Análise Epidemiológica: Algoritmos de Aprendizado de Máquina para Classificação de Doenças

Antônio Victor Ribeiro Morelli

Luciano Silva

Faculdade de Computação e Informática - Universidade Presbiteriana Mackenzie (UPM) – São Paulo, SP - Brasil

victormorelli6@gmail.com

Resumo. *O objetivo deste estudo foi identificar doenças através de um classificador com aprendizagem supervisionada, buscando a alta acurácia do mesmo, de forma que ele fosse capaz de classificar doenças com base nos sintomas apresentados pelos pacientes. Buscando esse fim, foram utilizadas informações epidemiológicas de algumas doenças bem conhecidas e foi criado um banco de dados com pacientes, seus respectivos sintomas e as doenças que lhes foram atribuídas no diagnóstico. Posteriormente, foi aplicado um modelo de aprendizado de máquina supervisionado SVM (support-vector machine) nesse banco de pacientes com a finalidade de classificar, com alta acurácia, as doenças com base nos sintomas apresentados.*

Palavras-chave: *SVM, aprendizado de máquina, epidemiologia, aprendizado supervisionado.*

Abstract. *The aim of this study was to identify diseases through a classifier with supervised learning, seeking its high accuracy, so it would be able to classify diseases based on the symptoms of the patients. For this purpose, epidemiological information of some well-known diseases was used, and a database with patients was created, including their respective symptoms and the diseases that were assigned to them in the diagnosis. Subsequently, a SVM supervised machine learning model (support-vector machine) was applied to the patient database in order to classify, with high accuracy, the diseases based on the presented symptoms found in the patients database.*

Keywords: *SVM, machine learning, epidemiology, supervised learning.*

1. Introdução

No mundo atual, o surgimento de uma vasta gama de tecnologias e suas aplicações geraram uma quantidade infindável de dados. Extrair um conhecimento analítico a partir destes dados se tornou muito estratégico, e as aplicações relacionadas à epidemiologia são de particular interesse para toda a humanidade. Para isso, é necessário o uso de ferramentas capazes de analisar estes dados com a precisão devida. Porém muitos sintomas dessas doenças são iguais ou semelhantes, o que dificulta o processo de aprendizado de máquina principalmente na identificação e classificação de doenças.

O aprendizado de máquina surgiu como um ramo da inteligência artificial devido à necessidade da automatização de análise de dados. Deste modo, originaram-se novas possibilidades para este campo analítico, tal como a construção de modelos analíticos que são, basicamente, sistemas que podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

Isto tornou possível a construção de modelos em uma escala infinitamente superior àquela dos humanos, algo que antes era inimaginável. A evolução advinda da inteligência artificial que busca não apenas compreender a realidade, mas como também construir entidades inteligentes. (RUSSELL e NORVIG, 2016, p. 3).

Este trabalho tem como finalidade construir um modelo de classificação epidemiológica. Para tal, foi utilizado como base um banco de dados de pacientes fictícios para fazer classificação das doenças, lançando mão do aprendizado de máquina supervisionado. Também foi demonstrado como se aplicou o teste de acurácia com matriz de confusão e *baseline*, e como se chegou ao resultado da acurácia dos modelos criados.

Este artigo está organizado em seções. Na seção 2 é mostrada a metodologia, com as ferramentas utilizadas para o desenvolvimento do projeto e a modelagem do banco de dados; na seção 3, é detalhado o treinamento dos modelos criados; na seção 4, é mostrado o teste do primeiro modelo; na seção 5, é mostrado o teste do segundo modelo; na seção 6, é mostrado o teste do terceiro modelo; na seção 7, é exposta a conclusão e são sugeridos rumos para os trabalhos futuros, e na seção 8, são apresentadas as referências bibliográficas utilizadas durante a confecção deste artigo.

2. Metodologia

O trabalho dividiu-se em dois momentos. Primeiramente foi realizada uma pesquisa bibliográfica para sustentação do trabalho, e posteriormente foi feita a escolha das ferramentas, linguagens e bibliotecas que seriam utilizadas pelo projeto.

A escolha de *python* 3 como linguagem para construção do algoritmo de aprendizado de máquina superviosonado foi feita devido ao *python* ser uma linguagem de alto nível, atualizada, com bibliotecas de ponta e tem um ótimo suporte para aprendizado de máquina. As bibliotecas escolhidas foram: o *sklearn* para o aprendizado de máquina, pandas e *numpy* para a análise de dados, e *seaborn* e *matplotlib* para a visualização dos dados. Para a modelagem do banco de dados, foram considerados valores binários para demarcar se há ou não os sintomas. Não foram utilizados valores bipolares devido a ser uma abordagem que demandaria mais adaptações e mais trabalho para o modelo desse projeto, uma vez que o *SVM* é um classificador binário linear.

No projeto foram utilizados alguns métodos para encontrar a acurácia de modelos de classificação. Um deles foi a matriz de confusão, também conhecida como matriz de erro, que é, basicamente, uma tabela que permite a avaliação da performance de um algoritmo através da contagem de seus erros e acertos.

No inicio do projeto, houve a criação do banco de dados. Neste, foram inseridas 4 doenças, a saber: Febre Amarela, Dengue, Chagas e Ebola. Seus respectivos números foram: 1, 2, 3, 4.

TIPO	DOENÇA
1	Febre Amarela
2	Dengue
3	Chagas
4	Ebola

Figura 1. Banco de Dados – Tipos e Doenças

	Tipo	Doença	Febre	Dor de Cabeça	Dores no corpo	Náuseas	Fadiga	Vômitos	Febre Alta	Hemorragias Em Geral	Forte Dor Retro Ocular	Icterícia	Olhos Amarelados	Inchaço no rosto e pernas	Diarreia
0	1	Febre Amarela	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	Febre Amarela	1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	Febre Amarela	1	1	1	1	1	1	1	1	0	1	1	0	0
3	1	Febre Amarela	1	1	1	1	1	0	1	1	0	1	1	0	0
4	1	Febre Amarela	1	1	1	0	0	0	0	0	0	0	0	0	0
5	1	Febre Amarela	1	0	1	1	1	0	0	0	0	0	0	0	0
6	1	Febre Amarela	1	0	0	0	1	1	1	0	0	0	0	0	0
7	1	Febre Amarela	1	0	0	0	1	1	1	0	0	1	0	0	0
8	1	Febre Amarela	1	0	0	0	1	1	1	0	0	1	0	0	0

Figura 2. Banco de Dados – Sintomas (Características)

Para o banco de dados dos pacientes fictícios e dos seus respectivos sintomas, as características possíveis (sintomas) para cada doença são: febre, dor de cabeça, dores no corpo, náuseas, fadiga, vômitos, febre alta, hemorragias em geral, forte dor retro ocular,

icterícia, olhos amarelados, inchaço no rosto e nas penas, e diarreia, conforme se pode verificar na figura 2.

Cada linha corresponde a um paciente e, caso ele tenha apresentado algum sintoma, este será identificado com o número 1, sendo utilizado o número 0 para a negativa do sintoma. A doença com a qual ele foi diagnosticado aparece também na tabela, e será utilizada para o treinamento do modelo.

Após a criação do banco de dados, foi notado que as características (sintomas) são as variáveis independentes, e que o tipo (doença) é a variável dependente.

Então, com a fórmula da regressão linear:

$$y = \theta_0 + \theta_1 . X_1$$

Nosso *y* será a variável de resposta que corresponde à coluna tipo. X_1 é a variável independente, que corresponde, no modelo, às colunas de características. θ_0 e θ_1 serão os parâmetros da reta.

3. Treinamentos

O algoritmo utilizado, *SVC* (*support vector classifier*), faz parte de uma classe maior de métodos de análise de dados chamada de *Machine Learning* (aprendizado de máquina). Para que o computador consiga aprender sobre os dados que ele vai precisar prever, nada mais justo do que fornecê-lo uma amostra desses dados, de forma que ele os analise previamente e consiga construir um modelo de predição para os mesmos.

Assim sendo, inicialmente utilizou-se uma amostra de tamanho variável e com pacientes diferentes, mas, como se era de esperar, não foi possível obter um resultado consistente quanto à acurácia, pois para cada teste se obtinha um valor diferente.

Para resolver esse problema, foi utilizado um recurso que permitiu escolher os mesmos pacientes para todas as iterações, de forma que se tornou possível reproduzir os resultados para qualquer execução do programa. Esse recurso é chamado de *seed*, e consiste em alimentar o algoritmo de geração de valores aleatórios com um número que vai determinar a seleção dos números aleatórios (que a partir de agora serão pseudoaleatórios, já que teremos sempre os mesmos resultados).

Para verificar o impacto da seleção de amostras na acurácia, foram utilizadas três técnicas de amostragem, e todas utilizaram como algoritmo de classificação o vetor de suporte linear.

As técnicas de amostragem utilizadas foram:

- Seleção aleatória (primeiro teste)

- Seleção sistemática (segundo teste)

- Seleção estratificada (terceiro teste)

Buscando otimizar o valor do *seed* para cada técnica de amostragem, foi criado um algoritmo no qual se variava o valor do *seed* e se observava o impacto dessa variação na acurácia, de forma que foi possível maximizar a acurácia para um determinado valor de *seed* sem causar *overfitting*.

3.1. Treinamento Primeiro Modelo

No primeiro modelo foi utilizada a técnica de amostragem aleatória, e os dados para o treinamento podem ser vistos na tabela 1.

Tabela 1. Proporção das características de treinamento do primeiro modelo

Doença	Quantidade Selecionada	Proporção do Total
Febre Amarela	12	80.00%
Dengue	11	73.33%
Chagas	12	80.00%
Ebola	9	60.00%

No primeiro modelo foi utilizado o método *rand* da biblioteca *numpy* para selecionar os elementos do banco de dados a serem utilizados no treinamento e no teste.

A amostragem aleatória consiste em selecionar dados sem nenhum critério, e tem como desvantagem a possibilidade de que alguma doença não seja selecionada para o treinamento ou para o teste.

3.2. Treinamento Segundo Modelo

No segundo modelo foi utilizada a técnica de amostragem sistemática, e os dados para o treinamento podem ser vistos na tabela 2.

Tabela 2. Proporção das características de treinamento do segundo modelo

Doença	Quantidade Selecionada	Proporção do Total
Febre Amarela	11	73.33%
Dengue	12	80.00%
Chagas	11	73.33%
Ebola	10	66.67%

No segundo modelo foi utilizado o método *train_test_split* da biblioteca *sklearn* para selecionar os elementos do banco de dados a serem utilizados no treinamento e no teste.

A amostragem sistemática consiste em selecionar cada n-ésimo elemento de uma amostra, e tem como desvantagem a possibilidade de que alguma doença não seja selecionada para o treinamento ou para o teste, e depende de a amostra ter sido embaralhada previamente para obter um bom resultado.

3.3. Treinamento Terceiro Modelo

No terceiro modelo foi utilizada a técnica de amostragem estratificada, e os dados para treinamento podem ser vistos na tabela 3.

Tabela 3. Proporção das características de treinamento do terceiro modelo

Doença	Quantidade Selecionada	Proporção do Total
Febre Amarela	11	73.33%
Dengue	11	73.33%
Chagas	11	73.33%
Ebola	11	73.33%

No terceiro modelo foi utilizado o método *train_test_split* da biblioteca *sklearn* com o parâmetro *stratify* em *y* para selecionar os elementos do banco de dados a serem utilizados no treinamento e no teste.

A amostragem estratificada consiste em selecionar a mesma proporção de elementos de uma amostra, e tem como vantagem evitar deixar alguma doença de fora do treinamento ou do teste.

4. Testes

Nesta etapa é verificada a acurácia dos modelos geradas na etapa de treinamento. Para tal comparamos o teste base conhecido como *baseline* com os resultados obtidos para cada modelo.

O *baseline* consiste na criação de uma matriz do tamanho do banco de dados, onde todos os valores das características (sintomas) são definidos como 1, ou seja, todos os pacientes apresentam todos os sintomas.

A matriz de confusão é uma ferramenta que mostra os erros e acertos do modelo gerado, fazendo distinção para os erros entre falso negativo e verdadeiro negativo, e igualmente para os acertos, entre falso positivo e verdadeiro positivo. É uma ferramenta de grande valia, pois permite analisar o desempenho de forma analítica.

4.1. Teste do Primeiro Modelo

Na execução do teste base a acurácia obtida foi de 13.33%, enquanto o primeiro modelo treinando obteve uma acurácia de 93.33%.

Portanto, o primeiro modelo teve uma melhora de 600% na acurácia comparado com o teste base.

A matriz de confusão do primeiro modelo é mostrada na figura 1, onde podemos verificar os acertos da máquina na diagonal, e qualquer valor fora dela são erros. Portanto, o único erro deste modelo foi diagnosticar um caso de febre amarela como dengue.

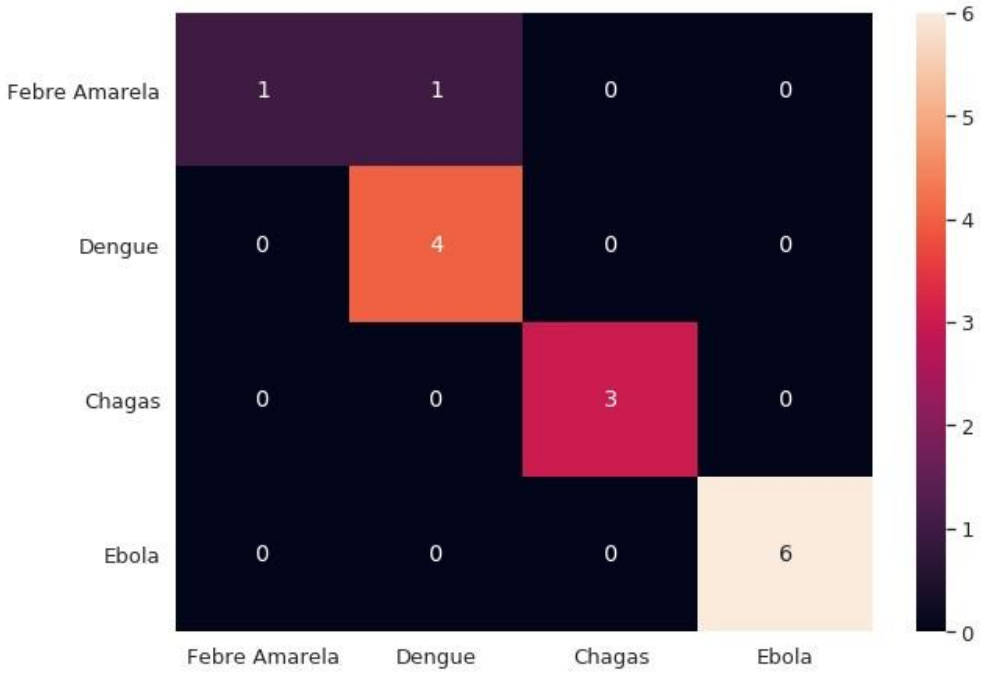

Figura 1. Matriz de Confusão do Primeiro Modelo da Aprendizagem de Máquina

4.2. Teste do Segundo Modelo

Na execução do teste base a acurácia obtida foi de 20%, enquanto o segundo modelo treinando obteve uma acurácia de 93.33%.

Portanto, o segundo modelo teve uma melhora de 366% na acurácia comparado com o teste base.

A matriz de confusão do segundo modelo é encontrada na figura 2, onde podemos verificar os acertos da máquina na diagonal, e qualquer valor fora dela são erros. Portanto, o único erro deste modelo foi diagnosticar um caso de febre amarela como dengue.

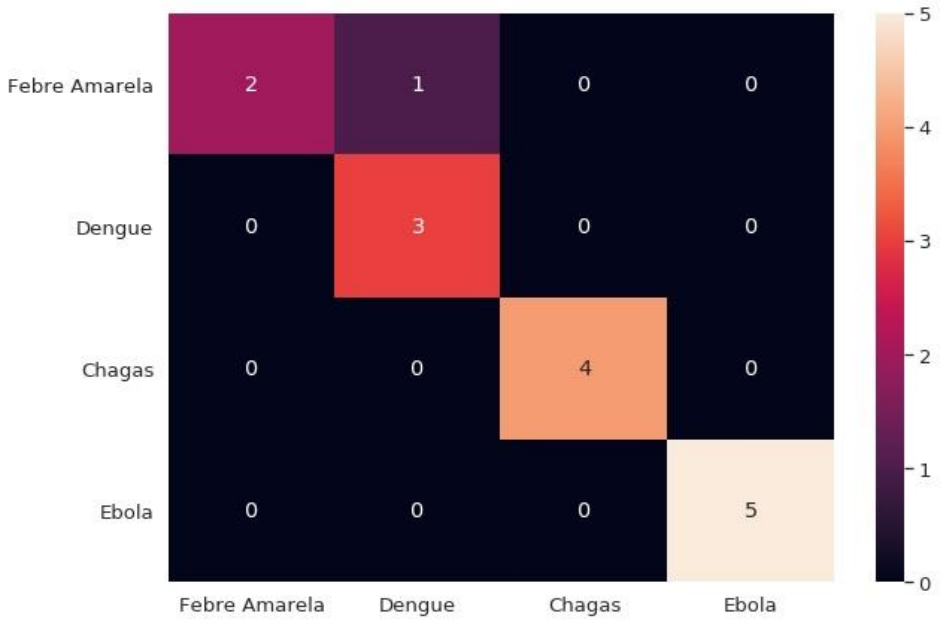

Figura 2. Matriz de Confusão do Segundo Modelo da Aprendizagem de Máquina

4.3. Teste do Terceiro Modelo

Na execução do teste base a acurácia obtida foi de 20%, enquanto o terceiro modelo treinando obteve uma acurácia de 93.33%.

Portanto, o terceiro modelo teve uma melhora de 366% na acurácia comparado com o teste base.

A matriz de confusão do terceiro modelo é mostrada na figura 3, onde podemos verificar os acertos da máquina na diagonal, e qualquer valor fora dela são erros. Portanto, o único erro deste modelo foi diagnosticar um caso de febre amarela como ebola.

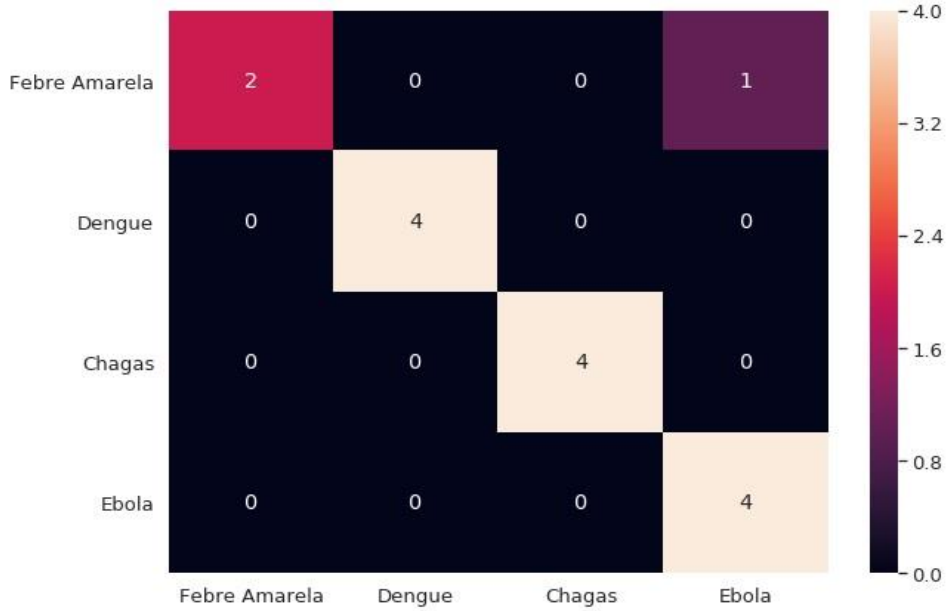

Figura 3. Matriz de Confusão do Terceiro Modelo da Aprendizagem de Máquina

7. Conclusões e Trabalhos Futuros

Este trabalho apresenta uma solução para classificação de doenças a partir dos sintomas do paciente. Tal solução visa propiciar um melhor entendimento do diagnóstico de doenças em um modelo de aprendizado de máquina, que neste trabalho foi implementado com a aprendizagem supervisionada.

Todos os modelos, do um ao três, tiveram a mesma acurácia quando o número de elementos para treinar era de 45 e o número de elementos para testar era de 15. A utilização do *seed* permitiu adicionar a característica de reprodutibilidade ao experimento. O algoritmo de escolha do *seed* chegou a um resultado que otimizou a acurácia através da seleção de uma proporção de 75% do banco de dados para treino e 25% do banco de dados para teste. A proporção de dados selecionados, para o treinamento e para teste, alterada pelo *seed* gerou a maior acurácia possível para os modelos aqui tratados.

Como possíveis trabalhos futuros, pode-se apontar:

- A implementação de um aplicativo para detecção de doenças auxiliando diagnósticos médicos em localidades com baixa disponibilidade de profissionais de saúde, principalmente médicos, com um banco de dados real e com pacientes reais;
- A implementação de um algoritmo para verificar o surto epidemiológico cíclico regional de doenças a partir da pesquisa dos seus sintomas nos mecanismos de busca na internet, e
- A implementação de um algoritmo não supervisionado para o diagnóstico de doenças e para a sugestão de tratamento, de acordo com os sintomas e necessidades de cada paciente.

8. Referências Bibliográficas

CHEN, D. Y. Análise de Dados com Python e Pandas. Tradução de Lúcia A. Kinoshita. 1ª. ed. [S.l.]: Novatec, 2018. ISBN 9788575226995.

COSTA, E.; SIMÕES, A. Inteligência Artificial. Fundamentos E Aplicações. 3ª. ed. [S.l.]: FCA, 2008. ISBN 9727223400.

COURNAPEAU, D. Scikit learn. Scikit learn, 2007. Disponivel em: <https://scikit-learn.org/stable/index.html>. Acesso em: 20 Janeiro 2019.

CVE - Centro de Vigilância Epidemiológica, 28 dez. 2018. Disponivel em: <http://www.saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/areas-de-vigilancia/doencas-de-transmissao-por-vetores-e-zoonoses/dados/famarela/famarela_seriehistorica.pdf>. Acesso em: 21 Janeiro 2019.

CVE - Centro de Vigilância Epidemiológica, 2019. Disponivel em: <http://www.saude.sp.gov.br/resources/cve-centro-de-vigilancia-epidemiologica/areas-de-vigilancia/doencas-de-transmissao-por-vetores-e-zoonoses/doc/famarela/2019/fa19_boletim_epid_1802.pdf>. Acesso em: 25 Fevereiro 2019.

DUA, D.; GRAFF, C. UCI Machine Learning Repository, 2007. Disponivel em: <http://archive.ics.uci.edu/ml>. Acesso em: 10 Fevereiro 2019.

GOVERNO FEDERAL. Ministério da Saúde. Disponivel em: <http://portalms.saude.gov.br/saude-de-a-z/doenca-de-chagas>. Acesso em: 15 Janeiro 2019.

GOVERNO FEDERAL. Ministério da Saúde. Disponivel em: <http://portalms.saude.gov.br/saude-de-a-z/ebola>. Acesso em: 15 Janeiro 2019.

GOVERNO FEDERAL. Ministério da Saúde. Disponivel em: <http://portalms.saude.gov.br/saude-de-a-z/febre-amarela-sintomas-transmissao-e-prevencao>. Acesso em: 15 Janeiro 2019.

GOVERNO FEDERAL. Ministério da Saúde. Disponivel em: <http://portalms.saude.gov.br/saude-de-a-z/dengue>. Acesso em: 15 Janeiro 2019.

RUSSELL, S.; NORVIG, P. Artificial Intelligence: A Modern Approach. 3ª. ed. [S.l.]: Pearson, 2016. ISBN 8535237011.

WOODWARD, M. Epidemiology: Study Design and Data Analysis. 3ª. ed. [S.l.]: Chapman and Hall/CRC, 2013.