Analise de pesquisa de opinão com usuários do transporte coletivo utilizando árvores de decisão.

André Fernandes Santana
PPGMNE - UFPR
andre.fsmi@gmail.com

Dr. Eloy Kaviski
Departamento de Hidráulica e Saneamento - UFPR

Dra. Liliana M. Gramani
Departamento de Matemática
UFPR

***Resumo* — Este trabalho analisa se o usuário do transporte coletivo de Curitiba tem informações suficientes para chegar ao seu destino. A pesquisa realizada no Terminal Boqueirão selecionou a linha mais congestionada da região na faixa de horário com início às 6h30min e término às 8h30min, ou seja, quando os passageiros estão apressados e tem pouco tempo para decidirem em qual linha embarcar.**

Palavras-chave* — *Transporte Coletivo; árvores de decisão; pesquisa de opinião; Weka.

I. INTRODUÇÃO

Curitiba é uma cidade modelo em transporte coletivo, reconhecida nacional e internacionalmente [8]. O município é referência no planejamento e fiscalização da mobilidade urbana [5]. Outro ponto explorado pela cidade é o de "*City Marketing*", um conceito estratégico na administração pública [3].

O presente artigo investiga se o usuário do transporte coletivo tem informações suficientes para chegar ao seu destino. A grande quantidade de conexões faz do complexo sistema curitibano um meio dinâmico, que quando bem utilizado traz benefícios, principalmente no quesito tempo de deslocamento [1].

O objetivo é verificar se os investimentos em mobilidade urbana, fiscalização e *City Marketing* refletem em usuários conscientes e aptos a explorarem o sistema de maneira segura.

II. ESTUDO DE CASO

O Terminal Boqueirão foi selecionado dentre os 21 terminais urbanos pertencentes à Rede Integrada de Transporte (RIT) [9], ou seja, o sistema de transporte coletivo de Curitiba.

O estudo foi realizado para analisar o fluxo de passageiros no sentido bairro centro e no horário do pico da manhã. No caso o horário das 6h30min até as 8h30min, isto é, quando os usuários estão apressados e tem pouco tempo para definirem suas rotas.

O trajeto investigado tem como origem o Terminal Boqueirão e os seguintes destinos: Terminal Carmo, Terminal Hauer e Centro, conforme representado na Fig. 1.

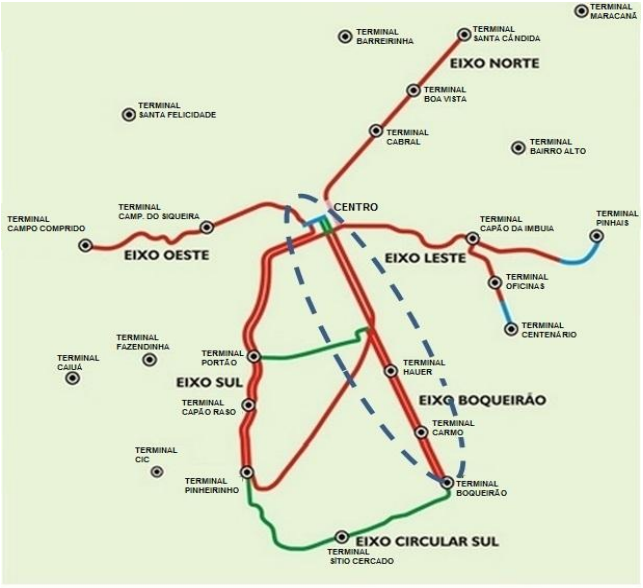

Fig. 1. Estrutura urbana da RIT, em destaque o eixo Boqueirão.
Fonte: URBS [9]

Para transitar no eixo Boqueirão o passageiro poderá optar pelas linhas listadas na Tabela I, não existindo um diferencial entre elas, pois todas possuem embarque em nível e nesta abordagem trafegam por vias exclusivas. Todas as outras linhas que circulam pelo eixo boqueirão e não possuem embarque em nível foram descartadas da análise.

TABELA I. OPÇÕES DE LINHAS PARA O DESLOCAMENTO

Cód. da linha	Nome da Linha	Categoria da Linha
206	Barreirinha/São José	Ligeirinho
500	Ligeirão Boqueirão	Ligeirão
503	Boqueirão	Expresso
505	Boqueirão/C. Cívico	Ligeirinho
508	Sitio Cercado (Anti-horário)	Ligeirinho
602	Circular Sul (Anti-horário)	Expresso

A referência [1] detalha as linhas listadas na Tabela I, inclusive com relatos históricos e peculiaridades das mesmas. Outras informações podem ser encontradas no *site* da URBS [9].

Para simplificar, o itinerário foi desmembrado em três segmentos (Tabela II) de mesma origem e destinos concentrados em apenas três pontos do eixo, os de maior fluxo. Fica evidente a quantidade de rotas disponíveis, por exemplo no segmento 1 no qual o usuário tem 6 opções de deslocamentos.

TABELA II. DESTINOS PROVÁVEIS DOS USUÁRIOS

Segmento	Destino	Opções de Linhas
1	T. Carmo	206 Barreirinha/São José
		500 Ligeirão Boqueirão
		503 Boqueirão
		505 Boqueirão/C. Cívico
		508 Sitio Cercado (Anti-horário)
		602 Circular Sul (Anti-horário)
2	T. Hauer	206 Barreirinha/São José
		500 Ligeirão Boqueirão
		503 Boqueirão
		505 Boqueirão/C. Cívico
		602 Circular Sul (Anti-horário)
3	Centro (UTFPR)	500 Ligeirão Boqueirão
		503 Boqueirão

A referência [1] constata que a linha Ligeirão Boqueirão é a que possui maior taxa de ocupação e consequentemente as maiores filas de embarque. Por esta razão foi selecionada para o estudo.

III. PESQUISA DE OPINIÃO

Na tentativa de entender os motivos de o Ligeirão Boqueirão aglomerar um maior número de usuários, realizou-se uma pesquisa de opinião com os passageiros desta linha sobre a origem e o destino dos mesmos.

A pesquisa de opinião tem como finalidade indagar a prioridade dos passageiros na linha Ligeirão Boqueirão e se os mesmos conseguem elaborar rotas alternativas. O LEA (Laboratório de Estatística Aplicada - UFPR) concedeu uma consultoria nesta etapa do trabalho, inclusive no processo de validação do questionário além de sugerirem formas eficientes para tabular os dados.

O questionário da Fig. 2 foi aplicado numa amostra de 77 pessoas nos dias 11, 12 e 13 de novembro de 2014. No decorrer dos dias, vários usuários abdicaram a responder o questionário, fazendo uma estimativa a cada cinco abordados apenas um estava disposto a participar.

Com a pesquisa concluída, os dados foram tabulados com o auxilio de uma planilha eletrônica do Excel.

Dentre as diversas técnicas de Mineração de Dados existentes — Análise de *Cluster*, Árvores de Decisão, Redes Neurais, Indução de Regras, Algoritmos Genéticos, Aprendizado Baseado em Casos —, optou-se por Árvores de Decisão [2].

A intenção na escolha das Árvores de Decisão consiste na exibição e analise dos resultados sem a utilização de um histograma.

Sexo (anote sem perguntar): ( )M ( )F
Faixa etária (anos):
( ) < 16 ( ) Entre 16 e 19 ( ) Entre 20 e 29 ( ) Entre 30 e 39
( ) Entre 40 e 49 ( ) Entre 50 e 59 ( ) Entre 60 e 65 ( ) > 65
1 – Com que frequência utiliza o transporte coletivo.
( ) Todos os dias ( ) Alguns dias ( ) Esporadicamente
2 – Qual a sua origem?
( ) São José dos Pinhais.
( ) Boqueirão.
( ) Sitio Cercado.
( ) Outra.
3 – Qual o seu destino?
( ) Terminal do Carmo.
( ) Terminal do Hauer.
( ) T.R.E.
( ) UTFPR.
( ) Praça Carlos Gomes.
4 – Conhece outras linhas para chegar ao seu destino?
( ) Sim. ( ) Não.
5 – Por que o Ligeirão Boqueirão?
( ) É mais rápido.
( ) O veiculo é mais bonito.
( ) Outra.
( ) Não existe outra opção.
6 – O que é mais importante:
( ) Ter um número maior de bancos para sentar.
( ) Ter corredores mais largos para facilitar a circulação.
7 – Você considera as condições de embarque:
( ) Adequadas ( ) Médias ( ) Ruins

Fig. 2. Questionário aplicado.

IV. ÁRVORES DE DECISÃO

Quinlan [4] é considerado o pai das árvores de decisão [2], contribuiu na elaboração de um algoritmo pioneiro chamado ID3 (1986).

Existem vários algoritmos de classificação que utilizam a representação sob o formato de Árvores de Decisão, além do já citado ID3, tem o C4.5, o CART (*Classification and Regression Trees*), o CHAID (*Chi Square Automatic Interaction Detection*), entre outros [7].

Árvore de decisão é uma maneira simples de classificar dados. Uma técnica eficiente para construir classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados [2].

No topo da árvore o atributo de maior valor é apresentado como o primeiro nó, e os atributos menos relevantes são exibidos nos nós seguintes. O diferencial das Árvores de Decisão está no fato do processo de tomada de decisões avaliarem os atributos mais relevantes, além de ser facilmente compreendido pela maioria das pessoas. Ao selecionar e ilustrar os atributos em ordem de importância, as Árvores de Decisão possibilitam identificar quais fatores mais se destacam na base de dados [6].

Uma árvore de decisão divide recursivamente um conjunto de treinamento, até que cada subconjunto seja de uma única classe. Para atingir o objetivo, o método analisa e verifica a distribuição de classes no processo de concepção da árvore.

Os dados, organizados de maneira compacta, são utilizados para classificar novos casos [2].

O método apresenta as seguintes vantagens: não adota uma distribuição particular para os dados; os atributos podem ser categóricos (qualitativos) ou numéricos (quantitativos); os modelos englobam qualquer função desde que o número de exemplos de treinamento seja suficiente; elevado grau de compreensão [6].

Estabelecida a Árvore de Decisão, é necessário verificar sua precisão utilizando dados que não tenham sido aplicados no treinamento. Desta forma é possível estimar como a árvore generaliza os dados e adapta-se a novas situações, além de determinar a proporção de erros e acertos [6].

V. Resultados

O *software* computacional WEKA (*Waikato Environment for Knowledge Analysis*) foi utilizado na confecção das Árvores de Decisão, devido sua praticidade e por ser um *software* de domínio público disponível em: http://www.cs.waikato.ac.nz/ml/weka/.

O formato do arquivo de entrada do WEKA pode ser uma extensão ARFF ou CSV. Neste artigo optou-se por CSV, pois é obtido de maneira direta por meio de uma planilha do Excel, *software* selecionado para armazenar os dados.

Para converter um arquivo XLS, é preciso salvar a planilha no tipo CSV (separado por vírgulas), na opção salvar como do Microsoft Excel. Um recurso simples que evita a edição completa dos dados. A Fig. 3 exibe uma parte da base de dados utilizada nesta pesquisa, já organizada com a extensão CSV.

```
Sexo,Idade,Quest01,Quest02,Quest03,Quest04,Quest05,Quest06,Quest07
M,Entre 40 e 49,Todos,SJP,Centro,Nao,Nao,Bancos,Med
F,Entre 20 e 29,Todos,BOQ,Centro,Nao,Nao,Bancos,Adeq
M,Entre 40 e 49,Todos,BOQ,Centro,Nao,Rapido,Corredor,Med
F,Entre 20 e 29,Todos,SJP,UTFPR,Nao,Nao,Bancos,Ruim
M,Entre 30 e 39,Todos,BOQ,TRE,SIM,Rapido,Corredor,Med
M,Entre 40 e 49,Alguns,BOQ,TH,SIM,Rapido,Corredor,Ruim
F,Entre 20 e 29,Todos,BOQ,TRE,Nao,Nao,Bancos,Adeq
F,Entre 30 e 39,Alguns,BOQ,UTFPR,Nao,Nao,Bancos,Med
M,Entre 20 e 29,Todos,SJP,TC,Nao,Nao,Corredor,Med
F,Entre 20 e 29,Todos,Outra,Centro,Nao,Nao,Bancos,Med
F,Entre 17 e 19,Todos,BOQ,Centro,SIM,Rapido,Bancos,Ruim
F,Entre 40 e 49,Todos,BOQ,TC,SIM,Rapido,Corredor,Adeq
F,Entre 20 e 29,Todos,SJP,UTFPR,Nao,Nao,Bancos,Adeq
M,Entre 20 e 29,Alguns,BOQ,UTFPR,SIM,Rapido,Corredor,Med
M,Entre 20 e 29,Todos,BOQ,UTFPR,Nao,Nao,Bancos,Med
M,Entre 50 e 59,Todos,Sitio,TH,Nao,Nao,Bancos,Adeq
M,Entre 17 e 19,Todos,Sitio,Centro,Nao,Nao,Corredor,Ruim
F,<= 16,Todos,BOQ,TC,SIM,Rapido,Bancos,Ruim
M,Entre 20 e 29,Todos,BOQ,UTFPR,SIM,Rapido,Corredor,Med
F,Entre 20 e 29,Todos,SJP,Centro,Nao,Nao,Bancos,Ruim
M,Entre 30 e 39,Todos,SJP,Centro,Nao,Nao,Bancos,Med
F,Entre 20 e 29,Todos,BOQ,UTFPR,SIM,Rapido,Corredor,Med
```

Fig. 3. Arquivo de entrada.

O arquivo é composto com as informações dos 77 usuários do transporte coletivo abordados na pesquisa de opinião. Após uma série exaustiva de testes foi empregado o algoritmo de classificação Random Tree. No processo de construção da árvore o algoritmo considera certa quantidade de atributos escolhidos aleatoriamente em cada nó e não executa nenhuma poda. Os outros algoritmos produzem uma quantidade menor de nós, pois realizam a poda automática da árvore.

A complexidade da mineração de dados consiste no fato que o pesquisador precisa ter um conhecimento prévio sobre sua base de dados, inclusive na escolha do algoritmo de classificação.

As configurações empregadas neste artigo podem ser visualizadas na Fig. 4. Como o objetivo da pesquisa é avaliar o grau de conhecimento que o usuário possui do sistema, o atributo Quest04 (questão 4) foi utilizado como referencia.

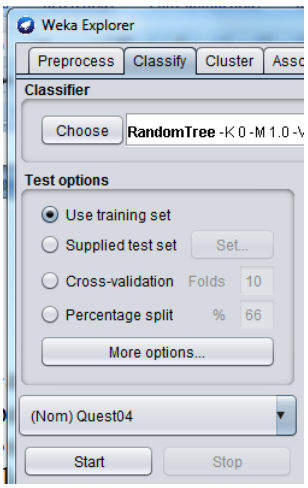

Fig. 4. Configuração do Weka.

A matriz de confusão (Tabela III) foi o indicador empregado para apurar a acurácia do método, obtendo uma taxa de 98,7% das instancias classificadas corretamente.

TABELA III. Matriz de Confusão

		Classificadas como	
		Não	**Sim**
Previstas	**Não**	47	1
	Sim	0	29

Para ilustrar um formato de resultado, na Fig. 5 pode-se observar uma árvore de decisão contendo apenas os atributos Quest04 e Quest05. Para esta configuração simples poderia ser utilizado o algoritmo J48 (C4.5).

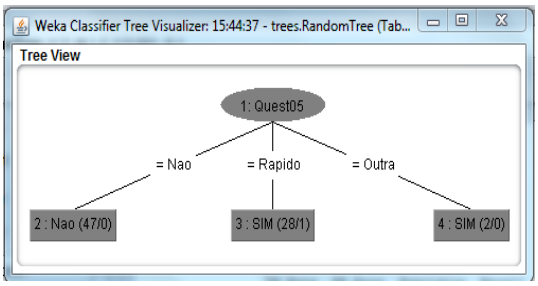

Fig. 5. Árvore de decisão, com 4 nós.

A vantagem do Random Tree está na utilização de um número maior de atributos. Por exemplo, caso a analise fosse um comparativo entre os usuários de diferentes faixas de idade ou verificar se determinada região da cidade é mais desinformada que outra. A visualização de uma árvore com todos os atributos foge do escopo deste trabalho. A finalidade é uma abordagem geral sobre o nível de conhecimento que os usuários possuem do sistema. Como critério comparativo a Tabela IV exibe a quantidade de nós que cada árvore de decisão possui, de acordo com a quantidade de atributos selecionados.

TABELA IV. TABELA COMPARATIVA

Algoritmo	Quantidade		Classificados corretamente
	Atributos	Nós	
J48	9	4	98,70%
Random Tree	9	43	98,70%

Da Fig. 5 tem-se que 47 usuários de um total de 77, desconhecem outra rota para chegarem ao seu destino, para eles existe apenas a linha Ligeirão Boqueirão. Conforme visto anteriormente, dependendo do trajeto escolhido o passageiro tem até 6 opções de linhas para o deslocamento.

Outra informação obtida da Fig. 5 é referente aos 28 usuários que conhecem outras rotas, porém optam pelo Ligeirão por ser a linha mais rápida do eixo Boqueirão.

VI. CONCLUSÃO

De acordo com os resultados obtidos da árvore de decisão pode-se constar que 61% dos passageiros entrevistados não possuem informações suficientes do sistema para planejarem uma rota eficiente. E aproximadamente 36% afirmam que ao embarcarem na linha Ligeirão chegarão mais rápido ao seu destino.

Segundo Santana [1], ambos os passageiros estão equivocados, já que existem opções de deslocamentos para a região que permitem ao usuário chegar ao seu destino no mesmo horário. Cabe ao passageiro explorar o sistema e buscar informações junto à URBS, para desta forma elaborar uma rota eficiente.

Curitiba prima por um sistema de transporte coletivo de qualidade, atualizando os itinerários e aumentando a quantidade de conexões, para que o passageiro tenha mais possibilidades para chegar ao seu destino. Entretanto o mesmo não está usufruindo de toda a rede, por comodidade utiliza sempre as mesmas linhas.

Fica como sugestão para trabalho futuro uma pesquisa que aponte o motivo do passageiro não querer explorar todas as alternativas do sistema.

Agradecimentos

Agradeço a CAPES e UFPR pela bolsa e estrutura de trabalho. Ao LEA, Laboratório de Estatística Aplicada pelo suporte na pesquisa de opinião. À URBS, por atenderem às minhas inúmeras solicitações. Todos aqueles que de alguma forma me ajudaram e incentivaram nesta longa caminhada.

Referências

[1] A. F. Santana. Simulação computacional da distribuição de usuários em sistemas de transporte coletivo via autômatos celulares. Dissertação (mestrado). Universidade Federal do Paraná – UFPR. 2015.

[2] E. P. Lemos. Análise de crédito bancário com o uso de Data Mining: redes neurais e árvores de decisão. Dissertação (mestrado) Administração, Universidade Federal do Paraná – UFPR. 2005.

[3] F. E. S. Garcia, O "City Marketing" de Curitiba diante das novas realidades mundiais, Anais: Encontros Nacionais da ANPUR, v. 6, 2013.

[4] J. C. Quinlan, C4.5: programs for machine learning. San Mateo: Morgan Kaufmann, 1993. 302p.

[5] J. L. Leal. Atuação dos agentes de fiscalização do transporte público e a qualidade do serviço – o caso de Curitiba. Revista dos Transportes Públicos – ANTP - Ano 38, 2015.

[6] M. H. Shiba et al. "Classificação de imagens de sensoriamento remoto pela aprendizagem por árvore de decisão: uma avaliação de desempenho." *Anais do XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia-GO* (2005): 4319-4326.

[7] S. C. Garcia, O uso de árvores de decisão na descoberta de conhecimento na área da saúde. Dissertação (mestrado) - Universidade Federal do Rio Grande do Sul, 2000.

[8] S. P. V. M. Branco. Estudo e aplicação de sistemas BRT – Bus Rapid Transit. Dissertação (mestrado) — Universidade do Porto, Novembro 2013.

[9] URBS. Urbanização de Curitiba – Rede Integrada de Transporte. 2016. Online e acesso em 29 de Julho de 2016. Disponível em: < http://www.urbs.curitiba.pr.gov.br/transporte/rede-integrada-de-transporte>.