**Modelo Preditivo do IDSUS 2010 a partir dos Indicadores Sociais Municipais do Censo 2010**

**Autor: Yuri Piffer**

**Disciplina: PCI510009 – Machine Learning e Deep Learning**

**Pós-Graduação Universidade Federal de Santa Catarina**

**Link para o repositório: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx**

1. **INTRODUÇÃO**
   1. **IDSUS 2010**

O Sistema Único de Saúde (SUS) do Brasil é um sistema de saúde pública universal e gratuita instaurado pela constituição de 1988. Sua atuação se dá desde atendimentos primários até procedimentos complexos, distribuição de vacinas e medicamentos e financiamento de pesquisas. Entre algumas ferramentas de avaliação do SUS, tem-se o Índice de desempenho do Sistema Único de Saúde (IDSUS). O objetivo do IDSUS é:

Avaliar o desempenho do SUS quanto à: universalidade do acesso, integralidade, igualdade, resolubilidade e equidade da atenção, descentralização com comando único por esfera de gestão, responsabilidade tripartite, regionalização e hierarquização da rede de serviços de saúde [REF].

Há diversos desafios que permeiam uma avaliação coerente do desempenho do Sistema Único de Saúde em toda a extensão territorial do Brasil devido à complexidade do processo demográfico, social e sanitário do país.

É importante destacar que, em função da grande diversidade (demográfica, cultural, socioeconômica, geográfica, etc.) dos territórios do nosso país, não seria adequado realizar uma classificação (ranqueamento) que apenas posicionasse, em ordem crescente ou decrescente, os municípios brasileiros. Assim, para realizar uma avaliação mais justa, a análise comparativa dos resultados do IDSUS deve levar em consideração a existência de Grupos Homogêneos [REF].

A formação dos Grupos Homogêneos (clusterização) utilizam três índices que possibilitam a análise do desempenho do SUS dos municípios brasileiros em seis grupos homogêneos, sendo esses índices o:

* Índice de Desenvolvimento Socioeconômico (IDSE),
* Índice de Condições de Saúde (ICS),
* Índice de Estrutura do Sistema de Saúde do Município (IESSM).

A partir da combinação destes três indicadores, estabelecem-se estes seis grupos, conforme a tabela [REF]. As melhores condições são encontradas no grupo 1. Os grupos seguintes apresentam condições inferires até chegar no grupo 6. Por Estrutura MAC (Média e Alta Complexidade) entende-se a estrutura de alta e média complexidade ou estrutura de atenção especializada, ambulatorial e hospitalar, urgência/emergência.

|  |  |  |  |
| --- | --- | --- | --- |
| **tABELA** | | | |
| **GRUPO** | **IDSE** | **ICS** | **IESSM** |
| **6** | baixo | baixo | Sem estrutura MAC (\*) |
| **5** | médio | médio | Sem estrutura MAC |
| **4** | baixo | baixo | Pouca estrutura MAC |
| **3** | médio | médio | Pouca estrutura MAC |
| **2** | alto | médio | Média estrutura MAC |
| **1** | alto | médio | Muita estrutura MAC |

Para a clusterização dos municípios brasileiros nestes seis grupos homogêneos no IDSUS de 2010 foram utilizados 24 indicadores de saúde: Acesso Potencial (que permite inferir a oferta potencial de atendimentos); Acesso Obtido (que pode ser medido por meio dos atendimentos realizados) e Efetividade (que mostra o grau com que serviços e ações atingem os resultados esperados). Utiliza-se a Análise de Componentes Principais (PCA), para calcular o peso desses 24 indicadores. Os subconjuntos ‘Índice de Acesso Potencial ou Obtidoo do SUS’ e ‘Índice de Efetividade do SUS’ são submetidos à metodologia de Análise de Cluster pelo método K-means para encontrar os grupos homogêneos de municípios passíveis de comparação, conforme pode ser visto na figura xxxxxxxxxxx.

|  |
| --- |
|  |
| Fig. xxxx |

O resultado da clusterização, por região do país, pode ser visto na tabela xxxxxxxx referente ao IDSUS de 2010. Os dados mostram que o grupo homogêneo 3 (ID3) representa 1.623 municípios, o grupo com mais municípios nele, muito embora represente apenas 13.6% da população brasileira de 2010. Os grupos homogêneos com melhor desempenho (ID1 e ID2) totalizam 1.776 municípios com 18.9% da população de 2010. O ID4 corresponde à 579 municípios que abrangem 41.1% da população de 2010, tratando-se das cidades de médio e grande porte. O ID4 e ID5, as classificações mais baixas, correspondem à 1.585 municípios com o total de 26.5% da população brasileira.

Outro fator importante apontado pela tabela é a disparidade de qualidade de atendimento do SUS por região do país. Enquanto sul e sudeste somam 1.631 municípios com IDSUS nos grupos homogêneos ID1 e ID2 e 201 municípios nos grupos ID5 e ID6, as regiões norte, nordeste e centro-oeste concentram 145 municípios nos grupos homogêneos ID1 e ID2 contra 1.384 municípios nos grupos ID5 e ID6, de pior desempenho.

|  |
| --- |
|  |
| tabela xxxxxxxx :  Distribuição dos municípios, população total e população exclusivamente SUS, por região brasileira e por cluster de municípios segundo o IDSUS 2010. |
|  |

<http://idsus.saude.gov.br/grupos.html>

* 1. **Censo 1010**

A última operação censitária no Brasil foi em 2010 realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A operação censitária percorre todas as cidades do país. No Censo 2010 foram 5.565 municípios e mais de 67 milhões de domicílios visitados. Os resultados do Censo retratam a população brasileira e suas características socioeconômicas que podem servir como base para o planejamento público.

Um resumo dos dados coletados pelo Censo pode ser obtido pelos Indicadores Sociais Municipais que contam com 17 tabelas para cada estado brasileiro [xxxxx]. Nesses dados são retratados para cada município tamanho da população, faixa etária, unidades domésticas e concentração de moradores, acesso à esgoto, etnia, renda e escolaridade, conforme a lista da tabela xxxx a seguir:

|  |  |
| --- | --- |
| Tabela xxxx:  Lista das 17 planilhas que compõem os pelos Indicadores Sociais Municipais do Censo 2010. | |
| **Tabela** | **Descrição do tipo de Dado** |
| 1 | Percentual de Distribuição Urbano / Rural - Homem / Mulher. |
| 2 | Distribuição percentual, por grupos de idade. |
| 3 | Pessoas de 15 anos ou mais de idade que não sabem ler e escrever, total e respectivas taxas de analfabetismo, por grupos de idade (%). |
| 4 | Unidades domésticas, por sexo das pessoas responsáveis pelos domicílios particulares permanentes. |
| 5 | Unidades domésticas, por tipo ( Único responsável / Com mais de Responsável ). |
| 6 | Unidades domésticas, por tipo (Estrutura familiar - Casal com filhos, homem com filhos etc.). |
| 7 | Domicílios particulares permanentes Distribuição percentual, por tipo de saneamento (Adequado, Semi-Adequado, Inadequado). |
| 8 | Rendimento mensal total domiciliar per capita nominal. |
| 9 | Valor médio e mediano do rendimento mensal total nominal, por sexo. |
| 10 | Valor médio do rendimento mensal total nominal, por cor ou raça. |
| 11 | Razão entre médias do rendimento mensal total nominal, por cor ou raça. |
| 12 | Proporção de pessoas residentes em domicílios particulares permanentes, por classes selecionadas de rendimento mensal total domiciliar per capita nominal. |
| 13 | População residente em domicílio particulares permanentes com saneamento inadequado Proporção de pessoas, por classes selecionadas de rendimento mensal total domiciliar per capita nominal. |
| 14 | Taxa de analfabetismo da população de 15 anos ou mais de idade por grupos de idade. |
| 15 | Proporção de domicílios particulares permanentes, por tipo de saneamento Ano 2000/2010. |
| 16 | Proporção de crianças de 0 a 5 anos de idade, residentes em domicílios particulares permanentes (com responsável ou cônjuge analfabeto, com saneamento inadequado, com responsável ou cônjuge analfabeto e saneamento inadequado). |
| 17 | Pessoas de 10 anos de idade que não sabem ler e escrever, total e respectiva taxa de analfabetismo. |

* 1. **Modelos Preditivos**

Um modelo preditivo, de forma simplificada, é um algoritmo com capacidade de ser aplicado a uma grande quantidade de dados e retornar padrões do dataset em questão. Entre vários algoritmos de aprendizado de máquina que atuam como modelos preditores, será abordado aqui dois algoritmos baseados em árvores de decisão, sendo eles Random Forest e Extra Trees. Embora os dois algoritmos utilizem árvores de decisão, a diferença está na quantidade de processos aleatórios empregados em cada [algoritmo](https://didatica.tech/algoritmos-de-machine-learning/).

O Random forest é um algoritmo que combina n árvores de decisão, sendo o total de árvores (n) passado como parâmetro. Cada uma dessas árvores utiliza um subconjunto das variáveis preditoras. A classificação acontece pela contagem da maioria dos votos das árvores e a regressão pela média dos votos de cada árvore. No Random forest há randomização da seleção de variáveis e da seleção de amostras, mas o threshold (split dos dados) é escolhido de acordo com a melhor divisão encontrada pelo algoritmo. No Extra Trees, sua diferença em relação ao Random Forest é que a escolha do threshold também é aleatória.

Em termos práticos, a escolha do threshold aleatório fará com que cada árvore seja mais fraca na sua capacidade de predição. Isso pode reduzir a capacidade de enviesamento do combinado de árvores, permitindo que o modelo em si se torne mais generalista. Consequentemente, é de extrema importância que o Extra Trees trabalhe com variáveis representativas e que, portanto, haja uma pré-seleção das mesmas.

* 1. **Justificativa**

Nos dois indicadores apresentados nos tópicos anteriores (IDSUS e Censo) há uma complexidade no processo de aquisição dos dados. No IDSUS há uma contagem de exames, procedimentos, internações e curas. O Censo exige a visitação dos domicílios brasileiros em todas as cidades. Ambos exigem um alto empenho de coleta de dados para obtenção dos resultados e que envolvem um alto custo operacional. Seria de grande valia que parte dos dados do Censo, especialmente os dados mais fáceis de serem coletados, pudessem ser utilizados para predizer a classificação do IDSUS.

* 1. **Objetivos**

São descritos abaixo o objetivos primário e secundários deste trabalho:

Objetivo primário:

* Desenvolver um sistema preditivo com aprendizado de máquina para a classificação do IDSUS 2010 dos municípios brasileiros utilizando os dados do Censo 2010.

Objetivos secundários:

* Avaliar a diferença de desempenho entre algoritmos de Randon Forest e Extra Trees para o desempenho da tarefa do objetivo principal,
* Apontar as variáveis (features) do Censo mais significativas para prever o IDSUS.

1. **MATERIAIS E MÉTODOS**

Trazer o github, falar das duas etapas abaixo

* 1. **Preparação dos dados**

Para a realização do modelo preditivo foi construído um dataset contendo todas as cidades brasileiras nas linhas e nas colunas os dados dos grupos homogêneos do IDSUS 2010 e os dados dos Indicadores Sociais Municipais do Censo 2010. Este dataset está disponível no repositório como ‘TABELA\_CENSO\_IDSUS\_2010.csv’ e conta com 5.560 linhas de cidades e 185 colunas de variáveis. É a partir dele que se executa o modelo preditivo. O total de 5.560 municípios já excluem Fernando de Noronha – PE e Nazária – PI por não terem dados do IDSUS, Brasília-DF por não constar nos Indicadores Sociais Municipais e São Valério da Natividade – TO, pois apresentou problemas na extração de dados.

Da planilha do IDSUS 2010 [] foram extraídas para uma planilha menor apenas as colunas com nome dos municípios, UF, ‘idsus\_fx’ e ‘Grupos Homogêneos’. O arquivo se encontra no repositório com o nome de ‘idSUS2010\_cidades.xlsx’.

http://i3geo.saude.gov.br/i3geo/sage\_tabelas/avaliacao/IDSUS\_2010\_Indic\_resultados%20e%20notas\_Pagina%20V30\_01\_2013.xls

Do Censo 2010 [REF] foram escolhidas as planilhas de Indicadores Sociais Municipais, as quais foram reorganizadas em pastas com os nomes dos referidos estados, cada estado contando com as 17 planilhas descritas na introdução. A planilha 6, que retrata a distribuição familiar, não foi utilizada. A pasta se encontra no repositório com o nome ‘indicadores\_sociais\_municipais’ que conta com subpastas nomeadas com os nomes dos estados em minúsculo e com as palavras separadas por ‘\_’.

O script em linguagem python para a extração desses dados pode ser encontrado no repositório como ‘Extração de Dados do Censo 2010. ipynb’. É importante destacar que o nome das cidades foram manipuladas para ter no nome da cidade concatenado à unidade federativa (p.ex.: ‘Nome Da Cidade\_UF’) tendo em vista que várias cidades do Brasil têm o mesmo nome em estados diferentes. Os valores faltantes apresentados como ‘-‘ pelos dados do Censo foram substituídos por NaN para melhor processamento.

* 1. **Algoritmos de classificação**

Foram escolhidos dois algoritmos de classificação para a realização deste trabalho sendo eles o Randon Forest e o Extra Trees. O script em linguagem python para a manipulação dos dados e predição pode ser encontrado no repositório como ‘Modelo de Predição IDSUS 2010 a partir do Censo 2010.ipynb’.

Duas variáveis advindas do IDSUS 2010 são utilizadas como classe (target) neste trabalho, são elas o ‘idsus\_fx’ que compreende as variáveis categóricas: '2,5 a 3,9', '4 a 4,9', '5 a 5,9', '6 a 6,9', '7 a 7,9', '8 e +' e a classe ‘Grupo Homogêneo’ com as variáveis categóricas: 'GH1', 'GH2', 'GH3', 'GH4', 'GH5', 'GH6'. Ambas as classes contam com 6 valores nominais.

Para a variável classificadora ‘idsus\_fx’, a classe ‘5 a 5,9’ apresentou maior frequência (47,8%, ou 2663 dos 5560 municípios). Para a variável classificadora ‘Grupo Homogêneo’, a classe ‘GH6’ apresentou maior frequência (39,2%, ou 2182 dos 5560 municípios). Em termos práticos, estas porcentagem atuam como ‘baselines’, ou seja, qualquer algoritmo preditor precisa ter acurácia maior que 47,8% para classificação do ‘idsus\_fx’ e 39,2% para classificação do ‘Grupo Homogêneo’. Caso contrário, é melhor abandonar o modelo preditor e apostar todos os casos no elemento da classe de maior frequência.

Alguns tratamentos foram executados no dataset:

* Exceto as duas variáveis classificadoras (‘Grupo Homogêneo’ e ‘idsus\_fx’), os outras 185 variáveis preditoras já se encontravam como valores numéricos. Assim sendo, só as primeiras precisaram do preprocessamento de labelEncoder.
* Na maioria das simulações, os valores faltantes (NaN) foram substituídos pela média da variável utilizando SimpleImputer;
* Na maioria das simulações, os variáveis preditoras sofreram escalonamento pelo StandardScaler;
* Em algumas das simulações, foi realizado escolha das variáveis (‘feature selection’) por meio de Selectkbest e qui-quadrado para os modelos de Random Forest ou o próprio método de feature selection das Extra Trees implementado pelo biblioteca Scikit learn.

A estudo foi dividido em duas etapas sendo a primeira a avaliação individual do desempenho dos modelos preditivos com Random Forest e Extra Tree variando-se a configuração de cada simulação. A segunda etapa utiliza validação cruzada a partir dos modelos mais representativos da etapa anterior.

1. **RESULTADOS**

Nas tabelas xxxx e xxxx são apresentados a acurácia dos modelos de avaliação individual do dataset em questão utilizando-se como classe target o ‘idsus\_fx’ ou o ‘Grupo Homogêneo’, ambas variáveis advindas do IDSUS 2010.

|  |  |  |
| --- | --- | --- |
| Algoritmo Random Forest com simulações utilizando diferentes configurações. Comparação de performance para ‘idsus\_fx’ e ‘Grupo Homogêneo’ (random state = 0) | | |
|  | **Acurácia** | |
| **Configurações de simulação** | **‘idsus\_fx’** | **‘Grupo Homogêneo’** |
| Com SimpleImputer (‘mean’) e StandardScaler: |  |  |
| * 15 árvores (train/test 20%) | 0.559 | 0.842 |
| * 60 árvores (train/test 30%) | 0.571 | 0.859 |
| * 60 árvores (train/test 20%) | 0.574 | 0.843 |
| * Sem StandardScaler , 60 árvores (train/test 20%) | 0.575 | 0.843 |
| * 60 árvores (train/test 20%), SelectKbest (chi2, k=20 variáveis) | 0.543 | 0.808 |
| * 60 árvores (train/test 20%), SelectKbest (chi2, k=60 variáveis) | 0.561 | 0.849 |
| Excluindo colunas com NaN (restou 64 variáveis), StandardScaler, 60 árvores (train/test 20%) | 0.569 | 0.843 |

|  |  |  |
| --- | --- | --- |
| Algoritmo Extra Trees com simulações utilizando diferentes configurações. Comparação de performance para ‘idsus\_fx’ e ‘Grupo Homogêneo’ (random state = 0) | | |
|  | **Acurácia** | |
| **Configurações de simulação** | **‘idsus\_fx’** | **‘Grupo Homogêneo’** |
| Com SimpleImputer (‘mean’) e StandardScaler: |  |  |
| * (train/test 20%) | 0.589 | 0.849 |
| * (train/test 30%) | 0.573 | 0.855 |
| * Sem StandardScaler , (train/test 20%) | 0.564 | 0.845 |
| * (train/test 20%), Selection feature importance > 0.018   (1 variável ‘Grupo Homogêneo’) | - | 0.560 |
| * (train/test 20%), Selection feature importance > 0.015   (3 variáveis ‘Grupo Homogêneo’) | - | 0.690 |
| * (train/test 20%), Selection feature importance > 0.014   (7 variáveis ‘Grupo Homogêneo’) | - | 0.82.4 |
| * (train/test 20%), Selection feature importance > 0.006   (45 variáveis ‘idsus\_fx’), (61 variáveis ‘Grupo Homogêneo’) | 0.573 | 0.849 |
| * (train/test 20%), Selection feature importance > 0.0055   (77 variáveis ‘idsus\_fx’), (75 variáveis ‘Grupo Homogêneo’) | 0.581 | 0.849 |
| * (train/test 20%), Selection feature importance > 0.005   ( variáveis ‘idsus\_fx’), (81 variáveis ‘Grupo Homogêneo’) | 0.591 | 0.847 |
| Excluindo colunas com NaN (restou 61 variáveis), StandardScaler (train/test 20%) | 0.571 | 0.853 |

Os resultados preditivos do ‘idsus\_fx’ foram bastante inferiores ao ‘Grupo Homogêneo’ para o mesmo dataset. Por tais razões e, considerando que todo o IDSUS se constrói a partir dos seis grupos homogêneos, serão descritos aqui apenas os resultados do ‘Grupo Homogêneo’.

De forma generalista, é possível dizer que a acurácia da predição dos dados do IDSUS utilizando o Censo 2010 atingiu uma média de 84% para simulações isoladas, seja por Random Forest ou Extra Trees, com ou sem escalonamento de dados e substituição de dados faltantes pela média.

Houve uma pequena melhora quando a divisão entre treinamento e teste foi mudada de 0.2 para 0.3, o que não serve como argumento para a escolha de um modelo uma vez que ainda serão apresentados os dados de validação cruzada. A exclusão de todas as colunas que apresentavam algum valor vazio, reduzindo o dataset de preditores de 185 para 61 colunas, não teve impacto significativo, seja para melhor ou para pior desempenho, muito embora tal prática não seja recomendada.

Por outro lado, o processo de seleção de variáveis apontou diferenças de desempenho. Nas simulações utilizando o algoritmo Random Forest, a utilização do selectKbest utilizando o qui-quadrado para escolher as 20 e 60 mais representativas variáveis resultou em acurácia de 80.9% e 84.9%, respectivamente. No Extra Trees, utilizando o método de Selection Feature Importance é possível ranquear as variáveis a partir da sua representatividade na predição do modelo e, a partir deste ranking, escolher quais devem ser utilizadas. Pelo filtro de variáveis com score mais alto que:

* 0.018, restando 1 variável preditora, resultou em acurácia de 56%;
* 0.015, restando 3 variáveis preditoras, resultou em acurácia de 69%;
* 0.014, restando 7 variáveis preditoras, resultou em acurácia de 82.4%;
* 0.006, restando 61 variáveis preditoras, resultou em acurácia de 81.4%;
* 0.0055, restando 75 variáveis preditoras, resultou em acurácia de 84.9%;
* 0.005, restando 81 variáveis preditoras, resultou em acurácia de 84.7%;

A variável preditora única, que resultou em acurácia do modelo de 56%, se refere à coluna 145, ‘Popu\_permanente\_GERAL\_renda\_até60%daMedianaBrasil(R$255.00)’ que compreende a proporção (%) de pessoas residentes em domicílios particulares permanentes de rendimento mensal total domiciliar per capita representando até 60% da mediana total do Brasil em 2010 (R$ 225,00), disponível na tabela 12 dos indicadores do Censo 2010.

As 7 variáveis resultantes quando o filtro da Selection Feature Importance foi configurado para 0.014 (acurácia de 82.4%) se referem às colunas 89, 113, 144, 145, 148, 149 e 150. Embora não caiba explicar cada uma das variáveis, é perceptível que seis das sete features referem-se à renda per capita. Tais colunas, respectivamente, são:

* “Unid\_doméstica\_HOMEM\_total”,
* “Renda\_mensal\_percapta\_GERAL\_valor\_médio”,
* “Popu\_permanente\_GERAL\_renda\_atéR$255.00(1/2)”,
* “Popu\_permanente\_GERAL\_renda\_atéR$255.00(1/2)”,
* “Popu\_permanente\_GERAL\_renda\_até60%daMedianaBrasil(R$255.00)”,
* “Popu\_permanente\_URBANA\_renda\_atéR$127.50(1/4)”,
* “Popu\_permanente\_URBANA\_renda\_atéR$255(1/2)”,
* “Popu\_permanente\_URBANA\_renda\_até60%daMedianaBrasil(R$249)”.

Na tabela xxxxxxx são apresentados os resultados da simulação utilizando o algoritmo de Extra Trees com validação cruzada (10 ‘folds’) para todas as 185 variáveis e também para apenas as 7 variáveis mais representativas listadas anteriormente.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tabela xxxx. | | | | | | | | | |
| Simulação por Extra Trees utilizando validação cruzada (todas as 185 variáveis)  Resultado de acurácia para cada uma das 10 ‘folds’: | | | | | | | | | |
| *1º* | *2º* | *3º* | *4º* | *5º* | *6º* | *7º* | *8º* | *9º* | *10º* |
| 0.682 | 0.861 | 0.841 | 0.897 | 0.874 | 0.879 | 0.881 | 0.841 | 0.782 | 0.633 |
| **Média dos ‘folds’:** | | | **0.817** | |  |  | Desvio Padrão: | | 0.086 |
| Simulação por Extra Trees utilizando validação cruzada (apenas 7 variáveis principais)  Resultado de acurácia para cada uma das 10 ‘folds’: | | | | | | | | | |
| *1º* | *2º* | *3º* | *4º* | *5º* | *6º* | *7º* | *8º* | *9º* | *10º* |
| 0.673 | 0.868 | 0.832 | 0.886 | 0.843 | 0.863 | 0.872 | 0.812 | 0.739 | 0.607 |
| **Média dos ‘folds’:** | | | **0.800** | |  |  | Desvio Padrão: | | 0.090 |

A utilização de validação cruzada permite uma avaliação mais precisa do modelo de predição, uma vez que o mesmo dataset é dividido entre treino e teste de diferentes maneiras de forma que toas as partes sejam utilizadas nas duas categorias em diferentes ‘folds’. A média dos resultados é também mais representativa para explicar o modelo, assim como o desvio padrão dos resultados de acurácia dos 10 folds ajuda a demonstrar a possibilidade de overfitting.

Os resultados da média para o modelo de Extra Trees utilizando validação cruzada com 10 folds para todas as variáveis e para somente as 7 variáveis principais foi de 81.7% e 80%. Os resultados tornam evidentes como, a partir deste dataset, é possível predizer o desempenho IDSUS dos municípios a partir de poucas variáveis uma vez que a diferença de 197 variáveis entre as duas simulações incrementou a eficiência do modelo em apenas 1.7%.

É preciso notar também que o primeiro e último fold das duas simulações apresentam acurácia mais baixa e refletem em um desvio padrão não tão baixo quando poderia se esperar. Talvez alguma outra abordagem estatística para estes dados poderiam ser realizadas, ou até a adição de outras variáveis que ajudem o modelo a desempenhar melhor.

1. **CONCLUSÃO E RECOMENDAÇÕES**

A utilização dos dados do Censo 2010, em especial a proporção de cidadãos de acordo com faixa de renda per capita, permitem elaborar um modelo de predição do IDSUS 2010 com a utilização de árvores de decisão. Não houve uma diferença de desempenho substancial entre os algoritmos Random Forest e Extra Trees. Por fim, a variável ‘idsus\_fx’ não é recomendada para a predição do desempenho dos municípios no IDSUS. Recomenda-se a futura execução deste modelo com os dados do Censo 2021 quando o mesmo for publicado.

**BIBLIOGRAFIA**