T2 - Tratamento de dados, viéses e privacidade

MO810A/MC959A

Marco Antonio J. Ticona
RA:241132
Mestrado em Ciência da Computação

Luís Carlos M. A. Júnior
RA:272572
Mestrado em Ciência da Computação

Mariana Ap. Ferreira
RA: 183670
Mestrado em Ciência da Computação

I. Introdução

O objetivo desse relatório é analisar e descrever as variáveis utilizadas em problemas relacionados com o tema escolhido para trabalhar durante a disciplina. Apenas relembrando, nosso objetivo é trabalhar com a detecção de anomalias congênitas em recém nascidos a partir de um banco de dados disponível *online* no site do DATASUS, do Ministério da Saúde [1]. Na nossa aplicação não faremos distinção entre os tipos de anomalias [2], porém o que é encontrado majoritariamente nas buscas bibliográficas são aplicações focando em um (ou alguns) tipos específicos, mas vamos tentar ficar o mais próximo possível das variáveis que trabalharemos.

II. Discussão sobre aplicações com outras bases de dados

O primeiro artigo a ser comentado é uma comparação entre bases de dados clínica e administrativa e suas limitações para o tratamento de pacientes com anomalia congênita cardíaca [4]. As bases administrativas avaliadas no artigo possuem informações voltadas para investigações dos custos e da qualidade do serviço do hospital em que o paciente esteve presente, como por exemplo:

1) Identificação do visitante e do hospital;
2) Informações cadastrais, como idade, gênero, raça e endereço;
3) Datas de admissão e saída no hospital;
4) Código do diagnóstico e dos procedimentos realizados;
5) Status vital na saída (vivo ou morto);

O autor do texto diz que as vantagens em utilizar esses dados na análise é que são fáceis de conseguir, podem abranger diversos hospitais e é possível avaliar doenças raras, já que cobre todos os atendimentos feitos pelos serviços de saúde que originam os dados. Já as desvantagens são que os códigos dos diagnósticos e/ou procedimentos podem não contemplar todas as novas doenças pela falta de atualização, além de poder ocorrer de o código ser preenchido de maneira errada (procedimentos similares para doenças opostas, por exemplo). Outro ponto negativo é que certos diagnósticos podem ser ”diminuídos” por outros, como por exemplo um paciente que acabou de descobrir que tem diabetes e depois pedra nos rins (não garantindo uma ordem de causa-efeito entre as doenças).

No contexto desse *dataset*, se assemelha bastante ao problema que temos pois também trabalharemos com dados sensíveis (como idade e gênero), que pode trazer um aumento no *bias* do modelo. Sobre os erros nos preenchimentos do diagnósticos e/ou falta de códigos para abranger novas doenças, como não faremos distinção entre as anomalias estamos **um pouco** menos propensos à identificar anomalias que na verdade podem não ser, porém **muito** suscetíveis à erros de digitação, já que nossas informações são digitadas por um humano. Já sobre privacidade, acredito que já fizeram um bom trabalho em transformar os identificadores dos pacientes e hospitais em códigos, porém precisaria avaliar o restante das variáveis sensíveis para que nenhum grupo possa ser identificado em sua totalidade.

Já os dados clínicos são majoritariamente coletados por pessoas da área da saúde, que possuem mais contexto sobre a doença avaliada em relação à áreas administrativas. As variáveis coletadas aqui incluem histórico de saúde do paciente, comorbidades, resultados de exames clínicos e outras observações que sejam necessárias. Em um ponto de vista médico, as variáveis podem ser mais ”acuradas” do que os dados mencionados acima, mas também estão suscetíveis aos mesmos problemas na coleta e transcrição das informações. Um ponto bem importante que o autor cita no texto é que a participação do paciente nessas tabelas é voluntária, ou seja, podemos acabar caindo em vieses amostrais já que não é uma representação fiel do todo. No nosso trabalho não utilizaremos dados clínicos, já que estamos avaliando principalmente informações pontuais e cadastrais da mãe e do pai e não o histórico clínico do pré-natal (por exemplo).

O segundo artigo que vamos comentar é sobre os resultados de procedimentos intervencionistas comuns em doenças cardíacas congênitas [5]. Nele, foi utilizado um *database* que concentra dados sobre os pacientes a partir de diversas fontes em um único lugar, com um programa de qualidade que garante completude e consistência: **The IMPACT Registry** [5]. Os pacientes que estão presente no *database* tiveram que previamente concordar com o compartilhamento das informações. Algumas das informações disponíveis na tabela não:

1) Variáveis demográficas como nome, endereço, CEP, documento de identidade, sexo, raça, idade e etnia;
2) Informações sobre plano de saúde;
3) Códigos de identificação do diagnóstico (doenças cardíacas pré e pós atendimento);
4) Informações clínicas como quantidade de procedimentos anteriores (como catéter) e data dos procedimentos;

5) Doenças pré-existentes e diversas informações clínicas gerais

Não chegamos a acessar os dados, para avaliar se de fato todas as colunas possuem valores não nulos, porém as informações acima foram tiradas do dicionário de dados do próprio site que o artigo cita e é possível notar logo de cara problemas gravíssimos de privacidade. O paciente, por mais que tenha concordado com a participação, pode ser completamente identificado (tem até documento de identidade!). Nos dados que vamos trabalhar ao longo do curso, não conseguimos identificar integralmente nenhuma das pessoas que aparecem nas informações (mãe, pai e criança). Esses dados deveriam ser excluídos totalmente do banco de dados. Também é possível observar que temos a presença das mesmas variáveis sensíveis: idade, sexo, raça, etc. Da mesma forma que foi comentado anteriormente, elas deveriam ser transformadas para ajudar na construção de privacidade dos dados e avaliadas para não gerar um *bias* desnecessário no estudo.

O último artigo avaliado utiliza o mesmo banco de dados que o nosso, o do DATASUS [1]. O objetivo do estudo é encontrar padrões que tenham influência no óbito do recém-nascido [3] na cidade de Santa Maria, no Rio Grande do Sul. O tratamento aplicado nos dados foi a exclusão de variáveis com dados em branco ou nulos e identificadores e um certo agrupamento de variáveis similares. Após o pré-processamento, os autores ainda tinham variáveis sensíveis como sexo e raça do recém-nascido, idade e escolaridade da mãe e o código do estabelecimento de saúde onde ocorreu o nascimento. Não fica claro no texto se os autores usaram alguma transformação para agregar mais privacidade às variáveis finais, principalmente na idade da mãe e nem se avaliaram os possíveis vieses do modelo resultante nas variáveis de raça, sexo e escolaridade.

Como é o mesmo banco de dados que utilizaremos, devemos estar cientes de que essas variáveis devem ser tradadas para garantir uma maior privacidade dos dados, além de garantir que não exista nenhuma variável identificadora, como nome ou número do documento de identidade. No nosso caso, temos variáveis sensíveis e quasi-identidicadoras. Como elas estão em sua forma bruta, será necessário utilizar técnicas que possibilitem a anonimização dos dados, como as técnicas de *t-closeness* para atributos como sexo, região, raça, etc. e adicionar um ruído nas variáveis numéricas (garantindo que a distribuição é a mesma da original).

Além disso, para outras variáveis que são quasi-identificadoras, é necessário avaliar a possibilidade de aplicar técnicas de anonimização e se caso não for possível, avaliar o real ganho da utilização da informação. Como em ambas as técnicas ainda mantemos tanto a distribuição original quando a representação da população original nos subgrupos, não há nenhuma perda na utilização das técnicas mencionadas.

A origem dos dados do Sistema de Informação sobre Nascidos Vivos (*SINASC*) tem origem nas secretarias de saúde, que coletam os dados de nascidos vivos nos estabelecimentos de saúde e nos cartórios, para casos de partos domiciliares. O Ministério da Saúde só considerará a base nacional completa quando todas as UFs enviarem seus dados [6]. Portanto, o viés social pode ocorrer durante a coleta dos dados nas secretarias, pois não há garantia de que os dados enviados pelas UFs representem toda a sociedade, o que constitui um viés social. Um tipo de viés social é o viés racial, onde o sistema pode refletir as desigualdades raciais existentes no Brasil. Por exemplo, os dados podem mostrar que as taxas de mortalidade infantil são mais altas entre crianças negras do que entre crianças brancas. Isso pode ser devido a fatores como a discriminação racial no acesso a serviços de saúde, a pobreza e a falta de acesso à educação.

O problema de viés social pode acabar gerando também um problema de viés de representação, já que a amostragem pode não representar toda a população, tornando assim a base de dados desbalanceada. O desbalanceamento da base não está necessariamente relacionado ao viés social, logo, esse problema pode ocorrer mesmo sem a presença do viés social.

O problema de viés temporal pode acontecer caso ocorram alterações em features importantes. No ano de 2011, os dados do *SINASC* passaram por algumas alterações para aumentar a representatividade. Com isso, houve alterações em variáveis que tiveram mudanças na forma de coleta. Portanto, mesmo que os dados anteriores tenham sido bem coletados, essa mudança na feature pode impactar o resultado do modelo, caracterizando assim um viés temporal.

O dataset do *SINASC* para o ano de 2022 possui uma quantidade grande de dados. Sem nenhum tratamento, o dataset possui 2.471.519 registros e 61 colunas. Infelizmente, o dataset possui muitos valores faltantes, sendo fundamental que a preparação desses dados ocorra de maneira correta para evitar problemas de viés de preparação. Portanto, outro viés com o qual é necessário se preocupar é o viés de preparação, que ocorre durante o processo de preparação dos dados. Para evitar esse problema, é necessário ter bastante cuidado com a remoção de features, registros ou qualquer outra manipulação da base de dados para não fazer uma alteração enviesada.

REFERENCES

[1] Ministério da Saúde, DATASUS (Departamento de Informática do SUS), "Sistema de Informação sobre Nascidos Vivos – Sinasc (2022)", https://opendatasus.saude.gov.br/dataset/sistema-de-informacao-sobre-nascidos-vivos-sinasc. Acessado em: 21 de Agosto de 2023.

[2] Ministério da Saúde, "Anomalias Congênitas", https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/anomalias-congenitas. Acessado em: 21 de Agosto de 2023.

[3] Pires, Márian Oleques, et al. "Óbitos infantis entre os anos 2000 e 2017 em uma cidade do sul do Brasil: técnicas de mineração de dados." Research, Society and Development 9.9 (2020): e550997489-e550997489.

[4] Welke, Karl F., Tara Karamlou, and Brian S. Diggs. "Databases for assessing the outcomes of the treatment of patients with congenital and paediatric cardiac disease–a comparison of administrative and clinical data." Cardiology in the Young 18.S2 (2008): 137-144.

[5] Moore, John W., et al. "Procedural results and safety of common interventional procedures in congenital heart disease: initial report from the National Cardiovascular Data Registry." Journal of the American College of Cardiology 64.23 (2014): 2439-2451.

[6] NASCIDOS Vivos: Notas Técnicas. Notas Técnicas. 2017. Disponível em: http://tabnet.datasus.gov.br/cgi/sinasc/Nascidos_Vivos_1994_2012.pdf.Acessoem:12set.2023.