Aprendizado de máquina na medicina: uso da classificação para diagnosticar pacientes

Gabriel Schrader Vilas Boas, João Pedro Assumpção Evaristo

Universidade Federal de São Paulo
São José dos Campos

gabriel.schrader@unifesp.br

joao.evaristo@unifesp.br

***Resumo*— Neste artigo, será visto todo o processo de manipulação de um *dataset* com dados que definem se um dado indivíduo, a partir de suas características, terá - ou não - derrame. Isso será feito por meio do uso de técnicas preditivas, como o aprendizado de máquina supervisionado e o *deep learning*, vistos durante a unidade curricular de Inteligência Computacional.**

I. Resumo

O acidente vascular cerebral é uma doença muito comum, que tem como alvo, geralmente, indivíduos com determinadas características. Dito isso, o objetivo deste trabalho é auxiliar, por meio do uso de técnicas computacionais de aprendizado supervisionado e de ciência de dados, o diagnóstico de pacientes que possivelmente podem ter um derrame.

Para alcançar o objetivo definido anteriormente, teremos que a extração de padrões dos dados disponibilizados pelo *dataset Stroke Prediction* será feita por meio das técnicas de *data science*, enquanto que a classificação de ter derrame - ou não - será feita com o uso de algoritmos de aprendizado supervisionado. Vale comentar que o uso, ao final, dos algoritmos de aprendizado só é capaz de produzir resultados satisfatórios após o emprego das técnicas de tratamento de dados, análise de padrões, etc.

Para a parte de extração de padrões, ficou claro quais são as características que melhor definem a potencial chance de um indivíduo apresentar um derrame posteriormente, sendo elas a idade, o valor do nível médio de glicose e o índice de massa corporal. Os resultados obtidos pelos algoritmos foram parecidos, obtendo uma acurácia média de 75%, o que é um bom resultado segundo a literatura de aprendizado de máquina.

Com os dados reunidos ao final do projeto, resultados satisfatórios foram obtidos. Os padrões observados, de fato, condizem com a realidade das pessoas que já tiveram derrames, segundo estudos feitos na literatura médica. Tratando a respeito das acurácias de predição, levando em conta a acurácia média de 75%, acreditamos que os modelos explorados podem ser utilizados para auxiliar um especialista da área quanto à rotulação, mas não pode ser utilizada unicamente para tal tarefa.

II. Introdução

O derrame - ou popularmente conhecido como AVC - é uma doença que se define como o entupimento ou rompimento de vasos que conduzem sangue ao cérebro, levando à uma paralisia da área do cérebro a qual não teve circulação sanguínea, sendo uma das principais causas de morte, incapacitações e internações mundo afora. Só no Brasil, mais de 150 mil casos de derrame por ano são registrados, segundo o hospital israelita A. Einstein.

Dito isso, vem à mente a pergunta: Por que não aliar a tecnologia, a ciência, a computação e os dados, de forma que seja possível tentar predizer um possível caso de AVC, dadas características de um indivíduo, e com isso tentar evitar a ocorrência de um derrame? Assim, surge a motivação deste trabalho e do emprego de técnicas computacionais para tentar auxiliar, de alguma forma, o diagnóstico e/ou o prognóstico do derrame cerebral.

Além do problema a ser tratado neste artigo, existem outros trabalhos que também aliam a inteligência artificial de forma a predizer doenças, como por exemplo o diagnóstico médico de pacientes com pneumonia a partir de imagens raios X do tórax. Esse exemplo e outros serão comentados posteriormente em um tópico dedicado a trabalhos semelhantes a este.

Partindo agora para a resolução do problema proposto anteriormente, será feito o uso de 3 técnicas de inteligência artificial, são essas: *Support Vector Machine*(SVM), técnica de aprendizado de máquina supervisionado, *Random forest*, outra técnica de aprendizado de máquina supervisionado, e *Multilayer Perceptron*(MLP), rede neural de *deep learning*. Todas as três técnicas foram escolhidas por serem boas técnicas de classificação, que julgamos serem adequadas para uma solução do problema com o *Stroke Prediction Dataset*.

III. Fundamentação Teórica

Para entender a proposta deste trabalho, é fortemente indicado que o leitor do documento possua alguns conhecimentos prévios, uma vez que, com tais conhecimentos, o entendimento do que foi feito e será mostrado seja pleno. Dito isso, nesta seção serão comentados quais tópicos são recomendados (ou até mesmo essenciais) para que o leitor tenha melhor compreensão do que será apresentado no decorrer deste artigo.

Inicialmente, é esperado que o legente entenda minimamente sobre o que é o acidente vascular cerebral, de sigla AVC. Tal fundamento é essencial para uma melhor compreensão dos estudos feitos sobre os dados e entender a finalidade do trabalho feito. Na seção anterior, de introdução deste documento, existe um breve resumo que pode auxiliar o leitor na melhor compreensão do que é tal doença que acomete tantas pessoas diariamente.

Dando continuidade aos fundamentos necessários para a melhor compreensão deste trabalho, é recomendado que o leitor tenha conhecimentos sobre o tratamento de dados, incluindo o conhecimento em: o que é o tratamento de dados, técnicas, análises, dentre outros. Este é o fundamento teórico mais importante para o pleno entendimento deste trabalho, pois a maior parte do que foi feito ao longo deste é, justamente, o tratamento e a análise dos dados existentes no *dataset*. Dito isso, técnicas como *oversampling*, *undersampling*, *discretização*, dentre outros, precisam ser já conhecidas pelo leitor.

Para finalizar os tópicos aconselhados a se ter conhecimento prévio, define-se que é importante conhecer o que é a classificação dentro do segmento da inteligência artificial, junto de uma visão geral de algumas técnicas para classificação de dados. Para esta recomendação de conhecimento, espera-se que o leitor conheça técnicas de aprendizado supervisionado (como o *Random Forest*).

Com as recomendações definidas, espera-se, agora, que o leitor entenda todo o processo realizado que será apresentado no decorrer deste artigo.

IV. Trabalhos Relacionados

Agora, serão vistos alguns trabalhos que, assim como este, também fazem uso de técnicas de classificação de IA com base em um *dataset*, de modo que se possa auxiliar no processo de diagnóstico médico para outras doenças. Estes trabalhos, em sua maioria, foram obtidos no *website Kaggle*, *site* no qual o próprio *dataset* trabalhado foi retirado.

Primeiro, temos o trabalho de *Mohit Baliyan* [1], o qual utiliza da técnica de rede neural convolucional para detectar, com base em imagens de raios X, se um paciente tem pneumonia, aliando o *deep learning* da medicina a fim de se diagnosticar problemas de saúde - trata-se de uma classificação.

Em *Stroke Data | Analysis and Prediction* [2], *Ruthvik PVS* mostra diversas perguntas que podem ser respondidas sobre o *dataset* utilizado neste trabalho com diversos conceitos de inteligência artificial, como matriz de confusão, dentre outros, além de utilizar a classificação por meio da árvore de decisão.

Joshua Swords propôs, em *Predicting a Stroke* [3], uma maneira de predizer se um indivíduo apresentará - ou não - um derrame cerebral com base no *Stroke Prediction Dataset*, por meio do *Random Forest* e utilizando como medida de eficácia o *F1 score*.

Em *Comparing performance of 13 Classifiers | F1 score* [4], o autor compara diversos algoritmos de classificação (como o SVM, também aqui utilizado) em um *dataset* para análise de falhas no coração de pacientes.

Por último, temos *HeartAttack prediction with 91.8% Accuracy* [5], de *Fahad Mehfooz*, que faz uso do método de Support Vector Machine para saber se um paciente terá um ataque do coração, com 91.8% de precisão.

Como pode-se ver, vários outros trabalhos tratam uma temática semelhante a que será vista e

trabalhada neste relatório, utilizando até a mesma abordagem (métodos/técnicas).

V. METODOLOGIA

Nesta seção, será vista e exposta a metodologia aplicada para tratar o problema de classificação de pacientes que podem - ou não - apresentar um acidente vascular cerebral (AVC), com base em determinadas características. Antes de partir para a explicação da metodologia aplicada em todo o trabalho, vamos a um fluxograma que a ilustra, tornando mais fácil a compreensão de tudo. Segue, abaixo, o fluxograma que define a metodologia aplicada:

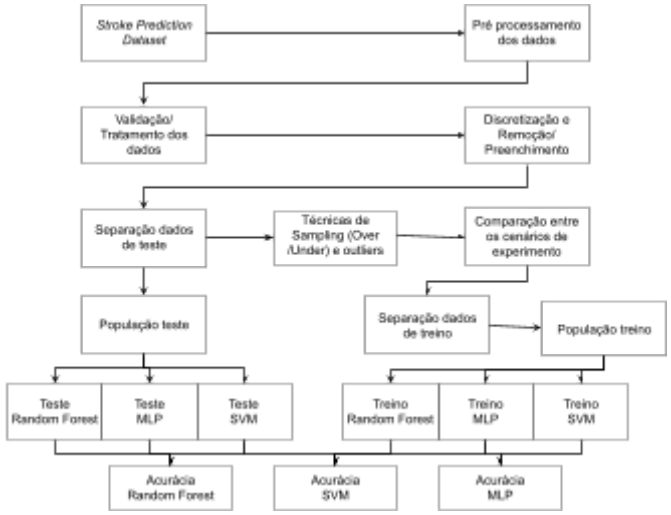

Fig. 1. Fluxograma da metodologia empregada.

Com o fluxograma da metodologia mostrado, vamos às explicações destas etapas. Inicialmente, foi desenvolvida a ideia por trás do planejamento da aplicação dos algoritmos, estabelecendo uma ordem de execução de passos para que o objetivo final - a classificação - fosse concretizado de forma satisfatória, organizada e de fácil aplicabilidade. Nesse sentido, a ideia de organização e metodologia é: manipular dados, aplicar métricas para adaptar quais tipos de dados precisam ser adaptados, utilizar as funções de separar indivíduos de treino e teste, e aplicar os algoritmos para a classificação, para que possa então ser calculada a acurácia de cada algoritmo empregado no *dataset*.

Com a ordem de execução de cada passo definida, parte-se para os detalhes de cada um dos passos supracitados. Assim, os passos seguidos para tratar e resolver o problema escolhido neste relatório, de forma aprofundada, são:

- Primeiro, manipular os dados, discretizando dados para números e separando em dois *datasets*, um no qual as linhas que possuíam dados faltantes foram removidas e, outro em que essas linhas foram preenchidas com a média das respectivas colunas. Esta etapa foi feita com o auxílio da biblioteca Pandas;
- Após a manipulação, foi feita a separação dos dados de testes em ambos os *datasets*, que serão utilizados posteriormente para os testes nos modelos aplicados. Essa etapa deve ser feita antes para que dados sintéticos do *Oversampling* ou faltantes no *Undersampling* não tenham efeitos na hora de qualificar os algoritmos. Para essa etapa, separamos 30% de cada *dataset* para testes.
- aplicar *Oversampling* e *Undersampling* em ambos *datasets*. O primeiro método funciona de maneira que, para os indivíduos em minoria, aumenta seu número de forma aleatória, de forma que atinjam um equilíbrio com a classe dominante. Já o segundo funciona de maneira contrária, reduzindo o número de amostras da classe dominante de forma aleatória até que haja um equilíbrio entre as classes. Os dois têm como objetivo deixar o conjunto de dados mais uniforme, o que evita que, em casos de *datasets* muito desbalanceados como o abordado aqui, haja um maior número de amostras de uma classe no momento do treino dos modelos, fazendo com que não saibam reconhecer uma das classes. No *dataset* específico, a classe que possui menor número é a 1 - pessoas que tiveram um AVC. São 167 amostras nas quais há a ocorrência de AVC contra 3410 em que não há. Ambas as técnicas podem ser aplicadas de forma fácil por meio de bibliotecas como a *imblearn*. Após essa etapa, ainda foram removidos os outliers de ambas abordagens, a fim de evitar que dados discrepantes impactam significativamente no treino;
- Com os *datasets* devidamente balanceados, foi feita uma comparação entre os diferentes cenários levantados, dados preenchidos e *Oversampling*, dados preenchidos e *Undersampling,* dados removidos e

Oversampling, e dados removidos e *Undersampling*. Para a comparação, utilizamos o modelo *Regressão Logística,* por ser um dos algoritmos que melhor explicitou as diferenças em nossos testes. Na média de cinco testes, obteve um melhor resultado para o cenário no qual os dados foram removidos e utilizado o *Undersampling*. Portanto, a partir daí, passamos a utilizar o dataset e os testes com essas configurações para o resto do pipeline de processos.

- É necessário então, separar os indivíduos que serão utilizados para os treinos. Como já havíamos separado o teste, em 30% dos dados, o restante, isto é, 70%, irão para treino. Essa separação é feita de maneira simples através da biblioteca *sklearn*. Nesta etapa, utilizamos os parâmetros padrões da função que faz essa separação. Além disso, vale lembrar que a separação desses indivíduos em tais subgrupos se dá de forma aleatória, conforme segue a função utilizada, isso significa que, para cada execução do algoritmo, diferentes populações de teste serão formadas;
- Por fim, aplicar os algoritmos Support *Vector Machine, Random Forest* e *Multilayer Perceptron*, por meio do uso de funções também disponibilizadas pela biblioteca *sklearn*. No caso do *SVM*, é possível escolher diversos parâmetros em sua função, como o tipo de *kernel*, parâmetro de regularização, grau, *gamma*, dentre outros. Quanto ao *Random forest*, podemos alterar parâmetros como o número de árvores, a profundidade máxima, a função para medir a qualidade de uma divisão, dentre vários outros. E para o *MLP*, pode-se alterar as camadas ocultas da rede neural, a função de ativação das camadas ocultas, o *alpha*, dentre outros parâmetros. Para esse experimento, foram usados os parâmetros que levaram a um melhor resultado, com base em testes de ajustes manuais.

Assim, conclui-se a definição da metodologia utilizada durante o desenvolvimento de todo o trabalho feito no decorrer deste relatório e estudo referente ao *dataset Stroke Prediction*.

VI. Experimentos e Discussão

Para finalizar este trabalho, vamos enfim para os experimentos realizados e discutir um pouco sobre eles.

Ao longo da execução do trabalho, diversas reuniões foram feitas semanalmente, de forma que era analisado o progresso que vinha sendo feito. Dentre as diversas reuniões, muitos experimentos foram feitos, principalmente na questão de ciência de dados (encontrando padrões entre os dados). Nestes experimentos comentados, diversos algoritmos de *over/undersampling, featurezação*, predição, *heatmap* e métodos para relacionar foram testados, experimentos esses que serão mais detalhados a seguir.

Os experimentos para saber a importância das *features* foram feitos de várias formas, sendo por uso de *Árvore de Decisão*, algoritmo de seleção de *features* (que apenas foi empregado), e teste de acurácia de algoritmos preditivos com a remoção de dados. Para a árvore de decisão, foram observados que as colunas *bmi, avg_glucose_level* e *age* eram as características do *dataset* que mais importam (aparecendo mais vezes na topologia da árvore com uma entropia maior), mesmo resultado obtido pelo algoritmo de *featurezação*. Tais resultados se mostraram verdadeiros posteriormente, quando uma série de experimentos foram realizados sobre cópias do *dataset* original, mas com remoções de colunas específicas, o que deu por verdade a importância dessas características.

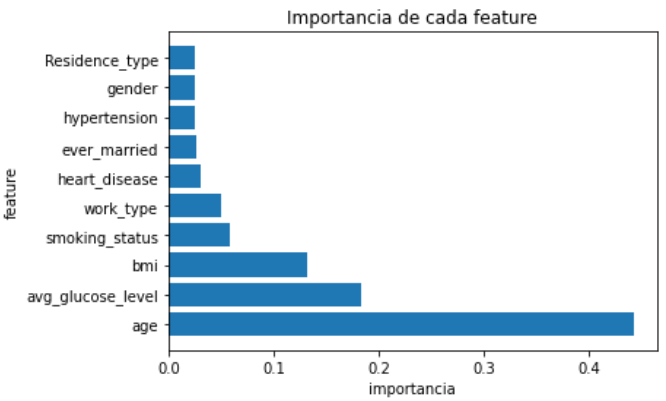

Fig. 2. Importância de cada *feature* segundo o algoritmo de *featurezação*.

	Features	Acuracias
0	gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status	77.846213
1	gender, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status	68.642316
2	gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, bmi, smoking_status	77.739980
3	gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, smoking_status	78.524242
4	age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status	77.026541
5	gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi	77.846213
6	gender, age, hypertension, heart_disease, ever_married, Residence_type, avg_glucose_level, bmi, smoking_status	78.807528
7	gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_statuss	67.716412
8	age, heart_disease, avg_glucose_level, bmi	79.060628

Fig. 3. Comparação de acurácias para testes de features

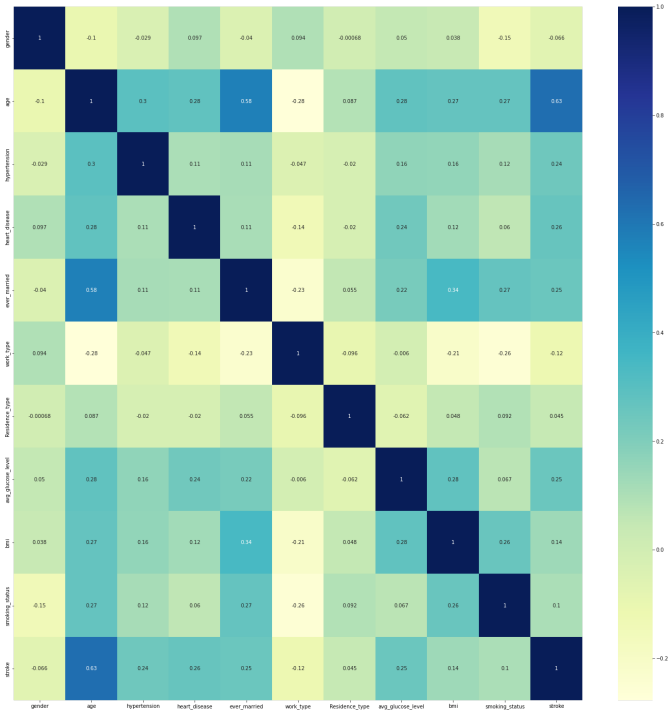

Fig. 4. Heatmap para ver a importância e correlação das features .

Tratando a respeito dos experimentos de tamanho do conjunto de amostras, foram testados diferentes cenários como abordado anteriormente. Aplicando os teste de predição sobre os quatro cenários, concluiu-se que o cenário no qual a técnica de *sampling* aplicada foi o *Undersampling,* junto com a configuração na qual as linhas que tinham ocorrências de valores nulos eram removidas, produziu os resultados mais satisfatórios para seguir com o tratamento dos dados.

```
Acurácia média com os dados preenchidos (over):  52.64 %.
Acurácia média com os dados preenchidos (under):  75.34 %.
Acurácia média com os dados removidos (over):  52.80 %.
Acurácia média com os dados removidos (under):  77.48 %.
```

Fig. 5. Acurácias com *over/undersampling e remoção de dados faltantes.*

Agora, vamos falar sobre os experimentos feitos na classificação final do trabalho. Antes de mais nada, é preciso dizer que a classificação foi a parte final do trabalho. Ao longo das reuniões sobre o progresso sendo feito no trabalho, o foco voltou-se muito aos experimentos de tratamento de dados e condições ideais dos dados, sobrando apenas uma semana para tratarmos a respeito dos algoritmos finais de classificação do derrame e de seus experimentos.

Apesar do que foi dito anteriormente, como o processo todo de *data science* envolvido neste trabalho envolveu muito a testagem dos dados com algoritmos preditivos, a classificação final não se diferenciou tanto daquelas usadas anteriormente. Desta forma, os únicos experimentos realizados para a classificação final foram referentes a alguns parâmetros dos algoritmos selecionados, visando encontrar aqueles mais próximos da otimalidade.

Com os testes e experimentos, foram obtidas uma média de acurácias de 77% para o *Support Vector Machine*, 73% para o *Multi Layer Perceptron*, e 76% para o algoritmo *Random Forest*.

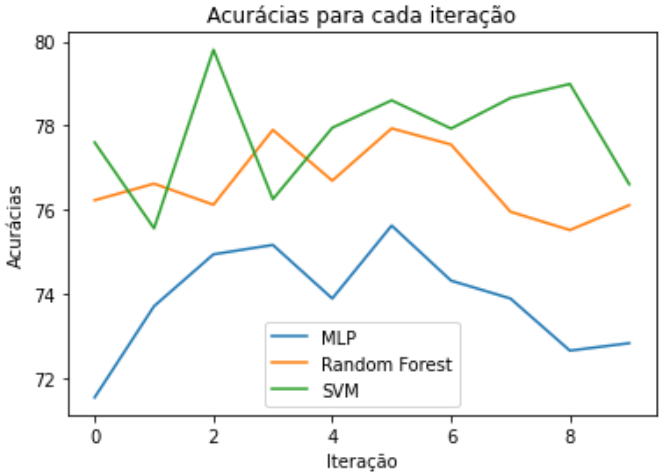

Fig. 6. Acurácias de cada algoritmo em diversas iterações.

VII. Conclusão

Para concluir este trabalho, temos que o uso de técnicas computacionais para extrair padrões de características presentes em indivíduos pode ajudar bastante o processo de predição de ocorrência ou não ocorrência do derrame cerebral em uma dada pessoa. Apesar de auxiliar, nota-se que é indispensável a supervisão de um profissional qualificado na área, pois ele será o responsável de realmente realizar o diagnóstico do paciente com determinadas características, utilizando técnicas como as aqui mostradas como uma ferramenta auxiliar apenas. Outra conclusão foi que, apesar de o *dataset* estudado conter diversas colunas de diferentes domínios, nem todas são de grande

importância, no que diz respeito aos modelos treinados, para um possível diagnóstico. A remoção dessas colunas não causaria um impacto considerável nas acurácias e ainda poderia tornar a base de dados mais enxuta, além de exigir menos informações pessoais dos pacientes.

VIII. Referências Bibliográficas

[1] Mohit. Baliyan. *Chest X Ray (Pneumonia) Diagnosis with CNN.* Disponível em: *https://www.kaggle.com/code/mohitbaliyan/chest-x-ray-pneumonia-diagnosis-with-cnn*

[2] *Ruthvik. PVS. Stroke Data | Analysis and Prediction.* Disponível em*:* https://www.kaggle.com/code/ruthvikpvs/stroke-data-analysis-and-prediction/notebook

[3] Joshua. Swords. *Predicting a Stroke.* Disponível em: https://www.kaggle.com/code/joshuaswords/predicting-a-stroke-shap-lime-explainer-eli5

[4] JP. *Comparing performance of 13 Classifiers.* Disponível em: https://www.kaggle.com/code/para24/comparing-performance-of-13-classifiers-f1-score

[5] Fahad. Mehfooz. *HeartAttack prediction with 91.8% Accuracy.* Disponível em: https://www.kaggle.com/code/fahadmehfoooz/heart attack-prediction-with-91-8-accuracy