Nanodegree Engenheiro de Machine Learning

Projeto Final

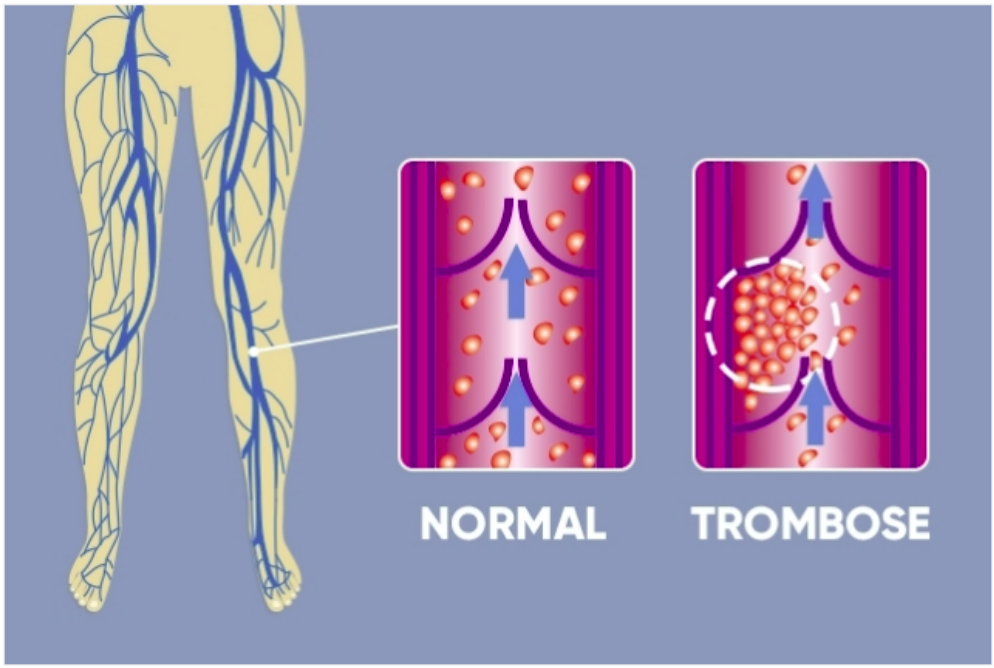
Proposta Capstone

Rosana Guimarães Ribeiro
21 Novembro 2017

Proposta: Aplicação de técnicas de Aprendizado de Máquina para diagnóstico da doença Trombose Venosa Profunda

Domínio do Background

A trombose venosa profunda é uma doença caracterizada pelo surgimento de um coágulo de sangue – o trombo – em um vaso do sistema venoso profundo, o que determina a obstrução parcial ou total dessa veia e impede a circulação local. Esse evento pode afetar qualquer parte do corpo, mas acomete com mais frequência as veias das pernas e coxas, às vezes de forma associada a uma inflamação venosa – a flebite –, quando recebe a denominação de tromboflebite. A formação do coágulo deriva de diversos fatores de risco, especialmente da imobilidade prolongada, sendo mais comum em pessoas com idade superior a 50 anos.
A trombose pode ser superficial ou profunda, como a trombose venosa profunda. Entretanto, em qualquer dos casos o tratamento com medicação deve ser urgente, porque o coágulo de sangue pode fluir através da corrente sanguínea alojando-se em órgãos como os pulmões, gerando uma embolia pulmonar, ou no cérebro, gerando uma trombose cerebral, por exemplo, situações graves que podem até levar à morte [1].



O desenvolvimento da trombose venosa profunda está principalmente relacionado com a diminuição da velocidade da circulação, também chamada de estase venosa, que ocorre quando o indivíduo é obrigado a ficar numa mesma posição por muito tempo, como em hospitalizações e imobilizações prolongadas e em viagens muito longas, particularmente as aéreas.
Da mesma forma, diversas situações e doenças que tornam o sangue mais viscoso levam o organismo a

produzir trombos com mais facilidade, como o uso de anticoncepcionais, a gravidez, o diabetes, a existência de tumores e as moléstias do sangue. Outros importantes fatores de risco para a trombose ainda incluem a presença de varizes, um episódio recente de infarto do miocárdio, a obesidade, o tabagismo e o próprio envelhecimento, só para citar alguns.

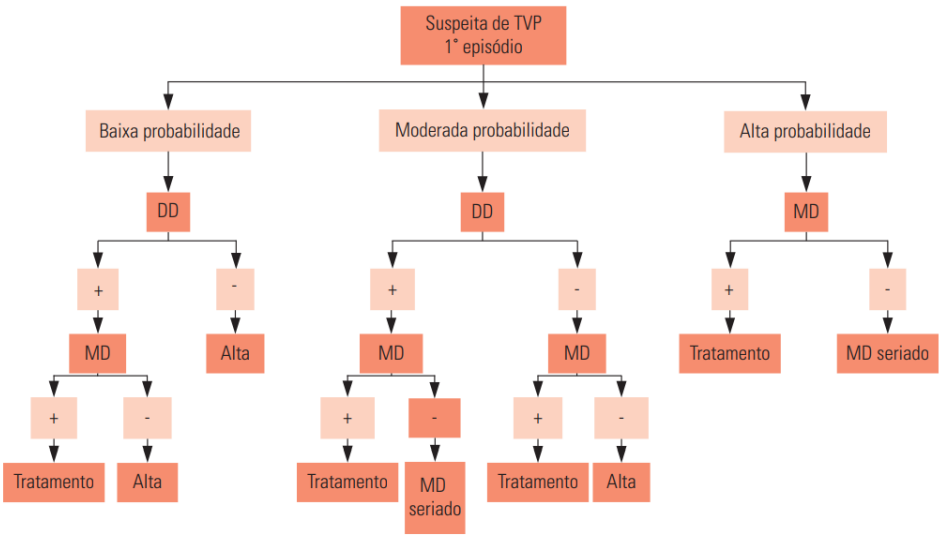
A trombose venosa profunda (TVP) é motivo de atenção especial dado seu elevado risco relativo e absoluto e importante morbimortalidade, principalmente em pessoas acima dos 50 anos ou aqueles que apresentam suscetibilidade genética para a coagulação sanguínea. Neste contexto, os exames radiológicos assumem papel propedêutico na abordagem da enfermidade.

Declaração do Problema

Paciente com edema e dor em membro inferior unilateral, de instalação súbita deve ser avaliado pensando-se no diagnóstico de TVP. Os principais sintomas são dor, edema, aumento da temperatura, hiperemia, sensação de peso no membro, palpação de cordão venoso e manifestações sistêmicas como febre baixa e taquicardia. O diagnóstico clínico da TVP pode ser difícil, pois grande parte dos casos é assintomática, e também devido ao fato de que 30 a 50% dos pacientes com sintomas e sinais de TVP não apresentam a doença. A sintomatologia é comum a outras afecções e, assim, apenas os sinais clínicos não são suficientes para o diagnóstico de TVP [2].

Embora alguns pacientes procurem o médico com um quadro clínico bastante evidente, existem condições onde o diagnóstico pode ser mais difícil. Mesmo assim, o cirurgião vascular, utilizando dados da história pessoal e familiar, análise de fatores de risco conhecidos, exame físico, probabilidades de outros diagnósticos de outras doenças e a avaliação por meio de um índice de score, são capazes de classificar o paciente num grupo de baixa, média ou alta probabilidade de ter trombose venosa profunda.

Para abordagem diagnóstica da trombose venosa profunda, o cirurgião vascular precisa analisar uma série de fatores de riscos para detectar o grau de trombose do paciente, como pode ser visto na figura abaixo, onde DD e MD são exames conhecidos como dímero–D e mapeamento dúplex, respectivamente.



A detecção e o reconhecimento de doenças com base em sistemas de aprendizagem de máquina podem fornecer indícios para identificar e tratar as doenças em seus estágios iniciais. Comparativamente, a identificação visual de doenças requer a experiência de um profissional da área, no modo geral, é cara, difícil e muitas vezes ineficiente.

A detecção automática de doenças, não somente da trombose é um tópico de pesquisa essencial, pois pode resultar em benefícios no monitoramento de grandes campos da área de saúde, e assim detectar automaticamente os sintomas de doenças logo que aparecem. Assim sendo, procurar um método rápido, automático, menos dispendioso e preciso para detectar casos da doença trombose é de grande significado realista.

Portando, visando a criação de uma aplicação para auxiliar os médicos no diagnóstico da doença trombose. Este projeto toma como base para diagnóstico desta doença um conjunto de dados tendo como formato de entrada, dados **numéricos** e **categóricos** e saída esperada **quatro categorias**, sendo esta **multi-label**. Sendo assim, a tarefa de aprendizagem aqui abordada faz uso de classificação multi-label.

Conjuntos de dados

No diagnóstico da doença trombose verificou-se que complicações em pacientes estão intimamente relacionadas aos anticorpos anticardiolipina. A trombose deve ser tratada como uma emergência e é importante detectar e prever as possibilidades de sua ocorrência. Para isto, uma base de dados contendo observações de pacientes é disponibilizada para investigar o grau de trombose.
Esta base de dados foi coletada do hospital da Universidade de Chiba, Japão. Tais dados podem ser encontrados nos sites:
http://lisp.vse.cz/pkdd99/Challenge/chall.htm (http://lisp.vse.cz/pkdd99/Challenge/chall.htm)
http://lisp.vse.cz/pkdd99/Challenge/tsumoto.htm (http://lisp.vse.cz/pkdd99/Challenge/tsumoto.htm)
Basicamente, a aplicação será desenvolvida com base em duas tabelas: TSUM_A.CSV, TSUM_B.CSV. Onde TSUM_A.CSV, apresenta informações básicas dos pacientes, ver tabela abaixo.

item	meaning	remark
ID	identification of the patient	
Sex		
Birthday		YYYY/M/D
Description date	the first date when a patient data was recorded	YY.MM.DD
First date	the date when a patient came to the hospital	YY.MM.DD
Admission	patient was admitted to the hospital (+) or followed at the outpatient clinic (-)	
Diagnosis	disease names	multivalued attribute

E TSUM_B.CSV apresenta os exames laboratoriais, ver tabela abaixo.

item	meaning	remark
ID	identification of the patient	
Examination Date	date of the test	YYYY/MM/DD
aCL IgG	anti-Cardiolipin antibody (IgG) concentration	
aCL IgM	anti-Cardiolipin antibody (IgM) concentration	
ANA	anti-nucleus antibody concentration	
ANA Pattern	pattern observed in the sheet of ANA examination	
aCL IgA	anti-Cardiolipin antibody (IgA) concentration	
Diagnosis	disease names	multivalued attribute
KCT	meassure of degree of coagulation	
RVVT	meassure of degree of coagulation	
LAC	meassure of degree of coagulation	
Symptoms	other symptoms observed	multivalued attribute
Thrombosis	degree of thrombosis	0: negative (no thrombosis) 1: positive (the most severe one) 2: positive (severe) 3: positive (mild)

A base de dados TSUM_A.CSV possui 1241 instâncias. Já a base de dados TSUM_B.CSV possui 807 instâncias, com 725 registros na classe **sem trombose** (0), 58 registros na classe **positivo, mais severo** (1), 18 registros na classe **positivo, grave** (2) e apenas 5 registros na classe **positivo leve** (3).

Enunciação da solução

Esta aplicação visa utilizar técnicas de apredizagem de máquina para diagnóstico da doença Trombose Venosa Profunda. Sendo assim, dado um conjunto de atributos relacionados às informações dos pacientes e exames laboratoriais, o modelo tem como variável-alvo o grau de trombose, a qual pode ser diferenciada por quatro classes diferentes:

0: negativo (sem trombose)
1: positivo (o mais severo)
2: positivo (grave)
3: positivo (leve)

Dessa fora, para a aplicação aqui abordada, as técnicas de apredizagem de máquina serão utilizadas para classificar os dados entre estas quatro classes.

Portanto, esta aplicação tem a finalidade de analisar esses dados e criar um modelo automatizado, levando em consideração as informações do paciente e seus dados laboratoriais. O modelo construido utilizando técnicas de aprendizagem de máquina pode então auxiliar profissionais de saúde no processo de diagnóstico.

A idéia é trabalhar no desenvolvimento de métodos para a classificação automática da doença trombose com base nos dados numéricos e categóricos resultando nas quatro classes.

Estudos mostram que métodos de aprendizagem de máquina podem ser aplicados com sucesso como um mecanismo eficaz na detecção de doenças. E para este projeto será aplicado em especial o método Árvore de Decisão, além de testes variando seus parâmetros. Outros métodos também serão abordados como Redes Neurais Artificiais (ANNs), KNN e Support Vector Machines (SVMs).

A Árvore de Decisão será abordado em especial pois, é um método fácil de entender e interpretar. Outras técnicas normalmente requerem normalização dos dados, variáveis de teste precisam ser criadas e campos em branco precisam ser removidos, enquanto para árvore de decisão a preparação dos dados pode ser dispensável ou desnecessário. Este método está apto a lidar tanto com dados nominais ou com dados categóricos. E por fim, é robusto, e tem bom desempenho com grandes quantidades de informação em pouco tempo.

Modelos de benchmark

A mesma base de dados apresentada na seção de Conjuntos de Dados foi utilizada por Shraddha Modi abordando métodos de classificação multirelacional para mineração de dados e machine learning. No artigo cujo título é, Relational Classification using Multiple View Approach with Voting, Shraddha, apresenta um algoritmo proposto e resultados experimentais para abordagem de visão múltipla com votação como técnica de combinação de visão. Ao final, Shradda apresenta tabelas de análises dos experimentos e resultados de precisão para esta base de dados.

Este artigo pode ser encontrado através do link:
http://research.ijcaonline.org/volume70/number16/pxc3888126.pdf
(http://research.ijcaonline.org/volume70/number16/pxc3888126.pdf)

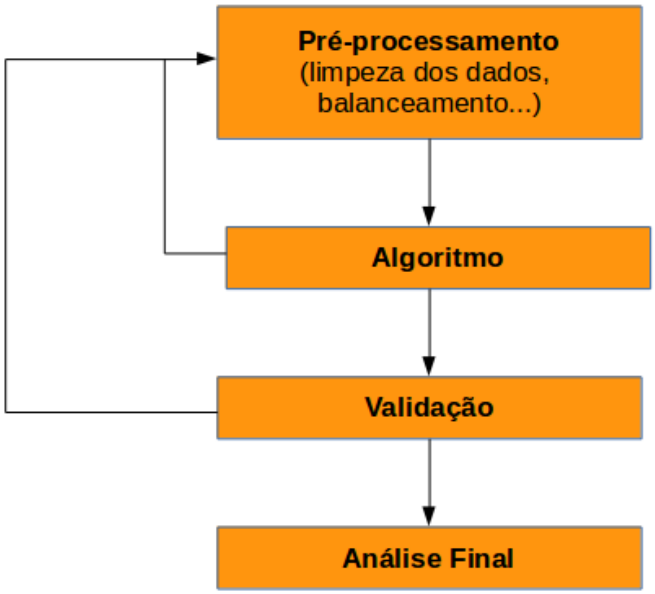
Conjunto de métricas de avaliação

Após a aplicação dos métodos de classificação, será feito o processo de validação para seleção do algoritmo que melhor generaliza o modelo. Será apresentada a matriz de confusão de cada algoritmo abordado avaliando as quatro classes junto com a medida de precisão (Precision), cobertura (Recall) e F-Measure para determinar o quão próximo do real cada algoritmo pode chegar.

Por fim, através de tabelas, gráficos e cálculos, será feita a análise dos dados para determinar qual modelo melhor generaliza para esta aplicação de diagnóstico da doença trombose venosa profunda. Um gráfico apresentando a curva ROC será gerado para avaliação dos classificadores.

Esboço do design do projeto

Para o desenvolvimento deste projeto e aplicação das técnicas de machine learning, serão utilizadas ferramentas simples e eficientes para mineração de dados e análise de dados. O código será desenvolvido em Python, fazendo uso do jupyter notebook. E como ferramenta principal será utilizada a Scikit-learn, sendo que os níveis de aprendizagem do sklearn possibilitam construir soluções sobre pacotes existentes em Python – como NumPy, SciPy e matplotlib.
De acordo as métricas citadas anteriormente, este projeto possui o seguinte fluxo de trabalho:



Uma das etapas que antecede o processo de aprendizagem de máquina é o de pré-processamento que engloba o tratamento e a preparação dos dados. Para que sejam descobertos padrões de qualidade é importante que essa etapa seja cuidadosamente executada. Ainda segundo, o desempenho dos algoritmos de aprendizado de máquina geralmente é afetado pelo estado em que os dados se encontram, ou seja, pela qualidade dos dados disponíveis. Podem ser mencionadas algumas das tarefas incluídas nessa fase, a saber: limpeza dos dados, tratamento de dados faltantes, seleção e construção de atributos, balanceamento, dentre outras.
Em seguida, serão feitas a seleção de alguns algoritmos como: Naive Bayes, KNN, SVM, Árvore de Decisão, Rede Neural e Adaboost. Para cada um dos algoritmos serão feitos ajustes de seus parâmetros que melhor se adequa à base de dados.

Referências Bibliográficas

[1] Tua Saúde. **Como identificar a trombose e o que fazer para evitar**. Website: https://www.tuasaude.com/trombose/ (https://www.tuasaude.com/trombose/). Acesso em 29 de Novembro de 2017.

[2] Pereira, M. Gomes, M. Madureira, N. Bersan, P. Silva, R. Vilela, R. Silva, S. Martins, T. Krettli, W. **Diagnóstico da trombose venosa profunda e particularidade na gravidez e puerpério**. Revista Med Minas Gerais. 2011. 21(4 Supl 6): S1-S143.