Replicação do estudo de Deo e Panigrahi, 2019

Rodolfo Bolconte Donato
Pós-Graduação em Ciência da Computação
Universidade Federal de Campina Grande (UFCG)
Campina Grande, Brasil
rodolfo@copin.ufcg.edu.br

***Abstract*—A replicação é um método capaz de melhorar o entendimento e compreensão de um trabalho, com o propósito de aumentar a discussão sobre o tema trabalhado e também os aspectos utilizados nos trabalhos originais. Visando tais discussões o presente trabalho tem o propósito de realizar uma replicação de um trabalho original que avalia o desempenho de modelos baseados em aprendizagem de máquina para a predição de dados de pessoas com ou sem diabetes.**

***Index Terms*—Replicação de Estudo, Aprendizagem de Máquina, Análise de Dados, Diabetes**

I. Estudo original

O estudo se trata de uma avaliação de desempenho de modelos de classificação baseados em Aprendizagem de Máquina, em que os autores utilizam como fonte de dados, informações de saúde que possam ser importantes para a realização de um diagnóstico se uma pessoa tem diabetes ou não. Os objetivos principais do estudo giram em torno de definir modelos com um bom desempenho de classificação de dados para o cenário de pessoas com diabetes ou não, e para isto são realizadas comparações das previsões realizadas, postos em evidência utilizando métricas como acurácia, sensibilidade, especificidade, etc [1].

A Aprendizagem de Máquina é um subcampo da Engenharia e Computação, que evoluiu a partir do estudo de reconhecimento de padrões e do aprendizado computacional em inteligência artificial, sendo uma área que permite aos computadores a capacidade de aprender a realizar determinadas tarefas a partir da percepção de experiências aprendidas. A aprendizagem é implementada a partir da construção de algoritmos computacionais que podem aprender de seus erros e fazer previsões sobre determinados dados, por exemplo. Estes algoritmos operam construindo um modelo a partir de entradas de amostras de dados para que decisões guiadas por estes dados possam ser tomadas, ao invés de resultados serem apresentados apenas através de instruções pré-programadas [5].

Quanto às informações sobre as pessoas que tem ou não diabetes, ela foram retidas a partir do *National Health and Nutrition Examination Survey (NHANES)*, um programa de estudos elaborado para avaliar a saúde e o estado nutricional das pessoas nos Estados Unidos.

O trabalho original intitulado "Performance Assessment of Machine Learning Based Models for Diabetes Prediction" pode ser conferido através da publicação no IEEEXplore: https://ieeexplore.ieee.org/document/8962811.

II. Objeto da reprodução

A presente replicação visa realizar uma execução o mais fiel possível ao estudo original, em que foram retidos dados sobre pessoas dos Estados Unidos que possuem diabetes ou não, para a execução de modelos de classificação utilizando dois algoritmos de aprendizagem de máquina, sendo eles das famílias de *Support Vector Machine (SVM)* e *Bagged Trees*, em que cada algoritmo foi testado e comparado utilizando 2 tipos de reamostragem de dados, a *K-Fold* e *Holdout*.

Os resultados obtidos pelos autores do estudo original foram discutidos de acordo com uma comparação de métricas para as previsões dos algoritmos, estas separadas pelos dois tipos de reamostragem dos dados. As métricas utilizadas para a comparação dos resultados foram: Acurácias Mínima, Média e Máxima, Desvio Padrão, Curva *ROC*, Sensibilidade e Especificidade. Os resultados obtidos no estudo original podem ser vistos na Figura 1:

Fig. 1. Resultados Métricos da Execução dos Modelos do Estudo Original

Model performance using 5-fold cross-validation (test data)

	Mean(%)	SD	Min(%)	Max(%)	AUC	Sens*(%)	Spec*(%)
Bagged Trees	88.43	3.69	85.71	95.24	0.88	83.00	93.87
Linear SVM	90.81	3.69	88.09	97.62	0.91	83.00	98.64

Model performance using hold out validation (test data)

	Mean(%)	SD	Min(%)	Max(%)	AUC	Sens*(%)	Spec*(%)
Bagged Trees	88.09	4.03	83.33	95.24	0.88	85.71	90.46
SVM	90.82	3.69	88.10	97.62	0.91	83.00	98.64

*sens - sensitivity; spec - specificity

Com os resultados, os autores concluem que o modelo utilizando um algoritmo de *SVM* se mostra ligeiramente com o melhor desempenho em ambos os métodos de validação, que no caso é a execução para os dois tipos de reamostragem, com valores distantes para acurácia, desvio padrão e especificidade, porém com valores próximos para Curva *ROC*, e até com *Bagged Trees* atingindo maior valor para Sensibilidade quanto utilizada a reamostragem Holdout. Porém os autores reconhecem que ambos os modelos se mostraram eficientes na previsão dos dados do conjunto e sugerem novas validações utilizando conjuntos de dados maiores e distintos.

III. Metodologia original

A. Coleta de dados

O estudo original utiliza um conjunto de dados com informações de pessoas que possuem diabetes ou não do país dos Estados Unidos. Tal conjunto de dados é de origem do *NHANES*, um programa de estudos que realiza pesquisas através de questionários sobre diversos aspectos da população estadunidense com relação à saúde e doenças, que para o cenário do estudo, foram utilizadas informações relacionadas à diabetes e as próprias pessoas, como níveis de ingestão de álcool, taxas de colesterol e calorias, índice de massa corporal, entre outros.

Realizando a retenção dos dados, são obtidos 384 amostras de pessoas sem diabetes e 14 amostras de pessoas com diabetes, então devido a este desbalanceamento de classes, são executadas técnicas de *oversampling* e *undersampling* para aumentar e diminuir, respectivamente, a quantidade de dados para cada classe (não diabético e diabético), que resultam em 70 amostras para cada classe, sendo assim o conjunto de dados original do estudo possui 140 amostras.

O conjunto de dados é então utilizado na execução de dois tipos de reamostragem: 1) *K-Fold*, que consiste em dividir o conjunto de dados em *k* subconjuntos (o conjunto foi dividido em 5 subconjuntos no estudo) do mesmo tamanho em que cada um é utilizado para teste e os k-1 restantes são utilizados para treino [4]; e 2) *Holdout*, que consiste em dividir o conjunto de dados em três subconjuntos exclusivos, um para treinamento, outro para validação e o último para teste dos modelos, em que no estudo foram utilizados, 56%, 14% e 30% dos dados para cada subconjunto, respectivamente [4].

Cada reamostragem foi realizada para a execução de dois algoritmos: 1) *SVM*, que basicamente define uma linha de separação dos dados, chamada de hiperplano. Essa linha busca maximizar a distância entre os pontos mais próximos em relação a cada uma das classes [3]; e 2) *Bagged Trees*, modelo com o propósito de reduzir a variância de um método de aprendizado estatístico, construindo árvores de regressão para a realização de previsões de dados, em que depois é calculada a média para tais previsões [2].

B. Dados gerados

O conjunto de dados utilizado para a execução das reamostragens e também dos modelos de aprendizagem constituem em 140 amostras, sendo 70 amostras de pessoas com diabetes e 70 de pessoas sem diabetes, com 8 informações sobre aspectos pessoais e de saúde das pessoas, como níveis de colesterol e calorias, quantidade de álcool e cigarro ingeridas por média, entre outras informações. Uma amostra fictícia do Conjunto de Dados utilizado no estudo original pode ser conferida na Tabela I.

Além do conjunto de dados utilizado para a execução dos modelos, tem-se também os dados produzidos a partir das previsões dos modelos, no caso valores referentes à cálculos métricos com o propósito de avaliar o desempenho dos modelos. As métricas utilizadas são Acurácia Média (%), Desvio Padrão (SD), Acurácia Mínima (Min %), Acurácia Máxima (Max %), Curva ROC (AUC), Sensibilidade (Sens*%) e Especificidade (Spec*%). Uma amostra destes resultados pode ser conferida na Tabela II.

TABLE I
Amostra Fictícia do Conjunto de Dados do Estudo Original

Variávies	Pessoa 1	Pessoa 2	Pessoa 3
Body Mass Index	5	4	7
Cholesterol (HDL)	19	0	19
High Calorie Consumption	0	19	19
Blood Pressure	17	17	0
Smoking	10	0	20
Drinking	2	3	14
Overall Diet	13	13	13
Overall Fitness	16	0	0
Diabetic	0	1	0

TABLE II
Amostra Fictícia dos Resultados Métricos dos Modelos de Aprendizagem

Mean(%)	SD	Min(%)	Max(%)	AUC	Sens*(%)	Spec*(%)
88.43	3.69	85.71	95.24	0.88	83.00	93.87
90.81	3.69	88.09	97.62	0.91	83.00	98.64
88.09	4.03	83.33	95.24	0.88	85.71	90.46

C. Análise de Dados

Os autores utilizam apenas os valores das métricas para definir o melhor modelo de aprendizagem e também técnica de reamostragem adequada para a previsão dos dados utilizados, ou seja, as conclusões são definidas com base nos maiores valores para cada métrica estatística utilizada, não há um cálculo de intervalos de confiança, por exemplo, para uma melhor conclusão sobre qual modelo pode ser melhor de fato em relação à outro com base num pensamento de aplicação para uma população.

IV. Diferenças metodológicas com o estudo original

É esperado que a replicação possua semelhanças e também diferenças com relação ao estudo original discutido nas seções anteriores, uma vez que para alguns aspectos não é possível executar fielmente alguns passos devido a aleatoriedade computacional, além também da replicação buscar uma adequação ao solicitado pela proposta do trabalho. Tais semelhanças e diferenças são discutidas abaixo:

- População estudada: a população utilizada é proveniente do *NHANES*, um programa de pesquisa dos Estados Unidos conduzido pelo *National Center for Health Statistics*, com o propósito de avaliar a saúde e o estado nutricional de crianças e adultos no país e para acompanhar as mudanças ao longo do tempo. Com relação aos números reais, são 140 amostras de pessoas com ou sem diabetes, utilizando 8 informações sociais e de saúde sobre as pessoas. No estudo original alguns dos dados foram construídos de forma sintética, utilizando técnicas de *Oversampling*, porém a utilização desta técnica não é

realizada na replicação, uma vez que os dados disponibilizados pelo NHANES é bem maior que a quantidade utilizada. Sendo assim, os dados utilizados na replicação são diferentes quanto ao seus valores, mas obtidos da mesma fonte de dados e com as mesmas informações sobre as pessoas;
- Pergunta de pesquisa: a pergunta do estudo original gira em torno de responder qual o melhor modelo baseado em aprendizagem de máquina para a predição de pessoas com diabetes. Tal pergunta é mantida para a replicação do estudo, através da comparação de métricas de previsões dos modelos e também de inferência a partir de intervalos de confiança;
- Metodologia de coleta de dados: os dados são coletados através do site do programa *NHANES*, em que são coletados também para a replicação;
- Dados gerados: os resultados do estudo original são gerados utilizando o MATLAB, software para a realização de cálculos numéricos, porém para a replicação é utilizada a linguagem de programação R, devida uma melhor praticidade em relação ao autor da replicação e também por ser o instrumento utilizado na disciplina do presente trabalho;
- Metodologia de análise de dados: os autores realizam uma análise dos resultados obtidos apenas comparando valores maiores e menores das métricas entre as previsões dos modelos. Para a replicação é feita uma análise a partir de intervalos de confiança das métricas acurácia, sensibilidade e especificidade das previsões dos modelos utilizando a reamostragem *holdout*, que são elaborados dois tipos de intervalos de confiança, o primeiro com relação as métricas de cada modelo de previsão e o segundo é uma diferença das métricas entre os dois modelos. Tais intervalos são construídos a partir da técnica *bootstrap*, que para a replicação são criadas 2000 conjuntos de amostras baseadas nos dados originais para que os intervalos sejam calculados;
- Código da análise de dados: os autores não disponibilizam os códigos de execução do estudo, porém como a ferramenta utilizada na replicação não é a mesma do estudo original, a ausência dos códigos originais não causa problemas na execução da replicação.

V. Resultados da Replicação/Reanálise

Na replicação são mantidas as mesmas quantidades de dados utilizados na reamostragem *holdout* do estudo original, sendo 56% dos dados para o treino dos modelos (78 amostras), 14% para validação (20 amostras) e os 30% restantes dos dados para o teste dos modelos treinados e validados (42 amostras).

Como mostrado anteriormente, são utilizadas 8 variáveis explicativas para a variável de resposta que indica se uma pessoa é diabética ou não, sendo elas: níveis de hipertensão, caloria, colesterol, dieta, exercícios físicos, consumo de álcool e cigarro e também o índice de massa corporal da pessoa.

Com os dados organizados, a criação e treino dos modelos são realizados com as 78 amostras, em seguida é executada uma primeira previsão utilizando as 20 amostras para validação, em que o *Random Forest* obteve 55%, 70% e 40% de acurácia, sensibilidade e especificidade, respectivamente, enquanto o *Support Vector Machine (SVM)* obteve 65%, 50% e 80% para as mesmas métricas. Após a previsão dos dados de validação, é realizada então a previsão em cima dos 42 dados para teste, com o *Random Forest* atingindo 61,9%, 71,42% e 52,38%, enquanto o *SVM* atinge 50%, 42,85% e 57,17% de acurácia, sensibilidade e especificidade.

Se for levado em consideração somente a acurácia das previsões para os dados de validação como forma de definir um melhor modelo (como o trabalho original realizou tal observação), o *SVM* atinge o maior valor, podendo ser considerado como o melhor modelo para tais dados, porém o *Random Forest* atinge uma acurácia maior em relação ao *SVM* para os dados de teste, o que inverte o cenário de melhor modelo, com uma diferença de quase 12% da acurácia entre os modelos, enquanto o *SVM* tem uma diferença de acurácia para os dados de validação, em torno de 10% de diferença.

Para a construção dos intervalos de confiança, é executado o *bootstrap* com 2000 conjuntos de amostras a partir das previsões realizadas pelos modelos, para que possam ser calculados novamente acurácia, sensibilidade e especificidade. Na Figura 2 é possível verificar algumas diferenças para os dois modelos, em que o *Random Forest* tem o seu menor valor para o intervalo de confiança da especificidade, enquanto o maior valor pode ser atingido pela sensibilidade, já no *SVM* os pontos extremos se invertem, em que o menor valor de um intervalo pode ser atingido pela sensibilidade, enquanto o maior pela especificidade. Quanto a acurácia, o *Random Forest* apresenta um intervalo com a possibilidade de atingir valor entre 0,429 e 0,738, enquanto o *SVM* pode atingir entre 0,310 e 0,619 de acurácia. Se for levado em conta que 50% seja um valor razoável para as 3 métricas, não é possível afirmar que existe um modelo que seja o melhor capaz, uma vez que as métricas de cada um podem ser tanto acima como abaixo deste limiar, logo não é possível definir também qual o modelo com o maior efeito na previsão de dados de uma população.

Visto que não é possível definir o melhor modelo com base nos valores obtidos para as métricas de uma população, o mesmo também acontece quando comparadas as diferenças das métricas de acurácia e especificidade (entre os dois modelos), porém sensibilidade apresenta um efeito positivo do *Random Forest* em relação ao *SVM*. De acordo com a Figura 3 o Intervalo de Confiança das diferenças de acurácia (entre 0 e 0,238) e especificidade (entre -0,206 e 0,115), mostra que tanto o *Random Forest* pode ter um desempenho melhor que o *SVM* quanto o contrário, já para a sensibilidade (entre 0,097 e 0.476) é claro um efeito maior do *Random Forest*. Porém, levando em conta os intervalos de confiança da diferença das 3 métricas, não é possível definir qual seria o melhor modelo para a previsão de pessoas diabéticas ou não em uma população, apenas que ambos os modelos possuem desempenhos semelhantes, já que não é correto definir que um modelo é melhor que outro utilizando apenas uma métrica estatística, que neste caso poderia ser a sensibilidade.

Fig. 2. Intervalos de confiança das métricas para as previsões dos modelos *Random Forest* e SVM para as amostras geradas pelo Bootstrap.

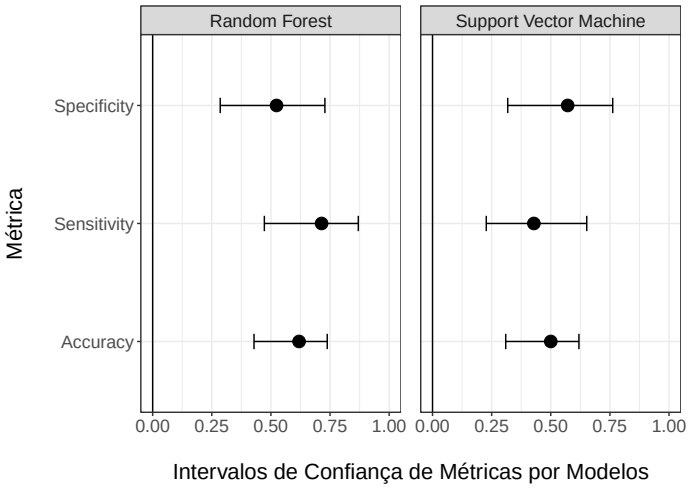
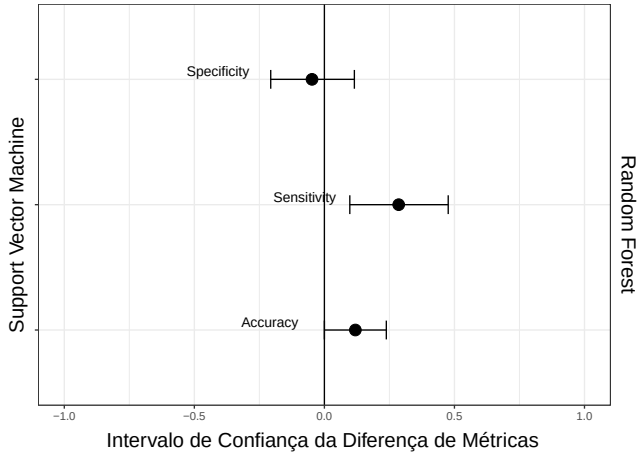


Fig. 3. Intervalos de confiança da diferença das métricas das previsões dos modelos *Random Forest* e *SVM* para as amostras geradas pelo Bootstrap.



VI. Replicação vs. estudo original

Comparando os resultados do estudo original com os resultados da replicação, a primeira diferença se encontra já nas acurácias dos modelos em relação aos dados de validação. No estudo original, os modelos atingem 90,2% e 87,2% de acurácia para o *Random Forest* e *SVM* respectivamente, uma grande diferença dos valores da replicação, 55% e 65%, tanto no tamanho dos valores quanto nos modelos que atingem, visto que o *Random Forest* atinge o maior valor no estudo original, enquanto na replicação o *SVM* atinge o maior valor.

Para os dados de teste, também há diferenças nos modelos. No estudo original o Random Forest 88,09%, 85,71% e 90,46%, enquanto o *SVM* atinge 90,82%, 83% e 98,64%, tais valores são para acurácia, sensibilidade e especificidade respectivamente, sendo assim é possível notar que o *SVM* atinge os maiores valores para acurácia e especificidade. Já na replicação o *SVM* atinge maior valor apenas na especificadade para os dados previstos, cerca de 57,17%, enquanto o *Random Forest* atinge os maiores valores de acurácia e sensibilidade, 61,90% e 71,42%. É importante evidencia também as diferenças de tais resultados, em que no estudo original, a maior diferença das 3 métricas entre os modelos é de 8,18% para a especificidade, sendo o *SVM* com o maior valor, enquanto nos resultados da replicação, a maior se dá na sensibilidade sendo o *Random Forest* 28,57% maior que o *SVM*.

Com relação aos intervalos de confiança, eles são calculados no estudo original porém não é descrita a forma de cálculo dos mesmos, apenas que são para as acurácias, logo não se tem informações sobre se foi utilizado o *bootstrap* com o objetivo de inferência para uma população ou se foram calculados apenas para os dados de teste. Os valores ficam entre 0,8333 e 0,9524 para o *Random Forest* e entre 0,881 e 0,9762 para o *SVM*, valores bem acima dos intervalos para a replicação visualizados na Figura 2. Uma observação a ser feita com relação aos intervalos de confiança das métricas dos modelos da replicação, é que os modelos podem ter um desempenho até pior que um "modelo burro", em que o mesmo seja capaz de classificar os dados em apenas um valor de classe, atingindo 50% de acurácia, por exemplo, que pode ser acima de fato dos valores atingidos pelos modelos da replicação.

Realizando discussões sobre tais diferenças apresentadas, a provável maior causa destas pode ser os dados, uma vez que no estudo original são realizadas técnicas de *oversampling* e *undersampling*, ou seja, a criação de dados sintéticos a partir dos dados originais e também desconsiderando determinados dados, logo, há uma alteração no conjunto de dados que não é feita na replicação, pois com base na quantidade dos dados disponibilizados pelo NHANES, não foi necessário realizar tais ajustes, apenas a coleta da mesma quantidade de dados de pessoas com diabetes ou não.

Outra possibilidade para a geração de diferenças consideráveis entre os dois estudos pode ser a configuração dos modelos de previsão, uma vez que não foram divulgados também as configurações de cada um no estudo original. Na replicação foram considerados valores padrão para o *SVM* e uma alteração apenas na quantidade de árvores para o *Random Forest*, sendo 50 árvores, já as demais configurações são mantidos os padrões da biblioteca.

VII. Material para replicação

O estudo original realiza todo o experimento utilizando o MATLAB, um software voltado para a realização de cálculos numéricos complexos, porém não são disponibilizados os códigos-fonte das execuções no artigo. Já a replicação é realizada utilizando – desde o tratamento dos dados até o cálculo dos intervalos de confiança – a linguagem de programação R voltada para a estatística. Os códigos tanto de tratamento quanto de execução de modelos e cálculo de métricas são disponibilizados no GitHub[1], além dos dados originais obtidos

[1]Link do Repositório: https://github.com/rodolfobolconte/replicacao-diabetes

no site do NHANES[2] que são disponibilizados em diversos conjuntos de dados.

REFERENCES

[1] R. Deo and S. Panigrahi. Performance assessment of machine learning based models for diabetes prediction. *2019 IEEE Healthcare Innovations and Point of Care Technologies*, 2019.

[2] Michael Foley. My data science notes, 2020.

[3] Wikipédia. Máquina de vetores de suporte — wikipédia, a enciclopédia livre, 2020. [Online; accessed 14-novembro-2020].

[4] Wikipédia. Validação cruzada — wikipédia, a enciclopédia livre, 2020. [Online; accessed 8-agosto-2020].

[5] Wikipédia. Aprendizado de máquina — wikipédia, a enciclopédia livre, 2021. [Online; accessed 7-maio-2021].

[2]Site do NHANES: https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015