

Aplicação de Técnica de Aprendizado de Máquina Supervisionado para Predição de Custos de Medicamentos Após Cirurgia Bariátrica

Vinícius Paes de Camargo¹, Josiane Melchiori Pinheiro¹, Heloise Manica Paris Teixeira¹

¹Departamento de Informática – Universidade Estadual de Maringá (UEM)
Bloco C56 – Maringá – PR

vinicamarggo@hotmail.com, jmpinhei@din.uem.br, hmpteixeira@uem.br

Abstract. *Bariatric surgery is a procedure offered by the Unified Health System (SUS) and is of great importance for the health of obese patients, as well as for reducing the use of drugs linked to the disease. This work applies machine learning algorithms, Linear Regression and Random Forest, to predict and classify medication costs used by patients after bariatric surgery. A prototype interface was developed for health experts. The results indicate that the Random Forest model allows the prediction of medication consumption after surgery with a good hit rate, since the model obtained an accuracy of 86%.*

Resumo. *A cirurgia bariátrica é um procedimento ofertado pelo Sistema Único de Saúde (SUS) e tem grande importância para a saúde dos pacientes obesos, bem como para diminuição do uso de medicamentos ligados à doença. Este trabalho aplica algoritmos de aprendizado de máquina, Regressão Linear e Floresta Aleatória, para predição e classificação de custos de medicamentos utilizados pelos pacientes depois da cirurgia bariátrica. Um protótipo de interface foi desenvolvido para usuários especialistas em saúde. Os resultados obtidos indicam que o modelo de Floresta Aleatória permite a predição do consumo de medicamentos após a cirurgia com um bom acerto, visto que o modelo obteve uma precisão de 86%.*

1. Introdução

Entre as atuais políticas públicas do Sistema Único de Saúde (SUS) para informatização hospitalar e construção de sistemas inteligentes está o DATASUS¹, que agrega e normaliza informações de todos os bancos de dados do SUS. Existem iniciativas de gestão total do hospital que utilizam e geram conhecimentos para o DATASUS, como por exemplo, o desenvolvimento do Sistema de Gestão da Assistência de Saúde do SUS (G-SUS). No entanto, no SUS, os recursos financeiros são limitados e alvo de disputa entre as diversas unidades que o compõe, fazendo com que o investimento em novas técnicas computacionais e até mesmo médicas, dependa de uma apurada comprovação de eficácia médico-financeira.

A Organização Mundial da Saúde (OMS) estima que existem meio bilhão de pessoas obesas ao redor do mundo [Organization et al. 2016], o que torna a validação da

¹<http://datasus.saude.gov.br/>

eficácia e eficiência da cirurgia bariátrica necessária para sua manutenção no rol de tecnologias do SUS. Assim, diversas técnicas computacionais podem ser utilizadas para analisar os dados existentes sobre este procedimento, permitindo assim sua validação.

Para análise e extração das informações a partir de dados brutos uma técnica normalmente utilizada é o Aprendizado de Máquina (AM), que segundo [Witten et al. 2011], é uma forma de adquirir descrições estruturais sobre algum fenômeno a partir de exemplos. Assim, a partir de um conjunto de dados brutos, pode-se modelá-los matematicamente, de modo que em exemplos futuros, um comportamento seja previsível por meio da aplicação do modelo adquirido.

A presente pesquisa tem como objetivo principal estudar e aplicar técnicas de aprendizado de máquina em uma base de dados na área cirurgia bariátrica. Foi desenvolvido um protótipo com os algoritmos estudados, cuja principal funcionalidade é possibilitar aos gestores realizar a predição de custo de medicamentos consumidos por pacientes após a cirurgia bariátrica.

A motivação da pesquisa é dada pelos seguintes questionamentos: 1) A influência das características de entrada em algoritmos de AM são estudadas nas mais diversas aplicações [Hauray et al. 2011, Lin et al. 2008] como fatores de alto impacto em recursos computacionais e temporais. Com um subconjunto de dados como características de entrada, é possível obter uma precisão na classificação dos custos que seja semelhante à relatada na literatura? 2) Estudos apontam que pacientes tendem a diminuir a ingestão de medicamentos após a cirurgia bariátrica [Geraldo et al. 2014]. A técnica de aprendizado de máquina utilizada no protótipo pode produzir dados sobre o impacto econômico que a cirurgia proporciona?

O artigo está organizado em 5 Seções. Na Seção 1 são descritas as premissas e motivações da pesquisa. Na Seção 2 é descrito o referencial teórico para compreensão do objeto de pesquisa. A seção 3 descreve o desenvolvimento e a Seção 4 os resultados. Por fim, na Seção 5 são feitas as considerações finais sobre o trabalho.

2. Referencial Teórico

2.1. Aprendizado de Máquina

Um algoritmo de AM procura encontrar uma função matemática h , também conhecida como hipótese ou modelo, que mapeia um conjunto de características de entrada em uma saída de maneira aproximada [Russell and Norvig 2009], seja esta função sobre saídas discretas (problemas de classificação) ou saídas contínuas (problemas de regressão). Antes da aplicação das técnicas de AM é necessário entender o problema a ser tratado e definir os objetivos a serem alcançados. De posse dos dados que o algoritmo de AM irá trabalhar, é importante limpá-los e padronizá-los, para só então analisar a qualidade da base e explorar os dados, visto que a qualidade do modelo de AM obtido dependerá da qualidade dos dados apresentados e da homogeneidade obtida por meio dos dados finais [Kalashnikov et al. 2005].

Quanto ao algoritmo de AM a ser utilizado é importante considerar se ele se adapta ao problema a ser resolvido e aos objetivos a serem alcançados, além de outros aspectos como: o número de parâmetros que podem ser ajustados no modelo; a precisão almejada e o tempo disponível; e a qualidade e o tamanho da base de dados; visto que todos esses

aspectos impactam nas taxas de predição. Além da escolha do algoritmo de AM, pode ser necessário melhorar os valores dos parâmetros usados no modelo, aplicando-se a técnica de hiperparametrização, na qual define-se uma faixa de valores para cada parâmetro do algoritmo e a técnica encarrega-se de aplicá-la ao algoritmo e avaliá-la.

Se ainda assim, o modelo continuar com baixa precisão e/ou alta variância, existe uma grande chance do modelo não ser adequado ao problema em questão, sendo necessário aplicar outros algoritmos de AM; ou que os dados da base não sejam representativos do problema, sendo importante, nesse caso, realizar a avaliação cruzada estratificada [Kohavi et al. 1995], para que o modelo seja testado em diversas fatias da base de dados a cada iteração da técnica de avaliação e assim consiga-se validar se apenas não houve um acaso estatístico que selecionou como teste os elementos mais difíceis. Após a avaliação do modelo de AM ter obtido uma precisão (número de elementos corretamente classificados em relação ao conjunto completo) e revocação (número de elementos de interesse corretamente classificados) consideradas adequadas, o modelo pode então ser colocado em produção.

As aplicações de AM podem ocorrer nas mais variadas áreas e problemas possíveis. Devido a esta variabilidade, [Witten et al. 2011] agrupa os problemas em alguns tipos: *classificação*, onde espera-se que a partir de exemplos vistos, seja possível classificar um exemplo nunca antes visto; *associação*, na qual deseja-se associar determinadas saídas com um grupo de classes de entrada; *agrupamento*, que agrupa dados semelhantes sem haver um rótulo de saída; e, *regressão*, onde a classe predita é do tipo numérica contínua.

Na presente pesquisa o modelo de aprendizagem para predição dos custos pós-cirurgia se caracteriza como classificação, pois há uma base de dados com as características do pacientes (idade, sexo, entre outras) e também rótulo de saída (custo de medicamentos pós-cirurgia), que permite ao algoritmo aplicado ter um objetivo claro para treinamento, bem como insumos que possibilitam a avaliação da qualidade do conjunto de técnicas e modelos para novos exemplos.

2.2. Regressão Linear, Floresta Aleatória e outros algoritmos

A Regressão Linear (RL) é um dos algoritmos mais básicos da estatística, pois tenta induzir um modelo baseado na equação de reta entre duas (Regressão Linear Simples) ou mais (Regressão Linear Múltipla) variáveis, uma dependente e as outras independentes [Werkema and Aguiar 1996]. A técnica de RL aparenta ser uma técnica demasiadamente simples, no entanto ela considera fatores como valor base e função perda (resíduos). Para treinamento e atualização do modelo normalmente é utilizada a função soma dos mínimos quadráticos, que calcula uma soma da distância quadrática entre cada ponto e a atual reta, de modo a permanecer a reta com a menor soma das perdas quadráticas.

Outra abordagem proposta na literatura é o aprendizado do modelo de Floresta Aleatória, o qual define diversos classificadores (árvores de decisão) escolhidos por amostragem aleatória, independente e uniforme e atribui um de atributos e elementos a cada classificador [Breiman 2001]. Ao final de cada iteração, o algoritmo faz uma combinação por voto majoritário das melhores árvores e constrói a floresta final que melhor representa os dados. Os principais parâmetros do modelo de Floresta Aleatória são: 1) estimadores - número de árvores que o modelo final possuirá; 2) critério - função de medida da qua-

lidade da solução; 3) máximo de atributos - número máximo de atributos que uma árvore de decisão pode ter.

Os dois critérios de decisão mais conhecidos para construção de uma árvore de decisão são Impureza Gini, que mede o grau de impureza de um nó interno da árvore e tende a isolar em um ramo da árvore a classe mais frequente; e Entropia (ganho de informação), que mede o grau de impureza dos dados baseando-se na falta de homogeneidade dos mesmos e tende a balancear o número de registros por ramo. Por meio da Floresta Aleatória, são obtidos valores que costumam fugir do mínimo local ao mesmo tempo em que evitam sobreajuste, resultando, em geral, em um bom modelo para AM.

2.3. Trabalho Relacionados

Farmacoeconomia [Packer and Resta 2014] aplica economia ao estudo dos medicamentos com objetivo de otimizar os gastos financeiros. Estudos interdisciplinares em farmacoeconomia [Packer and Resta 2014] e ciência da computação ainda são poucos explorados na academia. No entanto, observa-se que recentemente cresceu o interesse em pesquisas que aplicam AM com a predição de risco de doenças. A pesquisa de [Orozco 2013] descreve a utilização de Floresta Aleatória e Árvore de Decisão para previsão de falha anastomótica². Nesses casos em que a incidência de um evento é baixa, os autores utilizaram a técnica SMOTE³ para geração de dados das observações menos comuns e mais representativas e remoção dos valores mais comuns e menos representativos. A base de dados do trabalho foi obtida de um hospital de Eindhoven (Holanda) e as taxas de acerto da predição foram de 90% quando utilizou-se Floresta Aleatória e 78% com Árvores de Decisão.

As descobertas descritas em [Razzaghi et al. 2017] mostram que características representativas para a predição de risco de doenças após a cirurgia bariátrica, em combinação com a técnica SMOTE, levam a acertos próximos a 100% para 4 das 5 doenças analisadas. Em sua implementação, foi utilizada uma base de dados médico-econômica de pacientes de mais de 700 hospitais norte-americanos. O modelo de AM utilizado foi Floresta Aleatória, pois segundo este, apesar da dificuldade da visualização da representação interna, ela oferece um bom balanço entre facilidade de uso e taxas de acerto; afirmação provada pela taxa de 100% na predição de angina e AVC (“derrame cerebral”).

Nos trabalhos de [Pedersen et al. 2016] e [Thomas et al. 2017] foram aplicadas técnicas de AM de redes neurais às características como peso, pressão, consumo de insulina e outras, para avaliar a efetividade financeira da cirurgia bariátrica e do aparecimento de doenças correlacionadas. Segundo os autores, o modelo pode superar em até 5% os métodos da área de análise de efetividade cirúrgica e custo/benefício a longo prazo, chegando a taxas de acerto de até 84%. Redes neurais são modelos matemáticos inspirados no funcionamento dos neurônios biológicos e na estrutura do cérebro [Goldschmidt and Passos 2005], realizando simulação das sinapses cerebrais para inferir o comportamento dos sinais de ativação (saída), sendo poderosas ferramentas para problemas de grande complexidade.

A presente pesquisa tem como áreas de estudo a farmacoeconomia

²Vazamento de conteúdo de uma união cirúrgica entre duas vísceras.

³Synthetic Minority Over-sampling Technique, em resumo, balanceamento dos rótulos na base de dados.

[Packer and Resta 2014] em conjunto à cirurgia bariátrica, aplicando um algoritmo de AM a dados reais do SUS no estado do Paraná. Destaca-se também a prototipação de uma interface gráfica para o gestor de saúde.

3. Desenvolvimento

Para o desenvolvimento, foi utilizada uma base de dados do DATASUS anonimizada, com representação em formato CSV (*Comma-Separated Values*). A linguagem de programação usada foi a Python, que foi escolhida por ser código aberto, multiplataforma e possuir grande quantidade de bibliotecas para processamento científico. A biblioteca Scikit-learn do Python foi escolhida por possuir características que tornam o código manutenível para outros desenvolvedores, por ser de código aberto e seguir convenções de nomenclatura de funções das técnicas de AM, aumentando a interoperabilidade do código.

O problema investigado foi abordado como um tipo de AM supervisionada, aplicado como um modelo de classificação, pois, por meio das características de entrada, o gestor poderá saber qual a faixa de valor (classe) que um paciente consumirá em medicamentos após a cirurgia em um horizonte de tempo especificado. No contexto específico da predição do custo medicamentoso, diversas características são avaliadas, dentre elas idade, IDH da cidade do paciente e custo pré-cirurgia.

Cabe ressaltar que as decisões sobre a limpeza e pré-processamento foram validadas por uma especialista na área de farmácia. As seções seguintes descrevem cada etapa no desenvolvimento da pesquisa.

3.1. Pré-processamento

Os dados provindos do DATASUS estavam divididos em dois arquivos, ambos em formato CSV. Foi gerado um terceiro arquivo com dados sobre o IDHM (Índice de Desenvolvimento Humano Municipal), fornecido pela base de dados do Observatório do Plano Nacional de Educação⁴.

Um arquivo continha os dados do Sistema AIH (Autorização de Internação Hospitalar) e o outro os dados provenientes do sistema HÓRUS de gestão farmacêutica. O arquivo unificado foi obtido pela interseção dos campos *id* dos pacientes presentes nos arquivos do AIH e HÓRUS e *cod.cid* dos arquivos AIH e IDHM. Foram mantidas as colunas de dados do paciente e adicionadas colunas de meses e custos pré-cirúrgico e pós-cirúrgico e IDHM da cidade de residência. A etapa de pré-processamento dos arquivos gerou um arquivo unificado com dados representando 1.091 pacientes. As características presentes foram idade, sexo, raça, CID (código da doença), IDHM da cidade de residência, número de meses contabilizados antes da cirurgia, número de meses a prever, custo pré e pós-cirurgia.

Para limpeza de ruídos, foi utilizada a técnica de análise de interquartis, que busca os dados que são 1,5 vezes o valor da faixa interquartil. Como desejou-se remover apenas os ruídos extremos, adotou-se os dados 3 vezes maiores. A aplicação foi feita sobre a variável custos de medicamento pré-cirurgia, visto que esta é a característica que apresentava alta variação ao mesmo tempo que trazia o maior ganho de informação. Com sua aplicação, dois pacientes identificados como ruídos extremos foram removidos.

⁴<http://www.observatoriodopne.org.br/downloads>

Para a redução de dimensionalidade foi utilizada uma ordenação das características da base de dados e mantidas as $K=5$ que mais auxiliam as predições do modelo. Com a execução desta técnica, observou-se que três características da base eram insignificantes para o problema, são elas: raça, cid_principal e sexo. Acredita-se que a raça e sexo não influenciem a qualidade da base, pois o tratamento de cirurgia bariátrica é padronizado e pela legislação vigente, não se permite qualquer tipo de discriminação devido à essas características. A CID, é uma sigla usada para Classificação Internacional de Doenças e Problemas Relacionados à Saúde. A característica de entrada cid_principal representa a CID da principal doença que motivou o paciente a fazer cirurgia bariátrica. Como a cirurgia só é permitida nos casos de obesidade ou suas formas derivadas, sua inclusão na base não trouxe benefícios, já que todos os valores eram similares.

Ao final desta etapa, a base possuía 6 características, apresentadas na Tabela 1.

Campo	Tipo	Significado	Característica
idade	numérico discreto	idade na data da cirurgia	entrada
idh_cidade	numérico contínuo	IDHM da cidade de residência	entrada
meses_antes	numérico discreto	número de meses antes da cirurgia	entrada
meses_depois	numérico discreto	número de meses a prever	entrada
custo_pre_cirurgia	numérico contínuo	custo total antes da cirurgia	entrada
classe	numérico discreto	representação da faixa de custo pós-cirurgia	saída

Tabela 1. Características do modelo da base de dados após redução de dimensionalidade.

3.2. Aplicação dos algoritmos de AM

Após a limpeza e preparação dos dados, a simplicidade da base e a característica de saída indicaram o algoritmo de Regressão Linear (RL) como adequado. O algoritmo foi aplicado de modo a considerar a classe (característica de saída) como sendo um número contínuo, mesmo em essência não o sendo. Porém os resultados gerados pelo modelo de RL obtido não foram tão bons quanto a literatura indicava para problemas semelhantes e como o modelo não oferecia parâmetros, esses não foram aplicados à hiperparametrização, finalizando as tentativas com um modelo de regressão.

O fato da característica de saída ser discreta, levou então a ser usado um modelo de classificação, para o aprendizado de Floresta Aleatória, que foi aplicado sem nenhum parâmetro sobre uma divisão aleatória de 80% da base para treino e 20% para testes. A revisão da literatura indicou a possibilidade de aumento da precisão via hiperparametrização, sendo sua aplicação realizada por meio da técnica Busca em Grid com Validação Cruzada.

Para aplicação, foram escolhidos 4 parâmetros dos 12 parâmetros do modelo. A escolha foi motivada pela documentação da biblioteca, que afirma que esses 4 parâmetros afetam diretamente o funcionamento do modelo. Os parâmetros que compuseram o grid foram: número de estimadores, número máximo de características de entrada, número mínimo de elementos para dividir um nó interno da árvore e número mínimo de elementos para ser um nó folha.

O algoritmo de hiperparametrização, baseia-se na precisão atingida pela combinação de cada valor com todos os outros e então retorna os melhores valores para cada parâmetro, logo, o resultado final do algoritmo aplicado ao modelo de Floresta

Aleatória foi a seguinte: número de estimadores: 100; número máximo de características de entrada: 5; número mínimo de elementos para dividir um nó interno da árvore: 100; número mínimo de elementos para ser um nó folha: 10.

Para a avaliação dos modelos, utilizou-se a técnica conhecida como Validação Cruzada, que usa cada um dos $K=10$ grupos estratificados pelo K-Fold como base de teste em uma iteração e então armazena seus resultados em um vetor. A função matemática de avaliação da qualidade do modelo foi dada pela função precisão, que é representada pelo número de positivos verdadeiros dividido pela quantidade de elementos totais. A escolha foi realizada devido a característica da saída esperada, na qual apenas é importante quantos elementos são dados como sendo de uma faixa de custo, quando realmente devem ser desta faixa de custo.

3.3. Prototipação da interface

O protótipo da interface de usuário foi desenvolvido utilizando a biblioteca Tkinter. A motivação da sua escolha foi a facilidade e rapidez com que permite a criação de protótipos e também por vir embutida nas instalações padrão da linguagem Python.

O protótipo foi construído de modo que o gestor possa inserir diversos pacientes em um arquivo base (lote de dados) e o modelo fornecerá como resposta a predição sobre a faixa de custo medicamentoso de cada paciente e, então, o processador de arquivos retornará o novo arquivo com as classes preditas.

A interface gráfica do protótipo possibilita que um usuário com conhecimento prévio em editores de planilha eletrônica possa utilizá-la. Para gerar as classificações para um arquivo com mais de 1.000 exemplos, o algoritmo leva em média 30 segundos em um computador com configuração simples.

4. Resultados

A obtenção e avaliação dos resultados deste trabalho rotulou os dados em 4 classes, os quais foram assim divididos por recomendação do especialista. A Tabela 2 mostra as faixas de custo pós-cirurgia que serviram de delimitadores para as classes.

Custo pós-cirurgia	Classe	Quantidade	Porcentagem
custo = 0	0	313	28,8%
$0 < \text{custo} < 50$	1	625	57,4%
$50 \leq \text{custo} < 150$	2	104	9,6%
$150 \leq \text{custo}$	3	46	4,2%
TOTAL		1088	100%

Tabela 2. Tabela com critérios de rotulagem dos custos pós-cirurgia segundo recomendação do especialista.

A aplicação dos dados finais ao algoritmo de Regressão Linear permitiu um coeficiente de determinação de 46% com erro quadrado médio de 0,31, o que indica que o modelo consegue generalizar e representar menos da metade do conjunto de treino e ainda que tem uma tendência a classificar um novo dado com um rótulo menor que o rótulo real. A percepção obtida foi que sua utilização seria melhor se os valores estivessem de certa maneira agrupados ou mesmo se estivessem normalmente distribuídos. Como ponto forte

desse modelo, destaca-se a simplicidade de sua implementação e também seu tempo de treinamento, sendo o menor tempo dentre todos os modelos testados.

A decisão de utilizar o modelo de Floresta Aleatória deu-se por este ser mais adequado a natureza classificatória do problema, bem como pelo fato do modelo possuir uma baixa complexidade de parametrização, ao mesmo tempo em que oferece uma boa precisão e revocação. A avaliação das características de entrada aplicadas ao modelo de Floresta Aleatória permitiram a montagem da Figura 1, que ilustra a porcentagem média de separação de dados por característica sobre 10 execuções do modelo configurado com critério de entropia.

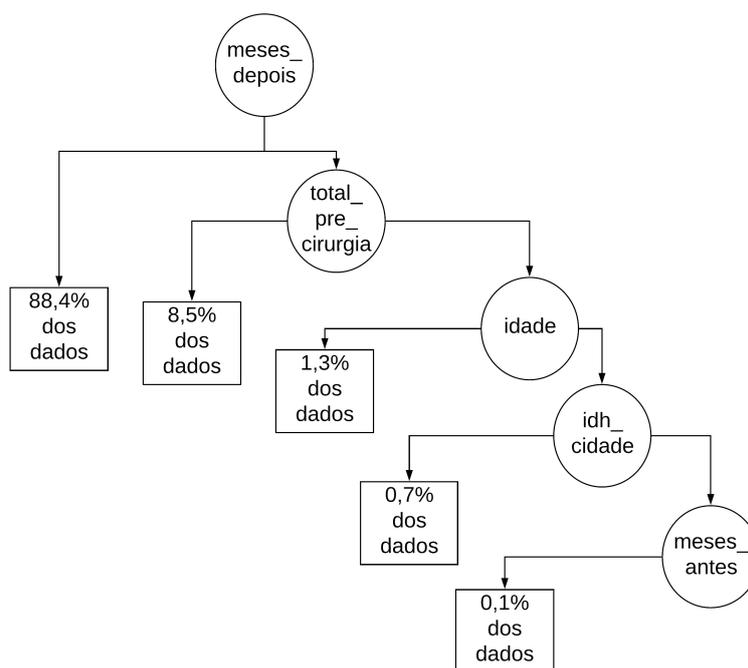


Figura 1. Árvore de ganho de informação percentual por característica.

A análise da Figura 1 nos permite entender que a característica mais importante da base de dados é *meses_depois*, que representa o número de meses à frente para o qual deseja-se prever o custo. Sua importância se dá pois o modelo tenta montar uma classificação temporal das faixas de custo e acaba decidindo unicamente baseado nesse número de meses em 88,4% dos casos da base de dados.

O *total pré-cirurgia* e *idade* são atributos que classificam 8,5% e 1,3% respectivamente pois pacientes com outras doenças (indicadas por maior custo medicamentoso pré-cirurgia) e pacientes em idades maiores tendem a possuir maior custo medicamentoso, indicando que o modelo os classifique como sendo das classes de custo mais elevadas.

Existem características que não conseguem separar muitos dados da base, como em *idh_cidade* e *meses_antes*, porém, como a etapa de redução de dimensionalidade apontou estas como sendo relevantes ao problema, estas foram mantidas e ajudam a diminuir a variância do resultado final da aplicação do modelo de Floresta Aleatória.

Com uma precisão de 81%, revocação de 86% e variância de apenas 1%, é factível que esse modelo consiga generalizar bem qualquer novo paciente. De qualquer forma, a

Tabela 3 mostra que o modelo só consegue acertar novos dados nas três primeiras classes, revelando assim que podem haver poucos exemplos de treinamento para a classe 3 ou mesmo que os dados não são suficientemente generalizáveis.

Classe	Precisão	Revocação
0	1,00	1,00
1	0,85	0,97
2	0,41	0,23
3	0,00	0,00
Média sobre o total de elementos	0,81	0,87

Tabela 3. Tabela com precisão e revocação para cada classe de custo.

Para correção do comportamento da pouca predição dos valores com classes maiores, foi aplicada a técnica SMOTE, preenchendo com um maior número de exemplos as classes 2 e 3. No entanto, a aplicação da técnica não mudou as métricas médias da aplicação do modelo de Floresta Aleatória.

A principal desvantagem desse algoritmo é seu tempo de treinamento que é superior ao algoritmo de RL. Como ponto forte, destaca-se sua alta precisão e revocação mesmo que com baixa parametrização. Outros pontos que contribuem para sua adoção em uma ampla gama de aplicações são a simplicidade de seu entendimento, seu alto número de parâmetros, a maioria derivado da árvore de decisão.

5. Considerações Finais

A escassez e limitação de recursos dos hospitais públicos brasileiros torna de grande relevância o estudo e a aplicação de técnicas computacionais, como AM, no sentido de contribuir com a tomada de decisão de gestores em saúde. Os questionamentos motivadores desta pesquisa foram respondidos com o desenvolvimento do trabalho. Com relação à primeira, “com um subconjunto de dados disponibilizados pelo DATASUS como características de entrada, é possível obter uma precisão na classificação dos custos semelhante à literatura?” os resultados indicaram que sim. A precisão de 86% obtida com a aplicação de Floresta aleatória foi condizente com as precisões da literatura, indicando que o DATASUS fornece boas características de entrada para o problema abordado neste trabalho.

Com relação à segunda pergunta, “A técnica de aprendizado de máquina utilizada no protótipo pode produzir dados sobre o impacto econômico que a cirurgia proporciona?”, o estudo mostrou que foi possível obter dados sobre o impacto econômico da cirurgia. No entanto, ressalta-se que é necessário uma avaliação dos resultados obtidos junto com especialistas e gestores em saúde, sendo este um objetivo futuro da pesquisa.

Adicionalmente, sugere-se como trabalhos futuros a aplicação do modelo de Floresta Aleatória em uma base de dados que possua informações sobre IMC e CID secundário dos pacientes, podendo também a pesquisa evoluir para análise da correlação entre os medicamentos pré e pós-cirurgia bariátrica. Quanto à interface prototipada, há uma possibilidade interessante de sua consolidação em um sistema Web.

Referências

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

- Geraldo, M. d. S. P., Fonseca, F. L. A., Gouveia, M. R. d. F. V., and Feder, D. (2014). The use of drugs in patients who have undergone bariatric surgery. *International journal of general medicine*, 7:219.
- Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações*, volume 1. Gulf Professional Publishing.
- Haurly, A.-C., Gestraud, P., and Vert, J.-P. (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLOS ONE*, 6(12):1–12.
- Kalashnikov, D. V., Mehrotra, S., and Chen, Z. (2005). Exploiting relationships for domain-independent data cleaning. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 262–273. SIAM.
- Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.
- Lin, S.-W., Ying, K.-C., Chen, S.-C., and Lee, Z.-J. (2008). Particle swarm optimization for parameter determination and feature selection of support vector machines. *Expert Systems with Applications*, 35(4):1817 – 1824.
- Organization, W. H. et al. (2016). *Global report on diabetes*. WHO Press, Gênova, Switzerland.
- Orozco, L. D. L. (2013). Performance of machine learning algorithms to predict anastomotic failure in bariatric surgery. Master's thesis, University of Technology Eindhoven.
- Packeiser, P. and Resta, D. (2014). Farmacoeconomia: Uma ferramenta para a gestão dos gastos com medicamentos em hospitais públicos. *Infarma - Ciências Farmacêuticas*, 26(4):215–223.
- Pedersen, H. K., Gudmundsdottir, V., Pedersen, M. K., Brorsson, C., Brunak, S., and Gupta, R. (2016). Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. *NPJ genomic medicine*, 1:16035.
- Razzaghi, T., Safro, I., Ewing, J., Sadrfaridpour, E., and Scott, J. D. (2017). Predictive models for bariatric surgery risks with imbalanced medical datasets. *Clemson University Research - TigerPrints Repository*.
- Russell, S. J. and Norvig, P. (2009). *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 3rd edition.
- Thomas, D. M., Kuiper, P., Zaveri, H., Surve, A., and Cottam, D. (2017). Neural networks to predict long-term bariatric surgery outcomes. *Bariatric Times*, 14(12):14–17.
- Werkema, M. C. C. and Aguiar, S. (1996). *Análise de regressão: como entender o relacionamento entre as variáveis de um processo*. UFMG, Escola de Engenharia, Fundação Christiano Ottoni.
- Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 3 edition.