Revista ARACÊ
ISSN: 2358-2472

APLICAÇÃO E ANÁLISE DE DESEMPENHO DE ALGORITMOS DE MACHINE LEARNING NA PREDIÇÃO DE CLASSE SOCIAL

APPLICATION AND PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS IN SOCIAL CLASS PREDICTION

APLICACIÓN Y ANÁLISIS DEL RENDIMIENTO DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO EN LA PREDICCIÓN DE CLASES SOCIALES

https://doi.org/10.56238/arev7n7-061

Data de submissão: 04/06/2025 **Data de publicação:** 04/07/2025

Carlos Eduardo da Silva Didrich
Graduando em Análise e Desenvolvimento de Sistemas. Instituto Federal de Rondônia (IFRO)
E-mail: carlosdidrich@gmail.com

Miguel Bonumá Brunet
Doutorando em Sociologia no PPGSA-UFRJ e na Europa-Universität Flensburg. Professor do Instituto Federal de Rondônia (IFRO)
E-mail: miguel.bonuma@ifro.edu.br

Elias Abreu
Mestre em Ciências da Computação. Professor do Instituto Federal de Rondônia (IFRO)
E-mail: elias.silva@ifro.edu.br

RESUMO

Esta pesquisa visou empregar algoritmos de *machine learning* para predizer a classe social com base em um conjunto de características individuais: renda, escolaridade, raça, sexo, idade, região geográfica e situação ocupacional. O objetivo foi obter a mesma categorização de classes sociais para dois conjuntos de dados distintos: a PNAD-C do IBGE e a pesquisa "Opinião sobre o Coronavírus", do Datafolha, fundamentando-se em uma categorização de referência internacional, para analisar a opinião pública brasileira segundo classe social. Para isso, treinamos e avaliamos seis algoritmos de *machine learning*: MLP Classifier, Random Forest, KNN, Regressão Logística, SVM e GaussianNB, utilizando a base de dados anual da PNAD-C, e posteriormente aplicamos o modelo que obteve melhor desempenho na base do Datafolha, ambos de 2021. A escolha do modelo baseou-se nos resultados de três métricas de validação: acurácia, F1-Score e área abaixo da curva do ROC. O modelo que apresentou melhor desempenho foi o Random Forest. A análise da aplicação deste modelo na base de dados do Datafolha revelou uma correspondência satisfatória com a distribuição original das features da PNAD-C, especialmente nas variáveis de maior peso: escolaridade, renda e situação ocupacional, corroborando com a literatura sobre estratificação social, além de fornecer novos insights sobre o tema.

Palavras-chave: Estratificação social. Classes sociais. *Machine learning*.

ABSTRACT

This research aimed to employ machine learning algorithms to predict social class based on a set of individual characteristics: income, education, race, sex, age, geographic region and occupational status. The objective was to obtain the same categorization of social classes for two distinct data sets:

Revista ARACÊ
ISSN: 2358-2472

the PNAD-C of IBGE and the research "Opinion on the Coronavirus", of Datafolha, based on a categorization of international reference, to analyze Brazilian public opinion according to social class. For this, we trained and evaluated six machine learning algorithms: MLP Classifier, Random Forest, KNN, Logistic Regression, SVM and GaussianNB, using the annual database of PNAD-C, and later applied the model that obtained better performance in the database of Datafolha, both of 2021. The choice of the model was based on the results of three validation metrics: accuracy, F1-Score and area below the ROC curve. The best performing model was Random Forest. The analysis of the application of this model in the Datafolha database revealed a satisfactory correspondence with the original distribution of the Features of the PNAD-C, especially in the variables of higher weight: schooling, income and literature on social stratification, and provide new insights on the subject.

Keywords: Social stratification. Social classes. Machine Learning.

RESUMEN
Esta investigación tuvo como objetivo emplear algoritmos de aprendizaje automático para predecir la clase social con base en un conjunto de características individuales: ingresos, educación, raza, género, edad, región geográfica y situación laboral. El objetivo fue obtener la misma categorización de clases sociales para dos conjuntos de datos distintos: la PNAD-C del IBGE y la encuesta "Opinión sobre el Coronavirus" de Datafolha, con base en una categorización de referencia internacional, para analizar la opinión pública brasileña según la clase social. Para ello, entrenamos y evaluamos seis algoritmos de aprendizaje automático: clasificador MLP, Random Forest, KNN, regresión logística, SVM y GaussianNB, utilizando la base de datos anual PNAD-C, y posteriormente aplicamos el modelo con mejor rendimiento en la base de datos Datafolha, ambos de 2021. La elección del modelo se basó en los resultados de tres métricas de validación: precisión, puntuación F1 y área bajo la curva ROC. El modelo con mejor rendimiento fue Random Forest. El análisis de la aplicación de este modelo a la base de datos Datafolha reveló una correspondencia satisfactoria con la distribución original de las características de la PNAD-C, especialmente en las variables con mayor peso: educación, ingresos y situación laboral, lo que corrobora la literatura sobre estratificación social y aporta nuevas perspectivas sobre el tema.

Palabras clave: Estratificación social. Clases sociales. Aprendizaje automático.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

Neste artigo, apresentamos o resultado de uma pesquisa sobre algoritmos de *machine learning* para predição da categoria classe social em um conjunto de dados da Pesquisa Nacional de Amostra por Domicílio Contínua (PNAD-C), do Instituto Brasileiro de Geografia e Estatística (IBGE), e a aplicação deste algoritmo na pesquisa "Opinião sobre o Coronavírus", do Datafolha, ambos de 2021. O principal objetivo é testar diferentes tipos de algoritmos de *machine learning* na base de dados da PNAD-C visando mensurar qual é mais adequado para essa predição, e após observar como o algoritmo de maior desempenho se comporta no conjunto de dados do Datafolha.

Esse objetivo se insere em um contexto mais amplo de uma pesquisa interdisciplinar envolvendo as áreas de sociologia e informática que intenta analisar as trajetórias ocupacionais da geração jovem brasileira em conexão com suas opiniões políticas e seu nível de confiança nas instituições democráticas. Para analisarmos as opiniões políticas de jovens de diferentes classes sociais, utilizamos uma estratificação específica que nos indica a classe social da cada indivíduo, conforme aponta a literatura sociológica sobre estratificação social (SCALON, 1999; PASTORE *et al*., 2000; SCALON, 2013; RIBEIRO, 2014). Essa estratificação é realizada tendo como base a Classificação Brasileira de Ocupações (CBO) do IBGE (IBGE, 2021), a qual é utilizada na PNAD-C.

Já as opiniões políticas e o nível de confiança em instituições são fornecidas por pesquisas do Datafolha. No entanto, a classificação ocupacional das pesquisas do Datafolha não permite uma comparabilidade adequada com a classificação ocupacional que utilizamos na PNAD-C, como demonstraremos a seguir. Neste contexto, um algoritmo que possa predizer satisfatoriamente a classificação ocupacional de um indivíduo pelos critérios da PNAD-C, tendo como base outras categorias que sejam comuns entre os dois conjuntos de dados, possibilitaria a comparabilidade adequada entre eles, permitindo então verificarmos a opinião política dos jovens brasileiros conforme a classe social. Esse é o desafio que nos propusemos nesta pesquisa.

Para realizar este objetivo, aplicamos seis modelos de *machine learning* para avaliar seus desempenhos na previsão da categoria classe social. Partimos de três critérios de análise de desempenho para medir qual modelo tem a melhor performance e, assim, definir qual algoritmo de *machine learning* utilizar: Accuracy, F1-Score e Curva ROC AUC. Para avaliar e visualizar o comportamento do melhor modelo, adotamos duas métricas: matriz de confusão e shap. Após verificar qual o melhor modelo, o aplicamos na base de dados do Datafolha, comparando seus resultados com a base de dados original da PNAD-C.

Este artigo é dividido em cinco seções, sendo a primeira essa introdução. Na segunda seção, apresentamos as classificações ocupacionais da PNAD-C e da pesquisa Opinião sobre o Coronavírus,

Revista ARACÊ
ISSN: 2358-2472

do Datafolha, além das diferenças que impedem sua comparação. Ainda nessa seção, expomos a base teórica que alicerça as categorias sociológicas utilizadas para a previsão da classe social. Em seguida, na terceira seção, evidenciamos a estrutura dos conjuntos de dados e demonstramos quais métodos de *machine learning* foram utilizados na pesquisa, bem como seus fundamentos teóricos. Também exporemos as três métricas de avaliação desses algoritmos e seus critérios. Após, na quarta seção, apresentamos os resultados obtidos com os algoritmos, comparando-os de acordo com as métricas de avaliação de desempenho. Demonstramos mais detalhadamente os resultados dos dois algoritmos que obtiveram maior desempenho, com resultados significativamente maiores que os demais. Além disso, apresentamos como o algoritmo de maior desempenho se comporta ao realizar a predição no conjunto de dados do Datafolha. Por fim, na última seção, resumimos as principais conclusões da pesquisa e analisamos suas contribuições para possibilitar a comparação entre os conjuntos de dados da PNAD-C e do Datafolha.

2 REFERENCIAL TEÓRICO

2.1 COMPARABILIDADE ENTRE CLASSIFICAÇÕES OCUPACIONAIS DA PNAD-C E DO DATAFOLHA

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD-C) é uma pesquisa do IBGE que tem como propósito principal gerar indicadores para monitorar tendências socioeconômicas de longo prazo, especialmente relacionadas à força de trabalho, por meio de uma amostragem probabilística de domicílios com representatividade geográfica. A pesquisa é realizada em cerca de 211.000 domicílios trimestralmente, permitindo a geração de estimativas trimestrais e anuais (IBGE, 2014). Nesse artigo, utilizamos a estimativa anual da PNAD-C do ano de 2021, como será descrito em detalhe na próxima seção.

A PNAD-C utiliza a Classificação Brasileira de Ocupações (CBO), criada com o objetivo de estabelecer uma hierarquia de ocupações, facilitando a agregação de informações relacionadas à força de trabalho, englobando características ocupacionais ligadas à natureza das funções desempenhadas e ao conteúdo das atividades laborais (IBGE, 2021). A última revisão da CBO ocorreu no final dos anos 1990 e início dos anos 2000, resultando na CBO-2002, a qual introduziu inovações conceituais, com cerca de 10 Grandes Grupos, 47 Subgrupos principais, 192 Subgrupos e 596 Grupos de base ou famílias ocupacionais. A nova versão da CBO foi alinhada com a última versão da International Statistical Classification of Occupations (ISCO-88). Essa revisão ocorreu devido a mudanças significativas no ambiente laboral nas últimas décadas, consequência da globalização, de avanços tecnológicos e de novas formas organizacionais, que reconfiguraram as relações de trabalho.

Revista ARACÊ
ISSN: 2358-2472

A classificação que utilizamos com o conjunto de dados da PNAD-C é baseada na proposta de estratificação social de Scalon (1999), referência internacional na área de estratificação social no Brasil. Seu esquema classificatório, delineado em sua tese de doutorado "Mobilidade social no Brasil" e depois renovado em outras pesquisas nas últimas décadas (SCALON; SALATA, 2012; SCALON, 2013; SCALON *et al*., 2021), parte do referencial teórico de Weber (1999). Este autor clássico da sociologia demonstra como as classes sociais são caracterizadas por situações de classe específicas definidas em torno das oportunidades de venda de um produto ou serviço no mercado. Nestes termos, o *status*, a escolaridade e o poder, por exemplo, são categorias chaves na definição das classes sociais.

Uma grande vertente de estudos sociológicos sobre estratificação social surge com base na teoria weberiana, em especial na Inglaterra (GOLDTHORPE *et al*., 1980; 1992), e se estabelece internacionalmente na sociologia, influenciando estudos nessa área no Brasil. Um dos seus principais fundamentos é justamente focar nas ocupações para entender a estrutura de classes, sem desconsiderar também outros fatores como renda, escolaridade e *status* como relevantes para definir as classes sociais. Tendo a ocupação como fundamento, as análises de classe alcançaram um patamar mais rigoroso de compreensão das dinâmicas de estratificação e mobilidade social, tanto inter, quanto intrageracionais, permitindo também comparações internacionais. Seguindo essa vertente teórica, nesse artigo nos desafiamos a alinhar dois conjuntos de dados pela situação ocupacional, posto que um deles - Datafolha - não permite a construção de um esquema de classes pois não houve coleta de dados precisos sobre a ocupação, apenas sobre algumas características da situação ocupacional, como demonstraremos adiante.

Ao analisar o contexto socioeconômico do Brasil, Scalon (1999) chega em uma proposta de nove estratos sócio-ocupacionais para representar a estrutura de classes brasileira, oriundos da estratificação proposta por Valle Silva (1992), sobre a estratificação da PNAD de 1988, que continha 342 tipos de ocupações. Utilizando técnicas de clusterização, Scalon chega nas seguintes classes sociais: Administradores e Gerentes, Profissionais, Proprietários Empregadores, Proprietários por Conta Própria, Trabalhadores Não-manuais de Rotina, Trabalhadores Manuais Qualificados, Trabalhadores Manuais Não-qualificados, Proprietários Rurais e Trabalhadores Rurais. Para analisar a emergência das classes médias no Brasil, esses estratos podem ainda ser agrupados em três grandes classes: classes médias altas (Administradores e Gerentes, Profissionais e Proprietários Empregadores), classes médias baixas (Proprietários por Conta própria e Trabalhadores Não-manuais de Rotina) e classes de trabalhadores manuais, contendo os demais estratos. Precisamente estas três grandes classes são o objeto de predição com *machine learning* nesta pesquisa.

Revista ARACÊ
ISSN: 2358-2472

Já a pesquisa "Opinião sobre o Coronavírus", do Datafolha, aborda a opinião da população brasileira sobre temas polêmicos durante a pandemia, tais como o hábito de uso de máscaras de proteção facial, a imposição de isolamento social, o fechamento obrigatório do comércio, a avaliação da atuação de governantes no combate ao coronavírus, a confiança nas instituições democráticas, dentre outros temas. A pesquisa conta com uma amostra mais reduzida em relação à PNAD-C, variando entre 5.000 e 2.000 entrevistados. Foi realizada de março de 2020 a janeiro de 2022, tendo sua frequência relacionada às ondas de aumento do contágio de Covid-19, sendo, em média, uma pesquisa a cada dois meses, totalizando 14 pesquisas no período (CESOP-DATAFOLHA, 2021).

Além da opinião, a pesquisa coleta informações sobre o perfil socioeconômico dos entrevistados, tais como sexo, idade, raça, renda e escolaridade, o que permite a comparabilidade de outras variáveis com a PNAD-C. No entanto, a classificação ocupacional proposta pela pesquisa não é adequada para relacionarmos seus resultados com outras pesquisas sobre estratificação social. Ao indagar os entrevistados sobre a ocupação, a pesquisa propõe as seguintes modalidades: Assalariado registrado, Assalariado sem registro, Funcionário público, Autônomo regular, Profissional liberal, Empresário, Free-lance / bico, Estagiário / aprendiz, Dona de casa, Aposentado, Estudante, Vive de rendas e Outros. Como é possível perceber, esta classificação não nos indica com precisão a ocupação dos indivíduos, apenas nos indica algumas características sobre a situação ocupacional.

Frente a esse cenário, mostrou-se necessário encontrarmos variáveis e modalidades em comum entre ambos os conjuntos de dados para que pudéssemos treinar algoritmos de *machine learning* com as mesmas informações para ambos, visando realizar a predição das três grandes classes sociais em estudo. Para explorar variadas possibilidades de análise, selecionamos sete variáveis: renda, situação ocupacional, escolaridade, idade, sexo, cor e região geográfica. Devido ao curto espaço, descreveremos sinteticamente a base teórica que sustenta a seleção dessas categorias para predição de classe social no contexto dessa pesquisa. Adotamos uma gama relativamente ampla de variáveis, em relação às tradicionalmente utilizadas para mensurar classes sociais, com o objetivo de observar o quanto elas também podem ser determinantes para a estratificação social brasileira.

O debate sociológico sobre classes sociais é muito amplo e possui diferentes abordagens teóricas (WRIGHT, 2015; SAVAGE, 2011). Embora a renda tenha sido historicamente a variável mais utilizada para mensurar estratos sociais (BARONE; HERTEL; SMALLENBROEK, 2022), a sociologia da estratificação passou a utilizar cada vez mais a ocupação e a escolaridade para compreender com maior acurácia a dinâmica das classes sociais (SCALON; SALATA, 2012). No caso da escolaridade, essa passa cada vez mais a definir o status de ocupações no mercado de trabalho conforme avança a capacidade de certificação e regulamentação da educação formal nas sociedades

Revista ARACÊ
ISSN: 2358-2472

modernas, levando ao surgimento de complexas estruturas institucionais que definem e restringem a atuação de profissões e ocupações, desde escolas e universidades até conselhos profissionais de classe (SILVA; HASENBALG, 2000).

A ocupação, por sua vez, está relacionada diretamente com a classe social, a categoria que pretendemos predizer nesta pesquisa, não possuindo correspondente no conjunto de dados do Datafolha. No entanto, para qualificar a predição, foi possível criarmos a categoria "situação ocupacional" agrupando em uma mesma variável características fundamentais da estrutura sócio-ocupacional brasileira. Nessa variável diferenciamos o trabalhador formal de informal; distinguimos as relações de trabalho entre empregados e autônomos; e separamos proprietários com empregados de sem empregados. Como é possível perceber, essa variável terá um peso significativo, pois indica algumas das principais características que definem as ocupações.

O recorte etário permite que analisemos como as diferentes gerações experimentam de maneira distinta as consequências das mudanças sociais ao longo do tempo, revelando como as oportunidades e recursos disponíveis são distribuídos de forma desigual entre grupos etários (CHAUVEL, 2014). Além disso, a análise da mobilidade social intergeracional permite rastrear a capacidade de uma geração em ascender ou declinar em relação à geração anterior (ERICKSON; GOLDTHORPE, 1993). No caso do Brasil, inúmeros estudos revelam como as gerações mais jovens vêm apresentando dificuldades em ocupações de maior qualificação, mesmo com o aumento do nível de educação (BRUNET et al., 2022; BRUNET; ANDRADE; CARDOSO, 2022).

Mais recentemente, a teoria da interseccionalidade vem evidenciando a necessidade de analisarmos conjuntamente as experiências de opressão e desigualdade considerando múltiplas dimensões, em especial gênero, raça e classe (COLLINS, 2020). O sexo e o gênero desempenham um papel fundamental na análise da classe social. As mulheres, em particular, enfrentam desafios adicionais na busca pela mobilidade social, devido às estruturas patriarcais que perpetuam a desigualdade de gênero (SCALON, 1999). A interseccionalidade reconhece que as experiências de mulheres de diferentes classes sociais podem variar significativamente, e a análise deve levar em consideração como as estruturas de poder de gênero se entrelaçam com as de classe.

As relações de classe também não são uniformes para todos os grupos raciais, de maneira que a raça pode atuar como um determinante significativo das oportunidades disponíveis, das redes de apoio social e das percepções sociais. No Brasil, grupos negros e pardos têm historicamente enfrentado racismo estrutural, o que afeta suas chances de ascensão social e suas experiências de trabalho (ALMEIDA, 2019). Da mesma forma, as oportunidades econômicas e sociais não são uniformes em todas as áreas geográficas. As diferenças regionais podem afetar significativamente as perspectivas de

Revista ARACÊ
ISSN: 2358-2472

mobilidade social e as experiências de classe, estando também intrinsecamente ligada à raça, o que torna essencial uma análise que leve em consideração todas essas dimensões interligadas.

Cabe ressaltar que não pretendemos adentrar especificamente na análise sociológica de cada uma dessas categorias. Temos apenas a intenção de aproximar a predição da classe social de um universo mais complexo de categorias, o qual as pesquisas sobre desigualdades sociais vêm apontando como relevante. Com essas sete categorias, esperamos fornecer o máximo de subsídio para que a predição seja realizada considerando múltiplos aspectos da realidade brasileira.

2.2 *MACHINE LEARNING*

O aprendizado de máquina é um conjunto de regras utilizadas para ensinar computadores a "aprenderem" de forma automática padrões e comportamentos a partir de dados de treinamento (MOLNAR, 2018). O aprendizado supervisionado é um tipo de aprendizagem que utiliza um conjunto de dados de entrada e seus respectivos valores de saídas. A partir desse conjunto de dados, o aprendizado supervisionado tenta aprender uma função para fazer o mapeamento da entrada para a saída e posteriormente utilizar essa função para predizer valores de saídas de dados de entrada não utilizados no treinamento.

O resultado da predição de um aprendizado supervisionado pode ser uma categoria, ou seja, uma classe dentro de um conjunto limitado de possibilidades. Nesse caso temos uma tarefa denominada de classificação. No entanto, se a predição for um valor numérico específico, então a tarefa é denominada de regressão (BISHOP, 2006).

Diversos algoritmos de aprendizado de máquina supervisionado foram propostos na literatura. Dentre eles, destacam-se a Floresta Aleatória (RF), a Perceptron de Múltiplas Camadas (MLP), a Máquina de Vetores de Suporte (SVM), o K-Vizinhos mais próximos (KNN), a Regressão Logística (RL) e o Gaussian Naive Bayes (GaussianNB).

O RF, do inglês *Random Forest*, é um algoritmo de aprendizado de máquina que utiliza múltiplas árvores de decisão geradas de forma aleatória para fazer a predição dos valores de saída (BREIMAN, 2001). A MLP, do inglês *multilayer perceptron*, é um tipo de rede neural artificial composta por uma camada de entrada, uma ou mais camada oculta e uma camada de saída. As MLPs consistem em redes totalmente conectadas, ou seja, cada neurônio de uma camada conecta-se a todos os neurônios da camada adjacente (HAYKIN, 2001). A SVM, do inglês *Support Vector Machine*, é um algoritmo que tem como principal objetivo encontrar o hiperplano ideal em um espaço N-dimensional que possa separar os pontos de dados em diferentes classes no espaço de características (SMOLA, 2004).

Revista ARACÊ
ISSN: 2358-2472

O KNN, do inglês *K-Nearest Neighbors*, utiliza os valores dos vizinhos mais próximos para fazer a sua predição, esse algoritmo olha para o conjunto de dados de treinamento e seleciona *k* vizinhos com base em uma distância pré-definida (como a distância de manhattan) e utiliza os valores desses vizinhos para fazer a sua predição (PEDREGOSA, 2011). A RL, do inglês *Logistic regression*, é um tipo de modelo estatístico que estima a probabilidade de ocorrência de um evento com base em um determinado conjunto de dados de variáveis independentes, tendo como resultado uma probabilidade, sendo a razão entre a probabilidade de um evento acontecer dividido pela probabilidade do evento não acontecer (LAVALLEY, 2008). O algoritmo GaussianNB é um algoritmo estatístico baseado no teorema de Bayes. Modelos bayesianos utilizam distribuições de probabilidade em vez de estimativas pontuais, para isso é introduzido uma distribuição probabilística *a priori* sobre um evento para obter a probabilidade *a posteriori* desse mesmo evento (BISHOP, 2006). Os algoritmos RF, MLP, SVM e KNN podem ser utilizados tanto para a tarefa de classificação quanto para a tarefa de regressão, já algoritmos RL são utilizados apenas para a tarefa de classificação (PEDREGOSA, 2011).

3 METODOLOGIA

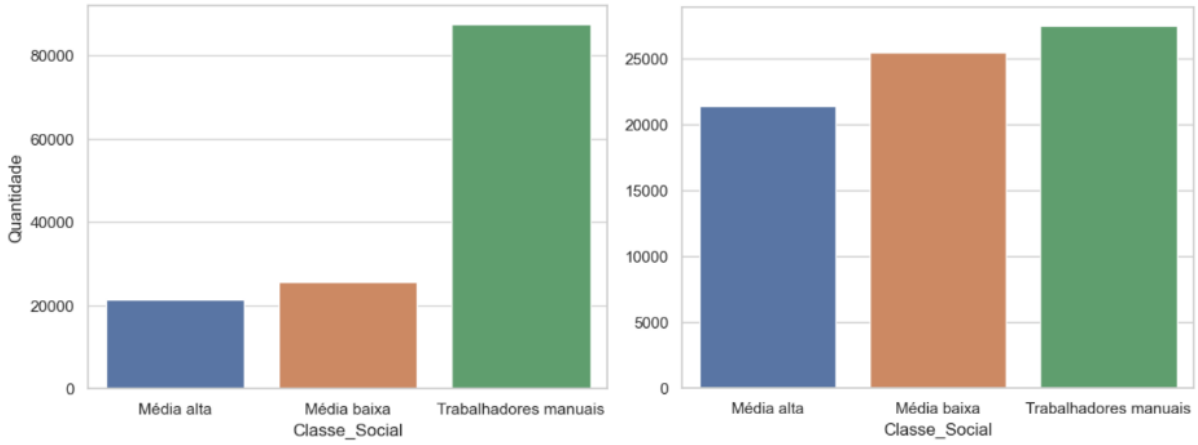
3.1 DESCRIÇÃO DOS CONJUNTOS DE DADOS

Para o treinamento e avaliação dos modelos, foi utilizado o conjunto de dados anual da PNAD-C de 2021. A base é composta por 134.414 indivíduos, pois só incluímos os indivíduos que compõem a População Economicamente Ativa (PEA). Desses, 84 não possuem renda informada, de forma que optamos por removê-los, reduzindo a quantidade de indivíduos para 134.330. A classe social se divide em três grupos sendo eles trabalhadores manuais com 87.452 indivíduos, média alta com 21.394 indivíduos e média baixa com 25.484 indivíduos.

A Figura 1a apresenta a quantidade de dados pertencentes a cada uma das três classes que desejamos predizer. Como é possível observar, tal base encontra-se desbalanceada, ou seja, a proporção de dados de uma determinada classe é muito diferente das demais. Optamos, então, por aplicar a técnica de subamostragem aleatória: as instâncias da classe majoritária são descartadas aleatoriamente até que uma distribuição mais equilibrada seja alcançada (HAIBO HE, YUNQIAN MA, 2013, p.45). Utilizando esta técnica, foram removidos aleatoriamente 60.000 indivíduos da classe trabalhadores manuais, restando 27.503. Com isso, obtivemos um nivelamento entre as classes, como apresentado na Figura 1b. Como as classes média alta e média baixa possuem pouco mais de 20 mil, a base final ficou com um total de 74.381 indivíduos.

Revista ARACÊ
ISSN: 2358-2472

Figura 1 - Classes sociais da PNAD-C segundo a classificação proposta



Fonte: Autores, 2023.

Para a aplicação do modelo selecionado, foi utilizado o conjunto de dados do Datafolha. O conjunto é composto por 2.018 indivíduos, sendo tratada e selecionada as *features*: idade, cor, sexo, região, ocupação, renda, escolaridade para as predições do modelo. Durante o tratamento dos conjuntos de dados, desenvolvemos um dicionário, disponível na Tabela 1, onde podemos observar a equivalência da situação ocupacional em ambas as bases no que tange as features selecionadas. As demais features não possuem distinção entre as bases.

Tabela 1 - Dicionário de base de dados PNAD-C X Datafolha

PNAD-C IBGE	**Datafolha**
Empregado no setor privado sem carteira de trabalho assinada	Assalariado sem registro
Trabalhador doméstico sem carteira de trabalho assinada	
Empregado no setor privado com carteira de trabalho assinada	Assalariado registrado
Trabalhador doméstico com carteira de trabalho assinada	
Empregado no setor público	Funcionário público
Militar e servidor estatutário	
Conta-própria Profissional	Profissional liberal
Conta-própria Contribuinte	Autônomo regular
Conta-própria Não contribuinte	Free-lance/ bico
Empregador	Empresário
Demais modalidades (Não ocupados)	Estagiário/ aprendiz
	Outros
	Desempregado (Procura emprego)
	Dona de casa

Revista ARACÊ
ISSN: 2358-2472

	Aposentado
	Estudante
	Desempregado (Não procura emprego)
	Vive de rendas

Fonte: Autores, 2023

3.2 ALGORITMOS DE *MACHINE LEARNING*

A fim de prever com eficiência a classe social de uma pessoa através de um conjunto de características, foram treinados e analisados os seguintes modelos de *machine learning*: Florestas Aleatórias (RF), Máquina de Vetor de Suporte (SVM), K-vizinhos mais próximos (KNN), Perceptron Multicamadas (MLP), Regressão Logística (RL) e Gaussian Naive Bayes (GaussianNB). Esses modelos foram selecionados devido a vasta utilização na literatura e a disponibilização em bibliotecas de aprendizado de máquina.

Para a implementação e testes desses modelos foi utilizado a biblioteca Scikit-learn que fornece implementações de última geração de muitos algoritmos de aprendizado de máquina bem conhecidos, mantendo uma interface fácil de usar e totalmente integrada com a linguagem Python (PEDREGOSA, 2011).

3.3 MÉTRICAS DE AVALIAÇÃO

A métrica utilizada para avaliação dos modelos de aprendizado de máquina foi a validação cruzada de 5 partições. Nessa métrica o conjunto de dados é dividido aleatoriamente em N partes, no caso usamos N igual a 5. Quatro delas são usadas no treinamento e a outra para teste, isso permite uma melhor avaliação do modelo e ajuda a evitar o *overfitting*. Para avaliar o desempenho dos modelos, foi calculada a média de três métricas: acurácia (a), F1-score (b) e AUC (c). A utilização de três métricas de avaliação deve-se ao fato de que a base de dados é desbalanceada, de maneira que apenas a utilização da acurácia, a métrica mais comum, geraria distorções.

A fim de obter modelos de classificação de alta fidelidade capazes de prever com eficiência a classe social das pessoas, o algoritmo *Grid Search* foi utilizado para realizar a seleção de parâmetros para os métodos de aprendizado de máquina. A partir destes experimentos, chegou-se aos seguintes parâmetros de configuração:

- **Random Forest:** o número de árvores na floresta foi configurado como igual a 300, a profundidade máxima da árvore foi definido como 10, o número mínimo de amostras necessárias para dividir um nó interno foi definido como 2.

Revista ARACÊ
ISSN: 2358-2472

- **Perceptron de Múltiplas Camadas (MLP):** a função de ativação da camada oculta foi definido como relu, o número de épocas foi fixado em 200, a rede possui uma camada oculta com 100 neurônios.
- **Máquina de Vetores de Suporte (SVM):** o kernel definido como Radial Basis Function (RBF), o parâmetro de penalidade C igual a 1.
- **K-Vizinhos mais Próximos (KNN):** o número de vizinhos foi definido como igual a 5;
- **Regressão Logística (RL):** a norma da penalidade foi definida como L2. Para solucionar problemas de otimização, foi selecionado o algoritmo lbfds.
- **Gaussian Naive Bayes (GaussianNB):** não foi configurado nenhum hiperparâmetro.

3.3.1 Accuracy

A métrica "Accuracy" (Acurácia) é uma medida de desempenho comum para modelos de classificação que avalia a proporção de instâncias corretamente classificadas em relação ao total de instâncias no conjunto de dados de teste. É especialmente útil quando as classes estão balanceadas no conjunto de dados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Onde: TP = *True Positive*, TN = *True Negative*, FP = *False Positive*, FN = *False Negative*.

3.3.2 F1-Score Macro

O F1-Score Macro é especialmente útil quando há desbalanceamento entre as classes, pois trata todas as classes igualmente e evita que o desempenho do modelo seja dominado pela classe majoritária. Ele fornece uma avaliação global do desempenho do modelo em todas as classes, independentemente do tamanho das classes individuais.

$$F1Macro = \frac{1}{k}\sum_{i=1}^{k} F1\,(C_1)$$

Onde: k é o número total de classes no conjunto de dados.

Revista ARACÊ
ISSN: 2358-2472

3.3.3 Roc Auc ovo

A fórmula matemática para calcular o 'roc_auc_ovo', que representa a área sob a curva ROC (Receiver Operating Characteristic) em problemas de classificação multiclasse usando a abordagem "One-Versus-One" (OvO), é a média aritmética das áreas sob as curvas ROC dos classificadores binários.

$$roc\_auc\_ovo = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} AUC(Ci, Cj)$$

Onde: k é o número total de classes e AUC(Ci, Cj) representa a área sob a curva ROC para o classificador binário treinado usando apenas as instâncias das classes *Ci* e *Cj*.

3.3.4 Confusion matrix

A matriz de confusão proporciona uma visão detalhada do desempenho do modelo ao comparar suas previsões com os valores reais dos dados. Como ilustrado na figura 2, a matriz identifica o valor real comparado ao valor predito pelo modelo.

Figura 2 - Lógica da confusion matrix

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: Diego Nogare, 2020.

3.3.5 Shap (SHapley Additive exPlanations)

SHAP é uma abordagem teórica de jogos para explicar a saída de qualquer modelo de aprendizado de máquina (SCOTT; SU-IN LEE, 2017). O gráfico SHAP com múltiplas classes permite entender quais características são mais influentes em cada classe individualmente e como elas afetam as probabilidades de decisão do modelo. Isso ajuda a obter insights mais específicos sobre as decisões do modelo para cada classe, conforme pode ser visto na Figura 4, na próxima seção.

Revista ARACÊ
ISSN: 2358-2472

4 EXPERIMENTOS E RESULTADOS

4.1 DESEMPENHO DOS ALGORITMOS

Iniciamos a apresentação dos resultados com a análise de desempenho dos seis algoritmos que foram treinados e aplicados na base de dados da PNAD-C, comparando as três métricas de avaliação descritas anteriormente. Em seguida, analisamos detalhadamente as variáveis e modalidades que exerceram maior peso nos modelos dos algoritmos que obtiveram melhores resultados. A Tabela 2 apresenta os resultados obtidos pelos modelos de aprendizado de máquina analisados nesta pesquisa de acordo com as três métricas. A acurácia mensura a proporção de predições corretas feitas pelo modelo em relação ao total de predições feitas. O F1-Score mede o equilíbrio entre precisão de acertos e *recall*. Por fim, a área abaixo da curva do ROC foi utilizada principalmente devido à distinção de tamanho de dados das três variáveis em análise, conforme apresentado anteriormente na Figura 1. Como a distinção entre classes é crucial, esta métrica auxilia ao indicar o quão bem o modelo separa as classes.

Tabela 2 - Desempenho dos modelos segundo as três métricas

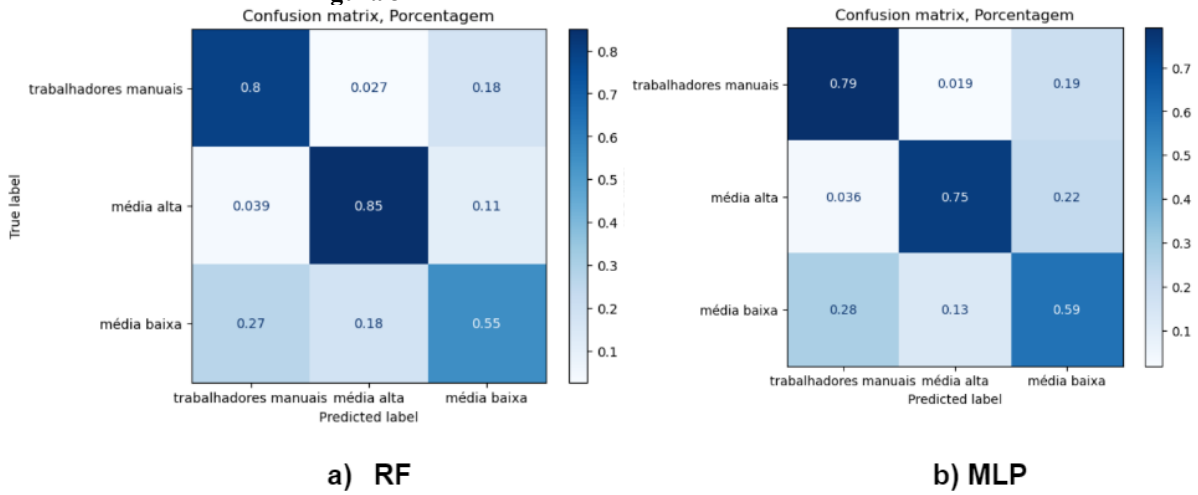
Modelo	Accuracy	F1-Score Macro	ROC AUC ovo
RandomForestClassifier	0.72	0.72	0.88
MLPClassifier	0.71	0.71	0.87
KNN	0.66	0.67	0.83
Regressão Logística	0.63	0.63	0.79
SVM	0.63	0.63	*NA*
GaussianNB	0.62	0.61	0.79

Fonte: Autores, 2023

Como é possível observar, os resultados do RF e da MLP se destacaram sobre os demais e obtiveram resultados bem próximos. Entretanto, o RF obteve melhores resultados nas três métricas. Tais resultados podem ser melhor analisados na matriz de confusão dos modelos, como podemos observar nas Figuras 3a (RF) e 3b (MLP). O RF novamente se destacou com 80% e 85% de assertividade na predição para as classes trabalhadores manuais e média alta, respectivamente. Porém, observamos que o modelo teve um percentual maior de erros na predição da classe média baixa, predizendo-a como trabalhadores manuais em 27% dos casos. Mais adiante avaliamos quais as principais variáveis relacionadas a essa confusão, o que, a partir do pensamento sociológico, nos permite mapear as características que estão presentes na margem entre a classe de trabalhadores manuais e a classe média baixa, considerando a configuração proposta de estratificação social.

Revista ARACÊ
ISSN: 2358-2472

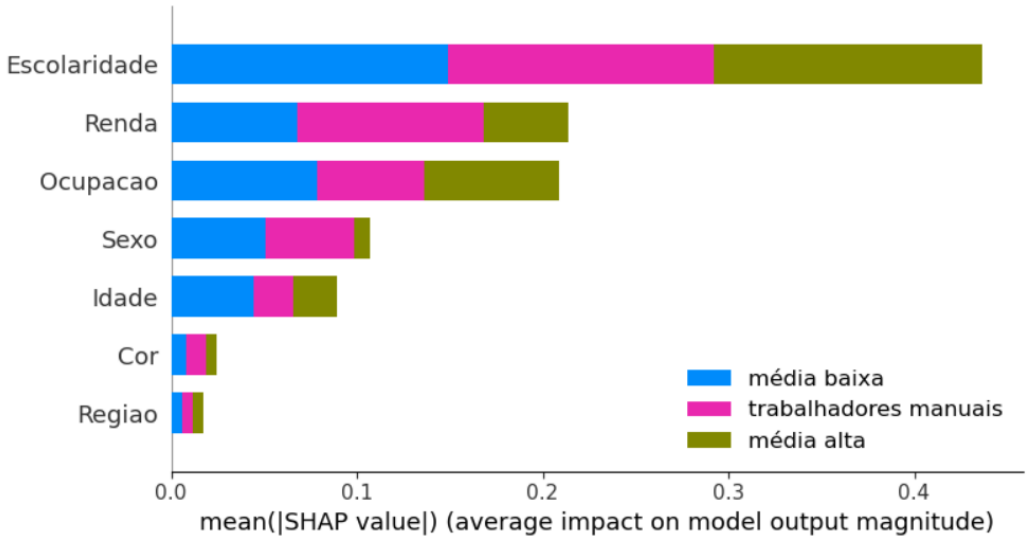
Figura 3 - Matriz de confusão dos modelos RF e MLP



Fonte: Autores, 2023

Através da função *shap* podemos analisar a influência das features na tomada de decisão do modelo utilizado. Na Figura 4, apresentamos essa distribuição para o RF, que foi o melhor modelo, conforme demonstrado na Tabela 2. Nota-se que o RF identificou que as features Escolaridade, Renda e Situação Ocupacional exerceram maior peso na predição das variáveis do modelo, destacando-se a primeira dessas variáveis. Entretanto, evidenciam-se diferenças entre as variáveis de maior peso para cada classe social. Para a predição das classes média baixa e média alta, as features Escolaridade e Ocupação revelaram ter maior peso, ao passo que, para trabalhadores manuais, Escolaridade e Renda exerceram um papel maior.

Figura 4 - Influência das features no modelo RF



Fonte: Autores, 2023.

Revista ARACÊ
ISSN: 2358-2472

Essa disparidade nos revela como as classes médias são mais marcadas pelo nível de escolaridade e pela situação ocupacional quando comparadas à classe de trabalhadores manuais, que demonstra maior relação com a faixa de renda. Tal evidência está de acordo com a literatura sobre a emergência das classes médias nas sociedades ocidentais, a qual aponta que uma nova classe média surge em meados do século 20 marcada pela regulamentação e certificação das profissões e dos diplomas que demarcam os níveis de escolaridade (BOURDIEU, 2011a; CHAUVEL, 2020; 2021; SCALON; SALATA, 2012). Na próxima subseção, detalharemos as modalidades de maior peso dentre estas três variáveis, para analisar em maior detalhe quais delas exercem maior influência sobre a decisão do modelo.

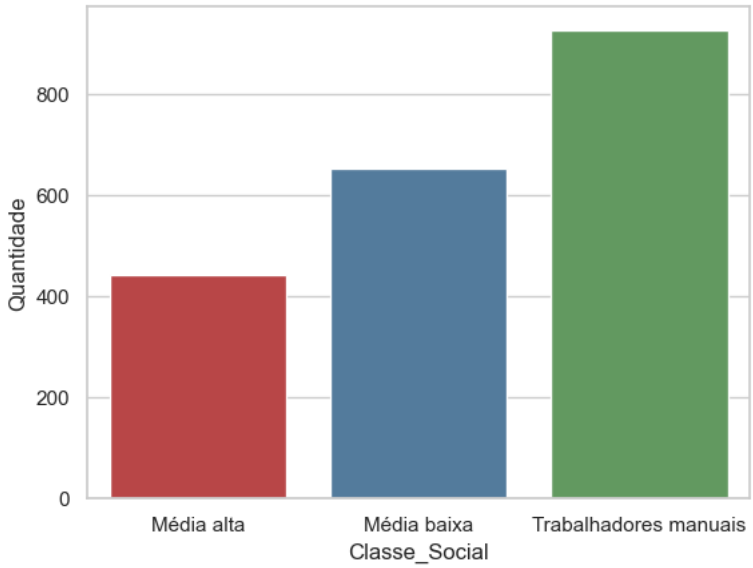
4.2 APLICAÇÃO DOS MODELOS NA BASE DE DADOS DO DATAFOLHA

Ao aplicarmos o modelo de RF treinado com a base IBGE PNAD-C na base de dados do DataFolha para a predição da classe social, pudemos finalmente conectar os conjuntos de dados por meio da categoria chave de nossa pesquisa, a saber, classe social. Foi possível obter um conjunto de dados do DataFolha com a opinião sobre a pandemia de coronavírus, opinião política, expectativa de futuro, dentre outras questões relevantes, e, ao mesmo tempo, as classes sociais devidamente classificadas segundo um esquema que permite comparabilidade internacional com estudos aprofundados sobre o tema.

A escolha do modelo deveu-se a sua melhor performance no conjunto de dados de treinamento e validação, como apresentado na Tabela 1. A Figura 5 apresenta a quantidade de cada classe social predita por esse modelo na base de dados do Datafolha. Como é possível observar, a predição obteve um balanceamento de dados semelhante à base de dados da PNAD-C, na qual foi realizado o treino, o que indica que a aplicação do algoritmo no conjunto de dados do Datafolha obteve correspondência com a distribuição original das features.

Revista ARACÊ
ISSN: 2358-2472

Figura 5 - Distribuição das classes social na predição do modelo RF aplicada na base de dados do Datafolha



Fonte: Autores, 2023.

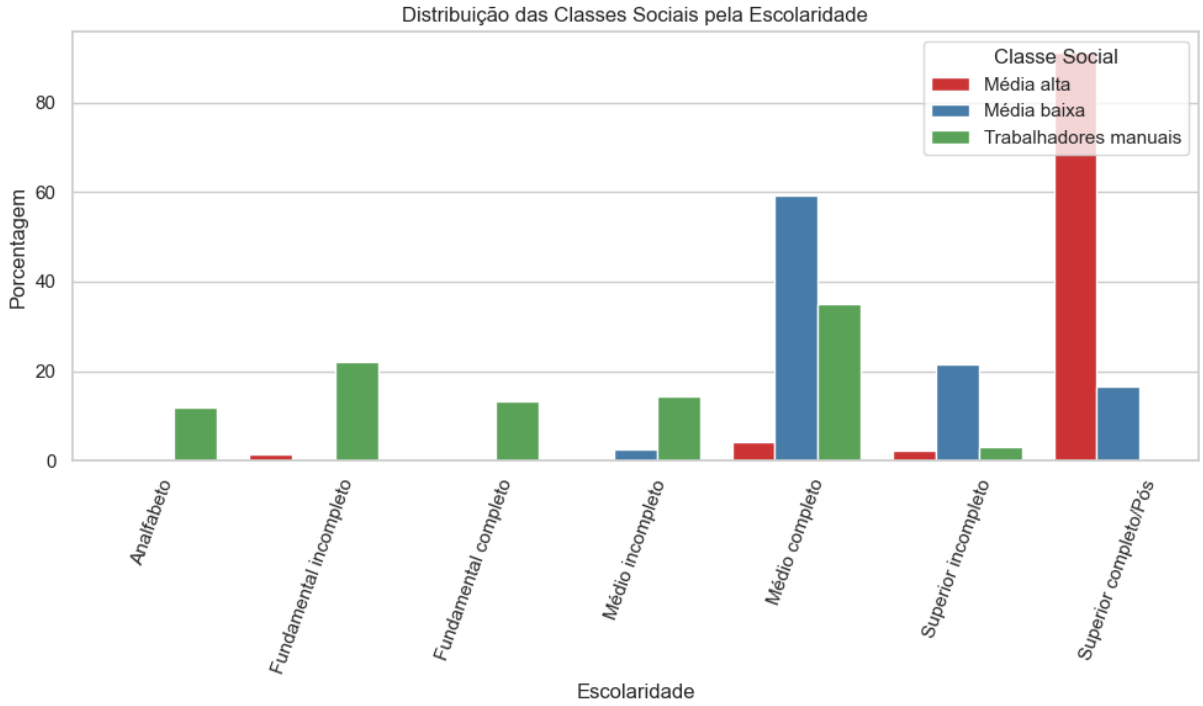
A distribuição das classes sociais preditas segundo as três variáveis de maior peso (Escolaridade, Renda e Situação Ocupacional), conforme apresentado na Figura 4, é evidenciada nas Figuras 6, 7 e 8, a seguir. A partir dessa distribuição, podemos avaliar o quanto a predição corresponde ao padrão esperado das classes sociais em comparação à base de dados original da PNAD-C, bem como segundo outras pesquisas sociológicas. Da mesma forma, podemos obter *insights* sobre as características das classes sociais por meio de análise sociológica. Cabe destacar, em especial, as características que demarcam os limites entre a classe de trabalhadores manuais e a classe média baixa, conforme apontado anteriormente.

A variável de maior peso na predição das três classes sociais é a escolaridade, como demonstrado na Figura 4, o que evidencia a crescente importância do nível de escolaridade para a divisão ocupacional da sociedade contemporânea. Este fenômeno já vinha sendo observado em estudos anteriores desde o processo acelerado de urbanização da sociedade brasileira, que contou com a complexificação da divisão social do trabalho, em especial no setor de serviços, demandando o aumento da especialização do conhecimento (SILVA; HASENBALG, 2000). Diversos estudos demonstram como a educação pode ser considerada como demarcador de classe nas sociedades modernas, desde autores clássicos na sociologia como Weber (1999) e Bourdieu (2011a) até pesquisas mais recentes voltadas à realidade brasileira (SCALON, 1999; SCALON; SALATA, 2012; VASCONCELLOS, 2016). No caso desta pesquisa, a escolaridade pode ter apresentado um peso maior do que a renda, que normalmente é a variável mais significativa em estudos de análises de classe, devido à classificação utilizada de três classes, na medida em que a fronteira entre trabalhadores manuais, classes médias e alta é mais demarcada pela escolaridade, o que poderia ser diferente se fosse

Revista ARACÊ
ISSN: 2358-2472

utilizado um esquema com maior número de classes. Podemos observar na Figura 6, a seguir, a distribuição das classes sociais segundo nível de escolaridade na predição realizada.

Figura 6 - Distribuição das Classes Sociais segundo Escolaridade



Fonte: Autores, 2023.

Claramente há uma forte correlação entre nível de escolaridade e classe social, bem mais demarcada do que as demais variáveis. A classe de trabalhadores manuais representa quase a totalidade da população sem ensino médio completo, com poucas exceções estatisticamente insignificantes. Já a classe média baixa possui quase 60% do total com ensino médio completo, ou mais de 80% com ensino médio, mas sem ensino superior, se acrescentarmos a população com ensino superior incompleto. Por fim, o ensino superior completo evidencia uma nítida fronteira entre a classe média alta e as demais classes, na medida em que esta classe apresenta praticamente sua totalidade demarcada por este nível de educação. Isso demonstra que o nível superior segue sendo um forte marcador de classe no Brasil, fato que está em consonância com inúmeras pesquisas sobre esse tema no Brasil (COSTA; KOLINSKI; COSTA, 2013; RIBEIRO, 2014; SALATA, 2018).

No entanto, ao mesmo tempo observamos um fenômeno que vem sendo objeto intenso de debate acadêmico: o aumento do acesso ao nível superior no Brasil e sua relação com a redução das desigualdades sociais. Como é possível observar na Figura 6, percebe-se que mais de um terço da classe média baixa se caracteriza por haver acessado o ensino superior, o que evidencia como parte dessa classe social tem contato com as instituições deste nível de ensino. Por um lado, tal fenômeno

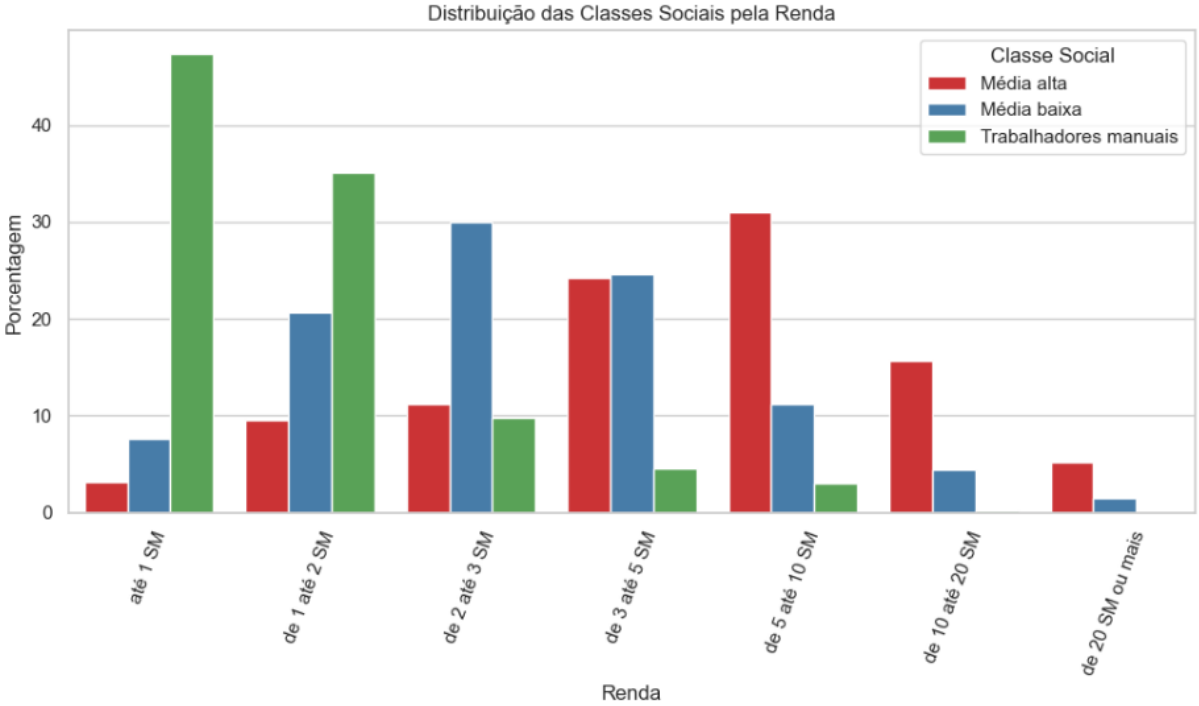
Revista ARACÊ
ISSN: 2358-2472

pode ser interpretado como um aumento positivo do acesso ao ensino superior, que se apresenta mais permeável àquela classe, a qual possuiria melhores oportunidades de ascender socialmente por meio do aumento do nível de escolaridade. Por outro lado, podemos conjecturar que o mercado de trabalho brasileiro pode não estar acompanhando o aumento do acesso ao nível superior, não havendo vagas suficientes para a quantidade de pessoas que vêm alcançando este nível de ensino, como apontado por outros estudos (BRUNET *et al*., 2021; BRUNET; ANDRADE; CARDOSO, 2022).

Conforme mencionado anteriormente, a variável renda domiciliar apresentou-se como a segunda feature com maior influência sobre a predição realizada, tendo maior peso em especial sobre a classificação de trabalhadores manuais. A distribuição das classes segundo a renda é apresentada na Figura 7, a seguir. Ao analisarmos as modalidades dessa variável, podemos perceber que a renda domiciliar de até 1 salário mínimo (SM) representa quase metade dos trabalhadores manuais, demarcando fortemente essa classe social, que decresce expressivamente conforme aumenta a faixa de renda. Já a classe média baixa apresenta um crescimento quantitativo até a faixa de renda entre 2 e 3 SM, da qual é a modalidade mais representativa, sendo ainda a maioria na faixa entre 3 e 5 SM. No entanto, decresce significativamente a partir daquela faixa de renda, dando lugar à classe média alta, que apresenta maior quantidade de indivíduos na faixa de renda entre 5 e 10 SM, sendo ainda maioria nas maiores faixas de renda, de mais de 10 SM.

Revista ARACÊ
ISSN: 2358-2472

Figura 7 - Distribuição das Classes sociais pela Renda



Fonte: Autores, 2023.

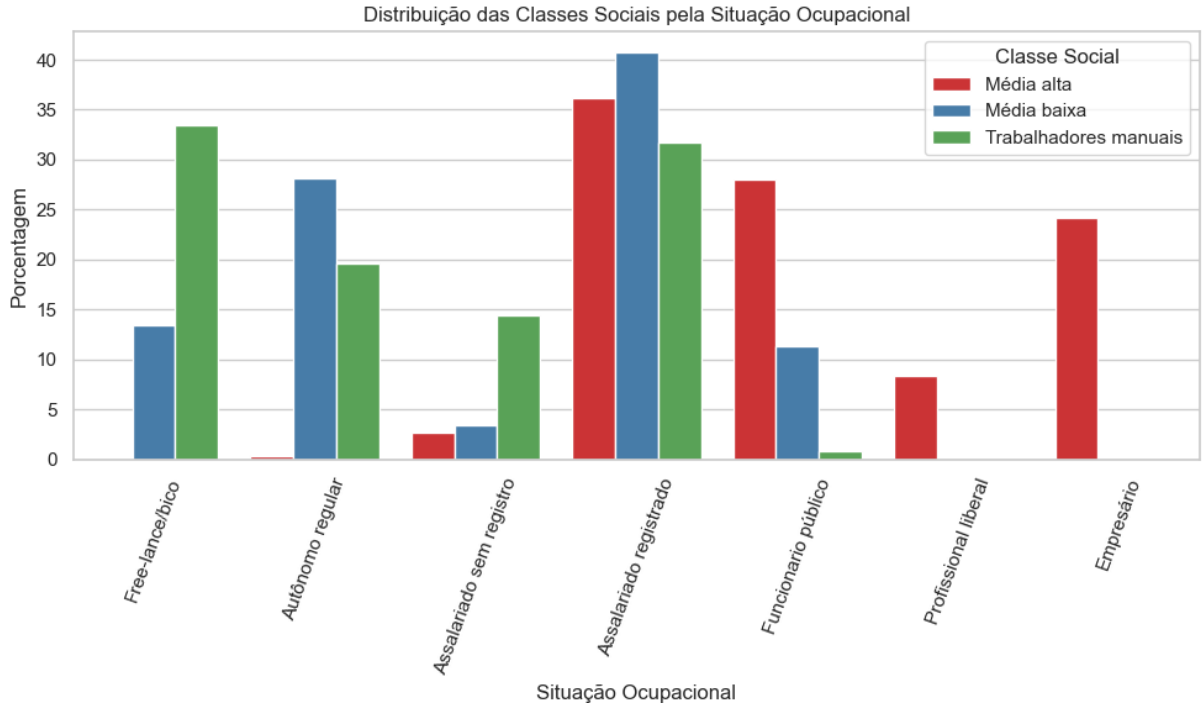
O padrão observado demonstra que a predição foi bem-sucedida na representação das classes sociais segundo faixas de renda, condizente com a base de dados da PNAD-C. Da mesma forma, revela as faixas de renda limítrofes entre as classes sociais em tela. Enquanto a classe de trabalhadores manuais é marcada fortemente pela renda domiciliar de até 1 SM, a classe média baixa caracteriza-se majoritariamente pela renda domiciliar de 2 a 5 salários mínimos, ao passo que a classe média alta tem predominância nas faixas de renda superiores a 5 salários mínimos. Diferentemente da escolaridade, tais demarcações não são estanques, na medida em que há proporções das três classes sociais em quase todas as faixas de renda (com exceção das duas maiores, nas quais não há trabalhadores manuais), mas permitem visualizar claramente a diferenciação contínua entre as três classes segundo a renda domiciliar.

Finalmente, a distribuição da Situação Ocupacional, terceira variável de maior peso na predição realizada, com segundo maior peso nas classes média alta e média baixa, é apresentada na Figura 8, a seguir. Cabe destacar que consta no gráfico apenas a população ocupada, já que estamos analisando a estrutura sócio-ocupacional. Além disso, é preciso considerar que a pesquisa do Datafolha não possui uma distribuição entre trabalhadores formais e informais em 2021 equivalente à PNAD-C do IBGE: os assalariados registrados estão superestimados nessa pesquisa, ao passo que os trabalhadores sem registro, autônomos e trabalhos temporários (bicos) estão em menor número. Segundo a PNAD-C do IBGE, mais da metade da população brasileira ocupada estava em situação de informalidade no Brasil

Revista ARACÊ
ISSN: 2358-2472

em 2021. Mesmo assim, tal descompasso não invalida a predição realizada, na medida em que o treino foi realizado no próprio conjunto de dados da PNAD-C, permitindo uma leitura adequada das classes sociais dos indivíduos no conjunto de dados do Datafolha.

Figura 8 - Distribuição das Classes sociais pela Situação Ocupacional



Fonte: Autores, 2023.

Como é possível observar, a classe de trabalhadores manuais afigura-se com mais de dois terços do total em situações ocupacionais marcadas pela informalidade, sendo um terço em trabalhos irregulares (bicos), caracterizados por uma situação ocupacional mais vulnerável. Já a classe média baixa, mesmo apresentando quase metade do total em situações ocupacionais marcadas pela informalidade, é composta em sua maioria por trabalhadores formais, o que revela uma heterogeneidade nesta classe social, a qual está presente também entre funcionários públicos. Por fim, a classe média alta está quase ausente de situações ocupacionais marcadas pela informalidade, além de compor a totalidade de profissionais liberais e empresários, caracterizadas por maior capital econômico e cultural, bem como a maioria dos funcionários públicos. Este padrão de distribuição das classes sociais segundo situação ocupacional condiz claramente com o esperado para cada classe conforme outras pesquisas, além de abrir caminhos para refletir sobre a realidade apresentada pelos dados, o que poderá ser realizado em outras pesquisas.

Em suma, ao analisarmos as três features de maior peso na predição realizada, podemos afirmar que o padrão de distribuição segundo escolaridade, renda e situação ocupacional está de acordo com o

Revista ARACÊ
ISSN: 2358-2472

conjunto de dados da PNAD-C, bem como com resultados de outras pesquisas que utilizam a mesma classificação (SCALON, 1999; SCALON; SALATA, 2012).

5 CONSIDERAÇÕES FINAIS

Os resultados aqui apresentados são uma parte de uma pesquisa mais ampla, a qual poderá, a partir dos resultados obtidos, efetuar uma análise da opinião pública brasileira segundo a classe social. De forma geral, o objetivo foi atingido com sucesso, na medida em que foi possível realizar uma boa predição da classe social com base nas características selecionadas, mesmo em um conjunto de dados distinto do original. Quanto à análise de desempenho dos algoritmos de *machine learning*, o algoritmo Random Forest (RF) obteve o melhor desempenho, seguido do algoritmo Multilayer Perceptron (MLP), o que pode ser observado nas três métricas utilizadas. Foi possível observar como o desempenho dos algoritmos está relacionado diretamente com a correlação entre as variáveis. Dependendo da maneira como estas fossem agregadas ou desagregadas, o desempenho poderia variar. Além disso, pudemos perceber como o balanceamento dos dados permitiu uma predição com maior qualidade, evitando o *overfitting*, que gera distorções na aplicação dos modelos em outros conjuntos de dados.

Os resultados também foram muito relevantes para a sociologia, pois evidenciaram a possibilidade do uso de técnicas de *machine learning* para a análise de dados socioeconômicos em múltiplos conjuntos de dados. A predição da classe social no conjunto de dados da PNAD-C e sua posterior aplicação com sucesso na base de dados do Datafolha permite a análise da opinião pública brasileira segundo um esquema de classes que possibilita a comparação com estudos da área de estratificação, tanto nacionais quanto internacionais. Tal análise poderá ser feita no escopo de outra pesquisa, devido ao curto espaço aqui disponível. Além disso, ficou claro que, ao compreendermos a lógica com a qual os algoritmos executam a predição, podemos utilizar a imaginação sociológica para analisar os resultados obtidos, propiciando diversos *insights* sobre a estratificação social brasileira. Em outras palavras, os algoritmos não apenas permitiram objetivamente conectar os conjuntos de dados por meio da predição, mas também revelaram nuances da realidade social ao cruzar diferentes variáveis e apontar suas correlações. No entanto, é sempre importante ressaltar que a análise de classe social é uma tarefa complexa, que envolve uma série de elementos também qualitativos. Os algoritmos de *machine learning* podem ser utilizados para auxiliar nesta tarefa, mas não devem ser utilizados como substitutos da análise sociológica.

Finalmente, esta pesquisa abriu novas possibilidades de investigação. A comparação entre os seis algoritmos de *machine learning* e sua avaliação por meio de três métricas distintas poderão ser

Revista ARACÊ
ISSN: 2358-2472

desenvolvidas relacionando os resultados aqui obtidos com outros executados em diferentes conjuntos de dados, atentando-se aos níveis de agregação e às correlações contextuais dos dados utilizados, o que poderá nos fornecer maiores detalhes sobre a aplicação de tais algoritmos em conjuntos de dados de pesquisas socioeconômicas. Além disso, a partir da aplicação do modelo de melhor avaliação no conjunto de dados do Datafolha, será possível prosseguir com a pesquisa que relaciona a opinião pública dos jovens brasileiros com sua classe social, inclusive comparando diferentes períodos e outras características.

AGRADECIMENTOS

Agradecemos à Pró-Reitoria de Pesquisa, Inovação e Pós-graduação (PROPESP) do Instituto Federal de Educação, Ciência e Tecnologia de Rondônia (IFRO), a qual, por meio do Edital nº 2/2025/REIT - PROPESP/IFRO, de 12 de maio de 2025, de apoio à comunicação científica e literária em áreas consideradas estratégicas à Pesquisa e à Pós-graduação do IFRO, subsidiou esta publicação científica e promoveu, por meio do Edital n° 5/2021/REIT - PROPESP/IFRO - Seleção de Novos Projetos de Iniciação Científica (Ciclo 2021-2022) e do Edital n° 10/2022/REIT - PROPESP/IFRO - Renovação de Projetos de Iniciação Científica (Ciclo 2022-2023), o Projeto de Pesquisa desenvolvido no IFRO, Campus Ji-Paraná, por docentes e discentes do Curso Técnico em Informática e do Curso de Graduação em Análise e Desenvolvimento de Sistemas, do qual este artigo é resultado.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

ALMEIDA, Silvio. Racismo estrutural. São Paulo: Pólen, 2019. Disponível em: https://www.scielo.br/j/bak/a/8R37NgQt56Sf5P58KRfMFzq/?format=pdf&lang=pt. Acesso em: 4 jul. 2025.

BARONE, Carlo; HERTEL, Florian R.; SMALLENBROEK, Oscar. The rise of income and the demise of class and social status? A systematic review of measures of socio-economic position in stratification research. Research in Social Stratification and Mobility, v. 78, p. 100678, 2022. Disponível em: https://doi.org/10.1016/j.rssm.2022.100678. Acesso em: 4 jul. 2025.

BISHOP, Christopher Michael. Pattern recognition and machine learning. Nova York: Springer Science Business Media, 2006.

BOURDIEU, Pierre. A distinção: crítica social do julgamento. 2. ed. Porto Alegre: Zouk, 2011a.

BOURDIEU, Pierre. O poder simbólico. 15. ed. Rio de Janeiro: Bertrand Brasil, 2011b.

BREIMAN, Leo. Random forests. Machine Learning, v. 45, p. 5-32, 2001.

BRUNET, Miguel Bonumá; ANDRADE, Leonardo Mota; CARDOSO, Nelson Andrade. Fratura geracional no Brasil no início do século XXI? Análise das oportunidades de vida da geração jovem no Brasil entre 2012 e 2019. Civitas: Journal of Social Sciences, Porto Alegre, v. 22, p. e41669, 2022. Disponível em: https://doi.org/10.15448/1984-7289.2022.1.41669. Acesso em: 4 jul. 2025.

BRUNET, Miguel et al. Análise da mobilidade escolar dos jovens segundo grupo de renda e escolaridade dos responsáveis no Brasil (2012-2020). Revista Edutec, Ariquemes, v. 3, n. 1, p. 77-86, jan./jun. 2022.

CESOP. Cesop/Datafolha. 2021. Disponível em: https://www.cesop.unicamp.br/por/banco_de_dados. Acesso em: 4 jul. 2025.

CHAUVEL, Louis. Le destin des générations: structure sociale et cohortes en France du XXe siècle aux années 2010. 2. ed. Paris: Quadrige, 2014.

COLLINS, Patricia Hill. Interseccionalidade. Tradução de Rane Souza. São Paulo: Boitempo, 2020.

COSTA, Luana; SCALON, Celi. Income inequality and social stratification in Brazil: key determining factors and changes in the first decade of the 21st century. In: PEILING, Li et al. (org.). Handbook on social stratification in the BRIC countries: change and perspective. Cingapura: World Scientific Publishing, 2013. p. 421-438.

ERICKSON, Robert; GOLDTHORPE, John H. The constant flux: a study of class mobility in industrial societies. Oxford: Clarendon Press, 1993.

GOLDTHORPE, John H.; LLEWELLYN, C.; PAYNE, C. Social mobility and class structure in modern Britain. Oxford: Clarendon Press, 1987.

HAYKIN, S. Redes neurais: princípios e práticas. 2. ed. Porto Alegre: Bookman, 2001.

Revista ARACÊ
ISSN: 2358-2472

HE, Haibo; MA, Yunqian. Aprendizagem desbalanceada: fundamentos, algoritmos e aplicações. [S.l.]: [s.n.], 2013.

IBGE. Tabela de Natureza Jurídica 2021 – Notas Explicativas. Rio de Janeiro: IBGE, 2021. Disponível em: https://concla.ibge.gov.br/images/concla/documentacao/CONCLA-TNJ2021-NotasExplicativas.pdf. Acesso em: 4 jul. 2025.

IBGE. Pesquisa Nacional por Amostra de Domicílios Contínua: notas metodológicas. Rio de Janeiro: IBGE, 2014.

LAVALLEY, Michael P. Logistic regression. Circulation, v. 117, n. 18, p. 2395-2399, 2008.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS, 31., 2017, Long Beach. Proceedings [...]. Long Beach: NIPS, 2017.

MOLNAR, C. Interpretable machine learning: a guide for making black box models explainable. [S.l.]: Lulu, 2018.

NOGARE, Diego. Performance de machine learning – Matriz de confusão. Disponível em: https://diegonogare.net/2020/04/performance-de-machine-learning-matriz-de-confusao/. Acesso em: 4 jul. 2025.

PASTORE, José; DO VALLE SILVA, Nelson; CARDOSO, Fernando Henrique. Mobilidade social no Brasil. São Paulo: Makron Books, 2000.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

RIBEIRO, Carlos Antonio Costa. Mobilidade e estrutura de classes no Brasil contemporâneo. Sociologias, Porto Alegre, v. 16, p. 178-217, 2014.

SAVAGE, Mike. Social classes and inequalities: sociability, culture and politics. Tempo Social, São Paulo, v. 28, n. 2, p. 1-25, maio/ago. 2016. Disponível em: https://doi.org/10.11606/0103-2070.ts.2016.110570. Acesso em: 4 jul. 2025.

SCALON, Celi. Mobilidade social no Brasil: padrões e tendências. Rio de Janeiro: Revan-Ipuerj-UCM, 1999.

SCALON, Celi. Social stratification and its transformation in Brazil. In: PEILING, Li et al. (org.). Handbook on social stratification in the BRIC countries: change and perspective. Cingapura: World Scientific Publishing, 2013. p. 3-19.

SCALON, Celi; SALATA, André. Uma nova classe média no Brasil da última década? O debate a partir da perspectiva sociológica. Sociedade e Estado, Brasília, v. 27, n. 2, p. 387-407, 2012. Disponível em: https://periodicos.unb.br/index.php/sociedade/article/view/5658. Acesso em: 4 jul. 2025.

SILVA, Nelson do Valle; HASENBALG, Carlos. Tendências da desigualdade educacional no Brasil. Dados, Rio de Janeiro, v. 43, n. 3, p. 423-445, 2000.

Revista ARACÊ
ISSN: 2358-2472

SILVA, Nelson do Valle; HASENBALG, Carlos A. (org.). Relações raciais no Brasil contemporâneo. Rio de Janeiro: Rio Fundo, 1992.

SMOLA, A.; SCHÖLKOPF, B. A tutorial on support vector regression. Statistics and Computing, v. 14, p. 199-222, 2004.

WEBER, Max. Economia e sociedade: fundamentos da sociologia compreensiva. Tradução de Regis Barbosa e Karen Elsabe Barbosa. Brasília: Editora Universidade de Brasília, 1999.

WRIGHT, Erik Olin. Análise de classes: abordagens. Petrópolis: Vozes, 2015.