UTILIZAÇÃO DE UM ALGORITMO DE CLASSIFICAÇÃO BASEADO EM TEOREMA DE BAYES PARA O ESTUDO SOBRE O CONSUMO DE ENERGIA ELÉTRICA

MARIANE COELHO AMARAL[1]; EDUARDO NUNES BORGES[2]; ANDERSON GARCIA SILVEIRA[3]; GRAÇALIZ PEREIRA DIMURO[4]

[1]Universidade Federal do Rio Grande– marianecamaral@gmail.com
[2]Universidade Federal do Rio Grande– eduardonborges@furg.br
[3]Instituto Federal Sul-rio-grandense– a.garcia.ee@gmail.com
[4]Universidade Federal do Rio Grande – gracaliz@gmail.com

1. INTRODUÇÃO

A economia dos países é movimentada por muitos setores e aspectos, dentre eles o consumo e geração de energia elétrica. Quando a demanda é superior à oferta, faltará eletricidade para os consumidores. Já quando a oferta é muito maior que a demanda por eletricidade, as empresas geradoras e distribuidoras de energia sofrem prejuízos.

A preocupação com o consumo de energia está cada vez mais presente na realidade brasileira através de políticas que incentivam a economia. A Pesquisa de posse de equipamentos e hábitos de uso (PPEHU), por exemplo, é uma pesquisa de campo realizada por questionários que tem como finalidade investigar, além da posse e utilização de equipamentos em nível nacional, algumas análises socioeconômicas e qualidade do fornecimento de energia. Essa pesquisa fornece embasamento para a elaboração de pesquisas mais simples, com menos questionamentos e menor complexidade nas questões.

As técnicas de mineração de dados consistem em descobrir conhecimentos em banco de dados, revelando informações que poderiam passar despercebidas por uma análise manual.

Assim, este trabalho mostra o início de um estudo sobre o consumo de energia elétrica utilizando um algoritmo classificador baseado no Teorema de Bayes. É objetivo deste trabalho iniciar os estudos sobre variáveis envolvidas no consumo de energia para futuros estudos mais complexos, passando pela elaboração de um instrumento, análise e formatação dos dados, montagem da base de dados e, finalmente, a etapa de mineração e interpretação dos resultados.

2. METODOLOGIA

Baseado nisso, foi aplicado um instrumento científico de maneira *online* entre estudantes de graduação da Universidade Federal do Rio Grande (FURG) e da Universidade Federal de Pelotas (UFPel), seguindo o procedimento de validação e constructo e circularidade do método científico, como exposto por RUTTER; SERTÓRIO (1994).

A mineração de dados consiste no processo de descoberta conhecimentos de interesse em banco de dados. Esse processo se divide em três passos: exploração - onde estão incluídas todas as pré-transformações necessárias, como limpeza, seleção e integração dos dados - construção do modelo e validação do modelo. Submete-se, então, o conjunto de treinamento a uma das técnicas de classificação, onde o modelo será construído e validado no conjunto de teste.

Os classificadores baseados no Teorema de Bayes rotulam as instâncias de um conjunto de dados com a classe que maximiza a probabilidade *a posteriori* calculada. Assim, o classificador Naïve Bayes é dito ingênuo porque supõe que os atributos são independentes e, ainda, é capaz de classificar um elemento de acordo com a probabilidade de este pertencer a uma classe que foi previamente determinada.

Este trabalho utiliza a plataforma livre *Waikato Enviroment for Knowledge Analysis (*WEKA*)*, que foi desenvolvida na Universidade de Waikato, para testar o desempenho do algoritmo Naïve Bayes na classificação de instâncias novas a partir de classes já existentes.

Foram consideradas e submetidas à validação de constructo nove variáveis que apontariam uma relação com o consumo de energia, sendo elas: escolaridade da mãe e do pai do respondente, renda familiar, material de construção da residência, valor da conta de energia, número de residentes, número de cômodos, bairro de localização da residência e se há aquecimento a gás no domicílio.

Após recolhimento e formatação das respostas do instrumento científico, foi dado início à formatação dos dados, considerando como variável principal o valor da conta de energia elétrica paga no último mês e considerando as respostas, aplicou-se o filtro *discretize* para dividir em três classes os respondentes, de acordo com a variável de estudo do trabalho: valor gasto em consumo de energia.

Os primeiros testes foram realizados com 100 respostas. Com isso, a primeira classe compreendeu os indivíduos que pagam menos de 100 reais, a segunda classe englobando os indivíduos que pagam entre 100 e 200 reais e, por fim, a última classe com os respondentes que pagam mais de 200 reais. As classes foram nomeadas de acordo com a Tabela 1.

Tabela 1: Divisão em três classes de respondentes

Classe	Limites de valores, em reais
A	(0,100]
B	(100,200]
C	Mais de 200

Fonte: elaborada pela autora

Será utilizada a validação cruzada, que se baseia em o algoritmo selecionar uma das partições (*folds*) para teste, enquanto as outras partições são utilizadas para treinamento, a fim de garantir uma amostragem aleatória, conforme abordado por KIRKBY (2004).

Para avaliação do modelo gerado, são consideradas as métricas de desempenho de acurácia geral do modelo, precisão e revocação, através da avaliação da taxa de verdadeiros positivos e falsos positivos.

A Acurácia diz respeito à porcentagem de acerto geral do classificador. Para seu cálculo, basta dividir quantidade de instâncias classificadas corretamente pela quantidade total. A Precisão identifica qual a porção de classificações positivas estão corretas. Essa medida é dada pela razão entre a frequência de verdadeiros positivos e todos os positivos (verdadeiros e falsos). Já a revocação (*recall*) identica qual a porção das instâncias que realmente são positivas foram identificadas corretamente.

A Estatística Kappa (K) é uma maneira de medir a concordância nos algoritmos e, para CASTRO (2010), mede a qualidade dos modelos de

classificação gerados e também o quanto a resposta se afasta do que era esperado, desconsiderando o acaso.

O número Kappa pode variar de 0 até 1 e quanto mais próximo de 0 ele for, maior será o nível de discordância, tendo sua interpretação de acordo com a classificação proposta no trabalho de LANDIS; KOCH (1977), onde, se o valor de K é, não há concordância entre os dados. Se o valor for entre 0,01 e 0,20, a concordância é considerada leve. Já se o valor de K estiver entre 0,41 e 0,60, a avaliação é moderada. Ainda, se o valor do coeficiente estiver entre 0,61 e 0,80, a concordância é boa ou substancial. Por fim, estes autores consideram concordâncias quase perfeitas os valores de K entre 0,81 e 1.

3. RESULTADOS E DISCUSSÃO

Finalizadas as etapas de pré-processamento, deu-se início às etapas de mineração de dados. Este trabalho mostra os resultados do primeiro algoritmo testado para essa base de dados, considerando a divisão dos respondentes em três classes e considerando a variável de investigação sendo o valor da conta de energia pago mensalmente, foi o classificador Naïve Bayes, implementado no *WEKA* como *NaiveBayes*.

A Tabela 2 representa a matriz de confusão gerada. Na diagonal principal, pode ser observado o número de instâncias classificadas corretamente que corresponde à soma algébrica dos termos. Se quisermos, portanto, confirmar a porcentagem de acerto do algoritmo, dividimos a soma dos elementos da diagonal principal pelo número total das instâncias consideradas pelo algoritmo, multiplicando-se por 100.

Tabela 2: Matriz de Confusão para 3 Classes

A	B	C	← Classificado como
53	8	2	A= (0,100]
9	**26**	2	B= (100, 200]
1	2	**7**	C= (200, +inf)

Fonte: elaborada pela autora

Da matriz de confusão, percebe-se que a maioria das instâncias de cada classe foi classificada corretamente. Percebe-se, ainda, que a maior quantidade de instâncias classificadas incorretamente ocorreu para as classes imediatamente ao lado, como seria o mais desejado ao ocorrer um erro de classificação.

O algoritmo também fornece as métricas de desempenho apresentadas neste trabalho, onde podem facilmente ser comprovados pelas definições correspondentes. A Tabela 3 mostra as métricas para o NaiveBayes quando os respondentes foram agrupados em três classes.

Tabela 3: Métricas de desempenho

Classe	FP rate	FP rate	Precision	*Recall*
A	0,841	0,213	0,841	0,841
B	0,703	0,137	0,722	0,703
C	0,700	0,040	0,636	0,700

Fonte: elaborada pela autora

Nota-se que, em todos as classes, a taxa de verdadeiros positivos foi de 0,700 ou mais, chegando até 0,841 na casse A. Nota-se também que a precisão e o *recall* apresentaram valores que podem chegar até 0,841. A acurácia geral do modelo, fornecida pelo registro do WEKA, é de 78,18%, contando com 86 instâncias classificadas corretamente e 24 instâncias classificadas de maneira incorreta.

O coeficiente Kappa, fornecido pelo software, foi de 0,6053, o que evidencia uma concordância substancial ou boa entre os dados, conforme escala definida pelo autor já mencionado.

4. CONCLUSÕES

Este trabalho parte da aplicação de um questionário em estudantes de cursos de graduação da Universidade Federal do Rio Grande e da Universidade Federal de Pelotas, com o objetivo principal de montar uma base de dados inicial para a aplicação de uma técnica de mineração de dados, denominada de classificação e dar início à análise do valor da conta de energia elétrica no domicílio de estudantes com as mesmas características dos entrevistados. Foi testado o algoritmo de classificação conhecido como Naïve Bayes, que é baseado no Teorema de Bayes.

Tendo em vista o resultado preliminar da classificação através do classificador utilizado neste trabalho, pode-se dizer que há uma evidência de que a classificação para maiores bases e dados seja ainda mais satisfatória.

Como trabalho futuro, espera-se a expansão da base de dados, para tornar-se uma amostra mais significativa. Também, espera-se testar outros classificadores e outras técnicas, a fim de chegar-se a resultados ainda melhores.

5. REFERÊNCIAS BIBLIOGRÁFICAS

CASTRO, D. Procedimentos de data mining na definio de valores para as análises de multicritérios como apoio a tomada de decisões e análises espaciais urbanas. In: XXIV Congresso Brasileiro de Cartogra. Aracaju, 2010.

KIRKBY, R. WEKA Explorer User Guide for Version 3- 4-3. University of Waikato, 2004. Disponvel em: http://weka.sourceforge.net/manuals/ExplorerGuide.pdf. Acesso em: 20 de junho de 2018.

LANDIS, R.; KOCH, G. The measurement of observer agreement for categorical data. Biometrics.Journal of International Biometric Society, v.33, p.159-174, 1977.

RUTTER, M.; SERTORIO, A. A. Pesquisa de Mercado. 2. ed. São Paulo: Editora Atica S. A., 1994.