UNIVERSIDADE FEEVALE

IGOR VIANA DO AMARAL

APLICAÇÃO DE MACHINE LEARNING PARA BUSCA DE PADRÕES E TENDÊNCIAS NA ÁREA DA SAÚDE

(título provisório)

Anteprojeto de Trabalho de Conclusão

Novo Hamburgo
2019

IGOR VIANA DO AMARAL

APLICAÇÃO DE *MACHINE LEARNING* PARA BUSCA DE PADRÕES E TENDÊNCIAS NA ÁREA DA SAÚDE

(título provisório)

Anteprojeto de Trabalho de Conclusão de Curso, apresentado como requisito parcial à obtenção do grau de Bacharel em Ciência da Computação pela Universidade Feevale.

Orientador: Prof. Dr. Juliano Varella de Carvalho

Novo Hamburgo
2019

RESUMO

Quando um profissional da saúde realiza atendimento a um paciente, e o diagnóstico não está evidente, são solicitados exames complementares para a confirmação de hipóteses e tratamento. As informações fornecidas por esses procedimentos são armazenadas, gerando um grande volume de dados. Esta grande massa de dados possibilita aplicações de técnicas de *machine learning* (aprendizado de máquina). O aprendizado de máquina é um segmento da inteligência artificial onde os sistemas aprendem com dados, identificam padrões e tomam decisões com o mínimo de intervenção humana. O presente trabalho consiste no estudo de *machine learning*, compreensão de uma base de dados relacionada a área médica e em sequência a busca de padrões e tendências. Os dados que serão coletados são referentes a procedimentos médicos e materiais usados nos mesmos, juntamente com dados anônimos de pacientes e suas doenças pré-existentes, fornecidos pela maior rede de assistência médica do brasil, a Unimed. Com o resultado proveniente do estudo, será realizado o pré-processamento dos dados e avaliação dos algoritmos e técnicas que melhor se relacionam com os objetivos específicos para a busca de padrões e tendências. Por fim, será criada uma interface para a visualização dos resultados obtidos da aplicação e será apresentado um relatório para um especialista de domínio da área da saúde.

Palavras-chave: Machine learning, padrões e tendências, saúde, procedimentos médicos.

SUMÁRIO

MOTIVAÇÃO

Nos dias atuais, o indivíduo ao precisar de auxílio médico ou até mesmo buscando a manutenção de sua saúde, procura um profissional da área da saúde. Quando atendido por esses profissionais são seguidas uma série de procedimentos, como a *anamnese* (entrevista clínica) juntamente com o exame físico geral para fornecer uma visão ampla do paciente. Quando o diagnóstico não está claro são solicitados os exames complementares de diagnóstico para a confirmação de hipóteses e tratamento. Como exemplo, pode-se citar um exame complementar de diagnóstico: dosagem de glicose, comumente usada como diagnóstico e monitoramento do diabetes mellitus e dos distúrbios da homeostase glicêmica, como também para o rastreamento do diabetes gestacional (PROTOCOLOS EXAMES LABORATORIAIS, 2009). Outro exemplo de exame é o de creatinina para detecção de lesão renal crônica, quando 50% ou mais dos néfrons estão comprometidos (PROTOCOLOS EXAMES LABORATORIAIS, 2009).

Essa informação, gerada a partir de exames complementares, produz um grande volume de dados, disponibilizado para consulta e processamento, ocasionando o fenômeno chamado de *Big Data*. Uma das definições mais populares para *Big Data* são os "3 Vs" conforme o pesquisador Laney, sustentando o aumento tridimensional de volume, velocidade e variedade (LANEY, 2001). Vários anos depois o modelo "3 Vs" foi estendido, adicionando outras características do *Big Data* como veracidade (SCHROECK ET AL., 2012), valor (DIJCKS, 2013) complexidade e desestruturação (INTEL, 2012).

Um médico consegue realizar um diagnóstico após verificar o conjunto de sintomas e resultados de exames clínicos de um paciente, utilizando o conhecimento adquirido e a experiência. Segundo Faceli (2011) é muito difícil escrever um algoritmo que, dados os sintomas e os resultados clínicos, consiga apresentar um diagnóstico que seja tão bom quanto de um médico experiente. Entretanto, de acordo com Konenko(2001), atualmente já existem trabalhos computacionais trazendo resultados promissores na área da saúde, equiparando o resultado de diagnósticos realizados por algoritmos, com porcentagem semelhante ao acerto dos médicos.

As dificuldades em interpretar dados e coletar informações eram tratadas computacionalmente por meio da aquisição de conhecimento de especialistas de um dado

domínio, como por exemplo da medicina, que era então codificado, frequentemente, por regras lógicas (CARVALHO, 2011). Entretanto, com o avanço computacional, hoje tendo como maior apoio a inteligência artificial é possível extrair informações, buscar por padrões e informações ocultas em dados. A área de inteligência artificial tem ganho destaque, em conjunto com Aprendizado de Máquina. Ambas áreas estão intimamente relacionadas ao *Big Data*, usando o grande acúmulo de dados para produzir informação e conhecimento (AMARAL, 2016).

Uma das definições de Aprendizado de Máquina é a capacidade de melhorar o desempenho na realização de alguma tarefa por meio de experiência (MITCHELL, 1977). Segundo Monard (2008), o Aprendizado de Máquina é uma área de inteligência artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o aprendizado, bem como a construção de sistemas capazes de adquirir conhecimento de forma automática.

Um exemplo da aplicação do aprendizado de máquina é a classificação e busca por padrões em uma caixa de e-mails, como por exemplo, para distinguir a diferença entre um e-mail legítimo e um e-mail de *spam*. Sabe-se que um e-mail, na maneira mais simplista, é um arquivo de caracteres, classificados em um cenário onde são considerados *spam* ou não, tendo assim somente duas alternativas, a afirmação ou a negação. Computacionalmente, um algoritmo de classificação de e-mails sem aplicações de técnicas de aprendizado de máquina e ausência de inteligência artificial, não terá um resultado duradouro que possa ser apresentado, pois e-mails de *spam* se alteram de tempos e tempos e de indivíduos para indivíduos, dificultando a classificação pelo algoritmo.

Mesmo com as periódicas modificações de padrões, os e-mails são armazenados, e facilmente são mostrados centenas de exemplos de mensagens que são consideradas *spam*. Assim, quando utilizado o aprendizado de máquina, é realizada a classificação, identificando uma grande massa de dados, aumentando a taxa de acerto e corrigindo classificações errôneas para que o programa reconheça e detecte um *spam* (ALPAYDIN, 2010) Este tipo de classificação será baseada em exemplos, uma prática comum usada no aprendizado de máquina (MONARD, 2008).

Outra prática aplicada no aprendizado de máquina é a Conexionista, onde o aprendizado é guiado por Redes Neurais. Redes Neurais são construções matemáticas altamente interconectadas espelhadas no modelo do sistema nervoso biológico. Muitos pesquisadores acreditam que as Redes Neurais possuem grande potencial na resolução de problemas que requerem intenso processamento sensorial humano, como a visão e o reconhecimento de voz (MONARD, 2008).

No cenário de operadoras de planos de saúde, a Unimed é a maior rede de assistência médica do Brasil, o maior sistema cooperativista no mundo. Em 2018 a cooperativa situada no Vale do Sinos gerou aproximadamente 3.500.000 (três milhões e quinhentos mil) registros referentes aos exames complementares, em conjunto com os materiais utilizados nos mesmos, como remédios controlados, materiais de laboratórios, entre outros. Todos esses registros estão armazenados em uma base de dados relacionando-os com os dados dos pacientes (FILHO, 2018).

O grande volume de informações aliado à ausência de análise e exploração de uma base de dados, nos proporciona um abrangente campo de pesquisa, contendo dados referentes a exames complementares e materiais médicos disponibilizados pela Unimed Vale do Sinos. Desta maneira, este trabalho propõe o estudo e a execução de algoritmos de aprendizado de máquina, a fim de buscar padrões e/ou tendências, descobrindo informações relevantes, que quando sintetizadas, possam ser apresentadas e analisadas por um especialista de domínio.

OBJETIVOS

Objetivo geral

Descobrir padrões e/ou tendências em uma base de dados contendo exames hospitalares, materiais usados em exames e consultas médicas, a partir de aplicação de técnicas de aprendizado de máquina (*machine learning*).

Objetivos específicos

- Compreender a base de dados onde estão os exames complementares de diagnóstico, juntamente com os dados dos materiais usados nos mesmos.
- Definir objetivos para encontrar padrões e/ou tendências nos dados.
- Extrair os atributos necessários para a análise dos dados.
- Construir *scripts* de pré-processamento dos dados.
- Investigar técnicas de *machine learning* adequadas para a resolução dos objetivos definidos.
- Escolher algoritmos adequados para implementação/aplicação.
- Criar uma interface para visualização dos resultados da aplicação de *machine learning*.
- Apresentar relatório com os resultados obtidos.

METODOLOGIA

A metodologia usada neste trabalho pode ser caracterizada como pesquisa de objetivo descritiva, pois demanda técnicas padronizadas de coleta de dados usando o aprendizado de máquina. A abordagem se classifica como quantitativa, pois será aplicada técnicas de estatística, traduzindo os resultados obtidos dos algoritmos em conhecimento. O método científico é classificado como dedutivo, pois ao identificar o problema e, com premissas verdadeiras e raciocínio lógico, será alcançado uma conclusão válida.

Caracteriza-se o trabalho como de natureza aplicada pois procura produzir conhecimento para aplicação prática dirigida à solução de um problema específico, como a descoberta de padrões usando técnicas de *machine learning* em uma base de dados na área da saúde. Será realizada a busca de material bibliográfico para embasamento teórico, assim como para entendimento do contexto da base de dados para o aprendizado de máquina.

Uma vez que compreendida a base de dados, serão definidos objetivos específicos para encontrar padrões e/ou tendências nos dados. A partir dos objetivos, serão investigadas técnicas adequadas para a resolução dos objetivos definidos, bem como os algoritmos a serem aplicados.

Após o conhecimento das técnicas de *machine learning*, serão feitas as classificações dos atributos relevantes e o pré-processamento das informações presentes na base de dados. Por fim, será criado uma interface para a visualização dos resultados obtidos da aplicação dos algoritmos de *machine learning* e apresentar um relatório com a conclusão de tais resultados. Ao final do trabalho pretende-se responder: Quais padrões e tendências são encontrados na base de dados de exames médicos da Unimed?

CRONOGRAMA

Trabalho de Conclusão I

Etapa	Meses			
	Mar	Abr	Mai	Jun
Escrita do anteprojeto.				
Revisão do anteprojeto.				
Entrega do anteprojeto.				
Compreensão da base de dados.				
Análise de técnicas para a resolução do problema.				
Extrair e tratar a base de dados preparando para a aplicação dos algoritmos				
Redação do TCC I				
Revisão do TCC I.				
Entrega do TCC I.				

Trabalho de Conclusão II

Etapa	Meses			
	Ago	Set	Out	Nov
Desenvolver uma solução aplicando as técnicas estudadas				
Analisar os dados obtidos e sintetizar os resultados.				
Redação do TCC II.				
Revisão do TCC II.				
Entrega do TCC II.				

BIBLIOGRAFIA

ALPAYDIN, Ethem, Introduction to Machine Learning: Second Edition 2010.

AMARAL, Fernando. Introdução à Ciência de Dados: mineração de dados e big data, 2016.

CARVALHO, André, Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina, 2011.

DIJCKS, Jean-Pierre, Oracle: Big data for the enterprise. Oracle White Paper. Redwood Shores, CA: Oracle Corporation, 2013.

FACELI, Katti; LORENA, Ana Carolina; GAMA, João; CARVALHO, André Carlos Ponce de Leon Ferreira de. Inteligência artificial: uma abordagem de aprendizado de máquina. [S.l: s.n.], 2011.

FILHO, Ubiratan, Análise de viabilidade técnica para coleta de informações de um monitor multiparamétrico, 2018.

INTEL, IT Center, Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data. Intel IT Center. Santa Clara, CA: Intel Corporation. 2012.

KONENKO, Igor, Machine learning for medical diagnosis: History, state of the art and perspective, 2001.

LANEY, Doug. . 3-D Data Management:Controlling Data Volume, Velocity and Variety. META Group Research Note, 2001.

MAURO, Andrea De MARCO. Greco, GRIMALDI, Michele, What is Big Data ? A Consensual Definition and a Review of Key Research Topics, AIP Conf. Proc., vol. 1644, 2014.

MONARD, Maria (2008). 4 Conceitos sobre Aprendizado de Máquina. Disponivel em <http://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf> acessado em: abril/2019

RESENDE. Leticia. (2009). Protocolos Exames Laboratoriais. Universidade Federal de Minas Gerais. Disponível em: <http://www.uberaba.mg.gov.br/portal/acervo/saude/arquivos/oficina_10/protocolos_exames_laboratoriais.pdf>. Acesso em: abril/2019.

SCHROECK, Michael., SHOCKLEY, Rebbeca, SMART, Janet, ROMERO-MORALES, Dolores, TUFANO, Peter, Analytics: The real-world use of big data. New York, NY: IBM Institute for Business Value, Said Business School, 2012.