Identificação de *outlier* no índice de aprovação no exame de direção da CNH

GABRIEL LIMA GOMES[1], LETICIA T. M. ZOBY[2]

[1]Graduando em Ciência da Computação, Bolsista PIBIC, Centro Universitário IESB, Brasília-DF, gabriel.lg08@gmail.com.
[2]Doutora em Engenharia de Sistemas Eletrônicos e de Automação, Centro Universitário IESB, Brasília-DF, letmaia@gmail.com

Área de conhecimento (Tabela CNPq): Banco de Dados 1.03.03.03-0

Apresentado no
7° Congresso de Iniciação Científica e Tecnológica do IFSP
29 de novembro a 02 de dezembro de 2016 - Matão-SP, Brasil

RESUMO: Para auxiliar na gestão de grande quantidade de dados, existem diferentes recursos computacionais para transformar dados em informações, o KDD (*Knowledge Discovery in Databases* - Descoberta de conhecimento em base de dados), é uma dessas metodologias existentes. Este trabalho descreve uma das etapas do KDD, a mineração de dados, através dela é possível extrair informações úteis para identificar possíveis atos ilícitos na aplicação do exame prático de direção para obtenção da carteira nacional de habilitação (CNH). Para aplicação deste trabalho, foi utilizado *outlier*, uma das tarefas da mineração de dados, como foco do trabalho, aplicando métodos estatísticos de forma automatizada para identificar valores atípicos na aprovação de candidatos neste exame. Com os resultados obtidos é possível observar que há índices de aprovações acima do normal, assim, estes casos merecendo uma análise mais aprofundada utilizando outras tarefas e técnicas de mineração de dados para identificar alguns padrões e confirmar se estes casos identificados são suspeitos de fraudes na aplicação do exame de direção do DETRAN.

PALAVRAS-CHAVE: CNH; KDD; MINERAÇÃO DE DADOS; *OUTLIER.*

Outiler identification in approval level in the exam direction of DETRAN

ABSTRACT: To support in manage of big data, there are different computational task to transform data in information, the KDD (Knowledge Discovery in Databases), are exist method. This paper describe one of the steps the KDD, the data mining, it possible extract useful information to identify possible illegal acts in exam direction to obtain of the national driving license. To apply this paper, was used outlier, a task of data mining, how focus of the paper, applying automatic statistics method to identify outlier values in approval of candidates. With the obtained results is possible notice there are approval level above normal else these cases merit a deeper analysis using others data mining task and techniques to identify some rules and to confirm if these case are suspect of frauds in the exam direction of DETRAN

KEYWORDS: DRIVING LICENSE; KDD; DATA MINING; OUTLIER.

INTRODUÇÃO

Devido à grande quantidade de dados coletados durante anos, a utilização do recurso de descoberta de conhecimento em base de dados tem ajudado na seleção de informações úteis para as companhias (HEKIMA, 2014), seja no marketing e propaganda, tomada de decisões entre outros.

Na área de segurança não é diferente, as empresas e órgãos públicos vêem investindo cada vez mais em métodos informatizados para identificação de fraudes, a fim de diminuir seus prejuízos.

Existe um grande esforço dos departamentos estaduais de trânsitos-DETRAN do Brasil para manter os seus sistemas robustos e seguros de modo a evitar fraudes nos processos para a emissão da carteira nacional de habilitação-CNH. Frequentemente a ousadia dos golpistas são manchetes de jornal, como em 24 de fevereiro de 2016 foi descoberta uma quadrilha que cobrava até R$ 6 mil para fraudar exames práticos do DETRAN/DF, colocando outras pessoas para realizarem a prova no lugar dos candidatos (STACCIARINI et al, 2016).

O processo para obtenção da CNH visa a capacitar o candidato com aulas e exame teórico e aulas e exame de direção veicular, assim, impedindo que pessoas não aptas conduzam veículos resguardando sua vida e a do próximo. O DETRAN tem despendido grande esforço e custo elevado para evitar que as fraudes aconteçam, isso porque o órgão vem adotando medidas preventivas com o objetivo de diminuir as probabilidades de uma ameaça se concretizar, contratando cada vez mais tecnologias de ponta para combate dessas ameaças, como a obrigatoriedade da biometria (DETRAN/GO, 2014), monitoramento em vídeo nos exames prático, dentre outras.

Assim, este trabalho propõe identificar valores atípicos de aprovação no exame prático utilizando KDD, utilizando a tarefa de mineração de dados, o *outlier*, e técnicas existentes nesta tarefa.

MATERIAL E MÉTODOS

Descoberta de conhecimento em base de dados / Mineração de dados

O processo de KDD é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. Este processo pode ajudar e/ou facilitar na formação de estratégias, por exemplo, na área de marketing, no aumento de lucratividade, auxilio na tomada de decisões, na segurança entre outros. Possui no total 5 etapas, que são: seleção (escolha do conjunto de dados), pré-processamento (eliminação de dados redundantes e inconsistentes), transformação (formatação e armazenamento dos dados), mineração de dados e pós-processamento ou interpretação do resultado (avaliação dos resultados obtidos) (FAYYAD, 1996).

Mineração de Dados - DM (Data Mining) é uma etapa do processo do KDD, onde se aplica algoritmos específicos para extração dos modelos de forma automática ou semiautomática (FAYYAD, 1996). Esta etapa é de extrema importância e uma atividade legítima para o processo KDD, desde que se entenda como realizá-la corretamente (CIOS et al. 2007).

O DM possui diversas tarefas, sendo uma selecionada de acordo com o que o usuário deseja, algumas delas são: classificação, *cluster*, *outlier* e associação. Dentre essas tarefas citadas anteriormente, o *outlier*, foi escolhido para este trabalho, pois consiste identificar valores atípicos na base de dados aplicando métodos estatísticos de forma semi-automatizada (AMO, 2004).

Dentro os métodos estatísticos para identificação de *outlier* têm-se:

- *z-score:* define quanto o valor se afasta da média utilizando desvio padrão. O resultado pode ser positivo ou negativo, e qualquer valor z > 3 ou z < -3 são considerados *outlier*. Para calcular valor de z, é utilizada a equação 1 (MARTINS, 2009).

$$Z_i = \frac{x_i - \bar{x}}{S} \quad (1)$$

Onde,

Xi é o valor a ser analisado, $\bar{x}$ a média amostral e S o desvio padrão que é dado pela equação 2.

$$S = \sqrt{\frac{\sum_{n=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (2)$$

- *boxplot:* representação gráfica para identificar valores que estão abaixo do limite inferior ou acima do limite superior, são considerados *outlier* (HAN; KABER; PEI, 2012), conforme a **figura 1**.

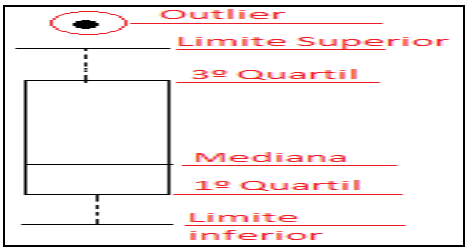

Figura 1. Exemplo de outlier com boxplot.

Aplicação

A base de dados utilizada para este trabalho foi fornecida por uma empresa prestadora de serviço para DETRAN chamada Search Tecnologia sediada em Brasília. Por motivos de segurança e sigilo, os dados identificatórios foram alterados, como também não será divulgado de qual DETRAN foram utilizados os dados. Foram aplicadas todas as etapas do processo KDD, citados anteriormente, mas será apresentada neste trabalho somente a etapa de mineração de dados.

A amostragem total da base de dados são de 312.919 e 75 colunas, dos anos 2010 à 2015. Após definir o objetivo a ser alcançado e realizando as etapas que antecedem a DM, o subconjunto de dados foi composto por 45.654 registros e 8 colunas. Em seguida foi realizada novamente as etapas que antecedem o DM houve uma redução que passou a ter 4 colunas e 640 registros no subconjunto.

A prova de exame de direção dos DETRAN é aplicada por uma banca de avaliadores, composta por 2 examinadores do DETRAN, escolhidos aleatoriamente dentre todos os escalados para aplicar o exame, e o resultado individual de cada examinador resultará na aprovação ou reprovação final do candidato.

Para análise de *outiler*, serão analisados os índices de aprovação por banca de avaliação.

Após finalizar a preparação do conjunto de dados a ser utilizado, a mineração de dados foi aplicada utilizando o software RStudio. Como dito anteriormente a tarefa escolhida para a realização deste trabalho foi a de detecção de *Outlier*, e utilizando as técnicas *z-score e boxplot*, já exemplificadas anteriormente.

RESULTADOS E DISCUSSÃO

Após aplicação do método *z-score* os resultados de Z variaram entre **-1.66 e 2.70**, ou seja, de acordo com a definição apresentada na seção anterior, não houve nenhum *outlier*,

Com aplicação do método *boxplot* é possível observar na figura 2 que foram identificado dois valores atípicos no índice de aprovação pela banca de avaliação.

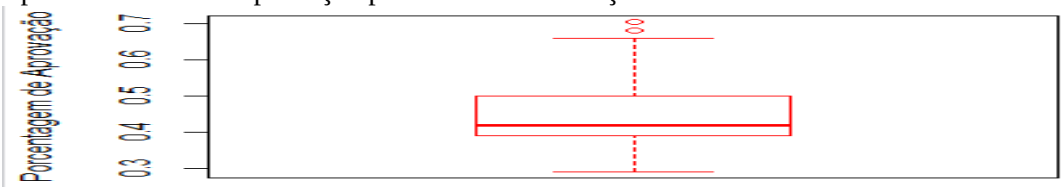

Figura 2. Resulado do box-plot aplicado na base de dados

De acordo com a definição apresentado na seção anterior, foram identificadas 2 bancas com índice atípico de aprovação. Após a identificação dessas duas bancas, foi realizada uma consulta no subconjunto de dados para verificar quais são essas bancas. Na tabela 1 é mostrada a identificação (alterada) dos examinadores que fazem parte de cada banca, a quantidade de exames, quantidade de exames aprovados e porcentual de aprovação desses dois casos identificados anteriormente.

TABELA 1. Descrição dos casos identificados como *outlier* no gráfico boxplot.

Código Examinadores	QTD_EXAMES	QTD_EXAM_AP	INDICE DE APROVAÇÃO
Banca 1: EX1520-EX1392	201	137	68,15%
Banca 2: EX1011-EX0465	237	167	70,46%

CONCLUSÕES

Ter informações tornou-se uma necessidade e um ponto importante para as organizações, o que pode fazer a diferença quando se trata de competitividade no mercado e/ou para segurança das organizações. O aumento cada vez mais rápido do volume de dados armazenados pode ocultar valiosas informações para alguns processos de tomada de decisão. Uma solução para este problema é aplicar o processo de KDD.

Foi demonstrado neste trabalho uma das etapas que compõem o processo KDD proposto para atender o objetivo deste trabalho, identificar valores atípicos no índice de aprovação no exame de direção veicular do DETRAN.

Após aplicação da tarefa de *outlier* e das suas técnicas de mineração de dados, é possível observar que existem casos suspeitos em ralação ao índice de aprovação. Com estes resultados é possível auxiliar as autoridades na tomada de decisão, assim, fornecendo informações úteis e concretas para abertura de auditoria, investigação ou qualquer outro procedimento estabelecido.

O processo de KDD é um poderoso método que pode auxiliar as organizações em diversas áreas, mesmo aparentando ser um processo fácil de ser realizado é um método complexo principalmente em bases de dados grandes, havendo erros no processo inicial pode prejudicar o restante do processo e gerando informações erradas, o que pode acarretar em diversas consequências, principalmente em grandes companhias.

Após a identificação destes casos, para trabalhos futuros, é possível realizar uma abordagem mais aprofundada sobre eles, utilizando outras tarefas e técnicas de mineração de dados, assim, tentando identificar alguns padrões de comportamento dessas bancas avaliadoras, então, podendo identificar possíveis casos de fraudes na aplicação de exame de direção por esses examinadores e se há mais alguma entidade e/ou pessoas envolvidas.

AGRADECIMENTOS

Os autores gostariam de agradecer ao Centro Universitário IESB pelo projeto de Iniciação Científica e ao CNPq pelo suporte financeiro com a bolsa CNPq/FUNTEL.

REFERÊNCIAS

CIOS, KRZYSZTOR J.; PEDRYCZ, WITOLD.; SWINIARSKI, ROMAN W.; KURGAN, LUKASZ A.; Data Mining, A Knowledge Discovery Approach. New York: Springer. 2007.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE. California, v. 17, n.3, p.37-54, nov. 1996.

HEKIMA. Por que a mineração de dados é essencial para as empresas que querem se destacar? - 2014. Disponível em: <http://bigdatabusiness.com.br/por-que-a-mineracao-de-dados-e-essencial-para-as-empresas-que-querem-se-destacar>. Acessao em: 01 agosto 2015.

MARTINS, GILBERTO DE ANDRADE. Estatística geral e aplicada. 3 ed. São Paulo. Atlas, 2009.

SOUTO, LIZIANE PRISCILA MARQUES. 2012. Aplicação de Mineração de dados para Levantamento de Perfis: Estudo de caso em uma Instituição de Ensino Superior Privada (FACISA - Faculdade de Ciências Sociais e Aplicadas). Encontro Nacional de Educação, Ciência e Tecnologia/UEPB.

STACCIARINI, ISA; SARAIVA, JACQUELINE; CARDIM, MARIA E.; Quadrilha cobrava até R$ 6 mil para fraudar exames do Detran-DF. 2016. Disponível em: < http://www.correiobraziliense.com.br/app/noticia/cidades/2016/02/24/interna_cidadesdf,519065/quadrilha-cobrava-ate-r-6-mil-para-fraudar-exames-do-detran-df.shtml>. Acesso em: 30 fev 2016.

Amo, Sandra de. **Técnicas de Mineração de Dados**. Universidade Federal de Uberlândia. XXIV CSBC. Salvador, 6 ago. 2004. Disponível em:<http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Acessado em: 01 nov. 2015.

DENTRAN/GO. **CFCs e clínicas terão que se adequar para biometria** - 2014. Disponível em: <http://www.detran.goias.gov.br/post/ver/180883/cfcs-e-clinicas-terao-que-se-adequar-para-biometria>. Acessado em: 25 nov. 2015.

Han, J.; Kamber, M; Pei, J. **Data Mining Concepts and Techniques**. p.703. USA: Elsevier. 2012.