Previsão de internações por doença pulmonar obstrutiva crônica através da análise de dados ambulatoriais com técnicas de aprendizado de máquina

Zilmar S. Silva[1], **Rogerio Salvini**[1]

[1]Universidade Federal de Goiás (UFG)
Alameda Palmeiras, Quadra D, Câmpus Samambaia
CEP 74690-900 - Goiânia - GO - Brasil
Fone: (62) 3521-1181 / Fax: (62) 3521-1182

[2]Instituto de Informática (INF) - UFG, Goiás, BR.

zilmarsousa@inf.ufg.br, rogeriosalvini@inf.ufg.br

Abstract. *Hospitalizations represent the highest per capita medical-hospital cost for public and private health insurance providers. As a result, they significantly increase the cost of healthcare and account for most of the cost of operators. Many of these hospitalizations can be predicted through the use of machine learning techniques applied in the analysis of patients' outpatient history data, as these patients have outpatient data available in the various public and private health information systems. One such hospitalization is hospitalization for Chronic Obstructive Pulmonary Disease (COPD). This study aims to find patterns in outpatient COPD data generating models for hospitalization prediction based on machine learning algorithms.*

Resumo. *Internações hospitalares representam o maior custo médico-hospitalar per capita para operadoras de planos de saúde públicos e privadas. Por isso aumentam consideravelmente o custo de assistência médica e representam a maior parte do custo das operadoras. Muitas dessas internações podem ser preditas através do uso de técnicas de aprendizado de máquina aplicadas na análise de dados de histórico ambulatorial dos pacientes, já que estes pacientes possuem dados ambulatoriais disponíveis nos vários sistemas de informações de saúde publica e privada. Uma dessas hospitalizações é a internação por Doença Pulmonar Obstrutiva Crônica (DPOC). Este estudo busca encontrar padrões em dados ambulatoriais de DPOC gerando modelos para a previsão de internação baseados em algoritmos de aprendizado de máquina.*

1. Introdução

Atualmente os custos de internações hospitalares estão entre os mais altos dentre todos os itens que compõem o custo médico hospitalar. É o que aponta o Índice de Variação de Custos Médico-Hospitalares de 2018 [VCMH/IESS 2018]. Um paciente ao sofrer uma internação ocupa um leito hospitalar, faz uso de medicamentos de alto custo, necessita de acompanhamento médico e de outros profissionais de saúde, necessita ter a disposição equipamentos hospitalares, vestuário, alimentação, lavanderia, etc., elevando assim o custo do serviço. Com o custo médico-hospitalar em crescente alta nos últimos

anos [VCMH/IESS 2018] e o número de beneficiários caindo, é notório o interesse de setores ligados ao serviço, o estudo de técnicas e métodos científicos e práticos que tenham potencial para diminuírem a ocorrência de internações hospitalares.

Dentre as várias patologias existentes e catalogadas que geram internações hospitalares, existem um grupo que são classificados internacionalmente como *Ambulatory Care Sensitive Condition (ACSCs)* e são foco de grande interesse no mundo todo [Billings et al. 1993]. O interesse neste grupo se refere ao fato de que tais patologias são passiveis de serem evitadas através do uso de boas práticas e tratamentos ambulatoriais adequados.

Pesquisas buscam apresentar métodos que tornem possíveis o diagnóstico precoce destas doenças, de forma que as mesmas possam ser tratadas ambulatorialmente. Existem modelos clínicos de referência, métodos estatísticos e mais recentemente foram desenvolvidos modelos utilizando algoritmos de aprendizado de máquina.

Segundo [Desikan et al. 2012], atualmente existem 16 identificadores de ACSCs no Estados Unidos. Estes identificadores são apresentados Tabela 1.

Tabela 1. Classificação Internacional - ICSAP [Desikan et al. 2012]

Nr	Grupo patológico
1	Complicação de curto prazo
2	Apêndice perfurado
3	Complicação de longo prazo
4	Asma pediátrica
5	Doença pulmonar obstrutiva crônica
6	Gastroenterite pediátrica
7	Hipertensão
8	Insuficiência cardíaca congestiva
9	Baixo peso ao nascer
10	Desidratação
11	Pneumonia bacteriana
12	Infecção do trato urinário
13	Admissão sem procedimento
14	Diabetes não controlada
15	Asma no adulto
16	Amputação de membros inferiores em pacientes com diabetes

Condições de saúde com estes identificadores são evitáveis e isso gera uma grande oportunidade para reduzir custos médicos. No Brasil segundo [Alfradique et al. 2009] estas condições estão organizadas em 20 grupos, conforme pode ser visto na Tabela 2 e são chamadas de Internações por Condições Sensíveis a Atenção Básica - ICSAB.

Dentro dessas categorias existem algumas centenas de doenças subclassificadas por seus CIDs [1]. Conforme a própria portaria ministerial 221 de 17 de abril de 2008

[1] *CID é um acrônimo da Classificação internacional de doenças (ICD - International Statistical Classification of Diseases and Related Health Problems). Atualmente está na décima versão (CID-10). Ele á*

Tabela 2. Classificação no Brasil - ICSAP

Nr	Grupo patológico
1	Doenças preveníeis imunizáveis
2	Condições evitáveis (sensíveis)
3	Gastroenterites infecciosas e complicações
4	Deficiências nutricionais
5	Anemia
6	Infecções de ouvido nariz e garganta
7	Pneumonias bacterianas
8	Asma
9	Doenças das vias aéreas inferiores
10	Hipertensão
11	Angina pectoris
12	Insuficiência cardíaca
13	Doenças cerebrovasculares
14	Diabetes mellitus
15	Epilepsias
16	Infecção no rim e trato urinário
17	Infecção da pele e tecido subcutâneo
18	Doença inflamatória de órgãos pélvicos femininos
19	Úlcera gastrointestinal
20	Doenças relacionadas ao pré-natal e parto

[Noronha 2008] que a estabeleceu as patologias da lista de ICSAB no Brasil, a intenção principal foi de que ela pudesse ser utilizada como instrumento de avaliação da atenção primária e/ou da utilização na atenção hospitalar, podendo ser aplicada para avaliar o desempenho do sistema de saúde brasileiro SUS.

Ainda sobre o Brasil, segundo [Morimoto and Costa 2017] é notório que estudos apontem não haver tendência de diminuição nos coeficientes de ICSAB, apesar de que em algumas regiões houve diminuição com o aumento de estratégias de saúde da família. Esses fatos evidenciam a necessidade de adoção de técnicas além das que já foram adotadas. Uma dessas técnicas é a predição de situações de risco através do uso da tecnologia, mais precisamente algoritmos de aprendizado automático. Abordagens usando mineração de dados e aprendizado de máquina além de outras técnicas, tem sido bastante discutidas e pesquisadas para predição de condições de saúde em pacientes com câncer, doenças cardíacas, fatores psicológicos, mal de *parkson*, hepatites, entre tantas outras, mas poucos discutidas no que diz respeito a predição de patologias relacionadas as *Ambulatory Care Sensitive Condition* (ACSC) [Morimoto and Costa 2017].

Segundo [Sarkar and Srivastava 2013] explicam, técnicas de aprendizado de máquina fizeram da tarefa de detecção de pacientes com o risco de eventos potencial-

a classificação e codificação das doenças e uma ampla variedade de sinais, sintomas, achados anormais, denúncias, circunstâncias sociais e causas externas de danos e/ou doença[Stefanie Weber 2018]. Está prevista uma nova versão (CID-11), que deve ser lançada em junho de 2018, para entrar em vigor a partir de 2022

mente evitáveis (PPEs) complicados, devido à natureza complexa da análise de um registro eletrônico de saúde (EHR), ou seja, o prontuário eletrônico do paciente, já que os dados normalmente são incompletos e dispersos e várias bases de dados. O aprendizado de máquina, quando aplicado à modelagem preditiva, pode determinar padrões de fatores de risco úteis para melhorar a qualidade da previsão.

Por ser uma das principais causas de morbidade e mortalidade [Orchard et al. 2018], várias técnicas tem sido exploradas em busca de modelos que possam detectar internações por doença pulmonar obstrutiva crônica - DPOC. Dentre as técnicas de aprendizado de máquina que vêm sendo empregadas na predição de internação por doença pulmonar obstrutiva crônica, podemos destacar: *support vector machine* e *neural networks* [Dias et al. 2014]; *neural networks* [Orchard et al. 2018]; *support vector machine*, *nayve bayes*, *decisions tree* [Swaminathan et al. 2017]; *random forest* [Spathis and Vlamos 2017]; *logistic regression* [Xie et al. 2013]

Outros grupos patológicos pesquisados também foram estudados utilizando uma variedade de outras técnicas, porém as tradicionais são as mais exploradas. Neste contexto, serão utilizados para a comparação deste estudo os modelos utlizando *support vector machine*, *multilayer perceptron*, *nayve bayes* e *decisions tree*.

O objetivo desde trabalho é de encontrar padrões em dados ambulatoriais para detecção de internações por doença pulmonar obstrutiva crônica em quatro técnicas tradicionais de aprendizado de máquina e compará-los.

O restante deste artigo está organizado da seguinte forma: a Seção 2 apresenta o método utilizado para geração do conjunto de dados; a Seção 3 aborda os resultados dos experimentos deste trabalho; a Seção 4 faz uma discussão sobre estes resultados; e a Seção 5 faz uma conclusão deste estudo.

2. Metodologia

Neste trabalho, propõe-se o uso de uma abordagem metodológica que busque equilibrar a engenharia de dados e a capacidade de representação das técnicas de aprendizado de máquina para prever a internação hospitalar. Para isso foram selecionadas as características de dois bancos de dados relacionais e depois as mesmas foram associadas em uma única base para a formação do conjunto de dados. Com uma quantidade de dados adequada é possível prever as internações e ao mesmo tempo deixar os modelos de predição com um custo computacional baixo. O que se observou durante os experimentos é que quando se utiliza uma quantidade muito grande de características, o modelo tende a um custo computacional demasiadamente alto. Então o que se buscou foi utilizar na metodologia uma quantidade de características que melhor representassem os dados, utilizando algorítimos de seleção de variáveis, selecionando apenas aquelas que melhor representam a seleção da variável dependente. A Figura 1 demonstra o fluxograma da metodologia utilizada para a pesquisa.

Figura 1. *Metodologia do experimento.*

2.1. Linguagens e bibliotecas

Os modelos propostos nestes experimentos foram implementados na linguagem *Python* utilizando a biblioteca *scikit-learn*.

A *scikit-learn* é uma biblioteca de aprendizado de máquina de código aberto para a linguagem de programação *Python*, onde inclui uma vasta quantidade de algoritmos de classificação, regressão e agrupamento incluindo *decisions tree*, *support vector machine*, *random forest*, *gradient boosting*, *k-means*, *nayve bayes*, *neural network*, etc.

A biblioteca foi utilizada em conjunto com a linguagem *Python* e outras bibliotecas cientificas como *NumPy*, *Pandas*, *Matplotlib*, *Seaborn*, *Plotly*. Os modelos foram implementados utilizando o ambiente computacional *web*, *Jupyter Notebook* em uma *Virtual Machine Debian GNU/Linux*.

2.2. Seleção de conjunto de dados

Para os experimentos desta pesquisa foram utilizados dados de pacientes do município de Mineiros - Goiás/Brasil, de dois sistemas de informações de saúde. Um sistema de informações com dados de histórico ambulatorial e outro com histórico de internações hospitalares. As informações foram coletadas de unidades básicas de saúde de gestão do SUS e possuem informações de dados clínicos, laboratoriais, epidemiológicos, sociodemográficas, ou seja, dados de prontuário eletrônico do cidadão e dados provenientes de fichas de atendimentos ambulatoriais. As internações são provenientes de 4 hospitais privados credenciados SUS. A Figura 2 demonstra o fluxograma do processo de formação do conjunto de dados, aplicado aos dois bancos de dados relacionais.

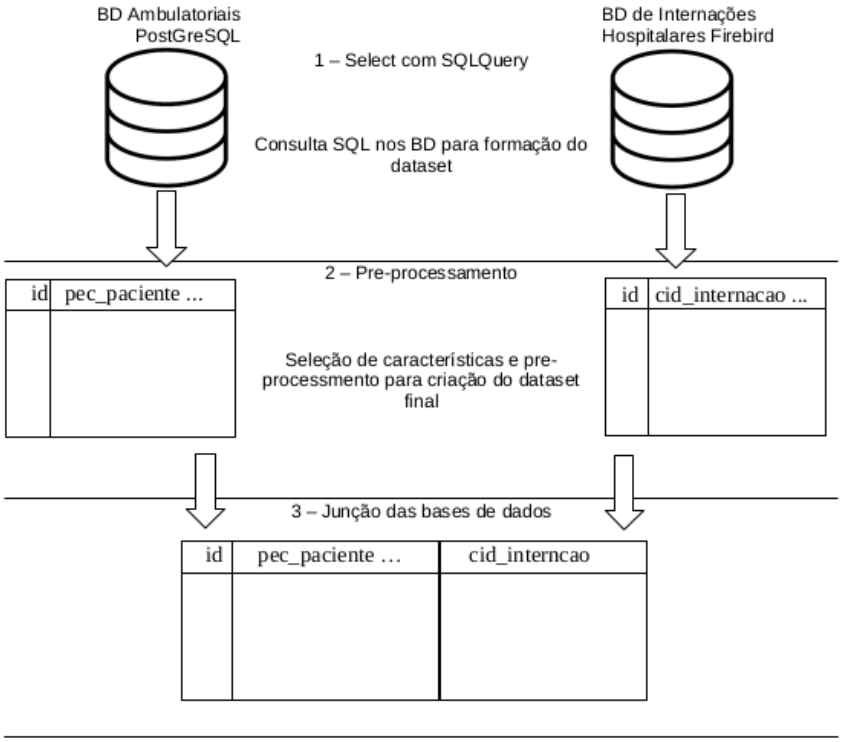

Figura 2. ***Fluxograma para formação do conjunto de dados.***

Dessas internações foram separadas apenas as patologias dentro do grupo de internações por condições sensíveis a atenção ambulatorial. E para este estudo foram selecionadas as internações por doença pulmonar obstrutiva crônica. Os registros das bases de dados ambulatoriais e hospitalares utilizados na pesquisa foram coletados nos dois bancos de dados no período de 01/2013 a 07/2018.

2.2.1. Seleção de amostras alvo

Após a seleção das características e amostras nos dois bancos de dados, foi realizada a junção das dois conjuntos de dados resultantes do *select SQL* e a associação das características dos mesmos para a identificação da variável dependente (internação) com os respectivos registros ambulatoriais, identificando nas amostras aqueles pacientes que tiveram internações após o atendimento ambulatorial. Neste processo de seleção, todas as amostras foram rotuladas como *true*, para registros de pacientes que tiveram uma internação após o atendimento ambulatorial, e *false* para paciente que não tiveram uma internação após o antedimento ambulatorial. No total o conjunto de dados ficou com 81 características e 189.208 registros, com 179.968 rotulados como não houve internação posterior ao registro (*false*) e 9.240 como houve internação. As internações por doença pulmonar obstrutiva crônica posterior ao registro representavam 981 do total de ICSAP conforme podemos observar na Tabela 3.

Como o foco do estudo são as internações por condições sensíveis a atenção primária, mais precisamente as internações por doença pulmonar obstrutiva crônica, todas as ocorrências com registros de internações por estas condições foram selecionadas. De-

Tabela 3. Resumo dos dados com as classes

Classe	Quantidade de dados	Percentual
Com internação	981	0,5%
Sem internação	179.968	99,5%

mais internações por condições, serão exploradas em trabalhos futuros e por isso foram removidas do conjunto de dados, restando apenas as internações por DPOC. A Tabela 3 mostra um resumo da quantidade de registros resultantes no conjunto de dados de treino.

2.3. Pre-processamento de dados

O conjunto de dados da pesquisa foram selecionados de dois bancos de dados relacionais e inicialmente foi realizado o pre-processamento desses dados.

2.3.1. Limpeza (seleção) de características

Características com valores duplicados foram removidas junto com aquelas que não continham nenhum valor. Por ser tratar de dados de um sistema legado, algumas características não obrigatórias são ignoradas pelos usuários e por isso resultam em registros vazios.

2.3.2. Tratamento de dados incompletos ou ausentes

Os registros da base de dados na maioria dos casos, são de dados binários para uma determinada característica. Estes dados são registrados como um série de perguntas das quais as respostas são sim ou não. Usuários tendem a ignorar estas entradas quando não são obrigatórias e por isso elas foram preenchidas com um caractere que indica que a informação foi omitida na hora do atendimento.

2.3.3. Tratamento de dados discrepantes ou atípicos

Por ser uma base em que o sistema legado utiliza se de campos pre-definidos para o preenchimento, não se encontrou dados discrepantes ou atípicos. Os dados de internações também não tinham discrepâncias, pois a fonte dos dados é de um sistema de faturamento, onde todos os dados são validados e por isso não tinham valores atípicos.

2.3.4. Criação de características derivados

Devido a natureza exploratória dos experimentos, algumas características novas foram criadas para uma melhor representação dos dados. Foram criados mais três campos:id_paciente, id_paciente_interna, e qt_paciente_interna. Para a id_paciente foi calculado a idade do paciente, utilizando a sua data de nascimento e o dia do registros no conjunto de dados. Para id_paciente_interna foi calculado com base na data de nascimento e o dia da internação. Para a qt_paciente_interna foram calculados quantos registros o paciente teve, anteriores a data de internação até o dia da internação.

2.3.5. Anonimização dos dados

As consultas SQL no banco de dados de atendimentos ambulatoriais selecionaram, apoś a geração do conjunto de dados, registros de dados brutos. O identificador utilizado para identificar unicamente o paciente foi o cartão nacional de saúde (CNS). Este atributo foi posteriormente anonimizado, por questões de privacidade com os dados do paciente, sendo assim impossível em qualquer circunstância a identificação do paciente.

2.4. Transformação

Os dados pré-processados foram então convertidos para uma estrutura compatível com o algoritmo de aprendizado.

2.4.1. Balanceamento das classes

As duas classes são extremamente desbalanceadas, conforme podemos observar na Tabela 3. Desta forma, foi utilizado a técnica de amostragem (*sampling*) para diminuir o efeito negativo desta diferença sobre os dados. Foi usado a técnica de *Random undersampling* na classe majoritária que representam os pacientes com registros ambulatoriais que não tiveram internações após o registro atendimento. Assim, a base de dados ficou balanceada com a mesma quantidade de registros de pacientes que tiveram e que não tiveram internação.

2.4.2. Seleção de variáveis

Técnicas de seleção para redução de dimensionalidade são importantes, pois tem a capacidade de selecionar os melhores recursos de forma que reduz o custo computacional e otimiza o modelo de predição. Nos experimentos utilizou-se do módulo *sklearn.feature_selection* do *python* para implementar a seleção automatizada de características.

A classe no módulo *sklearn.feature_selection* foi usada para seleção de recursos e redução de dimensionalidade no conjuntos de amostras, para melhorar as pontuações de precisão dos estimadores e para melhorar o desempenho do algorítimo no conjunto de dados.

A classe *sklearn.ensemble.ExtraTreesClassifier* implementa um meta-estimador que se encaixa em várias árvores de decisão aleatórias, em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar a adaptação excessiva. A Tabela 4 mostra as 25 variáveis mais relevantes calculadas pelo estimador.

Tabela 4. Variáveis com maior relevância

Característica	Relevância
co_exame	0,128
st_fumante	0,067
nu_altura	0,062
st_hipertensao_arterial	0,062
st_diabete	0,060
co_dim_tipo_orientacao_sexual	0,055
co_tipo_atend	0,050
nu_peso	0,050
st_infarto	0,045
tp_atend	0,040
st_problema_rins_nao_sabe	0,039
st_usa_planta_medicinal	0,037
co_dim_cbo	0,031
st_doenca_respira_n_sabe	0,031
st_deficiencia	0,025
st_solicitado_avaliadoA	0,020
st_problema_rins	0,020
st_doenca_respiratoria	0,018
st_doenca_cardiaca	0,017
st_solicitado_avaliadoS	0,015
st_domiciliado	0,014
st_doenca_respira_outra	0,012
st_proced_dab	0,012
st_avc	0,012
st_alcool	0,012

A técnica ao ser aplicada, calcula um percentual de relevância da característica para explicar a variável dependente e gerar uma classificação baseada na importância da característica (*feature_importances_*) sendo que quanto maior o valor, mais importante o recurso. Após o processamento, o algoritmo seleciona os atributos mais relevantes. Apenas atributos com *feature_importances_* maior que zero foram utilizados na construção de modelos. O *SelectFromModel* seleciona por padrão os recursos cuja importância é maior que a importância média de todos os recursos, esse limite pode ser alterado, mas nestes experimentos foi utilizado o valor padrão.

2.4.3. GridSearchCV

Os parâmetros utilizados nos modelos finais foram resultados de uma exaustiva pesquisa, utilizando uma classe em *python*, pertencendo ao pacote sklearn chamado *GridSearchCV*. O estimador GridSearchCV busca dentro de uma série de valores, os que desempenham o melhor ajuste e pontuação, para a base de dados em questão.

2.4.4. Avaliação dos classificadores

Em todos os modelos experimentadas foi utilizada a técnica de *k-fold cross validation* com o valor de k=30.

A avaliação do desempenho de uma classificação é baseada na contagem de registros de teste corretamente e incorretamente previstas pelo modelo e são normalmente organizadas em uma tabela conhecida como matriz de confusão. [Pang-Ning Tan 2005] explica que é mais conveniente resumir as informações de uma matriz de confusão em um único número para compararmos o desempenho de diferentes modelos. Isso pode ser feito com o uso da métrica de desempenho acurácia. Ela é definida como a proporção do número total de previsões que estavam corretas, ou seja, a porcentagem do total de registros classificados corretamente. A acurácia é determinada pela formula a seguir:

$$\text{Acurácia} = \frac{(TP+TN)}{(TP+FP+FN+TN)}$$

Onde:

- TP = Positivos verdadeiros (*True Positive*): quantidade de exemplos positivos corretamente classificados.
- FP = Falsos positivos (*False Positive*): quantidade de exemplos negativos erroneamente classificados como positivos.
- FN = Falsos negativos (*False Negative*): quantidade de exemplos positivos erroneamente classificados como negativos.
- TN = Negativos verdadeiros (*True Negative*): quantidade de exemplos negativos corretamente classificados.

Uma representação gráfica que é amplamente difundida para avaliar o desempenho de modelos de aprendizado é a curva ROC (*Receiver Operating Characteristic*). Ela foi usada pela primeira vez na área de medicina na década de 60 [Institute 2003].

A área sob a curva (AUC - *Area Under Curve*) ROC é frequentemente usada como medida de qualidade dos modelos de classificação [Theodoridis and Koutroumbas 2008]. Um classificador aleatório tem uma AUC ROC de 0,5 enquanto a AUC para um classificador perfeito é igual a 1. Na prática, a maioria dos modelos de classificação tem uma AUC ROC entre 0,5 e 1.

2.5. Algoritmos de Aprendizado de Máquina

Os algoritmos de aprendizado de máquina que foram utilizados para gerar os modelos classificadores neste trabalho foram: *Support Vector Machine* (SVM), Rede Neural *Multilayer Perceptron* (MLP), Rede Bayesiana *Naïve Bayes* (NB), e Árvore de Decisão (DT).

3. Resultados

Os seguintes resultados demonstrados na Tabela 5 consolidam os resultados dos experimentos.

Tabela 5. Resultados obtidos pelos quatro modelos classificadores gerados.

	DT	NB	SVM	MLP
Acurácia	0,89 (< 10E-4)	0,84 (< 10E-4)	0,97 (< 10E-4)	0,90 (± 0,05)
AUC ROC	0,91 (± 0,04)	0,87 (± 0,04)	0,97 (± 0,00)	0,90 (± 0,03)

Os resultados representam os valores de acurácia e AUC ROC com a base de testes. Foi calculado o valor de acurácia e AUC ROC e depois calculado a média e o desvio padrão do conjunto de dados. Após os treino, o modelo foi testado para verificar a capacidade de previsão de novos casos, com uma base de testes que não foi utilizada na validação cruzada, sendo reservado 30% do conjunto de dados para testes.

4. Discussão

Conforme é possível observar na Seção 3 e na Tabela 5, a técnica que obteve a melhor pontuação média de acurácia, chegando a 0,97 ($<$ 10E-4) foi o SVM, seguido pelo MLP com 0,90 ($\pm$ 0,05). Em seguida podemos observar que a DT com acurácia de 0,89 ($<$ 10E-4) teve um desvio padrão muito baixo, o que o coloca muito próximo do modelo MLP em eficácia para classificação de novos casos. E por último o NB teve o pior desempenho com acurácia de 0,84 ($<$ 10E-4).

A seguir podemos observar os resultados de AUC ROC e os valores de pontuação média de classificação. A técnica que na média obteve o melhor resultado foi o SVM alcançando valores de AUC ROC de 0,97 ($<$ 10E-4) com um desvio padrão bem baixo, seguido pelas DT com AUC ROC de 0,91 ($\pm$ 0,04), seguido pelo MLP com 0,90 ($\pm$ 0,03) e por último o NB com pontuação de 0,87 ($\pm$ 0,04).

Com os valores de acurácia e AUC ROC melhores que as demais técnicas e com desvio padrão baixo entre as amostras com SVM, pode-se entender que para este conjunto de dados e com os parâmetros testados esta técnica se mostra superior aos demais.

Os resultados apresentados na seção 3 com as quatro técnicas evidenciam modelos de qualidade e que abordagens de aprendizado de máquina utilizando dados ambulatoriais de prontuário eletrônico podem proporcionar modelos com bons níveis de assertividade para a previsão de internações por doença pulmonar obstrutiva crônica.

Neste estudo, é importante destacar o alto esforço para à coleta, modelagem e tratamento de dados de prontuário eletrônico que foram retirados de um sistema real em produção e depois pre-processados. Assim cabe registrar que o estudo diz respeito ao uso de dados ambulatoriais em um determinado período de tempo e em uma determinada população alvo. Portanto, a metodologia e a modelagem de dados utilizada, através da seleção das melhores variáveis discutido na Seção 2 e visto na Tabela 4, podem contribuir muito para gerar melhores decisões clinicas baseados na análise dos dados reais pregressos dos pacientes.

Apesar da importância incontestável de todas as fases da metodologia adotada, observa-se que se deve dar a devida importância à seleção e tratamento das características, pois são fundamentais para a geração de modelos de qualidade. Com isso, contando com uma base de dados históricos é possível com estas técnicas prever um cenário, prevendo as internações por DPOC, possibilitando assim a gestão informações decisivas para a tomada de decisão clínicas de tratamentos médicos. Apesar dos resultados, ainda não há um estudo para determinar a viabilidade de sua aplicação em um cenário real.

5. Conclusões

Internações por condições sensíveis a atenção ambulatorial é uma questão que traz grandes preocupações para os sistemas de saúde de todo o mundo. Custos operacionais tendem

a aumentar quando lidamos com pacientes de alto risco, com doenças graves. As técnicas de aprendizado de máquina são muito utilizadas na área médica pois trazem resultados com alto índice de acerto.

A utilização destas técnicas em dados ambulatoriais de prontuário eletrônico para detecção de internações se mostram promissoras, pois tem potencial para serem utilizadas pelas operadoras de saúde de todo o mundo. Os resultados destes experimentos ainda necessitam de validação prática posterior, para verificação se a técnica realmente pode ser funcional na prática do dia a dia, e se a mesma traria benefícios para as operadoras, equipes profissionais e ao paciente.

Os resultados dos experimentos disponíveis na Seção 3 demonstram que as técnicas de aprendizado de máquina podem fornecer modelos de previsões promissores, e verificando que técnicas de aprendizado de máquinas podem ter comportamentos diferentes dependendo da base de dados aplicada, assim sugerindo que a exploração de mais de uma técnica permite a escolha de uma modelo que melhor representa o conjunto de dados.

Referências

Alfradique, M. E., Bonolo, P. d. F., Dourado, I., Lima-Costa, M. F., Macinko, J., Mendonça, C. S., Oliveira, V. B., Sampaio, L. F. R., Simoni, C. d., and Turci, M. A. (2009). Ambulatory care sensitive hospitalizations: elaboration of brazilian list as a tool for measuring health system performance (project icsap-brazil). *Cadernos de saude publica*, 25(6):1337–1349.

Billings, J., Zeitel, L., Lukomnik, J., Carey, T. S., Blank, A. E., and Newman, L. (1993). Impact of socioeconomic status on hospital use in new york city. *Health affairs*, 12(1):162–173.

Desikan, P., Srivastava, N., Winden, T., Lindquist, T., Britt, H., and Srivastava, J. (2012). Early prediction of potentially preventable events in ambulatory care sensitive admissions from clinical data. In *Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on*, pages 124–124. IEEE.

Dias, A., Gorzelniak, L., Schultz, K., Wittmann, M., Rudnik, J., Jorres, R., and Horsch, A. (2014). Classification of exacerbation episodes in chronic obstructive pulmonary disease patients. *Methods Inf Med*, 53(2):108–114.

Institute, N. C. (2003). *Fourth National Forum on Biomedical Imaging in Oncology — Reports & Publications — Cancer Imaging Program (CIP)*. (Accessed on 05/29/2019).

Morimoto, T. and Costa, J. S. D. d. (2017). Hospitalization for primary care susceptible conditions, health spending and family health strategy: an analysis of trends. *Ciencia & saude coletiva*, 22(3):891–900.

Noronha, J. C. (2008). Ministério da saúde / secretaria de atenção à saúde - portaria nº 221, de 17 de abril de 2008. `http://bvsms.saude.gov.br/bvs/saudelegis/sas/2008/prt0221_17_04_2008.html`.
Acessado em: 14/07/2018 23:30.

Orchard, P., Agakova, A., Pinnock, H., Burton, C. D., Sarran, C., Agakov, F., and McKinstry, B. (2018). Improving Prediction of Risk of Hospital Admission in Chronic

Obstructive Pulmonary Disease: Application of Machine Learning to Telemonitoring Data. *J. Med. Internet Res.*, 20(9):e263.

Pang-Ning Tan, Michael Stenbach, V. K. (2005). *Introduction to Data Mining*. Addison-Wesley, 3th edition.

Sarkar, C. and Srivastava, J. (2013). Impact of density of lab data in ehr for prediction of potentially preventable events. In *2013 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 529–534. IEEE.

Spathis, D. and Vlamos, P. (2017). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics J*, page 1460458217723169.

Stefanie Weber, J. H. (2018). Who - international classification of diseases, 11th revision (icd-11). `http://www.who.int/classifications/icd/en/#`. (Acessado em: 14/07/2018 13:34).

Swaminathan, S., Qirko, K., Smith, T., Corcoran, E., Wysham, N. G., Bazaz, G., Kappel, G., and Gerber, A. N. (2017). A machine learning approach to triaging patients with chronic obstructive pulmonary disease. *PLoS ONE*, 12(11):e0188532.

Theodoridis, S. and Koutroumbas, K. (2008). *Pattern Recognition*. Academic Press.

VCMH/IESS (2018). Índice de variação de custos médico-hospitalares. `https://www.iess.org.br/cms/rep/historico_vcmh.pdf`. (Accessed on 05/24/2019).

Xie, Y., Redmond, S. J., Mohktar, M. S., Shany, T., Basilakis, J., Hession, M., and Lovell, N. H. (2013). Prediction of chronic obstructive pulmonary disease exacerbation using physiological time series patterns. *Conf Proc IEEE Eng Med Biol Soc*, 2013:6784–6787.