Os medos dos paulistanos - Uma análise social

João Davi Silva Mendes
RA 175904
Campinas - SP, Brasil
j175904@dac.unicamp.br

Marília Gabriela Rocha
RA 183881
Campinas - SP, Brasil
m183881@dac.unicamp.br

Pedro Aurélio Oliveira Morgado
RA 185560
Campinas - SP, Brasil
p185560@dac.unicamp.br

***Abstract*—O Projeto de Pesquisa consiste em analisar as respostas de um questionário aplicado pelo Datafolha com 1091 moradores da cidade de São Paulo, composto por 28 questões sobre medo, percepção de violência e características socioeconômicas. Esta análise será realizada por meio de métodos de Aprendizado Supervisionado de Máquina, tais como Árvores de Classificação, Floresta Aleatória, *Boosting* e Máquinas de Vetores de Suporte. Para redução do banco de dados, de modo a facilitar o estudo das respostas obtidas, serão usados métodos de Teoria de Resposta ao Item. O objetivo da análise consiste em verificar relações entre os medos dos paulistanos, suas percepções de violência e algumas características socioeconômicas com o sexo, etnia e renda familiar dos respondentes. O Projeto é de grande relevância devido a presença frequente do sentimento "medo" no cotidiano das pessoas e das assíduas exposições delas à situações de violência. Será utilizado no Projeto a ferramenta *RStudio*.**

***Index Terms*—TRI, medos, paulistanos, aprendizado supervisionado de máquina**

I. Introdução

Dados de junho de 2018 mostram que homicídios na cidade de São Paulo têm alta de 42% (ESTADÃO...,). Outros dados de outubro deste mesmo ano comprovam que o número de mortes em acidentes de moto aumentou 62% (G1...,). Levando em consideração o crescimento da violência na cidade de São Paulo juntamente com as mortes provindas de acidentes de trânsito, e partindo da premissa de que o meio em que as pessoas vivem interfere no modo como elas se comportam, um ponto interessante de ser levantado sobre esses aspectos é o quanto todos esses fatores podem influenciar nos medos dessas pessoas.

É de senso comum que os medos tem influência de fatores pessoais, sociais e econômicos. É possível que uma pessoa que more em um bairro mais vulnerável, sem políticas de segurança bem definidas, sinta mais medo de andar na rua a noite do que alguém de um bairro mais nobre, assim como uma mulher possuir mais receio em estar sozinha numa mesma rua escura do que um homem. Esses são fatores que são considerados terem influência na forma como as pessoas percebem as coisas ao seu redor e nas suas concepções de situações de risco.

Tendo como base um questionário sobre percepções do medo realizado em 2008 na cidade de São Paulo, busca-se analisar como os medos citados no questionário e algumas outras informações socioeconômicas estão correlacionados a situações pessoais (sexo, etnia) e econômicas e sociais (renda familiar) por meio de técnicas aprendidas no curso de Aprendizado Supervisionado de Máquina, assim como técnicas novas. As técnicas de Aprendizado Supervisionado de Máquina serão utilizadas na tentativa de prever quais serão o sexo, a etnia e a renda familiar dos entrevistados.

Instituto de Matemática, Estatística e Computação Científica - IMECC

II. Conjunto de Dados

A. Obtenção do conjunto de dados e descrição geral das variáveis

O conjunto de dados foi retirado do endereço eletrônico CIS (Consórcio de Informações Sociais) e contém o resultado de um questionário aplicado com 1091 moradores com idade acima de 16 anos da cidade de São Paulo pelo Datafolha - Instituto de Pesquisa em 15 de abril de 2008 (DATAFOLHA...,). O questionário é composto por 28 questões, que abarcam, além de características socioeconômicas dos respondentes, inúmeras perguntas sobre medo e percepção de violência, tendo um total de 86 variáveis, em sua maioria categóricas.

As perguntas que giram em torno do nível de medo dos respondentes a determinadas situações tem como modelo "Costuma sentir muito, um pouco ou não sente medo nenhum de...?" em que ao final é acrescentado o possível medo, tendo como possíveis respostas 0, 1, e 2, que significam, respectivamente, "Nenhum medo", "Pouco medo", e "Muito medo". Há também perguntas como "Dentre as situações abaixo, de qual você tem mais medo?", às quais possuem uma grande quantidade de possíveis respostas. Para fins de classificação, há também informações sobre sexo, idade, escolaridade, religião, renda familiar, estado conjugal, entre outras.

As possibilidades de respostas 97 e 99 significam, respectivamente, "Recusa-se a responder" e "Não sabe responder".

B. Manipulação dos dados

Foram retiradas do banco de dados as variáveis: Idade (em anos), mantendo-se somente a Idade (em faixas); Qual o maior medo do entrevistado, a qual foi retirado devido à possibilidade de 368 respostas diferentes, e portanto, não carregando muitas informações e de difícil análise; Qual país o entrevistado gostaria de nascer caso pudesse nascer de novo e Qual cidade brasileira gostaria de nascer caso pudesse nascer de novo, retiradas devido a grande diversidade de respostas possíveis e pouco relevância para os objetivos do projeto; Quantas vezes o entrevistado foi assaltado nos últimos doze meses, mantendo-se somente se ele foi assaltado ou não nos

últimos doze meses; Se o entrevistado joga ou não no bicho, retirada devido a irrelevância para os objetivos da pesquisa; Opinião do entrevistado sobre o trânsito de São Paulo, retirada por irrelevância; Idade de cada um dos filhos, retirada devido à diversidade de respostas possíveis, o que dificultaria a análise proposta. Além disso, as variáveis referentes a quantidade de eletrodomésticos, cômodos na residência, automóveis, etc, foram alteradas para tornarem-se categorias, sendo ”0” - O entrevistado não tem o que é perguntado, ”1” - O entrevistado tem um, ”2” - O entrevistado tem dois ou mais.

III. Metodologia

Para realização do projeto, utilizou-se principalmente o programa estatístico *R*. Além disso, para entendimento dos métodos usados, alguns endereços *online*, artigos e livros foram acessados e estão citados nas referências.

A. Análise Descritiva

Como todas as variáveis do banco de dados, após as manipulações iniciais, são categóricas e diferem-se em dicotômicas e politômicas, nominais e ordinais, não foi possível realizar um análise descritiva tão detalhada, não sendo possível verificar as correlações entre as variáveis. Analisou-se, portanto, o comportamento das variáveis de modo mais individual.

B. Redução da dimensão dos dados

O método escolhido para redução da dimensão dos dados foi Teoria de Resposta ao Item (TRI), comumente utilizado para a análise de questionários e listas de itens e sendo uma vertente da Análise Fatorial. Tal metodologia sugere formas de representar a relação entre a probabilidade de um indivíduo dar uma certa resposta a um item e seus traços latentes. (ANDRADE; TAVARES; VALLE, 2000)

Devido ao fato de algumas das respostas do questionário que são relacionadas ao medo não serem dicotômicas, ou seja, são itens de múltipla escolha avaliados de forma gradual com níveis intermediários, tornou-se necessário usar duas diferentes abordagens para a definição do traço latente referente a estas questões.

Escolheu-se por utilizar o modelo unidimensional, já que havia interesse em mensurar um único traço latente que fosse representativo da variável latente ”medo”.

1) ***Modelo Logístico de 2 Parâmetros (L2P)****:* Utilizou-se do Modelo Logístico de 2 Parâmetros para aplicação nas 6 variáveis que apresentavam valores de 0 e 1, representando ”sim” e ”não”, respectivamente, sendo classificadas, assim, como dicotômicas. O uso do modelo que apresenta dois parâmetros, considera a dificuldade e a discriminação, e está definido por:

$$P_{ij} = P(U_{ij} = 1|\theta_{ij}) = \frac{1}{1+e^{-a_i(\theta_j - b_i)}}$$

em que U_{ij} representa a resposta do indivíduo j ao item i, θ_j o traço latente do indivíduo j e a_i e b_i são os parâmetros de discriminação e dificuldade, respectivamente (ANDRADE; TAVARES; VALLE, 2000).

2) ***Modelo Gradual Parcial Generalizado****:* Para as 30 variáveis de respostas politômicas utilizou-se o Modelo Gradual Parcial Generalizado de Muraki, permitindo uma análise de itens de múltipla escolha que são avaliados de forma graduada. O modelo está definido por:

$$P_{ik}(\theta_j) = \frac{exp[\sum_{u=0}^{k=} Da_i(\theta_j - b_{i,u})]}{\sum_{u=0}^{m_i} exp[\sum_{v=0}^{u} Da_i(\theta_j - b_{i,v})]}$$

com $i = 1, 2, ..., I$, $j = 1, 2, ..., n$ e $k = 0, 1, ..., m_i$ e em geral $b_{i,0} = 0$ e onde

$P_{ik}(\theta_j)$ - probabilidade do indivíduo j selecionar uma opção k (de $m_i + 1$ categorias possíveis);

D - fator de escala, constante e igual a 1. Utiliza-se o valor 1,7 quando deseja-se que a função logística forneça resultados semelhantes ao da função ogiva normal;

θ_j - habilidade do indivíduo j;

$b_{i,k}$ - parâmetro de dificuldade da k-ésima categoria do item i;

a_i - parâmetro de discriminação (ou inclinação) do item i (ANDRADE; TAVARES; VALLE, 2000).

C. Classificação dos dados

Com a finalidade de testar modelos de classificação para as variáveis respostas “sexo”, “renda familiar” e “etnia”, tendo como base o traço latente gerado anteriormente e as demais covariáveis do banco de dados que não foram usadas no cálculo do traço latente, foram realizadas classificações usando diferentes técnicas de Aprendizado Supervisionado de Máquina, tendo sempre como indicador de comparação a acurácia de cada modelo. Os métodos aplicados estão descritos brevemente abaixo.

1) ***Árvore de Decisão****:* A construção desse método se dá por particionamentos recursivos no espaço das covariáveis. Cada particionamento é chamado de nó e cada nó terminal recebe o nome de folha. Na Fig. 1 isso pode ser visualizado de melhor forma com um exemplo:

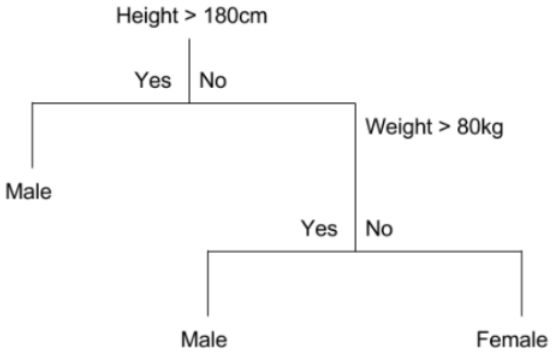

Fig. 1. Exemplo usando covariáveis altura e peso

Para realização do ajuste utilizou-se do pacote *rpart* do *R*.

Utilizou-se um Parâmetro de Complexidade (cp) igual à 0.015 para poda das árvores de decisão. A poda da árvore foi utilizada para evitar superajuste dos dados ao banco de treinamento.

Seja T_0 uma árvore muito grande. Considera-se o hiper-parâmetro não negativo α que corresponde a uma sub-árvore $T \subset T_0$, tal que

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - y_{\hat{R}_m})^2 + \alpha|T|$$

seja o menor possível. $|T|$ indica o número de nódulos terminais da sub-árvore T, R_m é o retângulo (isto é, um subconjunto do espaço preditor) correspondente ao m-ésimo nódulo terminal, e $y_{\hat{R}_m}$ é a resposta predita associada à R_m - ou seja,a média das observações de treino em R_m. (JAMES et al., 2013)

*2) **Floresta Aleatória**:* O método consiste em construir uma árvore para cada uma das amostras de treinamento *bootstrap*. A cada divisão realizada para a construção de cada árvore, uma amostra aleatória de m preditores é escolhido como candidatos do conjunto de p preditores possíveis ($m \approx \sqrt{p}$).

Para cada árvore de decisão criada, são realizadas as classificações para os dados e a partir de uma decisão da maioria se determina a etiqueta dos dados. O esquema pode ser bem ilustrado pela Fig. 2.

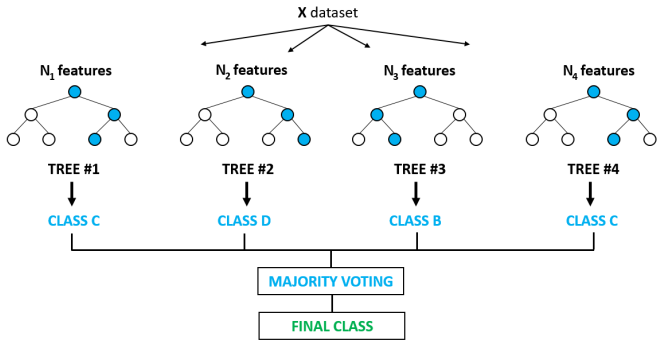

Fig. 2. Esquematização do processo de classificação da Floresta Aleatória.

- **Erro Out-of-Bag**: Método usado para estimar o erro de um modelo bagged, sem utilizar de métodos como validação cruzada. Comumente, as árvores utilizam 2/3 das observações como treino e o restante acaba ficando como observações de teste. As observações não utilizadas são chamadas de out-of-bag (OOB). Assim, se prediz a resposta para as observações de teste. Daí, é possível estimar o erro de teste do modelo como sendo o erro OOB, neste caso o erro de classificação.

*3) **Boosting**:* A aplicação desse método deu-se somente para a variável resposta sexo devido a ineficiência computacional em relação ao tempo de processamento para ajuste do método. O método consiste em ajustar uma sequência de algoritmos fracos a versões modificadas dos dados nas quais mais peso é dado a observações mais difíceis de classificar. Todo o processo se encontra ilustrado na Fig. 3

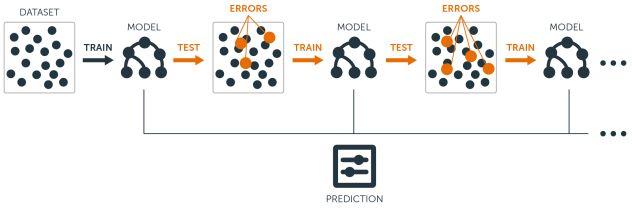

Fig. 3. Esquematização do processo de classificação do Boosting.

*4) **Máquinas de Vetores de Suporte (MVS)**:* O método tem como objetivo determinar o hiperplano que gere a maior margem entre as classes. A margem é definida como sendo a distância entre o hiperplano e o ponto mais próximo da classe A somado à distância do hiperplano ao ponto mais próximo da classe B. Ao maximizar a margem, o poder de generalização do classificador aumenta, ou seja, a chance do classificador acertar ao definir a qual classe um dado desconhecido pertence aumenta.

IV. Resultados

Com a utilização de técnicas de Teoria de Resposta ao Item, foi possível reduzir 36 variáveis relacionadas a medo, das quais 6 eram dicotômicas e 30 politômicas graduais (admitiam respostas do tipo 0 = "Nenhum medo", 1 = "Pouco medo", 2 = "Pouco medo" para a situação apresentada), em um único traço latente representativo do sentimento "medo".

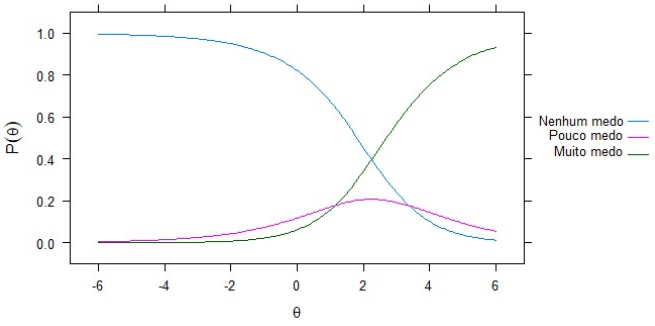

Fig. 4. Linhas de traço para medo de assombração.

Analisando-se a Fig. 4, percebe-se que quanto maior o traço latente, maior a probabilidade de escolha da categoria "Muito medo". Além disso, como o medo de assombração é um dos únicos medos não tão condizentes com situações palpáveis, supõe-se que esse seja o motivo do deslocamento para a direita da curva referente à categoria "Pouco medo" e do decaimento leve para a curva da categoria "Nenhum medo", visto que o traço latente é calculado conjuntamente para as questões e não individualmente.

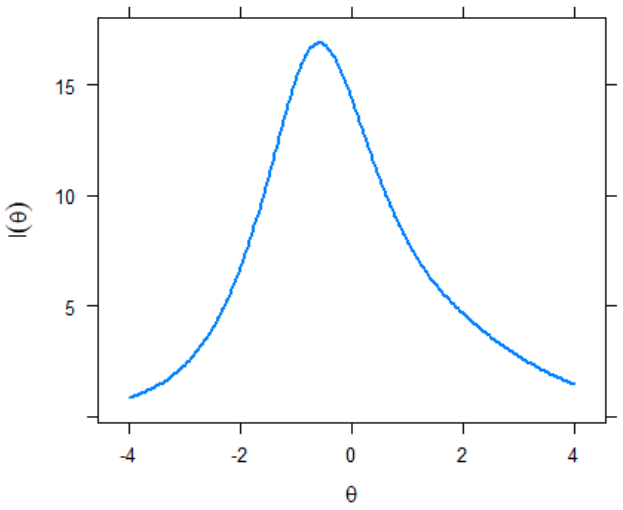

Fig. 5. Informação do teste de acordo com traço latente

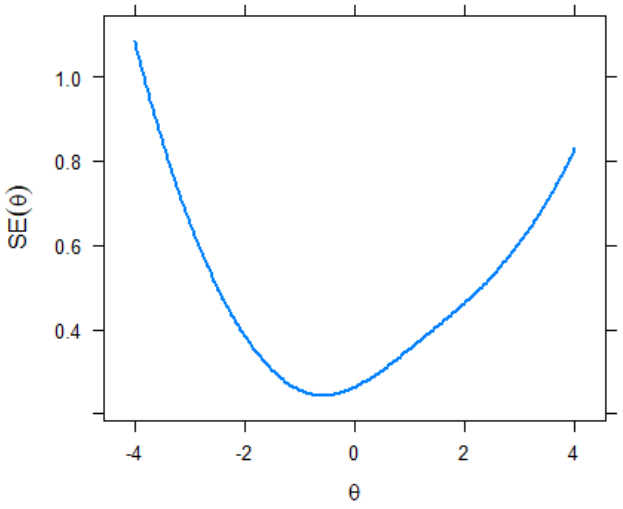

Fig. 6. Erro padrão do teste de acordo com traço latente

Com a Fig. 5 pode-se notar que a informação contida no teste, ou seja, nas variáveis do questionário utilizadas para o cálculo do traço latente, aumenta para valores centrais do traço latente e diminui para valores extremos. Ou seja, quanto mais central o traço latente do entrevistado, mais informações podem ser retiradas das questões referentes ao medo. Na Fig. 6 percebe-se que o erro padrão do teste tem funcionamento inverso ao da função de informação do teste, sendo menor para valores centrais de traço latente.

A partir do traço latente gerado anteriormente, ajustou-se quatro modelos de classificação, sendo eles Árvore de Classificação (AC), Floresta Aleatória (FA), Maquinas de Vetores de Suporte (MVS) e *Boosting*, sendo que, por questões de tempo demorado de execução, o método *Boosting* foi utilizado somente para a predição da variável resposta sexo. A acurácia encontrada para cada método aplicado em cada uma das variáveis resposta são expostas na Tabela I

TABLE I
ACURÁCIAS PARA CADA VARIÁVEL RESPOSTA POR MODELO UTILIZADO

	AC	FA	MVS-Linear	Boosting
Sexo	0.59	0.63	0.62	0.65
Cor	0.50	0.52	0.45	
Renda	0.39	0.40	0.41	

A. *Variável resposta "sexo"*

Para a predição da variável sexo no banco de dados de treino, obteve-se melhor resultado, em relação à acurácia, para o método *Boosting*, sendo esta de 0.65. No entanto, o método apresentou-se computacionalmente ineficiente, comparado aos demais métodos, devido ao tempo de execução. Além disso, a covariável de maior importância para a predição no método Boosting foi o traço latente, seguido pela categoria na população economicamente ativa. O método MVS-Linear apresentou acurácia de 0.63 e demonstrou-se computacionalmente eficiente.

TABLE II
IMPORTÂNCIA DAS CLASSES DAS COVARIÁVEIS PARA O MÉTODO BOOSTING PREVENDO SEXO

	Importância
Traço latente	100.00
pea11 - Dona de casa	40.78
idade2 - 25 a 34 anos	33.34
idade4 - 45 a 59 anos	26.17
cor3 - Parda	25.88
freezer 1 - 1 freezer	18.67

A Tabela II mostra a importância das variáveis para a predição no método *Boosting*. Ela mostra que o traço latente é a variável mais importante, o que já era esperado, uma vez que ela resume a informação de outras 36 variáveis. Em segundo lugar vem a categoria "dona de casa" na população economicamente ativa. Já era de se esperar que essa classe discriminasse bem o sexo, pois no geral as donas de casa são mulheres, e até mesmo no próprio questionário o termo apresenta-se no feminino, o que pode ter viesado as respostas. A terceira e a quarta classes mais importantes são as faixas etárias 25 a 34 anos e 45 a 59 anos da variável idade. Em seguida vem a classe parda da variável etnia, e por fim a classe de pessoas com um *freezer* em casa.

Para a predição da variável etnia, as acurácias para Árvore de Classificação, Floresta Aleatória, e MVS-Linear foram respectivamente 0.50, 0.52 e 0.45. Para os mesmos métodos para a variável renda, foram de 0.39, 0.40 e 0.41 respectivamente. Acredita-se que um dos motivos para as acurácias serem mais baixas para essas duas variáveis em relação à variável sexo é por essas serem politômicas e não dicotômicas.

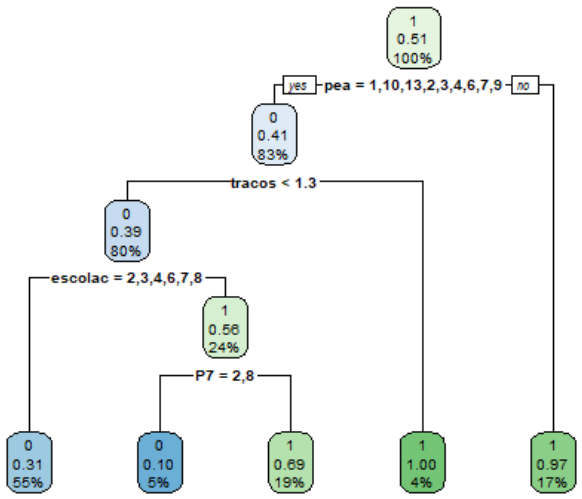

Fig. 7. Árvore de Classificação para a variável sexo

De acordo com a Fig. 7 [1], é possível perceber que a variável pea (Categoria na população economicamente ativa) é muito importante para a discriminação do sexo. Do primeiro nó à direita, a árvore já chega em uma folha que indica o sexo feminino. Isso acontece porque a variável pea considera também a categoria "Donas de casa". O segundo nó indica que o traço latente também é muito importante na discriminação, seguido da escolaridade do chefe da família (escolac), seguido por cidade natal do respondente (P7).

B. Variável resposta "etnia"

TABLE III
IMPORTÂNCIA DAS CLASSES DAS COVARIÁVEIS PARA O MÉTODO FLORESTA ALEATÓRIA - ERRO VALIDAÇÃO CRUZADA - PREVENDO ETNIA

	1	2
P78 - Cidade natal na Bahia	100.00	41.16
escolac7 - Ensino superior completo	43.89	60.45
religiao5 - Espirita Kardecista	72.87	40.44
comp2 - 2 computadores	36.72	19.01
aspira1 - 1 aspirador de pó	66.21	20.24
P20a3 - Mora com outros parentes	72.72	25.83

Para a classificação da variável etnia, a melhor acurácia obtida foi de 0.52 utilizando Floresta Aleatória com Validação Cruzada. A Tabela III mostra a importância das variáveis para as duas primeiras Árvores da Floresta. Não é possível fazer um ranking das variáveis porque a importância varia dependendo da Árvore.

Analisando-se a importância das classes das variáveis, como mostra a Tabela III, tem-se que a classe Cidade natal na Bahia, da variável cidade natal, é a mais importante. Em segundo lugar vem a escolaridade do chefe da família, em que o Ensino superior completo é a classe importante.

Observando-se a acurácia média, temos que as classes das covariáveis de maior impotância são: religiao5 - Espírita Kardecista, rendaf4 - Renda familiar de 2.076,00 reais até 4.150,00 reais, aspira1 - 1 aspirador de pó, e escola6 - Ensino superior incompleto.

Para a média de Gini, uma medida de pureza do nó (um valor pequeno indica que um nó contém predominantemente observações de uma única classe), o traço latente foi a variável mais importante, seguido por aspira1 - 1 aspirador de pó, P78 - Cidade natal na Bahia, e carro1 - 1 carro.

C. Variável resposta "Renda Familiar"

Conforme os resultados gerados para a variável resposta renda familiar, o melhor método em termos de acurácia foi Máquina de Vetores de Suporte com o valor de 0.41.

TABLE IV
IMPORTÂNCIA DAS CLASSES DAS COVARIÁVEIS PARA O MÉTODO FLORESTA ALEATÓRIA - ERRO OUT-OF-BAG - PREVENDO ETNIA

	Importância
Traço latente	100.00
comp1 - 1 computador	40.24
tvcor2 - 2 TV coloridas	26.21
aspira1 - 1 aspirador de pó	25.95
carro1 - 1 carro	24.01
freezer1 - 1 freezer	20.19

Tratando-se da importância das variáveis para o método floresta aleatória, notou-se altos valores para aquelas que estão diretamente ligadas à renda familiar, como número de bens existentes na casa dos respondentes. Isso pode ser explicado pela relação clara que quanto maior a renda, mais bens o indivíduo pode possuir. Isso pode ser melhor visualizado na Tabela IV.

No geral, não obteve-se boas predições para a variável resposta renda familiar.

V. CONCLUSÃO

Os resultados obtidos indicam que, no geral, não se obteve boas predições para as variáveis respostas sexo, etnia e renda dos entrevistados. As covariáveis serem todas categóricas e distinguirem-se entre nominais e ordinais, dicotômicas e politômicas, limitou uma melhor análise descritiva dos dados, como o estudo das correlações entre as variáveis, e o uso de métodos mais sofisticados de Aprendizado Supervisionado de Máquina, já que muitos exigem suposições de normalidade e independência das variáveis respostas. Além disso, acredita-se que as melhores acurácias para a variável resposta sexo em relação às demais seja devido à essa ser dicotômica e não politômica.

Apesar dos baixos valores de acurácia, notou-se que para a faixa de renda, os erros de predição ocorreram entre categorias próximas.

[1]O modelo ajustado para predição foi construído com parâmetro de complexidade igual a 0.015, mas para melhor visualização, a imagem representa uma árvore com parâmetro igual a 0.05.

Em relação à importância das variáveis para predição, uma das principais covariáveis para as predições tanto de sexo, quanto etnia e renda, foi o traço latente, o que faz sentido já que o mesmo resume informações presentes em 36 outras variáveis. Para a variável resposta sexo, a categoria do entrevistado na população economicamente ativa mostrou-se uma das mais importantes. Para as variáveis respostas etnia e renda, a cidade/estado de nascimento e as variáveis referentes à quantidade, em faixas (0 - Não tenho, 1 - Tenho um, 2 - Tenho dois ou mais), de pertences como eletrodomésticos, cômodos em casa, etc, mostram-se, respectivamente, com grande importância.

Para um próximo projeto, sugere-se a realização de predições com métodos multivariados e que sejam retiradas as covariáveis relacionadas à condições socioeconômicas para uma melhor análise da relação do sentimento "medo" com as variáveis respostas a serem preditas.

REFERENCES

ANDRADE, D. F. de; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. *ABE, Sao Paulo*, 2000. Citado na página page.22.

DATAFOLHA Instituto de Pesquisas. Medo dos Paulistanos, 2008. (Banco de dados). São Paulo, Datafolha, 2008. In: Consórcio de Informações Sociais, 2012. ⟨www.cis.org.br⟩. Acessado: 2018-11-24. Citado na página page.11.

ESTADÃO Homicídios na cidade de São Paulo têm alta de 42%, a maior desde 2012. ⟨https://sao-paulo.estadao.com.br/noticias/geral, homicidios-sobem-na-capital-paulista-pela-1-vez-em-2018-crimes-caem-no-estado, 70002368430⟩. Acessado: 2018-11-24. Citado na página page.11.

G1 Homicídios na cidade de São Paulo têm alta de 42%, a maior desde 2012. ⟨https://g1.globo.com/sp/sao-paulo/noticia/2018/10/19/ mortes-em-acidentes-de-moto-aumentam-62-em-sp-diz-infosiga.ghtml⟩. Acessado: 2018-11-24. Citado na página page.11.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado na página page.33.