UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

PEDRO DURAYSKI SACCILOTTO

Desenvolvimento de modelos de aprendizado de máquina para a detecção de fenótipos humanos com base em assinaturas de microbioma

Monografia apresentada como requisito parcial para a obtenção do grau de Bacharel em Engenharia da Computação

Orientador: Prof[a]. Dra. Mariana Recamonde Mendoza
Co-orientador: M.Sc Camila Gazolla Volpiano

Porto Alegre
2021

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
Reitor: Prof. Carlos André Bulhões Mendes
Vice-Reitora: Profª. Patricia Helena Lucas Pranke
Pró-Reitoria de Ensino (Graduação e Pós-Graduação): Profª. Cíntia Inês Boll
Diretora do Instituto de Informática: Profª. Carla Maria Dal Sasso Freitas
Diretora da Escola de Engenharia: Profª. Carla Schwengber Ten Caten
Coordenador do Curso de Engenharia de Computação: Prof. Walter Fetter Lages
Bibliotecária-Chefe do Instituto de Informática: Beatriz Regina Bastos Haro
Bibliotecária-Chefe da Escola de Engenharia: Rosane Beatriz Allegretti Borges

“The frog in the well
knows nothing of the sea.”
— ZHUANG ZHOU

AGRADECIMENTOS

Agradeço aos meus pais Vanice Terezinha Durayski e Mario Lúcio Carmanim Saccilotto pelo incentivo, paciência e carinho todos esses anos. Obrigado à minha orientadora Mariana Recamonde Mendoza pela confiança, indicações e conhecimento, que fizeram grande diferença no resultado final. Estendo os meus agradecimentos a todos os professores que contribuíram com a minha formação acadêmica e profissional até agora. Obrigado também pelo suporte da Agrega e da Adriana Ambrosini, o qual foi fundamental para a realização deste trabalho. Por fim, sou grato a Camila Gazolla Volpiano pelo apoio e inspiração.

RESUMO

Alterações do microbioma intestinal têm sido associadas a diversas condições de saúde em humanos. O presente trabalho teve como objetivo a aplicação de técnicas de aprendizado de máquina supervisionado para a construção de modelos preditivos, por meio da utilização de dados de fenótipos e microbiomas intestinais humanos. Utilizou-se como conjunto de dados as amostras publicadas pelo *American Gut Project*, as quais foram filtradas para a realização da predição de seis fenótipos: doença de refluxo gastroesofágico (GRDE), doença inflamatória intestinal, síndrome do intestino irritável, doenças autoimunes, doenças pulmonares e transtornos mentais. Como atributos foram utilizados os valores da tabela de abundância de *Amplicon Sequence Variants* das amostras obtidas, os quais foram tratados com PERFect, imputação de zero e transformação logarítmica. Aos modelos foram agrupados dados fenotípicos adicionais (e.g. idade, sexo e uso de probióticos). Os modelos foram treinados utilizando-se validação cruzada com 10 *folds*, técnicas de redução de desbalanceamento entre as classes e seleção de atributos *embedded*. Foi realizada uma comparação de diferentes modelos binários de classificação, como florestas aleatórias, regressão penalizada e *gradient boosting*. O modelo mais bem avaliado foi GRDE com regressão logística penalizada e valor de MCC 0,27. Foi confirmada a suposição de melhora na predição dos modelos com a inclusão de dados fenotípicos adicionais no conjunto de dados. Ainda que baixos valores de precisão tenham sido obtidos, impossibilitando a utilização dos modelos preditos na pesquisa clínica, considera-se, como perspectiva futura, que possa ser realizada análise dos atributos com maior contribuição e integração de novos dados para aperfeiçoar as predições dos modelos.

Palavras-chave: Aprendizado de máquina. microbioma. análise preditiva em saúde.

Development of machine learning models for detection of human phenotypes using microbiome signatures

ABSTRACT

Changes in the gut microbiome have been shown to be associated with several health conditions and other characteristics of the human host. The objective of the present work was to apply supervised machine learning techniques to build predictive models using human phenotypes and data from the gut microbiome. Samples published by the American Gut Project were used as dataset, which were filtered to predict six phenotypes: gastroesophageal reflux disease (GERD), inflammatory bowel disease, irritable bowel syndrome, autoimmune diseases, lung diseases and mental disorders. The values of the Amplicon Sequence Variants abundance table of the samples obtained treated with PERFect, imputation of zero and logarithmic transformation were used as features. Additional phenotype data (e.g. age, sex and use of probiotics) were added to the models. The models were trained using 10 fold cross validation, techniques to reduce imbalance between classes and embedded feature selection. A comparison of different binary classification models, such as random forests, penalized regression and gradient boosting was carried out. The best evaluated model was GERD with penalized logistic regression and a MCC value of 0.27. It was confirmed that the inclusion of additional phenotype data in the dataset improves the prediction of the models. Although low precision values have been obtained, making it impossible to use the models for clinical assistance, it is considered, as a future perspective, that the analysis of the attributes with greater contribution and further integration of new datasets can be performed to improve the predictions of the models.

Keywords: machine learning, microbiome, predictive analytics in healthcare.

LISTA DE ABREVIATURAS E SIGLAS

AGP	*American Gut Microbiome Project*
ASV	*Amplicon Sequence Variant*
AUC	*Area Under the ROC Curve*
CLR	*Center Log Ratio*
CNN	*Convolutional Neural Network*
DFL	*Differences in Filtering Loss*
DII	Doença Inflamatória Intestinal
DRGE	Doença do Refluxo Gastroesofágico
FA	Florestas Aleatórias
FB	F-Beta *Score*
FCBF	*Fast Correlation-Based Filter*
FN	Falso Negativo
FPR	*False Positive Rate*
FP	Falso Positivo
HCOFs	Highly Contributing OTU Features
IECBiot	Incubadora Empresarial do Centro de Biotecnologia
IMC	Índice de Massa Corporal
NCBI	*National Center for Biotechnology Information*
NGS	*Next-Generation Sequencing*
NIH	*National Institutes of Health*
MCC	*Matthews Correlation Coefficient*
ML	*Machine Learning*
OTU	*Operational Taxonomic Unit*
PCA	*Principal Component Analysis*

RSWR *Random Sample with Replacement*

ROC *Receiver-Operating Characteristic*

RP *Random Projection*

SMOTE *Synthetic Minority Over-sampling Technique*

SRA *Sequence Read Archive*

SU *Symmetrical Uncertainty*

SVM *Support Vector Machine*

TPR True Positive Rate

VN Verdadeiro Negativo

VP Verdadeiro Positivo

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

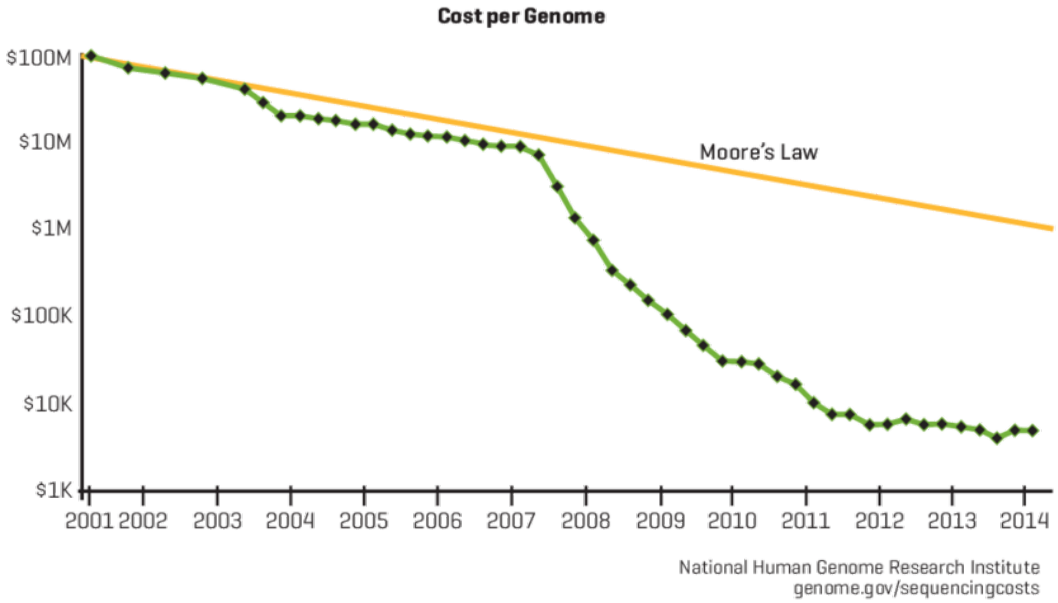
O termo "microbioma"se refere à comunidade de microrganismos dominantes que habita determinados locais ou espaços ecológicos. Desta forma, o microbioma intestinal humano se refere, de modo geral, ao agrupamento de microrganismos encontrados no intestino (ROBINSON; PFEIFFER, 2014), sendo formado, essencialmente, de bactérias que possuem um estreito relacionamento com o hospedeiro humano.

O microbioma humano (como um todo, incluindo o intestinal e de demais partes do corpo, como da pele e da boca, por exemplo) pode ser afetado pela idade, dieta alimentar, uso de medicamentos, sistema imunológico e outros fatores (TUDDENHAM; SEARS, 2015). Além disso, diversas condições clínicas têm sido associadas a diferenças da composição do microbioma entre indivíduos de grupos de caso e grupos controle, a exemplo da doença inflamatória intestinal (DII) (HALFVARSON et al., 2017), doença do refluxo gastroesofágico (DRGE) (MANOR et al., 2020), síndrome do intestino irritável (MENEES; CHEY, 2018), doenças autoimunes (BALAKRISHNAN; TANEJA, 2018), doenças pulmonares (BOWERMAN et al., 2020), diarreia (CHANG et al., 2008), pólipo colorretal (DADKHAH et al., 2019), obesidade (TURNBAUGH et al., 2006) (ZHANG et al., 2019), diabetes tipo 2 (QIN et al., 2012), depressão (PEIRCE; ALVIÑA, 2019) e autismo (VUONG; HSIAO, 2017).

As bactérias presentes no microbioma podem ser identificadas por meio do DNA com o uso de NGS (do inglês, *Next-Generation Sequencing*) aplicado em diferentes metodologias. Com o estudo da composição taxonômica e das funções destes microrganismos, é possível avaliar o potencial de relação entre o microbioma intestinal e diferentes condições de saúde e doenças humanas. Embora a análise do microbioma intestinal já seja comercializada como uma possibilidade de auxílio ao diagnóstico de diferentes condições de saúde, se tornando cada vez mais acessível devido à redução dos custos de sequenciamento de DNA (Figura 1.1), a extração de informações úteis, a partir desse grande conjunto de dados esparsos e com característica composicional, ainda é um desafio. Frente à isso, algoritmos de ML (do inglês, *Machine Learning*) podem auxiliar no diagnóstico destes dados (WU et al., 2018), já que podem realizar a extração de relações entre as composições de microrganismos no microbioma intestinal e a característica fenotípica humana desejada.

Na área comercial, há exemplos de empresas estrangeiras como a Viome e Thryve que têm disponibilizado testes de microbioma para o consumidor final *at-home*. O cliente

Figura 1.1: Evolução do custo de sequenciamento do genoma humano



A Lei de Moore se aplica principalmente a componentes de *hardware* de computadores, prevendo uma duplicação da capacidade de computação a cada dois anos (SCHALLER, 1997). O custo de sequenciamento de DNA seguiu um padrão semelhante por muitos anos, entretanto, passou a cair mais rápido do que a lei previa. Dados do *National Human Genome Research Institute*, incluindo o gráfico apresentado acima, mostram que o sequenciamento de um genoma humano inteiro custava pouco mais de 10 mil dólares em 2014, bem abaixo do valor a cerca de 100 milhões de dólares no início do milênio. Fonte: (LIPPERT; HECKERMAN, 2015).

recebe um kit para a coleta da amostra (como fezes, no caso do microbioma intestinal), que deverá ser posteriormente enviada à empresa responsável. Algumas empresas já utilizam algoritmos de ML na análise das amostras recebidas, os quais permitem a detecção de fenótipos humanos variados, assim como sexo biológico, idade, uso de antibióticos, qualidade da evacuação, dieta alimentar e diabetes tipo 2, entre outros. Com isso em mãos, as empresas podem disponibilizar recomendações pessoais de estilo de vida, nutrição e probióticos para seus clientes. Apesar disso, é importante salientar que a complexidade biológica evita a total implementação do uso do microbioma para medicina de precisão na prática clínica (CAMMAROTA et al., 2020), o qual não substitui um diagnóstico clínico profissional.

O presente trabalho foi desenvolvido em parceria com a empresa Agrega Pesquisa e Desenvolvimento em Biotecnologia, uma *startup* incubada na Incubadora Empresarial do Centro de Biotecnologia (IECBiot) da UFRGS, que possui experiência em serviços que envolvem a análise do DNA de microrganismos e microbiomas. A empresa pretende

oferecer um produto para a coleta *in-house* de fezes e a realização do sequenciamento e análise do DNA de bactérias presentes no intestino humano. Na aquisição do kit, a cliente receberá a opção de preenchimento de um questionário de saúde que deverá representar o seu perfil. Atualmente, a empresa já disponibiliza um laudo com informações relevantes sobre a caracterização do microbioma intestinal de indivíduos, mas estes resultados ainda se resumem a dados descritivos da composição taxonômica das bactérias do intestino e de suas consequências a partir de estudos científicos, a exemplo da caracterização da dieta alimentar empregada. Diante disso, existe o interesse de expansão da abrangência de uso do produto para o auxílio ao diagnóstico com aplicação de ML.

O presente Trabalho de Graduação teve como objetivo a implementação e a comparação de três diferentes modelos (“glmnet”, “gbm” e “rf”) para a predição de seis diferentes fenótipos humanos (DRGE, DII, síndrome do intestino irritável, doenças autoimunes, doenças pulmonares e transtornos mentais), utilizando-se dados provenientes de sequências de DNA, os quais foram gerados para a identificação e quantificação de ASVs (do inglês, *Amplicon Sequence Variants*) individuais através do gene *barcoding* 16S rRNA, presente em todas as espécies de bactérias. Para isso, foram utilizados os dados de NGS e fenótipos disponibilizados pelo AGP (do inglês, *American Gut Microbiome Project*), o qual apresentava mais de 15 mil amostras com variabilidade geográfica mundial e de livre acessibilidade para pesquisas de dados de saúde. Além disso, esse trabalho apresenta a base teórica sobre os formatos de dados de microbioma e conceitos fundamentais para a análise de ML dos mesmos, bem como trabalhos relacionados a esses domínios.

2 BASE TEÓRICA

2.1 Repositórios de dados de microbioma

A composição do microbioma de uma amostra é obtida após a extração, o sequenciamento e a análise das sequências de DNA de todos os microrganismos detectados nesta amostra. Utilizando-se essas sequências, ou seja, a ordem de bases nitrogenadas do DNA, é possível obter a taxonomia correspondente. Para a obtenção de sequências, a "metagenômica" e a "metataxonômica" são diferentes técnicas que podem ser utilizadas. O metagenoma é a coleção de genomas dos membros de um microbioma. Essa coleção é obtida utilizando-se a metagenômica, que é o sequenciamento *shotgun* de todo o DNA extraído de uma amostra seguido de análise de bioinformática. A análise metataxonômica ou *metabarcoding*, por sua vez, se baseia na amplificação, sequenciamento e análise de genes específicos, chamados de marcadores taxonômicos (genes *barcoding*), assim como o 16S rRNA bacteriano. O 16S rRNA é um *barcoding* padrão para a classificação de bactérias devido à sua alta informação filogenética e presença em bancos de dados de referência.

Para a realização de trabalhos deste tipo, sem a produção de dados próprios de NGS, existem duas opções principais para a obtenção de dados: repositórios públicos ou privados. A principal dificuldade envolvida em repositórios privados, como aqueles disponibilizados em banco de dados de empresas ou de projetos de pesquisa fechados, é a limitação de acesso. Além de restrições devido à questões comerciais, também existem aquelas com intuito de proteger a privacidade de pacientes, já que se tratam de dados possivelmente sensíveis.

Uma iniciativa reconhecida de geração de dados de microbioma vem do *Human Microbiome Project* do NCBI (do inglês, *National Center for Biotechnology Information*), criado em 2007 (PETERSON et al., 2009) com o objetivo de realizar associações entre doenças e mudanças no microbioma e disponibilizar conjuntos de dados para pesquisadores. Lá são encontradas uma quantidade significativa de dados (aproximadamente 15 TB em abril de 2021), porém, com restrições de acesso a dados de saúde, os quais só podem ser utilizados com justificativa a partir de um projeto escrito e liderado por um *principal investigator*. Outra iniciativa é o Projeto de Microbioma Brasileiro (PYLRO et al., 2014), o qual ainda encontra-se em processo de desenvolvimento com a expectativa deque nos próximos anos esteja acessível para pesquisadores e estudantes.

Outra iniciativa notável é o AGP (MCDONALD et al., 2018), o qual foi lançado em novembro de 2012 a partir de uma colaboração entre o *Earth Microbiome Project* e o *Human Food Project*. Em maio de 2017, o AGP incluiu dados de sequências microbianas de 15.096 amostras de 11.336 participantes humanos, totalizando mais de 467 milhões (48.599 únicos) de fragmentos do gene 16S rRNA (região chamada de V4) sequenciados com Illumina, depositando todos os dados não identificados em domínio público de forma contínua, sem restrições de acesso. O repositório do AGP se mostra ideal para trabalhos semelhantes a esse, uma vez que dispõe de uma enorme quantidade de dados públicos e, apesar de possuir uma grande quantidade de amostras de pessoas estadunidenses, também conta com dados de indivíduos que moram em pelo menos 130 países diferentes, o que contribui para o treinamento de um possível modelo geral para cada fenótipo. Isso é relevante porque, uma vez que o microbioma varia de pessoa para pessoa e também de acordo com as diferentes regiões geográficas, o uso de dados oriundos de amostras de microbiomas intestinais de brasileiros, apesar de importante, se torna uma tarefa difícil. Entretanto, uma desvantagem é que muitos atributos não foram preenchidos por todos os indivíduos do AGP, já que os estudos cadastrados possuem objetivos e metodologias diferentes. Considerando todos esses fatores, o AGP mostrou-se como o conjunto de dados mais homogêneo e o mais facilmente disponível para a realização do trabalho proposto.

2.2 Tabelas de ASVs

Por um longo período o método OTU (do inglês, *Operational Taxonomic Unit*) era majoritariamente utilizado para a análise de microbiomas. Com esse tipo de método, antes da determinação da taxonomia, deverá ocorrer o agrupamento de sequências segundo um *threshold* de similaridade, normalmente 97%. Porém, atualmente, o emprego de ASVs (do inglês, *Amplicon Sequence Variants*) vem substituindo o uso de OTU (CALLAHAN; MCMURDIE; HOLMES, 2017), possibilitando o agrupamento em sequências exatas, ou seja, sempre com 100% de similaridade, ao invés de algum valor como 97% dos OTUs. O uso de ASVs resulta na obtenção de mais dados (maior dimensionalidade), menor probabilidade de erros e portabilidade entre diferentes trabalhos. Por essa razão, o presente trabalho utiliza ASV como unidade taxonômica.

O pacote de software de código aberto “DADA2” (CALLAHAN et al., 2016) é a principal ferramenta utilizada para modelagem de ASVs. Os arquivos produzidos por NGS são arquivos de texto com extensão FASTQ, que armazenam a sequência de bases de

DNA, bem como a sua qualidade na forma de probabilidade de erro de sequenciamento. Tais arquivos são entrada de *scripts*, geralmente escritos em R, baseados no DADA2. O processo que infere a composição da amostra resume-se a: i) remoção de sequências com baixa qualidade; ii) modelagem de erros; iii) transformação dos arquivos de sequência em uma tabela de abundância dos ASVs, contendo a frequência de vezes que a sequência apareceu em cada amostra; e iv), remoção de erros de sequenciamento (quimeras). A partir destes dados é possível que outros pacotes sejam utilizados para a atribuição de taxonomia (e.g., algoritmo IDTAXA do DECIPHER (WRIGHT; YILMAZ; NOGUERA, 2012) e o algoritmo *naïve Bayes* do DADA2 (WANG et al., 2007)) utilizando-se bases de dados de referência como o SILVA SSU r138 ((YILMAZ et al., 2014) e o RDP (VILO; DONG, 2012). As saídas desses *scripts* podem ser tabelas de abundância de ASV e/ou abundância de ASVs classificadas e aglomeradas conforme *ranks* taxonômicos (i.e. reino, filo, classe, ordem, família, gênero e espécie), além de transformações dos dados, as quais devem lidar com a sua natureza composicional e inflada de zero. Tabelas tratadas levando em conta essas características podem ter seus atributos como entradas de modelos de ML para realizar a predição fenotípica, como ilustrado nas Figuras 2.1 e 2.2, que exemplificam os tipos de dados de entrada para a modelagem.

Figura 2.1: Exemplo de tabela de abundância de ASV.

	Amostra 1	Amostra 2
GGAATCTTCCGCAATGGACGAAAGTCTGACGGAGCAACGCCGCGTGAGTGATGAAGGATTTCGGTCTGTAAAGCTCTGTTGTTTATGACGAACGTGCAGTGTGTGAACAATGCATTGCAATGACGGTAGTAAACGAGGAAGCCACGGCTAACTACGTGCCAGCAGCCGCGGTAATACGTAGGTGGCGAGCGTTGTCCGGAATTATTGGGCGTAAAGAGCATGTAGGCGGCTTAATAAGTCGAGCGTGAAAATGCGGGGCTCAACCCCGTATGGCGCTGGAAACTGTTAGGCTTGAGTGCAGGAGAGGAAAGGGGAATTCCCAGTGTAGCGGTGAAATGCGTAGATATTGGGAGGAACACCAGTGGCGAAGGCGCCTTTCTGGACTGTGTCTGACGCTGAGATGCGAAAGCCAGGGTAGCGAACGG	0,71128	0
GGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTATGGATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTGGGTATCAAACAG	0,28872	0,1221
AGAATCATTCACAATGGGGGCAACCCTGATGGTGCGACGCCGCGTGGGGGAATGAAGGTCTTCGGATTGTAAACCCCTGTCATGTGGGAGCAAATTAAAAAGATAGTACCACAAGAGGAAGAGACGGCTAACTCTGTGCCAGCAGCCGCGGTAATACAGAGGTCTCAAGCGTTGTTCGGAATCACTGGGCGTAAAGCGTGCGTAGGCTGTTTCGTAAGTCGTGTGTGAAAGGCAGGGGCTCAACCCCTGGATTGCACATGATACTGCGAGACTAGAGTAATGGAGGGGGAACCGGAATTCTCGGTGTAGCAGTGAAATGCGTAGATATCGAGAGGAACACTCGTGGCGAAGGCGGGTTCCTGGACATTAACTGACGCTGAGGCACGAAGGCCAGGGGAGCGAAAGG	0	0,001
GGAATATTGGTCAATGGACGAGAGTCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATACGGGAATAAAGTGGAGTATGCATACTCCTTTGTATGTACCGTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACGGAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGGTGCTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTACCTTGAGTGCAGCATAGGTAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTCCGATTGCGAAGGCAGCTTACTGGACTGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAAACAG	0	0,8769

Na primeira coluna se encontram os ASVs, ou seja, as sequências de DNA que são estatisticamente comprovadas como estando presentes nas amostras. O primeiro ASV foi encontrado apenas na amostra 1, assim como o terceiro e quarto foram encontrados apenas na amostra 2, demonstrado a característica inflada de zero dos dados. Observe que a soma das abundâncias dos ASVs em cada uma das amostras é igual à 1, demonstrando a característica composicional dos dados de abundância de ASV.

Figura 2.2: Exemplo de tabela de abundância taxonômica em nível de gênero

	Amostra 1	Amostra 2
Phascolarctobacterium	0,71128	0
Bacteroides	0,28872	0,1221
family_Oscillospiraceae	0	0,877

Na primeira coluna se encontram as anotações taxonômicas para os ASVs. Os ASVs que compartilham da mesma taxonomia são aglomerados levando à redução da dimensionalidade da tabela. Os ASVs pertecentes à família "*Oscillospiraceae*" não puderam ser classificados em nível de gênero, neste caso é mantida a última classificação obtida segundo a hierarquia taxônomica.

2.2.1 Remoção de contaminantes e filtragem de ASVs raros

Os altos níveis de esparsidade nos conjuntos de dados do microbioma são um dos maiores desafios na análise de dados. Os resultados geralmente contém um grande número de ASVs ou OTUs raros, causados em sua maioria por contaminação e/ou erros de sequenciamento (KNIGHTS et al., 2011; RAVEL et al., 2011; SINHA et al., 2015). Atualmente, a filtragem é muito usual, e a maior parte dos autores a realiza por meio do emprego de regras práticas ("*rules of thumb*"), em que os limites para filtragem são determinados heurísticamente. Entretanto, recentemente uma nova abordagem chamada PERFect (SMIRNOVA; HUZURBAZAR; JAFARI, 2019) foi introduzida de forma a decidir quais atributos devem ser removidos com base em permutação e em uma função de perda, excluindo-se os atributos com uma contribuição insignificante para a covariância total. A proposta se baseia em classificar a importância de cada atributo, medindo sua contribuição para a covariância total e quantificando a chance de que o aumento da perda para um conjunto de atributos filtrados seja devido à aleatoriedade (SMIRNOVA; HUZURBAZAR; JAFARI, 2019).

Sendo assim, a filtragem com PERFect realizada no trabalho reduz a complexidade dos dados do microbioma enquanto preserva sua integridade na análises de ML, incluindo a retenção de atributos importantes para a preservação da capacidade de classificação de modelos, conforme demonstrado por Cao et al. (2021), que utilizaram comparações entre as AUCs (do inglês, *Area Under the ROC Curve*) obtidas de modelos de FA (Florestas Aleatórias) em conjuntos de dados filtrados e não filtrados.

2.2.2 Dados composicionais

Um sequenciador de DNA é limitado por um número máximo de leituras que ele tem capacidade de realizar, não conseguindo obter a totalidade da quantidade de sequências existentes de fato, já que possui uma capacidade limitada de dados gerados. Por essa razão, em microbiomas diversos as abundâncias (e.g. o número de contagens por amostra) representam uma fração aleatória do conteúdo original de DNA na amostra. Portanto, dados de microbioma são composicionais (GLOOR et al., 2017), já que a informação relevante está contida na relação entre as sequências. Composição é o ato de agrupar partes para formar uma só. Dados composicionais existem em proporções e carregam informações relativas. Eles possuem as seguintes características (AITCHISON, 1982): i) cada linha no conjunto de dados corresponde a uma réplica, ou experimento; ii) cada coluna no conjunto de dados corresponde a um ingrediente, ou parte em cada composição; iii) nenhum registro será negativo; e iv) a soma de todas entradas em uma linha é 1, ou 100%. Dados composicionais são vetores discretos que representam número de possibilidades em várias categorias mutuamente exclusivas, e induzem sua própria geometria de Aitchison.

A relação composicional de dados de abundância do microbioma pode ser observada nas Figuras 2.1 e 2.2 com a soma das proporções de ASVs de cada amostra resultando em 1. Percebe-se também que mudanças na abundância de um ASV induzem mudanças nas abundâncias observadas para outros ASVs.

Devido a limitação de sequenciamento, existe uma quantidade considerável de ASVs com nenhuma ocorrência nas amostras, mas que poderiam ser valores positivos se houvesse um maior número de tentativas. Como a contagem de ASVs não é confiável (i.e. sequenciar a mesma amostra duas vezes não resulta na mesma contagem), pode-se utilizar métodos de razão logarítmica para permitir a comparação entre as proporções de diferentes amostras, como por exemplo a transformação CLR (do inglês, *Center Log Ratio*). Transformações logarítmicas requerem valores diferentes de zero, portanto estes devem ser previamente substituídos removidos ou tratados, por exemplo, com um tratamento multiplicativo Bayesiano (MARTÍN-FERNÁNDEZ et al., 2015). Esse tratamento causa apenas uma pequena distorção na estrutura de covariância, tornando-se o tratamento mais adequado para a contagem de zero. Devido à isso, o tratamento de zeros com o método Bayesiano, seguido da transformação CLR dos dados de abundância foram realizados para preparação dos dados de sequenciamento do presente trabalho.

2.3 ML Supervisionado

Modelos preditivos de classificação complexos estão cada vez mais comuns na ciência (KUHN et al., 2008). Dentre os principais tipos de aprendizado dentro dos algoritmos de ML, o "ML Supervisionado" se refere à construção de algoritmos capazes de produzir padrões gerais e hipóteses a partir do uso de instâncias previamente conhecidas para predizer futuras instâncias desconhecidas (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Para isso, as instâncias do conjunto de dados a ser treinado já devem ter todos seus atributos conhecidos, incluindo o atributo alvo, e devem ser subdivididos em conjunto de dados independentes. O conjunto de dados de treinamento será a fundação para o programa de predição, enquanto o conjunto de dados de teste será aplicado no modelo treinado para avaliar a predição de instâncias desconhecidas. O "Princípio de Pareto", utilizado no presente trabalho e que afirma que aproximadamente 80% dos efeitos vêm de 20% das causas, pode ser utilizado arbitrariamente nas divisões entre subconjuntos de treinamento e teste. Existem estudos que definem *frameworks* para realizar a divisão segundo os parâmetros ajustáveis (GUYON et al., 1997), porém se o conjunto de dados possuir muitas instâncias, pode se atentar na possibilidade de alocar uma proporção menor de dados para teste.

O erro total do modelo de ML é uma soma entre o viés e a variância, e desejamos minimizar ambos, embora na prática normalmente exista um *trade-off*. Um modelo que não aprende bem o suficiente, ou que é demasiadamente simples para a hipótese a ser modelada, erra muito, tal que suas predições possuem um alto viés. Uma variância alta ocorre quando pequenas variações no conjunto de dados geram significativas mudanças na saída predita, o que torna o modelo instável em suas predições. Adicionalmente, os modelos também podem apresentar *underfit* (sub-ajuste), quando o algoritmo de aprendizado não é complexo o suficiente para capturar com precisão as relações entre atributos e saída, ou *overfit* (sobreajuste), quando um modelo se ajusta muito bem ao conjunto de dados de treinamento mas se mostra ineficaz para prever novos resultado, pois modela inclusive os ruídos inerente aos dados de treinamento.

2.3.1 Seleção de atributos

Modelos de ML podem ter dificuldades no aprendizado quando possuem dimensionalidade alta de atributos, causando *overfit*. O processo de seleção de atributos envolve

encontrar a quantidade de atributos aceitável dado uma função objetivo, como acurácia de predição e menor uso possível de atributos (JOHN; KOHAVI; PFLEGER, 1994), já que o aumento excessivo de atributos selecionados prejudica o classificador. Essa etapa auxilia na construção dos modelos de ML mais úteis possíveis para dado fenômeno, a fim de não possuir atributos irrelevantes que induzem a conclusões errôneas.

Métodos de seleção de atributos *embedded*, que foram aplicados no presente trabalho, são específicos para alguns modelos, já que estão embutidos no treinamento do modelo de classificação (LIU; ZHOU; LIU, 2019). Métodos *filter* definem um *score* de relevância dos atributos observando relações entre os atributos e as classes. Esse método possui como vantagem a escalabilidade em conjuntos de dados com alta dimensionalidade, sendo computacionalmente simples e independente de escolhas de modelos de classificação (SAEYS; INZA; LARRANAGA, 2007). Como desvantagem, esse método geralmente ignora as dependências de um atributo com outro(s) ou pode selecionar atributos redundantes. Finalmente, métodos de seleção de atributos *wrapper* são intensivos computacionalmente, e avaliam múltiplos modelos usando métodos que adicionam ou removem atributos, combinando-os. A combinação selecionada será a que possuir as melhores avaliações, portanto esse método considera as dependências entre os atributos. Como uma dimensionalidade maior pode aumentar exponencialmente as combinações, heurísticas são propostas para limitar o espaço de combinações.

"*Peeking phenomenon*" ocorre quando há um estágio preliminar ao treinamento, como a seleção de atributos. Ele ocorre se o conjunto de dados de teste é utilizado em alguma etapa de treino, por exemplo no treinamento do modelo de seleção de atributos. O efeito é uma acurácia de classificação tendenciosa e otimista. Kuncheva and Rodríguez (2018) afirma que é aceito que métodos *wrapper* dão melhores resultados que métodos *embedded* ou *filter*. Porém, considerando conjuntos de dados de grande dimensionalidade, métodos de validação cruzada K "*leave-one-out*" darão K+1 possibilidades distintas de acurácia e um risco alto de *overfit*. Logo, é proposto usar métodos *filter* e *embedded* de estado da arte para avaliação em conjuntos de dados extremamente largos, por exemplo, FCBF (do inglês, *Fast Correlation-Based Filter*), ReliefF e SU (do inglês, *Symmetrical Uncertainty*). A SU é uma medida de correlação baseada na entropia de Shannon. Quando um dos dois atributos corresponde à variável da classe, esse valor de medida obtido, que não leva em conta outras correlações entre as variáveis, pode ser usado para fazer um *rank* dos atributos. FCBF utiliza SU, porém leva em consideração as correlações entre os atributos. Esse método busca selecionar atributos com alta correlação com a variável de

classe e baixa correlação entre elas. O ReliefF é um método de seleção de atributos baseado nas instâncias. Um subconjunto de instâncias é aleatoriamente selecionado diversas vezes, e o peso das instâncias é atualizado com base na proximidade das instâncias da mesma classe nas amostras selecionadas.

2.3.2 Desbalanceamento de classes binárias

Conjuntos de dados com desequilíbrio na porcentagem de distribuição de suas classes são ditos desbalanceados, afetando o sucesso de generalização do modelo a ser treinado para instâncias desconhecidas. Diferentes métodos de amostragem foram propostos para quando não é possível obter mais dados para reduzir o desbalanceamento. O aumento artificial da quantidade de instâncias da classe minoritária é chamado de *oversampling*, enquanto a redução de instâncias da classe maioritária é chamado de *undersampling* (LIU; WU; ZHOU, 2008). Ambas técnicas introduzem viés na tentativa de compensar o desbalanceamento. O SMOTE (do inglês, *Synthetic Minority Over-sampling Technique*)(CHAWLA et al., 2002) possui várias modificações, mas de modo geral é realizado o *oversampling* de amostras considerando os *k-nearest neighbors*. Por fim, técnicas de pesagem de classes podem substituir a amostragem permitindo uma definição de maior importância das classes minoritárias com com uso de diferentes valores de pesos em cada classe a ser predita.

2.3.3 Algoritmos para classificação binária

O modelo de regressão logística permite, dado um conjunto de variáveis independentes de entrada, predizer uma variável dependente de valor categórico. Ao contrário de um modelo de regressão linear, que tem como saída um valor contínuo, a regressão logística transforma a saída com o uso de uma função sigmóide em um valor de 0 a 1, correspondente à probabilidade da saída predita ocorrer (CRAMER, 2002). Para a redução do erro deve ser realizada a penalização das predições confiantes e incorretas a partir da execução do algoritmo de gradiente descendente, que consiste em encontrar, de forma iterativa, valores de parâmetros que minimizam determinada função de custo (erro) do modelo (nesse caso *Log Loss* para saída binária). Modelos de regressão logística penalizada têm como objetivo diminuir a “sensibilidade” da mudança de dados de treinamento

para os dados de teste, fazendo a seleção de atributos de forma *embedded* (CASELLA et al., 2010). O modelo de regressão LASSO (também conhecido como L1) permite diminuir a variância, combatendo o *overfit* no treinamento. Além disso, quando há múltiplos atributos correlacionados, o modelo pode selecionar apenas um deles, zerando os outros, a partir do hiperparâmetro lambda na sua função de custo: quanto maior, mais atributos serão desconsiderados, com um aumento de viés (e quanto menor o lambda, maior a variância). O modelo de regressão RIDGE (também conhecido como L2) diminui a complexidade do modelo e é similar ao L1, podendo ter hiperparâmetros configuráveis. Entretanto, L2 tende a ser melhor quando a maioria dos atributos são úteis para a predição, já que o algoritmo não zera atributos que não afetam na classificação, apenas minimiza-os. A regressão *elastic net* combina L1 e L2, se tornando útil quando não está clara a escolha entre um ou outro. Desse modo, há dois hiperparâmetros lambdas, correspondentes a cada regularização. Devido à alta variação fenotípica do presente trabalho, o *elastic net* foi um dos algoritmos aplicados.

No modelo de árvore de decisão, o dado a ser predito percorre as observações de um item (i.e. os galhos), o qual foi dividido pelos valores de corte de cada atributo no modelo, até chegar à conclusão da classificação (i.e. as folhas). No treinamento, existem diversas etapas heurísticas, como o ganho de informação e o índice Gini (algoritmo CART), que permitem definir os critérios de seleção de atributos (QI, 2012). De forma geral, os algoritmos calculam o quão pura a divisão se torna com dado ponto de corte. É importante notar que pode ocorrer *overfit* se a árvore crescer até sua profundidade máxima. A "poda" da árvore permite mitigar esse efeito a partir da definição de limites de altura ou quantidade de folhas. O modelo de FA tem como composição uma série de árvores de decisão que foram treinadas com diferentes subconjuntos a partir de *bootstrap aggregating* (também conhecido como *bagging*), usada para estimar uma população através de amostragem com substituição, e cujas previsões são combinadas usando votação majoritária. A alta diversidade entre as árvores ocorre com a combinação de seleção aleatória de atributos em cada árvore treinada. Desta forma, a instabilidade e alta variância do treinamento de uma única árvore de decisão é mitigada, aumentando a acurácia (HO, 1995). Por essa razão, FA foi um dos três algoritmos de ML escolhidos. Um modelo pode ser avaliado pela proporção de amostras "*out-of-the-bag*" corretamente classificadas, que não foram utilizadas no treinamento de cada *bag* das árvores de decisão. O cálculo do erro *out-of-the-bag* permite a escolha do modelo que possui maior acurácia, além de não necessitar de alta demanda computacional e possibilitar a realização de teste enquanto o

modelo está sendo treinado.

O *gradient boosting* é uma técnica de ML que inicia criando uma árvore com uma única folha. No caso de classificação binária, o valor dessa única folha seria o *log odds* das predições convertido para uma probabilidade com uma função logística. Com isso, o algoritmo produz árvores de decisão a partir dos erros (resíduos) da árvore de decisão anterior, porém com um maior número de folhas. A variância possivelmente alta em cada árvore de decisão, após ajustar os parâmetros de acordo com a árvore anterior, é atenuada com um valor numérico de taxa de aprendizagem quando é empregada uma estratégia estocástica (RIDGEWAY, 2007). Friedman (2002) afirma que levar pequenos passos na direção correta dos resultados resulta em menor variância e melhores predições no conjunto de dados de teste. Devido à isso, esse algoritmo foi o último escolhido a ser aplicado no presente trabalho. A predição de um valor será uma equação da soma dos resíduos das árvores anteriores, e os resíduos devem ter seu valor reduzido com o aumento de árvores de decisão. O algoritmo continuará criando árvores até chegar a um máximo especificado, ou se a adição de novas árvores não reduzir os resíduos.

2.3.4 Métricas de avaliação

Mesmo com um classificador treinado e disponível para teste, a avaliação do modelo se faz necessária para a seleção que possuir os melhores resultados, já que não é possível estabelecer *a priori* o método que melhor resolve o problema. Há diversas métricas que avaliam o quanto o modelo treinado erra na classificação de novas instâncias. Em classificadores binários, podemos nomear as classes como positivas e negativas, podendo assim dividir as predições em 4 casos, como observado na Tabela 2.1, em que a classe positiva é a classe "Sim". Uma predição correta de uma classe positiva é um caso verdadeiro positivo (VP), enquanto a predição incorreta é um caso falso negativo (FN). A predição correta de uma classe negativa é um caso verdadeiro negativo (VN), enquanto a predição incorreta é um caso falso positivo (FP). O número total de instâncias sendo avaliadas (N) é a soma de todos esses casos descritos.

A "acurácia" (acc) avalia a taxa de acerto das predições do modelo, sendo possível equacionar como (FLACH, 2003):

$$acc = \frac{VP + VN}{N}$$

Tabela 2.1: Tabela de confusão para um classificador binário

	Referência Sim	*Referência Não*
Predição Sim	VP	FP
Predição Não	FN	VN

As métricas da tabela de confusão são baseadas em VP, VN, FP e FN.

A acurácia não é eficiente para a avaliação em casos com classes altamente desbalanceadas entre si no conjunto de dados, como acontece em muitos experimentos clínicos. Considere um caso hipotético de 18 pessoas que não tivessem diabetes tipo 2, e 6 que tivessem essa doença. Se um modelo tivesse predito que todas essas instâncias não possuem diabetes tipo 2, ele teria uma acurácia de 75%, enganando o avaliador que esse modelo é eficaz em predizer quem não possui a doença. A métrica de *recall* (rec), também conhecida como TPR (do inglês, *True Positive Rate*) ou sensibilidade, além da métrica de precisão ($prec$), são capazes de resolver esse problema. A precisão de uma classe define a taxa de acerto da predição, portanto dentre todas as predições de uma classe (incluindo as que não foram identificadas corretamente), qual a fração de acerto da mesma. O *recall* define a taxa de acerto da instância, portanto, dentre as instâncias de uma classe (incluindo as que deveriam ter sido identificadas como essa classe), quantas das predições esperadas da mesma foram corretas. Precisão e *recall* da classe positiva podem ser equacionadas como (FLACH, 2003):

$$prec = \frac{VP}{VP + FP}, rec = TPR = sens = \frac{VP}{VP + FN}$$

Modelos orientados à precisão tentam minimizar FP, ou erros do tipo 1, não cometendo erros nas predições da classe. Já modelos orientados à *recall* visam minimizar os FN, ou erros do tipo 2. O avaliador de um modelo de ML que prediz doenças tem como opção dar maior preferência a um taxa de *recall* do que de precisão. Considerando que a classe positiva é a posse da doença, talvez seja preferível uma maior taxa de acerto para a classe positiva neste domínio. Porém, se a doença for contagiosa, talvez a precisão seja mais importante. O F-Beta *Score* (FB) é uma métrica que se utiliza de um fator que permite definir quanto mais importante é o *recall* comparado à precisão. Assim, F1-*Score* (Beta=1) é a média harmônica entre as taxas de precisão e *recall*, visando o balanço entre as duas métricas. Estas métricas são equacionadas como (FLACH, 2003):

$$FB = (1 + \beta^2) * \frac{prec * rec}{(\beta^2 * prec) + rec}, F1 = \frac{2 * prec * rec}{prec + rec}$$

Num diagnóstico clínico em que a presença da doença é a classe positiva, a especificidade ($espec$) pode ser definida como a proporção das predições corretas de realmente não possuir a doença (LALKHEN; MCCLUSKEY, 2008). Portanto, um teste com 70% de especificidade corretamente reporta 70% dos pacientes sem doença corretamente, porém 30% são preditos incorretamente como possuindo a doença (FP). A métrica de *recall* e especificidade estão relacionadas: enquanto *recall* avalia a proporção dos casos positivos preditos corretamente, a especificidade avalia a proporção dos casos negativos preditos corretamente.

$$espec = \frac{VN}{VN + FP}$$

O F1-*Score* pode ser uma métrica eficiente, porém, segundo Chicco and Jurman (2020), ela ainda é capaz de apresentar resultados super otimistas e inflados em conjuntos de dados desbalanceados. Exemplos e vantagens do MCC (do inglês, *Matthews Correlation Coefficient*) em classificações binárias de dados desbalanceados são citados no estudo. O valor dessa métrica varia de -1 a 1, onde o valor 1 é uma classificação perfeita, -1 é a pior classificação possível e o valor 0 corresponde a uma capacitação de predição próxima à aleatória. Esta métrica é mais informativa e verdadeira, já que um modelo pontua bem apenas se houverem bons resultados nos 4 casos possíveis citados anteriormente, como pode ser observado na equação abaixo:

$$MCC = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}}$$

Por fim, a curva ROC (do inglês, *Receiver-Operating Characteristic*) traça uma curva de probabilidade referente à TPR e FPR (do inglês, *False Positive Rate*) em diferentes limiares de classificação referentes à distribuição das instâncias (FLACH, 2003): sendo 0 todas as entradas como positivas, e 1 todas como negativas. Um classificador que retorna apenas a classe predita oferece apenas um ponto no espaço ROC. Em classificadores probabilísticos, que retornam a probabilidade que dada instância pertença a uma classe, podemos criar a curva pela variação dos limiares. O valor da AUC permite especular a performance de um modelo resumindo a curva ROC em um cálculo da área

sob a curva.

$$FPR = \frac{FP}{FP + VN}$$

Por considerarem diferentes problemas de performance de modelos de ML, diversas dessas métricas possuíram seus valores avaliados no presente trabalho, principalmente *recall*, precisão, F1-*score*, especificidade, AUC e MCC em conjuntos de dados de teste, devido à seu desbalanceamento.

2.3.5 Técnicas para avaliação e comparação dos modelos

Validação cruzada é uma técnica que objetiva avaliar o poder de generalização de um modelo de ML e auxiliar na seleção de hiperparâmetros. Isto pode ser feito de diversas maneiras, mas, de forma geral, o método envolve o particionamento do conjunto de dados em subconjuntos de treino, teste, e, se desejado, validação. No *holdout*, o conjunto de dados é dividido em conjunto de treinamento e de teste, com tamanhos arbitrários (mas em geral com um subconjunto maior de treinamento). Após o treinamento, o conjunto de teste avalia a performance do modelo. É um método extremamente simples, envolvendo apenas uma única corrida e resultados potencialmente enganosos. O método *K Fold* envolve o cálculo da média de diferentes corridas de avaliação do modelo treinado a partir de uma divisão diversa do conjunto de dados em cada corrida. Por essa razão, esse método tende a ser menos tendencioso que o *holdout*. O *K Fold* é iniciado com o embaralhamento do conjunto de dados e divisão dos mesmos em K subconjuntos. O subconjunto terá seu papel definido no início de cada corrida: treinamento ou validação, em apenas uma das corridas. O poder de predição do modelo treinado em cada corrida será calculado K vezes. Após o fim de todas as corridas, o desempenho do modelo é calculado com a média e desvio padrão das K corridas avaliadas (KOHAVI et al., 1995). A fim de diminuir o viés do conjunto de dados, o *K Fold* pode ser repetido diversas vezes, com um novo embaralhamento inicial do conjunto de dados. A escolha de K é arbitrária, sendo tradicionalmente utilizado um valor de 5 ou 10. O aumento do mesmo afeta o tamanho do conjunto de treinamento, deixando-o mais próximo ao conjunto de dados original, porém, aumenta também o tempo de processamento. A performance estimada da validação cruzada K *Fold* possui um viés (VARMA; SIMON, 2006), visto que a validação está encapsulada no treinamento, portanto esse método não substitui a avaliação final de

performance com o conjunto de teste. Devido a isso, o presente trabalho procura avaliar os valores de performance da validação cruzada seguido dos valores de performance da aplicação do modelo no conjunto de dados de teste. Com isso, podemos confirmar as suposições de melhores modelos que são obtidas da validação cruzada e observar se os valores da predição no conjunto de dados de teste estão em um desvio padrão aceitável.

2.3.6 ML no R

O R apresenta um grande número de pacotes para ML, entretanto, cada pacote foi projetado de forma independente, com sintaxe, entradas e saídas muito diferentes. O pacote "caret" (do inglês, "*Classification And Regression Training*"), utilizado no trabalho, facilita a criação de diferentes modelos por estar encapsulado numa linguagem com diversas funções de modelagem estatística e gráfica (TEAM, 2000). O pacote também introduz uma sintática comum em diferentes funções de construções de modelos, assim como uma possível abordagem automática para a otimização dos valores de hiperparâmetros.

3 TRABALHOS RELACIONADOS

A predição de dados de saúde humana é um assunto complexo. Há a complexidade de um domínio relativamente novo e que requer um modelo de ML rigorosamente validado. Além disso, implicações éticas relacionadas às chances de um algoritmo classificar doenças, sem autoridade e conhecimento previamente reconhecidos para tal, são questões relevantes do ponto de vista social, devido ao potencial para causar impactos negativos à pessoa analisada e sua família. Por outro lado, o valor de uma ferramenta de predição de doenças pode ser de grande importância para diversas vidas, sobretudo nos casos em que o diagnóstico não pode ser facilmente determinado. Assim, nesse domínio é importante predizer com probabilidades calibradas: métricas de performance usuais podem não ser o suficiente (NICULESCU-MIZIL; CARUANA, 2005). Modelos de ML probabilísticos são necessários para o auxílio ao diagnóstico de doenças, os quais, muitas vezes, são envoltos por incertezas. Nesses casos, além de predizer se alguém possui ou não uma doença, o modelo terá como saída as probabilidades das diferentes classes. Com isso em mãos, um profissional capacitado poderia utilizar os resultados de modo complementar aos diagnósticos tradicionais, ampliando seu entendimento e aprimorando suas conclusões.

A aplicação e estudos de técnicas de classificação de ML para dados de microbioma humano são relativamente recentes, tendo poucos estudos científicos até 2011 (KNIGHTS; COSTELLO; KNIGHT, 2011). Apesar da literatura apresentar resultados bem sucedidos, Dave et al. (2012) cita algumas das dificuldades envolvidas, assim como desenhos dos estudos, erros de sequenciamento, contaminações, definições de tabela de OTUs (ainda amplamente utilizadas), instabilidade do microbioma humano em jovens, variabilidade entre populações geográficas, dieta alimentar, consumo de drogas e perturbações ocasionadas por medicações, sobretudo antibióticos. Atualmente, parte destes problemas são atenuados com a remoção de sequências com baixa qualidade ou quimeras, uso de ASVs e levantamento dos fenótipos do hospedeiro.

Zhou and Gallins (2019) realizou uma análise dos modelos e métricas empregadas em trabalhos publicados envolvendo ML. Percebe-se na área uma preferência por algoritmos como FA, seguido de regressão penalizada e SVM (do inglês, *Support Vector Machine*). Quanto às métricas, AUC, taxa de erro e acurácia predominam, com pouco uso da métrica de F1-*Score* ou MCC. A Tabela 3.1 ilustra um resumo do seu levantamento agrupado a outros estudos relevantes de microbioma intestinal. Não há muito detalhamento

quanto à natureza dos resultados, ou seja, se estes são referentes à predição no conjunto de dados de teste ou de validação cruzada.

Aryal et al. (2020) utilizou conjuntos de dados do AGP com tabelas de abundância de OTUs para a predição de doenças cardiovasculares a partir de dados de microbioma intestinal. Diversos modelos foram treinados usando os packages "caret", "kernlab", "randomForest", "rpart" e "glmnet" do R. Estes foram a árvore de decisão, FA, regressão penalizada, redes neurais e SVM. As 500 OTUs com maior variância entre todas as amostras foram selecionadas no modelo. Foi utilizada validação cruzada 10-*fold* repetida 10 vezes. As métricas utilizadas foram especificidade, *recall* e AUC. Após o treinamento dos modelos, foram identificadas os HCOFs (do inglês, *Highly Contributing OTU Features*) com o método varImp do caret. Os AUCs reportados dos modelos foram: FA, 0.58; redes neurais, 0.58; regressão penalizada, 0.57; SVM, 0.55; e árvore de decisão, 0.51. FA e redes neurais tiveram menor *recall*, em torno de 0.59, e maior especificidade, de aproximadamente 0.51, comparado aos demais modelos. Com uso de apenas HCOFs, a AUC da FA subiu para 0.65, e houve um aumento do *recall* para 0.7, porém a especificidade continuou com um valor próximo a 0.51. Os demais modelos não possuíram uma melhora significativa. Para diminuir ainda mais a dimensionalidade, foram treinados diferentes modelos com número de HCOFs diferentes. Um modelo com 25 HCOFs apresentou AUC de 0.7, *recall* de 0.7 e especificidade de 0.6. Dong et al. (2020) também utilizou seleção de OTUs para aperfeiçoar as predições, porém, com métodos de seleção de atributos *filter*. É afirmado que o número de OTUs das amostras, assim como o número de amostras e a escolha do modelo, afetam o bom ou mau funcionamento do método *filter*. Além disso, o autor afirma que métricas de regressão penalizada bem modeladas são o suficientes para modelagem de um modelo com boa performance.

Pietrucci et al. (2020) segue como passo-a-passo de preparo de dados o uso da biblioteca DADA2, utilizando tabelas de abundância de ASVs classificadas taxonomicamente em nível de família, dados fenotípicos das amostras de diferentes nacionalidades e amostras de fezes sequenciadas utilizando o 16S rRNA, ou seja, utilizando a metataxônomica. Diferentes modelos foram treinados, utilizando-se validação cruzada 5-*fold*: FA, rede neural e SVMs. Dos três, o modelo de FA resultou em um AUC de 0.8 e acurácia de 0.71 para a predição de mal de Parkinson. Concluiu-se que a obtenção de uma acurácia maior poderia ser realizada com um aumento no conjunto de dados de treinamento.

Yang and Zou (2020) introduziu a ferramenta mAML, que realiza ML automatizando as etapas de processamento dos dados e selecionando o modelo mais adequado para

os conjuntos de fenótipos. Em geral, quatro métodos de seleção de atributos são adotados para lidar com grandes conjuntos de dados de microbioma. O problema de desbalanceamento de classes é compensado usando diferentes métodos, como o RSWR (do inglês, *Random Sample with Replacement*), SMOTE ou ADASYN (do inglês, *Adaptive Synthetic Sampling Approach for Imbalanced Learning*). A ferramenta seleciona os melhores hiperparâmetros para cada classificador com validação cruzada, disponibilizando uma série de métricas para analisar o desempenho (F1-*Score*, precisão, *recall*, curva ROC). Pasolli et al. (2016) introduziu o MetAML, uma ferramenta de ML que ajuda a levantar a possibilidade de associações entre fenótipos e microbiomas. A ferramenta inclui a seleção automática de modelos, dentre eles FA, L1, SVM e *elastic net*. A ferramenta também inclui seleção de atributos automática.

O modelo de rede neural artificial é definido por um conjunto de neurônios em uma camada que se interliga com todos os neurônios em uma próxima camada. Após percorrer os neurônios, até o neurônio de saída, calcula-se o erro e inicia-se a fase de adaptação dos coeficientes de peso de cada neurônio, com gradiente descendente até que se chegue a um limiar pré-definido. O número total de camadas e neurônios em cada camada são especificados pelo usuário, e esse número distingue a rede neural de um algoritmo de *deep learning*, que em geral possui mais que três camadas. *Frameworks* com técnicas de *deep learning* são promissoras e já foram experimentadas. Oh and Zhang (2020) utilizou vários modelos de *autoencoder*, os quais são utilizados para aprender uma representação de baixa dimensionalidade do microbioma. O estudo também testou algoritmos tradicionais de redução de dimensionalidade, como PCA (do inglês, *Principal Component Analysis*) e RP (do inglês, *Random Projection*), substituindo o *representation learning*. O estudo concluiu que, em 2 conjuntos de dados, o PCA teve uma pequena melhoria comparado ao *representation learning*, porém, nos 5 restantes, o *encoder* resultou numa melhor performance de predição e, devido ao *hardware* disponível, menor tempo decorrido. LaPierre et al. (2019) também utilizou *autoencoder*, além de CNN (do inglês, *Convolutional Neural Network*), designados para processar imagens. Como o CNN é poderoso para esse tipo de processamento, foram desenvolvidos métodos para converter diferentes tipos de dados para imagens, como por exemplo dados de predição de doenças baseados em metagenômica. O autor comparou diferentes modelos tradicionais com modelos *deep learning* em diversas métricas, enfatizando tal necessidade nessa área.

Estudos de dados de microbioma são recentes, com protocolos distintos e usualmente uma fraca documentação, execução e robustez, como Topçuoğlu et al. (2020)

afirma. Alguns dos problemas da área são: i) falta de transparência nos métodos utilizados e como eles são utilizados; ii) possíveis avaliações de performance dos modelos sem um conjunto de dados de teste separado do treinamento; iii) variações não relatadas na performance de predição em diferentes *folds* da validação cruzada; e, iv) comparações não relatadas entre performance de validação cruzada e de teste. A área, que mostra um crescimento motivante com suas ferramentas de código aberto, principalmente com métodos de *deep learning*, ainda precisa de avanços para a melhora da reprodutibilidade e minimização da superestimação de performances.

Tabela 3.1: Resultados de estudos de ML com dados de microbioma intestinal

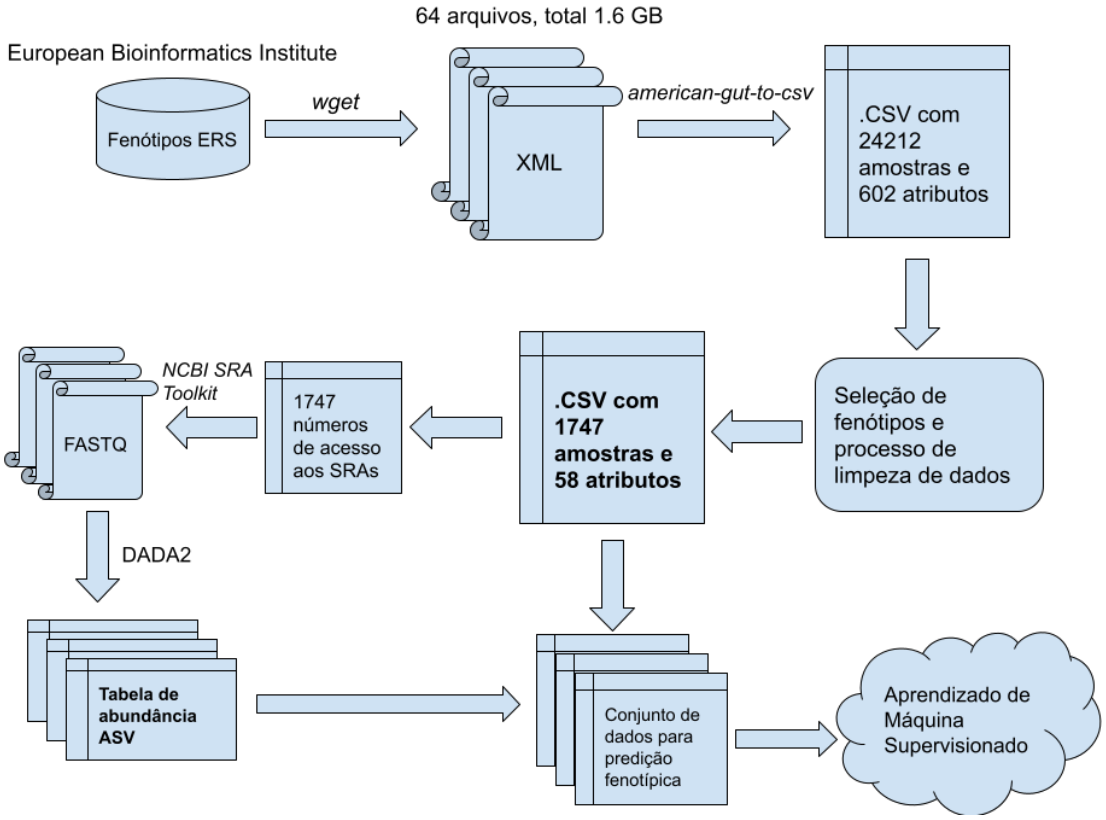
Estudo	Fenótipo	Casos	Controles	Nível	Método	Métrica	Valor
Zeller et al. (2014)	CRC	118	114	Espécie	FA	AUC	0.87
Chatelier et al. (2013)	Obesidade	164	89	Espécie	FA	AUC	0.65
Qin et al. (2012)	Diabetes 2	170	174	Espécie	FA	AUC	0.74
Ai et al. (2017)	CRC	44	99	Espécie	FA	AUC	0.94
Aryal et al. (2020)	Cardiovascular	478	473	N/A	FA	AUC	0.7
Dong et al. (2020)	Parkinson	197	130	N/A	LASSO	AUC	0.87
Pietrucci et al. (2020)	Parkinson	472	374	Família	FA	AUC	0.8
Topçuoğlu et al. (2020)	CRC	229	261	N/A	FA	AUC	0.69
Oh and Zhang (2020)	IBD	25	85	N/A	Deep	AUC	0.95
	Diabetes 2	170	174	N/A	Deep	AUC	0.76
	Diabetes 2	53	43	N/A	Deep	AUC	0.89
	Obesidade	164	89	N/A	Deep	AUC	0.65
	Cirrose	118	114	N/A	Deep	AUC	0.94
	CRC	48	73	N/A	Deep	AUC	0.67

Estudos levantados realizaram predição de diferentes fenótipos. Em sua maioria, a OTU foi utilizada como unidade taxonômica e diferentes modelos foram treinados, contudo, em geral, a melhor performance foi reportada com o uso de FA. CRC se refere à câncer colorretal e IBD se refere à DII. Fonte: (ZHOU; GALLINS, 2019) e autor.

4 METODOLOGIA

As próximas seções detalham como a obtenção dos conjuntos de dados fenotípicos e sua devida filtragem resultou nos dados de sequenciamento para obter os conjuntos de dados de entrada dos algoritmos de ML (Figura 4.1).

Figura 4.1: Processo de obtenção do conjunto de dados



Processo de obtenção dos dados. Primeiramente, os dados de fenótipos foram obtidos e filtrados, resultando também em uma lista de acessos aos SRAs para a obtenção dos arquivos FASTQ. Após o processamento dos FASTQ, uma tabela de abundância dos ASVs foi obtida, filtrada e tratada para a imputação de zeros e transformação CLR para iniciar o processo de ML.

4.1 Obtenção e filtragem de dados

A obtenção dos dados fenotípicos em formato XML foi realizada a partir do repositório de dados de sequenciamento EMBL-EBI (do inglês, *European Bioinformatics Institute*). Para isso, foi realizado um *download* recursivo utilizando comandos wget de

protocolo HTTPS no sistema operacional Linux. Utilizando APIs de XML e pandas para agrupamento tabular, o programa de código aberto em Python "american-gut-to-csv"[1] foi desenvolvido para agrupar os arquivos ERS em formato XML em um único arquivo de formato tabular CSV.

Foi desenvolvido um script em R com o propósito de remoção de características consideradas irrelevantes, assim como os dados do VioScreen (um questionário de nutrição para pesquisadores) e outras vastas opções de características de dieta. Após cada filtragem, um arquivo texto de *log* foi criado com o intuito de mapear os atributos e seus valores e frequências. A metodologia "*garbage in, garbage out*" foi aplicada para a remoção de instâncias de amostras que não eram de fezes, ou que pertenciam a indivíduos com idade menor que 18 anos, uma vez que estes podem possuir um microbioma intestinal instável (DERRIEN; ALVAREZ; VOS, 2019). Além disso, foram removidas as instâncias de indivíduos que possuíam doenças autodiagnosticadas ou diagnosticadas com base em abordagens de "medicina" alternativa, bem como as de indivíduos que não relataram a posse (ou não) da doença ou desconheciam-na. Atributos em branco e amostras com mais de um sequenciamento também tiveram suas instâncias removidas.

O *download* dos arquivos de sequenciamento foi realizado de acordo com o número de acesso SRA (do inglês, Sequence Read Archive) para cada uma das instâncias obtidas. Para isso, o SRA *Toolkit* foi instalado em um Linux WSL2, sendo utilizado para o *download* de acordo com os números de acesso dos SRA listados e, por fim, para a conversão dos dados em arquivos de formato FASTQ.

4.2 Escolha de fenótipos

Foram selecionados seis fenótipos com classes binárias que possuem estudos comprovando a relação entre seu diagnóstico e modificações do microbioma, sendo eles DRGE (MANOR et al., 2020), DII (HALFVARSON et al., 2017), síndrome do intestino irritável (MENEES; CHEY, 2018), doenças autoimunes (exceto DII e diabetes tipo 1) (BALAKRISHNAN; TANEJA, 2018), doenças pulmonares (BOWERMAN et al., 2020) e transtornos mentais (ROGERS et al., 2016). O desbalanceamento de classes também foi considerado para essa seleção. Os fenótipos com baixa variação foram excluídos. Fenótipos multiclasse, que são fenótipos de doenças com diferentes variações que influenciam em diferentes sintomas, apesar de interessantes, não foram considerados neste estudo.

[1] Disponível no endereço <https://github.com/sacci/american-gut-to-csv>

Outros dados fenotípicos de entrada também foram definidos considerando sua afinidade entre o questionário de fenótipo aplicado pela empresa, com o objetivo de aprimorar os modelos a serem treinados. Esses dados, chamados agora de "dados de questionário", são: idade, gênero biológico, tipo de dieta, intolerância à lactose, IMC (Índice de Massa Corporal) e frequências relacionadas à ingestão de bebidas alcoólicas, à rotina de exercícios, à ocorrência de diarreias, ao uso de probióticos e à ingestão de pelo menos um litro de água ao dia.

4.3 Processamento das leituras

A empresa Agrega disponibilizou seu conhecimento e *scripts* em R para a realização das etapas necessárias de conversão dos dados de saída do sequenciador de DNA, em formato FASTQ, em tabela de abundância no formato ASV, além da tabela de classificação taxonômica bacteriana. Para isso, o método DADA2 do pacote R "dada2" v1.18.0 foi utilizado para inferir as sequências biológicas verdadeiras dos arquivos FASTQ *single-end* de cada amostra. As leituras *forward* foram truncadas na posição 140 e os primeiros 10 nucleotídeos do início de cada leitura foram removidos. As leituras foram filtradas de acordo com o máximo recomendado de 2 erros esperados por leitura (EDGAR; FLYVBJERG, 2015) e as sequências contendo nucleotídeos não identificados foram removidas. As taxas de erro foram aprendidas para as amostras dispostas em cada experimento de sequenciamento. As tabelas de abundância obtidas foram mescladas, seguido da remoção de ASVs quiméricos. Ao final da análise, foi validado que todas as amostras apresentavam mais de 10 mil leituras.

As atribuições de níveis taxonômicos de filo até gênero foram feitas utilizando ambas as orientações com a função IdTaxa disponível através do pacote "DECIPHER" v2.14.0 (MURALI; BHARGAVA; WRIGHT, 2018) em sua confiança padrão ($\geq 60\%$), utilizando-se o classificador treinado SILVA SSU r138 ((YILMAZ et al., 2014), link para a licença: https://creativecommons.org/licenses/by/4.0/legalcode). Para garantir que nenhum artefato foi incluído na análise *downstream*, os ASVs anotados com um filo de "NA" ou com um domínio de "*Archaea*" ou "*Eukaryota*" foram removidos. As demais classificações taxonômicas anotadas com "NA" foram substituídas pela última classificação taxonômica atribuída. O pacote "phyloseq" v1.30.0 (MCMURDIE; HOLMES, 2012) foi utilizado para dispor os dados de contagem e objetos de atribuição taxonômica em um único objeto "phyloseq".

A fim de remover ASVs espúrias e reduzir a complexidade do conjunto de dados, foi realizada a filtragem não supervisionada de duas etapas implementada no pacote "PERFect" v1.0.0 do R (SMIRNOVA; HUZURBAZAR; JAFARI, 2019). Para isso, os ASVs foram ordenados de acordo com os valores *p* e filtrados com base em uma etapa de permutação de 1.000 repetições. Após, foi realizada a imputação de zeros com um tratamento multiplicativo Bayesiano (MARTÍN-FERNÁNDEZ et al., 2015), seguido de transformação CLR, sendo esta tabela utilizada como entrada para a modelagem.

Um *script* em R foi desenvolvido com o propósito de agrupar os dados fenotípicos à tabela de abundância de ASVs transformada por CLR. Ao todo foram gerados doze conjuntos de dados, já que cada fenótipo apresenta um conjunto de dados com e sem dados de questionário. O agrupamento foi realizado com o pacote "dplyr" v1.4.4. O método "dummyVars" do pacote "caret" (KUHN et al., 2008) v6.0-86 permitiu a tradução de dados textuais em dados numéricos com propósito de entrada para os modelos.

4.4 Treinamento dos modelos

Um *script* em R foi desenvolvido para facilitar o *loop* completo de treinamento do modelo e obtenção dos resultados de performance. Para isso, é selecionado o fenótipo e a utilização do conjunto de dados com ou sem questionário. Como as predições dos modelos possuem um caráter de auxílio clínico, foi alocado 80% do conjunto de dados para treinamento e 20% para teste de forma isolada para avaliação final do modelo (VARMA; SIMON, 2006).

O método "train" do caret foi utilizado para realizar o treinamento de cada fenótipo. Os dados foram pré-processados com o método "*Mean Centered*" (SINGH; SINGH, 2020), escolhendo-se como argumento "center" e "scale" no parâmetro "preProcess". Foi escolhido heuristicamente o "ROC" como valor de métrica para selecionar o modelo final, fazendo com que os hiperparâmetros escolhidos sejam os de maior valor nas amostras *holdout*. O parâmetro *grid* de *tuning* dos hiperparâmetros dos modelos foi configurado ao modificar o parâmetro "tuneLength" para o valor 20, selecionando, assim, o melhor parâmetro avaliado para cada modelo. Como parâmetro de controle "trainControl" foi selecionada a validação cruzada K-*fold*, com 10 *folds* e retornada as probabilidades das classes. Para tratar o desbalanceamento das classes, os modelos foram treinados com *undersampling* (LIU; WU; ZHOU, 2008) e SMOTE (CHAWLA et al., 2002).

Foram treinados três diferentes modelos com seleção de atributos *embedded* (LIU;

ZHOU; LIU, 2019): “glmnet”, um método de regressão logística penalizada *elastic net* (CASELLA et al., 2010); “gbm”, um método de *gradient boosting* estocástico (RIDGEWAY, 2007) com processamento extenso, e “rf”, um método de FA (QI, 2012), também com longa duração de processamento. Ao todo, para cada um dos modelos, foram treinadas 3 diferentes variações: *undersampling* sem dados de questionário, *undersampling* com dados de questionário e SMOTE sem dados de questionário. Para os seis diferentes fenótipos foi portanto obtido o total de 54 modelos treinados.

As 400 variações combinadas de valores nos hiperparâmetros em cada *fold* dos modelos “glmnet” modificaram o parâmetro “alpha”, referente à elasticidade do modelo entre L1 e L2, e “lambda”, a taxa de regularização que influencia na complexidade do modelo. Já as 400 iterações no “gbm” modificaram os parâmetros “n.trees”, referentes ao número de árvores, e “shrinkage”, referente à taxa de aprendizado, entre outros (KUHN et al., 2008). As 20 variações do hiperparâmetro “mtry” nos *folds* dos modelos “rf” modificaram o número de variáveis randomicamente amostradas em cada *split* de geração de árvore. Os erros de predição de cada árvore portanto são diferentes e menos correlacionados, e as predições finais são a média das árvores construídas, resultando em taxa de erros de predição mais acurada (LIAW; WIENER et al., 2002).

O pacote “DMwR”[2] v0.4.1 foi utilizado para realizar a amostragem SMOTE no “gbm”. O pacote “pROC”[3] v1.17.0.1 foi utilizado para curvas de probabilidades das classes fenotípicas. O pacote “ModelMetrics”[4] v1.2.2.2 foi utilizado para a obtenção das performances de predição no conjunto de teste. O pacote “MLeval”[5] v0.3 foi utilizado para interpretar rapidamente os resultados da função “train” do “caret”, agilizando a análise e gerando curvas ROC para a validação cruzada.

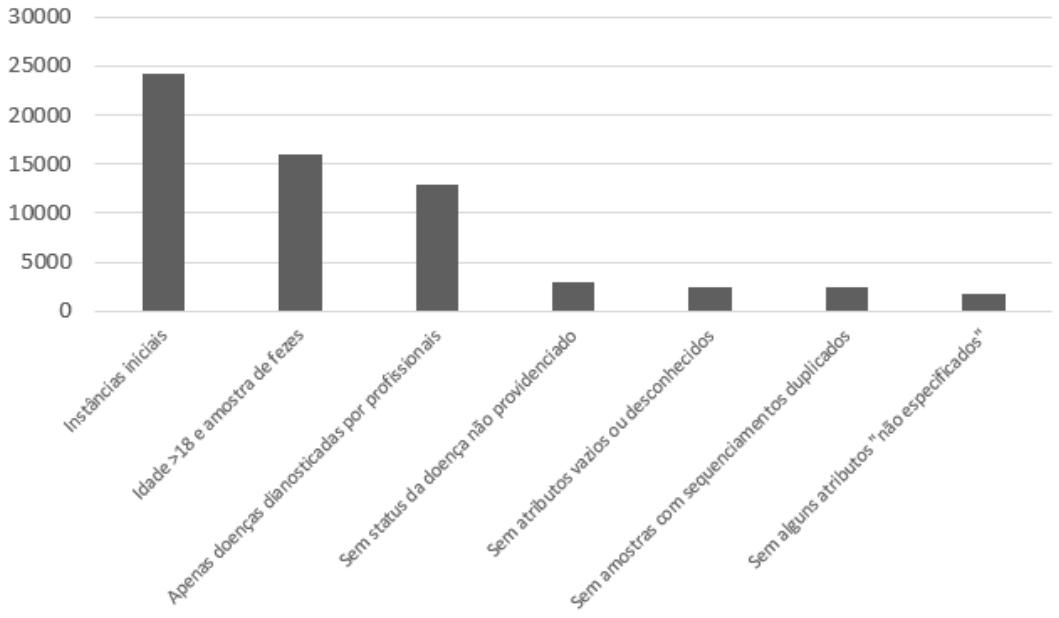
[2]Disponível no endereço <https://github.com/cran/DMwR>
[3]Disponível no endereço <https://github.com/cran/pROC>
[4]Disponível no endereço <https://github.com/cran/ModelMetrics>
[5]Disponível no endereço <https://github.com/crj32/MLeval>

5 RESULTADOS E DISCUSSÃO

5.1 Obtenção e processamento de dados

Para a realização deste trabalho, foram obtidos do EMBL-EBI 64 arquivos de metadados nomeados como “ERS”, totalizando 1.62 GB. Após a transformação do XML em CSV, o arquivo obtido apresentou 24.212 linhas correspondentes às amostras, além de 602 colunas equivalentes aos atributos. A partir das filtragem realizadas, o arquivo tabular com informações fenotípicas foi reduzido para 1.747 instâncias (cerca de 93% de redução) e 58 atributos (Figura 5.1). Cerca de 52% das instâncias são de moradores dos Estados Unidos da América, enquanto 38% do Reino Unido.

Figura 5.1: Instâncias obtidas após cada comando de limpeza de dados.



Número de instâncias restantes durante o processo de limpeza de dados, indo desde as 24.212 amostras iniciais até as 1.747 amostras finais.

A filtragem aplicada nos fenótipos poderia ser considerada severa, todavia, devido a limitações de *hardware* disponível, essa redução de amostras se fez necessária para o treinamento dos modelos em tempo hábil e para o *download* de SRAs.

Os fenótipos escolhidos para a modelagem são explorados quanto ao seu desbalanceamento na Tabela 5.1. DRGE exibiu o menor nível de desbalanceamento (17,3% de casos) seguido de transtorno mental (14,5% de casos). Depressão, apesar de possuir um dos menores desbalanceamentos, não foi selecionada por estar contida no fenótipo de

transtornos mentais. Câncer também foi desconsiderado por ser uma doença multiclasse onde não há estudos concretos que esse fenótipo "geral" possua relações com mudanças no microbioma. DII foi selecionado por apresentar uma alta correlação entre variações de micobioma e apresentação da doença. Abuso de substância, autismo e bulimia apresentaram os maiores níveis de desbalanceamento (<0,7% de casos).

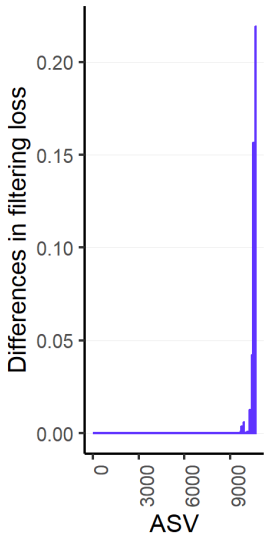
Após o processamento dos dados de sequenciamento, foi obtido um conjunto de dados com 10.697 ASVs e 38.755.631 leituras totais, distribuídos entre as amostras. Após a filtragem com o PERFect, 74% dos ASVs foram removidos, restando 2.799 ASVs, como pode ser observado na Figura 5.2. Entretanto, apenas 0,8% das leituras foram removidas em consequência disso.

Tabela 5.1: Desbalanceamento dos fenótipos binários

Fenótipo	Controle	Caso
DRGE	82,7%	17,3%
Transtorno mental	85,5%	14,5%
Síndrome do intestino irritável	87,8%	12,2%
Depressão	87,8%	12,2%
Doenças autoimunes	88,9%	11,1%
Doença pulmonar	89%	11%
Câncer	91,3%	8,7%
Doença cardiovascular	96,1%	3,9%
SIFO	96,5%	3,5%
DII	96,6%	3,4%
Diabetes	96,7%	3,3%
TDAH	97%	3%
Doença renal	98,2%	1,8%
SIBO	98,2%	1,8%
Doença hepática	98,3%	1,7%
Bipolaridade	98,5%	1,5%
Clostridium difficile	98,8%	1,2%
Epilepsia	98,9%	1,1%
Anorexia nervosa	99%	1%
TEPD	99%	1%
Esquizofrenia	99%	1%
Bulimia	99,3%	0,7%
Abuso de substância	99,3%	0,7%
Autismo	99,5%	0,5%

Em negrito estão os fenótipos escolhidos para a modelagem.

Figura 5.2: Perda de leitura após filtragem com método PERFect.



Curva de perda de filtragem do conjunto de ASVs. Os valores de DFL (em inglês, *Differences in Filtering Loss*) são apresentados. O gráfico demonstra que o conjunto ordenado até o número de 7.898 ASVs não apresenta perda da covariância total.

5.2 Resultados de desempenho dos modelos

Os modelos apresentados são orientados a *recall*, uma vez que a intenção do teste é servir meramente como um auxílio ao diagnóstico, ou seja, um resultado predito positivo para as doenças analisadas implica na recomendação de que o paciente busque auxílio médico para o diagnóstico por outras técnicas, e, em caso de confirmação da predição por estas, realize o tratamento adequado. Desta forma, decide-se minimizar FN, ao invés da minimização dos FP. Mesmo que o paciente não possua a doença (FP), a busca de auxílio médico desnecessária é menos prejudicial quanto predizer que o paciente não possui determinada condição, quando ele a possui de fato (FN).

5.2.1 Resultados avaliados na validação cruzada

As Figuras 5.3 e 5.4 apresentam *boxplots* das métricas *recall* e F1-*score* na validação cruzada dos modelos treinados. O ponto preto que secciona a caixa representa a mediana, e as linhas azuis extremas na esquerda e na direita representam o limite inferior e superior, respectivamente. A primeira altura, à esquerda da caixa, representa o primeiro quartil, e a segunda altura, à direita da caixa, representa o terceiro quar-

til. Os modelos foram divididos por 1. “glmnet”, 2. “gbm” e 3. “rf”. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, transtorno mental. As amostragens são “down”, *undersample*; “smote”, SMOTE. O agrupamento de dados é dividido por “semquest” quando não há dados adicionais fenotípicos e “comquest” quando há dados adicionais fenotípicos. Por exemplo, “gbm.smote.semquest” representa um modelo de *gradient boosting* com amostragem SMOTE sem dados de questionário. A linha vermelha na Figura 5.3 foi heuristicamente aplicada no valor 0,6 pra facilitar a visualização do método que possui mais valores medianos acima do mesmo, assim como na Figura 5.4 o valor de 0,3. As restantes métricas possuem seus valores apresentados no Apêndice.

Nessa etapa de investigação inicial, foram obtidos valores de recall com alta variabilidade (Figura 5.3), com cerca 5 a 6 modelos acima de 0,6 de recall em cada fenótipo. Observa-se que todos modelos apresentaram uma taxa de acerto da predição, ou precisão, bem baixa, com a mediana não ultrapassando 30%, por conseguinte prejudicando a métrica F1-*Score* (Figura 5.4) e a utilização dos modelos. O fenótipo DII, o mais desbalanceado dos seis, foi o mais prejudicado nessa métrica, sugerindo que necessita-se de mais casos. Os modelos possuem boa especificidade, conseguindo identificar bem a classe negativa de uma pessoa “saudável”. A regressão logística penalizada foi o algoritmo modelo mais bem avaliado quanto ao valor de F1-*score*.

Figura 5.3: *Recall* dos modelos na validação cruzada

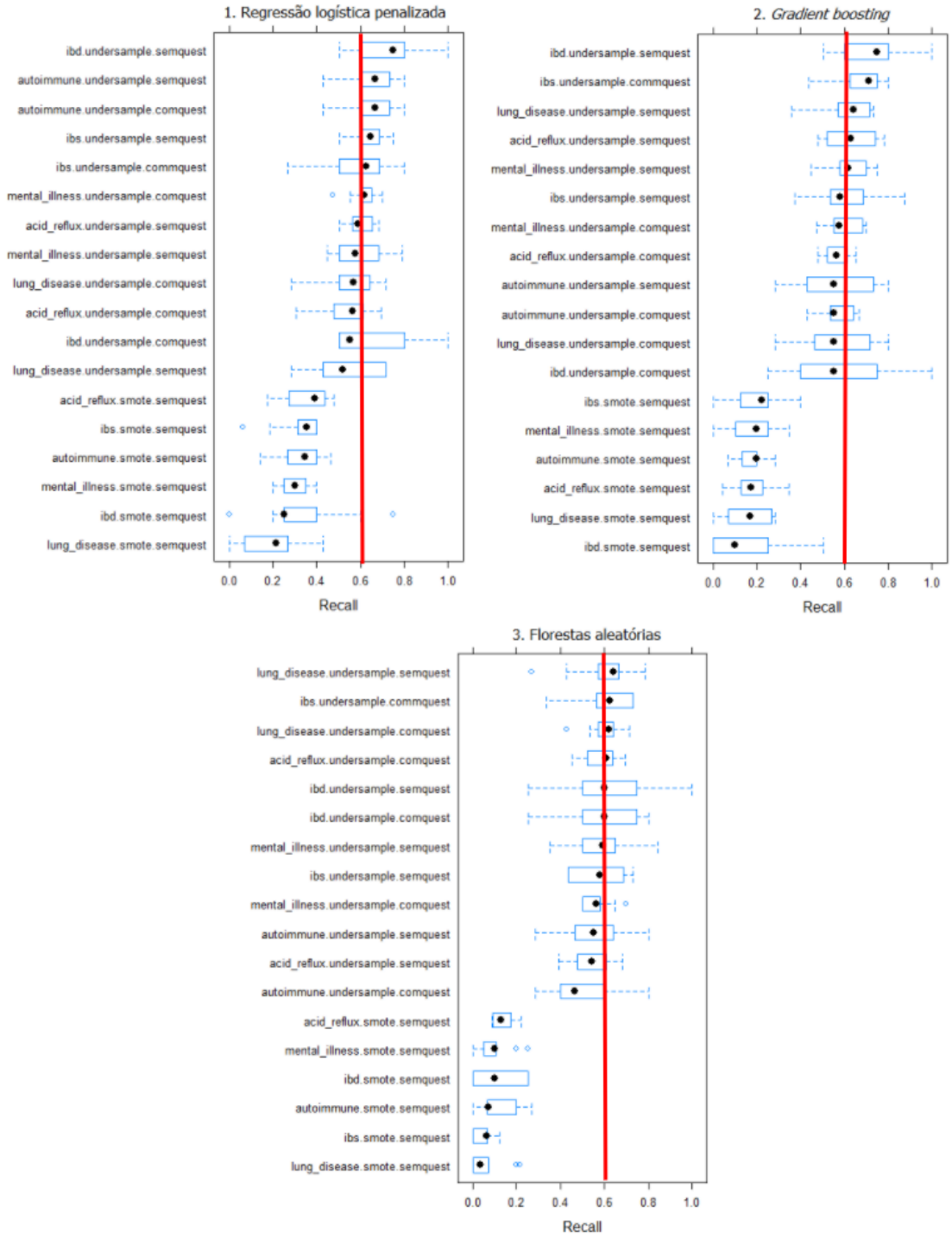


Diagrama de caixa dos valores de *recall* na validação cruzada de cada modelo.

Figura 5.4: F1-*Score* dos modelos na validação cruzada

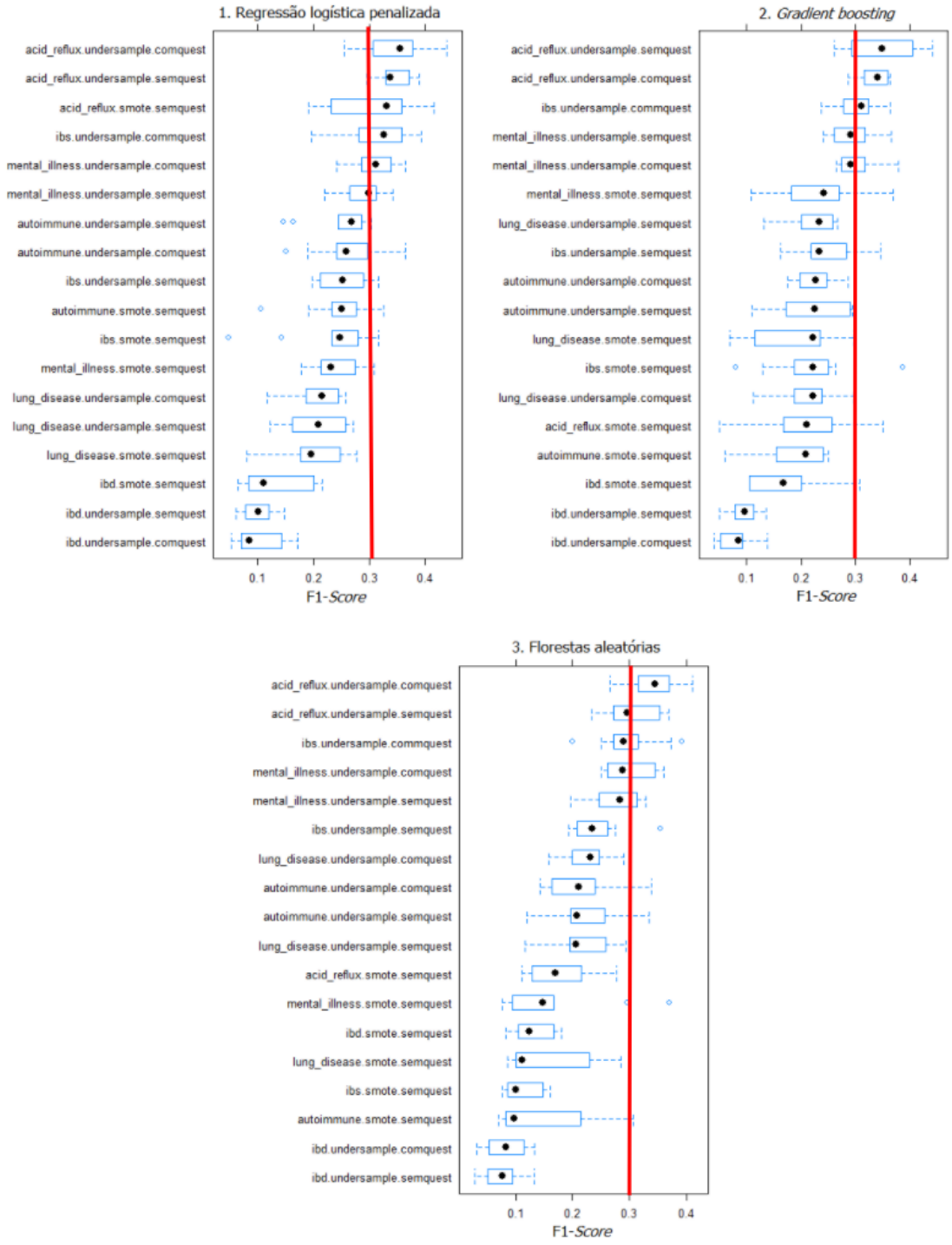
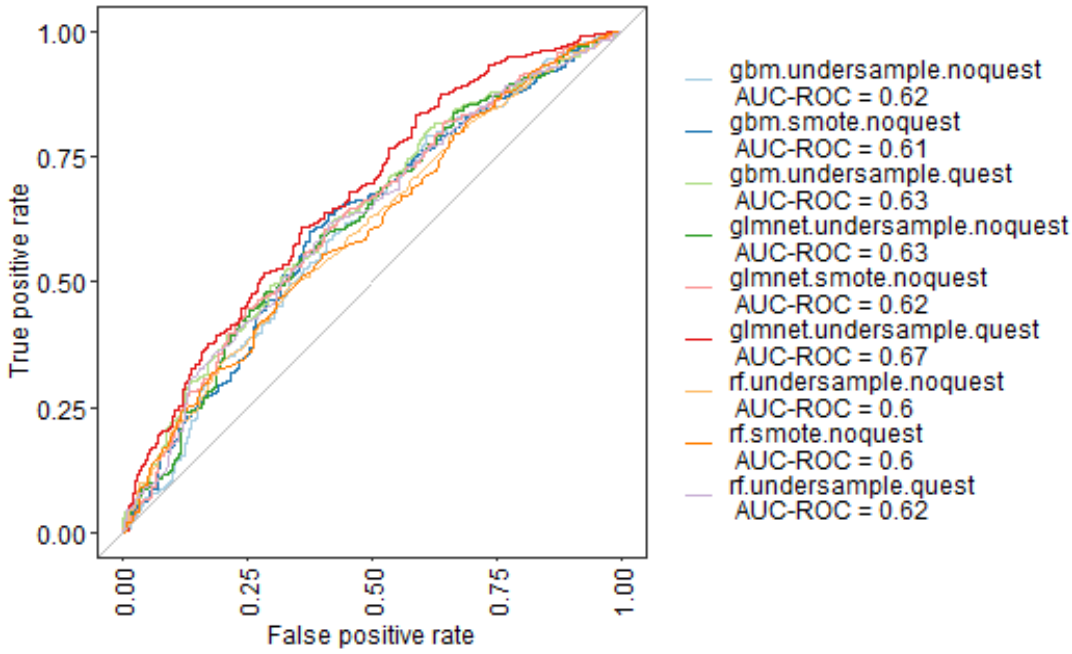


Diagrama de caixa dos valores de F1-*score* na validação cruzada de cada modelo.

As Figuras 5.5 - 5.10 apresentam as curvas ROC da validação cruzada dos diferentes modelos treinados. Dos seis fenótipos analisados, cinco tiveram “glmnet” como o melhor AUC, corroborando as afirmações de Dong et al. (2020) que regressão logís-

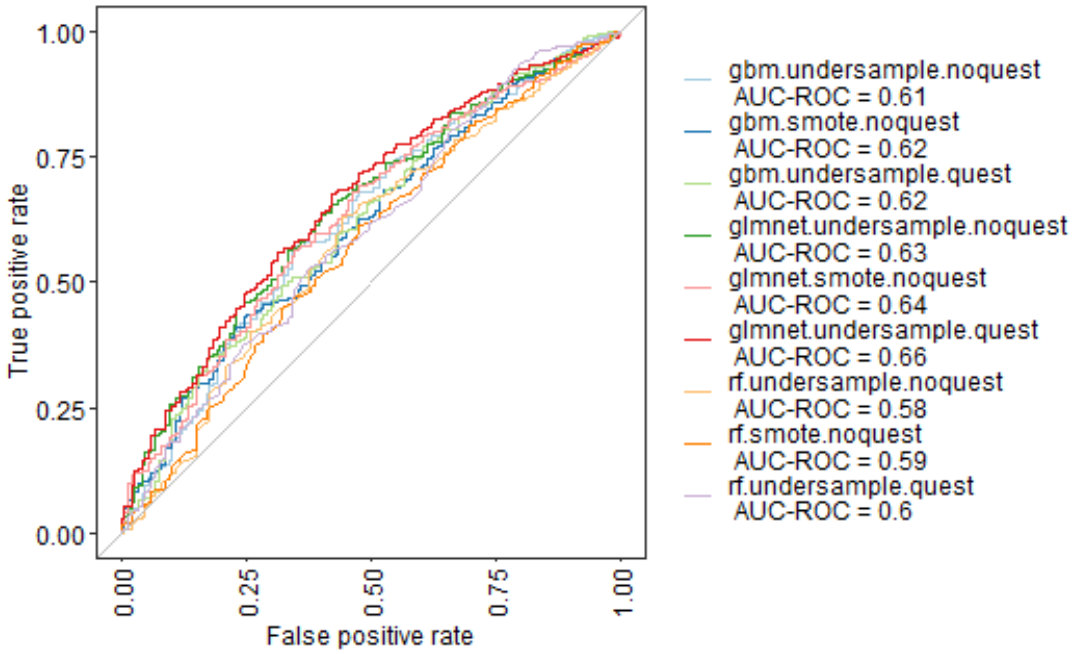
tica possui uma performance tão boa quanto modelos mais complexos, como florestas aleatórias. A exceção à regra foi o fenótipo de doenças pulmonares, que possui o *gradient boosting* como o modelo de melhor performance nessa métrica, no valor de 0.62. Além disso, o fenótipo foi o de pior performance geral com o restante, levando a concluir que o conjunto de dados obtido não foi capaz de extrair relações suficientes que levassem a conclusões parecidas às de Bowerman et al. (2020). Os valores de métricas AUC na escolha de diferentes amostragens não resultaram em notável diferença, com exceção do fenótipo de doenças pulmonares. Afirma-se também que conjunto de dados com questionário foram os que possuíram melhor AUC, apesar de não terem sido executados treinamentos com amostragem SMOTE e dados de questionário, os quais serão executados futuramente. Além disso, os piores AUCs foram obtidos com os modelos de florestas aleatórias.

Figura 5.5: AUC do fenótipo GRDE na validação cruzada



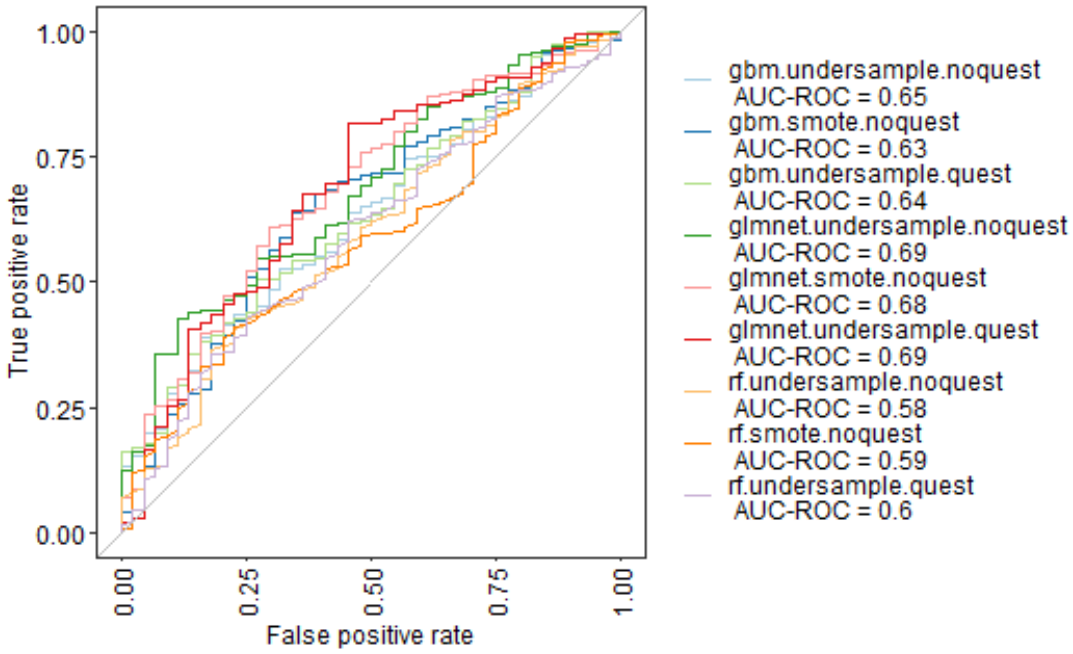
Curva ROC na validação cruzada de cada modelo treinado para predizer GRDE. Menor valor AUC obtido foi 0,6 e o maior foi 0,67.

Figura 5.6: AUC do fenótipo de doenças autoimunes na validação cruzada



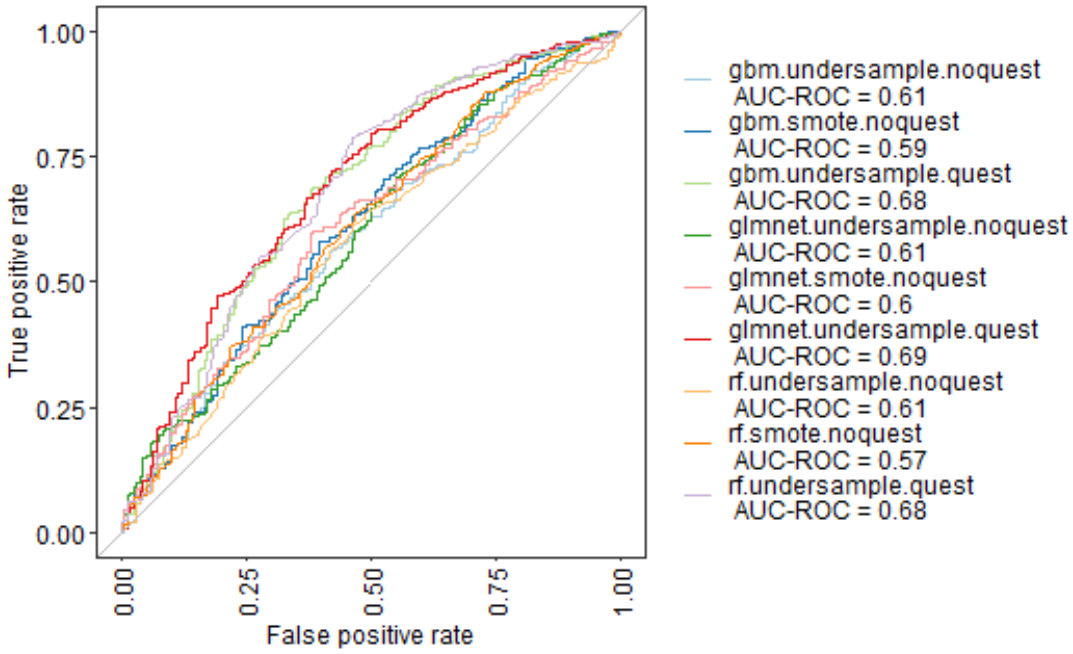
Curva ROC na validação cruzada de cada modelo treinado para predizer doenças autoimunes. Menor valor AUC obtido foi 0,58 e o maior foi 0,66.

Figura 5.7: AUC do fenótipo DII na validação cruzada



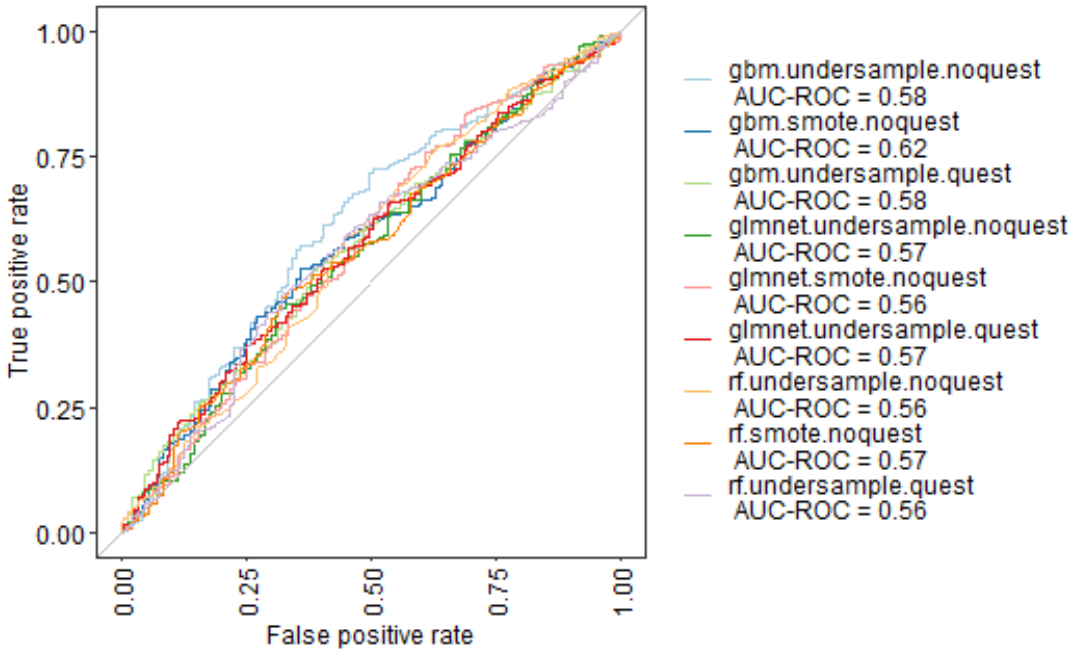
Curva ROC na validação cruzada de cada modelo treinado para predizer DII. Menor valor AUC obtido foi 0,58 e o maior foi 0,69.

Figura 5.8: AUC do fenótipo de síndrome do intestino irritável na validação cruzada



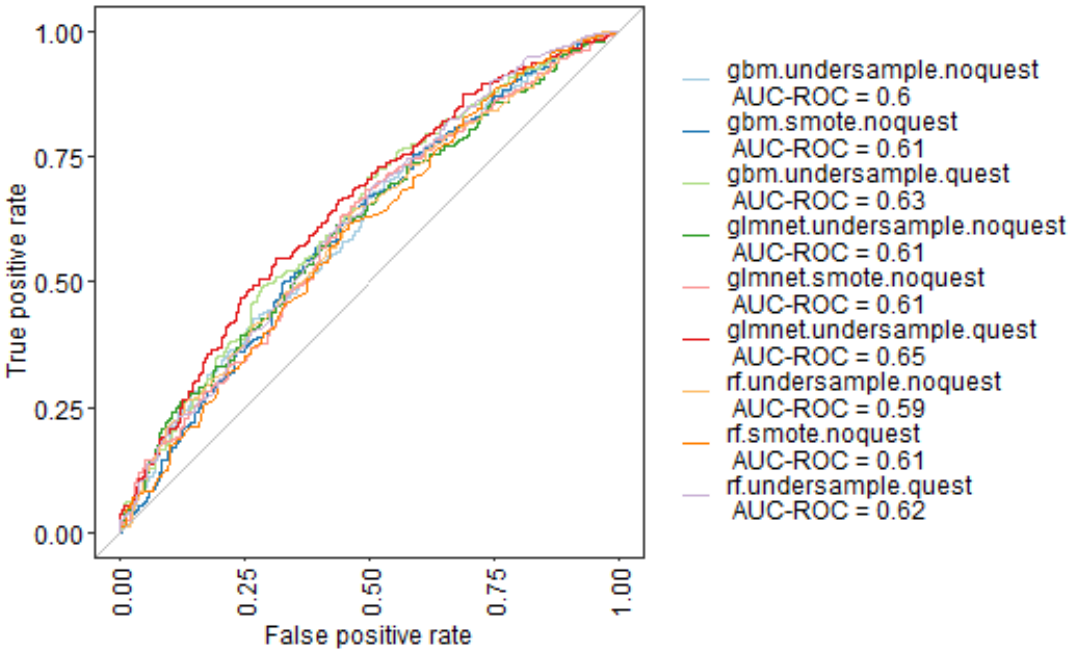
Curva ROC na validação cruzada de cada modelo treinado para predizer síndrome do intestino irritável. Menor valor AUC obtido foi 0,57 e o maior foi 0,69.

Figura 5.9: AUC do fenótipo doenças pulmonares na validação cruzada



Curva ROC na validação cruzada de cada modelo treinado para predizer doenças pulmonares. Menor valor AUC obtido foi 0,56 e o maior foi 0,62.

Figura 5.10: AUC do fenótipo de transtornos mentais na validação cruzada



Curva ROC na validação cruzada de cada modelo treinado para predizer transtornos mentais. Menor valor AUC obtido foi 0,59 e maior foi 0,65.

O fenótipo de DII e doenças pulmonares apresentaram as piores métricas de desempenho ao analisar o F1-*Score* em conjunto com *recall* dos modelos na Tabela 5.2, corroborando com as afirmações anteriores. Foi possível observar que o desbalanceamento das classes afetou o treinamento dos modelos, pois os fenótipos mais bem avaliados foram os menos desbalanceados. Percebe-se também uma preferência pela integração de dados adicionais fenotípicos.

Tabela 5.2: Indicadores de desempenhos médios dos melhores modelos na validação cruzada

Nome do modelo	acc	prec	rec	F1	AUC
acid_reflux.glmnet.undersample.comquest	0,64	0,25	0,55	0,34	0,67
mental_illness.glmnet.undersample.comquest	0,6	0,21	0,61	0,31	0,65
ibs.gbm.undersample.semquest	0,63	0,2	0,67	0,3	0,68
autoimmune.glmnet.undersample.comquest	0,59	0,17	0,65	0,26	0,66
lung_disease.gbm.undersample.comquest	0,55	0,13	0,57	0,21	0,57
ibd.gbm.undersample.semquest	0,53	0,05	0,72	0,1	0,65

Indicadores de desempenhos médios dos melhores modelos na validação cruzada. As métricas a serem maximizadas nessas escolhas foram uma combinação de F1-*Score* e *recall*.

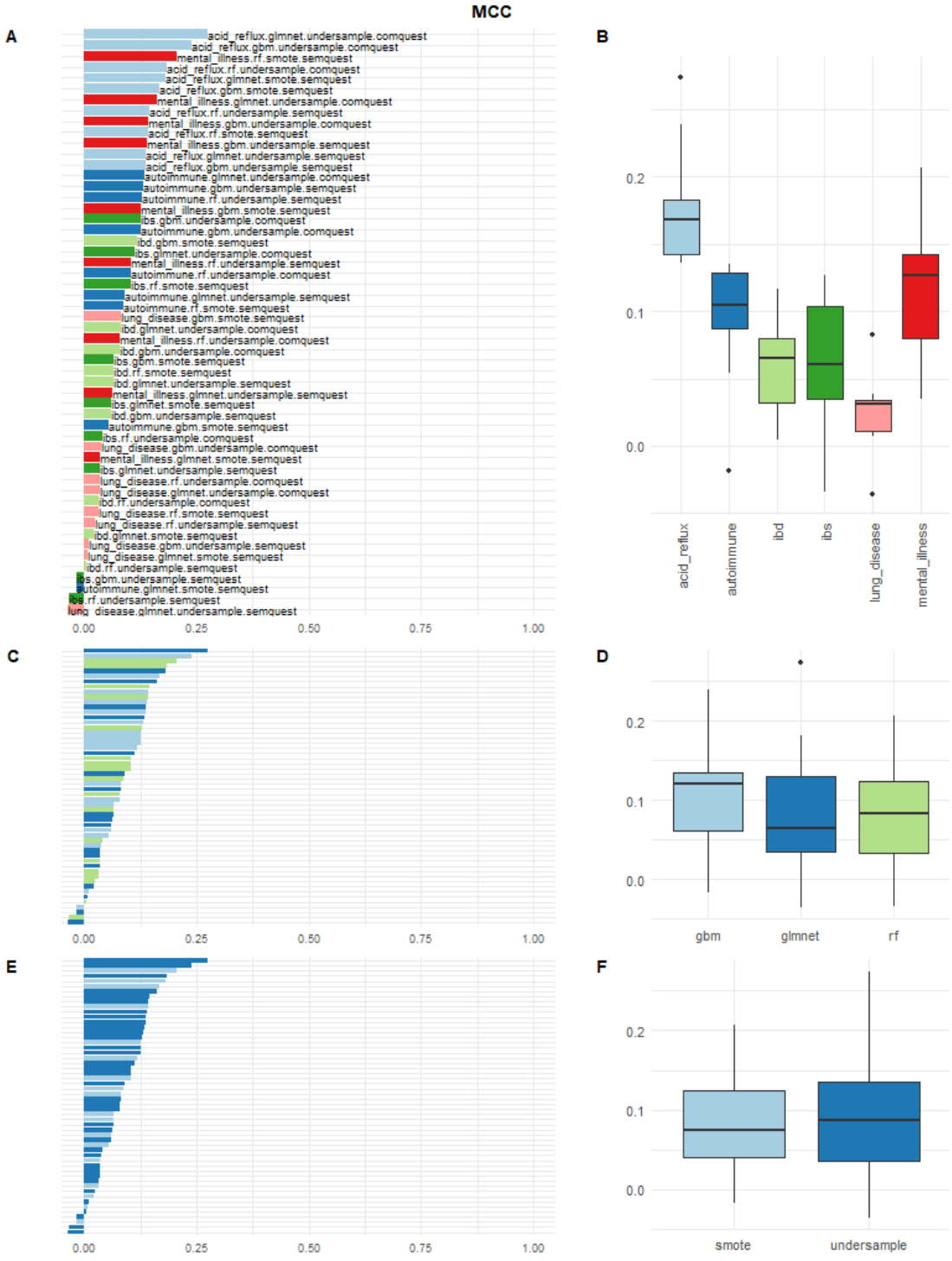
5.2.2 Resultados avaliados nas aplicações aos conjuntos de teste

A próxima análise apresentada envolve a aplicação dos conjuntos de dados de teste, que preservaram as proporções de desbalanceamento de cada fenótipo, nos modelos treinados. Durante o processo de *tuning* dos hiperparâmetros da validação cruzada, as instâncias *holdout* são avaliadas e podem não refletir o desbalanceamento que futuras predições encontrarão, resultando em estimativas de performance superestimadas. Na avaliação dos modelos, a acurácia não é a métrica apropriada, já que não distingue as classificações corretas de diferentes classes. O MCC, que só apresenta altos valores quando as quatro categorias da matriz de confusão possuem bons resultados, está apresentado na Figura 5.11), assim como *recall* na Figura 5.12. Estas duas foram as principais métricas escolhidas para a obtenção da Tabela 5.3. O fenótipo DRGE, o menos desbalanceado, foi o que possuiu maior valor MCC, de 0,27. Nas Figuras 5.11- 5.12 foram apresentados diferentes dados coloridos por A) fenótipo, C) tipo de modelo e E) tipo de amostragem. Há também o boxplot da métrica para cada B) fenótipo, D) tipo de modelo e F) tipo de amostragem.

Observa-se que os dados adicionais fenotípicos foram importantes para aprimorar a performance dos modelos, e que *undersampling* prevaleceu como a melhor forma de amostragem. Apesar de possuir uma ótima especificidade em geral, o SMOTE falha

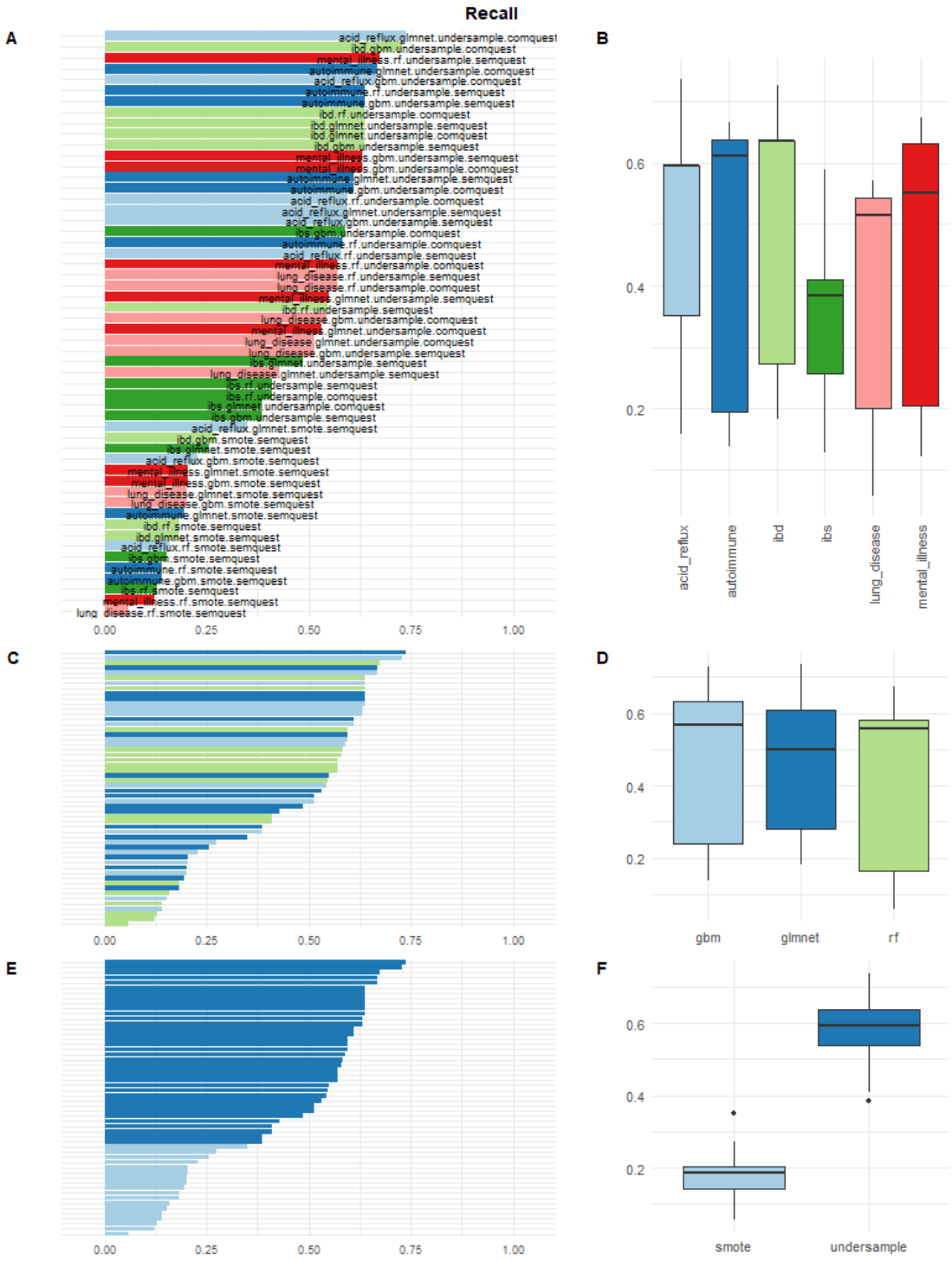
na métrica *recall*, descartando-se seu uso como o melhor modelo. Nessa métrica, o *undersampling* possuiu melhores resultados. Ao contrário de outros trabalhos relacionados, modelos de FA não aparecem como o melhor modelo em nenhum dos fenótipos, sendo o método com pior desempenho geral.

Figura 5.11: MCC considerando dados do conjunto de teste



Métricas de MCC para o conjunto de dados de teste de cada um dos modelos obtidos.

Figura 5.12: *Recall* considerando dados do conjunto de teste

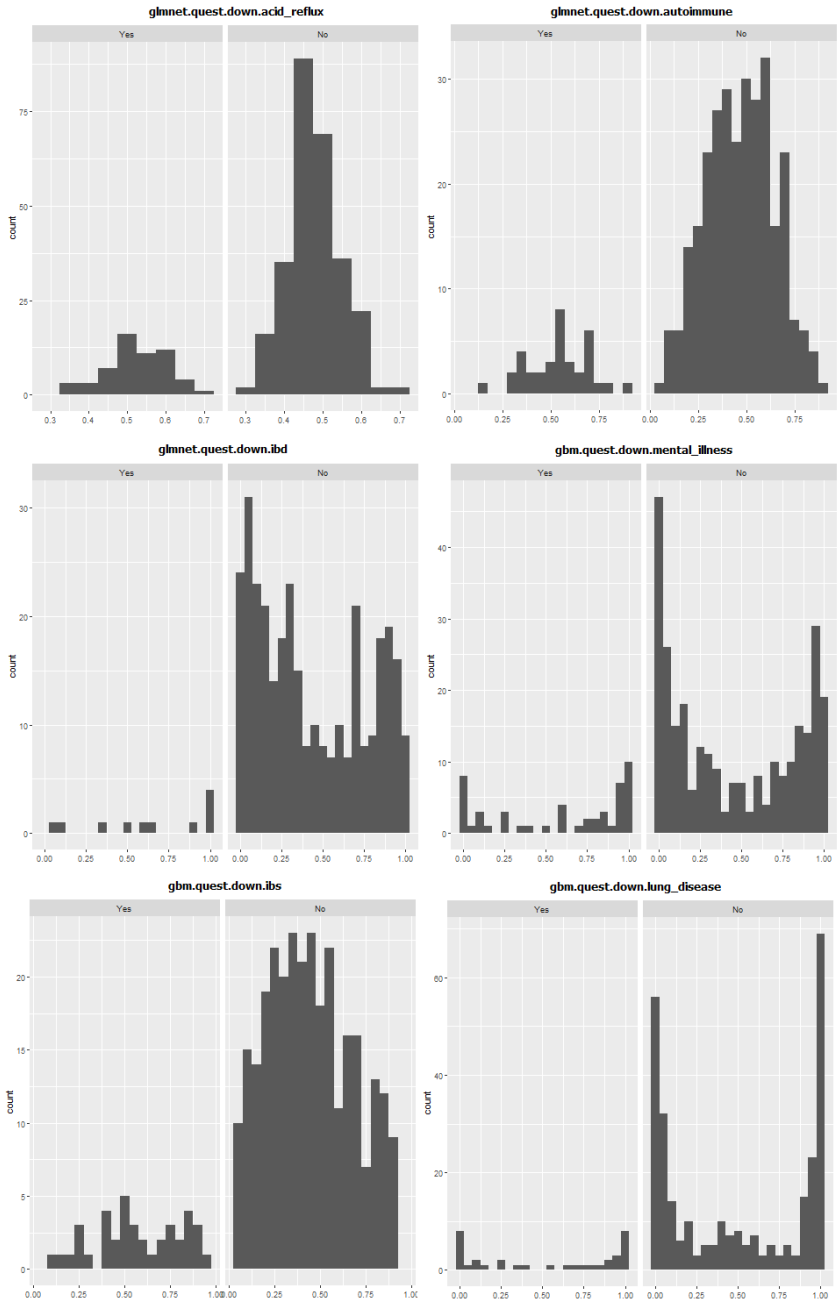


Métricas de *recall* para o conjunto de dados de teste de cada um dos modelos obtidos.

A Figura 5.13 ilustra as curvas de probabilidade da classe positiva e o quanto de certeza os modelos possuem para predizer a classe, ou seja, o quanto mais afastado de uma escolha aleatória o valor de probabilidade foi. O eixo x representa a probabilidade

da classe “Yes” (possuir o fenótipo). Quanto mais negativa for a assimetria da curva para a decisão por “Yes”, bem como, quanto mais positiva para a decisão por “No”, melhor é o desempenho do modelo. Porém nos fenótipos apresentados, nenhum se destacou como o de melhor desempenho nesse quesito.O limiar de decisão para classificação foi de 0,5 para todos os modelos.

Figura 5.13: Curvas de probabilidade da predição para a classe positiva de cada fenótipo



Idealmente, as curvas devem apresentar assimetria, mas nenhuma das predições obtidas apresentou a assimetria esperada nas suas curvas.

A tarefa de encontrar assinaturas de microbioma globais para a deteção de fenótipos humanos é bastante complexa, pois aspectos geográficos locais não podem ser descartados. A ótica de predição de doenças com intuito de auxílio médico, ao invés de total automatização e substituição do profissional, poderia levar a considerar como aceitáveis modelos de ML com performance média de predição. Porém, no presente trabalho, isso não foi obtido. O baixo desempenho geral dos modelos treinados com doenças pulmonares leva a crer que atualmente no conjunto de dados do AGP não exista correlação entre os dados de microbioma e a presença desse fenótipo.

Tabela 5.3: Indicadores de desempenhos dos melhores modelos na aplicação aos conjuntos de teste

Nome do modelo	MCC	prec	rec	F1	AUC	espec
acid_reflux.glmnet.undersample.comquest	0,27	0,29	0,74	0,42	0,7	0.62
mental_illness.gbm.undersample.comquest	0,14	0,2	0,63	0,31	0,61	0.57
autoimmune.glmnet.undersample.comquest	0,13	0,16	0,67	0,25	0,71	0,55
ibs.gbm.undersample.comquest	0,13	0,17	0,59	0,26	0,65	0.6
ibd.glmnet.undersample.comquest	0,08	0,05	0,64	0,09	0,69	0.59
lung_disease.gbm.undersample.comquest	0,04	0,12	0,54	0,19	0,5	0,52

Indicadores de desempenhos dos melhores modelos na aplicação aos conjuntos de teste. As métricas a serem maximizadas nessas escolhas foram MCC e *recall*.

5.2.3 Importância dos atributos

A correlação com o restante dos fenótipos não foi descartada devido às Tabelas 5.4 - 5.9 apresentarem resultados potencialmente relevantes na área de microbioma intestinal. Diferentes bactérias intestinais não influenciam de forma equivalente os processos que ocorrem no microbioma, e algumas bactérias específicas são capazes de ter um impacto muito maior que outras, mesmo não estando em maiores quantidades. Essas bactérias são como “espécies-chave”, e teriam a capacidade de influenciar todo o microbioma por estarem altamente conectadas com outras espécies. Por exemplo, uma molécula produzida por uma espécie-chave pode ser metabolizada por outra espécie (KLITGORD; SEGRÈ, 2010). *Ruminococcus* aparece como um atributo importante para a predição de GRDE. *Ruminococcus bromii* é um candidato a espécie-chave por atuar na liberação de

energia de fibra dietética e amido resistente no cólon humano, possuindo uma capacidade excepcional quando comparada com outras bactérias amilolíticas (ZE et al., 2012).

Vários membros da família *Lachnospiraceae* foram detectados como relevantes para todas as doenças avaliadas. Diferenças de abundância em *Lachnospiraceae* já foram encontradas em portadores de diferentes condições de saúde, tanto em humanos quanto em outros modelos animais, por exemplo em DII (SASAKI et al., 2019), transtorno depressivo (JIANG et al., 2015), diabetes (KOSTIC et al., 2015), doenças metabólicas e outros (VACCA et al., 2020). Os gêneros *Bifidobacterium* (encontrado em DII e doenças autoimunes) e *Akkermansia* (encontrado em síndrome do intestino irritado) vem sendo considerados como importantes marcadores de saúde do sistema gastrointestinal. *Bifidobacterium* é amplamente utilizado em probióticos, a estirpe BB536 de *Bifidobacterium longum* subsp. *longum*, por exemplo, é descrita como um probiótico clinicamente eficaz e com longa história de uso humano no alívio de doenças gastrointestinais, imunológicas e infecciosas (WONG; ODAMAKI; XIAO, 2019). *Akkermansia*, por sua vez, é reconhecida por colonizar a camada de mucosa do intestino humano, atuando no aumento da espessura do muco e no aumento da função de barreira intestinal (OTTMAN et al., 2017).

Tabela 5.4: Importância dos atributos para “glmnet.quest.down.acid_reflux”

Taxonomia	Importância da variável
age_years	100
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Coprococcus	72,4
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Colidextribacter	48,6
lactoseYes	46,7
Firmicutes - Bacilli - RF39 - order_RF39 - order_RF39	41,3
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Subdoligranulum	37,4
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	34,9
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Ruminococcus	31,6
Desulfobacterota - Desulfovibrionia - Desulfovibrionales - Desulfovibrionaceae - Bilophila	26
Firmicutes - Bacilli - Lactobacillales - Streptococcaceae - Streptococcus	25,4
Firmicutes - Clostridia - Clostridia UCG014 - order_Clostridia UCG014 - order_Clostridia UCG014	22,3
antibiotic_historyI.have.not.taken.antibiotics.in.the.past.year.	19,5
Desulfobacterota - Desulfovibrionia - Desulfovibrionales - Desulfovibrionaceae - Desulfovibrio	15,3
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	14,1
Firmicutes - Clostridia - Christensenellales - Christensenellaceae - Christensenellaceae R-7 group	13,9
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - UCG-003	13,7
Proteobacteria - Gammaproteobacteria - Enterobacterales - order_Enterobacterales - order_Enterobacterales	11,4
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Fournierella	7,4
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - NK4A214 group	6,4
Proteobacteria - Gammaproteobacteria - Enterobacterales - order_Enterobacterales - order_Enterobacterales	5,4
Firmicutes - Clostridia - Oscillospirales - [Eubacterium coprostanoligenes group - family_[Eubacterium] coprostanoligenes group	2,9
Firmicutes - Bacilli - RF39 - order_RF39 - order_RF39	0,7
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Anaerotruncus	0,5

Importância dos atributos para o mais bem avaliado modelo para o fenótipo GRDE.

Tabela 5.5: Importância dos atributos para “glmnet.quest.down.autoimmune”

Taxonomia	Importância da variável
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	100
bowel_movement_qualityI.tend.to.have.normal.formed.stool...Type.3.and.4	74,8
probiotic_frequencyNever	61,9
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospiraceae UCG-001	59,1
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - [Eubacterium] siraeum group	53,9
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Flavonifractor	51,3
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	46,7
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	42,3
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	39,8
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	35,9
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	35
Firmicutes - Clostridia - Lachnospirales - Defluviitaleaceae - Defluviitaleaceae UCG-011	33,4
Firmicutes - Clostridia - Clostridia vadinBB60 group - order_Clostridia vadinBB60 group - order_Clostridia vadinBB60 group	32,6
exercise_frequencyNever	31,5
Bacteroidota - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes	30,2
Proteobacteria - Alphaproteobacteria - Rhizobiales - Rhizobiaceae - family_Rhizobiaceae	28,8
Actinobacteriota - Actinobacteria - Bifidobacteriales - Bifidobacteriaceae - Bifidobacterium	28,5
Firmicutes - Clostridia - Oscillospirales - [Eubacterium] coprostanoligenes group - family_[Eubacterium] coprostanoligenes group	28,3
Desulfobacterota - Desulfovibrionia - Desulfovibrionales - Desulfovibrionaceae - Bilophila	27,6
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - family_Ruminococcaceae	26
Proteobacteria - Gammaproteobacteria - Pseudomonadales - Moraxellaceae - family_Moraxellaceae	24
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	23,4
Firmicutes - Bacilli - Bacillales - Bacillaceae - family_Bacillaceae	23,1
Proteobacteria - Alphaproteobacteria - Caulobacterales - Caulobacteraceae - Brevundimonas	22,5
diet_typeVegetarian	22,2
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	21,8
Actinobacteriota - Actinobacteria - Bifidobacteriales - Bifidobacteriaceae - Bifidobacterium	21,8
Proteobacteria - Gammaproteobacteria - Pasteurellales - Pasteurellaceae - family_Pasteurellaceae	21,6
Actinobacteriota - Coriobacteriia - Coriobacteriales - Eggerthellaceae - Eggerthella	20,2

Importância dos atributos para o mais bem avaliado modelo para o fenótipo “doenças autoimunes”.

Tabela 5.6: Importância dos atributos para “glmnet.quest.down.ibd”

Taxonomia	Importância da variável
Firmicutes - Negativicutes - Acidaminococcales - Acidaminococcaceae - Phascolarctobacterium	100
bowel_movement_qualityI.tend.to.have.diarrhea..watery.stool....Type.5..6.and.7	93,6
bowel_movement_qualityI.tend.to.have.normal.formed.stool...Type.3.and.4	73
Firmicutes - Negativicutes - Veillonellales-Selenomonadales - Veillonellaceae - Dialister	66,2
probiotic_frequencyRarely..a.few.times.month.	62,8
Actinobacteriota - Actinobacteria - Bifidobacteriales - Bifidobacteriaceae - Bifidobacterium	62,6
probiotic_frequencyNever	60,3
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Roseburia	60
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospiraceae UCG-001	59,6
Firmicutes - Bacilli - Lactobacillales - Lactobacillaceae - Lactobacillus	59,4
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Colidextribacter	56
Firmicutes - Clostridia - Peptostreptococcales-Tissierellales - Anaerovoracaceae - Family XIII UCG-001	53
Firmicutes - Clostridia - Oscillospirales - UCG-011 - family_UCG-011	53
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - family_Ruminococcaceae	48,8
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Intestinimonas	47,4
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	45,7
Proteobacteria - Gammaproteobacteria - Xanthomonadales - Xanthomonadaceae - Pseudoxanthomonas	44,4
Firmicutes - Bacilli - RF39 - order_RF39 - order_RF39	42,6
Firmicutes - Clostridia - Oscillospirales - UCG-011 - family_UCG-011	42,3
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	42,2
Firmicutes - Clostridia - Clostridia UCG-014 - order_Clostridia UCG-014 - order_Clostridia UCG-014	40,5
Desulfobacterota - Desulfovibrionia - Desulfovibrionales - Desulfovibrionaceae - Bilophila	39,9
Bacteroidota - Bacteroidia - Bacteroidales - Barnesiellaceae - Barnesiella	39,7
Bacteroidota - Bacteroidia - Bacteroidales - Prevotellaceae - Paraprevotella	38,7
Firmicutes - Clostridia - Christensenellales - Christensenellaceae - Christensenellaceae R-7 group	34,8
Firmicutes - Bacilli - Erysipelotrichales - Erysipelotrichaceae - Holdemanella	34,7
Proteobacteria - Gammaproteobacteria - Enterobacterales - Morganellaceae - Morganella	34,6
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	34,5
Firmicutes - phylum_Firmicutes - phylum_Firmicutes - phylum_Firmicutes - phylum_Firmicutes	34,4

Importância dos atributos para o mais bem avaliado modelo para o fenótipo DII.

Tabela 5.7: Importância dos atributos para “gbm.quest.down.mental_illness”

Taxonomia	Importância da variável
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - family_Ruminococcaceae	100
bmi	85,5
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospira	81
alcohol_frequencyNever	61,2
Proteobacteria - Gammaproteobacteria - Pseudomonadales - Pseudomonadaceae - Pseudomonas	60,2
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	48
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	46,2
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - family_Oscillospiraceae	42
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	37,8
Firmicutes - Negativicutes - Veillonellales-Selenomonadales - Veillonellaceae - Allisonella	37,8
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnoclostridium	36,5
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	34,5
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	30,4
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	29,5
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Intestinimonas	29,4
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	28,2
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	27,7
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Oscillibacter	27,4
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	27,4
Proteobacteria - Gammaproteobacteria - Enterobacterales - Enterobacteriaceae - family_Enterobacteriaceae	27
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Coprococcus	26
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Faecalibacterium	25,6
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	24,4
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Roseburia	23,4
Bacteroidota - Bacteroidia - Bacteroidales - Prevotellaceae - Prevotella	23,1
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Colidextribacter	22,5
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	22,4
Bacteroidota - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes	22,2
Proteobacteria - Gammaproteobacteria - Enterobacterales - Morganellaceae - family_Morganellaceae	21,9

Importância dos atributos para o mais bem avaliado modelo para o fenótipo “transtornos metais”.

Tabela 5.8: Importância dos atributos para “gbm.quest.down.ibs”

Taxonomia	Importância da variável
bowel_movement_qualityI.tend.to.have.normal.formed.stool...Type.3.and.4	100
Verrucomicrobiota - Verrucomicrobiae - Verrucomicrobiales - Akkermansiaceae - Akkermansia	39,7
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospiraceae NC2004 group	34,2
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - family_Oscillospiraceae	28,6
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	27,9
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	25,5
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	22,8
lactoseYes	22,3
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	21,7
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	21,2
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Subdoligranulum	20
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - Colidextribacter	20
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospiraceae UCG-001	19,3
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	17
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospira	16,7
Proteobacteria - Gammaproteobacteria - Enterobacterales - Enterobacteriaceae - family_Enterobacteriaceae	16,6
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	16,4
Firmicutes - Bacilli - Erysipelotrichales - Erysipelotrichaceae - Holdemanella	16
Bacteroidota - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes	15,7
Firmicutes - Bacilli - Erysipelotrichales - Erysipelatoclostridiaceae - Catenibacterium	15,4
Desulfobacterota - Desulfovibrionia - Desulfovibrionales - Desulfovibrionaceae - Bilophila	14,9
Proteobacteria - Gammaproteobacteria - Enterobacterales - order_Enterobacterales - order_Enterobacterales	14,8
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	14,6
Firmicutes - phylum_Firmicutes - phylum_Firmicutes - phylum_Firmicutes - phylum_Firmicutes	14
Bacteroidota - Bacteroidia - Bacteroidales - Marinifilaceae - Odoribacter	13,9
Firmicutes - Clostridia - Peptostreptococcales-Tissierellales - Peptostreptococcaceae - family_Peptostreptococcaceae	13,7
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	13,6
Verrucomicrobiota - Verrucomicrobiae - Verrucomicrobiales - Akkermansiaceae - Akkermansia	13,6
Firmicutes - Clostridia - Oscillospirales - Butyricicoccaceae - Butyricicoccus	13,6

Importância dos atributos para o mais bem avaliado modelo para o fenótipo síndrome do intestino irritado.

Tabela 5.9: Importância dos atributos para “gbm.quest.down.lung_disease”

Taxonomia	Importância da variável
bmi	100
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	55,5
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	45,2
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Subdoligranulum	38,6
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	38,3
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Ruminococcus	38,2
Firmicutes - Clostridia - Oscillospirales - Oscillospiraceae - family_Oscillospiraceae	37,2
Verrucomicrobiota - Verrucomicrobiae - Verrucomicrobiales - Akkermansiaceae - Akkermansia	36,5
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - Lachnospiraceae FCS020 group	34,9
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	34,3
Bacteroidota - Bacteroidia - class_Bacteroidia - class_Bacteroidia - class_Bacteroidia	32,6
Firmicutes - Clostridia - Lachnospirales - Lachnospiraceae - family_Lachnospiraceae	32,3
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	31,4
Bacteroidota - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes	31,2
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	30,7
Firmicutes - Clostridia - Oscillospirales - Butyricicoccaceae - Butyricicoccus	29,6
Firmicutes - Clostridia - Oscillospirales - UCG-011 - family_UCG-011	29,2
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	27,7
exercise_frequencyOccasionally..1.2.times.week.	27,6
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	27
Actinobacteriota - Coriobacteriia - Coriobacteriales - Eggerthellaceae - Eggerthella	26
Proteobacteria - Gammaproteobacteria - Enterobacterales - order_Enterobacterales - order_Enterobacterales	24,8
Firmicutes - Clostridia - class_Clostridia - class_Clostridia - class_Clostridia	24,4
Firmicutes - Clostridia - Oscillospirales - Ruminococcaceae - Subdoligranulum	23,7
Firmicutes - Bacilli - Lactobacillales - Streptococcaceae - Lactococcus	23,7
Firmicutes - Bacilli - Erysipelotrichales - Erysipelotrichaceae - Holdemanella	23,7
Bacteroidota - Bacteroidia - Bacteroidales - Rikenellaceae - Alistipes	23,6
Bacteroidota - Bacteroidia - Bacteroidales - Bacteroidaceae - Bacteroides	23,3
Bacteroidota - Bacteroidia - Bacteroidales - Tannerellaceae - Parabacteroides	21,5

Importância dos atributos para o mais bem avaliado modelo para o fenótipo “doenças pulmonares”.

6 CONCLUSÕES

Este estudo demonstra que a composição do microbioma intestinal pode ser utilizada para a elaboração de modelos de ML, apresentando estatísticas de performance de modelo que ultrapassam a predição ao acaso para GRDE, DII, síndrome do intestino irritável, doenças autoimunes, doenças pulmonares e transtornos mentais. Entre os diferentes classificadores testados, “glmnet” e “gbm” se mostraram mais precisos para prever os fenótipos com base no microbioma intestinal e dados de saúde, em comparação com FA. Mais trabalhos são necessários para compreender a relação do microbioma intestinal e os fenótipos testados, seja aplicando técnicas adicionais de ML ou avaliando novos conjuntos de dados, a fim de elaborar estratégias que pavimentem o caminho para futuras terapias orientadas para o microbioma, bem como diagnósticos baseados no mesmo.

7 TRABALHOS FUTUROS

Descartar a análise de fenótipos onde o microbioma intestinal tem pouca influência aparente é recomendado. Além disso, duas possibilidades surgem para a continuidade dos trabalhos. A primeira é persistir na tentativa de obter modelos de ML que detectem certos fenótipos humanos de forma "globalizada". O AGP ainda poderia contribuir para essa visão com a filtragem menos severa do conjunto de dados de fenótipos, de forma que resulte na obtenção de mais instâncias positivas. A pesquisa e a integração com outros conjuntos de dados, a exemplo do *Human Microbiome Project*, também é outra forma de obter mais instâncias positivas, e aumentar a performance geral dos modelos. A segunda possibilidade é continuar a pesquisa de conjuntos de dados brasileiros, iniciando um movimento para treinar modelos regionais de cada área do Brasil.

De forma geral, realizar o *tuning* mais abrangente dos hiperparâmetros também permite obter um ganho de performance. Retreinar os modelos apenas com os atributos mais contribuintes dos modelos anteriores parece ser uma forma de obter melhores performances. Em fenótipos multiclasse pode ser realizado o treinamento de dois tipos de modelos: um que faça a predição da classe geral, como "mental_illness", e outro que faça a predição de uma das subclasses. Dessa forma, a classificação final obteria maior precisão no diagnóstico final. A investigação e execução de diferentes técnicas de seleção de atributos, como PCA e *autoencoder* também são estratégias promissoras.

Por fim, um aumento de performance permite uma análise mais detalhada sobre a importância dos atributos e espécies de microorganismos por taxonomistas.

REFERÊNCIAS

AI, L. et al. Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. **Oncotarget**, Impact Journals, LLC, v. 8, n. 6, p. 9546, 2017.

AITCHISON, J. The statistical analysis of compositional data. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 44, n. 2, p. 139–160, 1982.

ARYAL, S. et al. Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. **Hypertension**, Am Heart Assoc, v. 76, n. 5, p. 1555–1562, 2020.

BALAKRISHNAN, B.; TANEJA, V. Microbial modulation of the gut microbiome for treating autoimmune diseases. **Expert review of gastroenterology & hepatology**, Taylor & Francis, v. 12, n. 10, p. 985–996, 2018.

BOWERMAN, K. L. et al. Disease-associated gut microbiome and metabolome changes in patients with chronic obstructive pulmonary disease. **Nature communications**, Nature Publishing Group, v. 11, n. 1, p. 1–15, 2020.

CALLAHAN, B. J.; MCMURDIE, P. J.; HOLMES, S. P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. **The ISME journal**, Nature Publishing Group, v. 11, n. 12, p. 2639–2643, 2017.

CALLAHAN, B. J. et al. Dada2: high-resolution sample inference from illumina amplicon data. **Nature methods**, Nature Publishing Group, v. 13, n. 7, p. 581–583, 2016.

CAMMAROTA, G. et al. Gut microbiome, big data and machine learning to promote precision medicine for cancer. **Nature Reviews Gastroenterology & Hepatology**, Nature Publishing Group, p. 1–14, 2020.

CAO, Q. X. et al. Effects of microbiome rare taxa filtering on statistical analysis. 2021.

CASELLA, G. et al. Penalized regression, standard errors, and bayesian lassos. **Bayesian analysis**, International Society for Bayesian Analysis, v. 5, n. 2, p. 369–411, 2010.

CHANG, J. Y. et al. Decreased diversity of the fecal microbiome in recurrent clostridium difficile—associated diarrhea. **The Journal of infectious diseases**, The University of Chicago Press, v. 197, n. 3, p. 435–438, 2008.

CHATELIER, E. L. et al. Richness of human gut microbiome correlates with metabolic markers. **Nature**, Nature Publishing Group, v. 500, n. 7464, p. 541–546, 2013.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. **BMC genomics**, Springer, v. 21, n. 1, p. 6, 2020.

CRAMER, J. S. The origins of logistic regression. Tinbergen Institute Working Paper, 2002.

DADKHAH, E. et al. Gut microbiome identifies risk for colorectal polyps. **BMJ open gastroenterology**, BMJ Specialist Journals, v. 6, n. 1, p. e000297, 2019.

DAVE, M. et al. The human gut microbiome: current knowledge, challenges, and future directions. **Translational Research**, Elsevier, v. 160, n. 4, p. 246–257, 2012.

DERRIEN, M.; ALVAREZ, A.-S.; VOS, W. M. de. The gut microbiota in the first decade of life. **Trends in microbiology**, Elsevier, v. 27, n. 12, p. 997–1010, 2019.

DONG, M. et al. Predictive analysis methods for human microbiome data with application to parkinson's disease. **Plos one**, Public Library of Science San Francisco, CA USA, v. 15, n. 8, p. e0237779, 2020.

EDGAR, R. C.; FLYVBJERG, H. Error filtering, pair assembly and error correction for next-generation sequencing reads. **Bioinformatics**, Oxford University Press, v. 31, n. 21, p. 3476–3482, 2015.

FLACH, P. A. The geometry of roc space: understanding machine learning metrics through roc isometrics. In: **Proceedings of the 20th international conference on machine learning (ICML-03)**. [S.l.: s.n.], 2003. p. 194–201.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational statistics & data analysis**, Elsevier, v. 38, n. 4, p. 367–378, 2002.

GLOOR, G. B. et al. Microbiome datasets are compositional: and this is not optional. **Frontiers in microbiology**, Frontiers, v. 8, p. 2224, 2017.

GUYON, I. et al. A scaling law for the validation-set training-set size ratio. **AT&T Bell Laboratories**, Citeseer, v. 1, n. 11, 1997.

HALFVARSON, J. et al. Dynamics of the human gut microbiome in inflammatory bowel disease. **Nature microbiology**, Nature Publishing Group, v. 2, n. 5, p. 1–7, 2017.

HO, T. K. Random decision forests. In: IEEE. **Proceedings of 3rd international conference on document analysis and recognition**. [S.l.], 1995. v. 1, p. 278–282.

JIANG, H. et al. Altered fecal microbiota composition in patients with major depressive disorder. **Brain, behavior, and immunity**, Elsevier, v. 48, p. 186–194, 2015.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. In: **Machine Learning Proceedings 1994**. [S.l.]: Elsevier, 1994. p. 121–129.

KLITGORD, N.; SEGRÈ, D. Environments that induce synthetic microbial ecosystems. **PLoS Comput Biol**, Public Library of Science, v. 6, n. 11, p. e1001002, 2010.

KNIGHTS, D.; COSTELLO, E. K.; KNIGHT, R. Supervised classification of human microbiota. **FEMS microbiology reviews**, Blackwell Publishing Ltd Oxford, UK, v. 35, n. 2, p. 343–359, 2011.

KNIGHTS, D. et al. Bayesian community-wide culture-independent microbial source tracking. **Nature methods**, Nature Publishing Group, v. 8, n. 9, p. 761–763, 2011.

KOHAVI, R. et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: MONTREAL, CANADA. **Ijcai**. [S.l.], 1995. v. 14, n. 2, p. 1137–1145.

KOSTIC, A. D. et al. The dynamics of the human infant gut microbiome in development and in progression toward type 1 diabetes. **Cell host & microbe**, Elsevier, v. 17, n. 2, p. 260–273, 2015.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. **Emerging artificial intelligence applications in computer engineering**, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

KUHN, M. et al. Building predictive models in r using the caret package. **J Stat Softw**, v. 28, n. 5, p. 1–26, 2008.

KUNCHEVA, L. I.; RODRÍGUEZ, J. J. On feature selection protocols for very low-sample-size data. **Pattern Recognition**, Elsevier, v. 81, p. 660–673, 2018.

LALKHEN, A. G.; MCCLUSKEY, A. Clinical tests: sensitivity and specificity. **Continuing education in anaesthesia critical care & pain**, Oxford University Press, v. 8, n. 6, p. 221–223, 2008.

LAPIERRE, N. et al. Metapheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction. **Methods**, Elsevier, v. 166, p. 74–82, 2019.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. **R news**, v. 2, n. 3, p. 18–22, 2002.

LIPPERT, C.; HECKERMAN, D. Computational and statistical issues in personalized medicine. **XRDS: Crossroads, The ACM Magazine for Students**, ACM New York, NY, USA, v. 21, n. 4, p. 24–27, 2015.

LIU, H.; ZHOU, M.; LIU, Q. An embedded feature selection method for imbalanced data classification. **IEEE/CAA Journal of Automatica Sinica**, IEEE, v. 6, n. 3, p. 703–715, 2019.

LIU, X.-Y.; WU, J.; ZHOU, Z.-H. Exploratory undersampling for class-imbalance learning. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 39, n. 2, p. 539–550, 2008.

MANOR, O. et al. Health and disease markers correlate with gut microbiome composition across thousands of people. **Nature communications**, Nature Publishing Group, v. 11, n. 1, p. 1–12, 2020.

MARTÍN-FERNÁNDEZ, J.-A. et al. Bayesian-multiplicative treatment of count zeros in compositional data sets. **Statistical Modelling**, Sage Publications Sage India: New Delhi, India, v. 15, n. 2, p. 134–158, 2015.

MCDONALD, D. et al. American gut: an open platform for citizen science microbiome research. **Msystems**, Am Soc Microbiol, v. 3, n. 3, p. e00031–18, 2018.

MCMURDIE, P. J.; HOLMES, S. Phyloseq: a bioconductor package for handling and analysis of high-throughput phylogenetic sequence data. In: **Biocomputing 2012**. [S.l.]: World Scientific, 2012. p. 235–246.

MENEES, S.; CHEY, W. The gut microbiome and irritable bowel syndrome. **F1000Research**, Faculty of 1000 Ltd, v. 7, 2018.

MURALI, A.; BHARGAVA, A.; WRIGHT, E. S. Idtaxa: a novel approach for accurate taxonomic classification of microbiome sequences. **Microbiome**, BioMed Central, v. 6, n. 1, p. 1–14, 2018.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. In: **Proceedings of the 22nd international conference on Machine learning**. [S.l.: s.n.], 2005. p. 625–632.

OH, M.; ZHANG, L. Deepmicro: deep representation learning for disease prediction based on microbiome data. **Scientific reports**, Nature Publishing Group, v. 10, n. 1, p. 1–9, 2020.

OTTMAN, N. et al. Action and function of akkermansia muciniphila in microbiome ecology, health and disease. **Best Practice & Research Clinical Gastroenterology**, Elsevier, v. 31, n. 6, p. 637–642, 2017.

PASOLLI, E. et al. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 12, n. 7, p. e1004977, 2016.

PEIRCE, J. M.; ALVIÑA, K. The role of inflammation and the gut microbiome in depression and anxiety. **Journal of neuroscience research**, Wiley Online Library, v. 97, n. 10, p. 1223–1241, 2019.

PETERSON, J. et al. The nih human microbiome project. **Genome research**, Cold Spring Harbor Lab, v. 19, n. 12, p. 2317–2323, 2009.

PIETRUCCI, D. et al. Can gut microbiota be a good predictor for parkinson's disease? a machine learning approach. **Brain Sciences**, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 242, 2020.

PYLRO, V. S. et al. Brazilian microbiome project: revealing the unexplored microbial diversity—challenges and prospects. **Microbial ecology**, Springer, v. 67, n. 2, p. 237–241, 2014.

QI, Y. Random forest for bioinformatics. In: **Ensemble machine learning**. [S.l.]: Springer, 2012. p. 307–323.

QIN, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. **Nature**, Nature Publishing Group, v. 490, n. 7418, p. 55–60, 2012.

RAVEL, J. et al. Vaginal microbiome of reproductive-age women. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 108, n. Supplement 1, p. 4680–4687, 2011.

RIDGEWAY, G. Generalized boosted models: A guide to the gbm package. **Update**, v. 1, n. 1, p. 2007, 2007.

ROBINSON, C. M.; PFEIFFER, J. K. Viruses and the microbiota. **Annual review of virology**, Annual Reviews, v. 1, p. 55–69, 2014.

ROGERS, G. et al. From gut dysbiosis to altered brain function and mental illness: mechanisms and pathways. **Molecular psychiatry**, Nature Publishing Group, v. 21, n. 6, p. 738–748, 2016.

SAEYS, Y.; INZA, I.; LARRANAGA, P. A review of feature selection techniques in bioinformatics. **bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2507–2517, 2007.

SASAKI, K. et al. Construction of a model culture system of human colonic microbiota to detect decreased lachnospiraceae abundance and butyrogenesis in the feces of ulcerative colitis patients. **Biotechnology journal**, Wiley Online Library, v. 14, n. 5, p. 1800555, 2019.

SCHALLER, R. R. Moore's law: past, present and future. **IEEE spectrum**, IEEE, v. 34, n. 6, p. 52–59, 1997.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, Elsevier, v. 97, p. 105524, 2020.

SINHA, R. et al. The microbiome quality control project: baseline study design and future directions. **Genome biology**, Springer, v. 16, n. 1, p. 1–6, 2015.

SMIRNOVA, E.; HUZURBAZAR, S.; JAFARI, F. Perfect: Permutation filtering test for microbiome data. **Biostatistics**, Oxford University Press, v. 20, n. 4, p. 615–631, 2019.

TEAM, R. C. R language definition. **Vienna, Austria: R foundation for statistical computing**, 2000.

TOPÇUOĞLU, B. D. et al. A framework for effective application of machine learning to microbiome-based classification problems. **Mbio**, Am Soc Microbiol, v. 11, n. 3, 2020.

TUDDENHAM, S.; SEARS, C. L. The intestinal microbiome and health. **Current opinion in infectious diseases**, NIH Public Access, v. 28, n. 5, p. 464, 2015.

TURNBAUGH, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. **nature**, Nature Publishing Group, v. 444, n. 7122, p. 1027, 2006.

VACCA, M. et al. The controversial role of human gut lachnospiraceae. **Microorganisms**, Multidisciplinary Digital Publishing Institute, v. 8, n. 4, p. 573, 2020.

VARMA, S.; SIMON, R. Bias in error estimation when using cross-validation for model selection. **BMC bioinformatics**, Springer, v. 7, n. 1, p. 1–8, 2006.

VILO, C.; DONG, Q. Evaluation of the rdp classifier accuracy using 16s rrna gene variable regions. **Metagenomics**, Ashdin Publishing, v. 1, n. 235551, p. 104303, 2012.

VUONG, H. E.; HSIAO, E. Y. Emerging roles for the gut microbiome in autism spectrum disorder. **Biological psychiatry**, Elsevier, v. 81, n. 5, p. 411–423, 2017.

WANG, Q. et al. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. **Applied and environmental microbiology**, Am Soc Microbiol, v. 73, n. 16, p. 5261–5267, 2007.

WONG, C. B.; ODAMAKI, T.; XIAO, J.-z. Beneficial effects of bifidobacterium longum subsp. longum bb536 on human health: Modulation of gut microbiome as the principal action. **Journal of Functional Foods**, Elsevier, v. 54, p. 506–519, 2019.

WRIGHT, E. S.; YILMAZ, L. S.; NOGUERA, D. R. Decipher, a search-based approach to chimera identification for 16s rrna sequences. **Applied and environmental microbiology**, Am Soc Microbiol, v. 78, n. 3, p. 717–725, 2012.

WU, H. et al. Metagenomics biomarkers selected for prediction of three different diseases in chinese population. **BioMed research international**, Hindawi, v. 2018, 2018.

YANG, F.; ZOU, Q. maml: an automated machine learning pipeline with a microbiome repository for human disease classification. **Database**, Oxford Academic, v. 2020, 2020.

YILMAZ, P. et al. The silva and "all-species living tree project (ltp)" taxonomic frameworks. **Nucleic acids research**, Oxford University Press, v. 42, n. D1, p. D643–D648, 2014.

ZE, X. et al. Ruminococcus bromii is a keystone species for the degradation of resistant starch in the human colon. **The ISME journal**, Nature Publishing Group, v. 6, n. 8, p. 1535–1543, 2012.

ZELLER, G. et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. **Molecular systems biology**, v. 10, n. 11, p. 766, 2014.

ZHANG, Z. et al. Impact of fecal microbiota transplantation on obesity and metabolic syndrome—a systematic review. **Nutrients**, Multidisciplinary Digital Publishing Institute, v. 11, n. 10, p. 2291, 2019.

ZHOU, Y.-H.; GALLINS, P. A review and tutorial of machine learning methods for microbiome host trait prediction. **Frontiers in Genetics**, Frontiers, v. 10, p. 579, 2019.

APÊNDICE A — FENÓTIPOS PRÉ-SELECIONADOS

Cabeçalho	Descrição
Run	ERR, usado para sincronizar com SraAccList.txt
XML.File	Arquivo .xml que possui essa informação
acid_reflux	Doença do refluxo gastroesofágico (DRGE)
add_adhd	Distúrbios neurocomportamentais (Hiperatividade e TDAH)
age_years	Idade reportada
alcohol_frequency	Frequencia em que bebe alcool
antibiotic_history	Última vez que o participante ingeriu antibiótico
asd	Foi diagnosticado com autismo
autoimmune	Doença autoimune além de DII e diabetes I
bmi_cat	Categoria IMC
bmi_corrected	IMC calculado e com valores incorretos removidos
bowel_movement_quality	O participante costuma ter diarréia ou outras dificuldades
cancer	Foi diagnosticado com cancer
cancer_treatment	Tratamento de cancer
cardiovascular_disease	Foi diagnosticado com doença cardiovascular
cdiff	Clostridium difficile
country_residence	País de residência
csection	Nasceu de cesariana
diabetes	Diabetes
diabetes_type	Tipo de diabetes
diet_type	Categorização alto nível da dieta alimentar
epilepsy_or_seizure_disorder	Epilepsia ou convulsão
exercise_frequency	Qual a frequência de exercicios
fungal_overgrowth	Foi dianosticado com SIFO
gluten	É seguido uma dieta livre de gluten
height_cm	Altura
ibd	Foi dianosticado com doença inflamatória intestinal
ibd_diagnosis	Tipo de DII: Colite ou Crohn
ibd_diagnosis_refined	Subclassificação da DII
ibs	Diagnosticado com síndrome do intestino irritável
kidney_disease	Foi diagnosticado com doença renal
lactose	É intolerante a lactose
liver_disease	Foi dianosticado com doença hepática
lung_disease	Foi dianosticado com doença pulmonar
mental_illness	Possui doença mental
mental_illness_type_anorexia_nervosa	Anorexia
mental_illness_type_bipolar_disorder	Bipolaridade
mental_illness_type_bulimia_nervosa	Bulimia
mental_illness_type_depression	Depressão
mental_illness_type_ptsd_posttraumatic_stress_disorder	PTSD
mental_illness_type_schizophrenia	Esquizofrenia
mental_illness_type_substance_abuse	Abuso de substância
one_liter_of_water_a_day_frequency	Quantas vezes bebe ao menos 1 litro de água
pregnant	A participante está grávida
probiotic_frequency	Em qual frequência ele consome probióticos
race	Categorização de alto nível da etnia
sample_type	Origem da amostra
sex	Sexo
sibo	Supercrescimento bacteriano (SIBO)
smoking_frequency	Em qual frequência o participante fuma
subset_age	Participante entre 20 e 69 anos
subset_antibiotic_history	Participante usou antibiótico até 1 ano atrás
subset_bmi	BMI entre 18.5 e 30
subset_diabetes	Participante não possui diabetes
subset_healthy	Participante saudável (’and’ lógico dos outros subsets)
subset_ibd	Participante diagnosticado com IBD
weight_change	Peso mudou mais de 4.5 quilos no último ano
weight_kg	Peso

Exemplos de atributos fenotípicos obtidos de cada amostra do projeto *AGP*.

APÊNDICE B — ACURÁCIA NA VALIDAÇÃO CRUZADA

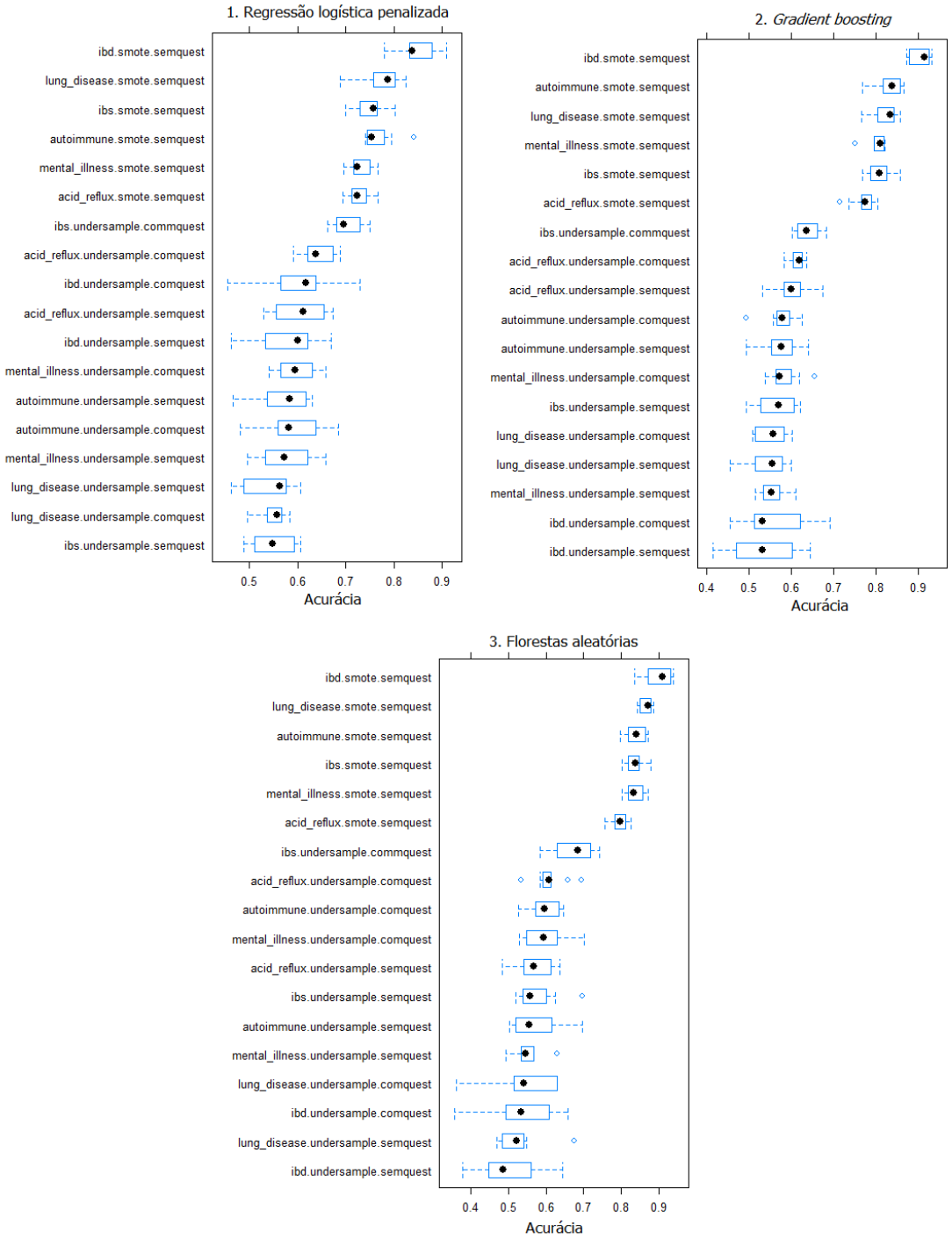


Diagrama de caixa dos valores de acurácia na validação cruzada de cada modelo. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, transtorno mental.

APÊNDICE C — PRECISÃO NA VALIDAÇÃO CRUZADA

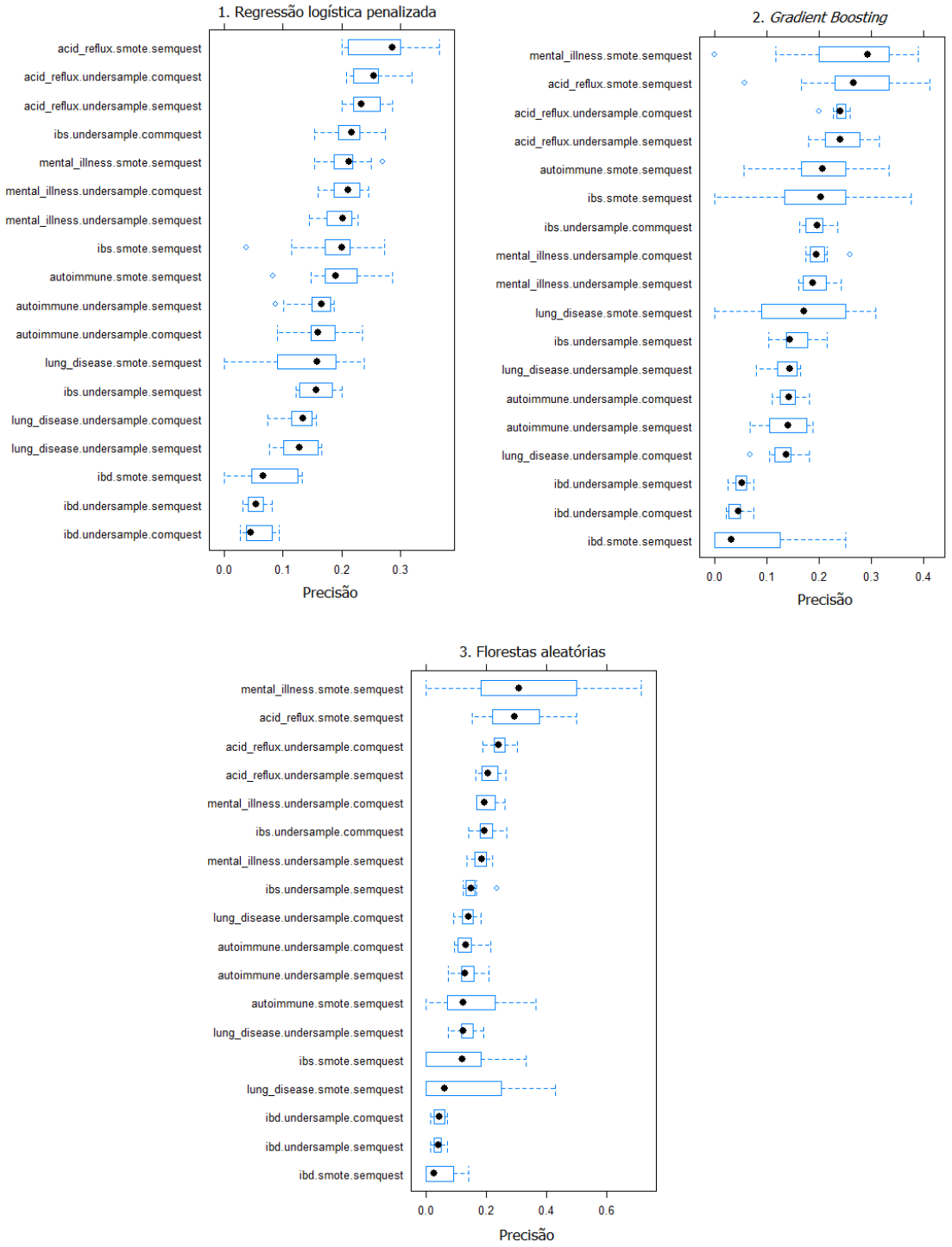


Diagrama de caixa dos valores de precisão na validação cruzada de cada modelo. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, transtorno mental.

APÊNDICE D — ESPECIFICIDADE NA VALIDAÇÃO CRUZADA

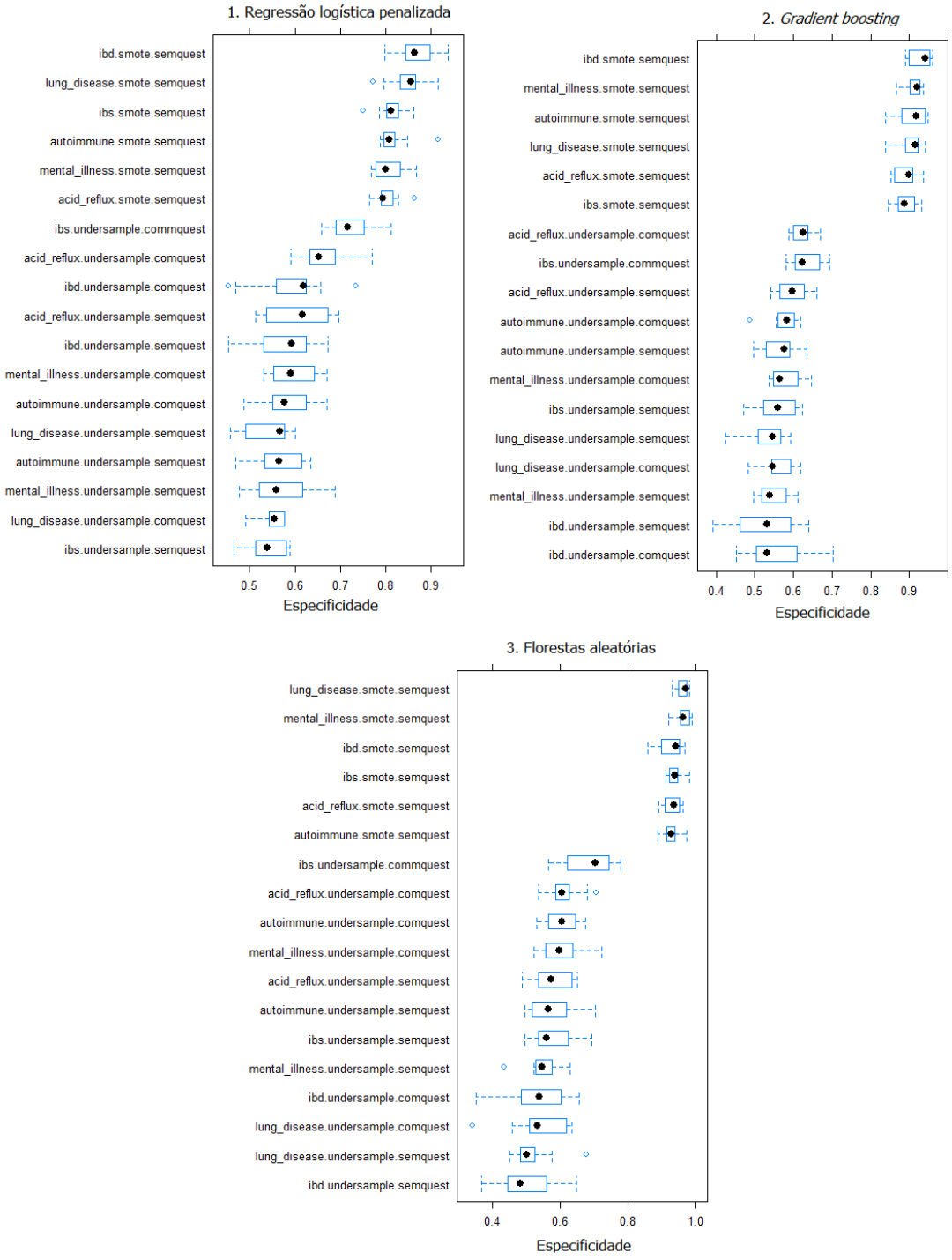
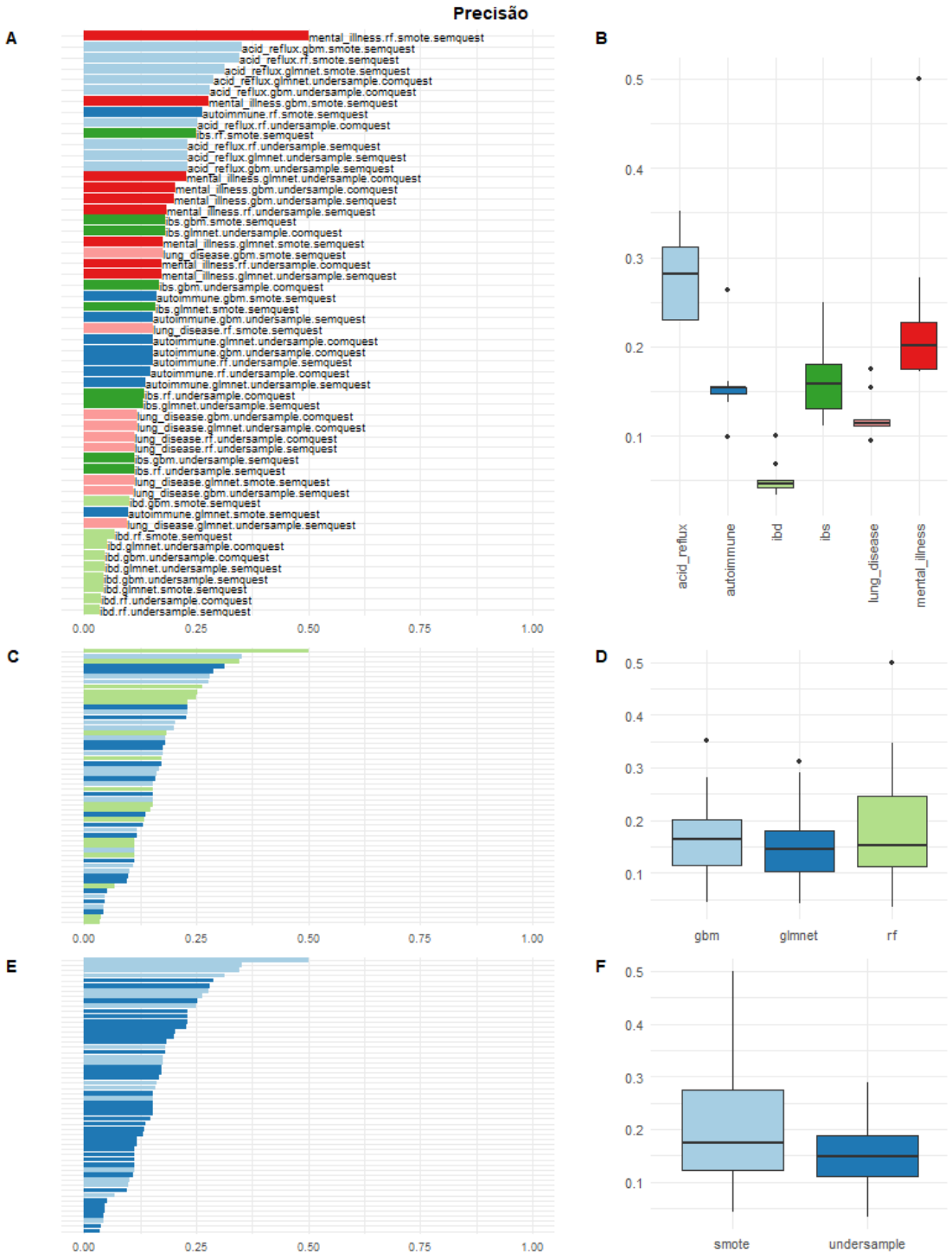


Diagrama de caixa dos valores de especificidade na validação cruzada de cada modelo. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, transtorno mental.

APÊNDICE E — ACURÁCIA NO CONJUNTO DE TESTE

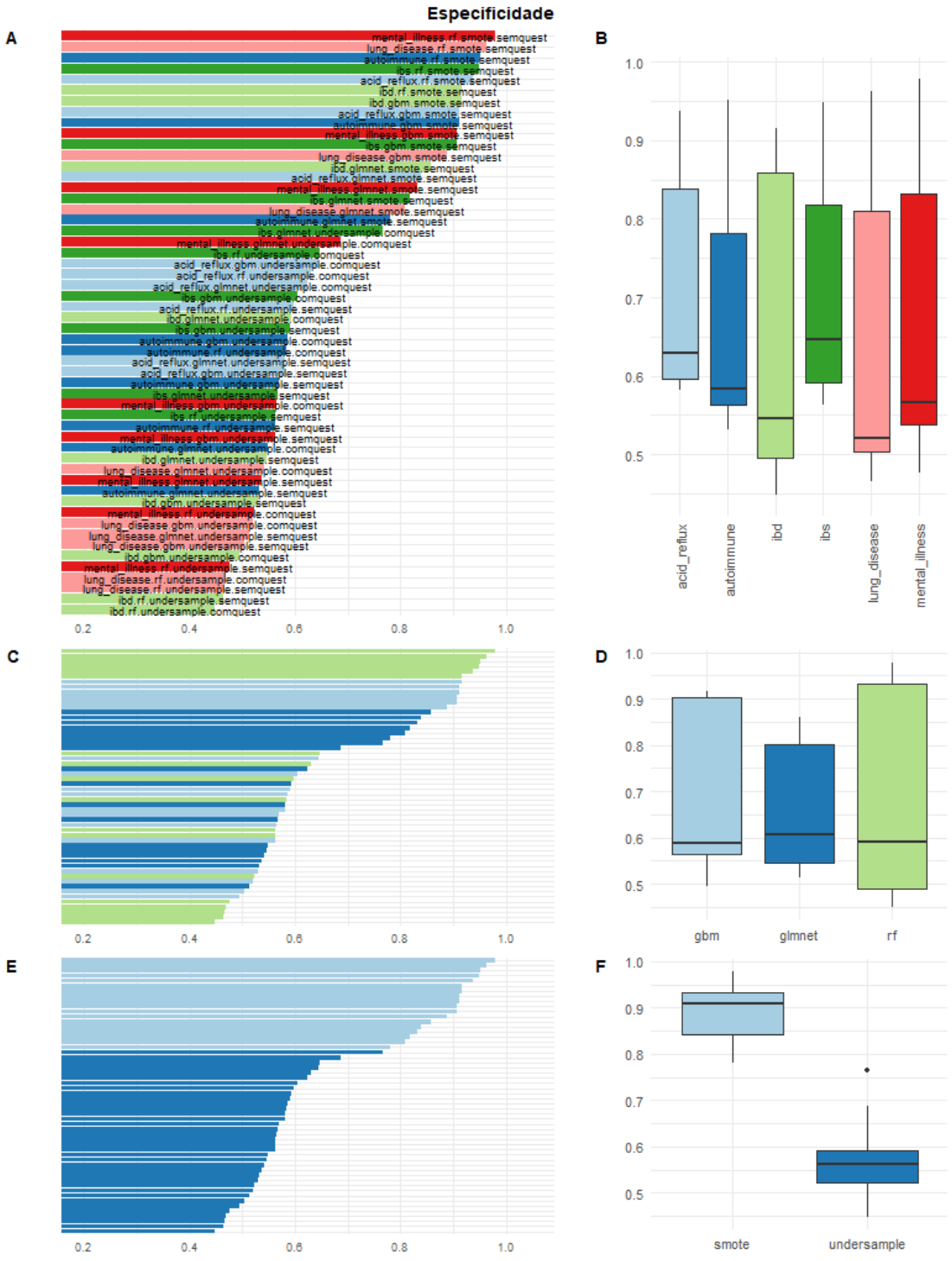
Métricas de acurácia para o conjunto de dados de teste de cada um dos modelos obtidos. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, doença mental.

APÊNDICE F — PRECISÃO NO CONJUNTO DE TESTE



Métricas de precisão para o conjunto de dados de teste de cada um dos modelos obtidos. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, doença mental.

APÊNDICE G — ESPECIFICIDADE NO CONJUNTO DE TESTE



Métricas de especificidade para o conjunto de dados de teste de cada um dos modelos obtidos. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, doença mental.

APÊNDICE H — F1-*SCORE* NO CONJUNTO DE TESTE

Métricas de F1-*score* para o conjunto de dados de teste de cada um dos modelos obtidos. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, doença mental.

APÊNDICE I — AUC NO CONJUNTO DE TESTE

Métricas de AUC para o conjunto de dados de teste de cada um dos modelos obtidos. Os fenótipos são “acid_reflux”, GRDE; “autoimmune”, doenças autoimunes (exceto DII e diabetes I); “ibd”, DII; “ibs”, síndrome do intestino irritável; “lung_disease”, doença pulmonar; e, “mental_illness”, doença mental.