USO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINAS PARA AVALIAÇÃO DA PROPENSÃO AO USO DO TRANSPORTE PÚBLICO

CLÁUDIA CRISTINA BAPTISTA RAMOS NAIZER
Universidade de São Paulo
ccbrnaizer@gmail.com

Marcela Navarro Pianucci
Universidade de São Paulo
manavarropg@gmail.com

Cira Souza Pitombo
Universidade de São Paulo
cirapitombo@usp.br

USO DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINAS PARA AVALIAÇÃO DA PROPENSÃO AO USO DO TRANSPORTE PÚBLICO

C.C.B.R.Naizer, M.N.Pianucci e C. S. Pitombo

RESUMO

Os principais modelos de previsão de demanda por viagens possuem uma abordagem paramétrica, a qual associa, a partir de uma função matemática calibrada, variáveis independentes a variáveis relativas à demanda por viagens. No entanto, esses modelos apresentam suposições matemáticas e restrições relativas à sua aplicação. Técnicas de Aprendizagem de Máquinas (AM) são "livres" de tais suposições, mostrando-se adequadas também para aplicação a problemas relativos à demanda por transportes. O presente trabalho tem como objetivo comparar o uso de dois algoritmos de AM (Redes Neurais Artificiais e Árvores de Decisão) para um problema de investigação da propensão à mudança modal (transporte individual motorizado ou ônibus – sistema metroviário) a partir de dados de Preferência Revelada e Declarada. Através dos resultados, comprovou-se a adequação de ambos os algoritmos para análise de problemas de escolha modal. Além disso, observou-se uma dependência espacial na região. Pessoas residentes em bairros mais carentes estão mais propensas à mudança do modo de transporte com objetivo de aumentar a qualidade dos seus deslocamentos.

1. INTRODUÇÃO

Demanda por viagens é fortemente relacionada a características individuais, dos domicílios de residência, das viagens, do meio urbano e do sistema de transporte (Kitamura *et al.*, 1997; Ortúzar e Willumsen, 2011). Tradicionalmente, os principais modelos de previsão de demanda por viagens possuem uma abordagem paramétrica agregada ou desagregada.

Modelos agregados tradicionais utilizam variáveis explicativas agregadas por unidades de áreas (população, empregos, custo médio de viagem, etc.), associando-as, através de função matemática calibrada, à quantidade de viagens produzidas por zona de tráfego, quantidade de viagens entre zonas de tráfego ou municípios, percentual de preferência modal por distrito, etc. No entanto, tal abordagem desconsidera que as escolhas relativas às viagens são feitas individualmente, podendo levar em conta características individuais e domiciliares, além daquelas agregadas, usualmente investigadas.

A análise desagregada mais tradicional de demanda por viagens baseia-se nos modelos de escolha discreta (Fotheringham, 1983; Ben - Akiva e Lerman, 1985). Tais modelos estimam os parâmetros que compõem as funções utilidades aleatórias das alternativas a partir da maximização da função verossimilhança. No entanto, eles implicam limitações relacionadas às suposições matemáticas. A principal limitação está relacionada à Independência das Alternativas Irrelevantes (IIA). O atributo IIA envolve a restrição de que os termos de erro aleatório são independentes e igualmente distribuídos (Koppelman e Wen, 2000), seguindo uma distribuição de *Gumbel*.

Tais suposições não fazem parte dos algoritmos de Aprendizagem de Máquinas (AM), os quais são técnicas semi-paramétricas ou não paramétricas que identificam padrões e classificam indivíduos. Este conjunto de algoritmos traz uma abordagem livre de suposições para a área de modelagem de demanda por transportes, já que não há limitações

de tipo de variáveis de entrada, multicolinearidade, distribuição de probabilidade, etc. A Figura 1 ilustra a justificativa para o uso de algoritmos de AM na área de modelagem de demanda por transportes.

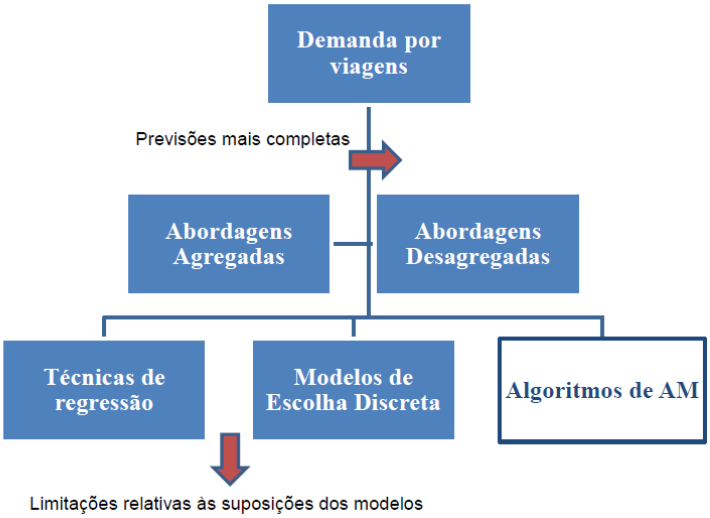


Fig. 1: Aplicação de algoritmos de AM em problemas de escolha modal (Adaptado de Pitombo *et al.*, 2017)

Dessa forma, a literatura recente apresenta a aplicação de algoritmos de Aprendizagem de Máquinas em problemas de demanda por transportes, tais como análise de viagens baseadas em atividades (Pitombo *et al.*, 2011; Ichikawa *et al.*, 2002; Arentze e Timmermans, 2007), escolha modal (Senbil *et al.*, 2005; Lu e Kawamura, 2010; Rasouli e Timmermans, 2014; Linder *et al.*, 2017) e distribuição de viagens (Pitombo et al., 2017, Rasouli e Nikras, 2013; Mozolin *et al.*, 2015). Tais autores exploraram o potencial de técnicas diversas, como Árvores de Decisão e Classificação, Redes Neurais Artificiais, Aprendizado por Regras de Associação, Máquinas de Vetores de Suporte, Algoritmos Genéticos, etc.

O presente artigo tem o objetivo de comparar o uso de algoritmos de AM, tais como Árvore de Decisão e Classificação (CART – *Classification And Regression Tree*) e Redes Neurais Artificiais (RNA) para um problema de investigação da propensão à mudança modal (transporte individual motorizado ou ônibus – sistema metroviário) a partir de dados de Preferência Revelada e Declarada. Para validação metodológica, é proposta calibração de modelo logit binomial para avaliação do desempenho dos algoritmos de AM.

A principal lacuna, e consequente contribuição deste trabalho, está focada na análise da escolha "não observada", mas sim na possível escolha frente diferentes cenários propostos na pesquisa realizada. A pesquisa foi realizada na cidade de Salvador, Bahia, Brasil, em 2017 e foi avaliada a propensão ao uso do metrô a partir de atributos como tarifa, tempo em descolamento e distância de caminhada até a estação.

2 ALGORITMOS DE APRENDIZAGEM DE MÁQUINAS (AM)

Aprendizagem de Máquinas é o campo que se dedica ao uso de ferramentas computacionais, aptas a assimilar e modificar conhecimentos e habilidades, aprimorando seus desempenhos baseando-se em experiências bem sucedidas anteriormente (Mitchell, 1997; Alpaydin, 2010). Explora a análise e construção de algoritmos que podem aprender a partir de dados de entrada amostrais e fazer previsões relativas a tais dados.

Os algoritmos de AM podem ser classificados em três grupos: aprendizado supervisionado,

aprendizado não supervisionado e aprendizado por reforço. A aprendizagem supervisionada se dá por meio do fornecimento de um conjunto de exemplos com a saída conhecida para cada conjunto de dados de entrada apresentado, de modo que o sistema de aprendizado aprenda uma regra geral por meio do mapeamento dos dados de entrada com os de saída (Russel *et al.*, 2002). Os algoritmos utilizados neste trabalho fazem parte da aprendizagem supervisionada: Árvore de Decisão (Quinlan, 1993), Redes Neurais Artificiais (Haykin, 1999).

2.1 CART: *Classification And Regression Tree*

Algoritmos de Árvore de Decisão fazem parte das ferramentas não paramétricas de AM. Possuem caráter preditivo para variáveis dependentes contínuas ou discretas ou classificatório para variáveis dependentes categóricas. São geradas regras de decisões, segregando o banco de dados inicial (nó raiz) em diversos grupos (nós filhos), até que os mesmos, a partir das regras de parada, não sofram mais nenhuma divisão (nós terminais ou folhas) (Breiman *et al.*, 1984). Os principais algoritmos são C4.5 (Quinlan, 1993), CHAID (Kass, 1980) e CART (Breiman *et al.*, 1984). Neste trabalho optou-se pela aplicação do algoritmo CART.

O algoritmo CART (*Classification And Regression Tree*) baseia-se na realização de divisões binárias do conjunto de dados. A partição é realizada de forma a maximizar a pureza ou homogeneidade dos nós filhos. O critério de divisão do conjunto de dados é verificado quando a seleção de determinada variável independente atinge maior *aprimoramento* (ou melhor aumento de homogeneidade dos nós filhos). Inicialmente, o nó raiz apresenta grau de impureza máximo por ser o nó no qual está contido o conjunto de dados completo. As categorias dentro deste nó serão definidas de acordo com a variável dependente estabelecida para o problema. Dessa forma, considerando que a variável dependente apresente n (1,2,3,i,...,n) categorias, a probabilidade da categoria i aparecer no nó raiz, por exemplo, definido como nó inicial 0, será p(i/0). Cabe ressaltar que a soma das probabilidades de todas as categorias da variável dependente em um dado nó é equivalente a 1. Após sucessivas divisões, irá ocorrer a diminuição da medida de impureza dentro dos nós filhos gerados. A máxima homogeneidade (Impureza=0) em um nó t será alcançada quando um nó contiver uma única categoria com 100% do conjunto de dados (p(i/t)=1). O cálculo mais comum da heterogeneidade dos dados do algoritmo CART é realizado pelo Índice Gini, dado pela Equação 1.

$$G(t) = 1 - \sum_{i=1}^{n} p^2(i/t) \qquad (1)$$

A diferença entre o Índice Gini para o nó pai e a soma dos valores para os nós filhos, ponderados pela proporção de casos em cada filho, é apresentada na árvore como *aprimoramento*. A escolha da melhor variável explicativa e melhor valor de corte se dá pela combinação (de variável e valor de corte) que produz maior valor de *aprimoramento*. Neste trabalho, a aplicação do algoritmo CART ocorreu como um problema de classificação, já que a variável dependente é categórica, com 2 categorias associadas (0- modo atual (automóvel ou ônibus); 1 – sistema metroviário). As variáveis explicativas, provenientes do banco de dados, são descritas na seção de Materiais e Método.

2.2 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNA) conseguem reproduzir o comportamento de qualquer função matemática, inclusive não-lineares (Smith, 1996). Sua estrutura de funcionamento é baseada em sistemas de equações, em que o resultado de uma equação é o valor de entrada para várias outras da rede. As RNA são sistemas de computação adaptativos, inspirados

nas características de processamento de informação encontradas nos neurônios reais e nas características de suas interconexões, com o objetivo de resolver problemas reais. O neurônio forma a base para as RNA, sendo o neurônio artificial o objeto que simula o comportamento do neurônio biológico, uma unidade de processamento matematicamente simples. Este recebe uma ou mais entradas, que correspondem às conexões sinápticas com outras unidades similares a ele, com seus respectivos pesos, transformando em saídas, cujos valores dependem diretamente da somatória ponderada de todas as saídas dos outros neurônios a esse conectado (Haykin, 1999).

A estrutura da rede neural vai depender do algoritmo de aprendizado usado para treinar a rede, pois os neurônios podem estar dispostos de diversas formas, De acordo com Haykin (1999) a arquitetura das redes se apresenta em três diferentes tipos: as redes progressivas de única camada, as redes progressivas de camadas múltiplas e as redes recorrentes. As redes progressivas de camadas múltiplas, *Multi-Layer Perceptron* (MLP), corresponde a um processador paralelo, constituído de neurônios (unidades de processamento) que são dispostos em uma ou mais camadas interligadas por muitas conexões. O aprendizado da rede MLP é denominado de treinamento e ocorre através do ajuste dos pesos. O aprendizado comumente é realizado utilizando algum algoritmo de treinamento, como por exemplo, o algoritmo *backpropagation* uma técnica de aprendizado supervisionado que utiliza pares (entrada e saída desejada) para, através do cálculo do erro, ajustar os pesos da rede e adquirir conhecimento (Haykin, 1999).

3 MATERIAIS E MÉTODO

A cidade de Salvador é a capital do estado da Bahia, localizado na região nordeste do Brasil. Para o ano de 2017, a população do município corresponde a 2.675.656 de habitantes, sendo o terceiro município mais populoso do país. Para o mesmo ano, a arrecadação foi de R$ 5.345.811.000,00 e o PIB per capita de R$ 19.812,07 (IBGE,2017). Devido ao crescimento da cidade de Salvador, na década de 90 decidiu-se pela construção de duas linhas de metrô para suprir as carências de mobilidade da cidade. O processo de licitação para a construção das linhas iniciou-se em 1997, mas devido a inúmeros atrasos, o primeiro trecho da Linha 1 foi terminado somente em 2014. Atualmente a Linha 1 opera em oito estações, entre Lapa e Pirajá, enquanto a Linha 2 opera em onze estações, entre Acesso Norte e Mussurunga (CCR METRÔ BAHIA, 2018).

Foi realizada uma pesquisa de Preferência Declarada e Revelada, com o objetivo de compreender quais eram os principais fatores que poderiam incentivar os moradores de Salvador a realizar a troca modal (automóvel, motocicleta ou ônibus) para o modo metroviário, em cenários hipotéticos, onde se variavam distância até a estação, tarifa e tempo de viagem. Os cenários utilizados são representados na Tabela 1. Além dessas informações, foram coletadas informações que caracterizassem o usuário e a viagem realizada pelo mesmo atualmente (parte da Pesquisa relativa à Preferência Revelada). Na Tabela 2, em seguida, são apresentadas as variáveis e as classes utilizadas. Os aplicativos utilizados foram: *SPSS 24.0 – IBM* e *Excel 2007 – Microsoft*.

Tabela 1: Cenários Hipotéticos da Pesquisa de Preferência Declarada

Cenário	Tempo de Viagem	Tarifa	Distância de caminhada ao ponto
1	Maior que 15 min	Menor que R$ 3,30	Maior que 400 m
2	Maior que 15 min	Menor que R$ 3,30	Menor que 400 m
3	Maior que 15 min	Maior que R$ 3,30	Menor que 400 m
4	Menor que 15 min	Maior que R$ 3,30	Menor que 400 m
5	Menor que 15 min	Maior que R$ 3,30	Maior que 400 m
6	Menor que 15 min	Menor que R$ 3,30	Maior que 400 m

Tabela 2: Variáveis do banco de dados

Variável	Classes	Tipo	Variável	Classes	Tipo
Idade	Até 18 anos	Ordinal	Modo atual	A pé ou bicicleta	Nominal
	19 a 25 anos			Automóvel ou motocicleta	
	26 a 40 anos			Ônibus convencional	
	41 a 60 anos			Outros	
	60 ou mais		Tempo de viagem atual	Até 10 minutos	Ordinal
Escolaridade	Superior Incompleto	Nominal		De 10 a 20 minutos	
	Superior Completo			De 20 a 30 minutos	
Automóveis na residência	Nenhum	Ordinal		De 30 a 40 minutos	
	1			Acima de 40 minutos	
	2		Escolha modal no Cenário Hipotético	Modo usado atualmente	Nominal
	3			Metrô	
	4 ou mais		Bairro de origem	163 bairros	Nominal
Posse de Motocicleta	Não	Nominal	Bairro de destino	163 bairros	Nominal
	Sim				

3.1. Método

O método utilizado neste trabalho corresponde ao fluxograma apresentado na Figura 2. As etapas do procedimento metodológico são sumariadas em seguida. Vale ressaltar que a amostra baseou-se em cada um dos cenários hipotéticos e que 70% da amostra foi selecionado aleatoriamente para calibração e treinamento, enquanto 30% restante foi selecionado para validação ou teste. As comparações entre as ferramentas basearam-se nas amostras de validação/teste.

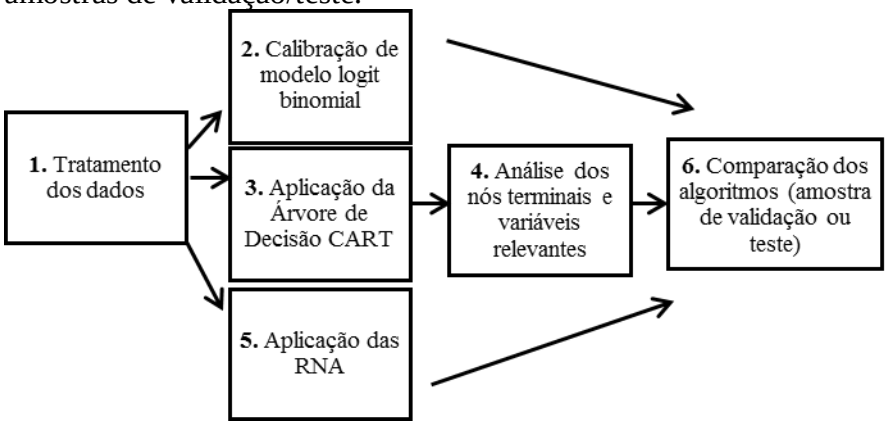


Fig. 2 Fluxograma metodológico

3.1.1. Tratamento dos dados

A pesquisa foi respondida por 64 usuários. Na etapa de tratamento dos dados, optou-se por exclusão do cenário 3 – Tabela 1 - (delimitado na etapa de planejamento de experimentos), considerando a baixa variabilidade das respostas. O banco de dados de usuários foi transformado em um banco de dados por respostas/cenários, sendo cinco para cada usuário. Dessa forma, o banco de dados final possui 320 elementos, cada linha possui as características dos usuários, da viagem atual e a resposta do mesmo sobre a troca modal em cada cenário hipotético proposto. A variável dependente do banco de dados é a resposta a respeito da troca modal (0-Não; 1 – Sim). As variáveis independentes são aquelas descritas na Tabela 2. Qualitativas Ordinais ou Nominais.

3.1.2 Calibração do modelo logit binomial

Após tratamento dos dados e divisão da amostra inicial, é calibrado o modelo logit binomial com 70% da amostra inicial. A regressão logística (*logit*) permite o uso de um modelo (curva em S) para prever a probabilidade π de um evento categórico. A modelagem da curva S é dada por uma transformação logística da probabilidade π, conforme a

Equação 2, da função logística g(x). A partir dela, derivam-se as equações de calibração (Equação 3, Equação 4). A qualidade do ajuste pode ser mensurada por medidas estatísticas apropriadas, tais como: testes de regressão de Cox & Snell e Nagelkerke (Hair *et al.*, 2010).

$$g(x) = \ln\left(\frac{\pi}{1-\pi}\right) \quad (2)$$

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (3)$$

$$\pi = \frac{1}{1+e^{-g(x)}} \quad (4)$$

Para x_0=1 e $\beta 0, \beta 1, \ldots, \beta n$: os coeficientes da equação de calibração e *n* o número de variáveis independentes. Π = probabilidade para o valor 1 da variável binária

Neste trabalho, as variáveis independentes, utilizadas para calibração do modelo logit binomial, estão representadas na Tabela 2, enquanto a variável dependente caracteriza a propensão à mudança do transporte público e é binária.

3.1.3 Aplicação da Árvore de Decisão CART

Aplicou-se o algoritmo CART ao banco de dados utilizando as variáveis dependente e independentes, mencionadas anteriormente, e as seguintes personalizações: Optou-se por utilizar os critérios de crescimento de profundidade máxima de cinco níveis e número mínimos de casos para "nós pais" de 20 e para "nós filhos" de 10; A homogeneidade em cada nó foi calculada baseada no índice Gini; Foi utilizado o valor de 0,0001 relativo ao *aprimoramento* mínimo.

3.1.4. Análise dos nós terminais e variáveis relevantes

Após a aplicação da Árvore de Decisão, pode-se observar o número de nós resultantes e as variáveis relevantes para as divisões dos ramos da árvore. As variáveis relevantes, assim como o valor de corte de cada uma delas, são selecionados baseados nos valores de *aprimoramento*, conforme mencionado anteriormente. A partir da seleção das variáveis é possível mensurar a sua importância normalizada para segregação do banco de dados em classes homogêneas. Podem-se comparar também as variáveis independentes selecionadas pelo modelo CART aos parâmetros associados às variáveis independentes para calibração do logit binomial e à importância associada às variáveis independentes para RNA.

3.1.5. Aplicação das RNA

Aplicou-se o algoritmo de RNA ao banco de dados utilizando as seguintes personalizações. Para a modelagem de RNA existem algumas opções de redes, porém as redes mais usuais e, disponíveis no *software* IBM SPSS 24.0, são as redes *perceptrons* multicamadas (MLP) e as redes de base radial. A rede MLP é função de variáveis de previsão (variáveis independentes) que minimizam o erro de predição da variável de saída. É composta por uma camada de entrada (variáveis independentes), em que as informações são recebidas; por nenhuma, uma ou mais camadas ocultas e uma camada de saída. A camada de saída fornece a solução do problema.

Utilizou-se a MLP e, em seguida, foi escolhido o tipo de arquitetura da rede pelo método de tentativa e erro. Existem três pontos importantes na concepção da arquitetura de uma RNA, sendo eles, o número de camadas escondidas, o número de neurônios nas camadas escondidas e a função de ativação. Os dois primeiros pontos determinam a complexidade do modelo neural. O *Backpropagation* é o algoritmo padrão do *software* IBM SPSS 24.0, além de ser normalmente o mais utilizado no treinamento das redes MLP. Esse algoritmo é do tipo supervisionado e utiliza pares de valores (entradas, saídas desejadas) que, através

de correções de erros, ajusta os pesos da rede. A seleção da melhor rede foi baseada essencialmente na habilidade para previsões corretas para os dados de validação.

3.1.6. Comparação dos algoritmos

Para comparação dos algoritmos de AM e validação metodológica, foram utilizados os resultados provenientes da amostra com 30% dos elementos. Assim, considerando valores estimados pelo algoritmo CART, pelas RNA, pelo modelo *logit* e valores observados, são realizados testes de hipóteses para testar as similaridades de distribuições entre pares de valores (observados x CART, observados x RNA e observados x *logit*) – *Kolmogorov-Smirnov*, além das similaridades de medidas típicas (teste da mediana). Outra forma de comparação de desempenho dos algoritmos seria o percentual de acertos a partir da amostra de validação.

4. RESULTADOS E DISCUSSÕES

4.1 Modelo *Logit* Binomial

Na calibração do modelo *logit*, foram considerados significativos os parâmetros associados às variáveis bairros da Origem (Bairro_O_ID) e tempo de viagem atual (Tempo_ID). O *pseudo* R^2 apresentou valores baixos, indicando baixa acurácia do presente modelo (Cox & Snell = 0,104; Nagelkerke = 0,140). A partir do modelo obtido pela amostra de calibração, observou um percentual de acertos equivalente a 61,6%. A partir da amostra de validação, observa-se um percentual de acertos equivalente a 51%. A Figura 3 apresenta os principais resultados relativos ao modelo *logit*.

Variáveis na equação

		B	S.E.	Wald	df	Sig.	Exp(B)
Etapa 1[a]	Bairro_O_ID	,022	,011	3,836	1	,050	1,022
	Tempo_ID	,530	,117	20,390	1	,000	1,698
	Constante	-1,977	,528	14,029	1	,000	,138

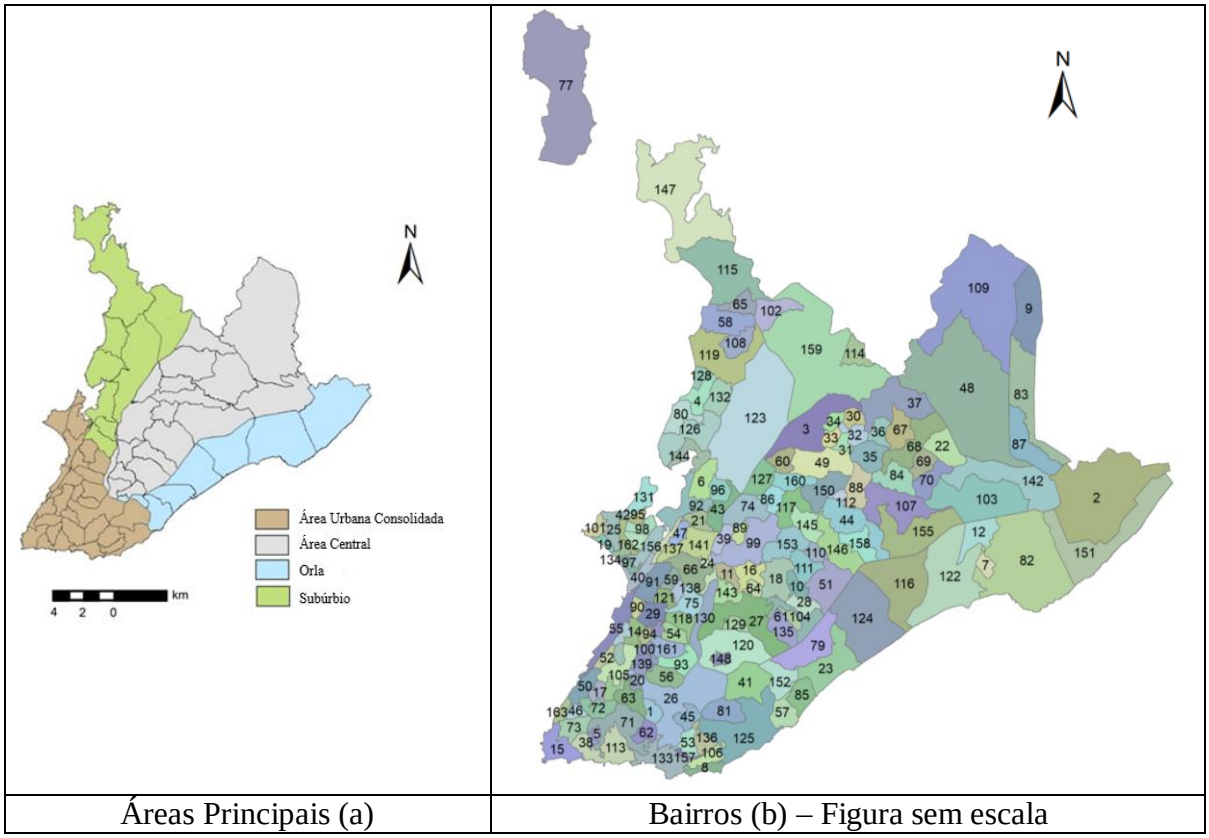
a. Variável(is) inserida(s) na etapa 1: Bairro_O_ID, Tempo_ID.

Fig. 3 Resultados na calibraçao da regressão logística

Observa-se a importância do bairro de origem. Para Salvador, os números mais altos de identificadores dos bairros estão associados àqueles bairros de baixa renda. Nota-se, então, pela importância e sinal dos parâmetros estimados, que residentes em bairros de renda mais baixa (valores altos dos números identificadores) estão mais propensos à mudança modal (Ônibus – Metrô). Além disso, indivíduos com tempos de viagens maiores, também seguem a mesma tendência de comportamento de escolha. Uma explicação para maior tendência ao uso pela população de mais baixa renda está no projeto da rede, que inicialmente foi delineado para atender aos principais corredores de transporte público da cidade. A Figura 4b ilustra a distribuição dos bairros em salvador, com seus respectivos identificadores, enquanto que a Figura 4a ilustra as regiões da cidade. A região Central e Suburbana são aquelas compostas com maior número de bairros de renda mais baixa.

4.2 Árvore de Decisão

A partir da amostra de treinamento, das variáveis mencionadas anteriormente e das regras de paradas adotadas pelos autores, foram obtidos 18 nós filhos, 10 nós terminais e profundidade de 4 níveis. A divisão dos dados foi baseada, principalmente, em três variáveis: tempo de viagem atual (Tempo_ID), bairro de origem (Bairro_O_ID) e distância de caminhada à estação no cenário hipotético (C_Dist), conforme Figura 5. Observa-se que, para a amostra de treinamento (224 observações), 58% estão propensos a utilizar a rede metroviária, enquanto 42% não estão propensos em determinados cenários.



Áreas Principais (a)	Bairros (b) – Figura sem escala

Fig. 4 Regiões Principais da cidade de Salvador (a) e identificadores de bairros (b). Fonte: Rocha *et al.* (2017), CONDER e Prefeitura Municipal de Salvador.

A primeira divisão da árvore é realizada através da variável tempo de viagem atual, dividida em viagens inferiores a 30 minutos e superiores a 30 minutos. O segundo e terceiro nível da divisão dizem respeito aos bairros de origem das viagens. No quarto nível de decisão observa-se novamente a variável bairro de origem e a variável distância de caminhada até a estação nos cenários hipotéticos, podendo ser menor (0) ou maior (1) do que 400 m.

De modo geral, observa-se que em bairros de baixa renda e tempo de viagem superior a 30 minutos, a adesão ao metrô é maior do que em bairros de alta renda e menor tempo de viagem. Além disso, para bairros de alta renda observa-se que a distância da residência à estação de metrô é um fator importante na decisão do modo utilizado. Sendo assim, viagens mais curtas acontecem nos bairros de alta renda, enquanto viagens mais longas são uma característica mais frequente nos bairros de baixa renda (Figura 4 e Figura 5). Pessoas residentes em bairros de baixa renda estão mais suscetíveis a trocar o modo atual (no caso ônibus) pelo sistema metroviário em implantação (nó 2, 5, 6, 11, 12, 13 e 14). Já para aquelas pessoas residentes em bairro de renda média ou alta (nó 4, 9, 10, 15, 16, 17 e 18) estão menos propensas a trocar o modo atual (predominantemente automóvel) pelo sistema metroviário. Um atributo que pode ocasionar a mudança de comportamento daqueles classificados como renda média ou alta é a distância de caminhada à estação (nó 15 e 16). A Figura 6, em seguida, apresenta a importância normalizada das variáveis independentes, sendo o bairro de residência dos indivíduos predominantemente importante no comportamento relativo à escolha modal, seguido do tempo de viagem. As duas variáveis mais importantes na análise da árvore também foram aquelas com coeficientes significativos na regressão logística. As análises atuais apontam uma associação espacial forte relacionada à propensão ao uso do sistema metroviário de Salvador. O algoritmo

CART apresentou um alto poder preditivo, com percentual de acertos equivalente a 83,5% na amostra de treinamento. A amostra de teste apresentou um percentual de acertos equivalente a 63%.

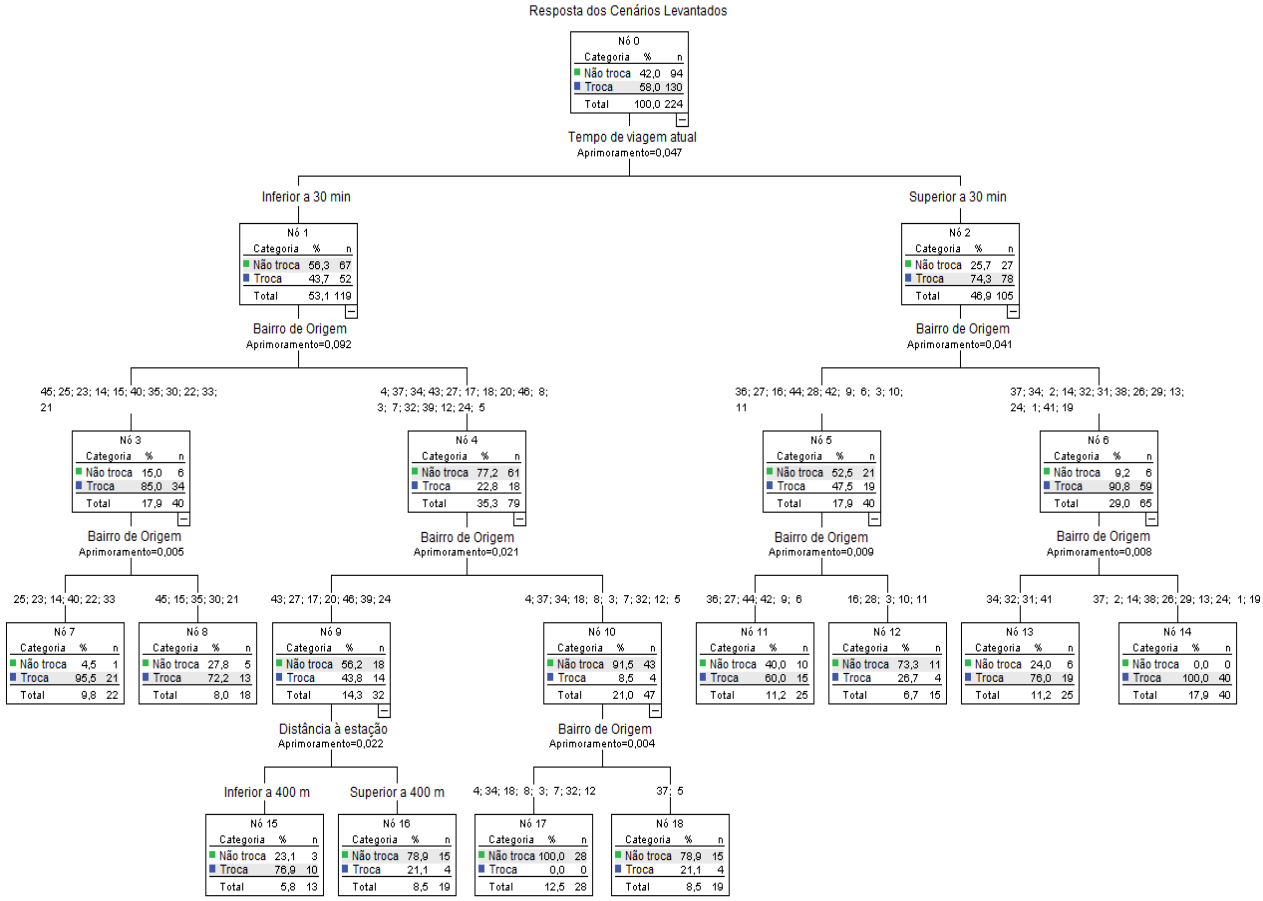


Fig. 5 CART (amostra de treinamento) obtida para análise de propensão à mudança para modo de transporte metroviário

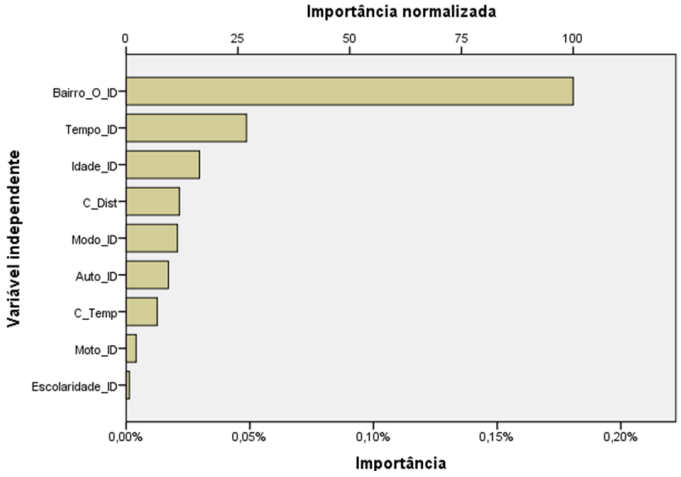


Fig. 6 Importância normalizada das variáveis independentes - CART

4.3 Redes Neurais

As redes neurais foram utilizadas para identificar os principais fatores que incentivam a mudança do transporte privado motorizado ou ônibus para o metrô. O processo de treinamento da rede foi realizado 68 vezes de forma a realizar todas as possíveis combinações do IBM SPSS 24.0. Após as 68 combinações concluídas, foram analisados

os resultados e escolhida a rede que apresentou a menor porcentagem de previsões incorretas. Optou-se por utilizar uma rede de arquitetura personalizada, composta por 2 camadas ocultas e com função de erro, a entropia cruzada, utilizada para minimizar o erro.

Similarmente ao algoritmo CART, a rede definiu como principal variável preditora, que incentiva a mudança do modo atual para o metrô, a variável Tempo de viagem (Tempo_ID), seguida da variável Bairro de origem (Bairro_O_ID). Essas mesmas variáveis foram àquelas associadas a parâmetros estimados significativos no modelo de regressão logística. A Figura 7 apresenta a importância normalizada das variáveis independentes a partir do algoritmo de RNA. Os resultados obtidos, através da ferramenta de RNA para a amostra de treinamento e teste, foram de 84,4% e 70,8% de previsões corretas, respectivamente. Observa-se que o modelo de RNA permite selecionar um número maior de variáveis, o que pode explicar sua maior acurácia quando comparado às abordagens prévias.

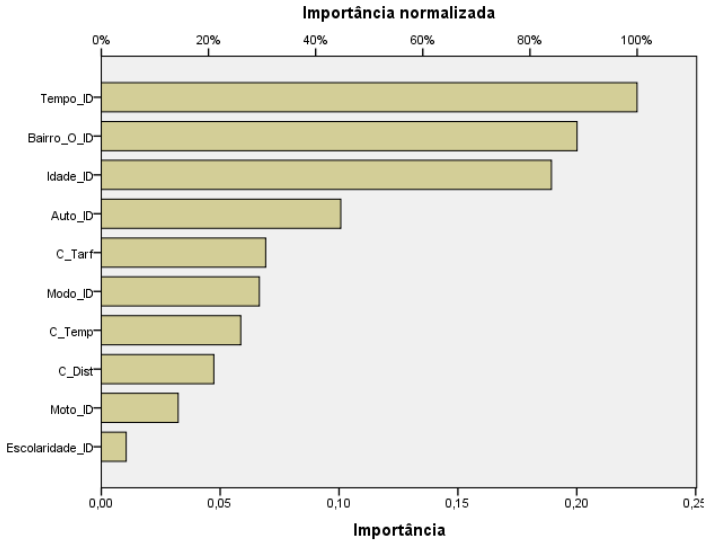


Fig. 7 Importância normalizada das variáveis independentes – RNA

4.3 Comparação dos algoritmos

Para comparação entre as abordagens, testes não paramétricos, como os testes estatísticos *Mann-Whitney* e *Kolmogorov-Smirnov*, foram realizados a fim de comparar as distribuições dos valores observados e estimados a partir dos três algoritmos. O teste da mediana foi realizado para testar se as amostras têm medidas centrais semelhantes. A Tabela 3 ilustra os resultados dos testes, além do percentual de acertos. Toda análise comparativa foi realizada com a amostra de validação. Com base nos resultados, pode-se afirmar que as RNA, seguidas pelo algoritmo CART, forneceram as estimativas mais precisas. Ambos os algoritmos de AM têm a mesma distribuição de probabilidade dos valores observados, bem como, as mesmas medidas centrais. No entanto, para o modelo *logit*, estas hipóteses foram rejeitadas.

Tabela 3 Comparação entre abordagens

	% acertos	Kolmogorov-Sminorv	Mann Witney	Mediana
Logit x obs	51%	Rejeita Ho*	Rejeita Ho*	Rejeita Ho*
CART x obs	63%	Retém Ho*	Retém Ho*	Retém Ho*
RNAs x obs	71%	Retém Ho*	Retém Ho*	Retém Ho*

Ho As distribuições dos valores são similares (testes de distribuições); As medianas dos valores são similares (teste da mediana).

5. CONCLUSÕES

O estudo comparativo das técnicas de Aprendizagem de Máquinas e da abordagem tradicional de escolha modal (*logit*) permitiu a comprovação de que a análise desagregada, a partir de modelos "livres" de restrições e suposições matemáticas, é eficaz na previsão de escolhas de modo de transporte, a partir de cenários hipotéticos. Os melhores resultados foram obtidos pela ferramenta de RNA, seguida do algoritmo CART. No entanto, ambos os algoritmos podem ser considerados adequados, especialmente quando há banco de dados multicolinear, com diferentes tipos de variáveis de entrada.

No estudo de caso observado, verificou-se que o bairro de residência, caracterizado implicitamente pela renda, bem como tempo de viagem em deslocamento, foram as variáveis consideradas mais importantes para mudança de comportamento individual, considerando as três abordagens. As análises mostraram uma dependência espacial em Salvador. Pessoas residentes em bairros mais carentes estão mais propensas à mudança do modo de transporte com objetivo e aumentar a qualidade dos seus deslocamentos.

AGRADECIMENTOS
Os Autores agradem às agências de fomento FAPESP e CNPq.

6 REFERÊNCIAS

Alpaydin, E (2010) Introduction to Machine Learning. MIT Press, 2a edição, 2010. ISBN10026201243X.

Arentze, T., Timmermans, H. (2007). Parametric action decision trees: incorporating continuous attribute variables into rule-based models of discrete choice. Transport. Res. B: Methodol. 41 (7), 772–783.

Ben-Akiva, M.E.; Lerman, S.R. (1985) Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge, MA.

Breiman, L.; Friedman, J.H; Olshen, R.A.; Stone, C.J. (1984) Classification and Regression Trees. Wadsworth International Group, Belmont, CA.

CCR Metrô Bahia (2018) Mapa das Linhas. Disponível em < http://www.ccrmetrobahia.com.br/linha-1/mapa-da-linha>
Acesso em: 15 de abril 2018.

Fotheringham, A.S. (1983) Some theoretical aspects of destination choice and their relevance to production-constrained gravity models. Environment and Planning A, v.15, n.8, p. 1121–1132. doi: 10.1068/a151121

Hair Jr, J. F.; Black, W. C.; Babin, B. J.; Anderson, R. E. (2010) Multivariate Data Analysis. Prentice Hall. 7a ed. 785 p.

Haykin, S. (1999) Neural networks - a comprehensive foundation. Ontario, Canada: Pretince Hall.

IBGE, Instituto Brasileiro de Geografia e Estatística (2017) Brasil em Síntese, Panorama Cidade de São Paulo. Disponível em < https://cidades.ibge.gov.br/brasil/ba/salvador/panorama > Acesso em: 15 de abril 2018.

Ichikawa, S.M., Pitombo, C.S., Kawamoto, E. (2002) Aplicação de Minerador de dados na obtenção de relações entre padrões de viagens encadeadas e características socioeconômicas. Anais do XVI do Congresso de Pesquisa e Ensino em Transportes,

Anpet, Natal (RN), vol2, p175-186.
Kass, G.V. (1980) An exploratory technique for investigating large quantities of categorical data. Applied Statistics, v.29, p. 119–127. doi: 10.2307/2986296

Kitamura, R., Mokhtarian, P.L., Laidet, L. (1997). A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. Transportation 24(2), 125–158.

Koppelman, F. S.; Wen, C.H. (2000) The paired combinatorial logit model: properties, estimation and application. Transportation Research Part B: Methodological, v. 34, n. 2, p. 75-89. doi: 10.1016/S0191-2615(99)00012-0

Lindner, A. ; Pitombo, C. S. ; Cunha, A. L. (2017) Estimating motorized travel mode choice using classifiers: An application for high-dimensional multicollinear data. Travel Behaviour and Society, v. 6, p. 100-109, 2017.

Lu, Y., Kawamura, K. (2010). Data-mining approach to work trip mode choice analysis in Chicago, Illinois, area. Transportation Research Record: Journal of the Transportation Research Board 2156, 73–80

Mitchell, Tom M. (1997) Machine Learning. McGraw-Hill. ISBN 0070428077

Mozolin, M.; Thill, J.C.; Linn, U.E. (2015) Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. Transportation Research Part B: Methodological, v. 34, p.53-73. doi: 10.1016/S0191-2615(99)00014-4

Ortúzar, J. D.; Willumsen, L. G. (2011) Modelling Transport. Wiley, London.

Pitombo, C.S.; Kawamoto, E.; Sousa, A.J. (2011) An exploratory analysis of relationships between socioeconomic, land use, activity participation variables and travel patterns. Transport Policy, v. 18, 347-357. doi: 10.1016/j.tranpol.2010.10.010

Pitombo, C. S.; De Souza, A.D.; Lindner, A. (2017) Comparing decision tree algorithms to estimate intercity trip distribution. Transportation Research Part C 77, p. 16-32

Quinlan, R. (1993) Learning efficient classification procedures and their application to chess end-games. Machine Learning: An Artificial Intelligence Approach, Tioga, Palo Alto, pp. 463-482.

Rocha, S. S. ; Lindner, A. ; Pitombo, C. S. (2017). Proposal of a geostatistical procedure for transportation planning field. Boletim de ciências geodésicas, v. 23, p. 636-653, 2017.

Rasouli, M.; Nikraz, H. (2013) Trip Distribution Modelling Using Neural Network. Transport Research Forum, Brisbane, Australia.

Rasouli, S., Timmermans, H.J.P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: effects on predictive success and uncertainty estimates. EJTIR 14 (4), 412–424.

Russel, S.J.; Norvig (2002) P. Artificial Intelligence: A Modern Approach. 2aedição. Prentice-Hall, 2002. ISBN10 0137903952

Senbil, M., Fujiwara, A., Zhang, J., Asri, D.U. (2005). Development of a choice model for evaluating sustainable urban form. In: Proc. Eastern Asia Soc. Transport. Stud., pp. 2164–2178

Smith, M. (1996) Neural Networks for Statistical Modeling. International Thomson Computer Press, Londres, Inglaterra.