PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
NÚCLEO DE EDUCAÇÃO A DISTÂNCIA
Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data

Gabriel Fernando dos Santos Ramos

Predição da capacidade de obtenção de renda mínima necessária no Brasil utilizando técnicas de aprendizado de máquina supervisionado

Belo Horizonte
2026

Gabriel Fernando dos Santos Ramos

Predição da capacidade de obtenção de renda mínima necessária no Brasil utilizando técnicas de aprendizado de máquina supervisionado

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

Belo Horizonte
2026

SUMÁRIO

1. Introdução

1.1. Contextualização

A renda é um dos principais determinantes das condições de vida do indivíduo, pois influencia diretamente o acesso a bens e serviços essenciais — moradia, alimentação, educação, saúde, transporte e previdência social. Em consonância com essa centralidade, a Constituição Federal de 1988 estabelece, no artigo 7º, inciso IV, o conceito e a finalidade do salário-mínimo:

> salário-mínimo, fixado em lei, nacionalmente unificado, capaz de atender a suas necessidades vitais básicas e às de sua família com moradia, alimentação, educação, saúde, lazer, vestuário, higiene, transporte e previdência social, com reajustes periódicos que lhe preservem o poder aquisitivo, sendo vedada sua vinculação para qualquer fim (BRASIL, 1988).

Na prática, porém, o valor legal do salário-mínimo costuma ser insuficiente para garantir integralmente esse conjunto de necessidades. Com o objetivo de estimar um patamar remuneratório compatível com uma vida digna, o Departamento Intersindical de Estatística e Estudos Socioeconômicos (DIEESE) divulga periodicamente o indicador denominado Salário-Mínimo Necessário (SMN).

Para realizar esse cálculo, o DIEESE considera uma família padrão composta por dois adultos e duas crianças, assumindo-se que as duas crianças consomem, em conjunto, o equivalente a um adulto, resultando em três adultos equivalentes em termos de consumo alimentar. Em seguida, toma-se o valor da cesta básica mais elevada entre as capitais pesquisadas e multiplica-se por três, obtendo-se o custo mensal da alimentação familiar.

Com base na proporção que a alimentação representa no orçamento das famílias de baixa renda — aproximadamente 35%, segundo levantamento citado pelo próprio DIEESE a partir da Pesquisa de Orçamentos Familiares — estima-se o orçamento total necessário para cobrir, além da alimentação, despesas com moradia, vestuário, transporte, saúde, educação e lazer (DIEESE, 2016).

Apesar de sua relevância como parâmetro socioeconômico, o SMN é calculado considerando a renda de um único trabalhador responsável pelo sustento familiar. Contudo, observa-se uma transformação significativa na estrutura econômica dos domicílios, marcada pela redução do modelo tradicional de provedor único masculino e

pela expansão de arranjos familiares com dupla renda. De acordo com a Organização para a Cooperação e Desenvolvimento Econômico (OCDE, 2024), o aumento da participação feminina no mercado de trabalho tem contribuído diretamente para o crescimento de lares nos quais homens e mulheres compartilham a responsabilidade financeira. Dessa forma, a utilização direta do valor integral do SMN pode superestimar a insuficiência de renda quando se analisa a renda individual de cada trabalhador.

Diante disso, torna-se pertinente adaptar esse indicador à realidade observada nos dados, especialmente quando se trabalha com microdados individuais, como os da PNAD Contínua. Nesse contexto, a aplicação de técnicas de aprendizado de máquina surge como uma abordagem promissora para analisar, de forma integrada, múltiplas características demográficas, educacionais e ocupacionais, permitindo identificar padrões associados à capacidade de obtenção de renda suficiente para atender às necessidades básicas definidas constitucionalmente.

1.2. O problema proposto

De acordo com o Relatório da Desigualdade Global 2026, produzido pelo World Inequality Lab e coordenado por pesquisadores como Thomas Piketty, o Brasil figura como o quinto país com maior desigualdade de renda entre 216 nações analisadas. No contexto brasileiro, os 10% mais ricos concentram cerca de 59,1% da renda nacional, ao passo que os 50% mais pobres detêm apenas cerca de 9,3% dessa renda, evidenciando um elevado grau de concentração de recursos e uma distribuição altamente assimétrica (WORLD INEQUALITY LAB, 2025).

Nesse contexto, a identificação objetiva de indivíduos em situação de insuficiência de renda assume papel central para a análise da vulnerabilidade socioeconômica. A mensuração adequada dessa condição permite não apenas dimensionar a magnitude do problema, mas também compreender os fatores associados à incapacidade de atingir um patamar mínimo de bem-estar econômico. Trata-se de questão relevante para a formulação, focalização e avaliação de políticas públicas, especialmente programas de transferência de renda e estratégias de inclusão produtiva, em consonância com os princípios constitucionais de redução das desigualdades sociais e promoção da dignidade da pessoa humana.

Diante desse cenário, o problema investigado neste trabalho consiste em classificar indivíduos adultos quanto à sua capacidade de obtenção de renda suficiente para atender ao patamar de necessidades básicas definido pelo Salário Mínimo Necessário (SMN), conforme metodologia do DIEESE, ajustado à realidade de domicílios com dois provedores de renda. A variável alvo assume natureza binária, distinguindo indivíduos que atingem, classe 1, ou não atingem, classe 0, esse nível mínimo considerado adequado.

A análise considera indivíduos com 18 anos ou mais, após filtragem, conforme os registros disponíveis nos microdados da PNAD Contínua, abrangendo uma ampla diversidade de perfis socioeconômicos, incluindo diferentes regiões do país, níveis de escolaridade, setores de ocupação e condições domiciliares. Essa abrangência permite capturar heterogeneidades estruturais presentes na população brasileira.

A definição do limiar de renda adotado baseia-se no valor do Salário Mínimo Necessário, no valor de R$ 6.641,58, divulgado pelo DIEESE para janeiro de 2023, ajustado à hipótese de existência de dois provedores por domicílio, obtendo-se R$ 3.320,79, com o objetivo de avaliar se a renda individual atinge um patamar mínimo considerado compatível com a manutenção das despesas básicas do núcleo familiar (DIEESE, 2023). A escolha do ano de referência 2023 justifica-se pela disponibilidade de dados aprofundados sobre renda e caracterização socioeconômica no âmbito da PNAD Contínua, uma vez que não foram encontradas publicações equivalentes para 2024 ou anos posteriores com cobertura detalhada das variáveis de trabalho e rendimento necessárias para a modelagem adotada. Os microdados utilizados neste estudo foram obtidos a partir dos arquivos disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no repositório oficial de dados da PNAD Contínua, os quais contêm as informações referentes às variáveis socioeconômicas consideradas na análise (IBGE, 2023).

O estudo possui abrangência nacional, contemplando todo o território brasileiro, o que possibilita a análise de desigualdades regionais por meio da incorporação de variáveis geográficas e características do domicílio. Essa perspectiva permite identificar padrões socioeconômicos distintos entre regiões, áreas urbanas e rurais, bem como entre diferentes contextos habitacionais.

O problema é abordado por meio da aplicação de técnicas de aprendizado de máquina supervisionado, com o treinamento de modelos de classificação a partir dos dados da PNAD Contínua.

1.3. Objetivos

O objetivo geral deste trabalho é desenvolver e avaliar um modelo de aprendizado de máquina capaz de classificar indivíduos adultos quanto à sua capacidade de obtenção de renda suficiente para atingir o SMN, conforme metodologia definida pelo DIEESE e ajustada à realidade de domicílios brasileiros sustentados, em média, por dois provedores de renda, utilizando dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua).

2. Coleta de Dados

Os dados utilizados neste trabalho foram obtidos a partir dos microdados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A PNAD Contínua é uma pesquisa amostral de abrangência nacional, com periodicidade contínua, cujo objetivo é produzir indicadores sobre características demográficas, educacionais, ocupacionais e de rendimento da população brasileira.

Os microdados da PNAD Contínua são disponibilizados publicamente no portal do IBGE, em formato de arquivos de dados estruturados, acompanhados de dicionário de variáveis e documentação metodológica. Os dados são disponibilizados em formato de texto de largura fixa (fixed-width text file), no qual cada linha do arquivo representa um indivíduo entrevistado e cada variável ocupa posições pré-definidas de início e fim dentro da linha. Diferentemente de formatos delimitados, como CSV, esse tipo de arquivo exige que a leitura seja realizada com base nas posições exatas de cada campo, conforme especificado no dicionário de variáveis fornecido pelo IBGE.

Após o download, o arquivo de microdados foi armazenado no Google Drive, possibilitando o acesso e processamento dos dados em ambiente Google Colab, que utiliza arquivo com extensão ipynb. Essa estratégia facilita a manipulação de arquivos de grande volume e assegura a persistência dos dados ao longo das etapas de

análise. Em seguida, foi definida uma estrutura contendo as posições iniciais e finais das variáveis de interesse (colspecs) e uma lista com os respectivos nomes das colunas (names), ambas construídas a partir do dicionário oficial da PNAD Contínua.

A leitura do arquivo foi realizada utilizando a biblioteca pandas da linguagem Python, por meio da função read_fwf, que é específica para arquivos de largura fixa. Esse procedimento permite converter o arquivo texto bruto em um DataFrame, estrutura tabular que facilita o tratamento, a filtragem e a análise estatística dos dados. O comando utilizado é ilustrado a seguir:

Figura 1 - Criação do DataFrame principal

```
df = pd.read_fwf("/content/drive/MyDrive/Arquivo_IBGE/PNADC_2023_trimestre1.txt",
                 colspecs=colunas_especificas,
                 names=colunas_nomes)
```

Devido à grande quantidade de variáveis disponíveis nos microdados, foi realizada uma seleção de colunas específicas, com base na relevância para o problema proposto.

Figura 2 - Criação das colunas específicas e posição das features no banco de dados

```
colunas_especificas = [
    (5, 7), # Unidade da Federação
    (32, 33), # Situação do domicílio, urbano ou rural
    (33, 34), # Domicílio em Região Metropolina
    (88, 90), # Número de pessoas no domicílio
    (94, 95), # Sexo
    (103, 106), # Idade de 0 a 130
    (106, 107), # Cor ou raça
    (151, 155), # Qual era o seu trabalho?
    (155, 156), # Qual setor trabalha?
    (124, 126), # Qual foi o curso mais elevado que frequentou anteriormente?
    (240, 243), # Quantas horas você trabalhava no serviço principal por semana?
    (246, 247), # Quanto tempo já trabalhou neste último emprego? (até 2 anos)
    (251, 253), # Quanto tempo já trabalhou neste último emprego? (anos)
    (579, 587) # Total do Rendimento Individual
]
```

As variáveis selecionadas abrangem aspectos:

a) Demográficos: esse grupo compreende características individuais básicas dos respondentes, como idade, sexo e cor ou raça. Essas variáveis são amplamente utilizadas por influenciarem diretamente as oportunidades no mercado de trabalho, os

níveis de remuneração e as desigualdades estruturais observadas na distribuição de renda. No contexto de modelos preditivos, tais atributos auxiliam na captura de padrões demográficos associados à probabilidade de um indivíduo alcançar rendimentos iguais ou superiores ao SMN.

b) Educacionais: o nível de escolaridade representa um dos principais determinantes da renda do trabalho, estando fortemente associado à qualificação profissional, produtividade e acesso a ocupações mais bem remuneradas. A inclusão dessa variável permite ao modelo incorporar o impacto do capital humano na geração de renda, além de possibilitar análises comparativas entre diferentes níveis educacionais.

c) Ocupacionais: as variáveis ocupacionais abrangem informações relacionadas ao vínculo e às condições de trabalho do indivíduo, como setor de atuação, tipo de ocupação, número de horas trabalhadas por semana, quantidade de trabalhos exercidos e rendimento individual total. Esse conjunto é central para o problema proposto, uma vez que a renda é diretamente determinada pelas características do trabalho exercido. A partir da variável de rendimento individual total, foi derivada a variável alvo possui_smn, que indica se o indivíduo alcança ou não o patamar do SMN.

d) Domiciliares: as características do domicílio, como tamanho da família, situação urbana ou rural e localização regional, fornecem contexto socioeconômico adicional ao indivíduo. Essas variáveis são relevantes porque influenciam tanto o custo de vida quanto as oportunidades de emprego disponíveis, além de refletirem desigualdades regionais e estruturais presentes no país. Sua inclusão contribui para uma modelagem mais realista e contextualizada do fenômeno estudado.

Segue tabela com descrição e tipo de variável:

Tabela 1 - Descrição detalhada de todas as variáveis

Nome da Coluna	Descrição	Tipo de Variável
uf	Unidade da Federação, posteriormente agrupada por região geográfica	Categórica nominal
zona_domicilio	Situação do domicílio (urbano ou rural)	Categórica nominal

regiao_domicilio	Indicação de domicílio em região metropolitana ou não	Categórica nominal
tam_domicilio	Número de pessoas residentes no domicílio (agrupado em faixas)	Categórica nominal
sexo	Sexo do indivíduo	Categórica nominal
idade	Idade do indivíduo em anos completos	Numérica discreta
cor/raca	Cor ou raça declarada	Categórica nominal
qual_trabalho	Tipo de ocupação principal, agrupada por grandes categorias	Categórica nominal
setor_trabalha	Setor de ocupação (formal, informal, público ou privado)	Categórica nominal
escolaridade	Nível mais elevado de escolaridade frequentado	Categórica ordinal
qnt_hrs_trabalha	Quantidade de horas trabalhadas semanalmente (agrupadas em faixas)	Categórica nominal
qnts_anos_trabalha	Tempo total de permanência no trabalho atual (agrupado em faixas)	Categórica nominal
renda_individual_total	Rendimento individual total mensal	Numérica contínua
possui_smn	Indicador binário de suficiência de renda em relação ao SMN ajustado	Binária

3. Processamento/Tratamento de Dados

3.1. Variável Alvo (Target)

A variável alvo originalmente apresentava natureza numérica contínua, correspondente ao valor da renda individual total declarada pelo respondente. Entretanto, como o presente estudo tem por objetivo a construção de um modelo de classificação supervisionada, tornou-se necessária a transformação dessa variável em uma categoria nominal binária, indicando se o indivíduo pertence ou não ao grupo de alta renda, conforme critério definido em relação ao SMN.

Considerando que, em modelos de aprendizado de máquina supervisionado do tipo classificação, não é admissível a existência de valores ausentes na variável dependente, comprovado na etapa de exploração inicial dos dados, optou-se pela remoção direta dos registros incompletos referentes à variável alvo (*renda individual total*).

Figura 3 - Formatação da variável alvo

```
# Salário mínimo necessário de acordo com DIEESE em janeiro de 2023: R$ 6.641,58
smn_dieese = 6641.58
smn_dieese_2provedores = smn_dieese / 2

# Criação da nova coluna, possui_smn, e remoção de linhas nas quais y (target) é ausente
df = df.dropna(subset='renda_individual_total')
df['possui_smn'] = (smn_dieese_2provedores <= df['renda_individual_total']).astype(int)
df = df.drop(columns='renda_individual_total') # remoção da coluna com dados brutos
```

3.2. Features

Em relação às demais variáveis que apresentaram valores ausentes, adotou-se uma estratégia de imputação diferenciada, de acordo com a natureza dos dados. Para as variáveis numéricas, empregou-se a imputação pela mediana, em virtude de sua robustez frente à presença de valores extremos. Para as variáveis categóricas, utilizou-se a moda, por representar a categoria mais frequente e preservar a distribuição original dos dados.

Com o objetivo de garantir a padronização do pré-processamento, evitar vazamento de dados (*data leakage*) e assegurar a aplicação consistente das transformações tanto no conjunto de treinamento quanto no de teste, optou-se pela utilização de *pipelines* ao longo de todo o processo de preparação dos dados, uma vez que essa abordagem favorece a modularidade, a autonomia dos componentes e a robustez do sistema como um todo (GÉRON, 2019, p. 38).

No tratamento das variáveis categóricas, realizou-se inicialmente o agrupamento de categorias com baixa frequência, visando reduzir o ruído nos dados e mitigar o risco de *overfitting* durante o treinamento dos modelos. Em seguida, as variáveis categóricas nominais foram codificadas por meio do método One-Hot Encoding, com a exclusão de uma categoria de referência, a fim de evitar o problema de multicolinearidade, conhecido como *dummy variable trap*.

Figura 4 - Criação de *pipeline* das variáveis categóricas nominais

```
# colunas categóricas nominal
cat_cols = ['uf', 'cor/raca', 'setor_trabalha', 'tam_domicilio',
            'sexo', 'zona_domicilio', 'regiao_domicilio', 'qual_trabalho',
            'qnt_hrs_trabalha', 'qnts_anos_trabalha']

# pipeline para cat_cols
cat_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(drop='first', sparse_output=False))
])
```

Ademais, as formatações aplicadas para essas variáveis categórica são descritas detalhadamente na tabela a seguir:

Tabela 2 - Variáveis antes e após formatação

Nome da Variável	Dados Brutos	Formatação
uf	Separados por estado	Agrupados por região do Brasil
zona_domicílio	[1] para urbano e [2] para rural	Sem alteração
regiao_domicílio	Separados por capital, região metropolitana, RIDE e interior	Binário sobre ser ou não região metropolitana
Tam_domicílio	Dados numéricos discretos	Agrupados: 1, 2-3, 4-5, 6+ indivíduos
sexo	[1] para masculino e [2] para feminino	Sem alteração
cor/raca	Dados categóricos em forma de números	Dados categóricos com nomes de cor/raca
qual_trabalho	Trabalho Específico	Agrupamento por 10 (dez) grande área de conhecimento do trabalho
setor_trabalha	Modelo de trabalho, sendo servidor público, informal, empregador, setor privado	Agrupamento de trabalhador domésticos em informal e militar para servidor público

qnt_hrs_trabalha	Variável numérica discreta	Reorganizado em 4 (quatro) grupos: 30hrs, 40hrs, 50hrs ou +50hrs.
qnts_anos_trabalha	Duas variáveis numéricas discreta, sendo uma com até 24 meses de trabalho e a outra com 2 anos ou mais	Reorganizado em 4 (quatro) grupos e em uma variável: 2anos, 5anos, 10anos ou +10anos

Destaca-se a variável categórica ordinal, escolaridade, para a qual foi empregado o método Ordinal Encoding, uma vez que suas categorias apresentam uma relação hierárquica natural. Essa abordagem permite preservar a ordem intrínseca entre os níveis educacionais, possibilitando que o modelo capture adequadamente a progressão existente entre as categorias.

Figura 5 - Criação de *pipeline* das variáveis categóricas ordinais

```
# colunas categóricas ordinal
ord_cols = ['escolaridade']

# pipeline para ord_cols
ord_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OrdinalEncoder(categories=[['fundamental', 'medio', 'superior']]))
])
```

Para a variável numérica, idade, aplicou-se um pipeline específico de pré-processamento. Os valores ausentes foram tratados por meio de imputação pela mediana em razão de sua robustez frente à presença de valores extremos. Em seguida, realizou-se uma transformação logarítmica, com o objetivo de reduzir a assimetria da distribuição e a influência de valores extremos. Então, aplicou-se a padronização dos dados, assegurando média zero e desvio padrão unitário, de modo a garantir que a variável estivesse em uma escala padronizada e adequada ao treinamento dos modelos.

Figura 6 - Criação de *pipeline* das variáveis numéricas discretas

```
# colunas númericas discretas
num_cols = ['idade']

# pipeline para num_cols
num_pipeline = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('log1p', FunctionTransformer(func=np.log1p, validate=False)),
    ('scaler', StandardScaler())
])
```

Por fim, realizou a ordenação e o agrupamento dos *pipelines* em um *pipeline* maior: *preprocessing*.

Figura 7 - Criação de *pipeline* agrupando os outros *pipelines*

```
# junção dos pipelines para cada tipo de dados
preprocessing = ColumnTransformer(
    transformers=[
        ('num', num_pipeline, num_cols),
        ('cat', cat_pipeline, cat_cols),
        ('ord', ord_pipeline, ord_cols)
    ]
)
```

4. Análise e Exploração dos Dados

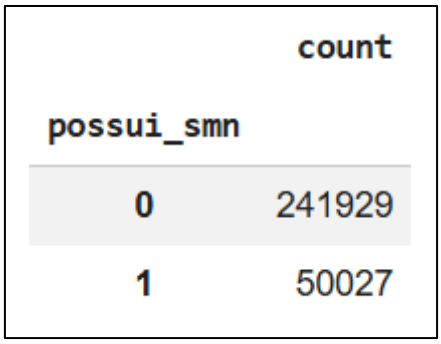
Para uma realização eficiente de tratamento dos dados, realizou-se uma análise exploratória inicial dos registros, com o objetivo de compreender a estrutura do conjunto de dados, identificar estatísticas descritivas e verificar a presença de valores ausentes.

O conjunto de dados analisado é composto por 473.335 registros, abrangendo informações socioeconômicas e laborais dos indivíduos. Observou-se a presença significativa de valores ausentes, sendo que 273.229 registros apresentam dados faltantes em variáveis relacionadas à situação de trabalho, o que demandou cuidados específicos durante a etapa de pré-processamento visto anteriormente. Além disso, identificou-se que 173.288 registros possuem ausência na variável alvo, denominada renda individual total, o que implicou em remoção direta dos registros incompletos.

Após o tratamento dos dados e a definição da variável alvo, constatou-se que esta apresenta desbalanceamento de classes, com aproximadamente 17% dos

indivíduos pertencendo à classe positiva, de acordo com a figura a seguir, correspondente às pessoas que recebem até um Salário Mínimo Necessário (SMN), modificado para a condição de dois provedores de renda no domicílio. Essa assimetria na distribuição das classes reforça a inadequação do uso exclusivo da acurácia como métrica de avaliação e justifica a adoção de métricas mais informativas, conforme demonstrado no capítulo seguinte.

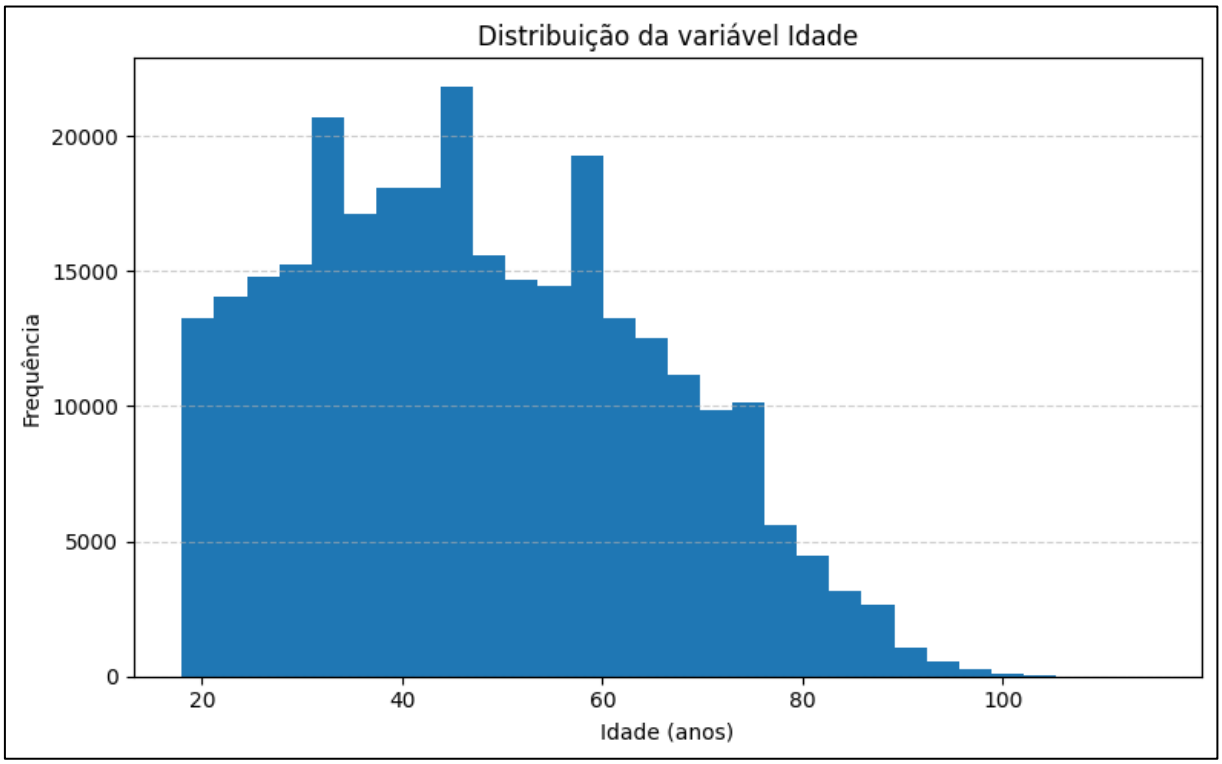
Figura 8 – Contagem entre as classes



possui_smn	count
0	241929
1	50027

Além disso, a análise da variável idade evidenciou uma distribuição assimétrica à direita, caracterizada pela presença de uma cauda longa, conforme ilustrado no gráfico apresentado adiante. Também foi constatada a existência de valores de renda individual de pessoas com menos de 18 anos, o que demandou filtragem para evitar ruídos.

Gráfico 1 - Frequência da variável idade



No que se refere às variáveis categóricas, identificou-se a presença de categorias com baixa frequência, sendo necessário o seu agrupamento, bem como a existência de uma variável categórica ordinal, escolaridade, cujas categorias apresentam uma relação hierárquica natural. Essa estratégia visa mitigar problemas de não representatividade e ruído amostral no conjunto de treinamento, os quais comprometem a capacidade de generalização dos modelos de aprendizado de máquina (GÉRON, 2019, p. 25).

5. Criação de Modelos de Machine Learning

Antes da criação dos modelos de machine learning, dividiu-se os dados por meio da função train_test_split de maneira estratificada, conforme apresentado a seguir:

Figura 9 - Separação dos dados em treino e teste

```
x_train, x_test, y_train, y_test = train_test_split(x, y,
                                                    train_size=0.8,
                                                    random_state=42,
                                                    stratify=y)
```

Essa etapa tem como objetivo separar o conjunto de dados original em dois subconjuntos distintos: treinamento e teste, permitindo avaliar a capacidade de generalização dos modelos em dados não vistos durante o treinamento.

A divisão dos dados em conjuntos de treinamento e teste foi realizada de modo a garantir tanto a capacidade de aprendizado dos modelos quanto a confiabilidade da avaliação de desempenho. Para isso, definiu-se que 80% das observações comporiam o conjunto de treinamento, enquanto os 20% restantes seriam reservados para o conjunto de teste.

Além disso, empregou-se a estratificação da variável alvo de forma a preservar a proporção original das classes tanto no conjunto de treinamento quanto no de teste. Esse procedimento é particularmente relevante em cenários de desbalanceamento de classes, pois evita que determinadas categorias fiquem sub-representadas em um dos subconjuntos. Ao manter a distribuição das classes, reduz-se o risco de viés na avaliação do modelo, garantindo que as métricas de desempenho reflitam de maneira mais fiel o comportamento do classificador em dados reais.

5.1. Regressor Logistic, Random Forest e Gradient Boosting

Com o objetivo de explorar diferentes perspectivas de modelagem e reduzir limitações inerentes a algoritmos individuais, optou-se pela utilização de três classificadores com naturezas distintas: a Regressão Logística, o Random Forest e o Gradient Boosting. A Regressão Logística fornece uma abordagem linear e interpretável, servindo como um modelo de referência robusto; o Random Forest é capaz de capturar relações não lineares e interações complexas por meio da agregação de múltiplas árvores independentes; enquanto o Gradient Boosting apresenta elevada capacidade preditiva ao ajustar modelos sequenciais que corrigem erros anteriores.

Vale ressaltar, ainda, que, no modelo de regressão logística, foi adotada a penalidade L2 como mecanismo de regularização, sendo o parâmetro C = 0,1

responsável por controlar a intensidade dessa regularização. Nesse contexto, valores menores de C implicam uma penalização mais forte sobre os coeficientes do modelo, favorecendo soluções mais simples e reduzindo o risco de *overfitting*. A escolha de C = 0,1 busca, portanto, um maior controle da complexidade do modelo, contribuindo para uma melhor capacidade de generalização em dados não observados.

Figura 10 - Criação do modelo de regressão logística

```
log_reg = LogisticRegression(
    penalty='l2',
    C=0.1,
    class_weight='balanced',
    max_iter=1000,
    random_state=42
)
```

Adicionalmente, os modelos Random Forest e Gradient Boosting utilizaram o hiperparâmetro, definido antes do treinamento, max_depth com o objetivo de controlar a complexidade das árvores de decisão que compõem esses algoritmos. A limitação da profundidade máxima das árvores atua como um mecanismo de regularização, ao restringir o grau de liberdade do modelo durante o processo de treinamento. Conforme destacado por Géron (2019, p. 178), árvores excessivamente profundas tendem a se ajustar de forma demasiada aos dados de treinamento, capturando ruídos e padrões específicos da amostra, o que aumenta o risco de sobreajuste. Dessa forma, ao reduzir o valor de max_depth, busca-se obter modelos mais simples e estáveis, capazes de generalizar melhor para dados não observados, mantendo um equilíbrio adequado entre viés e variância.

5.2. Ensemble por Votação (Voting Classifier)

A combinação desses classificadores foi realizada por meio de um *Voting Classifier* com votação suave (*soft voting*), estratégia que se baseia na agregação das probabilidades previstas por cada modelo. Essa abordagem fundamenta-se no princípio de que a combinação de múltiplos classificadores, mesmo que individualmente apresentem desempenho apenas moderadamente superior ao acaso, pode resultar em ganhos significativos de acurácia, desde que seus erros não sejam perfeitamente

correlacionados. De acordo com Géron (2019, p. 187), ensembles tendem a apresentar melhor desempenho estatístico à medida que agregam decisões de modelos distintos, pois a votação majoritária atua como um mecanismo de redução de erro e aumento da estabilidade das previsões, desde que haja diversidade suficiente entre os classificadores.

Figura 11 - Criação do modelo de Voting Classifier

```
voting_clf = VotingClassifier(
    estimators=[
        ('log_reg', log_reg),
        ('rf', rf),
        ('gb', gb)
    ],
    voting='soft',
    weights=[1, 1, 1]
)

model = Pipeline(steps=[
    ('preprocessing', preprocessing),
    ('voting_clf', voting_clf)
])
```

A avaliação dos modelos na base de treinamento foi conduzida por meio de validação cruzada estratificada, conforme figura a seguir, utilizando o objeto StratifiedKFold com três partições. Essa estratégia divide o conjunto de treinamento em três subconjuntos, garantindo que, em cada divisão, a proporção das classes da variável alvo seja preservada. A opção por embaralhar os dados antes da divisão contribui para reduzir possíveis vieses decorrentes da ordem original das observações, enquanto o uso de um valor fixo para random_state assegura a reprodutibilidade dos resultados obtidos ao longo das diferentes execuções do experimento.

Figura 12 - Validação cruzada estratificada

```
skf = StratifiedKFold(
    n_splits=3,
    shuffle=True,
    random_state=42
)
```

5.3. Avaliação dos modelos na base de treinamento

Para a mensuração do desempenho dos modelos, foram selecionadas as métricas de precisão (precision), revocação (recall) e F1-score, definidas no dicionário scoring. A escolha dessas métricas se justifica pelo fato de o problema envolver um cenário de classificação em que o simples uso da acurácia é insuficiente, pelo fato do desbalanceamento entre as classes. A precisão avalia a proporção de previsões positivas corretas, o recall mede a capacidade do modelo em identificar corretamente as instâncias da classe positiva, e o F1-score representa a média harmônica entre precisão e recall, fornecendo uma medida balanceada do desempenho.

Os modelos avaliados — Regressão Logística, Random Forest, Gradient Boosting e o classificador ensemble baseado em votação — foram organizados em uma estrutura de dicionário, o que permitiu a iteração sistemática sobre cada algoritmo durante o processo de validação. Para os modelos individuais, foi construído um *Pipeline* específico que integra as etapas de pré-processamento e o classificador propriamente dito, assegurando que todas as transformações sejam aplicadas corretamente em cada partição da validação cruzada. No caso do modelo de votação, utilizou-se diretamente o pipeline previamente definido, uma vez que este já encapsula tanto o pré-processamento quanto o ensemble de classificadores.

Figura 13 - Métodos de mensuração de desempenho e modelos

```
scoring = {
    'precision': 'precision',
    'recall': 'recall',
    'f1': 'f1'
}

modelos = {
    'log_reg': log_reg,
    'rf': rf,
    'gb': gb,
    'voting': model
}
```

A função cross_validate foi então empregada para executar a validação cruzada sobre o conjunto de treinamento, aplicando as métricas definidas e utilizando paralelização (n_jobs=-1) para otimizar o tempo de processamento. Para cada modelo, os valores de precisão, recall e F1-score obtidos nas diferentes partições foram agregados por meio do cálculo da média, resultando em uma estimativa mais estável e confiável do desempenho. Por fim, esses resultados foram armazenados em uma estrutura de dados que permite a comparação direta entre os modelos, subsidiando a análise do classificador mais adequado para o problema em estudo.

Figura 14 - Validação cruzada na base de treinamento

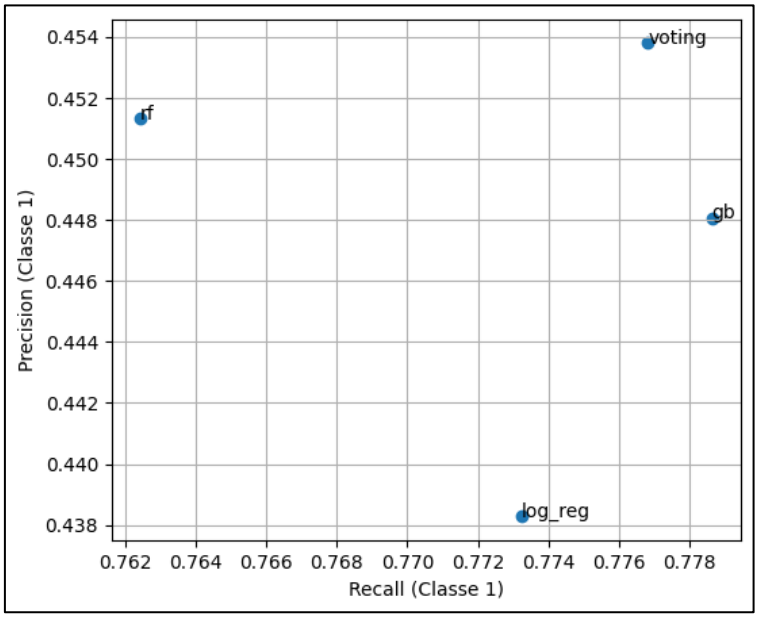
```
cv_results = cross_validate(
    pipe_individual,
    x_train,
    y_train,
    cv=skf,
    scoring=scoring,
    n_jobs=-1,
    return_train_score=False
)
```

Cabe destacar que todo o processo de validação e comparação entre os modelos foi realizado exclusivamente sobre a base de treinamento. Essa escolha metodológica tem como objetivo evitar o chamado data leakage, isto é, impedir que informações da base de teste influenciem direta ou indiretamente o ajuste dos modelos ou a seleção de hiperparâmetros. Ao conduzir os testes apenas no conjunto de treinamento, garante-se que a base de teste permaneça completamente isolada e seja utilizada apenas na etapa final de avaliação, funcionando como uma estimativa imparcial do desempenho do modelo em dados verdadeiramente não observados.

Essa abordagem é fundamental para assegurar a validade dos resultados obtidos, uma vez que observar repetidamente a base de teste durante o desenvolvimento do modelo pode levar a decisões excessivamente ajustadas a esse conjunto específico, resultando em uma superestimação do desempenho. Dessa forma, a utilização da validação cruzada no conjunto de treinamento permite explorar diferentes partições dos dados para avaliação interna, mantendo a base de teste como uma referência final e independente.

Continuamente, o gráfico a seguir ilustra a relação entre precisão e recall para a classe positiva (Classe 1) dos diferentes modelos avaliados na base de treinamento, permitindo uma análise comparativa do desempenho sob a ótica do compromisso entre essas duas métricas. Cada ponto representa um classificador, posicionado de acordo com sua capacidade de identificar corretamente as instâncias positivas (recall) e, ao mesmo tempo, manter um baixo número de falsos positivos (precisão). Quanto mais próximo o ponto estiver do canto superior direito do gráfico, melhor tende a ser o desempenho do modelo nesse equilíbrio.

Gráfico 2 - Precisão x Revocação com os quatro modelos na base de treinamento



Observa-se que a Regressão Logística apresenta o maior valor de recall entre os modelos individuais, indicando maior sensibilidade na identificação da classe positiva. No entanto, esse ganho em recall ocorre à custa de uma menor precisão, o que sugere um aumento na taxa de falsos positivos. Esse comportamento é típico de modelos lineares em cenários de fronteiras de decisão menos complexas, nos quais a priorização da sensibilidade pode comprometer a exatidão das previsões positivas.

O Random Forest apresenta um comportamento mais conservador, com recall ligeiramente inferior ao da regressão logística, porém com ganho em precisão. Esse resultado indica que o modelo tende a realizar previsões positivas mais confiáveis, ainda que identifique uma proporção menor de todos os casos positivos existentes.

Tal padrão é coerente com a natureza do algoritmo, que busca reduzir variância por meio da agregação de múltiplas árvores, favorecendo previsões mais estáveis.

O Gradient Boosting destaca-se por alcançar um dos maiores valores de recall, mantendo, ao mesmo tempo, uma precisão superior à da regressão logística. Esse resultado evidencia a capacidade do método em capturar padrões mais complexos nos dados a custo de interpretabilidade, ajustando-se de forma iterativa para corrigir erros anteriores, o que contribui para um melhor equilíbrio entre sensibilidade e exatidão.

O Voting Classifier, por sua vez, apresenta o melhor desempenho global entre os modelos avaliados, conforme indicado pelos maiores valores de precisão e F1-score, além de um recall elevado e estável. A posição desse modelo no gráfico reflete a vantagem da abordagem de ensemble, na qual a combinação de classificadores com características distintas permite reduzir erros individuais e produzir previsões mais robustas. Esse resultado corrobora a fundamentação teórica de que a agregação de modelos diversos tende a melhorar o desempenho preditivo, especialmente quando seus erros não são perfeitamente correlacionados.

A análise dos valores de F1-score reforça essa conclusão, uma vez que o Voting Classifier atinge o maior valor entre os modelos (0,5729), indicando o melhor compromisso entre precisão e recall. Assim, os resultados obtidos justificam a escolha do modelo ensemble como a alternativa mais adequada para a etapa final de avaliação no conjunto de teste, por apresentar maior equilíbrio e estabilidade no desempenho preditivo.

Figura 15 - Resultado dos quatro modelos na base de treinamento

	modelo	precision	recall	F1-score
0	log_reg	0.438285	0.773244	0.559458
1	rf	0.451310	0.762425	0.566993
2	gb	0.448052	0.778641	0.568793
3	voting	0.453799	0.776817	0.572912

6. Apresentação dos Resultados

A interpretação dos resultados do Voting Classifier foi realizada a partir da avaliação no conjunto de teste. A acurácia global obtida foi de aproximadamente 84,56%, indicando que o modelo classificou corretamente a maior parte das observações. Contudo, considerando o desbalanceamento da variável alvo, em que a classe positiva representa cerca de 17% da amostra, a acurácia isoladamente não é suficiente para uma avaliação adequada do desempenho, sendo necessário analisar métricas complementares que evidenciem o comportamento do modelo em cada classe.

Figura 16 - Resultado do Voting Classifier na base de teste

```
Acurácia:  0.8455610357583231
Classificação Geral:
              precision    recall  f1-score   support

           0       0.93      0.88      0.90     48386
           1       0.54      0.66      0.60     10006

    accuracy                           0.85     58392
   macro avg       0.73      0.77      0.75     58392
weighted avg       0.86      0.85      0.85     58392
```

Ao observar os resultados por classe, verifica-se que o modelo apresenta desempenho elevado para a classe 0 (indivíduos que não se enquadram na condição de até um SMN ajustado para dois provedores). Nessa classe, a precisão foi de 0,93, indicando que a grande maioria das previsões realizadas como classe 0 está correta. O recall de 0,88 evidencia que o modelo consegue identificar corretamente uma parcela significativa dos indivíduos pertencentes a essa classe, resultando em um F1-score de 0,90, valor que reflete equilíbrio entre precisão e sensibilidade. Esse comportamento era esperado, uma vez que se trata da classe majoritária no conjunto de dados, o que naturalmente favorece sua aprendizagem pelos algoritmos.

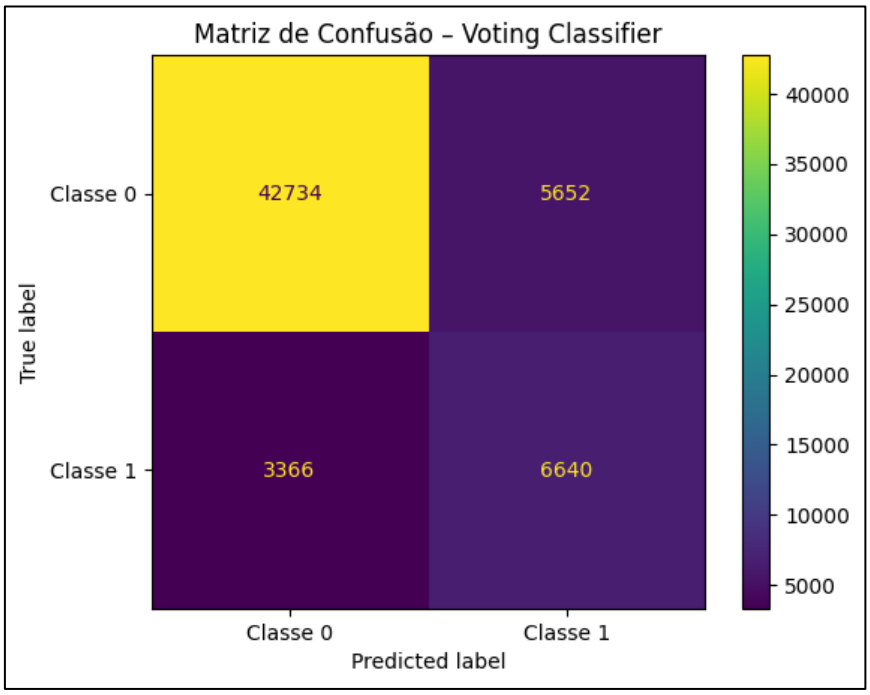
Em relação à classe 1, que representa o oposto da classe majoritária, observa-se um desempenho mais desafiador, porém consistente com a complexidade do problema. A precisão de 0,54 indica que, entre as observações classificadas como

pertencentes à classe 1, pouco mais da metade corresponde efetivamente a indivíduos dessa categoria. Por outro lado, o recall de 0,66 demonstra que o modelo é capaz de identificar aproximadamente dois terços dos casos reais da classe positiva, o que é particularmente relevante em um contexto de interesse social, no qual a não identificação de indivíduos em situação de maior vulnerabilidade pode ser mais crítica do que a ocorrência de falsos positivos. O F1-score de 0,60 sintetiza esse compromisso entre precisão e recall, evidenciando um desempenho equilibrado, ainda que inferior ao observado para a classe majoritária.

A análise das médias das métricas reforça essa interpretação. A média macro, que atribui o mesmo peso a ambas as classes, resultou em valores de 0,73 para precisão, 0,77 para recall e 0,75 para F1-score, revelando que, quando consideradas de forma equitativa, as classes apresentam um desempenho moderado a bom. Já a média ponderada, que leva em conta a proporção de observações em cada classe, apresentou valores elevados, com precisão, recall e F1-score em torno de 0,85, refletindo o bom desempenho geral do modelo e sua adequação ao padrão predominante dos dados.

A matriz de confusão fornece uma visão mais detalhada do comportamento do classificador. Observa-se que o modelo classificou corretamente 42.734 instâncias da classe 0 e 6.640 instâncias da classe 1. Em contrapartida, ocorreram 5.652 falsos positivos, nos quais indivíduos da classe 0 foram classificados como pertencentes à classe 1, e 3.366 falsos negativos, representando indivíduos da classe 1 que não foram identificados corretamente pelo modelo. Em um cenário de desbalanceamento, esse padrão indica que o modelo adota uma postura relativamente sensível à classe minoritária, aceitando um número maior de falsos positivos como forma de reduzir a quantidade de falsos negativos, o que está alinhado com a priorização do recall da classe de interesse.

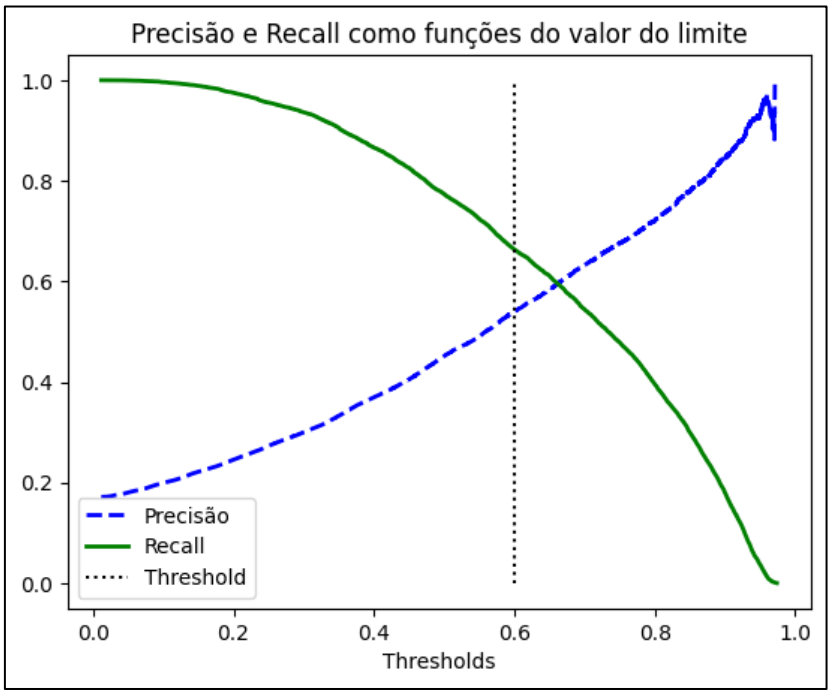
Gráfico 3 – Matriz de Confusão na base de teste



Posteriormente, analisou-se a relação entre precisão e recall em função da variação do limiar de decisão (*threshold*) utilizado pelo modelo para classificar uma observação como pertencente à classe positiva. Diferentemente da classificação padrão, que adota um limiar fixo de 0,5, essa análise permite compreender como o comportamento do classificador se altera à medida que se torna mais ou menos rigoroso na atribuição da classe de interesse.

Com base nessa avaliação, optou-se pela adoção de um limiar de decisão igual a 0,6, tornando o critério de classificação mais restritivo. Em termos práticos, isso significa que o modelo apenas classifica um indivíduo como pertencente à classe positiva quando a probabilidade estimada é igual ou superior a 60%. Essa estratégia foi adotada com o objetivo de aumentar a confiabilidade das classificações positivas, reduzindo a ocorrência de falsos positivos — isto é, casos em que o modelo indicaria que o indivíduo atinge o SMN quando, na realidade, não atinge.

Gráfico 4 - Precisão e Revocação em relação ao *Threshold*



Observa-se que, para valores baixos de *threshold*, o modelo tende a classificar um maior número de observações como pertencentes à classe positiva. Nesse cenário, o recall assume valores elevados, próximos de 1, indicando que a maioria dos indivíduos da classe positiva é corretamente identificada. Entretanto, essa maior sensibilidade ocorre à custa de uma precisão reduzida, uma vez que o número de falsos positivos aumenta, ou seja, muitos indivíduos da classe negativa passam a ser incorretamente classificados como positivos.

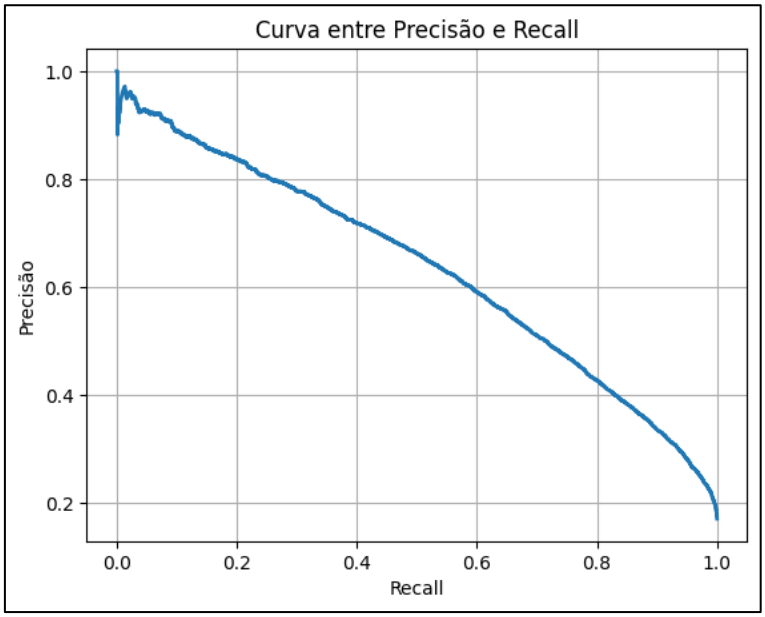
À medida que o valor do *threshold* aumenta, o modelo torna-se mais conservador na atribuição da classe positiva. Como consequência, a precisão cresce progressivamente, refletindo uma maior proporção de previsões positivas corretas, enquanto o recall diminui, indicando que parte dos casos reais da classe positiva deixa de ser identificada. Esse comportamento evidencia o *trade-off* clássico entre precisão e recall, no qual a melhoria de uma métrica implica, geralmente, a deterioração da outra.

A linha vertical pontilhada destaca o valor de *threshold* selecionado para a análise final do modelo, aproximadamente em 0,6. Nesse ponto, observa-se um equilíbrio relativamente adequado entre precisão e recall, com ambas as métricas apresentando valores intermediários e próximos entre si. A escolha desse limiar reflete uma decisão metodológica orientada à busca de um compromisso entre a redução de falsos positivos e a manutenção de uma sensibilidade aceitável para a classe minoritária,

especialmente relevante em um contexto de dados desbalanceados. Dessa forma, o gráfico evidencia que a definição do limiar de decisão exerce influência direta sobre o desempenho do classificador e deve ser realizada à luz dos objetivos do problema em estudo.

Uma outra maneira de analisar precisão e recall é criando um gráfico entre eles. De forma análoga ao que foi discutido na figura anterior, observa-se que, para *thresholds* mais elevados, o modelo torna-se mais conservador: apenas observações com alta probabilidade estimada são classificadas como pertencentes à classe positiva. Nesse cenário, a precisão permanece elevada, pois há poucos falsos positivos, porém o recall é reduzido, uma vez que muitos casos positivos reais não atingem o limiar exigido. Esse comportamento é equivalente à região inicial da curva, caracterizada por alta precisão e baixo recall.

Gráfico 5 – Curva entre Precisão e Revocação



Sendo assim, a Curva Precisão–Recall reforça visualmente as conclusões obtidas na análise do *threshold*, mostrando que o desempenho do modelo é fortemente dependente da escolha do limiar de decisão. A seleção de um *threshold* intermediário permite encontrar um ponto de equilíbrio entre precisão e *recall*, alinhando-se aos objetivos do estudo e às consequências práticas associadas a falsos positivos e falsos negativos.

6.1. Contextualização dos resultados à problemática

Os resultados apresentados a seguir contribuem para a formulação de estratégias voltadas, por exemplo, ao enfrentamento das desigualdades socioeconômicas. Os coeficientes estimados pelo modelo de regressão logística foram extraídos para análise interpretativa, estando detalhadamente apresentados na Tabela A1, no Apêndice.

Figura 17 - Obtenção dos coeficientes das variáveis do modelo de regressão logística

```
log_reg_fitted = model.named_steps['voting_clf'].named_estimators_['log_reg']

preprocessor = model.named_steps['preprocessing']

feature_names = preprocessor.get_feature_names_out()

coef_df = pd.DataFrame({
    'feature': feature_names,
    'coeficiente': log_reg_fitted.coef_[0],
    'odds_ratio': np.exp(log_reg_fitted.coef_[0])
})

coef_df = coef_df.sort_values(by='coeficiente', ascending=False).reset_index(drop=True)
coef_df
```

A principal motivação para o uso da regressão logística está relacionada à sua alta interpretabilidade, característica essencial para compreender os fatores que o explicam. Diferentemente de modelos baseados em árvores e em ensembles, que operam como estruturas altamente não lineares e de difícil decomposição analítica, a regressão logística fornece uma relação direta e transparente entre cada variável explicativa e a probabilidade do evento de interesse.

Em termos formais, a regressão logística modela o logaritmo das chances (log-odds) da classe positiva como uma combinação linear das variáveis independentes. Isso permite interpretar cada coeficiente como o impacto marginal de uma variável sobre a chance de ocorrência do evento, mantendo as demais constantes. Ao aplicar a exponenciação dos coeficientes, obtêm-se os odds-ratios, que indica o efeito multiplicativo de cada variável explicativa sobre as chances de ocorrência do evento de interesse. Valores superiores a 1 indicam aumento das chances, enquanto valores inferiores a 1 indicam redução, mantendo-se constantes as demais variáveis do modelo.

Os resultados indicam que características associadas a ocupações de maior qualificação apresentam os maiores efeitos positivos sobre a probabilidade de suficiência de renda. Destacam-se, nesse sentido, os indivíduos inseridos em atividades de ciências e ocupações intelectuais, como médicos, engenheiros, advogados, dentre outros, cujas chances de atingir o rendimento mínimo necessário são aproximadamente 7,9 vezes maiores em comparação à categoria de referência, que é apoio administrativo. Resultados semelhantes são observados para militares e diretores ou gerentes, com odds-ratios superiores a 6, evidenciando o papel central da posição ocupacional e do capital humano na determinação do nível de renda.

Variáveis demográficas e territoriais também exercem influência relevante. O fato de residir em zona urbana e em região metropolitana aumenta as chances de suficiência de renda, refletindo a maior concentração de oportunidades de trabalho, diversificação ocupacional e remunerações relativamente mais elevadas nesses espaços. Da mesma forma, o aumento da idade e do nível de escolaridade — tratado de forma ordinal no modelo — está positivamente associado à variável alvo, indicando que trajetórias mais longas no mercado de trabalho e maior acúmulo educacional tendem a ampliar as chances de atingir o patamar mínimo de renda definido.

Por outro lado, os resultados evidenciam fortes fatores de vulnerabilidade socioeconômica, associados a reduções expressivas nas chances de suficiência de renda. A inserção no setor informal e no emprego privado, bem como a ocupação em atividades elementares, apresenta odds-ratios substancialmente inferiores a 1, indicando menor probabilidade de alcançar o SMN. Em particular, trabalhadores informais possuem chances cerca de 80% menores em relação à categoria de referência, empregador, o que reforça o caráter estrutural da informalidade como mecanismo de reprodução da precariedade econômica no país.

Aspectos regionais e raciais também se mostram relevantes. A residência nas regiões Norte e Nordeste está associada a uma diminuição significativa das chances de suficiência de renda, evidenciando desigualdades territoriais persistentes no mercado de trabalho brasileiro. De forma similar, indivíduos que se autodeclaram pretos ou pardos apresentam menores probabilidades relativas de atingir o nível de renda considerado adequado, o que aponta para a permanência de desigualdades raciais estruturais, mesmo após o controle por ocupação, escolaridade e demais características observáveis.

Além disso, variáveis relacionadas à estrutura domiciliar, como o tamanho do domicílio, apresentam efeitos negativos à medida que aumenta o número de moradores, sugerindo maior pressão sobre a renda disponível per capita. Esse resultado é particularmente relevante para o desenho de políticas de transferência de renda, uma vez que indica que a composição familiar exerce impacto direto sobre a suficiência econômica dos indivíduos.

Em síntese, os resultados do modelo empírico corroboram a hipótese central de que a insuficiência de renda no Brasil está fortemente associada a fatores estruturais, como ocupação, setor de inserção no mercado de trabalho, escolaridade, localização territorial e características sociodemográficas. A identificação desses determinantes, com base em evidências estatísticas robustas, fornece subsídios relevantes para a formulação e avaliação de políticas públicas voltadas à redução das desigualdades sociais, ao fortalecimento do emprego formal, à ampliação do acesso à educação e à focalização mais eficiente de programas de transferência de renda. Dessa forma, esta análise auxilia na compreensão da vulnerabilidade socioeconômica no país, alinhando-se aos objetivos constitucionais de promoção do bem-estar social e de redução das disparidades regionais e sociais.

7. Links

Link para o vídeo: https://youtu.be/belJue2zmJk

Link para o repositório: https://github.com/gabriel-ramos-24/tcc-ciencia-de-dados

REFERÊNCIAS

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil de 1988**. Brasília, DF: Presidente da República, [2016]. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em 09 fev. 2026.

DIEESE – Departamento Intersindical de Estatística e Estudos Socioeconômicos. **Salário mínimo nominal e necessário**. São Paulo, 2023. Disponível em: https://www.dieese.org.br/analisecestabasica/salarioMinimo.html#2023. Acesso em: 9 fev. 2026.

DIEESE – Departamento Intersindical de Estatística e Estudos Socioeconômicos. **Metodologia da Pesquisa Nacional da Cesta Básica de Alimentos: Janeiro de 2016.** São Paulo, fev. 2016. Disponível em: https://www.dieese.org.br/metodologia/metodologiaCestaBasica2016.pdf. Acesso em: 9 fev. 2026.

OECD – Organisation for Economic Co-operation and Development**. Society at a Glance 2024: OECD Social Indicators.** Paris: OECD Publishing, 20 jun. 2024. Disponível em: https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/society-at-a-glance-2024_08001b73/918d8db3-en.pdf. Acesso em: 9 fev. 2026.

IBGE – Instituto Brasileiro de Geografia e Estatística. **Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua): microdados**. Rio de Janeiro, 2023. Disponível em: https://ftp.ibge.gov.br/Trabalho_e_Rendimento/Pesquisa_Nacional_por_Amostra_de_Domicilios_continua/Anual/Microdados/Trimestre/Trimestre_1/Dados/. Acesso em: 9 fev. 2026

WORLD INEQUALITY LAB. **World Inequality Report 2026**. Paris: World Inequality Lab, 2025. Disponível em: https://wir2026.wid.world/www-site/uploads/2026/01/World_Inequality_Report_2026.pdf. Acesso em: 10 fev. 2026.

GÉRON, Aurélien. **Mãos à obra: aprendizado de máquina com Scikit-Learn & TensorFlow**: conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. Tradução de Rafael Contatori. Rio de Janeiro: Alta Books, 2019. 576 p. ISBN 978-85-508-0902-1.

APÊNDICE

Tabela A1 – Coeficientes estimados e odds-ratios

Feature	Coeficiente	Odds Ratio
cat__qual_trabalho_ciencias_intelectuais	2,070236	7,926691
cat__qual_trabalho_militares	1,934938	6,923615
cat__qual_trabalho_diretores_gerentes	1,889207	6,614122
cat__qual_trabalho_tecnicos_medio	0,870980	2,389252
cat__sexo_masculino	0,665714	1,945880
cat__zona_domicilio_urbano	0,662555	1,939742
num__idade	0,579511	1,785164
ord__escolaridade	0,553143	1,738709
cat__regiao_domicilio_metropolitana	0,498768	1,646691
cat__qual_trabalho_rural	0,061113	1,063019
cat__uf_sul	-0,005778	0,994239
cat__qual_trabalho_operadores_maquinas	-0,103456	0,901716
cat__cor/raca_outras	-0,148769	0,861768
cat__tam_domicilio_2-3	-0,156303	0,855300
cat__qnts_anos_trabalha_10anos	-0,161600	0,850781
cat__tam_domicilio_4-5	-0,245267	0,782495
cat__uf_sudeste	-0,266379	0,766148
cat__qual_trabalho_comercio	-0,280338	0,755528
cat__qnts_anos_trabalha_5anos	-0,385929	0,679818
cat__qual_trabalho_construcao	-0,395111	0,673605
cat__qnt_hrs_trabalha_50hrs	-0,468173	0,626145
cat__cor/raca_parda	-0,504764	0,603648
cat__qnt_hrs_trabalha_40hrs	-0,542227	0,581452
cat__tam_domicilio_6+	-0,626620	0,534395
cat__uf_norte	-0,656767	0,518525
cat__cor/raca_preta	-0,684407	0,504389
cat__qnts_anos_trabalha_2anos	-0,752793	0,471049
cat__setor_trabalha_emprego_publico	-0,953055	0,385561
cat__qnt_hrs_trabalha_30hrs	-1,020594	0,360381
cat__uf_nordeste	-1,105935	0,330901
cat__qual_trabalho_ocupacoes_elementares	-1,476865	0,228353
cat__setor_trabalha_informal	-1,569938	0,208058
cat__setor_trabalha_emprego_privado	-1,802255	0,164927