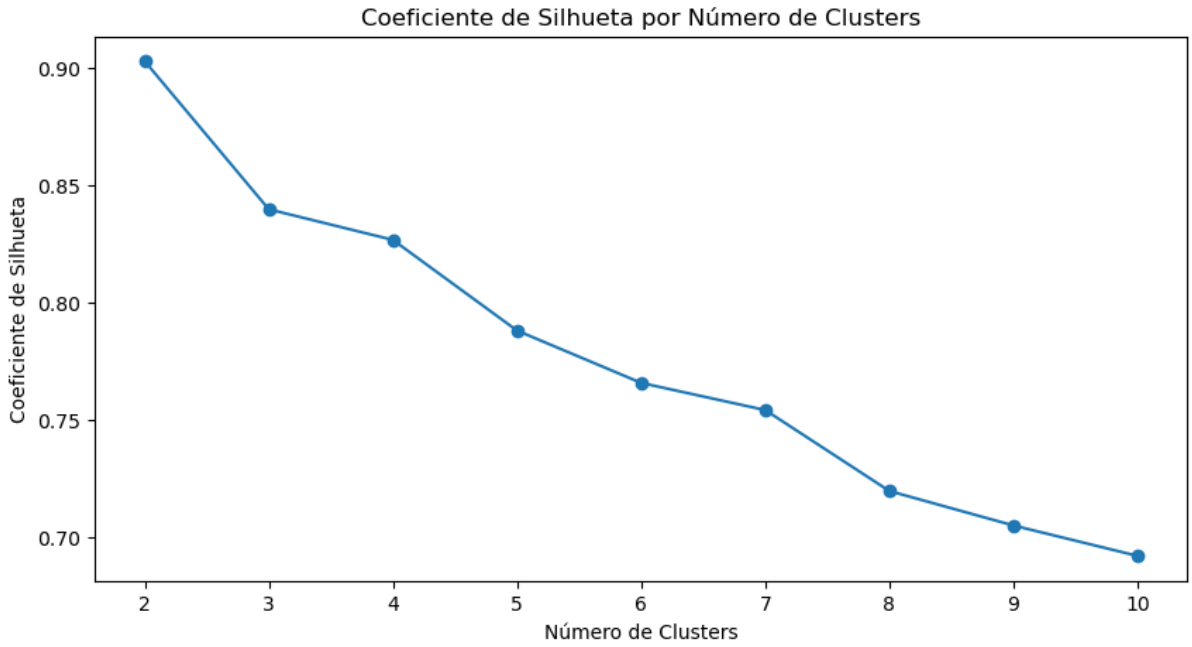


Fonte: Elaboração própria (2023).

O Gráfico 9 exibe o Coeficiente de Silhueta em relação ao número de *clusters*, fornecendo *insights* sobre a qualidade e coesão dos *clusters* em uma análise de agrupamento. O eixo horizontal representa o número de *clusters*, variando de 2 a 10, enquanto o eixo vertical mostra o coeficiente de silhueta, uma medida que avalia quão bem os objetos em um *cluster* estão separados em relação aos outros *clusters*.

Ao analisar o Gráfico 9, observa-se que a curva inicia em um ponto alto no eixo vertical (0.90) e decresce gradualmente à medida que o número de *clusters* aumenta. O coeficiente de silhueta próximo a 1, indica que os objetos estão bem agrupados e separados dos outros *cluster*s. Portanto, entende-se que um valor inicial alto sugere que a formação de três *clusters* é robusta e coesa.

Gráfico 9 – Coeficiente de silhueta por número de *clusters*



Fonte: Elaboração própria (2023).

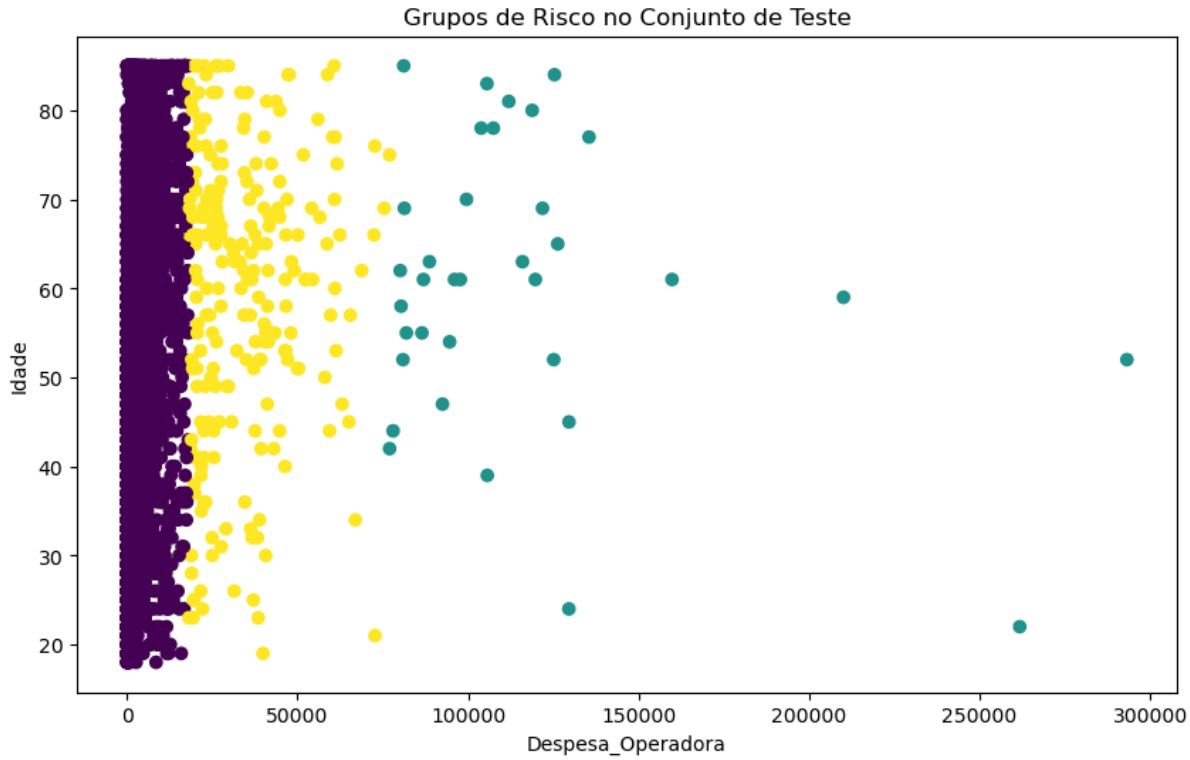
Após a definição da quantidade ótima de *clusters*, prossegue-se para a análise de *Cluster* ou aprendizagem não-supervisionada, a fim de que haja a segmentação e criação de classes por proximidade das informações. Especificamente nesse caso, a análise de *cluster* foi utilizada para segmentar a carteira de beneficiários da OPS Alfa em grupos de riscos, considerando o perfil e as variáveis apresentadas na Tabela 6. Para tanto, foram utilizados dois modelos: *K-Means* e o *Fuzzy C-Means*.

Ao analisar o Gráfico 10, que exibe os resultados obtidos pelo algoritmo K-*means*, foram identificados três *clusters* distintos representados por cores diferentes, roxo, amarelo e verde. O *cluster* predominantemente roxo, abrangendo a maior parte do Gráfico 10, sugere a presença de indivíduos com despesas assistenciais mais baixas, distribuindo-se de maneira uniforme por todas as faixas etárias. Este grupo pode ser interpretado como de "baixo risco financeiro", caracterizado por despesas assistenciais relativamente menores e sem uma correlação evidente com a idade.

Os outros dois *clusters*, embora menos densos, indicam grupos com despesas assistenciais mais expressivas. Esses grupos diferem em termos de custos assistenciais, mas não apresentam uma variação ou correlação clara em relação à idade. O *cluster* verde, em particular, destaca-se por conter indivíduos com as maiores despesas assistenciais, possivelmente representando um "grupo de alto risco financeiro". Visualmente, este grupo inclui alguns pontos dispersos, sugerindo a presença de outliers ou casos extremos de despesas assistenciais. A distribuição homogênea das idades em todos os *clusters* reforça a ideia de que a idade

isoladamente não é um fator determinante para o nível de despesas assistenciais nesta amostra específica. A detecção de *outliers*, especialmente no *cluster* de alto risco, sugere a necessidade de uma investigação mais aprofundada para compreender situações atípicas.

Gráfico 10 – Grupos de risco no conjunto de teste – K *means*



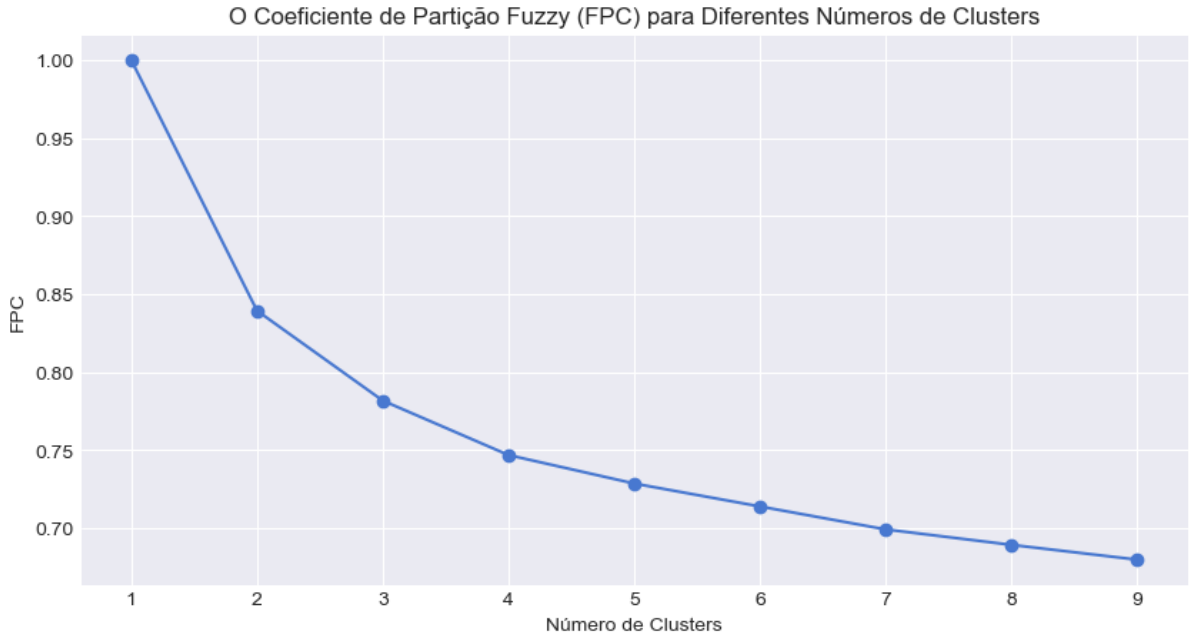
Fonte: Elaboração própria (2023).

O Gráfico 11 exibe o Coeficiente de Partição *Fuzzy* (FPC) em relação ao número de *clusters*, oferecendo *insights* sobre a adequação da partição *fuzzy* para diferentes configurações de agrupamento. O eixo horizontal representa o número de *clusters*, variando de 1 a 9, enquanto o eixo vertical mostra o FPC, uma medida que avalia a qualidade da partição *fuzzy*, indicando o quão bem os dados estão distribuídos entre os *clusters*.

Observa-se que a curva inicia em um ponto relativamente alto no eixo vertical (1) e decresce à medida que o número de *clusters* aumenta. O FPC próximo a 1 indica uma partição *fuzzy* mais precisa e eficaz. Desse modo, idealmente, busca-se um número de *clusters* que maximize o FPC, indicando uma partição *fuzzy* que melhor represente a estrutura dos dados.

Fuzzy C-Means

Gráfico 11 – Coeficiente de partição *Fuzzy* para diferentes números de *clusters*



Fonte: Elaboração própria (2023).

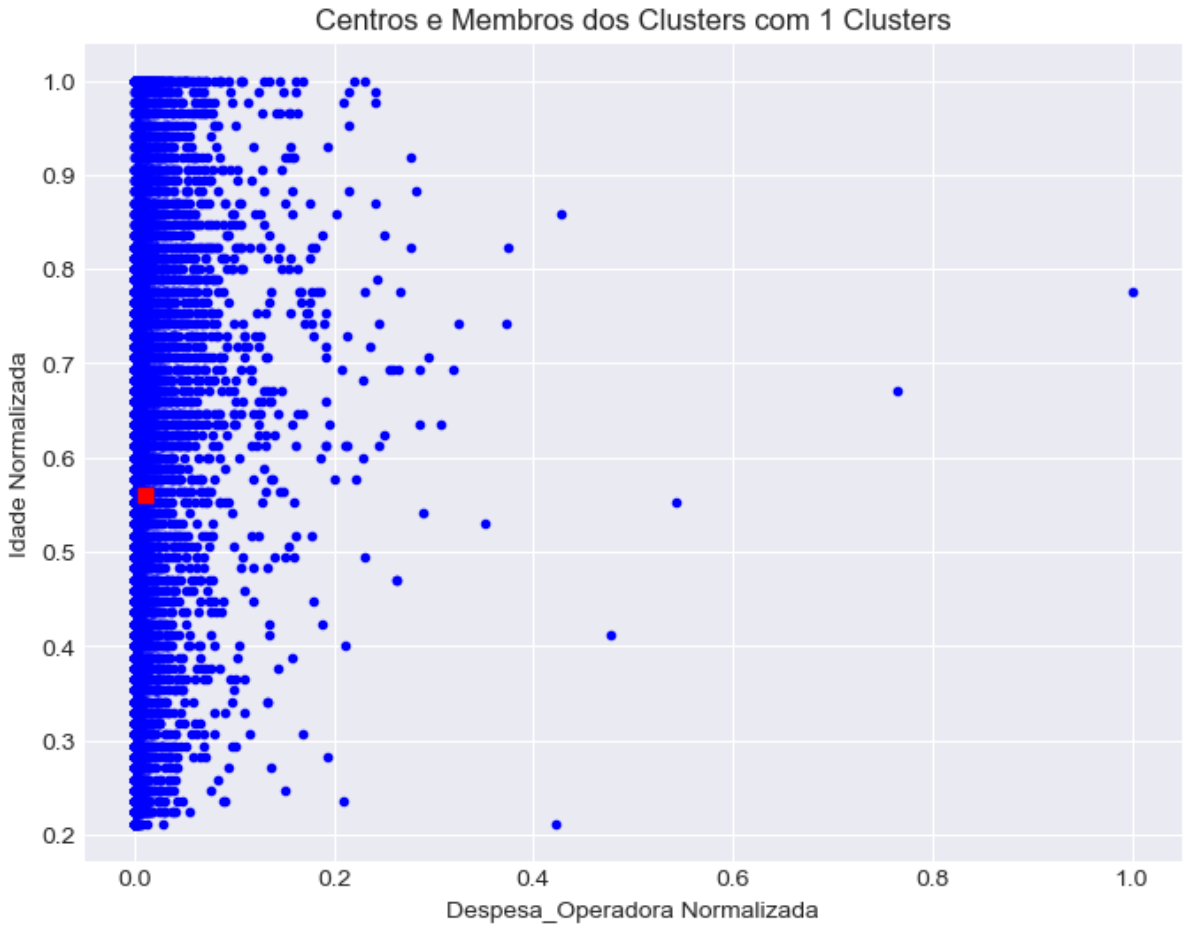
O Gráfico 12 apresenta o resultado da aplicação do algoritmo *Fuzzy C-means* para identificar padrões dentro do conjunto de dados normalizado, considerando as variáveis "Despesa_Operadora (assistenciais)" e "Idade".

Observa-se um único *cluster* (dado que o número ótimo de *clusters* foi identificado como 1 pelo critério utilizado), indicado pelos pontos azuis, com um centro de *cluster* marcado por um ponto vermelho. A distribuição dos pontos sugere uma alta densidade de dados com baixas despesas assistenciais e uma ampla dispersão de idades, sendo indicado pela concentração de pontos no lado esquerdo do Gráfico 12.

Além disso, há uma dispersão menos densa de pontos ao longo do eixo das despesas assistenciais da OPS Alfa, com poucos dados mostrando despesas maiores, mas sem um padrão claro de diferenciação com base na idade, como é evidente pela distribuição vertical dos pontos ao longo do eixo da idade.

A presença de um único *cluster* pode indicar que o conjunto de dados não possui uma variação significativa que justifique a separação em múltiplos grupos, ou pode ser um sinal de que o critério utilizado para determinar o número de *clusters* não foi sensível o suficiente para detectar agrupamentos mais sutis. O ponto central em vermelho representa o centroide do *cluster* único, que é um resumo das características médias do conjunto de dados baseado na lógica *fuzzy*, refletindo uma média ponderada das despesas operadoras e idades.

Gráfico 12 – Centros e membros dos *clusters*



Fonte: Elaboração própria (2023).

c) Algoritmos de aprendizagem de máquina – análise preditiva

Nesta seção estão expostos cada modelo de aprendizagem de máquina utilizados para alcançar os objetivos desta tese. Inicialmente a Tabela 8 expõe a acurácia dos modelos na base da OPS Alfa.

Os Gráficos desta seção mostram a distribuição das previsões feitas por três modelos de diferentes nos dados da base da OPS Alfa, indicando ainda o RMSE (*Root Mean Square Error*) para cada modelo. O RMSE é uma medida de quão bem um modelo pode prever a variável de resposta, possuindo a relação de que quanto menor o valor, melhor o modelo com relação ao conjunto de dados utilizado.

Tabela 8 - Acurácia dos modelos

Random Forest	XGBOOST	KNN

RMSE	RMSE	RMSE
10.991,95	11.381,52	15.774,09

Fonte: Elaboração própria (2023).

A Tabela 8 demonstra que o modelo KNN apresentou o maior RMSE na base da OPS Alfa, o que indica que este modelo teve o pior desempenho entre os três. Com um RMSE de 15.774,09 sugere que a abordagem baseada em vizinhos mais próximos não captura bem a complexidade dos dados ou está suscetível ao problema da "maldição da dimensionalidade", ou seja, base de dados com alta dimensionalidade.

O modelo XGBoost obteve um RMSE na base da OPS Alfa de 11.381,52. Embora tenha superado o KNN, não foi o modelo que mais se adequou ao objetivo desta tese. O XGBoost é conhecido por sua capacidade de modelar interações não lineares e complexas, mas pode não ter sido totalmente otimizado nesta base ou a natureza dos dados não se adequou às suas forças.

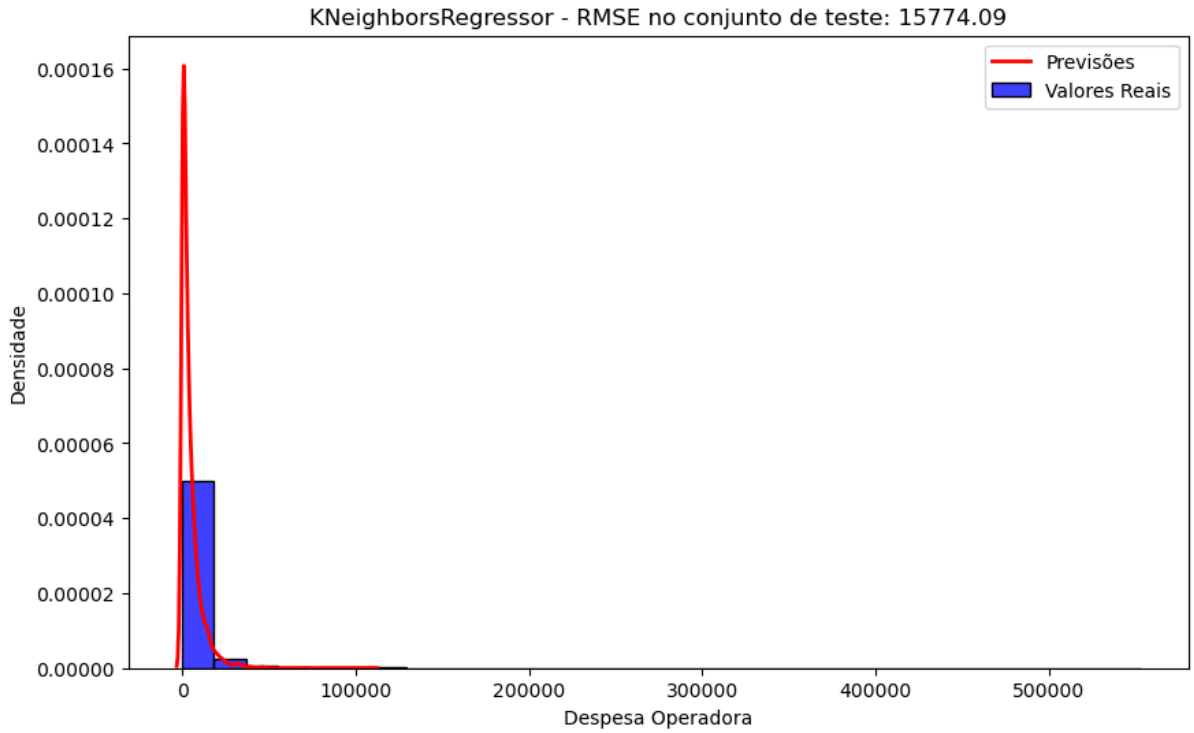
O *Random Forest* obteve o melhor desempenho com um RMSE de 10.991,95 na base da OPS Alfa. Isso sugere que o modelo foi capaz de capturar a complexidade dos dados de uma forma melhor do que os outros modelos. *Random Forests* são eficazes em lidar com variáveis categóricas e numéricas e podem capturar interações não lineares.

O Gráfico 13 apresenta a aplicação do algoritmo K-vizinhos próximos (KNN) para previsão das despesas assistenciais da OPS Alfa. No eixo horizontal, estão representados os valores das despesas assistenciais, variando de R$ 0,00 a R$ 80.000,00 enquanto o eixo vertical exibe a densidade associada a cada valor.

As previsões geradas pelo algoritmo KNN (linha vermelha), notavelmente, a curva segue o formato das colunas que representam as despesas reais, indicando uma concordância (em uma perspectiva mais geral) com os dados reais das despesas assistenciais. No entanto, há uma observação importante a respeito do ponto alto da linha, onde ela atinge uma densidade superior a 0,00012.

Essa discrepância sugere que o algoritmo KNN, ao realizar previsões, identificou uma tendência semelhante às despesas reais da OPS Alfa, contudo com algumas divergências, especialmente no ponto mais alto da curva. Essa discrepância pode indicar áreas em que o modelo de previsão está menos preciso, e que requeira alguma avaliação adicional do desempenho do algoritmo, podendo ser por meio de métricas de erro ou validação cruzada, para fins de aprimoramento da confiabilidade das previsões.

Gráfico 13 – K-vizinhos próximos ou KNN



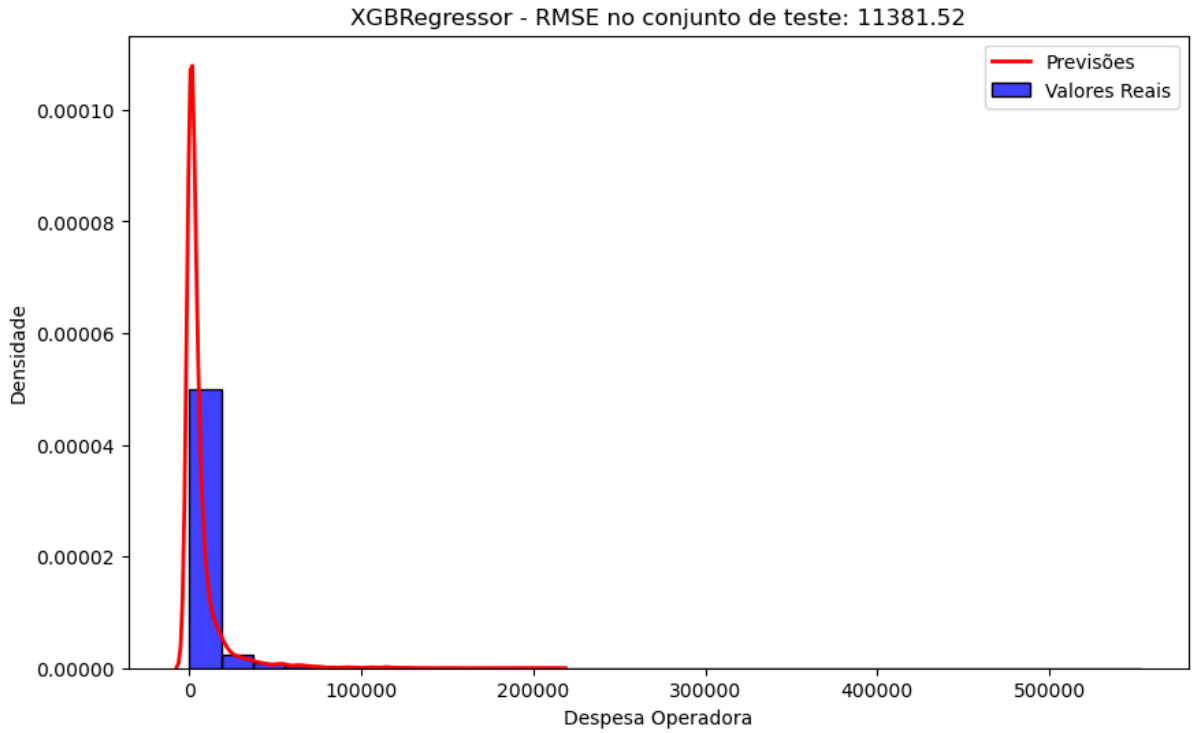
Fonte: Elaboração própria (2023).

O Gráfico 14 ilustra a aplicação do algoritmo XGB*oost* para a previsão das despesas assistenciais da OPS Alfa. No eixo horizontal, são apresentados os valores das despesas assistenciais, variando de R$ 0,00 a R$ 80.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, variando entre 0,00000 e 0,00016.

Cinco colunas azuis decrescentes indicam os valores reais das despesas assistenciais. A linha vermelha representa as previsões geradas pelo algoritmo XGBoost, apresentando uma curva que segue o formato das colunas que representam as despesas reais. No entanto, observa-se um pico na linha vermelha atinge uma densidade de 0,00016.

Essa discrepância sugere que o algoritmo XGBoost, ao realizar previsões, identificou uma tendência semelhante às despesas reais, mas com algumas divergências, especialmente no ponto de pico. Avaliar métricas de desempenho do algoritmo, como erro absoluto médio ou validação cruzada, podem ajudar na compreensão acerca da precisão do modelo e determinar se ajustes são necessários para melhorar a qualidade das previsões.

Gráfico 14 – XGBoost



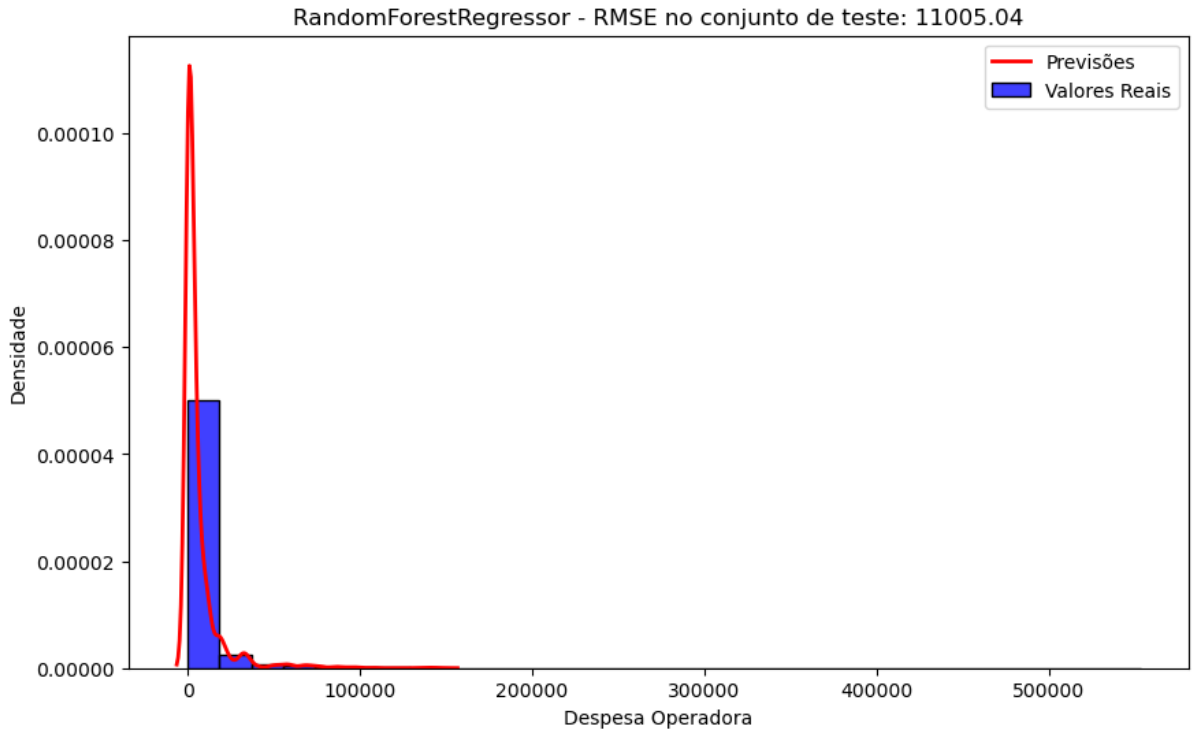
Fonte: Elaboração própria (2023).

O Gráfico 15 ilustra a aplicação do algoritmo Florestas Aleatórias (*Random Forest*) para prever as despesas assistenciais da OPS Alfa. No eixo horizontal, estão dispostos os valores das despesas assistenciais, variando de R$ 0,00 a R$ 60.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, variando entre 0,00000 e 0,00010.

Cinco colunas azuis decrescentes indicam os valores reais das despesas assistenciais, com destaque para um pico em R$ 10.000,00, apresentando uma densidade de 0,00010 no eixo vertical. A linha vermelha representa as previsões geradas pelo algoritmo de Florestas Aleatórias, apresentando uma curva que segue o formato das colunas que representam as despesas reais. Contudo, nota-se um pico na linha vermelha que atinge uma densidade de 0,00011 no eixo vertical.

Essa discrepância sugere que o modelo de Florestas Aleatórias identificou uma tendência semelhante às despesas reais, mas com algumas divergências, especialmente no ponto mais alto da curva. Avaliar métricas de desempenho do algoritmo, como erro absoluto médio ou validação cruzada, podem ajudar na compreensão acerca da precisão do modelo e determinar se ajustes são necessários para melhorar a qualidade das previsões.

Gráfico 15 – Florestas Aleatórias ou *Random Forest*



Fonte: Elaboração própria (2023).

Com o Gráficos 13, 14 e 15 das distribuições, foi possível então verificar que o modelo que melhor previu as despesas assistenciais da OPS Alfa foi o *Random Forest*, com menor nível de erro entre os modelos. Nesse sentido, este modelo de algoritmo preditivo fornecerá à OPS Alfa a possibilidade de gerenciar previamente os grupos de risco, aplicando então estratégias preventivas à saúde dos beneficiários, o que consequentemente refletirá na redução do valor das despesas assistenciais e sinistralidade da carteira da OPS Alfa.

Adicionalmente ao já exposto, a Tabela 9 a seguir demonstra em termos monetários o somatório das despesas reais e o somatório das despesas previstas, para mensurar e corroborar com o que fora estimado estatisticamente.

Tabela 9 - Mensuração das despesas

Modelos de algoritmos	**Despesa (R$)***	**Diferença (R$)***
Real	38.222.802,00	0
XGBoost	37.521.752,00	- 701.050,00
KNN	28.264.461,00	- 9.958.341,00
Random Forest	37.877.225,39	- 345.576,61 **

* Os valores financeiros não foram deflacionados.
Fonte: Elaboração Própria (2023).

Desse modo, em termos financeiros, o modelo que mais se aproximou da realidade foi o *Random Forest*, apresentando uma variação de **0,90% entre as despesas reais da OPS Alfa e a previsão do algoritmo, o que sustenta ainda mais a validação do resultado do modelo.

4.2.2 Operadora Beta

Os resultados da OPS Beta estão segmentados em: a) estatísticas descritivas; b) análise de *cluster* – grupos de risco e c) algoritmos de aprendizagem de máquina – análise preditiva.

a) Estatística descritiva

A Tabela 10 apresenta uma análise detalhada das variáveis relacionadas aos perfis sociodemográficos, indicadores econômicos e financeiros e indicadores de estado de saúde dos beneficiários da OPS Beta. A distribuição de gênero indica que 54.1% dos participantes são do sexo feminino, enquanto 45.9% são do sexo masculino.

Acerca da faixa etária, a média de idade da amostra é de 47.5 anos, com uma variação de 18 a 87 anos, apresentando uma diversidade de faixas etárias. A categoria de raça destaca uma predominância significativa de participantes brancos (70.6%). A amostra é totalmente localizada na região Nordeste, onde a OPS Beta está localizada. Quanto ao estado civil, a maioria é casada (49.0%).

A categoria econômico-financeira revela apresenta as informações sobre as despesas assistenciais, receitas e renda da carteira de beneficiários da OPS Beta. A média das despesas assistenciais é R$ 5.827,50, com uma variação significativa de até R$ 700.771,00. A receita média e a renda média são R$ 12.703,90 e R$ 34.658,80 respectivamente. Os dados abrangem um período de 2018 a 2023, com uma distribuição uniforme ao longo desses anos.

Os indicadores de saúde revelam que a grande maioria dos beneficiários da OPS Beta não possui câncer (90.6%), diabetes (88.3%) ou pressão alta (65.5%). Quanto à colesterol alto, a maioria também não apresenta esse quadro (69.2%), a maioria dos participantes não fuma (85.3%) e 60.0% realiza exames de rotina regularmente.

Tabela 10 - Sumário das variáveis OPS Beta

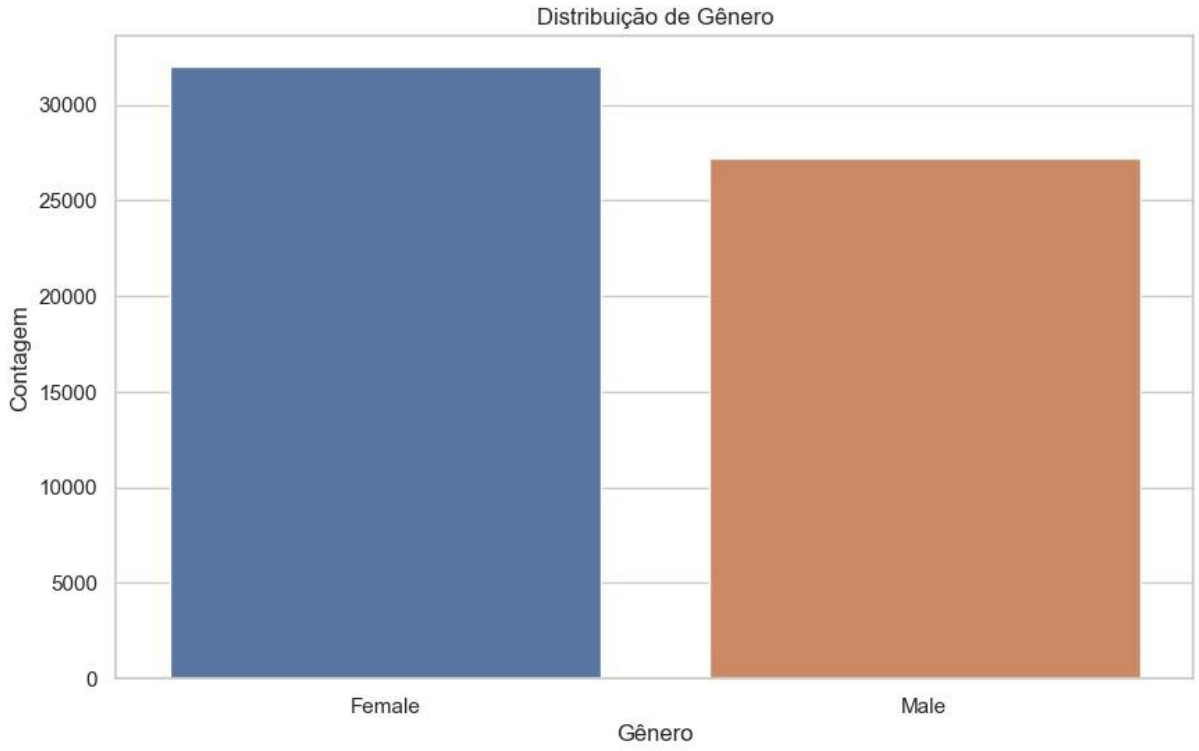
CATEGORIA	VARIÁVEL	ESTATÍSTICAS/VALORES	FREQUÊNCIA
PERFIL SOCIODEMOGRÁFICOS	1. Sexo	1. Feminino 2. Masculino	31982 (54.1%) 27188 (45.9%)
	2. Idade	Média (sd) : 47.5 (17.7) min < med < max: 18 < 47 < 87	68 valores distintos
	3. Raça	1.Branco 2.Negro 3.Pardo 4.Amarelo 5.Indígena 6.Multiplas raças	41788 (70.6%) 1318 (2.2%) 3137 (5.3%) 10763 (18.2%)

			1097 (1.9%) 1067 (1.8%)
	4. Região	1.Norte 2.Nordeste 3.Centro Oeste 4.Sudeste 5.Sul	2. 59170 (100)%
	5. Estado civil	1.Divorciado 2.Casado 3.Solteiro 4.Viúvo	8875 (15.0%) 28979 (49.0%) 17487 (29.6%) 3829 (6.5%)
ECONÔMICO E FINANCEIRO	6. Despesa_Operadora [assistencial]	Média (sd): 5827.5 (16354.4) min < med < max: 0 < 1194 < 700771 IQR (CV): 4687.8 (2.8)	15575 valores distintos
	7. Receita_Operadora	Média (sd): 12703.9 (51167.5) min < med < max: 0 < 1416 < 2542644 IQR (CV): 6671 (4)	18779 valores distintos
	8. Renda	Média (sd): 34658.8 (37820.7) min < med < max: 0 < 24000 < 409118 IQR (CV): 36671.2 (1.1)	19925 valores distintos
	9. Ano	Média (sd): 2020.5 (1.7) min < med < max: 2018 < 2021 < 2023 IQR (CV): 3 (0)	2018: 9860 (16.7%) 2019: 9861 (16.7%) 2020: 9861 (16.7%) 2021: 9861 (16.7%) 2022: 9861 (16.7%) 2023: 9866 (16.7%)
INDICADORES DE ESTADO DE SAÚDE	10. Câncer	1.Não 2.Sim	53614 (90.6%) 5556 (9.4%)
	11. Diabetes	1.Não 2.Sim	52276 (88.3%) 6894 (11.7%)
	12. Pressão_alta	1.Não 2.Sim	38729 (65.5%) 20441 (34.5%)
	13. Colesterol_alto	1.Não 2.Sim 3.Não diagnosticado	40923 (69.2%) 18231 (30.8%) 16 (0.0%)
	14. Fumante	1.Não 2.Sim	50486 (85.3%) 8684 (14.7%)
	15. Rotina_Exames	1.Não 2.Sim	23673 (40.0%) 35497 (60.0%)

Fonte: Elaboração Própria, (2023).

O Gráfico 16 ilustra a distribuição dos beneficiários conforme o gênero, destacando uma leve predominância do público feminino, que representa 54,1% do total, em comparação aos beneficiários masculinos, que compreendem 45,9%. Apesar da discrepância percentual, a distribuição equitativa entre os dois gêneros não evidencia um desequilíbrio significativo. No entanto, é importante observar que essa distribuição de gênero pode influenciar em distintos padrões de utilização dos serviços assistenciais, emergindo como um fator relevante a ser considerado na análise dos elementos que impactam as despesas assistenciais da OPS Beta.

Gráfico 16 – Distribuição de gênero

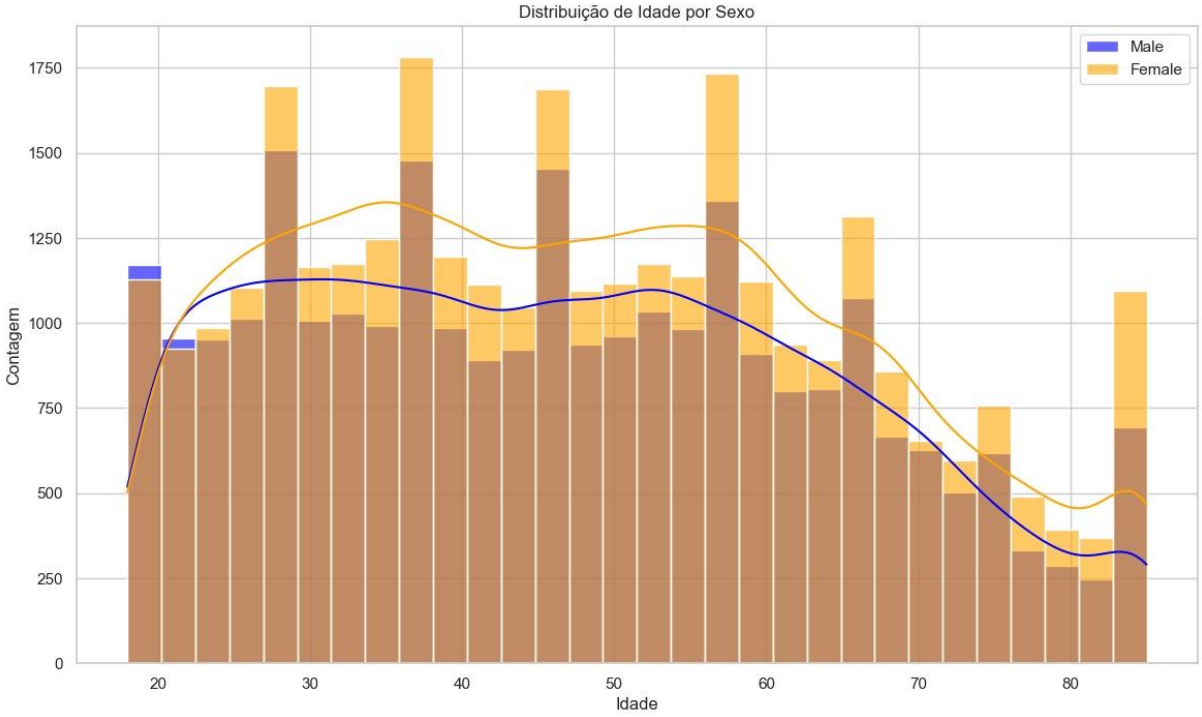


Fonte: Elaboração própria (2023).

A análise da distribuição da idade por gênero, apresentada no Gráfico 17, revela uma amplitude de idades entre 18 e 87 anos. Destaca-se uma concentração significativa de beneficiários nas faixas etárias de 28 a 58 anos, tanto para homens quanto para mulheres. Essa observação aponta para uma representação expressiva de adultos em meia-idade, um aspecto relevante que merece atenção, pois pode impactar as estratégias e políticas internas da OPS Beta voltadas para atender às necessidades específicas desse grupo etário, especialmente no que diz respeito a iniciativas de promoção à saúde.

Ao examinar o Gráfico 17, observa-se que as concentrações nas faixas etárias entre 28 e 58 anos podem sugerir padrões comportamentais distintos, desafios de saúde específicos ou mesmo influências socioculturais relevantes. Esses dados sugerem *insights* valiosos que poderiam ser explorados em estudos mais aprofundados, contribuindo para uma compreensão mais amplas dos fatores que moldam o perfil demográfico e de saúde dos beneficiários da OPS Beta.

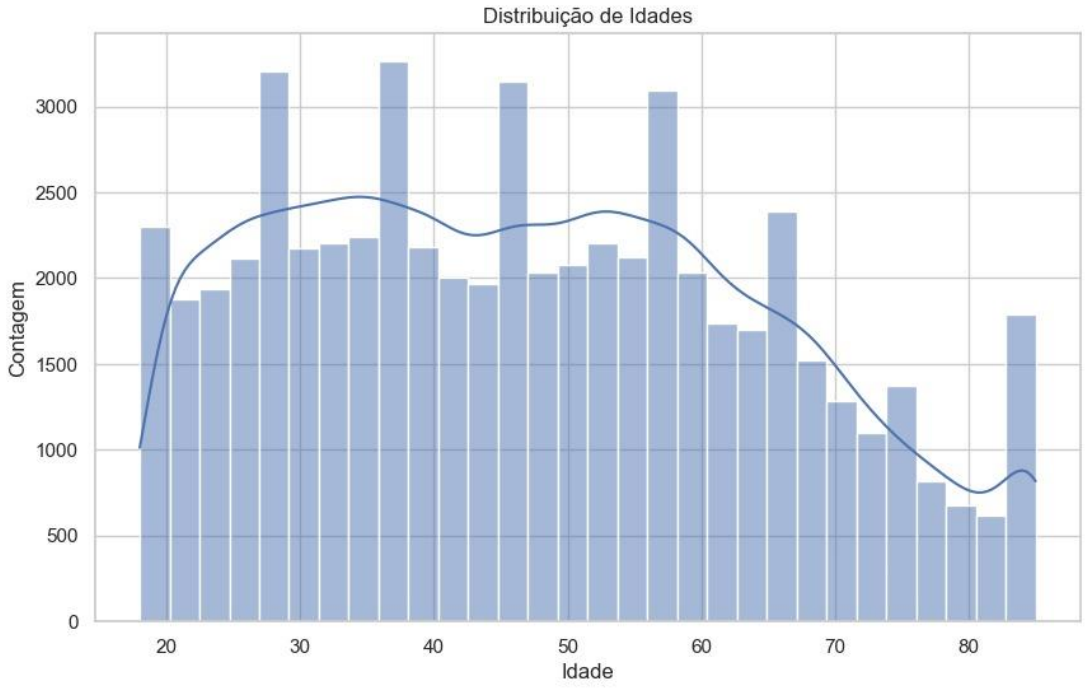
Gráfico 17 – Distribuição de idades por gênero



Fonte: Elaboração própria (2023).

O Gráfico 18 ilustra a distribuição por idade, indicando que a OPS Beta tem como público principal adultos em diversos estágios da vida, com uma significativa representação de beneficiários em faixas etárias consideradas maduras. A concentração notável de beneficiários na faixa etária entre 28 e 58 anos destaca a importância de considerar estratégias específicas na oferta de serviços de saúde preventiva e especializada, visando atender às demandas características dessa fase da vida. Essa observação oferece *insights* para o desenvolvimento de abordagens personalizadas e eficazes, alinhadas às necessidades predominantes dos beneficiários da OPS Beta.

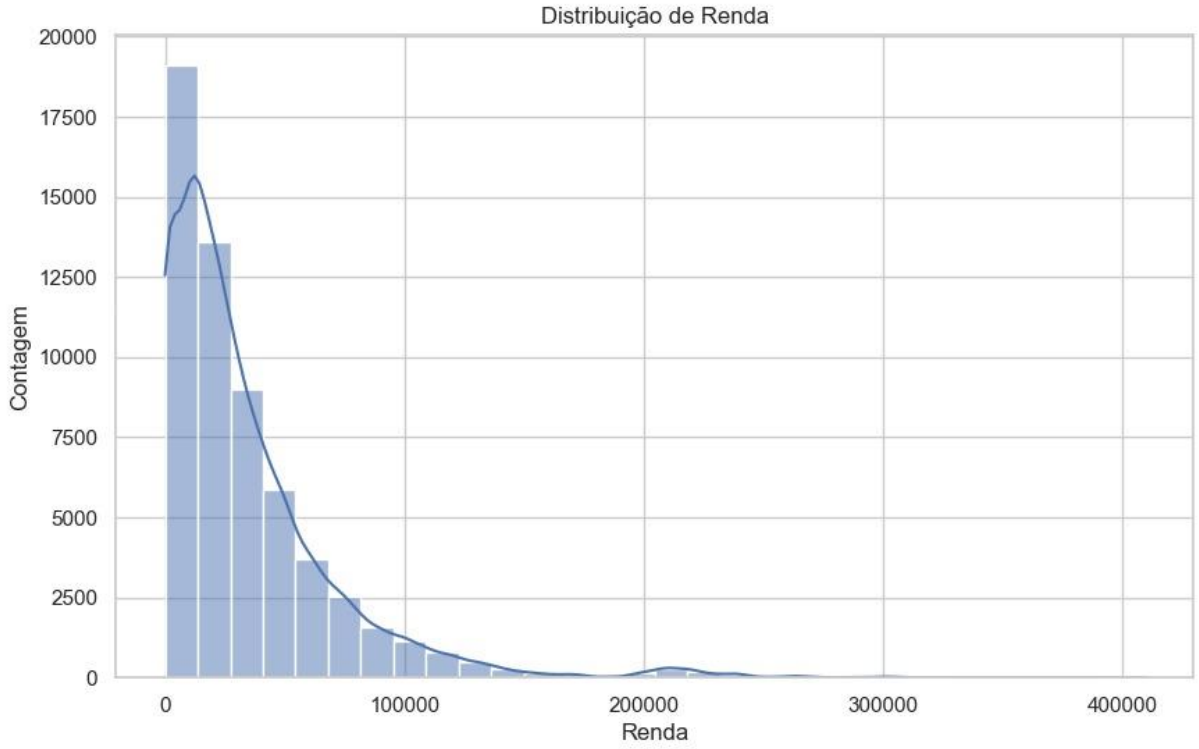
Gráfico 18 – Distribuição de idades



Fonte: Elaboração própria (2023).

O Gráfico 19 apresenta uma representação visual da probabilidade associada a diferentes faixas de renda entre os beneficiários da OPS Beta. No eixo horizontal, estão dispostos os intervalos de renda, enquanto o eixo vertical exibe a probabilidade associada a cada faixa. Destaca-se que a primeira faixa de renda concentra a maior probabilidade de representação dos beneficiários da OPS Beta, sugerindo uma distribuição significativa nesse intervalo específico. Essa visualização contribui para uma compreensão mais clara da distribuição de renda entre os beneficiários da OPS Beta.

Gráfico 19 – Distribuição de renda



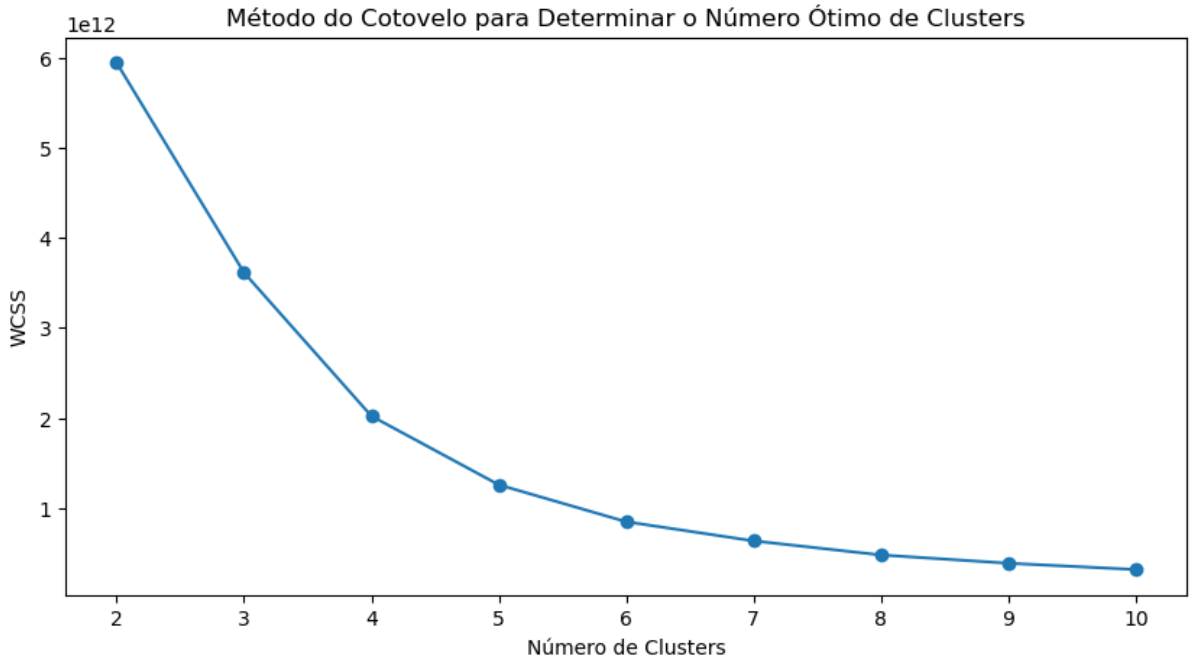
Fonte: Elaboração própria (2023).

b) Análise de *cluster* – grupos de risco

Antes da análise de *clusters*, empregou-se o método do cotovelo para determinar o número ideal de *clusters*, uma técnica que avalia a variância explicada em relação ao número de *cluster*s. O Gráfico 20, ilustra o eixo horizontal, que representa o número de *clusters* variando de 2 a 10, e o eixo vertical, que exibe a soma dos quadrados intra*cluster*s (WCSS), uma métrica indicativa da dispersão dos pontos dentro de cada *cluster*. A curva no gráfico representa a variação do WCSS em relação ao número de *clusters*, mostrando uma queda acentuada no início e diminuindo progressivamente à medida que o número de *clusters* aumenta. A identificação do "cotovelo" na curva, indicando a mudança na taxa de declínio, sugere que o número ótimo de *clusters* é 3.

Assim, conclui-se que a segmentação dos dados em três *cluster*s é a abordagem mais apropriada, proporcionando uma estrutura de agrupamento eficaz e representativa da variabilidade presente nos dados analisados. Esse resultado do método do cotovelo orienta a análise de *clusters*, contribuindo para uma interpretação mais precisa e significativa dos padrões subjacentes nos dados.

Gráfico 20 – Método do cotovelo para determinar o número ótimo de *clusters*

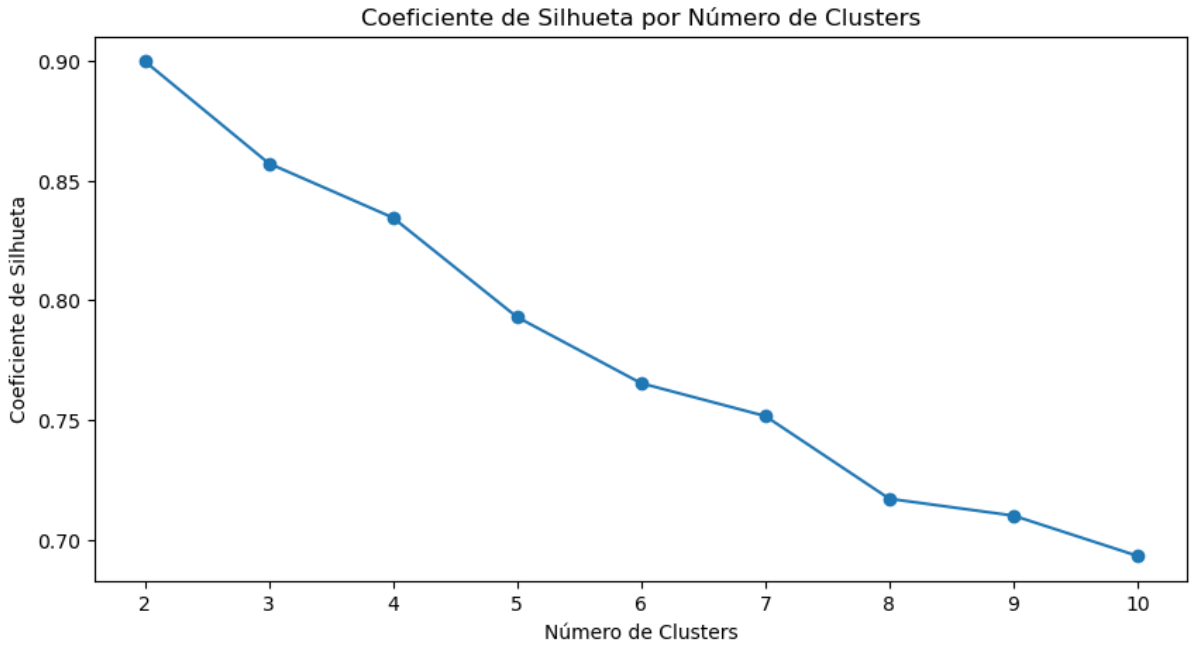


Fonte: Elaboração própria (2023).

O Gráfico 21 apresenta o Coeficiente de Silhueta em relação ao número de *clusters*, proporcionando sugestões sobre a coesão e qualidade dos *clusters* em uma análise de agrupamento. No eixo horizontal, encontramos o número de *cluster*s variando de 2 a 10, enquanto o eixo vertical exibe o coeficiente de silhueta, uma métrica que avalia a eficácia da separação dos objetos dentro de um *cluster* em relação aos outros *clusters*.

Ao analisar o Gráfico 21, destaca-se que a curva inicia em um ponto elevado no eixo vertical (0.90) e diminui gradualmente com o aumento do número de *clusters*. O coeficiente de silhueta em 0.90 indica uma boa coesão e separação dos objetos entre os *clusters*. Portanto, a presença de um valor inicial elevado sugere que a formação de três *clusters* é robusta e coesa, reforçando a escolha desse número como a configuração mais apropriada para a análise de agrupamento.

Gráfico 21 – Coeficiente de silhueta por número de *clusters*



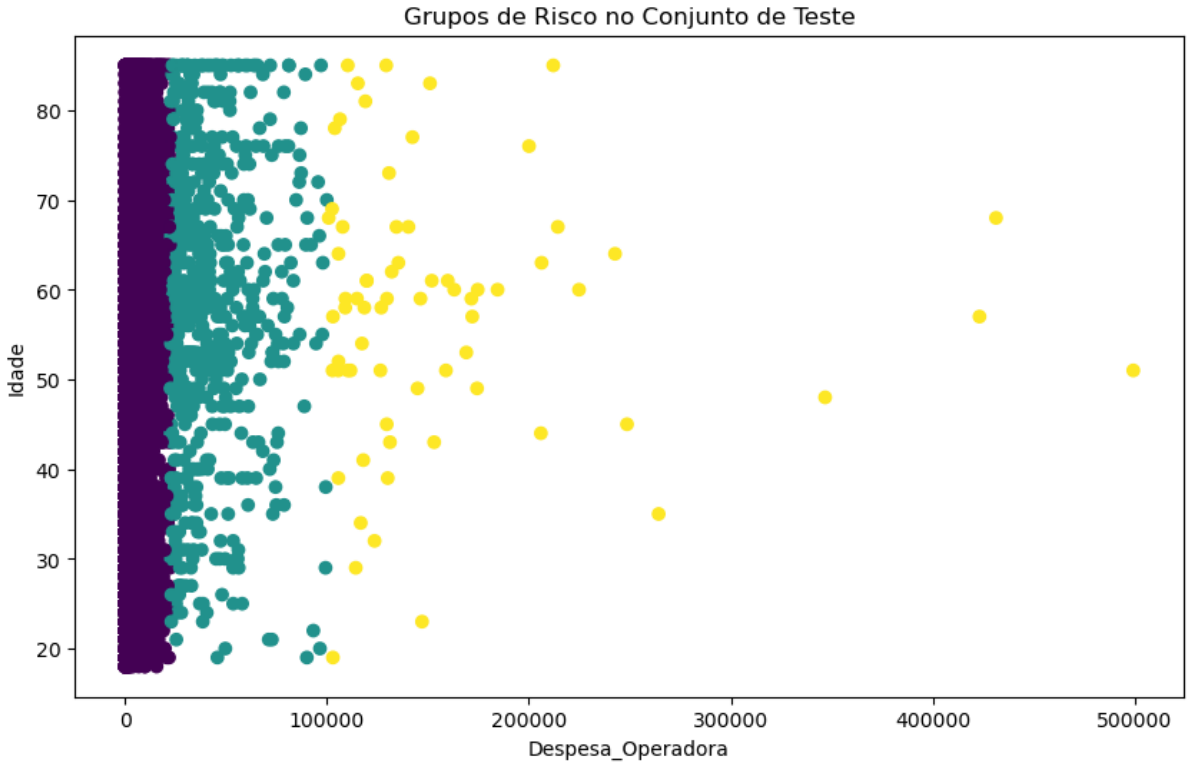
Fonte: Elaboração própria (2023).

Ao examinar o Gráfico 22 dos resultados do K-*means*, identificamos três grupos distintos indicados por cores diferentes. O *cluster* predominantemente roxo, que ocupa a maior parte do espaço do gráfico, parece consistir em indivíduos com despesas assistenciais mais baixas e que se distribuem de maneira uniforme por todas as faixas de idades. Este grupo pode ser interpretado como de "baixo risco financeiro", onde as despesas assistenciais são relativamente menores e não são influenciadas pela idade.

Os outros dois *clusters*, embora menos densos, representam grupos com despesas assistenciais mais relevantes. Estes grupos são diferenciados em termos de despesas assistenciais, mas sem uma variação ou correlação clara em relação à idade. O *cluster* amarelo, em particular, destaca-se por conter indivíduos com as maiores despesas assistenciais, possivelmente representando um "grupo de alto risco financeiro". Graficamente, este grupo inclui alguns pontos dispersos, sugerindo a presença de possíveis *outliers* ou casos extremos de despesas assistenciais.

Os dados apresentam uma distribuição homogênea de idades em todos os *cluster*s, reforçando a ideia de que a idade por si só não seria um fator determinante para o nível de despesas assistenciais, dentro desta amostra específica. A presença de *outliers*, especialmente no *cluster* de alto risco, pode ser um indicativo de situações atípicas que merecem uma investigação mais aprofundada.

Gráfico 22 – Grupos de risco no conjunto teste



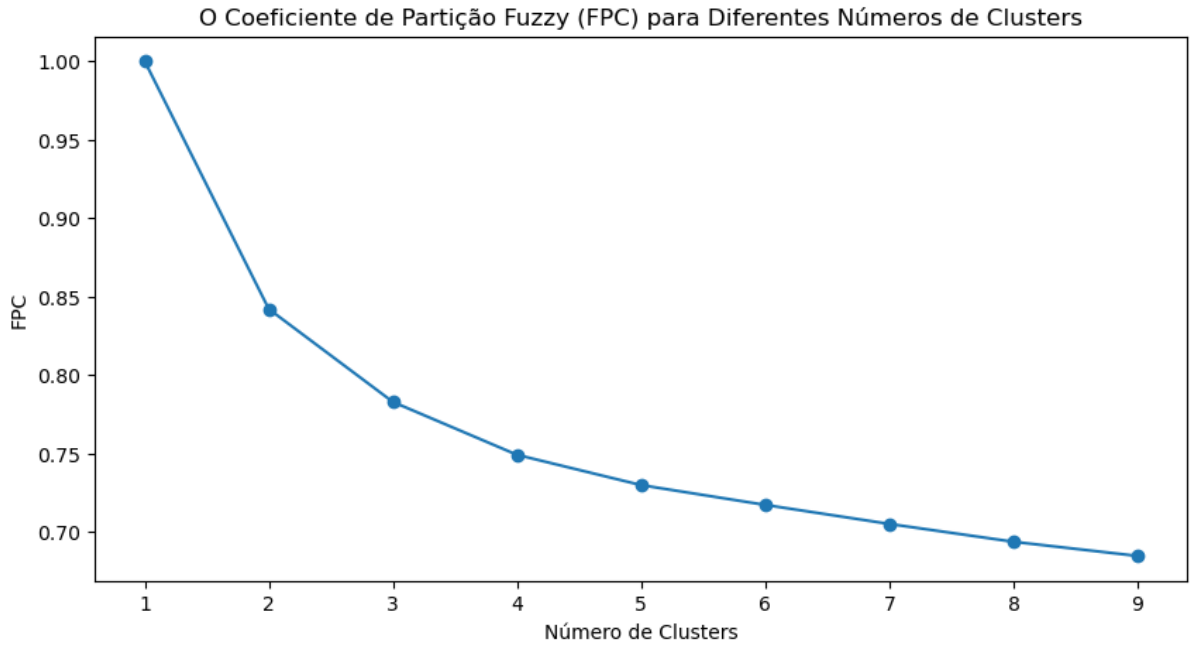
Fonte: Elaboração própria (2023).

Fuzzy C-Means

O Gráfico 23 exibe o Coeficiente de Partição F*uzzy* (FPC) em relação ao número de *clusters*, fornecendo sugestões sobre a adequação da partição *fuzzy* para diferentes configurações de agrupamento. O eixo horizontal representa o número de *clusters*, variando de 1 a 9, enquanto o eixo vertical mostra o FPC, uma medida que avalia a qualidade da partição *fuzzy*, indicando quão bem os dados estão distribuídos entre os *clusters*.

Ao analisar o gráfico, percebe-se que a curva inicia em um ponto relativamente alto no eixo vertical (1) e diminui à medida que o número de *cluster*s aumenta. Um FPC próximo a 1 sugere uma partição *fuzzy* mais precisa e eficaz. Dessa forma, a busca ideal é por um número de *clusters* que maximize o FPC, indicando uma partição *fuzzy* que melhor represente a estrutura dos dados.

Gráfico 23 – Coeficiente de partição Fuzzy para diferentes números de *clusters*



Fonte: Elaboração própria (2023).

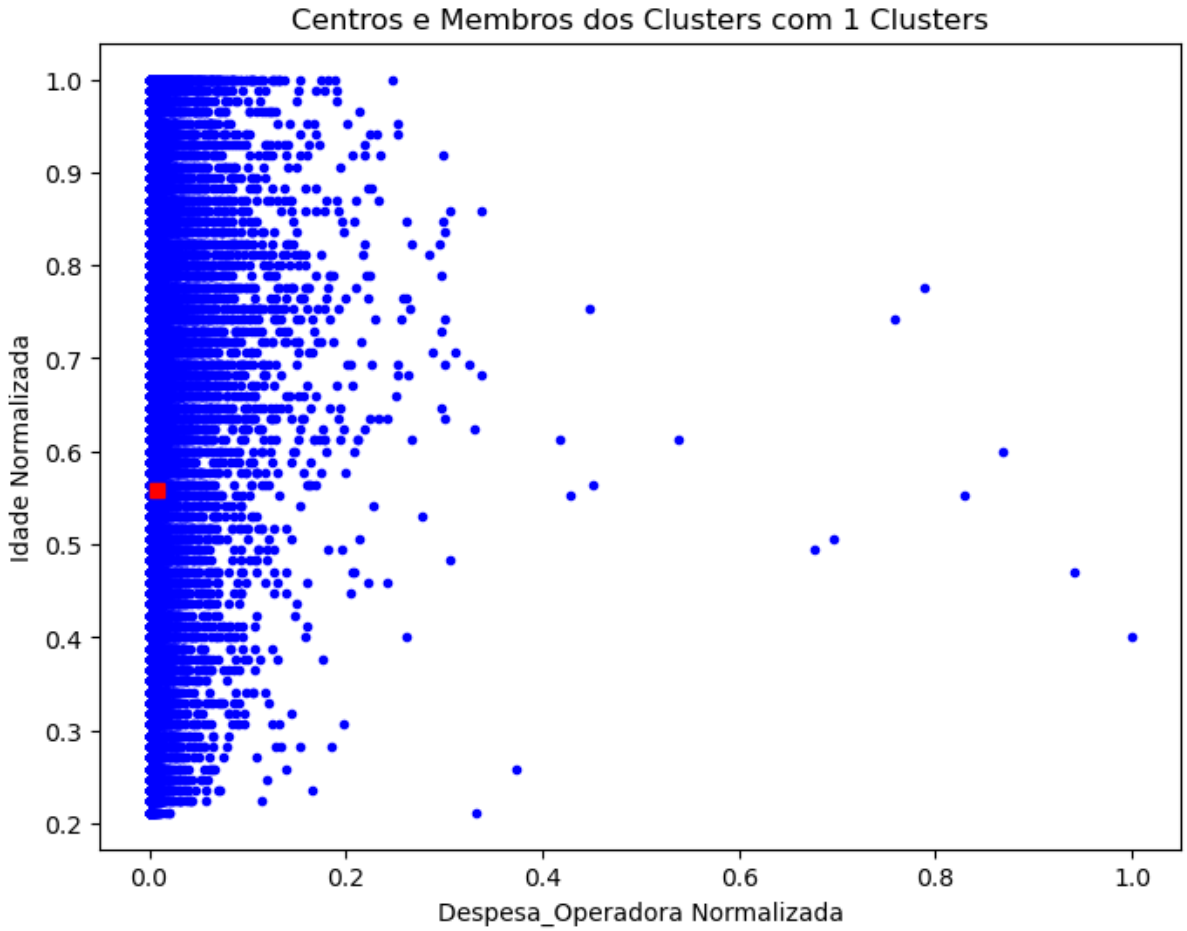
Observa-se um único *cluster* (dado que o número ótimo de *clusters* foi identificado como 1, pelo critério utilizado), indicado pelos pontos azuis, com um centro de *cluster* marcado por um ponto vermelho. A distribuição dos pontos sugere uma alta densidade de dados com baixas despesas assistenciais e uma ampla dispersão na faixa de idades, sendo indicado pela concentração de pontos no lado esquerdo do Gráfico 24.

Além disso, há uma dispersão menos densa de pontos ao longo do eixo das despesas assistenciais, com poucos dados mostrando despesas maiores, mas sem um padrão claro de diferenciação em sua relação com as faixas de idade, como é visto pela distribuição vertical dos pontos ao longo do eixo da idade.

A presença de um único *cluster* pode indicar que o conjunto de dados não possui uma variação significativa que justifique a separação em múltiplos grupos ou pode sinalizar que o critério utilizado para determinar o número de *clusters* não foi sensível o suficiente para detectar agrupamentos mais sutis.

O ponto central em vermelho representa o centroide do *cluster* único, que é um resumo das características médias do conjunto de dados baseado na lógica *fuzzy*, refletindo uma média ponderada das despesas assistenciais e idades.

Gráfico 24 – Centros e membros dos *clusters* com 1 *cluster*



Fonte: Elaboração própria (2023).

c) Algoritmos de aprendizagem de máquina – análise preditiva

Nesta seção, são apresentados os modelos de aprendizado de máquina utilizados para atingir os objetivos desta tese. Inicialmente, a Tabela 11 expõe a acurácia dos modelos na base de dados da OPS Beta.

Os gráficos desta seção exibem a distribuição das previsões realizadas por três modelos distintos nos dados da OPS Beta, destacando também o RMSE (*Root Mean Square Error*) para cada modelo.

Tabela 11 - Acurácia dos modelos

Random Forest	XGBOOST	KNN
RMSE	RMSE	RMSE
11556,04	11052,64	16001,27

Fonte: Elaboração própria (2023).

A Tabela 11 demonstra que o modelo KNN apresentou o maior RMSE na base da OPS Beta, o que indica que este modelo teve o pior desempenho entre os três. Com um RMSE de 16001,27, sugere que a estratégia baseada em k-vizinhos mais próximos não consegue capturar eficientemente a complexidade dos dados ou pode estar sujeita ao desafio da "maldição da dimensionalidade", especialmente em casos de bases de dados com alta dimensionalidade.

O modelo XGBoost obteve um RMSE na base da OPS Beta de 11556,04. Apesar de ter superado o desempenho do modelo KNN, o XGBoost não alcançou a mesma eficácia do modelo *Random Forest*. O XGBoost é reconhecido por sua habilidade em modelar interações complexas e não lineares. Contudo, é possível que o modelo não tenha sido totalmente otimizado para os dados em questão, ou que a natureza dos dados não favoreça plenamente suas características distintivas.

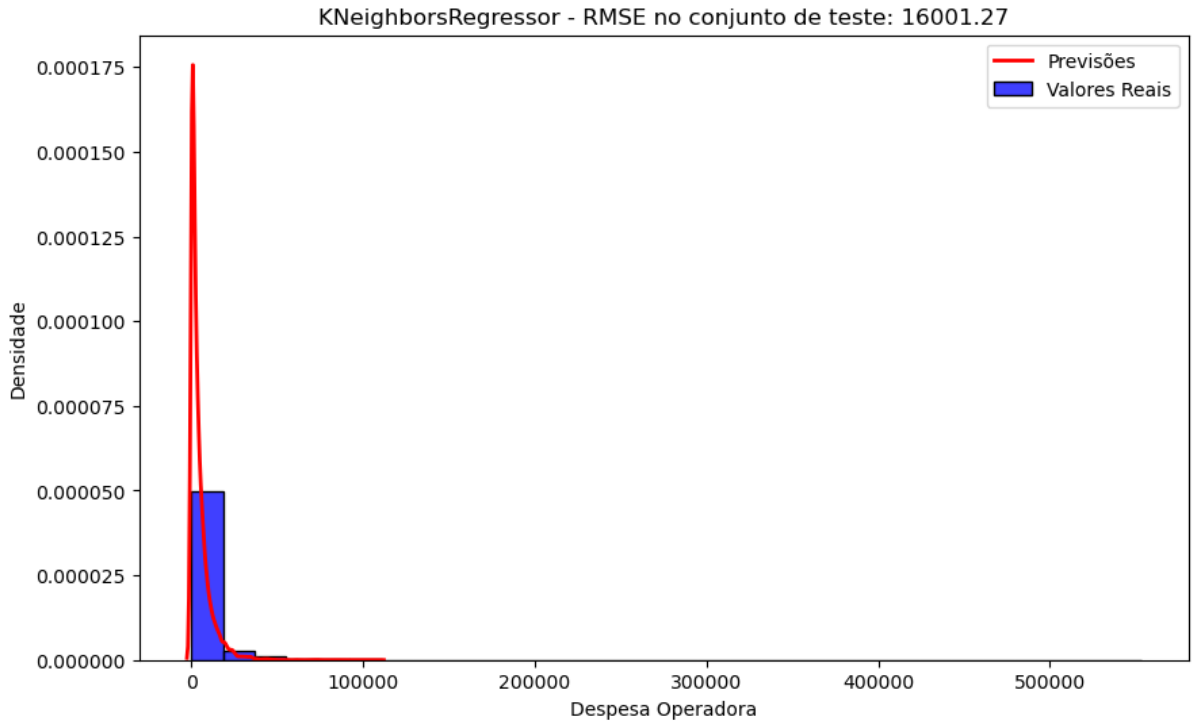
O *Random Forest* obteve o melhor desempenho com um RMSE de 11052,64. Isso indica que o modelo foi capaz de compreender de maneira mais precisa a complexidade dos dados em comparação com os outros modelos. As *Random Forests* demonstram eficácia ao lidar com variáveis tanto categóricas quanto numéricas, sendo capazes de capturar interações não lineares nos dados.

A partir destes resultados, parece que o modelo *Random Forest Regressor* seria a melhor escolha entre os três para prever a "Despesa assistencial da OPS Beta" neste conjunto de dados específico. No entanto, todos os modelos parecem ter dificuldade em prever com precisão os valores mais altos de despesa, sugerido pelo gráfico de dispersão que indicam um pico no extremo inferior da escala de despesa.

O Gráfico 25 ilustra a aplicação do algoritmo K-vizinhos próximos (KNN) na previsão das despesas assistenciais da OPS Beta. No eixo horizontal, encontram-se os valores das despesas assistenciais, variando de R$ 0,00 a R$ 500.000,00, enquanto o eixo vertical exibe a densidade associada a cada valor. As previsões geradas pelo algoritmo KNN, representadas pela linha vermelha, demonstram uma conformidade geral em descompasso com os dados reais das despesas assistenciais, conforme evidenciado pela discrepância de formato entre a curva prevista e as colunas que representam as despesas reais.

Essa discrepância sugere que o algoritmo KNN, ao realizar previsões, identificou áreas de menor precisão, sugerindo a necessidade de avaliação adicional do desempenho do algoritmo, por meio de métricas de erro ou validação cruzada, visando aprimorar a confiabilidade das previsões.

Gráfico 25 – K-vizinhos ou KNN



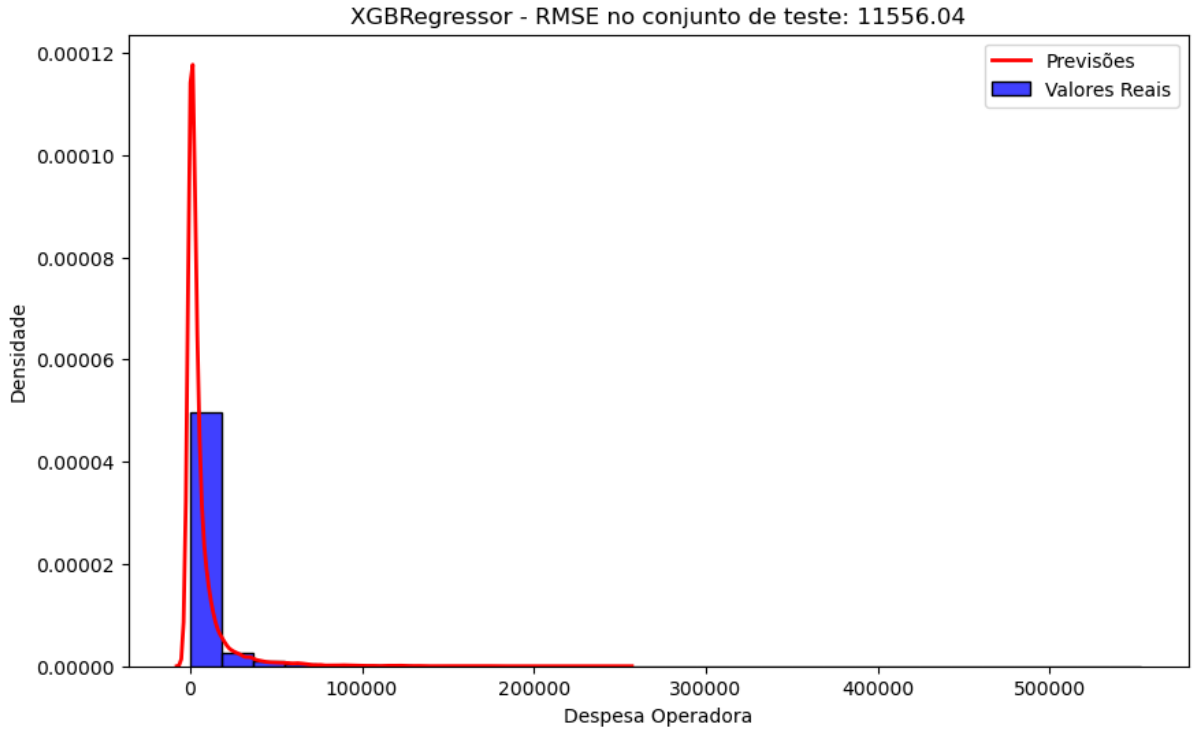
Fonte: Elaboração própria (2023).

O Gráfico 26 exemplifica a aplicação do algoritmo XGBoost na previsão das despesas assistenciais da OPS Beta. No eixo horizontal, são apresentados os valores das despesas assistenciais, variando de R$ 0,00 a R$ 500.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, variando entre 0,00000 e 0,00012. Cinco colunas azuis decrescentes indicam os valores reais das despesas assistenciais, proporcionando uma referência visual.

A linha vermelha traça as previsões geradas pelo algoritmo XGBoost, apresentando uma curva que segue o formato das colunas que representam as despesas reais. Contudo, um pico na linha vermelha atinge uma densidade de 0,00012, indicando uma discrepância notável. Essa observação sugere que o algoritmo XGBoost identificou uma tendência semelhante às despesas reais, mas apenas até a densidade 0,00006.

Para uma avaliação mais aprofundada da precisão do modelo, seria prudente examinar métricas de desempenho, como o erro absoluto médio ou a implementação de validação cruzada. Essas análises adicionais podem fornecer sugestões sobre o desempenho do algoritmo XGBoost, permitindo ajustes necessários para aprimorar a qualidade das previsões e garantir uma representação mais fiel das despesas assistenciais.

Gráfico 26 – XGBoost



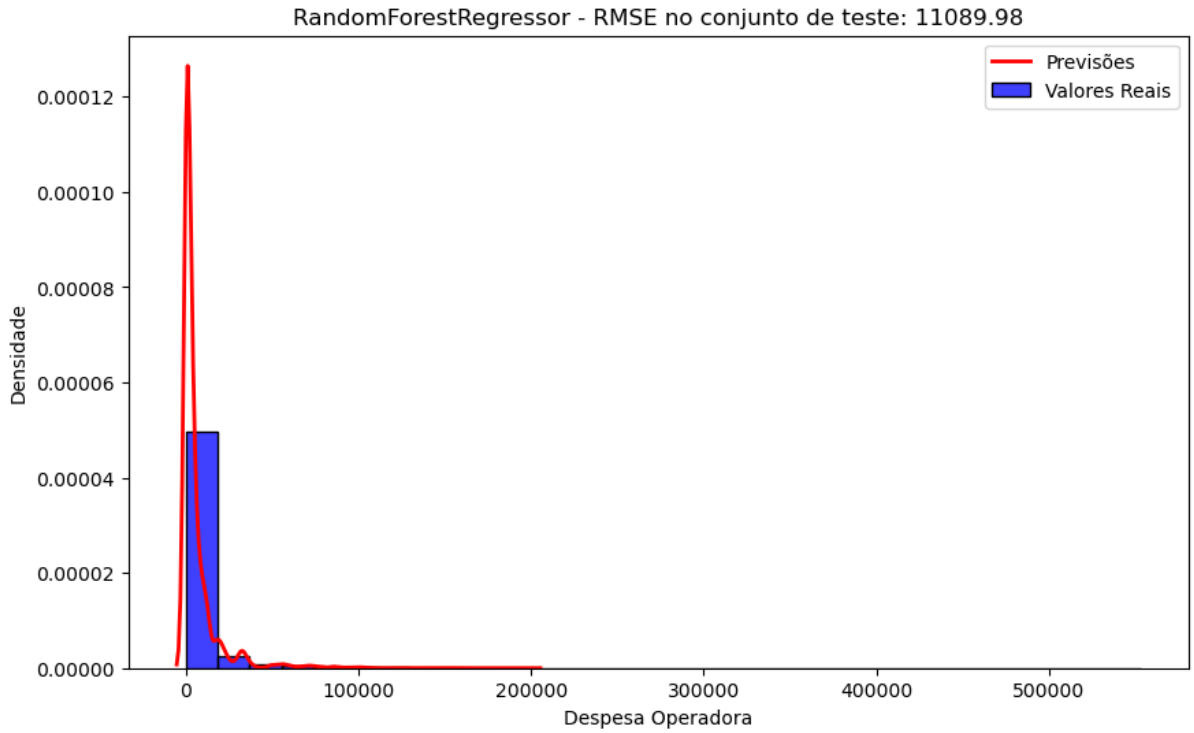
Fonte: Elaboração própria (2023).

O Gráfico 27 apresenta a aplicação do algoritmo de Florestas Aleatórias *(Random Forest*) na previsão das despesas assistenciais da OPS Beta. No eixo horizontal, são exibidos os valores das despesas assistenciais, variando de R$ 0,00 a R$ 600.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, variando entre 0,00000 e 0,00010.

A linha vermelha representa as previsões geradas pelo algoritmo de Florestas Aleatórias, exibindo uma curva que acompanha o formato das colunas que representam as despesas reais. No entanto, observa-se na altura da densidade 0,00005 uma discrepância notável. Essa observação sugere que o modelo de Florestas Aleatórias identificou uma tendência semelhante às despesas reais, mas com algumas divergências, especialmente no ponto mais alto da curva.

Para uma avaliação mais aprofundada da precisão do modelo, seria prudente examinar métricas de desempenho, como o erro absoluto médio ou a implementação de validação cruzada. Essas análises adicionais podem fornecer *insights* sobre o desempenho do algoritmo de Florestas Aleatórias, permitindo ajustes necessários para aprimorar a qualidade das previsões e garantir uma representação mais fiel das despesas assistenciais.

Gráfico 27 – Florestas aleatórias ou *Random Forest*



Fonte: Elaboração própria (2023).

Com base nas distribuições apresentadas nos Gráficos 25, 26 e 27 fica evidente que o modelo que melhor antecipou as despesas assistenciais da OPS Beta foi o *Random Forest*, destacando-se por apresentar o menor nível de erro em comparação com os demais modelos. Essa eficácia posiciona o algoritmo de Florestas Aleatórias como uma ferramenta para a OPS Beta, permitindo a antecipação e gestão proativa de grupos de risco. A implementação de estratégias preventivas de saúde direcionadas aos beneficiários tem o potencial de resultar na redução do montante das despesas assistenciais e da sinistralidade da carteira da OPS Beta.

Além disso, a Tabela 12, apresentada a seguir, detalha em termos monetários a soma das despesas reais e a soma das despesas previstas. Essa abordagem não apenas respalda estatisticamente as estimativas realizadas, mas também fornece uma visão concreta da concordância entre as previsões dos modelos e os valores reais das despesas assistenciais.

Tabela 12 - Mensuração das despesas

Modelos de algoritmos	**Despesa (R$)***	**Diferença (R$)***
Real	126.344.840,00	0
XGBoost	125.149.968,00	- 1.194.872,00
KNN	91.979.733,42	- 34.365.106,57
Random Forest	125.144.996,59	- 1.199.843,40 **

* Os valores financeiros não foram deflacionados.
Fonte: Elaboração Própria (2023).

Desse modo, em termos financeiros, o modelo que mais se aproximou da realidade foi o *XGBoost*, apresentando uma variação de 0,94% entre as despesas reais da OPS Beta e a previsão do algoritmo, o que sustenta ainda mais a validação do resultado do modelo.

4.2.3 Operadora Gama

Os resultados da OPS Gama estão segmentados em: a) estatísticas descritivas; b) análise de *cluster* – grupos de risco e c) algoritmos de aprendizagem de máquina – análise preditiva.

a) Estatísticas descritivas

A Tabela 13 proporciona uma análise pormenorizada das variáveis relacionadas aos perfis sociodemográficos, indicadores econômicos e financeiros, bem como indicadores de estado de saúde dos beneficiários da OPS Gama. A distribuição de gênero destaca que 56.5% dos participantes são do sexo feminino, enquanto 43.5% são do sexo masculino. No tocante à faixa etária, a média é de 49.5 anos, variando de 18 a 98 anos, demonstrando uma ampla diversidade de idades. A categoria de raça indica uma significativa predominância de participantes autodeclarados como brancos (74.6%). É relevante notar que a amostra está integralmente concentrada na região Sudeste, onde a OPS Gama está situada. Em relação ao estado civil, a maioria dos beneficiários é casada (52.0%).

No âmbito econômico-financeiro, são apresentadas informações cruciais sobre as despesas assistenciais, receitas e renda da carteira de beneficiários da OPS Gama. A média das despesas assistenciais é de R$ 7.532,50, com uma notável variação de até R$ 722.533,00. Quanto à receita média e à renda média, essas são de R$ 13.150,90 e R$ 29.854,80, respectivamente. Os dados abrangem um período de 2018 a 2023, apresentando uma distribuição equitativa ao longo desses anos.

Os indicadores de saúde revelam que a grande maioria dos beneficiários da OPS Gama não apresenta diagnóstico de câncer (84.4%), diabetes (85.1%) ou pressão alta (68.8%). Em relação ao colesterol alto, a maioria também não registra esse quadro (62.6%). Adicionalmente, a maior parte dos participantes não é fumante (84.1%), e 63.5% realiza exames de rotina regularmente. Essas informações delineiam um perfil abrangente do estado de saúde da população beneficiária da OPS Gama.

Tabela 13 - Sumário das variáveis OPS Gama

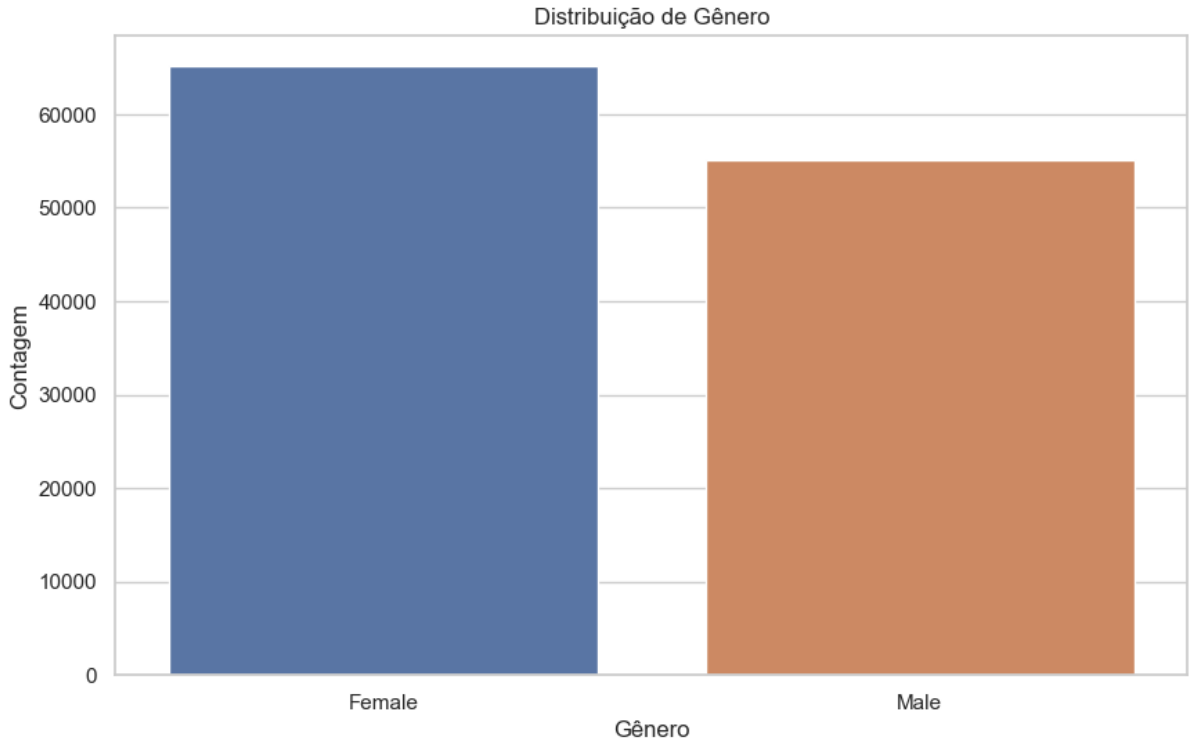
CATEGORIA	VARIÁVEL	ESTATÍSTICAS/VALORES	FREQUÊNCIA
PERFIL SOCIODEMOGRÁFICOS	1. Sexo	1. Feminino 2. Masculino	65134 (56.5%) 55175 (43.5%)
	2. Idade	Média (sd): 49.5 (17.7) min < med < max: 18 < 49 < 98	70 valores distintos
	3. Raça	1.Branco 2.Negro 3.Pardo 4.Amarelo 5.Indígena 6.Multiplas raças	89750 (74.6%) 2647 (2.2%) 6377 (5.3%) 17084 (14.2%) 2286 (1.9%) 2166 (1.8%)
	4. Região	1.Norte 2.Nordeste 3.Centro Oeste 4.Sudeste 5.Sul	4. 120309 (100%)
	5. Estado civil	1.Divorciado 2.Casado 3.Solteiro 4.Viúvo	18044 (13.0%) 58947 (52.0%) 35533 (29.5%) 7785 (6.5%)
ECONÔMICO E FINANCEIRO	6. Despesa_Operadora [assistencial]	Média (sd): 7532.3 (16399.7) min < med < max: 0 < 1194 < 722533 IQR (CV): 4692 (2.8)	15957 valores distintos
	7. Receita_Operadora	Média (sd): 13150.5 (51034.2) min < med < max: 0 < 1417 < 2542644 IQR (CV): 6674 (4)	18985 valores distintos
	8. Renda	Média (sd): 29854.8 (37788.5) min < med < max: 0 < 24000 < 409118 IQR (CV): 36664 (1.1)	32584 valores distintos
	9. Ano	Média (sd): 2020.5 (1.7) min < med < max: 2018 < 2021 < 2023 IQR (CV): 3 (0)	2018: 20250 (16.8%) 2019: 19851 (16.5%) 2020: 20051 (16.7%) 2021: 20087 (16.7%) 2022: 20015 (16.6%) 2023: 20055 (16.7%)
INDICADORES DE ESTADO DE SAÚDE	10. Câncer	1.Não 2.Sim	101614 (84.4%) 18695 (15.6%)
	11. Diabetes	1.Não 2.Sim	102382 (85.1%) 17926 (14.9%)
	12. Pressão_alta	1.Não 2.Sim	73194 (60.8%) 47115 (39.2%)
	13. Colesterol_alto	1.Não 2.Sim 3.Não diagnosticado	75365 (62.6%) 44923 (37.4%) 32 (0.0%)
	14. Fumante	1.Não 2.Sim	101174 (84.1%) 19135 (15.9%)
	15. Rotina_Exames	1.Não 2.Sim	43940 (36.5%) 76369 (63.5%)

Fonte: Elaboração Própria, (2023).

O Gráfico 28 apresenta a distribuição dos beneficiários conforme o gênero, evidenciando uma predominância do público feminino, representando 56,5% do total, em

comparação aos beneficiários masculinos, que compreendem 43,5%. Apesar da discrepância percentual, a distribuição equitativa entre os dois gêneros não evidencia um desequilíbrio significativo. No entanto, é importante observar que essa distribuição de gênero pode influenciar distintos padrões de utilização dos serviços assistenciais, emergindo como um fator relevante a ser considerado na análise dos elementos que impactam as despesas assistenciais da OPS Gama.

Gráfico 28 – Distribuição de gênero



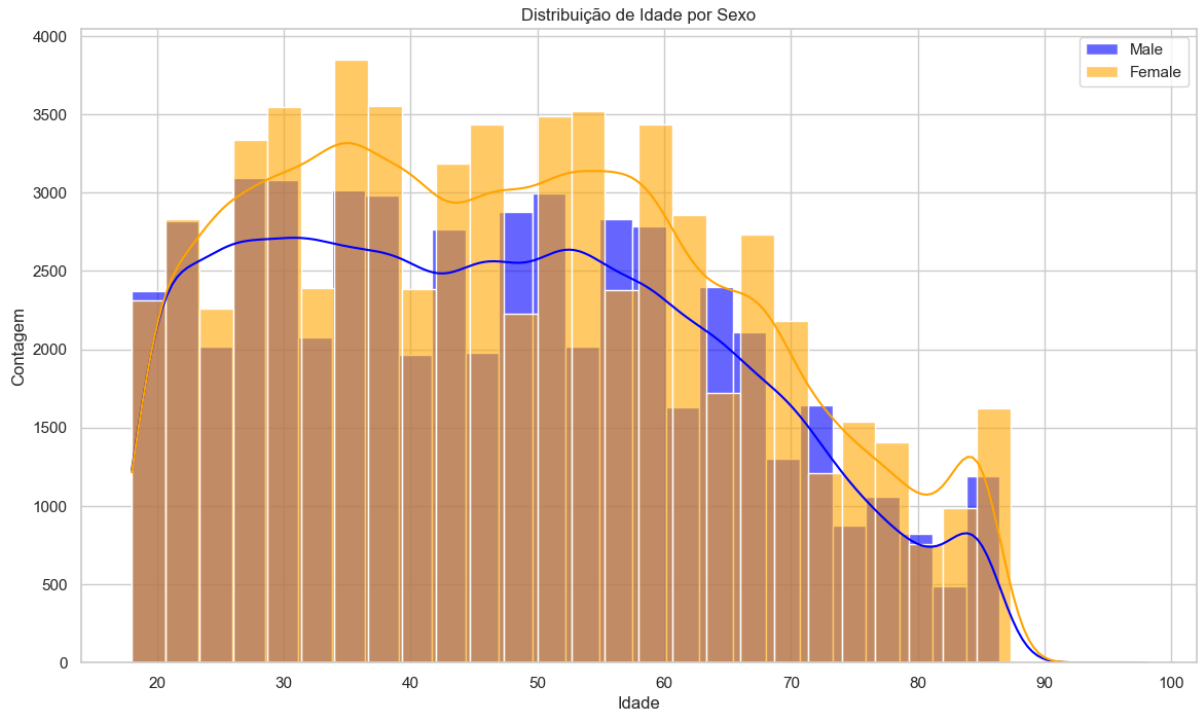
Fonte: Elaboração própria (2023).

A análise da distribuição da idade por gênero, apresentada no Gráfico 29, revela uma amplitude de idades entre 18 e 98 anos. Destaca-se uma concentração significativa de beneficiários nas faixas etárias de 28 a 60 anos, tanto para homens quanto para mulheres. Essa observação aponta para uma representação expressiva de adultos em meia-idade, um aspecto relevante que merece atenção, pois pode impactar as estratégias e políticas internas da OPS Gama voltadas para atender às necessidades específicas desse grupo etário, especialmente no que diz respeito a iniciativas de promoção à saúde.

Ao examinar o Gráfico 29, observa-se que as concentrações nas faixas etárias entre 28 e 60 anos podem sugerir padrões comportamentais distintos, desafios de saúde específicos ou mesmo influências socioculturais relevantes. Esses dados sugerem indicativos que poderiam

ser explorados em estudos mais aprofundados, contribuindo para uma compreensão mais ampla dos fatores que moldam o perfil demográfico e de saúde dos beneficiários da OPS Gama.

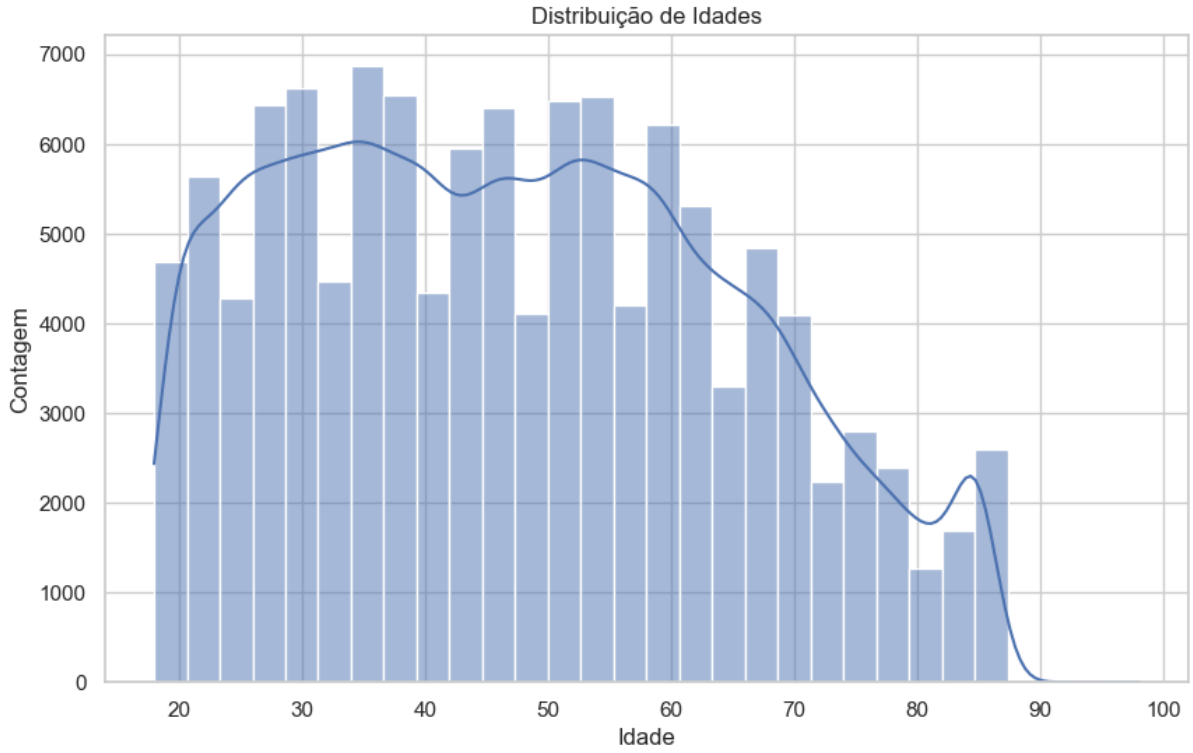
Gráfico 29 – Distribuição de idades por gênero



Fonte: Elaboração própria (2023).

O Gráfico 30 apresenta a distribuição etária, evidenciando que a OPS Gama tem como seu público principal adultos em variadas fases da vida, com uma expressiva representação de beneficiários em faixas etárias consideradas maduras. A notável concentração de beneficiários na faixa etária entre 28 e 60 anos destaca a relevância de considerar estratégias específicas na oferta de serviços de saúde preventiva e especializada, com o intuito de atender às demandas características dessa etapa da vida. Tal observação fornece perspicácia para a formulação de abordagens personalizadas e eficientes, alinhadas às necessidades predominantes dos beneficiários da OPS Gama.

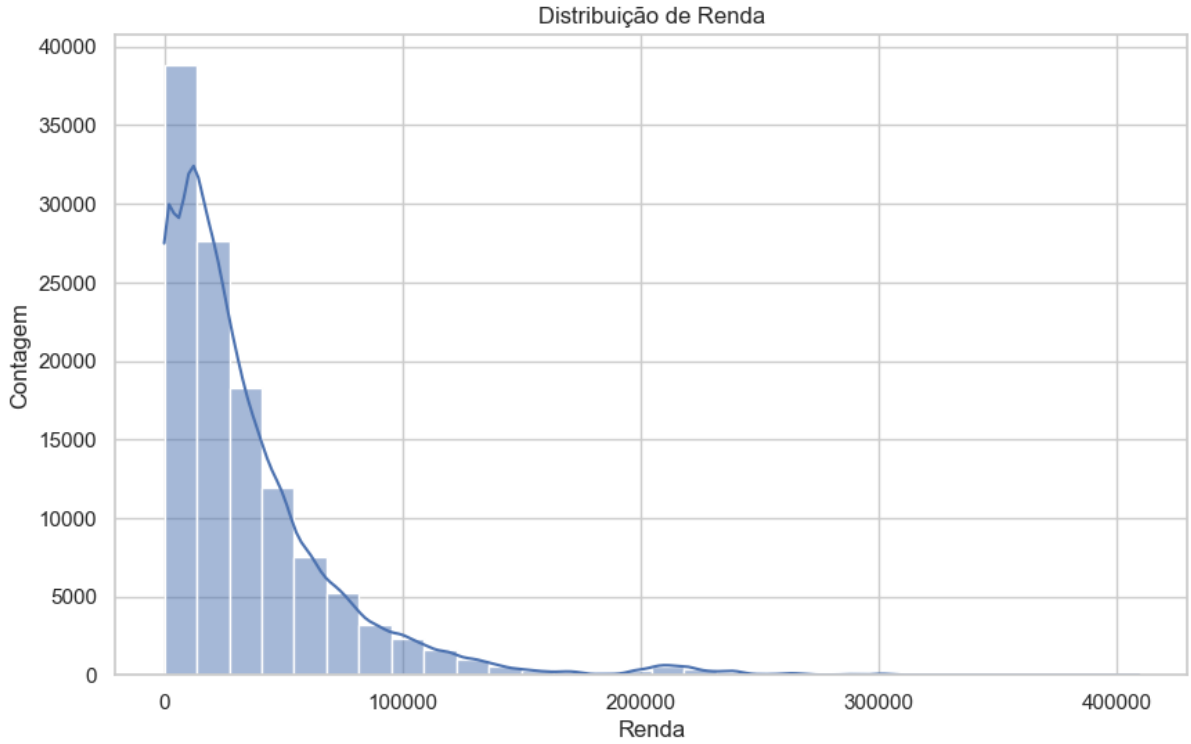
Gráfico 30 – Distribuição de idades



Fonte: Elaboração própria (2023).

O Gráfico 31 oferece uma representação gráfica da probabilidade associada a distintas faixas de renda entre os beneficiários da OPS Gama. No eixo horizontal, os intervalos de renda são explicitamente apresentados, enquanto o eixo vertical exibe a probabilidade relacionada a cada faixa. É relevante destacar que a primeira faixa de renda é caracterizada pela maior probabilidade de representação dos beneficiários da OPS Gama, indicando uma distribuição significativa nesse intervalo específico. Esta representação visual contribui substancialmente para uma compreensão mais precisa da distribuição de renda entre os beneficiários da OPS Gama, promovendo uma análise visual e interpretativa dos padrões associados.

Gráfico 31 – Distribuição de renda



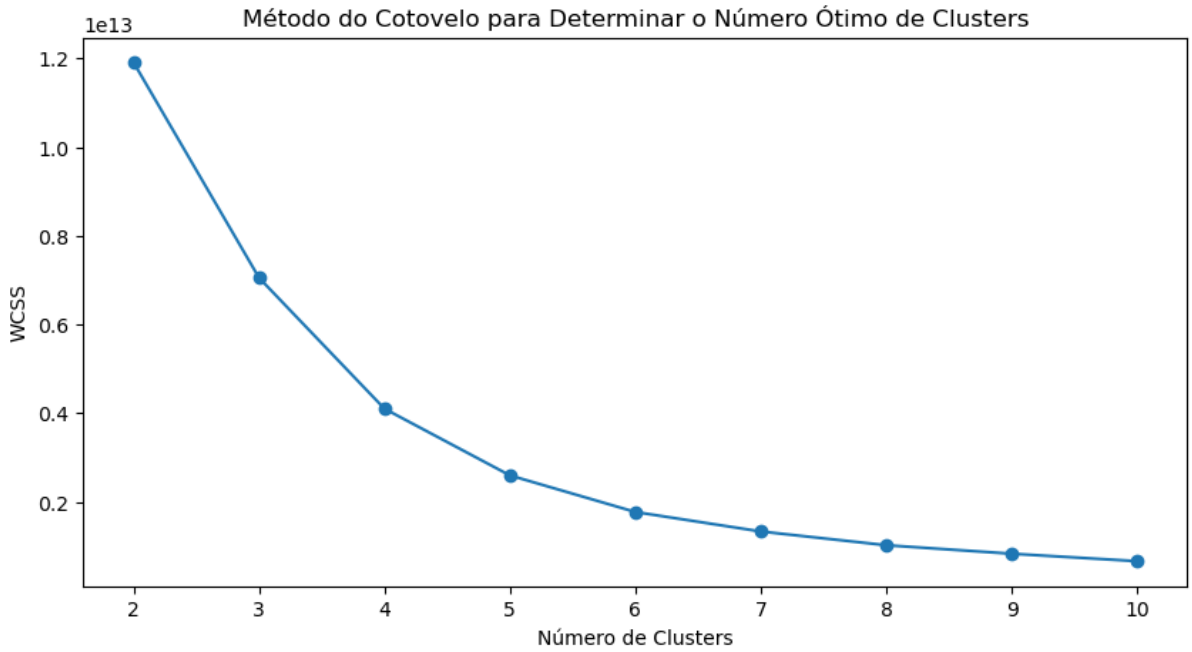
Fonte: Elaboração própria (2023).

b) Análise de *cluster* – grupos de risco

Antes da realização da análise de *clusters*, foi aplicado o método do cotovelo para determinar o número ideal de *clusters*, uma técnica que avalia a variância explicada em função do número de *clusters*. O Gráfico 32, exibido a seguir, apresenta o eixo horizontal representando o número de *cluster*s, variando de 2 a 10, enquanto o eixo vertical exibe a soma dos quadrados intra*cluster*s (WCSS), uma métrica que reflete a dispersão dos pontos dentro de cada *cluster*. A curva no gráfico representa a variação do WCSS conforme o número de *clusters*, evidenciando uma queda acentuada inicial e uma diminuição progressiva à medida que o número de *cluster*s aumenta. A identificação do "cotovelo" na curva, indicando a alteração na taxa de declínio, sugere que o número ótimo de *clusters* é 3.

Portanto, conclui-se que a segmentação dos dados em três *clusters* é a abordagem mais apropriada, proporcionando uma estrutura de agrupamento eficaz e representativa da variabilidade presente nos dados analisados. Este resultado obtido pelo método do cotovelo orienta a análise de *cluster*s, contribuindo para uma interpretação mais precisa e significativa dos padrões subjacentes nos dados.

Gráfico 32 – Método do cotovelo para determinar o número ótimo de *clusters*

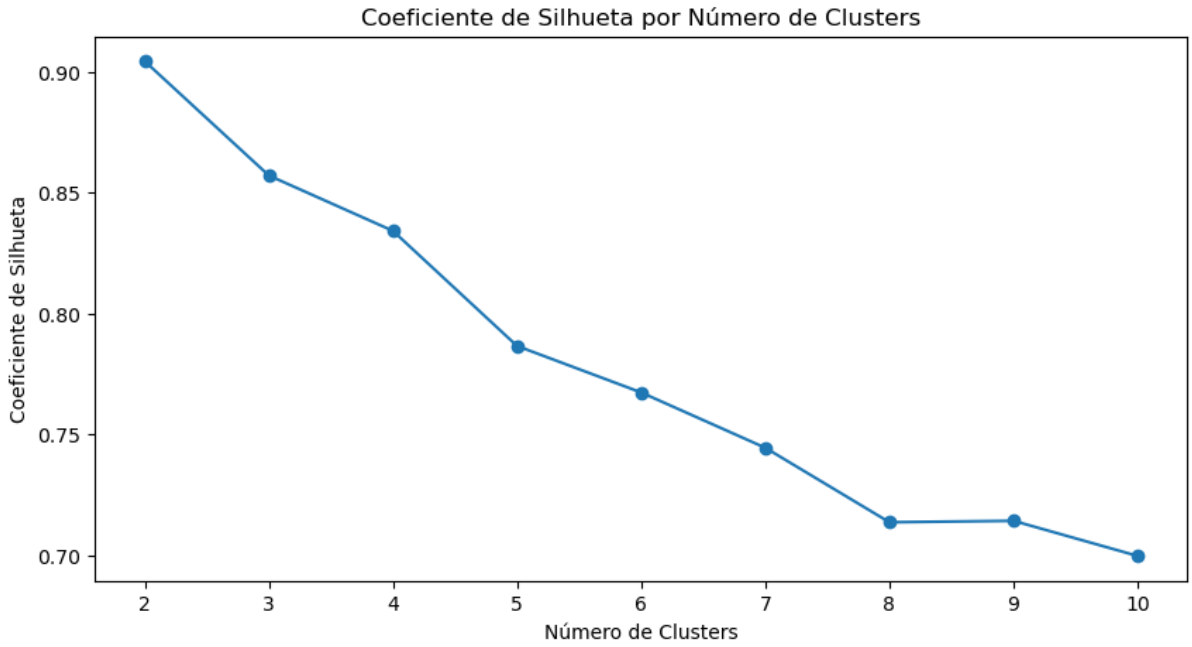


Fonte: Elaboração própria (2023).

O Gráfico 33 exibe o Coeficiente de Silhueta em relação ao número de *clusters*, oferecendo *insights* sobre a coesão e qualidade dos *clusters* em uma análise de agrupamento. No eixo horizontal, o número de *cluster*s varia de 2 a 10, enquanto o eixo vertical mostra o coeficiente de silhueta.

Ao examinar o Gráfico 33, é notável que a curva inicia em um ponto elevado no eixo vertical (0.90) e diminui gradualmente com o aumento do número de *cluster*s. O coeficiente de silhueta em 0.90 indica uma boa coesão e separação dos objetos entre os *clusters*. Portanto, a presença de um valor inicial elevado sugere que a formação de três *clusters* é robusta e coesa, fortalecendo a escolha desse número como a configuração mais apropriada para a análise de agrupamento.

Gráfico 33 – Coeficiente de silhueta por número de *clusters*



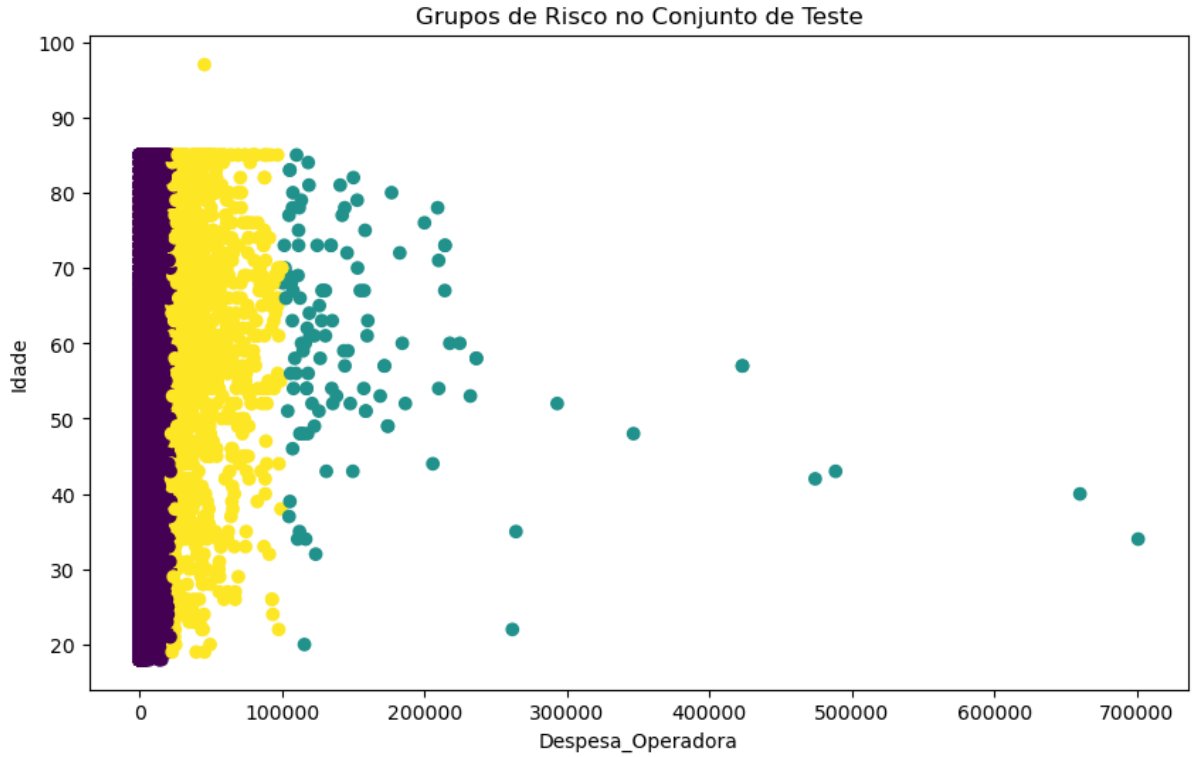
Fonte: Elaboração própria (2023).

Ao analisar os resultados do K-*means* no Gráfico 34, observa-se a identificação de três grupos distintos, representados por cores diferentes. O *cluster* roxo e o *cluster* amarelo, ocupam a maior parte do espaço no gráfico e parecem compostos por indivíduos com despesas assistenciais mais baixas, distribuídos de maneira uniforme em todas as faixas etárias. Estes grupos podem ser interpretados como de "baixo risco financeiro", indicando despesas assistenciais relativamente menores e não influenciadas pela idade.

O outro *cluster*, embora menos denso, representa um grupo com despesas assistenciais mais relevantes. Esse grupo difere em termos de despesas assistenciais, mas não apresenta uma variação ou correlação clara em relação à idade. O *cluster* verde, em particular, se destaca por conter indivíduos com as maiores despesas assistenciais, possivelmente representando um "grupo de alto risco financeiro". Visualmente, este grupo inclui alguns pontos dispersos, indicando a presença de possíveis outliers ou casos extremos de despesas assistenciais.

Os dados revelam uma distribuição homogênea de idades nos *clusters* roxo e amarelo, reforçando a ideia de que a idade, por si só, não seria um fator determinante para o nível de despesas assistenciais dentro desta amostra específica. A presença de *outliers*, especialmente no *cluster* de alto risco, sugere situações atípicas que demandam uma investigação mais aprofundada.

Gráfico 34 – Grupos de risco no conjunto teste



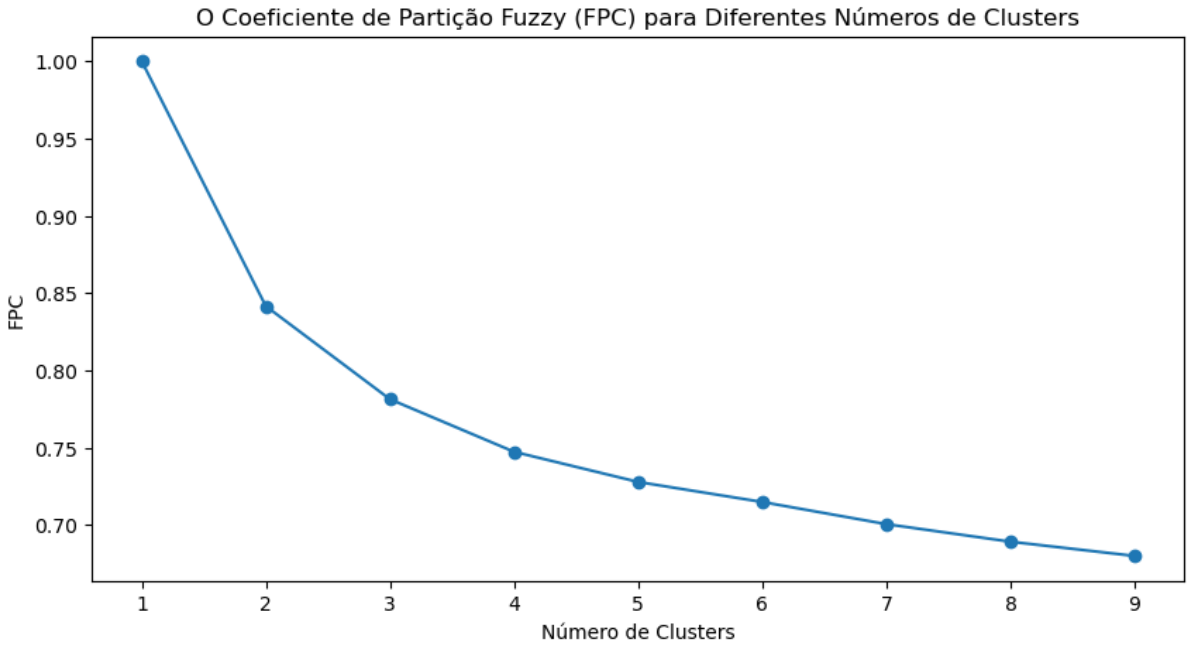
Fonte: Elaboração própria (2023).

Fuzzy means

O Gráfico 35 apresenta o Coeficiente de Partição Fuzzy (FPC) em relação ao número de *clusters*, oferecendo *insights* sobre a adequação da partição fuzzy para diferentes configurações de agrupamento. O eixo horizontal abrange o número de *clusters*, variando de 1 a 9, enquanto o eixo vertical exibe o FPC, uma medida que avalia a qualidade da partição fuzzy, indicando quão bem os dados estão distribuídos entre os *cluster*s.

Ao analisar o gráfico, observa-se que a curva inicia em um ponto relativamente alto no eixo vertical (1) e diminui à medida que o número de *cluster*s aumenta. Um FPC próximo a 1 sugere uma partição fuzzy mais precisa e eficaz. Assim, busca-se idealmente um número de *cluster*s que maximize o FPC, indicando uma partição fuzzy que melhor represente a estrutura subjacente dos dados. Este resultado contribui para a compreensão da qualidade da partição fuzzy em diferentes configurações de *clusters*.

Gráfico 35 – Coeficiente de partição Fuzzy para diferentes números de *clusters*



Fonte: Elaboração própria (2023).

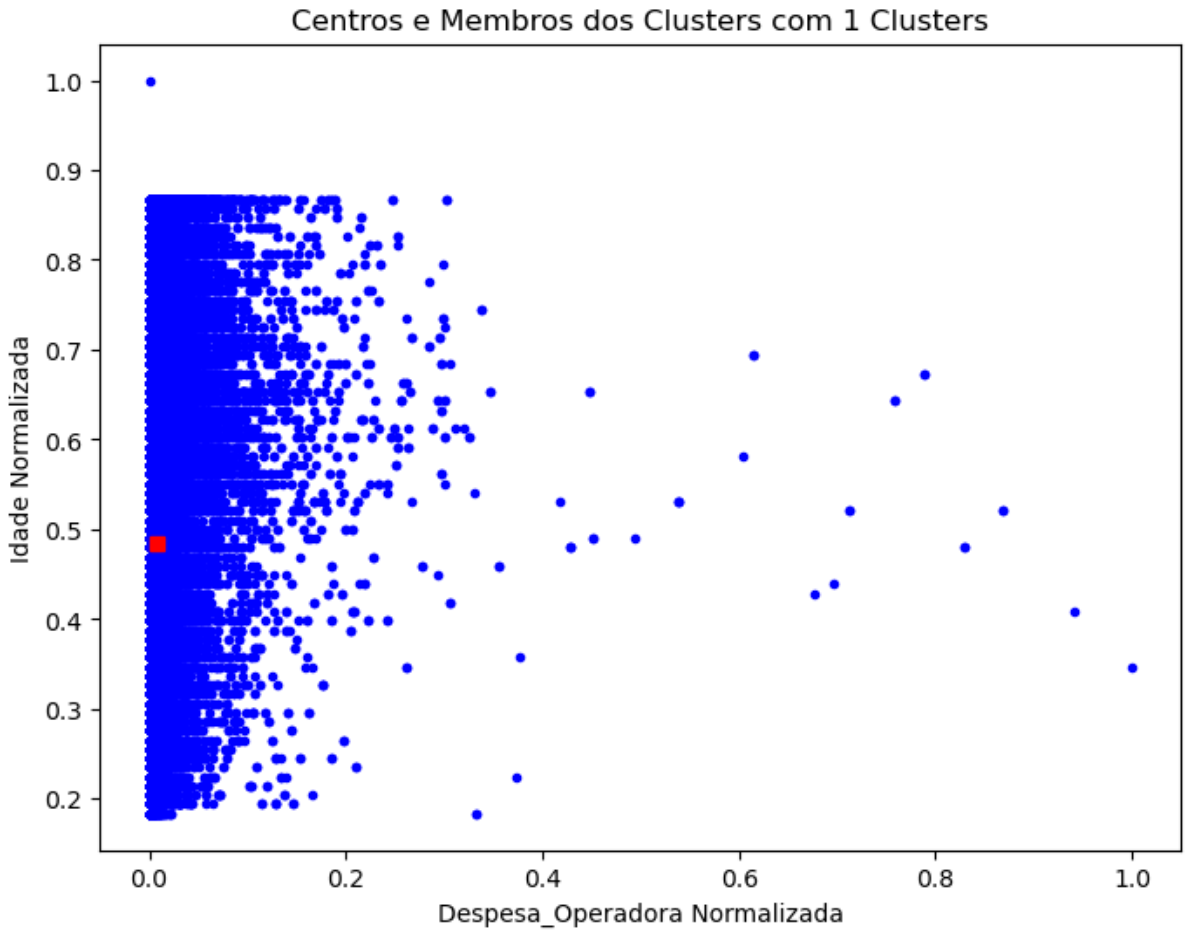
Observa-se um *cluster* único, identificado como 1 de acordo com o critério utilizado, representado pelos pontos azuis, com o centro do *cluster* marcado por um ponto vermelho no Gráfico 36. A disposição dos pontos indica uma alta densidade de dados caracterizada por despesas assistenciais mais baixas e uma ampla dispersão nas faixas etárias, evidenciada pela concentração de pontos no lado esquerdo do gráfico.

Além disso, há uma dispersão menos concentrada de pontos ao longo do eixo das despesas assistenciais, com poucos dados indicando despesas mais elevadas, mas sem uma diferenciação clara em relação às faixas etárias, como se observa na distribuição vertical dos pontos ao longo do eixo da idade.

A existência de um único *cluster* pode sugerir que o conjunto de dados carece de variação significativa que justificaria a subdivisão em grupos distintos, ou indica que o critério adotado para determinar o número de *cluster*s pode não ter sido suficientemente sensível para detectar agrupamentos mais sutis.

O ponto central em vermelho representa o centroide do *cluster* único, proporcionando um resumo das características médias do conjunto de dados com base na lógica fuzzy, refletindo uma média ponderada das despesas assistenciais e idades.

Gráfico 36 – Centros e membros dos *clusters* com 1 *cluster*



Fonte: Elaboração própria (2023).

c) Algoritmos de aprendizagem de máquina – análise preditiva

Nesta seção, são apresentados os modelos de aprendizado de máquina utilizados para atingir os objetivos desta tese. Inicialmente, a Tabela 14 expõe a acurácia dos modelos na base de dados da OPS Gama.

Os gráficos desta seção exibem a distribuição das previsões realizadas por três modelos distintos nos dados da OPS Gama, destacando também o RMSE (*Root Mean Square Error*) para cada modelo. Após a modelagem, a tabela 14 a seguir demonstra a acurácia dos modelos na base da OPS Gama.

Tabela 14 - Acurácia dos modelos

Random Forest	XGBOOST	KNN
RMSE	RMSE	RMSE
9412,94	7129,27	10536,40

Fonte: Elaboração própria (2023).

A Tabela 14 evidencia que o modelo KNN registrou o maior RMSE na base da OPS Gama, indicando seu desempenho inferior em comparação com os outros dois modelos. Com um RMSE de 10536,40, sugere-se que a abordagem fundamentada nos k-vizinhos mais próximos não consegue capturar de maneira eficaz a complexidade dos dados.

O modelo XGBoost, por sua vez, apresentou um RMSE de 7129,27 na base da OPS Gama, superando o desempenho dos modelos KNN e *Random Forest*. O XGBoost é reconhecido por sua capacidade de modelar interações complexas e não lineares e demonstrou o melhor desempenho.

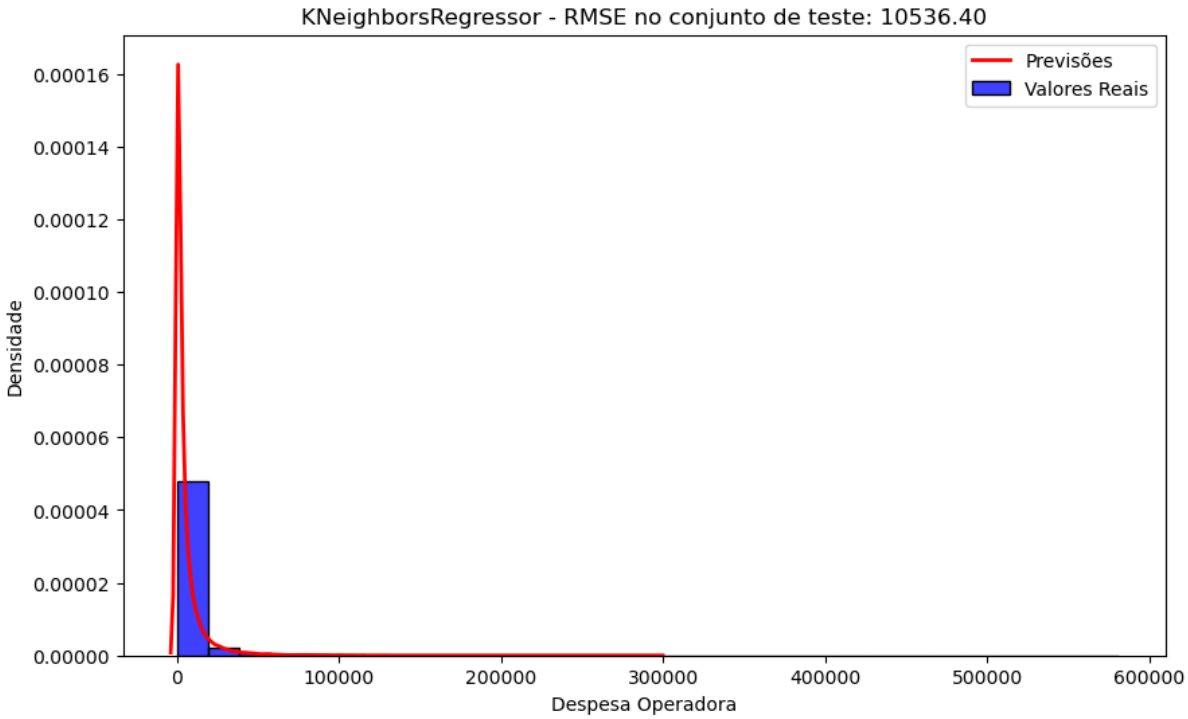
Esses resultados indicam que este modelo foi capaz de compreender de maneira mais precisa a complexidade dos dados em comparação com os outros modelos. Desse modo, com base nessas constatações, sugere-se que o modelo XGBoost é a escolha mais adequada entre os três para prever a "Despesa assistencial da OPS Gama" neste conjunto de dados específico.

No entanto, todos os modelos enfrentam desafios em prever com precisão os valores mais altos de despesa, como sugerido pelo gráfico de dispersão que revela um pico no extremo inferior da escala de despesa.

O Gráfico 37 apresenta a aplicação do algoritmo K-vizinhos próximos (KNN) na previsão das despesas assistenciais da OPS Gama. No eixo horizontal, são representados os valores das despesas assistenciais, variando de R$ 0,00 a R$ 700.000,00, enquanto o eixo vertical mostra a densidade associada a cada valor. As previsões geradas pelo algoritmo KNN, representadas pela linha vermelha, demonstram uma concordância geral que não corresponde completamente aos dados reais das despesas assistenciais, evidenciada pela discrepância de formato entre a curva prevista e as colunas que representam as despesas reais.

Essa diferença sugere que o algoritmo KNN, ao realizar previsões, identificou áreas de menor precisão, indicando a necessidade de uma avaliação adicional do desempenho do algoritmo, por meio de métricas de erro ou validação cruzada, visando aprimorar a confiabilidade das previsões.

Gráfico 37– K-vizinhos



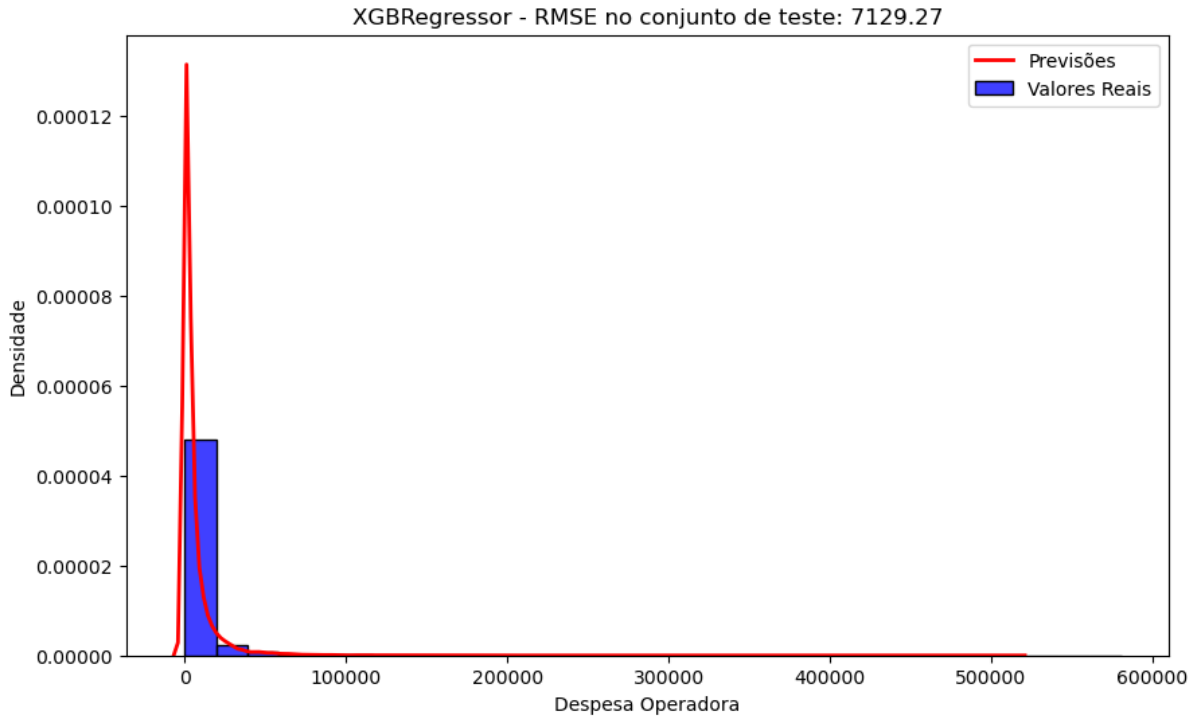
Fonte: Elaboração própria (2023).

O Gráfico 38 ilustra a implementação do algoritmo XGBoost para a previsão das despesas assistenciais da OPS Gama. No eixo horizontal, são delineados os valores das despesas assistenciais, variando de R$ 0,00 a R$ 700.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, oscilando entre 0,00000 e 0,00012.

A linha vermelha traça as previsões geradas pelo algoritmo XGBoost, exibindo uma curva que segue o formato das colunas representativas das despesas reais. Entretanto, um pico na linha vermelha atinge uma densidade de 0,00012, evidenciando uma discrepância notável. Esta observação sugere que o algoritmo XGBoost identificou uma tendência semelhante às despesas reais, mas apenas até a densidade 0,00004.

Para uma análise mais profunda da precisão do modelo, é sugerido realizar a avaliação de métricas de desempenho, tais como o erro absoluto médio ou a implementação de validação cruzada. Essas análises suplementares podem oferecer *insights* sobre o desempenho do algoritmo XGBoost, permitindo ajustes necessários para aprimorar a qualidade das previsões e assegurar uma representação mais fidedigna das despesas assistenciais.

Gráfico 38 – XGBoost



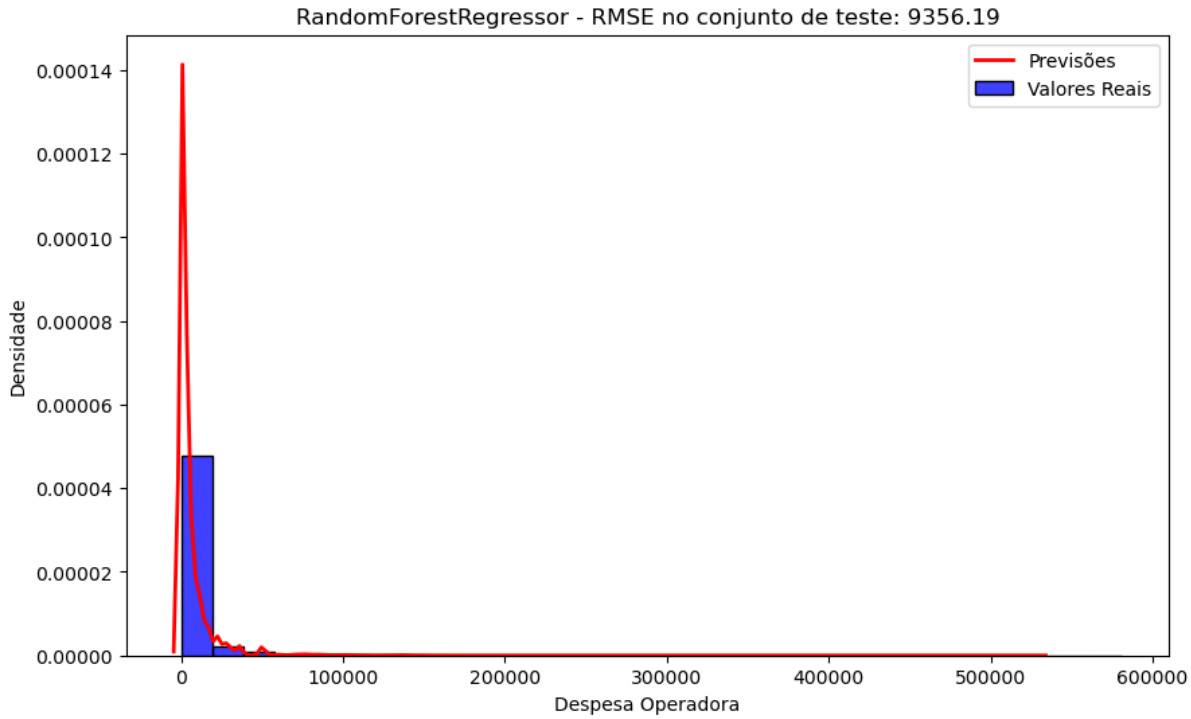
Fonte: Elaboração própria (2023).

O Gráfico 39 apresenta a aplicação do algoritmo de Florestas Aleatórias (*Random Forest*) na previsão das despesas assistenciais da OPS Gama. No eixo horizontal, são dispostos os valores das despesas assistenciais, variando de R$ 0,00 a R$ 700.000,00, enquanto o eixo vertical representa a densidade associada a cada valor, oscilando entre 0,00000 e 0,00012.

A linha vermelha delineia as previsões geradas pelo algoritmo de Florestas Aleatórias, apresentando uma curva que segue o formato das colunas que representam as despesas reais. Entretanto, observa-se, na altura da densidade 0,00004, uma discrepância notável. Esta observação sugere que o modelo de Florestas Aleatórias identificou uma tendência semelhante às despesas reais, mas com algumas divergências, especialmente no ponto mais elevado da curva.

Para uma avaliação mais aprofundada da precisão do modelo, seria recomendável a análise de métricas de desempenho, como o erro absoluto médio, ou a implementação de validação cruzada. Essas análises complementares podem proporcionar insights sobre o desempenho do algoritmo de Florestas Aleatórias, permitindo ajustes necessários para aprimorar a qualidade das previsões e assegurar uma representação mais fiel das despesas assistenciais.

Gráfico 39 – Florestas aleatórias ou *Random Forest*



Fonte: Elaboração própria (2023).

Com base nas distribuições apresentadas nos Gráficos 37, 38 e 39, fica evidente que o modelo que melhor antecipou as despesas assistenciais da OPS Beta foi o *XGBoost*, destacando-se por apresentar o menor nível de erro em comparação com os demais modelos. Essa eficácia posiciona o algoritmo *XGBoost* como uma ferramenta para a OPS Gama, permitindo a antecipação e gestão proativa de grupos de risco. A implementação de estratégias preventivas de saúde direcionadas aos beneficiários tem o potencial de resultar na redução do montante das despesas assistenciais e da sinistralidade da carteira da OPS Gama.

Além disso, a Tabela 15, apresentada a seguir, detalha em termos monetários a soma das despesas reais e a soma das despesas previstas. Essa abordagem não apenas respalda estatisticamente as estimativas realizadas, mas também fornece uma visão concreta da concordância entre as previsões dos modelos e os valores reais das despesas assistenciais.

Tabela 15 - Mensuração das despesas

Modelos de algoritmos	**Despesa (R$)***	**Diferença (R$)***
Real	240.653.287,00	0
XGBoost	238.881.376,00	- 1.171.911,00 **
KNN	208.454.099,00	- 32.199.188,00
Random Forest	238.823.988,73	- 1.829.298,26

* Os valores financeiros não foram deflacionados.
Fonte: Elaboração Própria (2023).

Desse modo, em termos financeiros, o modelo que mais se aproximou da realidade foi o *XGBoost*, apresentando uma variação de 0,74% entre as despesas reais da OPS Gama e a previsão do algoritmo, o que sustenta ainda mais a validação do resultado do modelo. Resultado que corrobora com os achados de Teixeira *et al*. (2023), que a partir de fatores socioeconômicos e de saúde, utilizou o *XGBoost* e alcançou uma alta taxa geral de previsibilidade da mortalidade por câncer no Brasil, confirmando que algoritmos de aprendizagem de máquina podem oferecer uma abordagem de modelagem superior em comparação com os modelos estatísticos convencionais.

4.3 Discussão sintética das hipóteses

A síntese dos resultados da investigação do impacto da IA na predição informacional das despesas assistenciais na saúde suplementar brasileira e sua relação com o risco de solvência das OPS, indica que os algoritmos demonstraram capacidade preditiva, produzindo impactos positivos no risco de solvência das OPS Alfa, Beta e Gama.

A primeira hipótese sugere que **a IA possui capacidade para prever as despesas assistenciais das OPS brasileiras**. Os resultados corroboram essa hipótese, demonstrando que os modelos de IA, ao integrarem dados contábeis, assistenciais e sociodemográficos, fornecem previsões com níveis de acurácia entre 99,06% a 99,26% o que representa a possibilidade real das OPS otimizarem o ciclo financeiro invertido e a capacidade de alocação de recurso. Acerca de resultados que comprovam os benefícios da utilização da IA, GFS Silva *et al* (2023), em seu estudo sobre predição informacional na saúde, conclui que os algoritmos de aprendizado de máquina são capazes de identificar padrões e tendências de dados que podem não ser tão aparentes, podendo contribuir para uma melhoria significativa nos diagnósticos e tratamentos dos pacientes.

A segunda hipótese **postula que a predição informacional das despesas assistenciais tem um impacto positivo no risco de solvência das OPS**. A tese traz validade para esta hipótese, evidenciando que a precisão nas previsões de despesas proporcionadas pelos algoritmos permite às OPS uma melhor gestão de seus recursos e riscos financeiros, contribuindo para a sustentabilidade e solvência do setor da saúde suplementar. Os níveis de acurácia dos algoritmos são diretamente proporcionais a capacidade de ação das OPS para redução da sinistralidade e consequentemente mitigação do risco de solvência ou até mesmo surgimento de novas OPS, conforme Araújo e Silva (2018) abordam em seu estudo.

Desse modo, a redução da sinistralidade, pode promover um ambiente de sustentabilidade para as OPS, reduzindo o risco de oligopolização da saúde suplementar no Brasil, corroborando com Coelho de Sá *et al* (2017) cujos os resultados encontrados reforçaram a importância dos gestores das OPS se anteciparem por meio de ações estratégicas para a manutenção da solvência das OPS.

Dado que estas ações que visam à manutenção da solvência das OPS são tomadas por Gestores financeiros, Diretores Executivos, CEO, CFO, Conselheiros de Administração, Conselheiros Fiscais, Auditores internos, Gestores de risco e de *compliance*, os resultados desta tese indicam que a utilização da IA, para predizer informações sobre as despesas assistenciais, pode ser uma fonte informacional para subsidiar melhores tomadas de decisão dos usuários internos acerca da alocação de capital das OPS e gestão do ciclo financeiro invertido. Segundo Araujo e Novinski (2009), estas decisões que visam alcance da manutenção da solvência das OPS podem ser utilizadas como sinais de qualidade, tendo maior probabilidade de atrair um *pool* de beneficiários.

5. CONCLUSÃO

As conclusões desta tese estão estruturadas em cinco seções principais para uma abordagem abrangente e detalhada:

1. Conclusões gerais: apresenta um resumo abrangente dos achados mais significativos da pesquisa;
2. Diretrizes para gestão de riscos de solvência das OPS: propõe diretrizes específicas para aperfeiçoar a gestão de riscos de solvência das OPS, a partir da proposta de utilização de análise preditiva baseada em IA;
3. Principais contribuições: destaque das contribuições significativas para a literatura e práticas na área de saúde suplementar, enfatizando as contribuições em abordagens econômicas, sociais e de Governança Corporativa.
4. *Insights* para pesquisas futuras: sugestões de pesquisas futuras com base nas descobertas e limitações identificadas, propondo caminho para aprofundamento e inovação contínuos no campo.
5. Limitações da pesquisa: limitações do estudo, fornecendo um panorama transparente das áreas que necessitam de mais investigação ou que apresentaram desafios durante a pesquisa.

Cada uma dessas seções desempenha um papel importante na contextualização das descobertas da tese, realçando seu valor teórico e prático, e ainda indicando possíveis futuras direções no campo da IA na saúde suplementar brasileira.

5.1 Conclusões gerais

Em um cenário permeado de conflitos de interesses, assimetria informacional e alto índice de insolvência, as OPS brasileiras, sob a forte pressão de relevantes impactos sociais, enfrentam desafios acerca da sustentabilidade das suas operações. A respeito do desequilíbrio informacional e demais conflitos entre os atores, a Teoria da Regulação Econômica surge como resposta, a fim de assegurar a eficiência, protegendo os interesses dos consumidores e promovendo práticas comerciais equilibradas.

Como resultado da regulação na saúde suplementar brasileira, as OPS possuem a obrigatoriedade de divulgação de informações à sociedade e governo (por meio da ANS), incluindo transparência em relação ao estado de solvência, alcançando a proposta da Teoria da sinalização em que a assimetria informacional é reduzida a partir da sinalização de informações que emanam das empresas para o mercado. A referida sinalização acerca da solvência das OPS é acompanhada por meio do índice de sinistralidade das OPS, que em 2023 alcançou 88,20%.

Desse modo, esta tese defendeu a premissa de que a utilização de modelagens preditivas informacionais da previsibilidade de gastos [despesas assistenciais], impacta positivamente o estado de solvência das OPS brasileiras, na medida em que reduz o índice de sinistralidade. Os resultados demonstram que, a partir da análise de *clusters Fuzzy C-Means* e *K-means*, utilizando as bases de dados das OPS Alfa, Beta e Gama, as carteiras de beneficiários foram segmentadas em grupos de risco, conforme classes por proximidade das informações.

Por sua vez, esta segmentação permite o acompanhamento próximo e preventivo dos grupos que apresentam maiores níveis de riscos de utilização, aumentando a probabilidade de redução do índice de sinistralidade e consequentemente a manutenção da solvência da OPS, surgindo como respostas às conclusões do estudo de Araújo e Silva (2018), que identificaram que o aumento da sinistralidade oferece riscos à sobrevida e à abertura de novas OPS.

No tocante as modelagens de previsibilidade informacional das despesas assistenciais, foram selecionados três modelos de aprendizagem de máquina, K-Vizinhos próximos, florestas aleatórias e *XGBoost*. Para as OPS Alfa e Beta, o modelo de florestas aleatórias demonstrou maiores níveis de acurácia, apresentando uma variação de 2,22% e 1,13%, respectivamente,

entre a despesa real e a predita pelo algoritmo. Contudo, para a OPS Gama, o modelo do *XGBoost* demonstrou maiores níveis de acurácia, apresentando uma variação de 0,40% entre a despesa real e a predita pelo algoritmo. A combinação entre análise de *cluster* e aprendizagem de máquina é definida por Teixeira *et al*. (2023), como uma metodologia promissora, adaptável e aplicável para o alcance de bons resultados.

Esse resultado implica em afirmar que a utilização destes algoritmos como mecanismo disruptivo para a gestão prévia do risco, minimizará os possíveis gastos futuros por grupo de risco segmentado, contribuindo ainda para que os gestores e tomadores de decisão das OPS consigam gerir o ciclo financeiro invertido das OPS com mais capacidade de fluxo de caixa para as decisões de otimização de alocação de recursos.

De modo geral, estes resultados permitem a não rejeição das hipóteses de pesquisa, pois os algoritmos utilizados demonstraram capacidade de predição informacional das despesas assistenciais das OPS participantes, bem como esta predição informacional pode impactar positivamente e em níveis muito significativos o risco de solvência das OPS.

Por fim, ao considerar a relevância do mercado da saúde suplementar, a utilização de algoritmos e modelagens preditivas não possui limites organizacionais, uma vez que sua utilização contribui para redução do risco de solvência das OPS no mercado, podendo se estender a um olhar mais preventivo e assistencial com a saúde dos beneficiários, trazendo então um relevante impacto social direto para 25% da população brasileira.

5.2 Diretrizes para gestão de riscos de solvência das OPS, a partir de análise preditiva por IA

Esta tese aborda o desafio crescente enfrentado pelas OPS no Brasil, particularmente no que se refere à gestão eficaz de risco de solvência. Diante da complexidade do ambiente de saúde suplementar, marcado por variáveis econômicas, assistenciais e sociodemográficas dinâmicas, emerge a necessidade de abordagens inovadoras. Nesse contexto, a IA surge como uma ferramenta que oferece novas perspectivas e soluções para predição informacional dos riscos na saúde suplementar.

A metodologia adotada enfatiza a aplicação de modelos preditivos de IA, integrando dados contábeis, assistenciais e demográficos. Esta abordagem multidimensional não apenas aprimora a precisão das previsões de despesas assistenciais, mas também proporciona *insights* sobre os fatores que podem influenciar o risco de solvência das OPS. Dessa forma, as diretrizes

propostas abaixo, buscam promover a orientação das OPS na implementação eficaz de soluções de IA em busca de uma gestão de riscos mais proativa.

Essas diretrizes sugeridas representam um passo significativo para a integração de tecnologias de IA nas práticas operacionais das OPS, contribuindo para um sistema de saúde mais resiliente e sustentável.

- **Desenvolvimento de modelos preditivos específicos**: Criar e implementar modelos de IA (*machine learning*) que sejam adaptados às particularidades de cada OPS, considerando fatores regionais, econômicos, demográficos, modalidade e modelo de gestão.
- **Integração de dados multidisciplinares [interoperabilidade]**: Fomentar a integração de dados contábeis, assistenciais e sociodemográficos para alimentar e retroalimentar os modelos de IA, aumentando os níveis de acurácia e aprimorando a precisão nas previsões.
- **Monitoramento e avaliação contínua**: Estabelecer mecanismos de monitoramento e avaliação periódicos dos modelos de IA para garantir sua eficácia, precisão e adequação às mudanças no mercado, legislação e dinâmica da gestão e do ambiente de negócio da OPS.
- **Colaboração, compartilhamento de conhecimento e fomento de parcerias estratégicas**: Promover parcerias entre OPS, instituições de pesquisa (Institutos e Universidades) e o setor governamental para compartilhar conhecimentos, experiências e inovações de aplicações da IA à saúde suplementar.
- **Capacitação e conscientização**: Investir na formação de profissionais de análise de dados e IA para tomadas de decisão estratégicas.

5.3 Principais contribuições

Esta tese apresenta contribuições significativas e aplicáveis às práticas de gestão de riscos e estratégia das OPS brasileiras, sendo agrupadas em abordagens estruturais abaixo descritos:

a) Contribuições Econômicas

- Sustentabilidade financeira das OPS: Todas as temáticas e discussões propostas nesta tese, proporcionam *insights* para a promoção da sustentabilidade financeira das OPS brasileiras, por meio da aplicação de modelos de IA na predição informacional das

despesas assistenciais. Desse modo, essa abordagem pode contribuir diretamente para a estabilidade econômica do setor, reduzindo os riscos de insolvência e ampliando as possibilidades dos gestores das OPS nas decisões de alocação de recursos.

- Impacto econômico da insolvência: A partir da relevância econômica das OPS, a abordagem acerca da relação direta entre a solvência das OPS e o impacto econômico no setor de saúde suplementar, proporciona o entendimento sobre como a insolvência pode afetar não apenas as OPS individualmente, mas também a cadeia produtiva e a geração e manutenção de empregos deste setor.
- Desenvolvimento de estratégias baseadas em risco: A partir da possibilidade de previsibilidade das despesas assistenciais, os resultados desta tese lançam luz sobre sugestões de estratégias baseadas em prêmios de risco, em que as OPS proporcionariam uma visão inovadora para atrair clientes e otimizar a alocação de recursos financeiros. Ainda de forma mais relevante, essa abordagem estratégica pode contribuir para a competitividade e longevidade das OPS no mercado.

b) Contribuições Sociais

- Acesso e qualidade dos serviços de saúde: Sob a ótica social, os resultados desta tese enfatizam a importância da sustentabilidade das OPS para garantir o acesso e a qualidade dos serviços de saúde para os beneficiários, com vistas a prevenção de colapso ou interrupções que impactariam diretamente a prestação de assistência à saúde a estes.
- Redução da concentração de mercado: Os resultados desta tese lançam luz para os possíveis impactos sociais decorrentes da concentração de mercado resultante da insolvência de OPS. A pesquisa destaca a necessidade de manutenção da dinâmica da regulação promovendo a livre concorrência, evitando a perda de poder de escolha dos beneficiários e a possível elevação dos custos devido à forte concentração de carteiras em grandes *players* do mercado.
- Sustentabilidade da cadeia produtiva da saúde suplementar: De maneira relevante, o conteúdo desta tese destaca a importância econômica e social da cadeia produtiva da saúde suplementar, evidenciando o papel das OPS na sustentabilidade do setor. Desse modo, a insolvência das OPS enseja em impacto na continuidade das operações das operadoras, o que consequentemente pode comprometer a geração/manutenção de empregos e a estabilidade da cadeia produtiva.

c) Contribuições para o sistema de Governança Corporativa

- Implementação de práticas de governança: De maneira clara, os resultados desta tese reforçam a importância das práticas de governança corporativa já editadas pelas Resoluções Normativas 443/2019 e 518/2022 da ANS, principalmente acerca da necessidade da redução da assimetria informacional da OPS em relação aos beneficiários, contribuindo para a promoção de decisões que considerem o potencial de entrega e saúde econômica e financeira das OPS.

Em síntese, essa tese não apenas avança o conhecimento teórico, mas também oferece aplicações práticas que têm o potencial de influenciar positivamente a economia, sociedade, meio ambiente e governança corporativa no contexto da saúde suplementar brasileira.

5.4 *Insights* para pesquisas futuras

Diante da profundidade e extensão do tema desta tese, sugiro *insights* para pesquisas futuras, iniciando pelo aprofundamento na análise regional, investigando as variações regionais na sinistralidade e nos riscos de solvência das OPS, considerando as particularidades socioeconômicas e epidemiológicas de diferentes regiões do Brasil. Essa abordagem pode fornecer percepções valiosas para estratégias específicas em diferentes contextos geográficos.

Outro subtema que merece destaque é a exploração do impacto das Práticas de Governança Corporativa, conforme estabelecidas nas Resoluções Normativas 443/2019 e 518/2022 da ANS, no desempenho das OPS e na gestão do risco de solvência. A partir desse *insight*, há o vislumbre acerca de possibilidades de aperfeiçoamento da regulação, por meio da redução da assimetria informacional.

Adicionalmente, há uma sugestão de análise do impacto social da insolvência de OPS, não apenas no acesso aos serviços de saúde, mas também em termos de empregabilidade no setor, concentração de mercado e possíveis efeitos colaterais sobre a população em geral. Essa pesquisa pode envolver estudos de caso em OPS específicas e análises mais aprofundadas do contexto organizacional de cada OPS participante.

Sugiro ainda o constante desenvolvimento e treinamento de modelos preditivos, aprimorando continuamente os modelos de inteligência artificial utilizados na predição informacional, explorando novas técnicas e abordagens de aprendizado de máquina. Essa sugestão de pesquisa pode incluir a integração de dados adicionais, como indicadores econômicos externos, para melhorar a precisão das previsões.

Há ainda a sugestão acerca da avaliação do impacto das recomendações contidas nesta tese. Dessa forma, seriam analisadas como as práticas sugeridas impactam efetivamente a gestão de riscos, a solvência e a tomada de decisões no ambiente de mundo real das OPS.

Vislumbrando uma sugestão com a ótica do longo prazo, sugiro realizar estudos longitudinais para entender como as variáveis preditivas e a gestão de riscos evoluem ao longo do tempo. Isso permitiria uma análise mais dinâmica das tendências e uma compreensão mais profunda dos fatores que influenciam a solvência das OPS ao longo dos anos.

Por fim, sugiro pesquisa futura com ênfase na experiência do beneficiário, a partir da investigação de como a aplicação de IA e aprimoramento das práticas de gestão de riscos impactam diretamente a experiência do beneficiário, avaliando se as melhorias na previsibilidade de gastos são traduzidas em benefícios tangíveis para os usuários, proporcionando mais cuidados preventivos com a saúde ou até mesmo ampliação da cobertura.

5.5 Limitações da pesquisa

Reconhecer e compreender as limitações desta tese contribuem para uma interpretação dos resultados de forma crítica e ainda para orientação das pesquisas futuras que se originarão desta. Abaixo indico algumas limitações identificadas:

- Disponibilidade dos dados: a ANS não dispõe de banco de dados estruturado ou equivalente aos existentes para empresas de capital aberto ou fechado. Sendo assim, a disponibilidade dos dados assistenciais e sociodemográficos foram coletados de maneira individual e manualizada, o que de certa forma impacta na possibilidade de elaboração de pesquisas mais abrangentes e generalização dos resultados.
- Fatores externos e contextuais: o estudo ocorreu de maneira transversal, atravessando os anos da pandemia da COVID-19. Esse acontecimento mundial provocou mudanças e fez surgir cenários imprevistos em toda a sociedade, ocasionando possíveis mudanças em algumas legislações, cenários econômicos imprevistos e impacto financeiro sobre a saúde pública e privada.
- Generalização dos resultados: a natureza preditiva da pesquisa implica em algumas incertezas sobre eventos futuros. Os modelos utilizados são baseados em dados históricos e, portanto, estão sujeitos a mudanças nas condições futuras que podem não ter sido previstas.

- Aspecto temporal dos Dados: os dados utilizados têm uma temporalidade específica, assim, mudanças relevantes no ambiente regulatório, econômico ou de saúde que ocorram após o período de coleta podem não ser refletidas na pesquisa.

REFERÊNCIAS

ALTMAN, E. I.; HOTCHKISS, E. Corporate financial distress and bankruptcy: Predict and avoid bankruptcy, analyze and invest in distressed debt. 3. ed. New York: Wiley Finance, 2010.

ANDRADE, M. V.; MAIA, A. C. Diferenciais de utilização do cuidado de saúde no sistema suplementar brasileiro. Revista de Estudos Econômicos, São Paulo, v. 39, n. 01, pp. 7-38, jan./mar. 2009. Disponível em: <http://www.scielo.br/pdf/ee/v39n1/v39n1a01.pdf>. Acesso em: 09 de fev. 2022.

ARAUJO, Â. A. S.; SILVA, J. R. S. Análise de tendência da sinistralidade e impacto na diminuição do número de operadoras de saúde suplementar no Brasil. Ciência saúde coletiva, Rio de Janeiro, v. 23, n. 8, p. 2763-2770, Ago. 2018. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S141381232018000802763&lng=pt &nrm=iso. Acesso em: 18 jun. 2022.

ARORA, P.; BOYNE, D.; SLATER, J.; GUPTA, A.; BRENNER, D. R.; DRUZDZEL, M. J.; Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. ISPOReThe Professional Society for Health Economics and Outcomes Research. v. 22, p. 439-445. 2019. Disponível em: https://doi.org/10.1016/j.jval.2019.01.006. Acesso em 01 de ago. 2022.

BARROS, A. J. S.; LEHFELD, N. A. S. Fundamentos de Metodologia: Um Guia para a Iniciação Científica. 2 ed. São Paulo: Makron Books, 2000.

BRAGANÇA, C. G.; PINHEIRO, L. E. T.; BRESSAN, V. G. F.; SOARES, L. A. DE C. F. Liquidação de operadoras de planos de assistência à saúde no Brasil. Enfoque: Reflexão Contábil, v. 38, n. 2, p. 33-47, 2019. Disponível em: https://doi.org/10.4025/enfoque.v38i2.43515. Acesso em 01 de ago. 2022.

AREIAS, C. A. C.; JOÃO V. F. C. O Resseguro Na Saúde Suplementar: Um Estudo Contrafactual Sobre Os Impactos Da Adoção De Tratados De Resseguros Por Operadoras De Planos De Saúde No Brasil. BBR Brazilian Business Review (Portuguese Ed.), v. 18, n. 2, p. 217–235, 2021. Disponível em: http://dx.doi.org/10.15728/bbr.2021.18.2.6. Acesso em 01 de set. 2022.

BAUMOL, W. J. On the Proper Cost Tests for Natural Monopoly in a Multiproduct Industry. The American Economic Review. v. 67, n. 5, p. 809-822, dez. 1977.

BEZDEK, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981. Disponível em: DOI:10.1007/978-1-4757-0450-1. Acesso em 03 de set. 2023.

BORDE, S.; RANE, A.; SHENDE, G.; SHETTY, S. Real Estate Investment Advising Using Machine Learning. International Research Journal of Engineering and Technology, v. 4, n. 3, p. 1821-1825, Mar. 2017.

BOULDING, W.; KIRMANI, A. A consumer-side experimental examination of signaling theory: Do consumers perceive warranties as signals of quality? Journal of Consumer Research, v. 20, n. 1, p. 111–123, Jun. 1993. Disponível em: https://doi.org/10.1086/209337. Acesso em 03 de set. 2023.

BRASIL. Lei nº 9.656, de 03 de junho de 1998. Dispõe sobre os planos de seguros privados de assistência à saúde. Casa Civil, subchefia para assuntos jurídicos, Brasília, DF, 1998.

BRASIL. Lei nº 9.961, de 28 de janeiro de 2000. Cria a Agência Nacional de Saúde Suplementar – ANS e dá outras providências. Casa Civil, subchefia para assuntos jurídicos, Brasília, DF, 2000.

CFM - Conselho Federal de Medicina. Resolução CFM nº 2.180/2018. Estabelece os dados de médicos que devem ser disponibilizados em consultas eletrônicas relacionadas aos registros dos profissionais médicos inscritos no Sistema Conselhos de Medicina e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, ano 155, n. 181, p. 128, 19 set.2018.

ANS – Agência Nacional de Saúde Suplementar. Resolução Normativa nº 205/2009. Estabelece novas normas para o envio de informações do Sistema de Informações de Produtos (SIP). ANS, 9 out. 2009.

ANS – Agência Nacional de Saúde Suplementar. Resolução Normativa 209/2009. Estabelece os critérios de manutenção de Recursos Próprios Mínimos e constituição de Provisões Técnicas. ANS, 15 out. 2009.

ANS – Agência Nacional de Saúde Suplementar. Resolução Normativa nº 443/2009. Dispõe sobre adoção de práticas mínimas de governança corporativa, com ênfase em controles internos e gestão de riscos, para fins de solvência das operadoras de plano de assistência à saúde. ANS, 25 jan. 2019.

ANS – Agência Nacional de Saúde Suplementar. Resolução Normativa, nº 518/2022. Dispõe sobre adoção de práticas mínimas de governança corporativa, com ênfase em controles internos e gestão de riscos, para fins de solvência das operadoras de plano de assistência à saúde. ANS, 29 abr. 2022.

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR – ANS. Caderno de Informação de Saúde Suplementar. Rio de Janeiro: ANS, dez. 2023.

AGÊNCIA NACIONAL DE SAÚDE SUPLEMENTAR – ANS. Caderno de Informação de Saúde Suplementar. Rio de Janeiro: ANS, mar. 2022.

BRASIL. [Constituição (1988)]. Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal, 2016. 496 p. Disponível em: https://www2.senado.leg.br/bdsf/bitstream/handle/id/518231/CF88_Livro_EC91_2016.pdf.Acesso em: 4 set. 2021

CAPITOLIO CONSULTORIA. Estudo sobre sinistralidade na saúde suplementar. Disponível em: https://www.sonhoseguro.com.br/2019/05/capitolio-lanca-estudo-sobre-sinistralidade-na-saude-suplementar-2018-2017/. Acesso em: 20 jun. 2021.

CHEN, H.; CHIANG, R. H. L.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. Management Information Systems Research Center, v. 36, n. 4, p. 1-22, Dez. 2012. Disponível em: https://doi.org/10.2307/41703503. Acesso em 24 set. 2022.

CHEN, T.; GUESTRIN, C.; XGBoost: A Scalable Tree Boosting System. Expert Systems With Applications. v. 9, n. 5, p. 785-794, Jun. 2016. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em 12 dez. 2023.

CHEN, H.; CHIANG, R. H. L.; STOREY, V. C. Business intelligence and analytics: From big data to big impact. MIS Quarterly: Management Information Systems, v. 36, n. 4, p. 1165–1188, Dez. 2012. Disponível em: https://doi.org/10.2307/41703503. Acesso em 12 dez. 2023.

COASE, R. H. The Nature of the Firm. Economica, [s.l.], v. 4, n. N.S., p. 386-405, Nov. 1937.

COELHO DE SÁ, M.; MACIEL JÚNIOR, J. N.; REINALDO, L. M. Processo de Ruína Finito: um Estudo de Caso na Saúde Suplementar no Brasil. Revista Evidenciação Contábil e Finanças. v. 5, n. 2, p. 88–103, 2017. Disponível em: DOI:10.18405/20170206. Acesso em: 01 jul. 2022.

CONNELLY, B. L.; HOSKISSON, R. E.; TIHANYI, L.; CERTO, S. T. Ownership as a form of corporate governance. Journal of Management Studies. v. 47, n. 8, p. 1561-1589, Dez. 2010. Disponível em: doi:10.1111/j.1467-6486.2010.00929.x. Acesso em 01 out. 2022.

CHUNG, K. C.; TAN, S. S.; HOLDSWORTH, D. K. Insolvency prediction model using multivariate discriminant analysis and artificial neural network for the finance industry in New Zealand. International Journal of Business and Management, v. 3, n. 1, p. 19-28, Jan. 2008. Disponível em: https://doi.org/10.5539/ijbm.v3n1p19. Acesso em 14 set. 2022.

DHAMIJA, P.; BAG, S. Role of artificial intelligence in operations environment: A review and bibliometric analysis. The TQM Journal, v. 32, n. 4, p. 869–896, Mar. 2020. Disponível em: 10.1108/TQM-10-2019-0243. Acesso em 11 ago. 2023.

DOLGUI, A.; IVANOV, D.; SETHI, S.; SOKOLOV, B.Scheduling in production, supply chain and Industry 4.0 systems by optimal control: fundamentals, state-of-the-art and applications. International Journal of Production Research, v. 57, n. 2, p. 411-432. Mar. 2018. Disponível em: 10.1080/00207543.2018.1442948. Acesso em 04 mai. 2022.

ESPEJO-GARCIA, B.; MARTINEZ-GUANTERB, J.; PÉREZ-RUIZB, M.; LOPEZ-PELLICERA F. J.; ZARAZAGA-SORIAA, F. J. Machine learning for automatic rule classification of agricultural regulations: A case study in Spain. Computers and Electronics in Agriculture, v. 150, p. 343–352. Jul. 2018. Disponível em: https://doi.org/10.1016/j.compag.2018.05.007. Acesso em 05 mai. 2022.

FONSECA, J. J. S. Metodologia da pesquisa científica. Fortaleza: UEC, 2002. Ebook.

GFS SILVA; DUARTE, L.S.; SHIRASSU, M.M.; Peres, S.V.; MORAES, M.A.; CHIAVEGATTO FILHO, A. Machine learning for longitudinal mortality risk prediction in

patients with malignant neoplasm in São Paulo. Artificial Intelligence in the Life Sciences, v. 3, n. 2023. Disponível em: https://doi.org/10.1016/j.ailsci.2023.100061. Acesso em 02 jan. 2024.

GONG, H.; SUN, Y.; SHU, X.; HUANG, B. Use of random forests regression for predicting IRI of asphalt pavements. Construction and Building Materials, v. 189, p. 890-897, 2018. Disponível em: https://doi.org/10.1016/j.conbuildmat.2018.09.017. Acesso em 03 jan. 2024.

GHOSH, S.; DUBEY, S. K. Comparative Analysis of K-Means and Fuzzy CMeans Algorithms, International Journal of Advanced Computer Science and Applications, v. 4, p. 35-39, 2013. Disponível em: DOI: 10.14569/IJACSA.2013.040406. Acesso em 03 dez. 2023.

GUIMARÃES, A. L. S.; ALVES, W. O. Prevendo a insolvência de operadoras de planos de saúde. Revista de Administração de Empresas (RAE), São Paulo, v. 49, n. 4, p. 459–471, out./dez. 2009. Disponível em: <http://www.scielo.br/pdf/rae/v49n4/v49n4a09.pdf>. Acesso em: 12 fev. 2022.

IGARASHI, W.; RAUTENBERG, S.; MEDEIROS, L. F.; PACHECO, R. C. D. S.; SANTOS, N. D.; FIALHO, F. A. P. Aplicações de Inteligência Artificial para Gestão do Conhecimento nas organizações: um estudo exploratório. Revista Capital Científico - Eletrônica, v. 6, n. 1, p. 239-256, 2008. Disponível em: < https://revistas.unicentro.br/index.php/capitalcientifico/article/view/816/925>. Acesso em 12 fev. 2022.

INSTITUTO DE ESTUDOS DE SAÚDE SUPLEMENTAR (IESS). Relatório de Emprego na Cadeia Produtiva da Saúde. Edição abril/22. São Paulo, 2022. Disponível em: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.iess.org.br/sites/default/files/2022-07/RECS_Abr22_SaoPaulo.pdf. Acesso em: 01 Jun. 2022.

KAUFMAN, D. A inteligência artificial irá suplantar a inteligência humana?. São Paulo: Estação das Letras e Cores, 2019.

LANGABEER, J. R.; HELTON, J. Financial Distress and the Role of Managed Care in Medicaid Insurance Markets. Health Services Management Research, v. 31, n. 2, p. 97-106. 2018.

KHALID, S.; KHAN, M. A.; MAZLIHAM, M.S.; ALAM, M. M.; AMAN, N.; TAJ, M. T.; ZAKA, R.; JEHANGIR, M. Predicting Risk through Artificial Intelligence Based on Machine

Learning Algorithms: A Case of Pakistani Nonfinancial Firms. Hindawi Complexity. v. 2022. Disponível em: https://doi.org/10.1155/2022/6858916. Acesso em 19 jun. 23.

LEVI, J.; LEE, J. J.; GIBSON, L. Measuring the impact of the economy on the health care system: The positive case for managed care. Journal of Health Care Finance, v. 35, n. 3, p. 33-41. 2009.

MAKRIDAKIS, S. The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. Futures, v.90, p. 46–60, Jun. 2017. Disponível em: DOI: https://doi. org/10.1016/j.futures.2017.03.006. Acesso em 19 jun. 2022.

MALTA, D. C.; MERHY, E. E. Buscando novas modelagens em saúde: as contribuições do Projeto Vida e do Acolhimento para a mudança do processo de trabalho na rede pública de Belo Horizonte, 1993-1996. Revista Eletrônica do DMPS, Unicamp, 01 dez. 2001. Disponível em: <http://www.uff.br/saudecoletiva/professores/merhy/artigos-02.pdf>. Acesso em: 18 jun. 2021.

MISHRA, D.; SHARMA, R.R.K.; GUNASEKARAN, A.; PAPADOPOULOS, T; DUBEY, R. Role of decoupling point in examining manufacturing flexibility: an empirical study for diferente business strategies. Total Quality Management and Business Excellence, v. 30, n. 9, p. 1126-1150. 2019. Disponível em: DOI:10.1080/14783363.2017.1359527. Acesso em 12 jun. 2022.

OLIVEIRA, F.T.; GOMES, M. R. A.; FERREIRA, L.; PIMENTA, T. L.; COSTA, L. H. O marco da solvência na saúde suplementar. Cadernos De Estudos Interdisciplinares, 3(1). 2019. Recuperado de: https://publicacoes.unifal-mg.edu.br/revistas/index.php/cei/article/view/534.

PAN, Y. Heading toward Artificial Intelligence 2.0. Engineering, v. 4, n. 2, p. 409–413, Dez. 2016. DOI: https:// doi.org/10.1016/J.ENG.2016.04.018.

PELTZMAN, S. A teoria econômica da regulação depois de uma década de desregulação. In: MATTOS, Paulo (Coord.). Regulação Econômica e Democracia: o debate norte-americano. São Paulo. cap. 34, p. 81-127, 2004.

PLANTIN, G.; ROCHET, J.-C. When Insurers Go Bust: An Economic Analysis of the Role and Design of Prudential Regulation. Princeton University Press. 2007. http://www.jstor.org/stable/j.ctt7sdg9

RAMEZAN, A. C.; WARNER, A. T.; MAXWELL, E. A. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sens, v. 11, n. 2. 2019. Disponível em: https://doi.org/10.3390/rs11020185. Acesso em 10 set. 2022.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3.ed. San Francisco: Prentice Hall, 2010.

SALVATORI, R. T.; VENTURA, C. A. A. A agência nacional de saúde suplementar - ANS: onze anos de regulação dos planos de saúde. Organizações & Sociedade (O&S), Universidade Federal da Bahia (UFBA), Salvador, v. 19, n. 62, p. 471-487, jul./set. 2012. Disponível em: <https://portalseer.ufba.br/index.php/revistaoes/article/view/11208/8117>. Acesso em: 12 jan. 2022.

SAMUELSON, P.;NORDHAUS, W. Economics. 12th Edition, McGraw-Hill, New York. 1985.

SANTOS, H. G.; ZAMPIERI, F. G.; SILVA, K. N.; SILVA, G. T.; LIMA, A. C. P.; CAVALCANTI, A. B.; CHIAVEGATTO FILHO, A. D. P. Machine learning to predict 30-day quality-adjusted survival in critically ill patients with câncer. Journal of Critical Care, v. 55, p. 73–78. 2020. Disponível em: https://doi.org/10.1016/j.jcrc.2019.10.015. Acesso em 02 jan. 2023.

SCHWAB, K. A Quarta Revolução Industrial. São Paulo: Edipro, 2016. Disponível em: <http://www. edipro.com.br/produto/a-quarta-revolucaoindustrial/>. Acesso em 21 jul. 2021.

SILVA, E. D. S.; SANTOS, J. F. D.; PEROBELLI, F. F. C.; NAKAMURA, W. T. Capital structure of Brazil, Russia, India and China by economic crisis. RAM – Revista de Administração Mackenzie, v.17, n. 3, p. 105-131, 2016. Disponível em: https://doi.org/10.1590/1678-69712016/administracao.v17n3p105-131. Acesso em 19 jun. 2022.

SPENCE, A. M. Market signaling: Informational transfer in hiring and related screening processes. Cambridge, MA: Harvard University Press, 1974.

STIGLER, G. J.; FRIEDLAND, C. What can regulators regulate? The case of electricity. Journal of Law and Economics, v. 5, p. 1-16, out. 1971. Disponível em: < http://www.jstor.org/stable/725003> . Acesso em 14 nov. 2023.

STIGLITZ, J. E. The contributions of the economics of information to twentieth century economics. The Quarterly Journal of Economics, v.115, n. 4, p. 1441-1478. Nov. 2000. Disponível em: < https://www.jstor.org/stable/2586930>. Acesso em 06 ago. 2022.

STREIT, R. E.; BORENSTEIN, D. An agent-based simulation model for analyzing the governance of the Brazilian Financial System. Expert Systems with Applications, v. 36, n. 9, p. 11489-11501. Disponível em: https://doi.org/10.1016/j.eswa.2009.03.043. Acesso em 06 ago. 2022.

TEIXEIRA, B. C.; TATIANA, N. T.; FRANCISCO, C. N.; CHIAVEGATTO FILHO, A. D. P. Spatial Clusters of Cancer Mortality in Brazil: A Machine Learning Modeling Approach. International Journal of Public Health. v. 68. Jul. 2023. Disponível em: doi: 10.3389/ijph.2023.1604789. Acesso em 09 set. 2023.

VELJANOVSKI, C. Economic approaches to regulation. In: The Oxford Handbook of Regulation. 2010.

WICHMANN, R.M.; FERNANDES, F.T.; CHIAVEGATTO FILHO, A.D.P. Improving the performance of machine learning algorithms for health outcomes predictions in multicentric cohorts. Sci Rep 13, v. 1022. Jan. 2023. Disponível em: https://doi.org/10.1038/s41598-022-26467-6. Acesso em Jun. 2023.

APRÊNDICE I

Código Phyton - Análise Exploratória

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Carregar os dados

data = pd.read_csv("C:/Users/igorl/OneDrive/Área de Trabalho/Marilia/data.csv")

# Estatísticas Descritivas para Variáveis Numéricas
print("\nEstatísticas Descritivas das Variáveis Numéricas:")
print(data.describe())

# Configurar estilo dos gráficos
sns.set(style="whitegrid")

# Função para plotar gráfico de barras
def plot_bar_chart(data, column, title, xlabel, ylabel, rotation=0):
    counts = data[column].value_counts()
    plt.figure(figsize=(10,6))
    sns.barplot(x=counts.index, y=counts.values)
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel(ylabel)
    plt.xticks(rotation=rotation)
    plt.show()

# Gráfico de Distribuição de Gênero
plot_bar_chart(data, 'Sexo', 'Distribuição de Gênero', 'Gênero', 'Contagem')

# Histograma de Distribuição de Idades
plt.figure(figsize=(10, 6))
sns.histplot(data['Idade'], kde=True, bins=30)
plt.title('Distribuição de Idades')
plt.xlabel('Idade')
plt.ylabel('Contagem')
plt.show()

# Histograma de Distribuição de Renda
plt.figure(figsize=(10, 6))
sns.histplot(data['Renda'], kde=True, bins=30)
plt.title('Distribuição de Renda')
plt.xlabel('Renda')
plt.ylabel('Contagem')
plt.show()

# Gráfico de caixa da relação entre Diabetes e Idade
```

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Diabete', y='Idade', data=data)
plt.title('Relação entre Diabetes e Idade')
plt.xlabel('Diabete')
plt.ylabel('Idade')
plt.show()

# Histograma de idade por sexo com KDE
plt.figure(figsize=(14, 8))
sns.histplot(data[data['Sexo'] == 'Male']['Idade'], color="blue", label='Male', kde=True,
bins=30, alpha=0.6)
sns.histplot(data[data['Sexo'] == 'Female']['Idade'], color="orange", label='Female', kde=True,
bins=30, alpha=0.6)
plt.title('Distribuição de Idade por Sexo')
plt.xlabel('Idade')
plt.ylabel('Contagem')
plt.legend()
plt.show()

# Pairplot com variáveis numéricas e categóricas
num_vars = ['Idade', 'Renda', 'Cobranca_Operadora', 'Despesa_Operadora']
cat_vars = ['Sexo', 'Fumante', 'Diabete']
pairplot_data = data[num_vars + cat_vars]
sns.pairplot(pairplot_data, hue="Sexo", palette="husl")
plt.show()

# Boxplot de Despesas de Operadora por Fumante
plt.figure(figsize=(10, 6))
sns.boxplot(x='Fumante', y='Despesa_Operadora', data=data)
plt.title('Despesas da Operadora por Fumante')
plt.xlabel('Fumante')
plt.ylabel('Despesa da Operadora')
plt.show()

# Heatmap de Correlação
corr = data[num_vars].corr()
plt.figure(figsize=(10, 8))
sns.heatmap(corr, annot=True, fmt=".2f", cmap="coolwarm")
plt.title('Heatmap de Correlação das Variáveis Numéricas')
plt.show()
```

 # Análise agrupamento de risco

Kmeans

```
import pandas as pd
from sklearn.cluster import KMeans
from sklearn.model_selection import train_test_split
from sklearn.metrics import silhouette_score
```

```
import matplotlib.pyplot as plt

# Carregar os dados
data = pd.read_csv("C:/Users/igorl/OneDrive/Área de Trabalho/Marilia/data.csv")

# Selecione as colunas 'Despesa_Operadora' e 'Idade'
selected_columns = ['Despesa_Operadora', 'Idade']
data_selected = data[selected_columns]

# Dividindo os dados em treino e teste
data_train, data_test = train_test_split(data_selected, test_size=0.2, random_state=42)

# Encontrando o número ótimo de clusters
wcss = []  # Soma dos quadrados intra-cluster
silhouette_coefficients = []

for k in range(2, 11):  # Testar para k de 2 a 10
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(data_train)
    wcss.append(kmeans.inertia_)
    score = silhouette_score(data_train, kmeans.labels_)
    silhouette_coefficients.append(score)

# Método do Cotovelo
plt.figure(figsize=(10, 5))
plt.plot(range(2, 11), wcss, marker='o')
plt.title('Método do Cotovelo para Determinar o Número Ótimo de Clusters')
plt.xlabel('Número de Clusters')
plt.ylabel('WCSS')
plt.show()

# Coeficiente de Silhueta
plt.figure(figsize=(10, 5))
plt.plot(range(2, 11), silhouette_coefficients, marker='o')
plt.title('Coeficiente de Silhueta por Número de Clusters')
plt.xlabel('Número de Clusters')
plt.ylabel('Coeficiente de Silhueta')
plt.show()

# Escolha o número de clusters com base nas análises anteriores
# Por exemplo, se 3 é o número ótimo de clusters:
kmeans_optimal = KMeans(n_clusters=3, random_state=42)
kmeans_optimal.fit(data_train)

# Aplicar o modelo treinado no conjunto de teste
labels_test = kmeans_optimal.predict(data_test)

# Gráfico dos Grupos de Risco no Conjunto de Teste
plt.figure(figsize=(10, 6))
plt.scatter(data_test.iloc[:, 0], data_test.iloc[:, 1], c=labels_test, cmap='viridis')
```

```
plt.title('Grupos de Risco no Conjunto de Teste')
plt.xlabel('Despesa_Operadora')
plt.ylabel('Idade')
plt.show()
```

Fuzzy C-Means

```
import numpy as np
import pandas as pd
import skfuzzy as fuzz
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt

# Carregar os dados
data = pd.read_csv("C:/Users/igorl/OneDrive/Área de Trabalho/Marilia/data.csv")

# Selecionar colunas 'Despesa_Operadora' e 'Idade'
selected_columns = ['Despesa_Operadora', 'Idade']
data_selected = data[selected_columns].values

# Dividir os dados em conjuntos de treino e teste
data_train, _ = train_test_split(data_selected, test_size=0.2, random_state=42)

# Normalizar os dados de treino
data_train_norm = data_train / np.max(data_train, axis=0)

# Definir parâmetros para o Fuzzy C-means
m = 2.0  # Exponente para o pertencimento 'fuzzy'

# Vamos testar o número de cluster*s* de 1 a 9
fpcs = []

for n_cluster*s* in range(1, 10):
    cntr, u, u0, d, jm, p, fpc = fuzz.cluster.cmeans(
        data_train_norm.T, c=n_cluster*s*, m=m, error=0.005, maxiter=1000, init=None
    )
    fpcs.append(fpc)

# Plotar o gráfico do Coeficiente de Partição Fuzzy (FPC)
plt.figure(figsize=(10, 5))
plt.plot(np.arange(1, 10), fpcs, marker='o')
plt.title('O Coeficiente de Partição Fuzzy (FPC) para Diferentes Números de Cluster*s*')
plt.xlabel('Número de Cluster*s*')
plt.ylabel('FPC')
plt.show()

# Encontrar o número ótimo de cluster*s* e imprimir
optimal_n_cluster*s* = np.argmax(fpcs) + 1  # Adicionamos 1 porque o índice do array começa
em 0
```

```
print(f'O número ótimo de clusters é: {optimal_n_clusters}')

# Aplicar Fuzzy C-means com o número ótimo de clusters
cntr, u_optimal, _, _, _, _, _ = fuzz.cluster.cmeans(
    data_train_norm.T, c=optimal_n_clusters, m=m, error=0.005, maxiter=1000, init=None
)

# Plotar os resultados com o número ótimo de clusters
fig, ax0 = plt.subplots(figsize=(8, 6))
colors = ['b', 'orange', 'g', 'r', 'c', 'm', 'y', 'k', 'Brown', 'ForestGreen']

# Atribuir as cores aos pontos com base em sua probabilidade de pertencimento
for j in range(optimal_n_clusters):
    ax0.plot(data_train_norm[u_optimal.argmax(axis=0) == j, 0],
             data_train_norm[u_optimal.argmax(axis=0) == j, 1],
             '.', color=colors[j])

# Marcar os centros dos clusters
for pt in cntr:
    ax0.plot(pt[0], pt[1], 'rs')

ax0.set_title(f'Centros e Membros dos Clusters com {optimal_n_clusters} Clusters')
ax0.set_xlabel('Despesa_Operadora Normalizada')
ax0.set_ylabel('Idade Normalizada')
plt.show()
```

Modelos de algoritmo de aprendizagem de máquina

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import OneHotEncoder, MinMaxScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.linear_model import LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
import xgboost as xgb
import matplotlib.pyplot as plt
import seaborn as sns

# Carregar os dados
data = pd.read_csv("C:/Users/igorl/OneDrive/Área de Trabalho/Marilia/data.csv")

# Separar a variável alvo das preditoras
X = data.drop('Despesa_Operadora', axis=1)
y = data['Despesa_Operadora']
```

```
# Identificar colunas categóricas e numéricas
categorical_features = X.select_dtypes(include=['object', 'bool']).columns.tolist()
numerical_features = X.select_dtypes(exclude=['object', 'bool']).columns.tolist()

# Dividir os dados em conjuntos de treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Criar o pré-processador que aplica OneHotEncoder às variáveis categóricas e
MinMaxScaler às numéricas
preprocessor = ColumnTransformer(
    transformers=[
        ('num', MinMaxScaler(), numerical_features),
        ('cat', OneHotEncoder(), categorical_features)
    ])

# Preparando o scaler para a variável alvo
scaler_y = MinMaxScaler()
y_train_scaled = scaler_y.fit_transform(y_train.values.reshape(-1, 1))
y_test_scaled = scaler_y.transform(y_test.values.reshape(-1, 1))

# Definição dos modelos com seus respectivos parâmetros para o Grid Search
models = {
    'LinearRegression': (LinearRegression(), {}),
    'KNeighborsRegressor': (KNeighborsRegressor(), {'model__n_neighbors': [3, 5, 7]}),
    'XGBRegressor': (xgb.XGBRegressor(objective='reg:squarederror'),
{'model__n_estimators': [50, 100], 'model__max_depth': [3, 5]}),
    'RandomForestRegressor': (RandomForestRegressor(), {'model__n_estimators': [50, 100],
'model__max_depth': [3, 5]})
}

# Realizar Grid Search, calcular RMSE e plotar gráficos para cada modelo
for name, (model, params) in models.items():
    pipeline = Pipeline(steps=[('preprocessor', preprocessor),
                               ('model', model)])

    grid_search = GridSearchCV(pipeline, params, cv=5, scoring='neg_mean_squared_error',
n_jobs=-1)
    grid_search.fit(X_train, y_train_scaled.ravel())

    best_model = grid_search.best_estimator_
    y_test_pred_scaled = best_model.predict(X_test)

    # Inverter a transformação para obter as previsões nos valores originais
    y_test_pred = scaler_y.inverse_transform(y_test_pred_scaled.reshape(-1, 1)).flatten()

    # Calcular RMSE nos valores originais
    rmse_test = np.sqrt(mean_squared_error(y_test, y_test_pred))

    # Plotar o gráfico de histograma dos valores reais e curva de densidade das previsões
    plt.figure(figsize=(10, 6))
```

```
sns.histplot(y_test, bins=30, kde=False, color='blue', stat='density', label='Valores Reais')
sns.kdeplot(y_test_pred, color='red', label='Previsões', lw=2)
plt.title(f'{name} - RMSE no conjunto de teste: {rmse_test:.2f}')
plt.xlabel('Despesa Operadora')
plt.ylabel('Densidade')
plt.legend()
plt.show()
```