Avaliação de técnicas de aprendizado de máquina supervisionado na classificação de obesidade

Diego de Almeida Miranda - Discente
Reginaldo Massanobu Kuroshu - Orientador

I. Resumo

Através de métodos de sequenciamento de nova geração (NGS - Next Generation Sequencing) podemos analisar e quantificar o material genético de comunidades microbiais presentes no intestino humano, com isso é possível observar com mais clareza o papel que a microbiota intestinal exerce na saúde. É sabido que indivíduos com obesidade apresentam a composição de bactérias alterada em relação ao microbioma intestinal de indivíduos magros. Este trabalho apresenta a aplicação de métodos de aprendizado de máquina supervisionados na classificação de indivíduos obesos através dos dados de composição do microbioma intestinal. De forma geral, os modelos *Random Forest* e *Gaussian Naive* Bayes se destacam, e as famílias *Lachnospiraceae*, *Ruminococcaceae* e *Bacteroidaceae* foram as mais frequentes dentre os atributos selecionados.

II. Introdução

A. O microbioma

O material genético de comunidades microbiais, formadas por fungos, bactérias, vírus e protozoários em diversas partes do corpo humano, é chamado de microbioma, ou microbiota, qu protagonizam diversos papeis na saúde e bem estar do indivíduo. Apenas no trato digestivo é possível que existam entre 10 e 100 trilhões de organismos que nos beneficiam de diversas maneiras [1].

Os métodos de sequenciamento de nova geração, do inglês next generation sequencing (NGS), tem permitido que os cientistas consigam realizar análises sob esses organismos. A abordagem mais comum para quantificar a composição do microbioma humano consiste em utilizar marcadores genéticos, mais especifiacamente a região 16S do RNA ribossomal, que é traduzida pelo sequenciamento do DNA para o estudo de composição dos mais diversos microbiomas. A região 16S contém cerca de 1500 pares de base, é um trecho em comum em todo o reino de bactérias que permite realizar a identificação de família e gênero, entretanto, tal região é pouco sensível para distinguir organismos a nível de espécie ou cepa. O cálculo de abundância de uma determinada unidade taxonômica operacional se dá a partir da contagem de leituras desta OTU numa amostra.

B. A obesidade

É natural que a composição de microrganismos varie entre populações ao redor do mundo, pois o microbioma está sujeito às condições na qual o indivíduo hospedeiro vivência, portanto o estilo de vida, a alimentação e o perfil da expressão gênica de cada indivíduo [2] pode também influenciar na ecologia microbial. Por exemplo, De Filippis [3] apresenta que uma alimentação mediterrânea, a base de vegetais, aumenta de forma notável os níveis da população do gênero de bactérias *Prevotella*.

Além disso, existem estudos que apontam alterações perceptíveis na ecologia microbial do intestino de pessoas que expressam alguma condição clínica, como é o caso da obesidade. De 2006 a 2019 a prevalência de obesidade no Brasil aumentou em 72%, saindo de 11,8% para 20,3% de obesos no país [4]. Num cenário global, segundo a Wold Health Organization (WWO), a prevalência de obesos quase triplicou desde 1975 [5]. Estamos encarando uma crise na saúde metabólica provocada por uma epidemia de obesidade. O sedentarismo e a facilidade de acesso a alimentos densamente energéticos levam a um consumo calórico excessivo quando comparado com as necessidades energéticas reais do indivíduo. Entretanto, diferenças entre a ecologia microbial podem ser um fator importante na homeostase energética. Em outros termos, indivíduos propensos a obesidade devem possuir comunidades microbiais do intestino características de sua condição. A mudança na composição bacterial pode não ser apenas uma consequência da obesidade, mas pode sim possuir um papel fundamental na obesidade patogênica. Ley [6] estudou mais de 5 mil bactérias, identificadas pela região 16S, encontradas no intestino de camundongos obesos e notou que os indivíduos obesos tiveram uma redução de 50% na abundância de *Bacteroides* e um aumento proporcional de *Firmicutes*, quando comparados com a microbiota intestinal de camundongos saudáveis. No ano seguinte [7] são apresentados resultados semelhantes na análise da microbiota humana.

O aumento de *Lachnospiraceae* e da família *Prevotellaceae* também se mostram fortemente relacionadas com a condição de obesidade e tolerância a glucose, ao menos na Coorte hispânica do condado de Cameron, como apresentado por Ross [8].

C. O volume de dados

Um único indivíduo pode ser fonte de um imenso volume de dados, isto graças aos métodos de NGS. Estes dados são chamados em inglês de multi-view data por possuírem informações heterogêneas complementares que ajudam a descrever um sistema ou um fenômeno biológico. Isto posto, um médico sozinho não conseguiria interpretar tamanho volume de informação, e com isso se faz necessário o uso de métodos

computacionais eficientes de compreender esse conteúdo. Os métodos atuais de aprendizado de máquina tem se mostrado uma ferramenta ideal para possibilitar análises mais completas, por serem particularmente uteis em reconhecer padrões em dados de alta dimensionalidade, gerando insights mais profundos sobre os sistemas biológicos analisados. Em especial modelos de aprendizado de máquina supervisionado podem nos ajudar a desenvolver modelos preditivos treinados a partir dos dados de abundância de organismos na microbiota intestinal, como é discutido em [9].

III. Trabalhos Relacionados

Em [10] é apresentada uma ferramenta computacional, MetAML de classificação baseado em meta-genômica e avaliação do potencial de associação entre microbioma e fenótipo. Esta ferramenta utiliza de classificadores por aprendizado de máquina, seleção de atributos, validação cruzada e usa como recurso perfis quantitativos de microbiota, como abundâncias relativas em nível de espécie e presença de marcadores específicos de cepas.

O software desenvolvido foi aplicado em um total de 2424 amostras metagenômicas publicamente disponíveis em oito estudos de larga escala. Todas as amostras foram processadas com a ferramente MetaPhlAn2[11] para quantificar o perfil de espécies e subespécies após o pré-processamento de dados de sequenciamento padrão.

Foram considerados seis conjuntos de associação metagenômicas de doenças, abrangendo um total de cinco doenças: cirrose hepática, câncer colorretal, doenças inflamatórias intestinais, obesidade e diabetes tipo dois (provenientes de dois estudos distintos). Cada dataset analisado de forma independente a partir de métodos de validação cruzada, que usa repetidamente subconjuntos da amostra com um fenótipo conhecido para treinar um modelo estatístico. Os classificadores baseados em *Support Vector Machines* (SVM) e *Random Forest* (RF) foram usados para esta avaliação.

Em [12] é feita a avaliação dos modelos RF, Extreme Gradient Boost (XGB), ENET, SVM na classificação de 29 conjuntos de dados de microbioma humano. Desses 29, 4 são referentes a indivíduos obesos. Aqui também são utilizadas técnicas validação cruzada, otimização de hiperparâmetros e seleção de atributos. Na seleção de atributos foram comparadas os 20, 50 e 100 principais atributos selecionados pelos modelos, onde foi analisado que mais de 25% dos atributos são cconsiderados importantes de forma simultânea pelos modelos.

IV. Métodos

A. Base de Dados

Os dados foram obtidos através do conjunto de dados MicrobiomeHD, com as tabelas de abundância absoluta já processadas. Cruzando as amostras das tabelas com os metadados de cada conjunto removemos as amostras que possuíssem valores NaN em algum dos atributos, e também amostras que possuíssem "sobre peso", i.e, $25 < \text{IMC} < 30$.

Conjunto de dados	ref	nº features	IMC ≥ 30	IMC ≤ 25
Escobar	[13]	2520	10	10
Ross	[8]	6406	6	37
Jumpertz	[14]	12775	35	26
Wu	[15]	28041	39	5
Gordon	[16]	20236	196	61
Goodrich	[17]	72256	193	451
Zupancic	[18]	105998	128	127

B. Os Modelos

Foram selecionados cinco modelos de aprendizado supervisionados de máquina para serem avaliados na tarefa de classificação dos indivíduos. Os modelos Gaussian Naive Bayes (NB) e Gaussian Process utilizam da pressuposição de que os dados possuem uma distribuição Gaussiana. Dentre estes modelos selecionados, quatro deles utilizam árvores de decisão, sendo eles Decision Tree(DT), Random Forest(RF), além dos outros dois modelos que também utilizam técnicas de boosting, como Adaptative Boosting(ADA) e Extreme Gradient Boosting(XGB). Além dos já citados anteriormente, há os modelos baseados em rede Peceptron e Multi Layer Perceptron(MLP). Por fim, são também utilizados os modelos Support Vector Machine(SVM), que utiliza de hiperplanos, e K-Nearest Neighbors(KNN), um algoritmo de classificação não paramétrico.

Todos os modelos de aprendizado de máquina foram obtidos através da linguagem de programação Python utilizando a biblioteca Scikit-learn [19], exceto o modelo XGBoost que possui sua própria biblioteca[20].

C. Otimização de parâmetros

Cada modelo possui seus próprios hiperparâmetros que controlam o processo de aprendizagem. Em geral, pode-se utilizar modelos com ou sem a otimização de parâmetros. Com o intuito de obter uma otimização dos resultados de acurácia - e outras métricas de avaliação -, foi realizada a otimização de hiperparâmetros através de uma busca exaustiva. Afim de reduzir o sobre-ajuste dos modelos a otimização de parâmetros utilizou uma parcela de um terço do conjunto de dados disponíveis, o qual não foi utilizado na validação cruzada dos modelos. Para cada modelo segue os hiperparâmetros disponíveis para a busca:

1) **Support Vector Machines**
   Regularização: 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13} e 2^{15};
   Kernel: sigmoid, poly, rbf e linear;
   Gamma: 2^{3}, 2^{1}, 2^{-1}, 2^{-3}, 2^{-5}, 2^{-7}, 2^{-9}, 2^{-11}, 2^{-13} e 2^{-15};
2) **Decision Tree**
   Critério: gini e entropia;
   Splitter: Melhor e Aleatório;
3) **Random Forest**
   Estimadores: 10, 50, 100, 200 e 500;
   Critério: gini e entropia;
4) **AdaBoost**
   Estimadores: 10, 50, 100, 200 e 500;

Taxa de aprendizado: 0.25, 0.5, 1, 1.5 e 2;

5) **Extreme Gradient Boosting**
   Profundidade máxima: 4, 6, 8, 10, 100 e 1000;
   Taxa de aprendizado: 0.001 e 0.01;
   Subamostra: 0.5, 0.75 e 1;
   Subamostra de coluna por árvore (*colsample_bytree*): 0.4, 0.6, 0.8 e 1.0;
6) **K-Nearest Neighbors**
   K: 1, 2, 3, 5 e 7;
   Pesos: uniforme e distância;
   Algoritmos: padrão, ball_tree, kd_tree e bruto;
   Parâmetro de potência para a métrica Minkowski(*p*): 1 e 2;
7) **Perceptron**
   Penalidade: l1, l2 e elasticnet;
   Alpha: 2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, 2^{7}, 2^{9}, 2^{11}, 2^{13} e 2^{15};
8) **Multilayer perceptron**
   Camadas ocultas: (50, 50, 50), (50, 100, 50) e (100,);
   Solucionador: sgd, adam e lbfgs;
9) **Gaussian Naive Bayes**
   Não possui hiperparâmetros a serem otimizados.
10) **Gaussian Process**
    Kernel: rbf, dot product, matern, rational quadratic e white kernel;

D. Validação Cruzada e Avaliação

Adotamos o método de validação cruzada *K-fold*, o qual consiste em dividir o conjunto de dados em *K* partições de forma que *K*-1 subconjuntos são usados para treino e um subconjunto para teste. Os treinos são repetidos *K* vezes para que todos os *folds* tenham oportunidade de serem usados para treino e teste. Dessa forma, a acurácia do modelo é dada pela acurácia média das *K* execuções. Como temos um conjunto de dados não equilibrados, i.e, temos uma quantidade significativamente maior de uma classe que de outra, usaremos o *Stratified K-fold*, que preserva em cada subconjunto a proporção de elementos das classes. Em outras palavras, em um conjunto de dados que possui um terço das amostras classificadas como A e o restante como B, o *Stratified K-fold* irá garantir que em cada *fold* seja composto de um terço de amostras da classe A.

Para avaliar o desempenho geral dos modelos serão observadas três principais métricas de avaliação: AUC, acurácia e pontuação F1.

Uma forma de avaliar a qualidade do modelo é utilizando a curva ROC e calculando a área debaixo da curva, ou *area under the curve* (AUC).

ROC é uma curva de probabilidade, enquanto AUC é representa o grau ou medida se separação. Quando temos o valor de AUC = 0.5, isso significa que o modelo não tem capacidade de classificar, enquanto AUC = 1.0 representa que o modelo faz cassificações perfeitas.

A curva ROC tem como eixo x o valor de *False Positive Rate* (FPR), dado por

$$FPR = \frac{FalsoPositivo}{VerdadeiroNegativo + FalsoPositivo}$$

já no eixo y temos o valor de *True Positive Rate* (TPR), obtido a partir de

$$TPR = \frac{VerdadeiroPositivo}{VerdadeiroPositivo + FalsoNegativo}$$

A acurácia é uma métrica simples que nos permite avaliar a qualidade da classificação de uma modelo. o valor da acurácia varia de 0 a 1, de forma que 1 representa que o modelo acertou em todas as predições e 0 o oposto. Dessa forma, temos

$$\text{acurácia} = \frac{VerdadeiroPositivo + VerdadeiroNegativo}{|amostras|}$$

A pontuação F1 é uma média harmônica entre a revocação e a precisão. Dessa forma, quando temos um F1 baixo, isso significa que temos uma precisão ou revocação baixa.

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

O resultado final da curva ROC e da acurácia são obtidos a partir da média de 5 execuções completas da validação cruzada $K - fold$ para que dessa forma tenhamos um resultado mais preciso do real desempenho do modelo.

E. Seleção de atributos

A seleção de atributos nos permite encontrar quais das OTUs encontradas nas amostras são mais significativas na classificação daquele conjunto de dados. Encontrar um subconjunto de atributos que ajude a otimizar o desempenho dos modelos de classificação nos permite identificar quais os organismos mais relevantes na condição clínica de obesidade.

Os modelos baseados em árvore fornecem um valor de importância para cada atributo baseado no critério de divisão na construção das árvores. Em geral, a soma das importâncias é de 1.0, assim, cada atributo possui uma importância que varia de 0 a 1. Com isso, podemos selecionar aquelas *feature*s mais importantes para a classificação e analisar se os modelos obtiveram um ganho de desempenho.

Os modelos baseados em árvore foram utilizados em conjunto para encontrar os atributos que são comumente mais relevantes. Portanto, para cada um dos modelos DT, RF, XGB e ADA, foram selecionados os atributos com uma importância maior que zero em pelo menos dois desses modelos.

V. Resultados

A. O desempenho

Para a análise mais completa do impacto da otimização de hiperparâmetros e da seleção de atributos, foram geradas tabelas com o desempenho dos modelos com e sem a otimização de hiperparâmetros, e com e sem a seleção de atributos. Como a quantidade de tabelas de resultantes da avaliação é muito grande, serão apresentadas a seguir algumas poucas que exemplificam o desempenho geral dos modelos nos possíveis cenários.

Na Tabela I podemos ver o desempenho dos modelos a partir da métrica f1 sem a otimização de hiperparâmetros e sem a seleção de atributos.

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.4	**0.92**	0.07	0.00	0.79	0.00	0.36
DT	0.55	0.89	0.51	0.2	0.72	0.26	0.61
RF	0.58	0.91	0.36	0.00	**0.80**	0.07	**0.65**
ADA	0.59	0.89	0.55	**0.06**	0.76	**0.28**	0.59
XGB	0.71	0.91	0.52	0.00	0.75	0.15	0.47
KNN	0.69	0.91	0.34	0.00	0.77	0.27	0.64
Perceptron	0.49	0.88	0.58	0.05	0.71	0.23	0.55
MLP	0.62	0.90	**0.64**	0.00	0.72	0.20	0.52
NB	**0.88**	0.91	0.61	0.00	0.68	0.24	0.56
GP	0.00	0.00	0.00	0.00	0.00	0.00	0.03

Tabela I
MÉTRICA F1 DE MODELOS NÃO OTIMIZADOS

É notável que o desempenho geral dos modelos nos conjuntos de dados [15] e [17] é demasiadamente baixo. No conjunto de dados [15] apenas 3 dos 10 obtiveram F1 diferente de 0, sendo o desempenho máximo para esse conjunto de dados F1 de 0.06 obtido pelo modelo ADA. Isso pode ter sido ocasionado pela pouca quantidade de amostras com IMC menor ou igual a 25. Algo que fortalece esta percepção se dá pelo fato de que o conjunto de dados [8] foi o melhor classificado pelos modelos e, em contraste ao [15], possui uma proporção similar, mas inversa, isto é, poucas amostras de IMC maior ou igual a 30.

Na Tabela II apresentado o desempenho em acurácia dos modelos com otimização de hiperparâmetros, mas sem seleção de atributos. Em destaque estão os melhores desempenhos para aquele conjunto de dados.

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.59	**0.86**	0.61	0.78	0.62	0.54	0.49
DT	0.59	0.81	0.61	0.86	0.60	0.57	0.54
RF	0.58	**0.86**	0.66	**0.9**	**0.68**	**0.68**	0.51
ADA	0.54	0.82	0.63	0.88	0.62	0.62	0.51
XGB	0.53	**0.86**	**0.68**	**0.9**	0.62	0.66	0.50
KNN	0.57	0.82	0.66	0.77	0.56	0.49	0.50
Perceptron	0.72	0.83	0.59	0.84	0.55	0.55	0.50
MLP	0.62	0.84	**0.68**	0.8	0.66	0.60	**0.56**
NB	**0.86**	0.83	0.66	**0.9**	0.59	0.61	**0.56**
GP	0.65	0.2	0.60	**0.9**	0.58	0.56	0.52

Tabela II
ACURÁCIA DE MODELOS OTIMIZADOS

Podemos notar que, para os conjunto de dados com mais amostras, o modelo RF é o que possui o melhor desempenho geral, atingindo até 0.68 de acurácia. Enquanto para conjuntos de dados com amostras mais reduzidas é possível obter uma acurácia de até 0.9 com o modelo RF. No geral é possível observar que para os os modelos com conjunto de dados reduzidos é mais frequente que mais modelos alcancem um desempenho máximo similar. Para o conjunto de dados [15] os modelos RF, XGB, NB e GP obtiveram o mesmo desempenho. Em sequência ao RF, o modelo que mais apresenta o melhor desempenho de acurácia é o XGB.

Agora, podemos ver o desempenho de AUC dos modelos com otimização de hiperparâmetros e com seleção de atributos realizada. Veja a Tabela III.

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.76	0.5	0.52	0.44	0.5	0.48	0.49
DT	0.74	0.48	0.55	0.42	0.48	0.50	0.49
RF	0.65	**0.52**	0.61	0.48	0.48	0.49	0.49
ADA	0.61	0.49	0.59	0.45	0.48	0.48	0.49
XGB	0.46	0.5	0.60	**0.5**	0.5	0.49	0.5
KNN	0.74	0.47	0.51	0.44	**0.54**	**0.51**	**0.50**
Perceptron	0.52	0.5	0.50	0.52	0.5	0.48	0.5
MLP	0.75	0.47	0.51	0.48	0.47	0.49	0.49
NB	**0.8**	0.46	0.58	**0.5**	0.51	0.45	0.49
GP	0.66	0.47	**0.64**	0.49	0.48	0.49	**0.50**

Tabela III
MÉTRICA AUC DE MODELOS OTIMIZADOS COM SELEÇÃO DE ATRIBUTOS

B. Atributos mais frequentes

De forma geral, a tabela a seguir apresenta o número de variáveis que foram selecionadas em cada um dos conjuntos de dados:

Conjunto de dados	a principio	selecionadas
Escobar	2520	14
Ross	6406	1
Jumpertz	12775	38
Wu	28041	39
Gordon	20236	3
Goodrich	72256	68
Zupancic	105998	15

É notável a redução abrupta de todos os conjuntos de atributos. Dessas poucas variáveis, foram realizadas contagens da frequência de *filo*, *família* e *gênero*. Os dois *filos* mais relevantes apresentados pela seleção de atributos são *Firmicutes* e, em seguida, *Bacteroidetes*. As famílias *Lachnospiraceae*, *Ruminococcaceae* e *Bacteroidaceae* foram as mais frequentes entre os atributos selecionados. Por fim, em nível de gênero, os atributos mais frequentes foram aqueles que não possuem classificação, mas entre os filogenéticamente identificados, os mais frequentes foram *Bacteroides*, *Alistipes*, *Faecalibacterium* e *Lachnospiracea incertae sedis*.

VI. CONCLUSÕES

Vizualindo os resultados obtidos, podemos notar que o desempenho médio geral dos modelos é relativamente baixo, principalmente nos maiores conjuntos de dados. É possível notar, também, uma tendência a um resultado vicioso entre os modelos em alguns casos, como podemos ver nos resultados obtidos pela métrica F1 no conjunto de dados [15] na Tabela I. Isso se dá, provavelmente, pela baixa quantia de amostras e pela alta similaridade entre elas. O volume de atributos presentes em cada conjunto de dados torna a tarefa de classificação ainda mais difícil, uma vez que selecionar os melhores atributos se torna uma tarefa complicada. Utilizando o método de seleção de atributos proposto não foi possível notar um ganho de desempenho significativo no processo de classificação.

Ainda assim, é possível averiguar que os modelos de aprendizado de máquina RF e XGB apresentam um melhor desempenho geral com todos os atributos preservados. Uma

vez aplicada a seleção, o modelo KNN passa a apresentar melhor desempenho nas métricas avaliadas.

A redução brusca do número de variáveis em cada conjunto de dados pode ser indicio de um sobre-ajuste, tornando o desempenho dos modelos ainda menos confiável. Ainda assim, foi possível notar a similaridade entre os principais atributos escolhidos e as OTUs relevantes já apresentadas na literatura. Destacam-se, principalmente, as famílias *Lachnospiraceae*, *Ruminococcaceae* e *Bacteroidaceae*.

Faz-se necessário que sejam propostos novos métodos de seleção de atributos para que então possamos realizar a tarefa de classificação de maneira mais eficiente. A construção de conjuntos de dados menos homogêneos também parece positivo para a investigação do problema.

REFERÊNCIAS

[1] Bäckhed F, Ley RE, Sonnenburg JL, Peterson DA, Gordon JI. Host-bacterial mutualism in the human intestine. Science. 2005 Mar 25;307(5717):1915-20. doi: 10.1126/science.1104816. PMID: 15790844.

[2] Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. Genetic Determinants of the Gut Microbiome in UK Twins. Cell Host Microbe. 2016 May 11;19(5):731-43. doi: 10.1016/j.chom.2016.04.017. PMID: 27173935; PMCID: PMC4915943.

[3] De Filippis F, Pellegrini N, Vannini L, Jeffery IB, La Storia A, Laghi L, Serrazanetti DI, Di Cagno R, Ferrocino I, Lazzi C, Turroni S. High-level adherence to a Mediterranean diet beneficially impacts the gut microbiota and associated metabolome. Gut. 2016 Nov 1;65(11):1812-21.

[4] Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Boletim Epidemiológico. Vigitel Brasil 2019: principais resultados. Vol. 51, nº16; Abril 2020.

[5] World Health Organization. Obesity and overweight. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.

[6] Ley RE, Bäckhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. Obesity alters gut microbial ecology. Proc Natl Acad Sci U S A. 2005 Aug 2;102(31):11070-5. doi: 10.1073/pnas.0504978102. Epub 2005 Jul 20. PMID: 16033867; PMCID: PMC1176910.

[7] Ley RE, Turnbaugh PJ, Klein S, Gordon JI. Human gut microbes associated with obesity. nature. 2006 Dec;444(7122):1022-3.

[8] Ross, M.C., Muzny, D.M., McCormick, J.B. et al. 16S gut community of the Cameron County Hispanic Cohort. Microbiome 3, 7 (2015). https://doi.org/10.1186/s40168-015-0072-y

[9] Knights D, Costello EK, Knight R. Supervised classification of human microbiota. FEMS microbiology reviews. 2011 Mar 1;35(2):343-59.

[10] Pasolli E, Truong DT, Malik F, Waldron L,Segata N (2016) Machine Learning Meta-analysis ofLarge Metagenomic Datasets: Tools and Biological Insights. PLoS Comput Biol 12(7): e1004977.doi:10.1371/journal.pcbi.1004977.

[11] Truong DT, Franzosa E, Tickle T, Scholz M, Weingart U, Pasolli E, et al. MetaPhlAn2 for enhancedmetagenomic taxonomic profiling. Nature Methods 2015; 12(10):902–903. doi:10.1038/nmeth.3589PMID:26418763

[12] Wang XW, Liu YY. Comparative study of classifiers for human microbiome data. Medicine in Microecology. 2020 Jun 1;4:100013.

[13] Escobar JS, Klotz B, Valdes BE, Agudelo GM. The gut microbiota of Colombians differs from that of Americans, Europeans and Asians. BMC microbiology. 2014 Dec;14(1):1-4.

[14] Jumpertz R, Le DS, Turnbaugh PJ, Trinidad C, Bogardus C, Gordon JI, Krakoff J. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. The American journal of clinical nutrition. 2011 Jul 1;94(1):58-65.

[15] Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R. Linking long-term dietary patterns with gut microbial enterotypes. Science. 2011 Oct 7;334(6052):105-8.

[16] Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. A core gut microbiome in obese and lean twins. Nature. 2009 Jan 22;457(7228):480-4. doi: 10.1038/nature07540. Epub 2008 Nov 30. PMID: 19043404; PMCID: PMC2677729.

[17] Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD. Human genetics shape the gut microbiome. Cell. 2014 Nov 6;159(4):789-99.

[18] Zupancic ML, Cantarel BL, Liu Z, Drabek EF, Ryan KA, et al. (2012) Analysis of the Gut Microbiota in the Old Order Amish and Its Relation to the Metabolic Syndrome. PLoS ONE 7(8): e43052. doi:10.1371/journal.pone.0043052

[19] Scikit-learn: Machine Learning in Python.Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E.; Journal of Machine Learning Research, vol 12; pages 2825–2830; 2011.

[20] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).

Tabelas complementares

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.59	**0.86**	0.61	0.78	0.62	0.54	0.49
DT	0.59	0.81	0.61	0.86	0.60	0.57	0.54
RF	0.58	**0.86**	0.66	**0.9**	**0.68**	**0.68**	0.51
ADA	0.54	0.82	0.63	0.88	0.62	0.62	0.51
XGB	0.53	**0.86**	**0.68**	**0.9**	0.62	0.66	0.50
KNN	0.57	0.82	0.66	0.77	0.56	0.49	0.50
Perceptron	0.72	0.83	0.59	0.84	0.55	0.55	0.50
MLP	0.62	0.84	**0.68**	0.8	0.66	0.60	**0.56**
NB	**0.86**	0.83	0.66	0.9	0.59	0.61	**0.56**
GP	0.65	0.2	0.60	0.9	0.58	0.56	0.52

Tabela 1: Acurácia de modelos otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.72	**0.92**	0.50	0.0	0.72	0.29	0.21
DT	0.57	0.89	0.53	**0.13**	0.71	0.27	0.65
RF	0.53	**0.92**	0.39	0.0	**0.79**	0.07	**0.66**
ADA	0.51	0.90	0.55	**0.13**	0.73	0.24	0.57
XGB	0.40	**0.92**	0.47	0.0	0.72	0.12	0.56
KNN	0.70	0.90	0.54	0.02	0.55	**0.34**	0.64
Perceptron	0.63	0.88	0.40	0.0	0.58	0.26	0.04
MLP	0.62	0.91	0.58	0.04	0.75	0.17	0.54
NB	**0.90**	0.90	**0.60**	0.0	0.71	0.26	0.55
GP	0.76	0.13	0.33	0.00	0.62	0.25	0.43

Tabela 2: Métrica f1 de modelos otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.59	0.5	0.59	0.43	0.55	0.48	0.49
DT	0.59	0.46	0.60	0.54	0.53	0.48	0.54
RF	0.58	0.5	0.62	0.5	0.55	**0.50**	0.51
ADA	0.54	0.47	0.63	**0.55**	0.52	**0.50**	0.51
XGB	0.53	0.5	0.64	0.5	0.53	0.49	0.50
KNN	0.57	0.47	0.65	0.45	**0.61**	0.48	0.50
Perceptron	0.72	0.5	0.57	0.47	0.55	0.48	0.50
MLP	0.62	0.48	**0.66**	0.47	0.6	0.47	**0.56**
NB	**0.86**	0.47	**0.66**	0.5	0.5	**0.50**	**0.56**
GP	0.65	**0.53**	0.56	0.5	0.58	0.47	0.52

Tabela 3: Métrica AUC de modelos otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.76	**0.86**	0.53	0.79	0.66	**0.66**	0.49
DT	0.74	0.81	0.56	0.77	0.63	0.57	0.49
RF	0.65	0.80	0.63	0.88	0.63	0.68	0.49
ADA	0.61	0.74	0.60	0.77	0.64	0.58	0.49
XGB	0.46	**0.86**	0.63	**0.9**	**0.66**	0.65	0.5
KNN	0.74	0.78	0.50	0.8	0.47	0.60	**0.50**
Perceptron	0.52	0.83	0.51	0.79	0.45	0.57	0.5
MLP	0.75	0.74	0.51	0.88	0.62	0.55	0.49
NB	**0.8**	0.79	0.57	**0.9**	0.37	0.40	0.49
GP	0.66	0.82	**0.65**	0.88	0.64	**0.68**	**0.50**

Tabela 4: Acurácia de modelos otimizados com seleção de atributos

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.74	**0.92**	0.44	0.0	**0.8**	0.05	0.58
DT	0.74	0.89	0.46	0.0	0.77	0.30	**0.66**
RF	0.62	0.88	0.47	0.0	0.77	0.04	**0.66**
ADA	0.56	0.84	0.52	0.04	0.77	0.25	**0.66**
XGB	0.24	**0.92**	0.43	0.0	**0.8**	0.11	0.37
KNN	**0.79**	0.87	0.48	0.0	0.37	0.31	0.40
Perceptron	0.41	0.88	0.35	**0.07**	0.28	0.23	0.10
MLP	0.76	0.83	0.44	0.0	0.76	0.32	0.29
NB	0.76	0.88	0.56	0.0	0.14	0.36	**0.66**
GP	0.75	0.89	**0.57**	0.0	0.78	0.03	**0.66**

Tabela 5: Métrica f1 de modelos otimizados com seleção de atributos

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.76	0.5	0.52	0.44	0.5	0.48	0.49
DT	0.74	0.48	0.55	0.42	0.48	0.50	0.49
RF	0.65	**0.52**	0.61	0.48	0.48	0.49	0.49
ADA	0.61	0.49	0.59	0.45	0.48	0.48	0.49
XGB	0.46	0.5	0.60	**0.5**	0.5	0.49	0.5
KNN	0.74	0.47	0.51	0.44	**0.54**	**0.51**	**0.50**
Perceptron	0.52	0.5	0.50	0.52	0.5	0.48	0.5
MLP	0.75	0.47	0.51	0.48	0.47	0.49	0.49
NB	**0.8**	0.46	0.58	**0.5**	0.51	0.45	0.49
GP	0.66	0.47	**0.64**	0.49	0.48	0.49	**0.50**

Tabela 6: Métrica AUC de modelos otimizados com seleção de atributos

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.69	**0.86**	0.59	**0.9**	0.65	**0.7**	0.45
DT	0.6	0.82	0.60	0.89	0.60	0.55	0.51
RF	0.57	0.85	0.62	**0.9**	**0.69**	0.65	0.51
ADA	0.57	0.81	0.63	0.86	0.65	0.60	0.52
XGB	0.66	0.85	0.65	**0.9**	0.64	0.65	0.50
KNN	0.54	0.85	0.61	**0.9**	0.66	0.6	0.50
Perceptron	0.68	0.79	0.65	0.72	0.63	0.58	0.55
MLP	0.67	0.82	**0.69**	0.86	0.59	0.59	0.57
NB	**0.84**	0.84	0.66	**0.9**	0.55	0.61	**0.58**
GP	0.5	0.13	0.57	**0.9**	0.33	**0.7**	0.50

Tabela 7: Acurácia de modelos não otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.4	**0.92**	0.07	0.00	0.79	0.00	0.36
DT	0.55	0.89	0.51	0.2	0.72	0.26	0.61
RF	0.58	0.91	0.36	0.00	**0.80**	0.07	**0.65**
ADA	0.59	0.89	0.55	**0.06**	0.76	**0.28**	0.59
XGB	0.71	0.91	0.52	0.00	0.75	0.15	0.47
KNN	0.69	0.91	0.34	0.00	0.77	0.27	0.64
Perceptron	0.49	0.88	0.58	0.05	0.71	0.23	0.55
MLP	0.62	0.90	**0.64**	0.00	0.72	0.20	0.52
NB	**0.88**	0.91	0.61	0.00	0.68	0.24	0.56
GP	0.00	0.00	0.00	0.00	0.00	0.00	0.03

Tabela 8: Métrica f1 de modelos não otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.69	0.5	0.52	0.5	0.49	0.5	0.45
DT	0.6	**0.51**	0.59	0.47	0.51	0.47	0.51
RF	0.57	0.49	0.59	0.5	0.56	0.48	0.51
ADA	0.57	0.46	0.62	**0.50**	0.56	**0.50**	0.52
XGB	0.66	0.49	0.63	0.5	0.54	**0.50**	0.50
KNN	0.54	0.49	0.57	0.5	0.55	**0.50**	0.50
Perceptron	0.68	0.48	0.65	0.46	**0.60**	0.48	0.55
MLP	0.67	0.47	**0.69**	0.47	0.48	0.47	0.57
NB	**0.84**	0.48	0.66	0.5	0.45	0.49	**0.58**
GP	0.5	0.5	0.5	0.5	0.5	0.5	0.50

Tabela 9: Métrica AUC de modelos não otimizados

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.68	**0.86**	0.57	**0.9**	0.65	**0.7**	0.49
DT	0.66	0.77	0.57	0.81	0.64	0.55	0.49
RF	0.73	0.77	0.62	0.89	0.63	0.67	0.49
ADA	0.64	0.78	0.61	0.8	0.64	0.60	0.49
XGB	0.36	0.84	0.59	**0.9**	**0.66**	0.61	0.5
KNN	0.53	**0.86**	**0.64**	**0.9**	**0.66**	0.62	0.50
Perceptron	0.49	0.62	0.56	0.82	0.51	0.58	0.49
MLP	**0.8**	0.65	0.53	0.85	0.64	0.55	**0.51**
NB	0.75	0.79	0.60	**0.9**	0.37	0.41	0.49
GP	0.71	0.65	0.56	0.86	0.61	**0.7**	0.50

Tabela 10: Acurácia de modelos não otimizados com seleção de atributos

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.45	**0.92**	0.00	0.0	0.78	0.00	0.60
DT	0.59	0.87	0.49	0.0	0.77	0.28	**0.66**
RF	0.71	0.86	0.50	0.0	0.77	0.12	0.60
ADA	0.68	0.87	0.53	0.09	0.78	0.24	**0.66**
XGB	0.29	0.91	0.48	0.0	**0.8**	0.18	0.00
KNN	0.59	**0.92**	**0.59**	0.0	**0.8**	0.29	0.41
Perceptron	0.31	0.66	0.36	**0.11**	0.53	0.25	0.10
MLP	0.76	0.74	0.46	0.0	0.77	0.29	0.64
NB	0.65	0.87	0.56	0.0	0.13	**0.35**	**0.66**
GP	**0.77**	0.77	0.00	0.0	0.76	0.00	**0.66**

Tabela 11: Métrica f1 de modelos não otimizados com seleção de atributos

	[12]	[8]	[14]	[15]	[16]	[17]	[18]
SVM	0.68	0.5	0.5	0.5	0.48	0.5	0.49
DT	0.66	0.49	0.57	0.45	0.48	0.28	0.49
RF	0.73	0.51	0.60	0.49	0.47	0.50	0.49
ADA	0.64	0.49	0.60	0.53	0.48	0.49	0.49
XGB	0.36	0.50	0.58	0.5	0.5	0.48	0.5
KNN	0.53	0.5	**0.64**	0.5	0.5	**0.52**	0.50
Perceptron	0.49	**0.52**	0.54	**0.54**	**0.51**	0.50	0.49
MLP	**0.8**	0.4	0.53	0.47	0.48	0.48	**0.51**
NB	0.75	0.45	0.60	0.5	**0.51**	0.44	0.49
GP	0.71	0.43	0.49	0.48	0.46	0.5	0.50

Tabela 12: Métrica AUC de modelos não otimizados com seleção de atributos